Universidade de Santiago de Compostela

Facultade de Medicina e Odontoloxía

Departamento de Anatomía Patolóxica e Ciencias Forenses



Ancestry analysis in Forensic Genetics: New markers and methodology

Memoria que presenta para optar al grado de doctor,

Carla Isabel Arcanjo dos Santos

Santiago de Compostela, Julio 2015

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

La Doctora María Victoria Lareu Huidobro, Catedrática del Departamento de Anatomía Patológica y Ciencias Forenses de la Facultad de Medicina y Odontología de la Universidad de Santiago de Compostela,

CERTIFICA:

Que la presente memoria que lleva por título: **"Ancestry analysis in Forensic Genetics: New markers and methodology"**, de la licenciada en Biología por la Universidade de Aveiro (Portugal), *Carla Isabel Arcanjo dos Santos*, ha sido realizada bajo mi dirección, considerándola en condiciones para optar al Grado de Doctor y autorizándola para su presentación ante el Tribunal correspondiente.

Y para que así conste, firmamos la presente en Santiago de Compostela, a 29 de Julio de 2015.

Prof. Dra. María Victoria Lareu Huidobro

Carla Isabel Arcanjo dos Santos

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

This work was partially funded by research projects INCITE09208-163PR from *Xunta de Galicia* and BIO2013-42188-R from *Programa estatal de investigación, desarrollo e innovación orientada a los retos de la sociedad, Ministerio de Economía y Competitividad (España)*; as well as a PhD grant (SFRH/BD/75627/2010) awarded by the *Portuguese Foundation for Science and Technology (FCT)* and co-financed by the European Social Fund (Human Potential Thematic Operational Program).

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA

POPH QUALIFICAR É CRESCER. QREN QUADRO DE REFERÊNCIA ESTRATÉGICO NACIONAL PORTUGAL 2007-2013 GOVERNO DA REPÚBLICA PORTUGUESA UNIÃO EUROPEIA Fundo Social Europeu

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Agradecimentos

Sempre pensei que, chegado este momento, seria capaz de dar asas à imaginação e agradecer a todos aqueles que de alguma forma me ajudaram sem ter de escrever os seus nomes. Tudo isto porque sempre me senti muito grata pela vossa presença na minha vida – a nível profissional, claro, mas principalmente a nível pessoal. E espero que, ao longo do caminho que percorremos juntos, eu tenha sido capaz de expressar, de forma pessoal (com palavras, atitudes e sorrisos) a minha gratidão. Assim, enfrentar-me a esta folha em branco revelou-se uma tarefa muito mais difícil do que aquilo que eu esperava. A minha queda artística sempre esteve mais virada para as linhas e cores do que para a literatura. De todas as formas, espero que as palavras que se seguem sejam capazes de refletir, de forma sincera, aquilo que sinto.

As primeiras palavras são para agradecer à minha querida família, em especial à minha mãe, ao meu pai e ao meu maninho (nunca te esqueças que apesar de te tratar assim, és o mais velho). Obrigada por cuidarem de mim, por me fazerem sentir querida, por apoiarem todas as minhas decisões mesmo quando não sabiam bem qual poderia ser o resultado final. Vocês estão sempre no meu coração e é a vocês que dedico (e sempre dedicarei) aquilo que alcanço.

Depois, um muito obrigada a ti, Maviki. Porque desde o início me aceitaste, confiaste em mim e pela oportunidade de formar parte da grande equipa que diriges.

Um muito obrigada a ti também Chris, por tudo o que me ensinaste, por estares sempre disposto a ajudar e pela amizade.

A todos vocês – independentemente do número de anos que partilhámos juntos, do como e onde nos conhecemos, da distância que nos separa – obrigada pela companhia, apoio e amizade. Obrigada por me lembrarem qual é a minha origem. Obrigada por me fazerem sentir em casa mesmo quando estava longe.

Obrigada! Gracias! Thank you!



Emanuel Bessa Monteiro (2015)
Um agradecimento especial, o teu talento não deixa de me surpreender!

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

"Uma única vida é pouco. Para se fazer aquilo que se sabe, se pode, se quer e se deve fazer é preciso deixar muitas outras coisas para trás."
(José Luis Peixoto, in Abraço)

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Summary

Title: ***Ancestry analysis in Forensic Genetics: New markers and methodology***

The analysis of human population variation is an area of considerable interest in the forensic genetics field – ancestry informative assays can provide investigative leads from DNA evidences in cases where the source individual is not known or is unable to declare its ancestry. Short tandem repeat (STR) markers, commonly used in human identification applications, are not ideal for ancestry prediction as they are highly variable between individuals and have low inter-population variation. Binary loci with large allele frequency differences between populations have been proposed as an alternative to estimate biogeographic origin. Autosomal markers have emerged as amongst the best ancestry markers, providing valuable complementary data to Y-chromosome and mitochondrial DNA.

This thesis describes five complementary ancestry informative assays that, when combined, can differentiate up to six population groups: Africa, East Asia, Oceania, Native America and two Eurasian groups (Europe and Central South Asia). Nevertheless, the differentiation of the more closely related Middle East group is prone to higher classification errors – higher numbers of markers have to be analysed. The described assays provide ancestry informative genotypes in sensitive tests designed to work with limited DNA samples. Three analyse single nucleotide polymorphisms (SNPs) using SNaPshot® technology. The others analyse insertion deletion polymorphisms (Indels) and STRs using direct PCR-to-CE techniques that offer the possibility to detect mixed profiles.

The emergence of next generation sequencing (NGS) technologies allows the analysis of more markers in a single assay, improving the differentiation of closely related populations. Furthermore, the enlarged multiplex scales enhance population divergence balance, enabling the analysis of admixed individuals without biased estimation of co-ancestry proportions. However, although initial validation of a NGS platform is presented in this thesis, further evaluations have to be performed prior to the inclusion of such technology in routine casework.

Keywords: ancestry informative markers; single nucleotide polymorphisms; insertion deletion polymorphisms; forensic genetics; next generation sequencing.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Resumen

Titulo: ***Análisis de ancestralidad en Genética Forense: Nuevos marcadores e metodología***

El análisis de variabilidad de las poblaciones humanas es un área de gran interés para el campo de la genética forense – conjuntos de polimorfismos informativos de ancestralidad, a partir de evidencias de ADN, pueden proporcionar pistas en los casos en que no se conoce la persona de origen o ésta no puede declarar su ascendencia. Los microsatélites (STR), comúnmente usados en aplicaciones de identificación humana, no son ideales para la predicción de ancestralidad ya que son muy variables entre individuos y tienen baja variabilidad entre poblaciones. Como alternativa para estimar origen biogeográfico, se han propuesto polimorfismos binarios con grandes diferencias en las frecuencias alélicas entre poblaciones. En particular, los marcadores autosómicos están entre los mejores marcadores de ancestralidad, proporcionando datos complementarios al cromosoma-Y y ADN mitocondrial.

Esta tesis describe cinco reacciones multiplex complementarias que, al combinarse, pueden diferenciar hasta seis grupos ancestrales: África, Asia Oriental, Oceanía, América y dos grupos de Eurasia (Europa y Centro/Sur de Asia). Sin embargo, la diferenciación de Oriente Medio, el grupo más cercano, es propensa a errores de clasificación más altos – un mayor número de marcadores tienen que ser analizados. Los ensayos descritos proporcionan genotipos informativos de ancestralidad en pruebas sensibles diseñadas para trabajar con muestras de ADN limitadas. Tres analizan polimorfismos nucleotídicos simples (SNP) a través de la tecnología SNaPshot®. Las demás analizan polimorfismos de inserción/deleción (Indels) y STRs a través de amplificación seguida de electroforesis capilar, ofreciendo la posibilidad de detección de muestras mezcladas.

La aparición de tecnologías de secuenciación de nueva generación (NGS) permite el análisis de más marcadores en una única reacción, mejorando la diferenciación de poblaciones cercanas. Por otra parte, el aumento del tamaño de las reacciones multiplex mejora el equilibrio de divergencia poblacional, permitiendo el análisis de individuos mestizos sin sesgo de las proporciones de co-ancestralidad. Sin embargo, aunque en esta tesis se presenta la validación inicial de una plataforma de NGS, nuevas evaluaciones deben ser realizadas antes de la inclusión de este tipo de tecnología en el análisis de casos de rutina forense.

Palabras clave: marcadores informativos de ancestralidad; polimorfismos nucleotídicos simples; polimorfismos de inserción/deleción; genética forense; secuenciación de nueva generación.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Resumo

Titulo: ***Análise de ancestralidade en Xenética Forense: Novos marcadores e metodoloxía***

A análise da variabilidade das poboacións humanas é unha área de gran interese para o campo da xenética forense – conxuntos de polimorfismos informativos de ancestralidade, partindo de evidencias de ADN, poden proporcionar pistas nos casos nos que non se coñece a persoa de orixe ou esta non pode declarar a súa ascendencia. Os microsatélites (STR), comunmente usados en aplicacións de identificación humana, non son ideais para a predición de ancestralidade xa que son moi variables entre individuos e teñen baixa variabilidade entre poboacións. Como unha alternativa para estimar orixe bioxeográfica, propuxéronse polimorfismos binarios con grandes diferenzas nas frecuencias alélicas entre poboacións. En particular, os marcadores autosómicos están entre os mellores marcadores de ancestralidade, proporcionando datos complementarios ó cromosoma-Y e ADN mitocondrial.

Esta tese describe cinco reaccións multiplex complementarias que, ó combinárense, poden diferenciar ata seis grupos ancestrais: África, Asia Oriental, Oceanía, América e dous grupos de Eurasia (Europa e Centro/Sur de Asia). Porén, a diferenciación de Oriente Medio, o grupo máis achegado, é propensa a erros de clasificación máis altos – un maior número de marcadores deben ser analizados. Os ensaios descritos proporcionan xenotipos informativos de ancestralidade en probas sensibles deseñadas para traballar con mostras de ADN limitadas. Tres analizan polimorfismos nucleotídicos simples (SNP) a través da tecnoloxía SNaPshot®. As demais analizan polimorfismos de inserción/deleción (Indels) e STRs a través de amplificación seguida de electroforese capilar, ofrecendo a posibilidade de detección de mostras mesturadas.

A aparición de tecnoloxías de secuenciación de nova xeración (NGS) permite a análise de máis marcadores nunha única reacción, mellorando a diferenciación de poboacións próximas. Por outra parte, o aumento do tamaño das reaccións multiplex mellora o equilibrio de diverxencia poboacional, permitindo a análise de individuos mestizos sen nesgo das proporcións de co-ancestralidade. Non obstante, aínda que nesta tese se presenta a validación inicial dunha plataforma de NGS, novas avaliacións deben ser realizadas antes da inclusión deste tipo de tecnoloxía na análise de casos de rutina forense.

Palabras chave: marcadores informativos de ancestralidade; polimorfismos nucleotídicos simples; polimorfismos de inserción/deleción; xenética forense; secuenciación de nova xeración

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

ABBREVIATIONS **IX**

1 INTRODUCTION **3**

1.1 HUMAN POPULATION GENETICS 4

1.1.1 A summary of the demographic history of the modern Homo sapiens *4*

1.1.1.1 Evidences for understanding evolution 6

1.1.1.1.1 Morphological evidences 6
1.1.1.1.2 Archaeological evidences 6
1.1.1.1.3 Linguistic evidences 7
1.1.1.1.4 Genetic evidences 7

1.1.1.2 Models that explain the origin of modern humans 9

1.1.1.3 Overview of the out-of Africa event 10

1.1.1.4 The particular case of the Pacific region 12

1.1.1.4.1 Geographic division of the Pacific region 14
1.1.1.4.2 Cultural evidences in the Pacific region 15
1.1.1.4.3 The colonization of Remote Oceania: the Austronesian expansion 16
1.1.1.4.4 Genetic evidences in the Pacific region 17

1.1.2 Evolutionary forces *18*

1.1.2.1 Mutation 19

1.1.2.2 Genetic drift 20

1.1.2.3 Natural selection 21

1.1.2.4 Migration and gene flow 21

1.1.3 Human population structure and ancestry informative markers *22*

1.2 THE HUMAN GENOME AND DNA POLYMORPHISMS 26

1.2.1 Nucleic acids *26*

1.2.2 Organization of the genetic information *28*

1.2.3 The human genome *29*

1.2.3.1 Nuclear genome structure 30

1.2.3.2 Inheritance of the genome 32

1.2.3.2.1 Recombination in the nuclear genome 33
1.2.3.2.2 Non-recombining segments of the human genome 35

1.2.3.3 Human genome variation 36

1.2.4 Short polymorphisms in the human genome *37*

1.2.4.1 Short tandem repeats 38

1.2.4.2 Single nucleotide polymorphisms 40

1.2.4.3 Insertion/deletion polymorphisms 42

1.2.4.4 Online databases of human genetic variation 44

1.2.4.5 Typing methods 48

1.2.4.5.1 Amplifying the signal: the polymerase chain reaction 48
1.2.4.5.2 Targeted typing of short polymorphisms 50
1.2.4.5.3 Sequencing 54
1.2.4.5.4 Amplification artefacts of short polymorphisms 60

1.2.4.6 Advantages and disadvantages of each short polymorphism class 63

1.2.4.7 General applications of short polymorphisms 64

1.2.4.7.1 Anthropological applications 65

1.2.4.7.2 Clinical applications 65
1.2.4.7.3 Forensic applications 66
1.3 FORENSIC GENETICS STUDIES 69
1.3.1 A short history of forensic genetics *69*
1.3.1.1 Early days of forensic analysis of genetic variation 69
1.3.1.2 The PCR revolution 71
1.3.1.3 New markers and their application in forensic genetics 72
1.3.2 Application of population genetics to the forensic field *76*
1.3.2.1 Selection of ancestry informative markers 80
1.3.2.2 Data analysis systems 82
1.3.2.3 An example of AIM markers forensic application – the 11-M Madrid bomb attack investigation 86
2 OBJECTIVES **91**
2.1 MAIN OBJECTIVES 91
2.2 SPECIFIC OBJECTIVES 91
3 BLOCK 1 – DATA ANALYSIS SYSTEMS **97**
3.1 MATERIALS AND METHODS 97
3.1.1 Webpages of interest *97*
3.1.2 Collection of ancestry reference data with the SPSmart browser *98*
3.1.3 STRUCTURE software *101*
3.1.3.1 An overview of the *STRUCTURE* program 101
3.1.3.2 Ancestry and allele frequency models 103
3.1.3.3 Preparation of a *STRUCTURE* input file 107
3.1.3.4 How to run *STRUCTURE* software 109
3.1.3.5 Estimate the optimum number of populations 111
3.1.3.6 *STRUCTURE* associated software 113
3.1.3.6.1 CLUMPP 113
3.1.3.6.2 distruct 114
3.1.4 The Snipper web portal *115*
3.1.4.1 Background on *Snipper* analysis 115
3.1.4.2 Preparation of a *Snipper* input file 116
3.1.4.3 How to run *Snipper* 118
3.1.5 Principle component analysis (PCA) *120*
3.1.5.1 Background on PCA 120
3.1.5.2 Preparation of PCA input files 121
3.1.5.3 Creating a PCA plot 122
3.1.6 Multinomial logistic regression (MLR) to analyse admixed individuals *125*
3.1.6.1 Background on MLR 125
3.1.6.2 Preliminary comparison of MLR and Bayesian methods 126
3.2 RESULTS AND DISCUSSION 126

3.2.1 What information can be obtained from STRUCTURE? *126*
3.2.2 Evaluating Snipper output *128*
3.2.3 What information can be obtained from PCA? *128*
3.2.4 Preliminary comparison of MLR and Bayesian methods *129*

4 BLOCK 2 – NEW AIM ASSAYS **133**

4.1 MATERIAL AND METHODS 134
4.1.1 Revision of the SNPforID 34-plex forensic ancestry test (Appendix 2A) *134*
4.1.1.1 Population samples 134
4.1.1.2 *34-plex* component SNP reconfiguration 135
4.1.1.3 Amplification and single base extension primer modifications 137
4.1.1.4 Modified PCR and SBE reaction protocols 138
4.1.1.5 Population analysis 139
4.1.2 Pacifiplex: an AIM-SNP panel centred on Australia and the Pacific region *140*
4.1.2.1 Population samples 140
4.1.2.2 Marker selection and development of the multiplexed SNP assay 141
4.1.2.3 PCR amplification and genotyping 143
4.1.2.4 Statistical analysis 145
4.1.3 Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing (Appendix 2B) *147*
4.1.3.1 Population samples 147
4.1.3.2 AIM-Indel selection and development of the multiplex reaction 148
4.1.3.3 Amplification and genotyping 148
4.1.3.4 Statistical analysis 151
4.1.4 Completion of a world wide reference panel of samples for an ancestry informative Indel assay (Appendix 2C) *152*
4.1.4.1 Population samples 152
4.1.4.2 Amplification reactions 152
4.1.4.3 Statistical analyses 153
4.1.5 Eurasiaplex: a forensic SNP assay for differentiating European and South Asian ancestries (Appendix 2D) *154*
4.1.5.1 Population samples and SNP genotype databases accesses 154
4.1.5.2 SNP selection and multiplex construction 154
4.1.5.3 Assessment of population divergence and ancestry assignment performance of *Eurasiaplex* and *34-plex* SNPs 158
4.1.5.4 AIM-SNP analysis applied to ancestry inference in a challenging identification case 159
4.1.6 Development of a novel forensic STR multiplex for ancestry analysis and extended identity testing (Appendix 2E) *160*
4.1.6.1 Selection of ancestry informative tetranucleotide STRs 160

4.1.6.2 Multiplex design, amplification, and CE detection of the AIM-STRs 163
4.1.6.3 Sequence analysis of example alleles and construction of allelic ladder 164
4.1.6.4 Bayesian analysis of STR data for inference of ancestry 165
4.1.6.5 Expanding *pop.STR*: An open online database of forensic STR population data 167
4.1.6.6 Forensic statistics 167
4.1.6.7 Genotyping of forensic standard reference material controls 167
4.2 RESULTS AND DISCUSSION 168
4.2.1 Revision of the SNPforID 34-plex forensic ancestry test (Appendix 2A) 168
4.2.1.1 Performance of the modified *34-plex* SNaPshot assay: SBE product mobility variation and peak height ratios 168
4.2.1.2 Population analyses 171
4.2.1.3 Reference/control DNA profiles and AIM-SNP genotypes in seventy populations 177
4.2.1.4 Tri-allelic AIM-SNPs rs4540055 and rs5030240 178
4.2.1.5 Concluding remarks 180
4.2.2 Pacifiplex: an AIM-SNP panel centred on Australia and the Pacific region 182
4.2.2.1 Assay performance 182
4.2.2.2 Characterization of HGDP-CEPH reference populations with 62 AIMs 186
4.2.2.3 Study population analyses 191
4.2.2.4 Concluding remarks 197
4.2.3 Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing (Appendix 2B) 198
4.2.3.1 Genetic characterization of reference populations 199
4.2.3.2 Inferring genetic ancestry 202
4.2.3.3 Discussion 206
4.2.4 Completion of a worldwide reference panel of samples for an ancestry informative Indel assay (Appendix 2C) 210
4.2.4.1 Completing the HGDP-CEPH reference database 210
4.2.4.2 Characterization of third alleles in component Indel MID-548 213
4.2.4.3 Genetic ancestry analysis 215
4.2.4.4 *forInDel*: an open access forensic Indel frequency browser 223
4.2.4.5 Concluding remarks 223
4.2.5 Eurasiaplex: a forensic SNP assay for differentiating European and South Asian ancestries (Appendix 2D) 224
4.2.5.1 Characteristics of the *Eurasiaplex* assay and component SNPs 224
4.2.5.2 Population divergences assessed with principal component analysis 228
4.2.5.3 Bayesian classification analysis 229
4.2.5.4 AIM-SNP and morphometric analysis of unidentified cranium 231
4.2.5.5 Discussion 233
4.2.6 Development of a novel forensic STR multiplex for ancestry analysis and extended identity testing (Appendix 2E) 235

4.2.6.1 AIM-STR allele frequency distributions 235
4.2.6.2 Multiplex performance and analysis of standard control DNAs 236
4.2.6.3 Sequence analysis of STR repeat structures 238
4.2.6.4 Population analysis and inference of ancestry of HGDP-CEPH samples 239
4.2.6.5 Forensic informativeness of the *12-plex* assay 243
4.2.6.6 Discussion 243

5 BLOCK 3 – INTO NGS 249

5.1 MATERIAL AND METHODS 249

5.1.1 A SNaPshot of next generation sequencing for forensic SNP analysis (Appendix 3A) 249

5.1.1.1 Samples 249
5.1.1.2 Preparation of PCR amplicons 249
5.1.1.3 SNaPshot analysis of PCR amplicons 250
5.1.1.4 Sanger sequencing of PCR amplicons 251
5.1.1.5 NGS library preparation 251
5.1.1.6 Template preparation 253
5.1.1.7 Ion Torrent data analysis 253
5.1.1.8 Allele call thresholds 253
5.1.1.9 Statistical analysis 254

5.1.2 Inter-laboratory evaluation of SNP-based forensic identification by massively parallel sequencing using Ion PGM™ (Appendix 3B) 254

5.1.2.1 DNA samples, extraction of DNA and preparation of artificial mixtures 255
5.1.2.2 Ion PGM™ library and template preparation, enrichment and sequencing 256
5.1.2.3 Data analysis 257

5.1.3 Building a forensic ancestry panel from the ground up: the EUROFORGEN Global AIM-SNP set (Appendix 3F) 258

5.1.3.1 Sources of AIM-SNPs and allele frequencies in the five main global population groups 258
5.1.3.2 Selection criteria for the final AIM-SNP set 260
5.1.3.3 Marker subsets for smaller-scale multiplexed forensic tests 262
5.1.3.4 Tri-allelic SNPs included in the selection process 262
5.1.3.5 Population analyses 263

5.2 RESULTS AND DISCUSSION 264

5.2.1 A SNaPshot of next generation sequencing for forensic SNP analysis (Appendix 3A) 264

5.2.1.1 Amplicon coverage bias 264
5.2.1.1.1 Effect of GC content on coverage 266
5.2.1.1.2 Effect of amplicon length on coverage 267
5.2.1.1.3 Effect of PCR assay on coverage 268
5.2.1.2 Negative controls 269
5.2.1.3 Allele read frequencies 270
5.2.1.4 Genotype concordance 272
5.2.1.5 Sensitivity 273
5.2.1.6 Discussion 274

5.2.1.7 Conclusions 278

5.2.2 Inter-laboratory evaluation of SNP-based forensic identification by massively parallel sequencing using Ion PGM™ (Appendix 3B) 279

5.2.2.1 Sequence coverage from Ion PGM™ 279

5.2.2.2 Sequencing characteristics Ion PGM™ that impact forensic SNP genotyping 284

5.2.2.2.1 Base misincorporation rates 284
5.2.2.2.2 Allele read frequency balance 286
5.2.2.2.3 Strand bias 287
5.2.2.3 Genotype concordance 288

5.2.2.3.1 Inter-run and inter-lab concordance 289
5.2.2.3.2 Coriell cell-line control DNA concordance between Ion PGM™ genotypes and online data 289
5.2.2.4 Outlier SNPs: HID SNP markers showing discordances or requiring data scrutiny 290

5.2.2.4.1 Discordant SNPs 291
5.2.2.4.2 SNPs with no-calls 293
5.2.2.4.3 SNPs with mean analysis parameter values deviating from defined thresholds 294
5.2.2.5 Assessments of Ion PGM™ sensitivity 294

5.2.2.6 Mixture analysis 296

5.2.2.7 Context sequence examinations with IGV 297

5.2.2.8 Concluding remarks 298

5.2.3 Building a forensic ancestry panel from the ground up: the EUROFORGEN Global AIM-SNP set (Appendix 3F) 300

5.2.3.1 Characteristics of the ancestry informative SNPs selected 300

5.2.3.2 Balancing population-specific Divergence levels and subsets of the full Global AIM-SNP set 305

5.2.3.3 Analysis of admixed populations with the Global AIM-SNP set 307

5.2.3.4 Assessment of within-population group heterogeneity of selected AIM-SNPs 309

5.2.3.5 Discussion 311

6 FINAL DISCUSSION 317

7 CONCLUSIONS 327

7.1 ABOUT THE DATA ANALYSIS SYSTEMS 327

7.2 ABOUT THE NEW ANCESTRY INFORMATIVE ASSAYS 327

7.2.1 Revision of the SNPforID 34-plex forensic ancestry test 327

7.2.2 Pacifiplex: a 29 AIM-SNPs assay dedicated to the Pacific region 328

7.2.3 A 46-plex AIM-Indel assay and the completion of the HGDP-CEPH panel of samples 329

7.2.4 Eurasiaplex: a 23 AIM-SNP assay for Europe versus South Asia differentiation 330

7.2.5 A 12-plex AIM-STR assay 331

7.3 ABOUT ANCESTRY INFERENCE AND NEXT GENERATION SEQUENCING 332

7.3.1 Use of already existing AIM-SNP assays in NGS 332

7.3.2 Evaluation of the HID-Ion AmpliSeq™ Identity panel with the Ion PGM™ 332

7.3.3 The EUROFORGEN Global AIM-SNP set 334

8 RESUMEN 339

9 APPENDIXES **349**

9.1 SUPPLEMENTARY TO THE INTRODUCTION 351

9.1.1 Appendix I *351*

9.1.2 Appendix II *371*

9.1.3 Appendix III *387*

9.2 BLOCK 1 APPENDIXES 403

9.2.1 Appendix 1A *403*

9.2.2 Appendix 1B *417*

9.2.3 Appendix 1C *441*

9.3 BLOCK 2 APPENDIXES 443

9.3.1 Appendix 2A *443*

9.3.2 Appendix 2B *455*

9.3.3 Appendix 2C *465*

9.3.4 Appendix 2D *471*

9.3.5 Appendix 2E *479*

9.4 BLOCK 3 APPENDIXES 491

9.4.1 Appendix 3A *491*

9.4.2 Appendix 3B *503*

9.4.3 Appendix 3C *515*

9.4.4 Appendix 3D *521*

9.4.5 Appendix 3E *527*

9.4.6 Appendix 3F *535*

10 REFERENCES **551**

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Abbreviations

~ – approximately

A – adenine

A-chr – autosomic chromosome

aDNA – ancient DNA

AIM – ancestry informative marker

AMH – anatomically modern humans

bp – base pairs

C – cytosine

cc – cubic centimeter

CE – capillary electrophoresis

CEPH – Centre d'Étude du Polymorphisme Humain

chr – chromosome

cM – centiMorgan

dbSNP – NCBI short genetic variation database

ddNTP – dideoxinucleotide

DNA – deoxyribonucleic acid

dNTPs – deoxynucleotides

e.g. – *exempli gratia* (for example)

EVC – externally visible characteristics

G – guanine

HGDP – Human Genome Diversity Project

HWE – Hardy Weinberg equilibrium

i.e. – *id est* (that is)

IGV – Integrative Genomics viewer

In – informativeness-for-assignment metric

Indel – insertion/deletion polymorphism

ISP – Ion Sphere™ particles

Kb – kilobase pairs

km – kilometre

KY – thousand years

KYA – thousand years ago

LD – linkage disequilibrium

LR – likelihood ratio

LT – Life Technologies

m – meter

Mb – megabase pairs

MCMC – markov chain monte carlo

min – minute

miRNA – micro RNA

mL – mililiter

mM – milimolar

mRNA – messenger RNA

mtDNA – mitochondrial DNA

MYA – million years ago

N – nucleotide

NCBI – National Center for Biotechnology Information

ng – nanogram

NGS – next generation sequencing

NN – no-call

PC – principal component

PCA – principal component analysis

PCR – polymerase chain reaction

pg – picogram

PGM – Personal Genome Machine®

PNG – Papua New Guinea

RFUs – relative fluorescence units

RNA – ribonucleic acid

rs – clustered RefSNPs

SBE – single base extension

sec – second

SNP – single nucleotide polymorphisms

STR – short tandem repeat

T – thymine

TFS – Thermo Fisher Scientific

U – uracil

X-chr – X chromosome

Y-chr – Y chromosome

µg – microgram

µL – microliter

µM – micromolar

In the reference list:

1000GPC – The 1000 Genomes Project Consortium

EPC – The Encode Project Consortium

HPAC – Hugo Pan-Asian SNP Consortium

IHMC – The International HapMap Consortium

IHGSC – International Human Genome Sequencing Consortium

THDCRG – The Huntington's Disease Collaborative Research Group

REGWG – Race Ethnicity and Genetics Working Group

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

INTRODUCTION



USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

1 Introduction

This thesis is a summary of more than four years of work centred on the application of genetics, especially of population genetics, to the forensic field. Defining a population is not simple. In fact, populations are usually described based on the presence of shared characteristics, which may be social, cultural, physical or even geographical, but that not always reflect possible genetic relationships between individuals. For this reason, we will first explore the demographic history of the modern *Homo sapiens*, as it is a key element explaining the current human genetic diversity (section 1.1.1). Special attention will be given to the diversity in the Pacific region (section 1.1.1.4), the specific study object in the central work described in the present thesis. Such diversity was shaped by the action of several evolutionary forces during thousands of years (section 1.1.2) and it can be detected using polymorphic positions present in the human genome (section 1.1.3). Particularly, we will focus in small-scale polymorphisms as they have characteristics that are beneficial to their use, among others, in forensic genetics applications (section 1.2.4). The last section of the introduction gives an overview of the core theme of this thesis: forensic genetics studies (section 1.3).

As a supplement to this introduction, three appendixes reviewing knowledge about autosomal ancestry informative STRs and binary markers in forensic analysis are presented. They include: i) a book chapter (Appendix I) that reviews new markers in forensic genetics, including ancestry informative markers; ii) a review article (Appendix II) about the continuous improvements in the research and development of forensic STR genotyping; and iii) a book chapter (Appendix III) that reviews some of the currently available forensically viable AIM panels, their optimized PCR multiplexes and the population differentiation power they offer (those assays will be described in detail in section 4 / block 2).

1.1 HUMAN POPULATION GENETICS

1.1.1 A summary of the demographic history of the modern *Homo sapiens*

Humans and chimpanzees are the most closely related ape species but their lineages diverted from each other approximately 6-8 million years ago (MYA) (Goodman et al. 1998, Chen and Li 2001, Marques-Bonet et al. 2009, Langergraber et al. 2012). Several intermediate hominid species existed (Wood and Richmond 2000): *Sahelanthropus tchadensis* (Brunet et al. 2002) is thought to be one of the first hominids and is dated very closely to the divergence between humans and chimpanzees; the later *Australopithecus* genus is probably one of the most known, particularly *Lucy* (A.L.288-1), an *Au. afarensis* representative (Kimbel and Delezene 2009).

As it will be described in section 1.1.1.1, morphological, archaeological, linguistic and genetic evidences all point out for an African origin of modern humans. Nevertheless, some of the intermediate species that existed before were able to migrate out of Africa. The first hominid specie to be found out of Africa was *Homo erectus*. It spread to western and East Asia and then to Indonesia (Stringer 2012). Its expansion was probably possible due to their bigger body size (including a brain with approximately 900 cc), allowing for better tolerance to heat and dehydration, and the use of stone tools. Such characteristics allowed them to survive in a variety of climates, helping the expansion outside of the African continent (Wood and Collard 1999).

It is thought that two *Homo* species descended from *H. erectus* (**Figure 1**): *Homo heidelbergensis* and *Homo floresiensis* (Wood and Richmond 2000, Brown et al. 2004, Stringer 2012, Jobling et al. 2014). The first, *H. heidelbergensis* (Wagner et al. 2010), was a less robust hominid that had a larger brain (approximately 1200 cc). *H. floresiensis* was characterized by a small body size (approximately 1 m tall, with 380 cc brain) usually explained by selective pressure due to environmental conditions – they lived in tropical rainforests in isolated islands: resources were scarce and thermoregulation would be easier (Brown et al. 2004). *Homo sapiens* is thought to have evolved from *H. heidelbergensis* (**Figure 1**). The same is true for *Homo neanderthalensis* (King 1864) and the Denisova hominid (Krause et al. 2010, Reich et al. 2010).

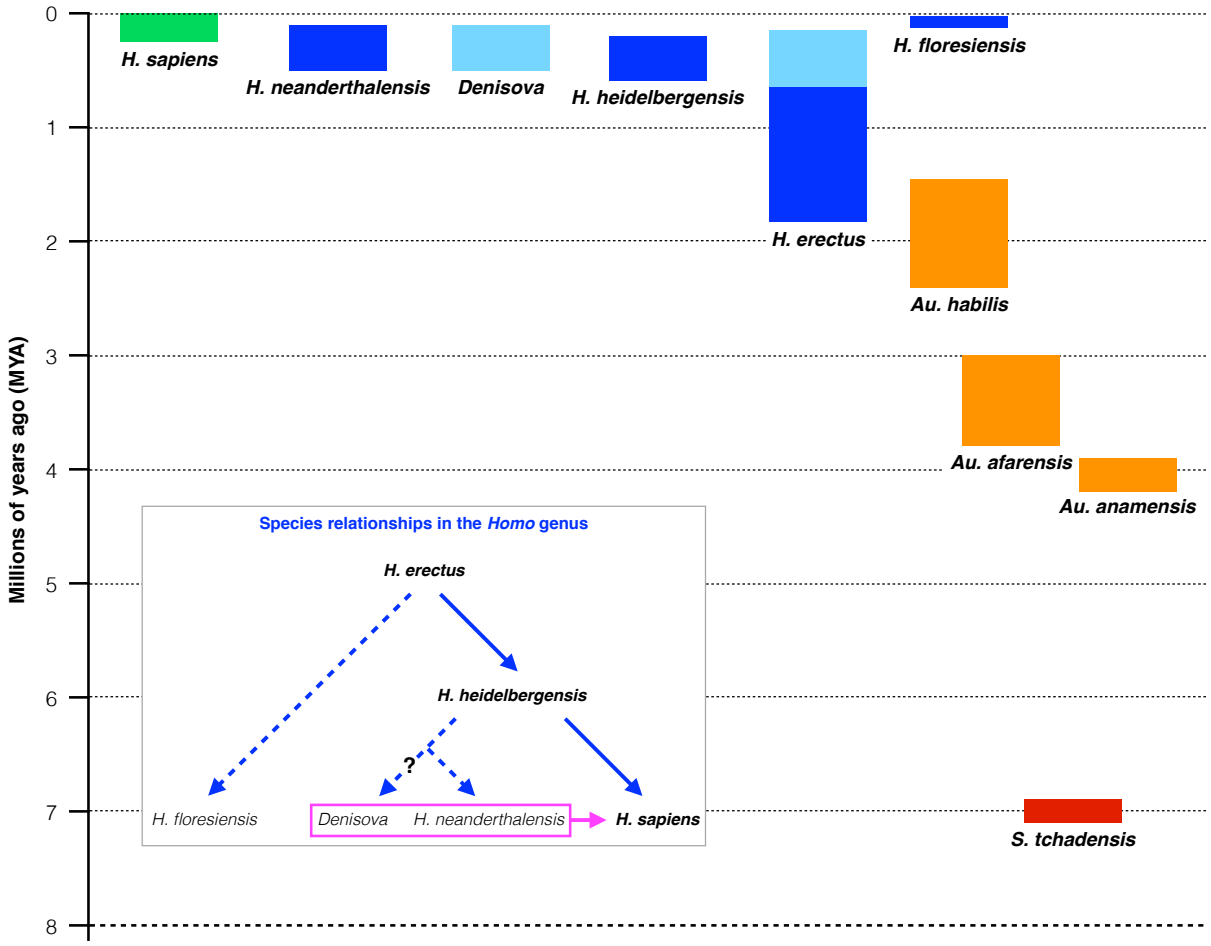


Figure 1: Hominid taxa and their probable time span (represented by the height of each bar). Note that, as the fossil record is not complete, the exact dates are difficult to infer. This is particularly true for the later date of *H. erectus* and the totality of Denisova time span (light blue). Anatomically modern humans belonging to *H. sapiens* are represented in green. Other *Homo* species are represented in blue. Some of the known *Australopithecus* species are in orange. And *S. tchadensis*, probably the first hominids after the divergence between humans and chimpanzees, is in red. The inset box shows the inferred relationships between *Homo* species, with solid arrows indicating the *H. sapiens* ancestral line. The origin of Denisova is not clear (indicated by a question mark). Archaic contribution from Denisova and *H. neanderthalensis* (pink arrow) to modern humans is considered. Adapted from Wood and Baker (2011), Stringer (2012) and Jobling et al. (2014).

It is important to keep in mind that such classifications are not universally accepted. For that reason, intermediate species between *H. erectus* and *H. sapiens* are usually designated *archaic H. sapiens*.

1.1.1.1 Evidences for understanding evolution

Several evidences have been collected to explain the evolution of modern *H. sapiens*. For example, fossils can potentially give information about environment, morphology and, to a certain extent, behaviour. Other types of evidences include archaeological and linguistic studies. Genetic evidences, both from modern populations and from ancient DNA (aDNA), provide information about the history of current lineages but also from those of the individuals that lived less than 100 thousand years ago (KYA).

1.1.1.1.1 Morphological evidences

It is not easy to differentiate *archaic Homo* individuals from anatomically modern humans (AMH, those that are unequivocally assigned to *H. sapiens*). When a complete fossil specimen is available, cranial features such as globular shape and the degree of facial retraction are generally used to achieve such differentiation (Lieberman et al. 2002). The first cranium described as from a fully modern human dates approximately 195 KYA (McDougall et al. 2005) and it was found in Ethiopia. Outside Africa, modern human fossils date less than 45 KYA (Barker et al. 2007), with fossils with approximately 40 KYA being found in Europe, East Asia and Australia. Furthermore, sub-Saharan populations are, at a morphological level, the most variable and such variation decreases with distance to this geographic region. In fact, distance to Africa explains 19-25% of the morphological variation (Manica et al. 2007).

Other evidences such as hand morphology (Susman 1998), neural development (Orban and Caruana 2014) and bipedalism (Ruff 2009, Wood and Baker 2011) can give information about human behaviour and evolution.

1.1.1.1.2 Archaeological evidences

Archaeological evidences complement fossil morphological analysis. In fact, such evidences can give information about hominid activity through the analysis of artefacts and physical remains.

It is known that non-human apes use several types of tools (Whiten et al. 1999, van Schaik et al. 2003, Pruetz and Bertolani 2007) and for that reason it is possible to infer that

our common ancestor also used tools. Nevertheless, there are no indications of tool usage by the first hominids. The first evidences, found in Ethiopia, were dated with approximately 2.6 MYA (Semaw et al. 2003). However, the authors were cautious and did not associate a particular hominid with their findings. Still, it was later recognized that the toolmakers had a good level of technological understanding and sophistication as they performed a quality based selection of the raw material used to prepare those tools (Stout et al. 2005). More recently, the date for tool use by hominids was extended – evidences of meat consumption assisted by stone tools were found, also in Ethiopia, and dated as older than 3.39 MYA (McPherron et al. 2010). In this case, they were attributed to *Au. aferensis*.

More advanced tools started to be used approximately 1.76 MYA and were associated with *H. erectus* (Lepre et al. 2011). With time, innovations in style and materials used to create tools occurred, as exemplified by a set of wood spears from 400 KYA (Thieme 1997). The last stage of the Paleolithic period, called Upper Paleolithic in Europe and Asia and roughly equivalent to the Later Stone Age in Africa, is defined by a variety of cultural artefacts and is mainly associated with modern humans – a particular culture from this period is associated with Neanderthals (Jobling et al. 2014).

1.1.1.1.3 Linguistic evidences

Although being more difficult to date, linguistic evidences can be informative about modern humans origin and their peopling of the world (Gray et al. 2011). A study of several languages showed that phonetic diversity is higher in Africa and that it decreases with distance to central/southern Africa. In fact, distance to Africa explains 19% of the phonetic diversity (Atkinson 2011), which is in accordance with morphological patterns described in section 1.1.1.1.1.

1.1.1.1.4 Genetic evidences

Genetic variability can be measured by several methods, from direct counting of the number of variants or nucleotide diversity to indirect statistics such as effective population size or linkage disequilibrium extent.

Studies involving the scrutiny of complete genome sequences, such as the 1000 Genomes Project (described in section 1.2.4.4), are ideal to evaluate genetic diversity. In that project, the number of single nucleotide polymorphisms (SNPs) and insertion/deletion polymorphisms (Indels) found in each of the three population groups considered and per individual was higher in Africa, intermediate in Europe and lower in Asia (1000GPC 2010), a pattern also observed for nucleotide diversity (Hernandez et al. 2011). Nevertheless, it is important to consider that the number of populations and the geographic coverage in such studies was reduced. For this reason, and despite the small number of samples included in some of the populations groups, the use of the HGDP-CEPH panel of samples (described in section 1.2.4.4) increases the geographic coverage. A study analysing microsatellite variability in the HGDP-CEPH panel demonstrated that African sub-Saharan populations have the highest heterozygosity values (Rosenberg et al. 2002). These conclusions were confirmed, in the same population samples, through a large-scale SNP genotyping study (Li et al. 2008). In fact, genetic diversity reduces in a linear fashion with the increase of the distance to Africa (Prugnolle et al. 2005, Ramachandran et al. 2005), a pattern also observed through haplotype diversity analysis, with the minimum value detected in South America (Conrad et al. 2006). On the other hand, genetic differentiation between populations increases linearly with geographic distance (Relethford 2004, Serre and Paabo 2004, Manica et al. 2005, Ramachandran et al. 2005).

Locus phylogeny also gives information about the time and place of its origin. For this to be possible, it is necessary to assume that a phylogeny can be reliably reconstructed and that geographic movement was limited in a way that the current distribution gives information about the ancestral distribution (Jobling et al. 2014). Uniparental phylogenies (mitochondrial DNA and Y-chromosome) (Ingman et al. 2000, Thomson et al. 2000) and phylogenies from several autosomal and X-chromosomal loci (Takahata et al. 2001) point to an African origin for most loci, and that most of our ancestors lived in Africa before 60 KYA. Unusual phylogenies with highly divergent non-African haplotypes can be found. In some cases it is possible that they are derived from archaic hominids (Abi-Rached et al. 2011) – they could represent selective advantage to humans migrating out of Africa as those alleles were already adapted to the Eurasian environment. In other words, such unusual phylogenies could represent an ancient non-African contribution from archaic humans to the modern human genetic pool.

1.1.1.2 Models that explain the origin of modern humans

As discussed in the previous section, *H. erectus* first migrated out of Africa (occupying great part of Europe and Asia) followed by the later appearance and expansion of the modern *H. sapiens*. There are several possible models to describe the latter events, the extreme ones being:

- *Multiregional model* – it considers that: i) the human species arose approximately 2 MYA; subsequent evolution occurred within a single specie that included all archaic humans (such as *H. erectus* and *H. neanderthalensis*) but also the modern forms; and iii) this continuous specie evolved worldwide to the diverse populations of modern *H. sapiens sapiens* found today (Wolpoff et al. 1984, Wolpoff et al. 1988, Wolpoff et al. 2000). This model should not be confused with separate or multiple origins for different populations at the same time (parallel evolution) – the multiregional model considers that, after *H. sapiens* left Africa, there were significant genetic exchanges with hominid species found during expansion across the globe (Wolpoff et al. 2000). Such gene flow accounts for the shared pattern of evolution across the human species and means that multiple independent origins in different regions are not possible (Wolpoff et al. 2000);
- *Out-of-Africa model* – it considers that: i) the transition from *H. erectus* to *H. sapiens* occurred in Africa; and ii) approximately 60 KYA these first modern humans migrated out of Africa and replaced the hominids present in other continents with no significant gene flow (Stringer and McKie 1996, Wood and Baker 2011). This model will be presented in further detail in section 1.1.1.3.

As shown in the previous section (1.1.1.1), morphological, archaeological, linguistic and genetic evidences favour the out-of-Africa model. This is especially true when considering genetic evidences: the out-of-Africa model predicts that diversity is higher in Africa and that the root of most of the genetic lineages is in Africa. If any other model was considered the current geographic distribution of genetic variability would be different.

Generally speaking, demographic models support an African sub-Saharan origin of the modern *H. sapiens* followed by a stepwise expansion to the other continents. This expansion

would involve several bottleneck events and is usually described by serial founder models (**Figure 2**). In such models, a single starting population is considered. A new population is formed by a small subset of individuals (termed founders) that migrate and expand. After population growth, a subset of individuals starts a new population, and so on. Such models imply that genetic diversity decreases at the same time that linkage disequilibrium and population differentiation increase with distance to the origin point, in this case, Africa (Prugnolle et al. 2005, Ramachandran et al. 2005).

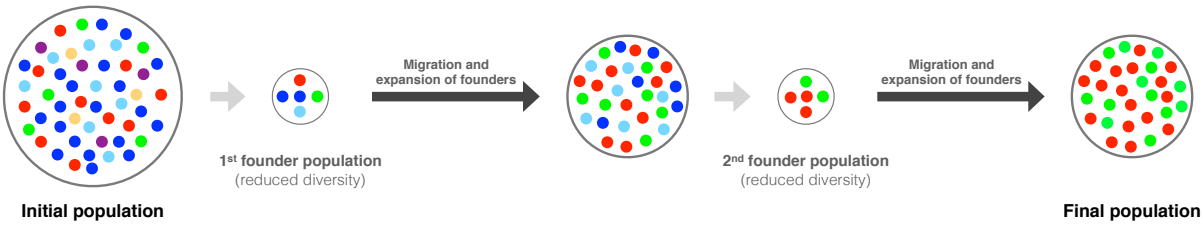


Figure 2: Graphical representation of the serial founder model. Coloured points represent individuals; colours represent population variability.

1.1.1.3 Overview of the out-of Africa event

Genetic evidences support the hypothesis that modern humans first migrated out of Africa approximately 90-120 KYA (Jobling et al. 2014). Furthermore, evidences such as the AMH fossils from 100-130 KYA found in Israel (Grün et al. 2005) and stone tools made by AMH found in Jebel Faya dated from approximately 125 KYA (Armitage et al. 2011), indicate human presence in the Levant region and in the Arabian Peninsula during the last interglacial period. The migration through the Bab al- Mandab strait to Jebel Faya might have successfully continued to Southeast Asia as shown by fossil evidence found in Southern China and dated as older than 100 KYA (Liu et al. 2010b, Beyin 2011). Nevertheless, as described by Dennell and Petraglia (2012) "*We need to recognise that not all dispersal events are successful, and the first appearance of a species such as H. sapiens in a region does not necessarily indicate permanent colonisation*".

Around 50-70 KYA, fully modern humans were able to migrate out of Africa to Asia following a southern coastal route, i.e, crossing the Bab al-Mandab strait (Oppenheimer 2012, Balme 2013, Jobling et al. 2014). The first evidences of modern humans in western Asia date approximately 47-49 KYA and, a few thousand years later, they can be found in southern Europe (~ 46 KYA) followed by a rapid expansion that reached as further west as Spain by 41

KYA (Mellars 2006, Mellars 2011). This period can be interpreted as a transition from Neanderthals to modern humans (Jobling et al. 2014). In fact, evidences indicate that during this expansion event, modern humans admixed with Neanderthals, probably in the Middle East region (**Figure 3A**) around 47-65 KYA (Sankararaman et al. 2012). Studies such as the ones from Green et al. (2010) and Sankararaman et al. (2012) showed that Neanderthals are closer to non-African than to African populations, with approximately 1.3% of the genome of all non-Africans being of Neanderthal origin. These migrant populations were subject to new selective pressures. Bottlenecks followed by rapid expansions originated founder effects, reducing diversity and increasing differentiation of new populations with the increase of the distance to Africa. Such events are inevitable during migrations and they shaped modern genetic diversity.

Further east, the first evidences of modern humans in East Asia were found in Tianyuan Cave, China. A set of 34 bones from a single individual were found and dated as approximately 39-42 KYA (Shang et al. 2007). Genetic analyses showed that modern humans present in that region had already diverged from the ancestors of present Europeans and that they were related to the ancestors of present Asians and Native Americans (Fu et al. 2013). Furthermore, this individual, as present Eurasian populations, had a genetic component related to Neanderthals but no Denisova component was detected (Fu et al. 2013). In Siberia, the majority of evidences found south of 55º N latitude point for a colonization by modern humans around the same time as the Tianyuan Cave (Goebel 1999). Further north, and despite the colder environments, modern human colonization is estimated to have occurred approximately 32 KYA (Goebel et al. 2008).

Siberia is also the origin point of the dispersal in direction to the Bering land bridge followed by the migration into the Americas around 18 KYA (Goebel et al. 2008). In fact, the earliest archaeological sites in Beringia date approximately 13-14 KYA (Goebel et al. 2008). Early fossil evidences in the American continent are dated as approximately 12.9 KYA in Idaho (Green et al. 1998) and from 12.7-16 KYA in Brazil (Feathers et al. 2010). Two migration routes have been proposed: through a coastal ice-free corridor that emerged approximately 15 KYA from Siberia to Beringia and finally North America; or through a land route that crossed the Beringia land bridge and then followed a ice-free corridor in North America (Jobling et al. 2014). Considering that around 18-21 KYA there was a full glacial period, the land route would be inhospitable and so an obstacle to human migrations.

Nevertheless, the corridor was present until 24 KYA and again after 14 KYA and it could represent an option (Goebel et al. 2008). It is still not clear which route was followed, or even if both were used but 13 KYA remains in California suggest that watercraft was used by early migrants (Goebel et al. 2008). Genetic evidences from fossils found in Alaska (dated from approximately 11 KYA) also support an early coastal dispersal. Its Y-chromosome was of haplogroup Q1a3-M3, the same as the majority of present day Native Americans, but its mitochondrial DNA was of a rare lineage (D4h3) that can be found, at low frequencies, in the Pacific coast from Mexico to Chile (Perego et al. 2009).

As mentioned in the beginning of this introduction, the Pacific region is of particular interest for this thesis. For this reason, a thorough overview of the colonization process by modern humans is presented in the next section.

1.1.1.4 The particular case of the Pacific region

Due to adverse climatic conditions, sea level has been lower than at present for most of the last 100 KYA, reaching -120 m during the Last Glacial Maximum around 20 KYA (Jobling et al. 2014). Only in the last 12 KYA (especially in the last 6-8 KYA) climatic conditions, particularly temperature and consequently sea level, changed and reached values similar to the ones observed in present times (Jobling et al. 2014). Due to the lower sea levels, the world map was considerably different than at present, with land bridges connecting islands in bigger landmasses. In the particular case of the Pacific region, such landmasses were only submerged around 8 KYA so, at the time modern humans expanded out of Africa, Australia was part of a larger continent that included New Guinea and Tasmania, known as Sahul, that was separated from Asia (Sunda) by more than 90 km of water (Jobling et al. 2014). The area between Sunda and Sahul limited by the Wallace and Lydekker's lines, respectively, includes several islands and it is known as Wallacea. The biological significance of such barrier can be acknowledged by the maintenance of very distinct faunas: typically Eurasian ecology and placental mammals in the Sunda side and distinct ecology and marsupial mammals in Sahul (Cooper and Stringer 2013, Jobling et al. 2014).

Wallacea also represented a barrier to hominid migrations. Evidences found in Lake Mungo (New South Wales) suggest that modern humans reached Australia approximately 50 KYA (Bowler et al. 2003, Rasmussen et al. 2011, Jobling et al. 2014) descending from the

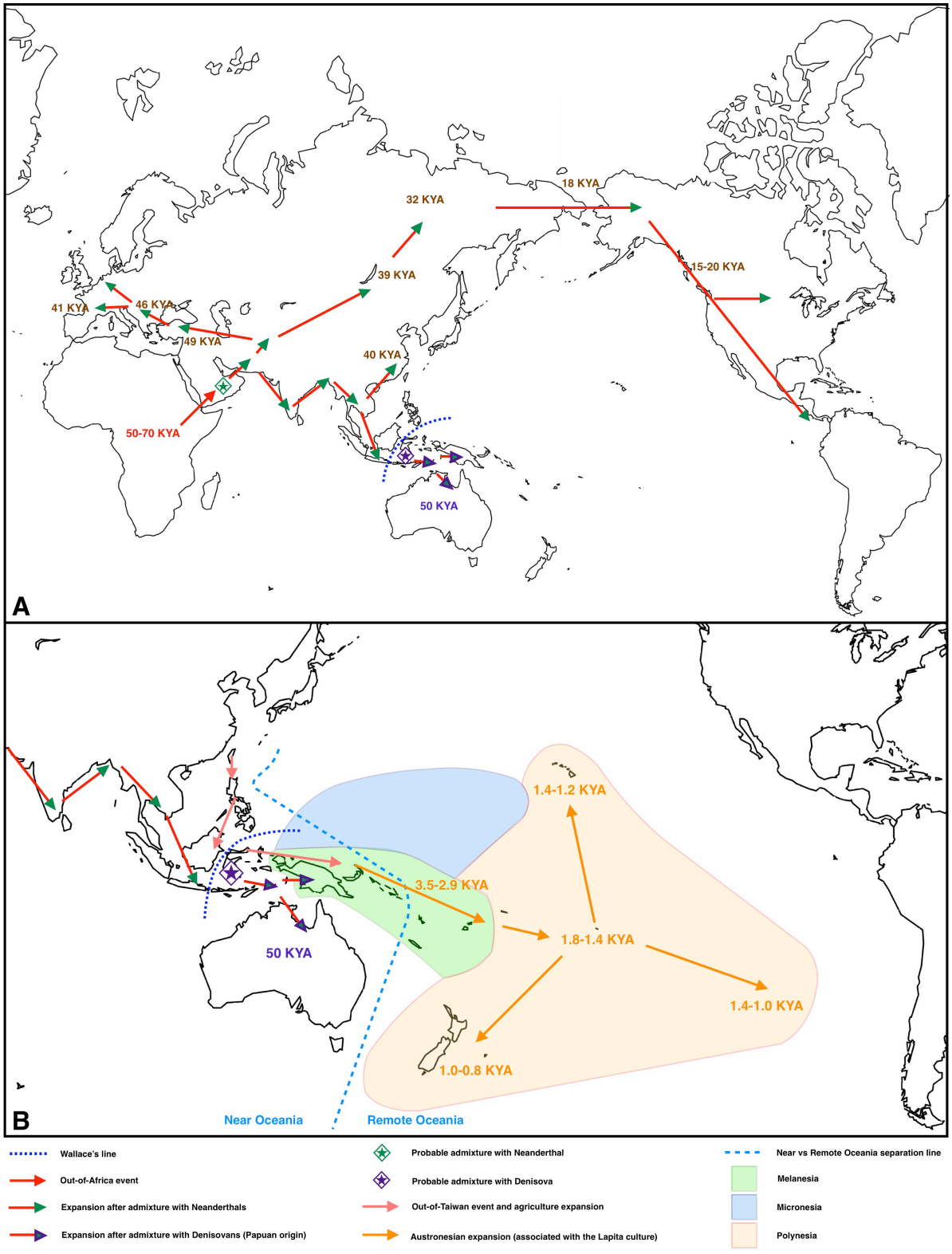


Figure 3: Main modern human migrations that shaped the current genetic landscape. A) Proposed migration routes that followed the out-of-Africa movement approximately 50-70 KYA. Possible location of admixture events with Neanderthals and Denisovans are indicated. B) Detail of the Pacific region - geographic (Near and Remote Oceania) and cultural (Melanesia, Micronesia and Polynesia) divisions are showed. The two main migration events, defined by both Papuan and Austronesian expansions, are included together with probable dates of settlement of key regions. [Maps adapted from Jobling et al. (2014), Cooper and Stringer (2013)]

individuals that firstly occupied Asia (~62-75 KYA), as revealed by Aboriginal Australian genome sequencing data (Rasmussen et al. 2011). During this migration, admixture with Denisovan individuals occurred (**Figure 3**), probably east of Wallace's line after the initial spread of modern humans (Reich et al. 2011, Cooper and Stringer 2013). The fact that this admixture event occurred east of Wallace line prevented back-migration of admixed individuals and explains the lack of Denisovan genetic signals in mainland East/Southeast Asia and its presence exclusively in the Pacific region (and in the eastern part of Southeast Asia such as Phillipines) (Reich et al. 2011, Cooper and Stringer 2013). In fact, approximately 4.8% of modern Papuan and Bouganvillean individual's genomes is related to Denisovans (Reich et al. 2010).

Rasmussen et al. (2011) study also showed that Aboriginal Australians are more closely related with Highlanders from Papua New Guinea (PNG). This is expected since at the time (50 KYA) Australia and PNG were part of Sahul (Rasmussen et al. 2011, Jobling et al. 2014). But to reach Sahul from mainland Sunda, hominids had to cross a significant water distance. As discussed by Derricourt (2005), crossing of large distances of water can be interpreted as a sign of modern human technology resulting from culture, social organization and need, and complex communication. The first confirmed maritime crossing by modern humans is dated for more than 40 KYA, a travel from the Indonesian/Malaysian archipelago to Sahul (Derricourt 2005), but it is believed that small voyages (<20 km) between Sunda and Wallacea were performed before the first true migration corresponding to the colonization of Australia 50 KYA (Balme 2013).

1.1.1.4.1 Geographic division of the Pacific region

The islands of the Pacific are classified into Near and Remote Oceania (**Figure 3B**). The former is defined by the intervisibility between islands and was first settled 45 KYA by hunter-gatherers while Remote Oceania was only settled during the last 3500 years by farmers of the Lapita complex (Terrell 2004, Jobling et al. 2014). Near Oceania includes the island of New Guinea and the islands lying off its northeast coast. As mentioned before, New Guinea formed part of the Sahul landmass when it was first settled, probably about 45 KYA. Further traveling into Near Oceania was usually done towards visible destinations but it could require voyages of 50-100 km. Early occupation evidences found in New Ireland (~35 KYA) and in

the Solomon Islands (~29 KYA) support a eastward migration from Sahul, which could help to explain the existence of Denisovan genes outside Australia and Philippines (Reich et al. 2011, Jobling et al. 2014). Further east, into Remote Oceania, the distances between islands increase to more than 350 km, rising the difficulty level of travels and settlement of modern humans, which only occurred 20 KYA (Jobling et al. 2014). In fact, the earliest settlements of Remote Oceania are all associated with a cultural package known as Lapita complex. The Lapita complex appeared 3.5 KYA in the islands off the north coast of New Guinea and is associated with agriculture, complex technology and navigational techniques that allowed them to rapidly expand east reaching as far as Hawaii, Easter Island and New Zealand (Jobling et al. 2014). Agriculture appeared independently in several locations approximately 10 KYA and its dispersal influenced the modern language pattern across the world (Bellwood 2001, Diamond and Bellwood 2003, Jobling et al. 2014). This can be explained by the fact that agriculture generated advantages over hunter-gatherers such as higher population densities, development of complex technology and acquired resistance to diseases, which ultimately lead to the replacement of hunter-gatherer societies and languages through farmers' expansion (Diamond and Bellwood 2003).

1.1.1.4.2 Cultural evidences in the Pacific region

A traditional division of Pacific islands distributes them into three groups: Polynesia, Melanesia and Micronesia (**Figure 3B**). After the initial spread of the Lapita culture from Melanesia to the western border of Polynesia around 2.9 KYA followed by a 1 KY pause, a distinct Polynesian culture started to develop and spread to previously uninhabited islands (probably to the Cook Islands and the Marquesas around 1.8-1.4 KYA). From there they expanded to Hawaii, Easter Island and New Zealand approximately 1.4-0.8 KYA (Jobling et al. 2014). This rapid expansion into East Polynesia explains the lack of cultural, biological and linguistic variation in this region (Wilmshurst et al. 2011, Knapp et al. 2012). Micronesia western islands were probably settled around 3 KYA directly from Southeast Asia and the rest of the islands were settled slightly later from islands to the south (Jobling et al. 2014).

Languages spoken on the Pacific islands are traditionally classified into two groups: Austronesian and Papuan. Austronesian languages appeared approximately 5 KYA and are highly diverse but they all belong to a singular language family widely distributed from

Madagascar to Easter Island and from Taiwan to New Zealand (Gray and Jordan 2000, Jobling et al. 2014). Papuan languages are older, dated as previous to the Austronesian inflow, and belong to many different families that may or may not share a common origin. Their presence is restricted to New Guinea and a few nearby islands, specifically from Timor to the Santa Cruz group of Eastern Solomon Islands - matching the eastern boundary of Near Oceania (Jobling et al. 2014).

1.1.1.4.3 The colonization of Remote Oceania: the Austronesian expansion

Evidences suggest that Papuan-speakers originally settled Near Oceania and that Austronesian languages arrived with farming through the expansion of the Lapita complex, which had its origin in Southeast China and arrived via Taiwan (Jobling et al. 2014). There are several evidences that point Taiwan as the Austronesian languages cradle and origin of the farmers that rapidly expanded into Polynesian: i) all Austronesian languages belong to a single family, with the highest diversity in Taiwan reducing further east into Remote Oceania; ii) Lapita-like pottery and artefacts give a clear gradient from Taiwan into Wallacea; iii) the oldest farming remains in Southeast Asia were found in Taiwan; and iv) genetic data obtained from autosomal short tandem repeats (STRs) and SNPs, Y-chromosome (Y-chr) and mitochondrial DNA (mtDNA) comparing mainland East Asian, Taiwan Aboriginals and Polynesian individuals trace back the Austronesian migration to Taiwan (Trejaut et al. 2005, Friedlaender et al. 2008, Kayser et al. 2008a, Mirabal et al. 2012, Mirabal et al. 2013, Jobling et al. 2014, Lipson et al. 2014). Nevertheless, it is important to point out that Taiwan Aboriginals ancestry can be traced back to mainland East Asia, that being the ultimate origin of the Austronesian expansion (Diamond 2001, Trejaut et al. 2005). This latter idea was recently challenged in a study based on 17 Y-STRs that showed Taiwanese tribes are not related to mainland or Island Southeast Asia and that it is possible that only some of the Taiwanese tribes were involved in the Austronesian migration out of Taiwan (Zeng et al. 2014). Other theories for the Austronesian origin are possible such as the one explored in the "Early train" hypothesis (Jinam et al. 2012). This is supported by mtDNA (Hill et al. 2006, Hill et al. 2007) and Y-chr (Karafet et al. 2010) analyses and states that Austronesians originated approximately 10-30 KYA in Indochina/South China dispersing south and west through Malaysia, Sumatra, Java and Borneo, which were then part of Sunda (Jinam et al.

2012). This hypothesis does not exclude the later migration from Taiwan but considers it cannot be the sole contributor of all Austronesian genetic diversity (Jinam et al. 2012).

Several other models were defined to explain how the dispersion in the Pacific region occurred, each reflecting a differing degree of time of contact and ancestral contribution from Southeast Asia and Near Oceania to Remote Oceania. The extreme models are the "Express Train" – described by Diamond (1988) and later expanded by Gray et al. (2009) to include a series of pulses and pauses, supported by linguistic evidences (Blust 1999, Cann 2000, Gray and Jordan 2000) – and the "Entangled Bank" (Terrell 1988). Examples of intermediate models are the "Triple I" (Green 1991) and the "Slow Boat" (Kayser et al. 2000a). Each of these models may not fit all the evidence available and they are not complete, i.e., they lack sufficient complexity (Jobling et al. 2014). Furthermore, it is also possible that contacts and genetic admixture between Asian and Melanesian populations occurred only in some areas but not in others, explaining differences between studies (Ricaut et al. 2010).

1.1.1.4.4 Genetic evidences in the Pacific region

Genetic evidence supports archaeological and linguistic findings that define the origin of Pacific islanders in the west and their slow migration towards east as described by the "Entangled Bank" and "Slow Boat" models. Despite not being representative of the complete genome, strong evidences can be obtained from mtDNA and Y-chr, especially when trying to clarify human population histories and possible sex-biased or socio-cultural events (Kayser et al. 2006, Wilkins and Marlowe 2006, Underhill and Kivisild 2007, He et al. 2012). There are clear lineage distribution differences in Asia and Near Oceania: uniparental lineage diversity reduces from west to east towards Remote Oceania and lineages present in Remote Oceania can be tracked back to both Asia and Near Oceania (Kayser et al. 2006, Kayser et al. 2008a, Mona et al. 2009, Karafet et al. 2010, Kayser 2010, Corser et al. 2012, Delfin et al. 2012, Mirabal et al. 2012, Duggan et al. 2014, Jobling et al. 2014). Furthermore, there are evidences of this being a sex-biased event with high numbers of Melanesian Y-chrs and Asian mtDNAs found in Polynesia (Kayser et al. 2008a). But, as discussed by Duggan et al. (2014), despite the great Austronesian influence, the migration history into Oceania is not straightforward, with contradicting information being given by complete mtDNA genomes. For example, frequency distribution differences between haplogroups B and E were observed, which might

be explained by other migrations into Near Oceania rather than the two previously discussed (Duggan et al. 2014). A very important marker indicative of the Austronesian expansion is the "Polynesian motif" (Melton et al. 1995), which is nested in the B4a1a clade. This is restricted to Austronesian-speaking populations and it is almost fixed in Remote Oceania (Melton et al. 1995, Kayser et al. 2006, Hill et al. 2007). Although haplogroup B4a1a1 was dated as previous to the out of Taiwan migration (Soares et al. 2011), a recent study suggests that this is not an evidence of a possible pre-Lapita Near Oceanian origin and that its dispersal is clearly associated with the Austronesian expansion (Duggan et al. 2014). Furthermore, Oceanian population complexity increases when considering evidences from uniparental (Redd and Stoneking 1999, Redd et al. 2002) and autosomal (Pugach et al. 2013) data that suggest a recent relationship between Indian subcontinent and Aboriginal Australian populations or an early Amerindian contribution to the Easter Island gene pool (Thorsby 2012).

Finally, despite the small amount of works dedicated to autosomal markers, genome-wide analysis is not consistent with the "Entangled Bank" model but gives different points of view. Studies analysing STRs, SNPs and Indels consider that the "Slow Boat" model is more appropriate (Friedlaender et al. 2008, Kayser et al. 2008b, Wollstein et al. 2010). Data indicated low genetic diversity in Remote Oceania and high diversity in Melanesia, with the majority of the Polynesian genomes being closer to East Asia than Melanesia. But this data is not informative regarding the pace and time of the dispersal and admixture events.

1.1.2 Evolutionary forces

The frequencies of the population variants of one polymorphic locus are regulated by a series of factors: mutation, genetic drift, natural selection and migration (**Figure 4**). But the variability of each region of the genome is affected differentially. Coding and regulatory regions are affected by selection through processes of natural selection, which can exert negative or positive effects. In the first case, the genetic variant creates changes to protein sequence, secondary structure or expression that can consequently lead to a reduction of individual fitness reducing the variant frequency in the population. But when the variant confers advantages for survival or reproduction, its frequency in the population will increase. On the other hand, DNA with no known function, including non-coding regions, are largely

free from selection, so any variant that appears would maintain, in theory, the same low frequency. However, this does not always happen since there are certain evolutionary forces, independent of natural selection, acting on them.

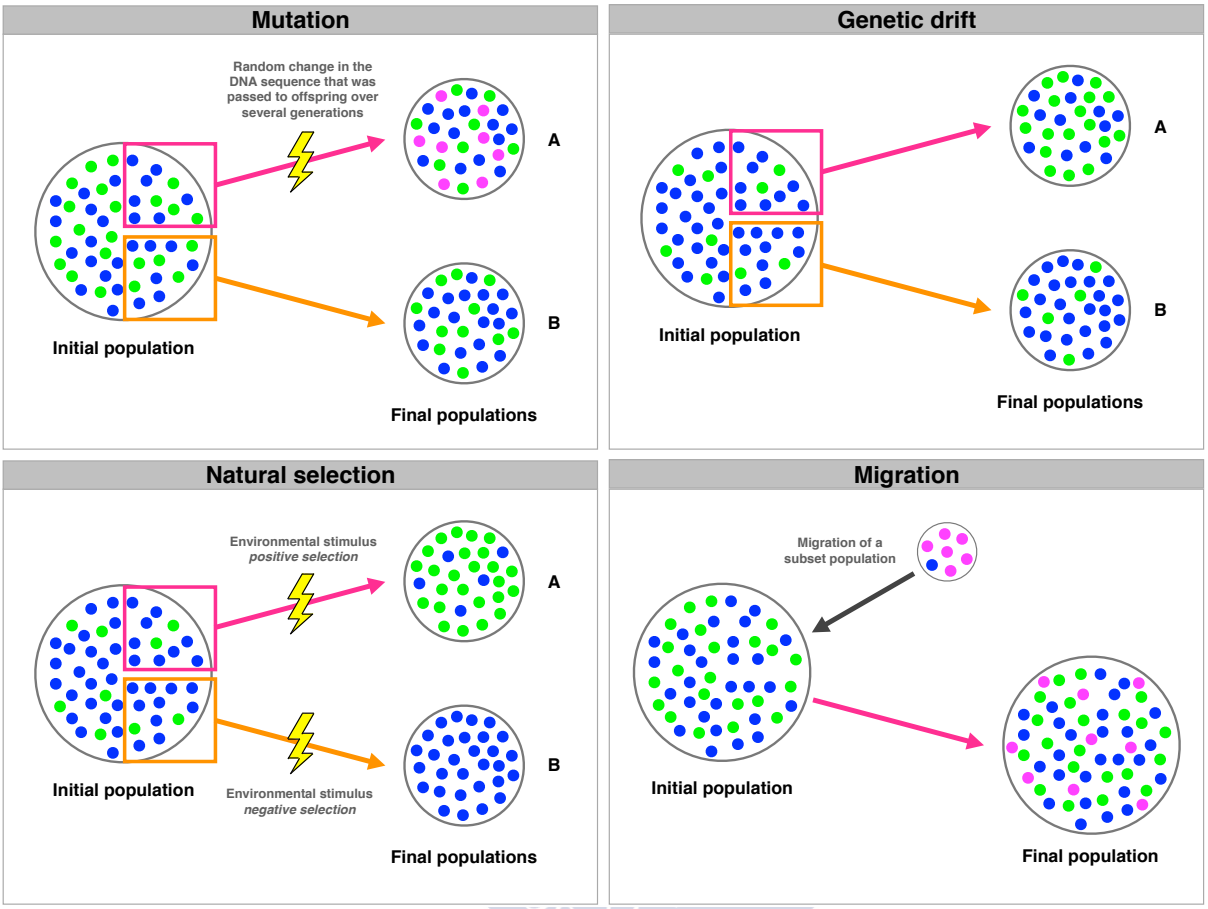


Figure 4: Schematic representation of the effects induced by evolutionary forces on a hypothetical initial population. In each case, the variability of a possible characteristic is represented in an initial population. The possible effects on the frequencies of such variants, after several generations, are represented in the hypothetical final populations. Mutation is a random event that generates new variability. Genetic drift corresponds to the random variation of allele frequencies across generations. Natural selection acts over existent phenotypes and it can increase the frequency of a variant (positive selection) or reduce it (negative selection). Migration, or gene flow, is the movement of subsets of populations that leads to the exchange of genetic variability.

1.1.2.1 Mutation

Mutations are probabilistic events that correspond to random changes in the genetic code. For that reason, they are the source of genetic variation as this is the only process that creates new alleles. By itself, mutation does not have a great impact on allele frequencies, but any new mutants that appear in a population will later be affected by other evolutionary forces.

They usually occur in a single cell and can cause damage in somatic tissues. But when present in the germ-line, that way being hereditable, mutations can be affected by evolution.

Mutations can be classified into three major categories: i) point mutations, i.e., the change of one DNA base for another, can have different effects depending on their location – neutral mutations occur in the non-coding region of DNA; silent mutations, present in coding regions, do not produce any effect in the amino acid sequence (due to redundancy of the genetic code, see Table 1); but when a mutation changes the amino acid sequence there can be noticeable effects; ii) insertions and deletions of DNA sequences can change the function of a gene; iii) chromosomal rearrangements affect large sections of chromosomes or even entire chromosomes.

The probability of a new mutation is generally low (due to the fidelity of the DNA polymerase and DNA repair mechanisms), so one mutation can appear in a population and not in other, changing the genetic pool of one population and increasing the variability between populations.

1.1.2.2 Genetic drift

Genetic drift corresponds to random changes in allele frequencies between generations, in the same population. Individually, each parent only passes half of their nuclear genome to their offspring, and this can be different between each child. But at a population level each generation represents a number of copies of each variant dictated by previous frequencies but subject to random fluctuations, so that the frequency of each allele can vary greatly in the genetic pool of the population. After several generations, a particular allele can be fixed, that is, it reaches a frequency of 1, or it can be extinct, i.e., its frequency is 0. In both cases, no changes will occur in the population unless new alleles are introduced through mutation or gene flow from other populations. This evolutionary force acts to increase the variation between groups and decrease variation within the group. The population size is of great importance: in smaller populations alleles are rapidly fixed or lost and, in larger populations, changes in allele frequencies are generally much more subtle.

1.1.2.3 Natural selection

Natural selection requires differences in survival and reproduction rates. An individual's survival can be affected by certain traits, which can vary with the surrounding environment. But in an evolutionary context, gene survival is more important, that is, natural selection also considers differences in an individual's fertility and its ability to pass his genes to the next generation.

A common type of selection is that against recessive homozygotes, that is, when two copies of a harmful recessive allele are present (as is the case in certain genetic disorders) the individual's fitness is reduced. As allele frequency reduces when selection acts against that allele, it is expected that such allele will be completely removed from a population, as shown in **Figure 4** (negative selection example). Nevertheless, other evolutionary forces are acting, at the same time, on that population. Mutation adds the negatively selected allele to the genetic pool, reaching equilibrium with selection against the mutation. Genetic drift can also influence the population allele frequency of a harmful mutation, mainly depending on the relative strength of both selection and drift. In the case of advantageous mutations, selection usually leads to the increase of their frequency in the population, either reaching near fixation or through balancing of the allele frequencies, that is, it selects for heterozygous individuals without favouring one allele over the other.

Natural selection acts through generations, but there is a particular case where an allele can be eliminated from a population in a single generation. That is the case of selection against dominant alleles, particularly if such allele is lethal. In this case, the fitness of both homozygous dominant and heterozygous would be reduced.

1.1.2.4 Migration and gene flow

Population migration and gene flow are two tightly linked concepts, but while migration refers to movement and change of location from one occupied area to another, gene flow specifically indicates the movement of genes, i.e, it is the outcome of migration that indicates reproductive success of the migrants in their new location. Gene flow in particular, represents a process capable of changing allele frequencies within a sub-population through the homogenization of the frequencies and the introduction of new variation absent or rare in the

genetic pool of the recipient population. As genetic drift, gene flow is a random process but while drift increases variation between groups, gene flow reduces genetic differences between populations.

Migration into previously unoccupied land by a set of founders representing a subset of the original population that may or may not be representative of variant frequencies in that population (as shown in **Figure 2**) is called colonization.

1.1.3 Human population structure and ancestry informative markers

The distribution of human genetic diversity has long been a subject of interest, and it has important implications for human evolution, forensics, and the distribution of genetic diseases in populations (Jorde et al. 2000). Commonly in the past, individuals were sorted into groups based on characteristics such as physical appearance or language. Human phenotypic variation can be organised into a geographical pattern. This is why members of the same local group are typically more closely related to each other than to members of groups who live in different geographical areas (Bamshad et al. 2004).

Genetic diversity in human populations is low, indicating a recent origin and a small size of the ancestral human population. Initially, within-population differences between individuals was estimated to be approximately 85% (Lewontin 1972). A more recent study analysing 377 STRs showed that while the proportion of diversity that exists between human continental populations is relatively low (around 5%), the within-population, inter-individual component of genetic variation accounts for approximately 95% proportion of human genetic diversity (Rosenberg et al. 2002). Although the proportion of genetic variation distributed among populations is low in comparison to other species, the genetic differences can be used to sort individuals into groups with each genotype contributing equally to the estimate of between-population and within-population variation. But the contribution of each genotype to the distribution of individuals into genetically inferred groups is cumulative and allele frequencies among loci within a population are correlated. The statistical power with which ancestry can be inferred reliably depends on the informativeness of each marker tested. For example, a bi-allelic marker such as a SNP has an informativeness value maximised when one allele is limited to a single population, i.e., if an individual has this allele, there is little uncertainty about their origin. At present, there are statistical methods that can detect and

correct population stratification and can be used to infer ancestry by combining information from markers that differ in frequency between ethnic groups (Bamshad et al. 2004, Enoch et al. 2006, Jobling et al. 2014).

Recent studies using several hundred markers had suggested that the world's population can be grouped into seven major discrete ancestral groups (Rosenberg et al. 2002, Enoch et al. 2006, Allocco et al. 2007, Li et al. 2008). In fact, when individuals are grouped on the basis of genetic similarity, group membership corresponds closely to continental origin. Also on those studies, it was possible to infer ancestry and ethnic affiliation in recently admixed populations in the context of the contributions of putative parental populations. Mixed ancestries inferred from genetic data can often be interpreted as arising from recent admixture among multiple founder populations. However, it can also be the result of a shared ancestry before the divergence of the two populations without subsequent gene flow between them. Genetic clusters often corresponded closely to predefined regional or population groups or to collections of geographically and linguistically similar populations.

The gene diversity results show higher levels of diversity in African populations than in non-African populations. A higher level of African diversity supports the hypothesis that modern humans first arose in Africa and then colonised other parts of the world. The study of Li et al. (2008) produced a population tree that included the major splits between different continents but also sublineages within continents – the obtained branching pattern largely agrees with the approximate order of human expansion and supports the out-of-Africa model of human origin.

The dispersal of humans throughout the world distributed genetic variation due to demographic events such as bottlenecks and effective population size (there is more fluctuation of the neutral markers frequencies in smaller populations). This increased the variance of neutral polymorphism frequencies among groups. Furthermore, polymorphisms that arose in one group had a greater chance of remaining restricted to that group because their spread was limited by the degree of gene flow between groups. The founding and dispersion of new populations eventually led to the sorting of different polymorphisms and polymorphisms with different frequencies among populations since founder populations typically carried with them only a subset of the genetic variation found in their ancestral population. As these founders became more widely separated from one another, the probability of two individuals mating with one another became more and more variable.

Individuals were more likely to mate with one another if they lived closer to each other and this assortative mating restricted the exchange of polymorphisms among individuals who lived in different geographical regions, and over time led to genetic differentiation between groups (Jorde et al. 2000, Bamshad et al. 2004).

As described by Enoch et al. (2006), the choice of markers can influence the ability to differentiate populations. Furthermore, the power to estimate the most probable geographical origin of a particular polymorphism variant or a sequence profile using appropriate algorithms and databases depends mainly on how informative the marker under study is (Salas et al. 2006). But as Allocco et al. (2007) have pointed out, the differences between continentally defined groups are large enough so that a randomly selected set of SNPs (which represents a minute fraction of the genetic variation in the human genome) can be used to characterize ancestral geographic origin in an accurate and reproducible manner.

By definition, ancestry informative markers (AIMs) are those that show alleles with large frequency differences between populations independently of them being associated with phenotypic traits. For this, neutral markers may also be useful for the estimation of ancestry. They can include microsatellites, short biallelic Indels and SNPs. AIMs are selected to maximize δ, the absolute difference in allele frequency between two ancestral populations, and the ideal AIM is monomorphic in one population and highly abundant in another ($\geq$ 60%) (Yang et al. 2005, Enoch et al. 2006, Salas et al. 2006).

Uniparental markers are also useful for ancestry assignment. The dispersal of human populations after the out of Africa event led to a differential distribution of the variants present both in the Y-chromosome as in the mitochondrial DNA. Both markers are represented by a series of well-characterized haplogroups (**Figure 5**), each defined by a phylogenetically stable set of relatively slowly mutating polymorphisms. Due to the action of evolutionary forces described in section 1.1.2, haplogroups are not equally distributed across human populations (**Figure 5**). For that reason, it is possible to infer the probable geographic origin of an individual from uniparental markers data.

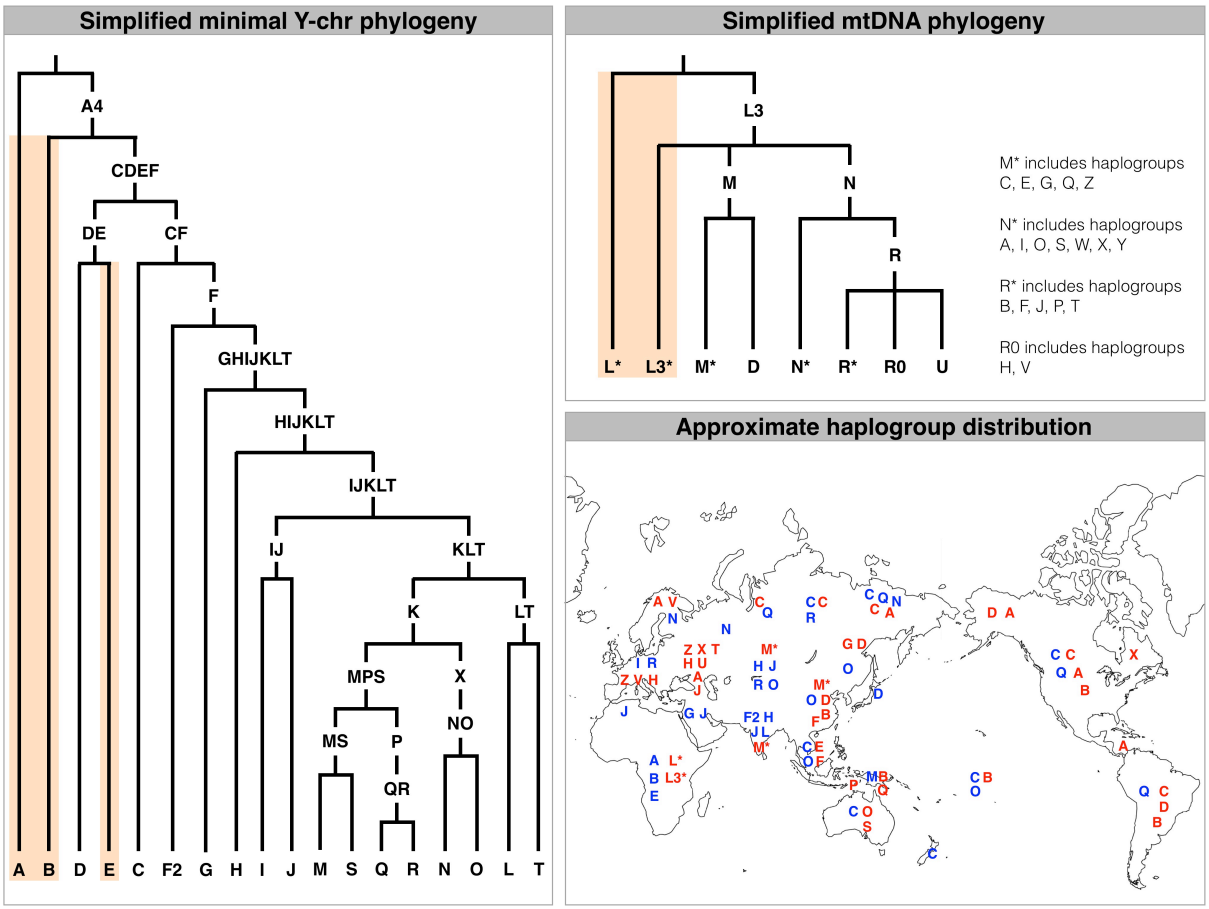


Figure 5: Y-chromosome and mitochondrial DNA phylogenies and haplogroup distribution. Simplified versions of the phylogenetic trees are represented – both include major clades. Orange shaded areas represent African lineages. The simplified minimal Y-chromosome phylogeny was adapted from the PhyloTree-Y tree version from November 30th 2014 (available at http://www.phylotree.org/Y/Y_tree_skeleton.pdf) and originally published by van Oven et al. (2014). The simplified mtDNA phylogeny tree was adapted from the PhyloTree-mtDNA tree build 16 from February 19th 2014 (available at http://www.phylotree.org/tree/main.htm). Approximate haplogroup geographical distribution is represented in the map – blue letters correspond to Y-chr and red to mtDNA haplogroups. The distribution map was adapted from frequency distribution maps from Jobling et al. (2014), available at http://www.scs.illinois.edu/~mcdonald/WorldHaplogroupsMaps.pdf.

1.2 THE HUMAN GENOME AND DNA POLYMORPHISMS

With the exception of some viruses, all organisms use deoxyribonucleic acid (DNA) as their genetic material. This macromolecule stores the information needed to construct all the components of a cell and the total content of genetic information in an organism is organised in its genome.

1.2.1 Nucleic acids

DNA is a polymer composed of monomers called nucleotides, each being constituted by three parts: one base, one sugar and three phosphate groups (**Figure 6A**). The base is the informational part of the nucleotide: the order in which the bases appear in the DNA sequence defines the genetic information encoded. There are four different bases: adenine (A) and guanine (G) are called purines and can be defined by the presence of two closed rings; cytosine (C) and thymine (T) are called pyrimidines and are single-ringed molecules (**Figure 6B**). The base is attached to the 1' carbon atom of the sugar molecule, in this case deoxyribose. The three phosphate groups are attached to the 5' carbon of the deoxyribose (**Figure 6A**). The sugar molecule and the phosphate groups have a structural role being the key components of the single-stranded DNA polymer: the sugar-phosphate backbone – the phosphate group of one nucleotide binds to the 3' carbon of the next nucleotide (two phosphate groups are released as a result of the polymerization). This particular structure is defined by its polarity: the polymerization process always occurs in the 5' to 3' direction, that is, the 5' end has a free hydroxyl (-OH) group on the deoxyribose 5' carbon and the 3' end has a free hydroxyl group on the 3' carbon (**Figure 6C**). But most of the time, DNA within a cell exists in the double-stranded form. This structure resembles a double helix with the complementary strands being kept together by hydrogen bonds between bases (**Figure 6C,D**). The base pairing is highly specific: thymine always pairs with adenine through two hydrogen bonds; cytosine and guanine are kept together by three hydrogen bonds (**Figure 6E**).

But DNA is not the only nucleic acid found within cells. Ribonucleic acid (RNA) has a different sugar molecule (ribose instead of deoxyribose) and thymine is replaced by uracil (U), a pyrimidine that pairs with adenine. RNA molecules can be divided into several species depending on their function. Non-coding RNAs (ncRNA) are not translated into protein

sequences but have crucial functions in translation (the production of a protein within the ribosome), such as transfer RNA (tRNA) and ribosomal RNA (rRNA), and regulation, such as micro RNA (miRNA). On the other hand, messenger RNA (mRNA) has a key role in protein production as it is the product of gene transcription (the production of a RNA molecule from a gene coded in DNA) and is the template for translation.

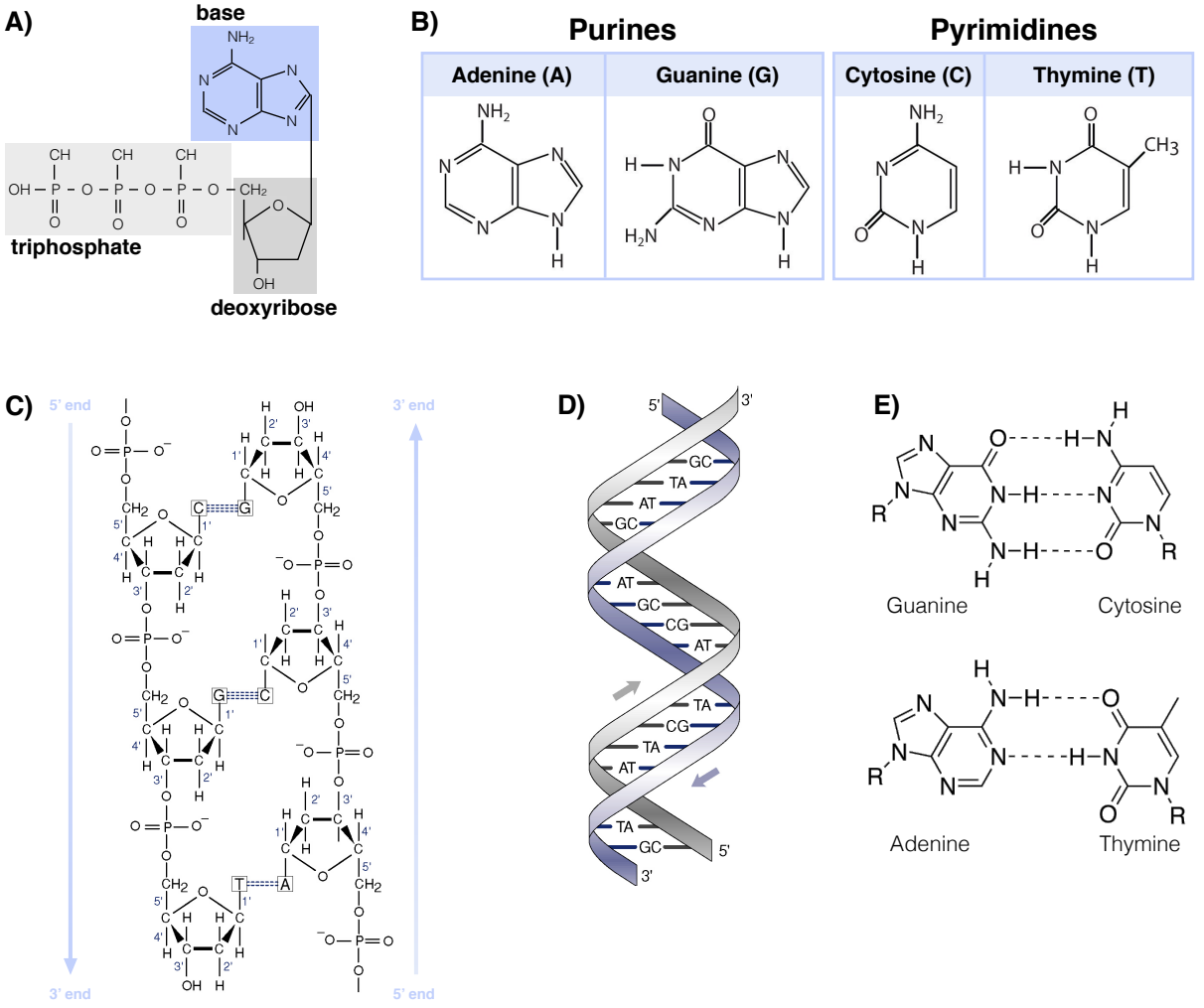


Figure 6: Nucleotides and DNA structure. A) Nucleotide composition. B) Bases found in DNA: purines (adenine and guanine) and pyrimidines (cytosine and thymine). C) Double-stranded DNA has two anti-parallel strands connected by pairing bases. D) The two strands form a double helix. E) Guanine and cytosine are paired through three hydrogen bonds while adenine and thymine by two hydrogen bonds.

1.2.2 Organization of the genetic information

As mentioned before, the order in which the bases appear in the DNA sequence defines the genetic information encoded. This information is mainly translated into proteins – macromolecules produced by the cells that participate in nearly every process within the cell and that assume diverse functions such as structural, mechanical, signalling or even metabolic. Despite the importance of such macromolecules, as little as 3% of the genome is used to code them (EPC 2012). And although no specific function is known for most of the human genome, approximately 75% is transcribed into RNA molecules (also known as transcriptome), including the previously mentioned ncRNA (Mattick and Makunin 2006, Jobling et al. 2014).

The information necessary to produce proteins and functional RNAs is organised in structures called genes. Those DNA segments are first copied into an mRNA molecule (transcription) that is then used to convert the nucleotide sequence into an amino acid sequence (i.e., a protein) in a process called translation. But a gene comprehends more DNA sequence than the necessary to code a protein. The 5' end of the gene usually includes sequences necessary for the initiation and regulation of the transcription process, such as the promoter and enhancer. On the other side, at the 3' end, termination signals and sequences responsible for mRNA stability are coded. But the produced mRNA molecule is only a precursor of the final protein template. In fact, genes include non-coding sequences (introns) in between the coding regions (exons). The processing of the pre-mRNA molecule includes, among other changes, the removal of the intron sequences, a process known as splicing. Only then the mRNA molecule is mature and ready to be transferred to the cytoplasm where it will be translated into a protein sequence.

Proteins are polymers composed by the combination of 20 different monomers, in this case amino acids. The genetic code (Table 1) is a nearly universal three-letter code that allows the translation of the nucleotide sequence of the mature mRNA into a specific amino acid sequence. There are 64 possible codons, i.e., sets of three adjacent nucleotides, each representing a specific amino acid. For this reason, the same amino acid can be represented by more than one codon, that is, the genetic code is redundant. Furthermore, some codons indicate the starting and stopping points of translation.

Table 1: The genetic code. There are 64 possible three nucleotide combinations, each being translated into a specific amino acid. The AUG codon represents the Met amino acid but it also indicates the translation starting point. On the other hand, UAA, UAG and UGA codons indicate the translation stopping point. There are 20 amino acids: Ala - alanine; Arg - arginine; Asn - asparagine; Asp - aspartic acid; Cyst - cysteine; Gln - glutamine; Glu - glutamic acid; Gly - glycine; His - histidine; Ile - isoleucine; Leu - leucine; Lys - lysine; Met - methionine; Phe - phenylalanine; Pro - proline; Ser - serine; Thr - threonine; Trp - tryptophan; Tyr - tyrosine; Val – valine. Note that this table represents the human nuclear genome code (there are some differences in comparison with the mitochondrial genome).

1st position (5' end)	2nd position: U	2nd position: C	2nd position: A	2nd position: G	3rd position (3' end)
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA Stop	UGA Stop	A
	UUG Leu	UCG Ser	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

Although this thesis will not address genes or their products, this section is important to understand their structure and, more importantly, the difference between introns and exons. Genetic variants located in genes can be particularly useful for some applications in the forensic field. Nevertheless, there are ethical considerations about the use of coding regions that should be contemplated (section 1.3.1.3).

1.2.3 The human genome

The human genome is divided into two main categories, each in a different cell compartment (**Figure 7**). First, the nuclear genome, contained in the nucleus, is present in two copies per diploid cell being composed by approximately 6.4 billion base pairs (bp) organised into 46 linear molecules called chromosomes. Those are organised in pairs: 22 pairs are common to all individuals (autosomes) and one pair differs between genders (sex

chromosomes) – while females have two X-chromosomes, males have one X and one Y copy. Second, the mitochondrial genome (mtDNA) is a circular molecule composed by 16,569 bp with 2-10 copies being found per mitochondrion (cytoplasmic organelle that can reach numbers of 100 to 1000 per cell). Variability, that is, differences between genome copies, can be found in both genomes but in the mtDNA this is mainly located in the control region – a segment of 1.1 Kb that includes two hypervariable segments: HVS I and HVS II.

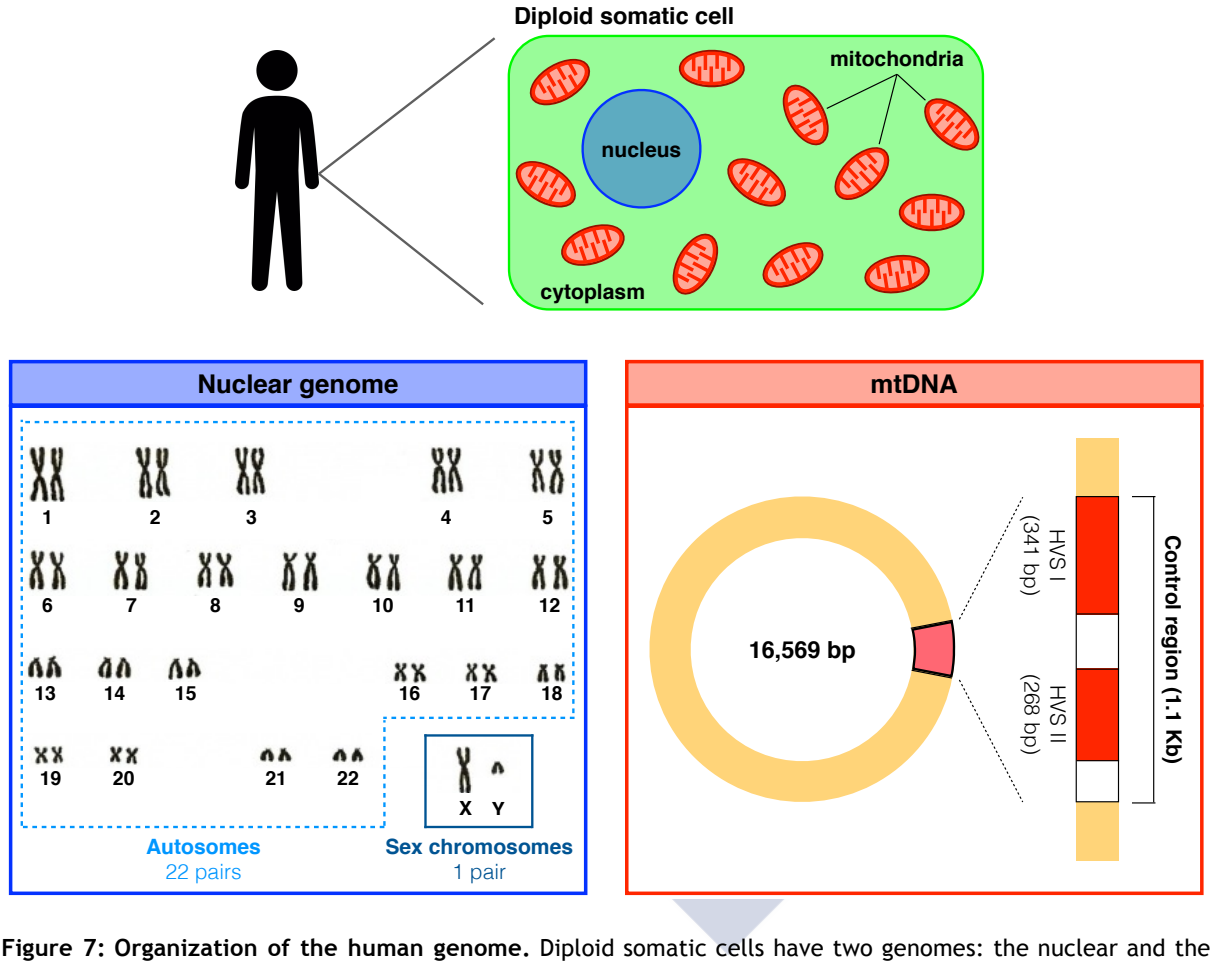


Figure 7: Organization of the human genome. Diploid somatic cells have two genomes: the nuclear and the mitochondrial (mtDNA). The nuclear genome is organised in 22 autosomic pairs and one gender defining pair (sex chromosomes). The mtDNA is a circular molecule comprising 16,569 base pairs. The control region has two hypervariable segments (HVS I and II) where the great majority of polymorphisms are located.

1.2.3.1 Nuclear genome structure

As mentioned before, the nuclear genome is organised in chromosomes, each corresponding to a highly condensed linear DNA molecule. In fact, DNA packaging is an essential feature that, together with DNA methylation, helps to regulate gene expression: the transcription activity is higher in the less condensed regions.

First, the linear DNA double helix is coiled around an octameric protein complex (histones) forming a nucleosome. The further coiling of the nucleosomal fibre generates the 30 nm fibre (or solenoid) that is then organised into chromatin. Chromosomes correspond to the higher packaging stage (**Figure 8**) – although their structure is highly recognizable, it is only present during small periods of the cell cycle, specifically during cell division.

Chromosomes have different characteristics, the most evident being the size difference observed between autosomes. As shown in **Figure 7**, their size increases from chromosome 1 to chromosome 22. Sex chromosomes are quite different between themselves: while the X-chromosome (X-chr) matches chromosome 7 with approximately 155 Mb, the Y-chromosome (Y-chr) is among the smallest with approximately 60 Mb. Another characteristic is the centromere (**Figure 8**). This is a highly condensed region that is involved in chromosome segregation during cell division. In fact, DNA replication during cell division originates temporary tetraploid cells: the two copies of each chromosome (known as sister chromatids, **Figure 8**) are kept together at their centromeres. Furthermore, the centromere divides the chromosome into two regions termed arms: the smallest is the *p* arm and the longest the *q* arm. The centromere position varies between different chromosomes allowing them to be grouped (**Figure 8**): i) metacentric chromosomes have two clearly distinguishable arms and the centromere is close to the centre; ii) when the *p* arm is present but it is short the chromosome is called acrocentric. Finally, the telomeres are specialised structures at each end of the chromosome (**Figure 8**). They have protective functions such as avoiding sequence loss and degradation but also prevent fusion with other chromosomes.

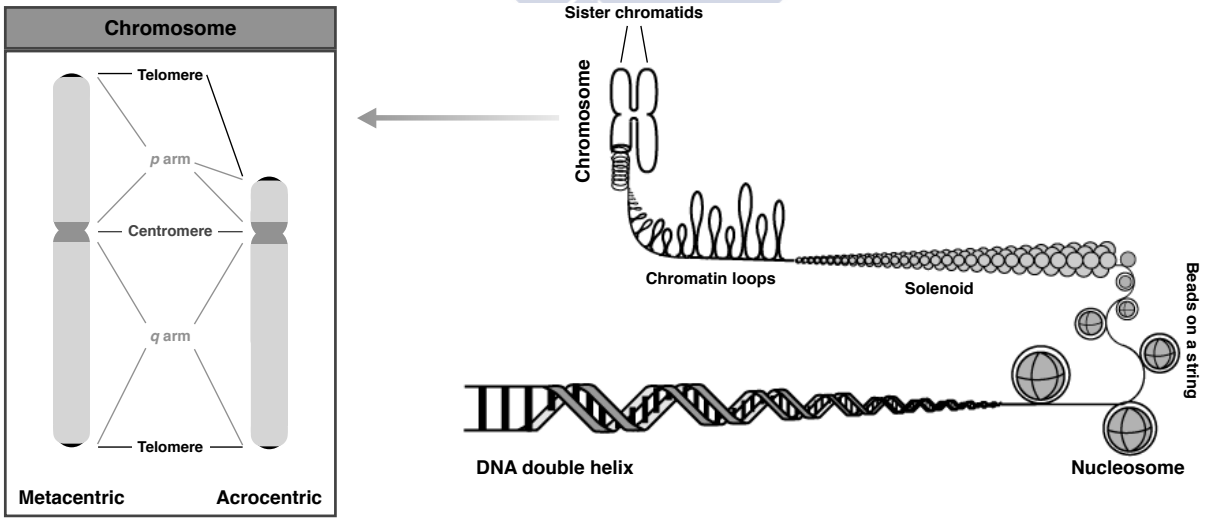


Figure 8: DNA packaging and human chromosome structure.

1.2.3.2 Inheritance of the genome

Cellular division is the basis of life of any organism. This process includes passing the genetic information from mother to daughter cells.

All living organisms that breed through sexual reproduction have two types of cellular division: mitosis and meiosis. Mitosis includes a series of events that culminate in the cell division. It occurs in all somatic cells and is responsible by the growth and maintenance of the organ and body structure. During mitosis the genome of the mother cell is passed, unchanged, to the daughter cell. But during this process some errors in DNA replication (mutations) can happen and sometimes contribute to disease development. On the other hand, meiosis is a specialised division process that occurs during gamete production and reduces the genomic content by half. Gametes (sperm and egg) are haploid cells and the fertilization process, i.e., the fusion of two gametes that initiates the development of a new organism, is necessary to maintain the number of chromosomes in the following generation. For this reason, the resulting zygote contains a mixture of paternal and maternal chromosomes. As in mitosis, mutations can occur during meiosis. But while the first are confined to the individual that carries them, mutations during meiosis can be passed to the offspring.

Considering the human genome previously described, autosomes are inherited from both parents, one haploid copy from each. The Y-chr, being responsible by male determination, is only passed down from father to son (paternal inheritance). This implies that the X-chr in men is always inherited from the mother. MtDNA can be found in both genders but it is only passed down from mother to all her children, so it is maternally inherited (**Figure 9**).

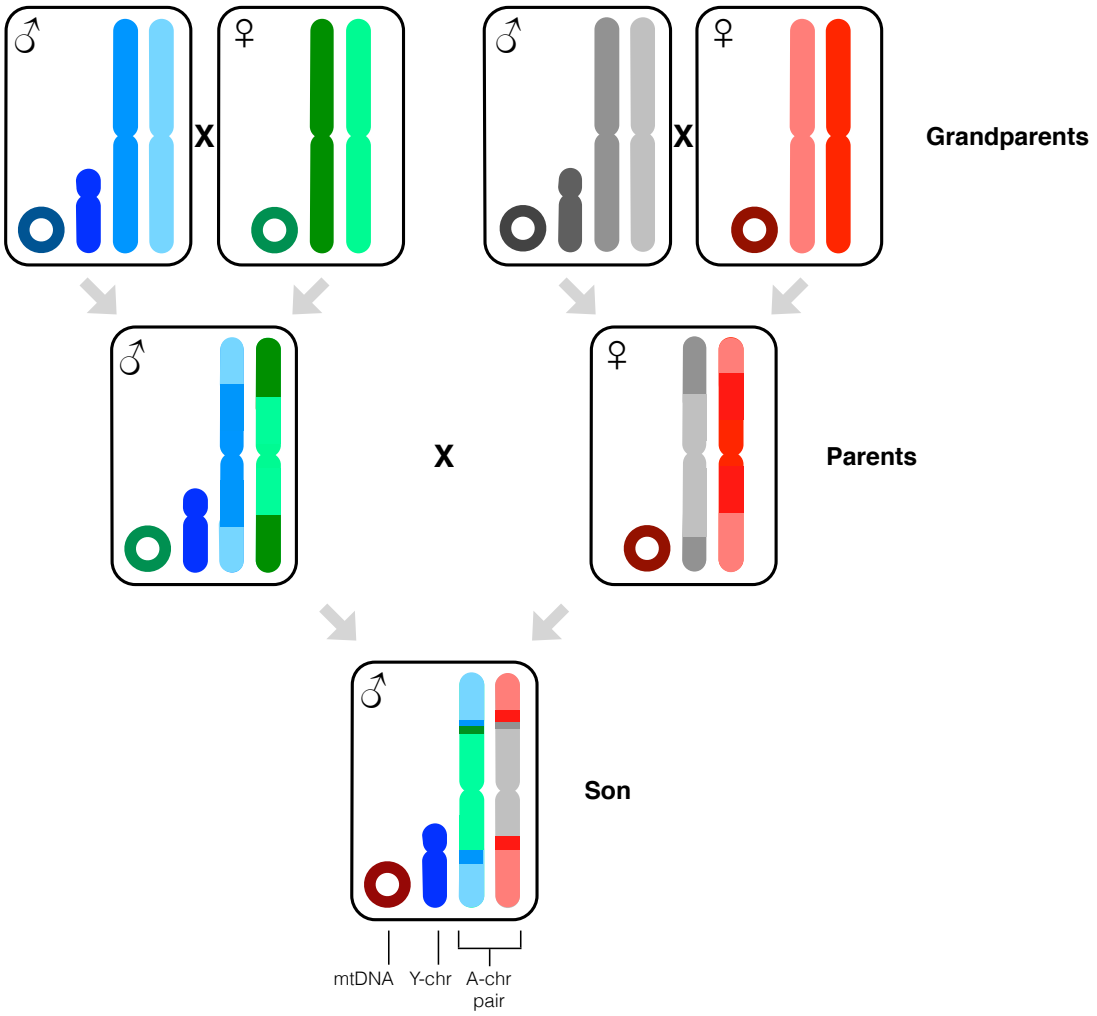


Figure 9: Inheritance of the genome. A-chr - autosomic chromosome pair.

1.2.3.2.1 Recombination in the nuclear genome

All cells produced by meiosis are genetically different due to independent assortment and recombination (also known as crossing over). Independent assortment is the random separation of each chromosome copy into a different gamete. Recombination is an event that takes place during the initial phase of the meiotic division. It occurs at the chiasmata (points at which the chromosome pairs remain in contact during the first phase of meiosis), leading to the exchange of DNA between homologous chromatids of a chromosome pair (**Figure 10**). This process has some sex-related differences, with more recombination in female meiosis (Collins 2009). Overall, gametes receive a recombined chromosome originated from both maternally and paternally derived homologs (Collins 2009).

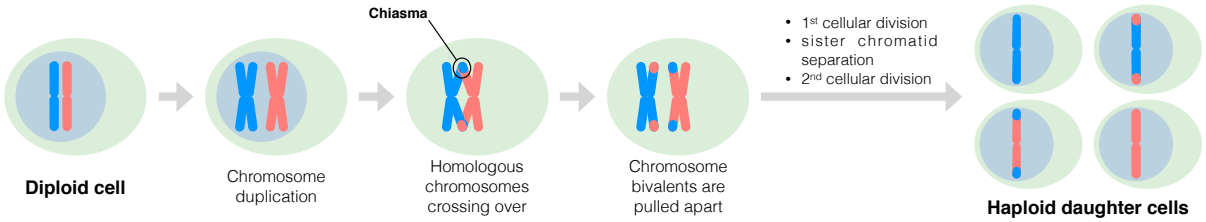


Figure 10: A summary of the meiotic division. During the first steps of meiosis, the diploid genome is duplicated and tightly condensed in chromosomes, which originates a tetraploid cell with chromosome bivalents. Then, each pair of homologous chromosomes exchanges genetic material through crossing over at the chiasma. After recombination, chromosome bivalents are pulled apart and the first cellular division occurs. In each of the resulting cells, sister chromatids are separated and the second cellular division takes place. The final product is two recombinant and two non-recombinant haploid cells. Note that, for clarity, only one chromosome is shown.

The likelihood of recombination between two autosomal markers depends on two factors: physical and genetic distance. Considering physical distance, markers lying far apart from each other are often separated by recombination, while markers close together are rarely separated (Collins 2009, Jobling et al. 2014). Genetic distance represents the recombination frequency between two points in the genome and is expressed in centiMorgans (cM). The existence of recombination "hotspots" or "cold domains" can affect the recombination rate between markers. Hotspots can be defined as short regions (around 2 Kb) where the recombination rates increase dramatically in comparison with the local rates (IHMC 2005). In fact, most of the fine scale recombination events that occur in the human genome are concentrated in such regions (Abecasis et al. 2005, IHMC 2005). Furthermore, recombination also varies across the full length of a chromosome following an expected pattern (**Figure 11**). It tends to be higher in the telomeric regions, especially in the terminal 20-35 Mb, and on shorter chromosome arms (IHGSC 2001).

Linkage disequilibrium (LD) reflects the tendency of particular alleles at separate loci to be co-inherited because of reduced recombination between them, leading to the establishment of haplotypes (Jobling et al. 2014). For the majority of markers not under natural selection, LD is gradually broken down by recombination and regenerated by genetic drift (Reich et al. 2001, Abecasis et al. 2005). So, the strength of allelic association depends, for a population, on the number of founding individuals, the time since founding, and other factors such as mutation, drift and natural selection (these evolutionary forces were described in detail in section 1.1.2). This can explain why isolated populations have larger linkage disequilibrium (Reich et al. 2001, Collins 2009). As recombination is responsible for the breaking of the linkage blocks, LD distribution across a chromosome is opposite to that of recombination

(**Figure 11**): LD blocks are usually near the centromere and correlated with chromosome length with low LD near the telomere (IHMC 2005, Collins 2009). The X-chr is a particular case since it has extensive LD blocks (IHMC 2005).

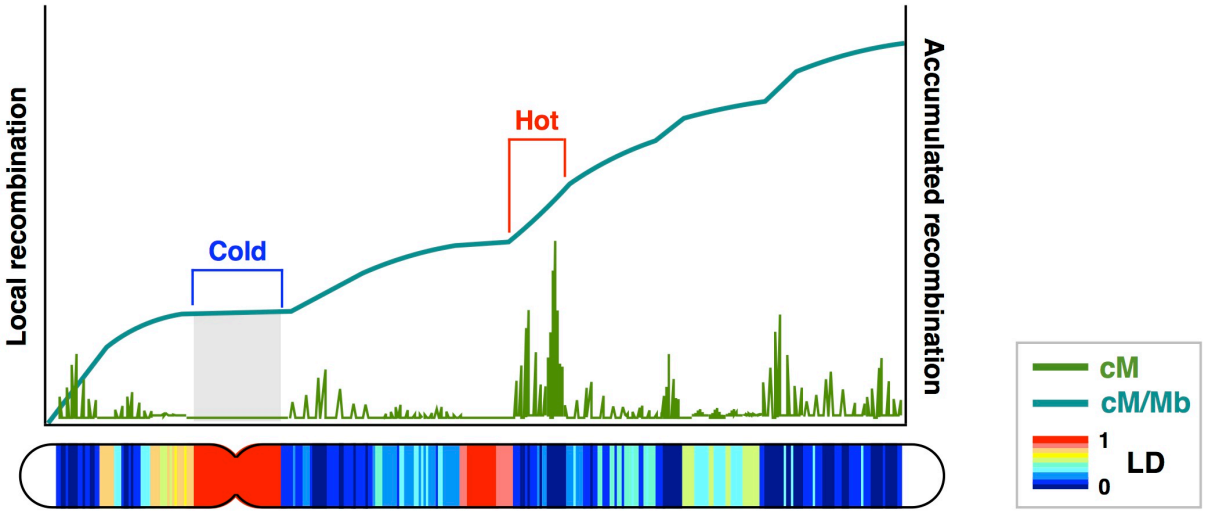


Figure 11: Recombination and its relationship with linkage disequilibrium. Local recombination (measured in cM) varies across a chromosome, with locations having low (cold) or high (hot) activity. This is reflected in flat and upslope regions, respectively, when considering the accumulated recombination (measured in cM/Mb and that corresponds to the total genetic distance). LD block are correlated with areas of low recombination such as the centromere Adapted from Phillips et al. (2012).

1.2.3.2.2 Non-recombining segments of the human genome

Due to their particular inheritance, both Y-chr and mtDNA are passed from one generation to the next without suffering recombination (**Figure 9**). For this reason, genetic variation in each of those non-recombining DNA segments is in complete linkage and so paternal and maternal haplotypes can be defined. Those are kept unchanged and maintain lineage records that are particularly useful in population genetics (see section 1.1.3).

The Y-chr is present in haploid state in male individuals. Although it is paired with a maternally inherited X-chr copy, those two DNA segments are not chromosome homologs. For this reason, it is expected that no Y-chr recombination will occur during meiosis. Nevertheless, recombination previous to chromosome segregation is an important stage of cell division and this explains the existence of the pseudoautosomal regions (PAR). PAR1 is a 2.6 Mb segment in the Yp telomeric end that suffers obligatory recombination with the X-chr in each meiotic division. This recombination event is considered an extreme example of higher recombination rates in shorter chromosome arms: in just 2.6 Mb this region has a genetic distance of 50 cM, nearly guaranteeing a crossover event (IHGSC 2001). On the other hand,

PAR2 is a 0.32 Mb segment in the Yq telomeric end but, contrary to the observed for PAR1, its recombination with the X-chr is not mandatory. The remaining Y-chr escapes meiotic recombination (known as non-recombining portion of the Y or NRY) and sequences in this region are considered to be truly male specific, including the sex-determining gene SRY.

When considering the mtDNA, although the necessary biochemical machinery to perform recombination is present in mitochondria, this could only happen if paternal inheritance of mtDNA occurs. Despite this has been shown to exist in humans (Ankel-Simons and Cummins 1996), paternal mitochondria are actively destroyed after fertilization (Sutovsky 2003, Sutovsky et al. 2004). Although possible, recombination is not likely to occur (Innan and Nordborg 2002, Pakendorf and Stoneking 2005) and so mtDNA should be considered a non-recombining molecule.

1.2.3.3 Human genome variation

The human genome is highly conserved between individuals but, with the exception of monozygotic twins, each individual has a unique genome. Differences can be found between single nucleotide positions to thousands or millions of base pairs. In fact, they can be classified into four major groups according to their size: i) a few base pairs, including base substitutions, small insertions or deletions or even the change in the number of repeats; ii) tens of base pairs to a few kilobases, which usually include large insertions, deletions or the expansion by several repeat units; iii) kilobases to megabases such as duplications, deletions and inversion of DNA segments; and iv) multi-megabase to whole chromosome scale, including chromosomal translocations or changes in ploidy (for example monosomy or trisomy). And while the majority of such differences can go unperceived as no visible effect arises from their variation, they are also responsible for the great variety of phenotypes – observable traits in a cell or organism that can go from armless visible characteristics such as hair or eye colour to serious diseases.

Several concepts regarding genome variation need to be defined. First, variation implies that two or more alternative forms of a particular gene or DNA sequence are present at a specific chromosomal location. Each of those forms is an allele and they are located at a particular locus (pl. loci). As the human nuclear genome is diploid, for a particular locus an individual can be homozygote (it has the same allele at each copy of the diploid locus),

heterozygote (it carries two different alleles) or even hemizygote (it only has one allele as a result of normal chromosome constitution; for example, X-chr loci in male individuals). Second, variation points can be called in a variety of ways, the most common being mutations, variants or polymorphisms. Mutation is commonly used to describe a de novo change, that is, a newly arising variation point, or a variation that is known to cause a disease. Variant is a more general term, although it sometimes refers to differences between genome copies that are present in a given population at a frequency lower than 1%. When the variant is present at frequencies higher than 1% it is called polymorphism.

1.2.4 Short polymorphisms in the human genome

The present work will focus on shorter variations, with particular interest on those located in the nuclear genome. Specifically, three different polymorphism classes (**Figure 12**) will be considered:

- *Length polymorphisms* – produced by the insertion or deletion of one or more nucleotides. Generally they can be found in repetitive regions like the mini or microsatellites (which can be differentiated by the number of copies of the repetitive unit, see section 1.2.4.1).
- *Sequence polymorphisms* – they result from the substitution of one or more nucleotides in the DNA sequence. Point mutations are the most frequent in the human genome (see section 1.2.4.2)
- *Insertion/deletion polymorphisms* – although they represent the insertion of deletion of one or more nucleotides, being equivalent to length polymorphisms, they can sometimes be considered to be sequence polymorphisms, especially when only one nucleotide is deleted (see section 1.2.4.3).

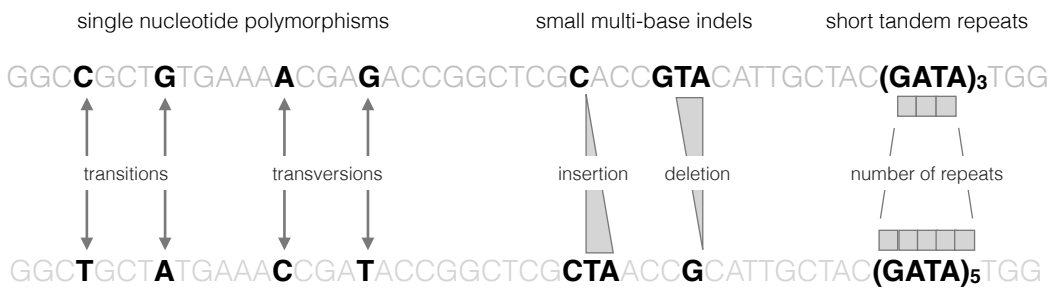


Figure 12: Different types of DNA polymorphisms.

Each group will be described in detail in the following sections.

1.2.4.1 Short tandem repeats

Repeated DNA sequences arranged in tandem, normally known as variable number of tandem repeats or VNTRs, are a common variation in the human genome. Such markers usually have high mutation rates, which imply that they have identity by state, that is, individuals can have the same allele although their common ancestor has a different allelic state. Depending on the number and size of the repeat units, VNTRs can be classified into different groups: microsatellites, minisatellites and satellites (arranged in increasing order of repeat unit length and repeat copy number). Other specialised repeats can be found in the genome, the best example being the hexanucleotide repeat TTAGGG found in the telomeres. Despite their small size they are not classified as microsatellites since they can reach 1,500-2,500 repeats.

Although the main subject of the present work are the binary markers described in the following sections, microsatellites (also known as short tandem repeats or STRs) have an important role in particular applications, specially due to their reduced copy number (usually less than 50 copies) of short repeat units (1-7 bp). STRs with 1 bp repeats correspond to poly(N) structures, with poly(A) and poly(T) being the more common. They are not generally used because they are highly susceptible to errors and alleles are difficult to differentiate. Di-nucleotide microsatellites were widely used in the past because they could be easily found but their analysis was prone to artefacts. On the other side, use of microsatellites with 5-7 bp repeats is limited as they are rare in the genome. STRs with 3-4 bp repeat units are, at the present, the most commonly used.

STR genotypes can be represented by the number of repeat units observed at a particular locus. For example, if an individual is homozygote 10 for locus X, this means that it has 10 repeat units in both chromosomes. Each allele can be represented by the repeat unit sequence followed by the number of repeats. In the given example, if locus X is a tetranucleotide with repeat unit GATA, the studied individual has two copies of allele $(GATA)_{10}$. Furthermore, microsatellites can present another variability level: differences in the repeat structure for the same allele, that is, some STRs are said to have a complex repeat structure. For example, locus X can have two alleles' 10: $(GATA)_{10}$ and $(GATA)_5TATA(GATA)_4$. Those have the

same length but their sequence composition is different. To differentiate such alleles the use of the apostrophe punctuation mark (') is recommended: allele 10 - $(GATA)_{10}$; 10' - $(GATA)_5TATA(GATA)_4$. Complex alleles can also include different length repeats in their structure.

Independently of the repeat unit size and structure, STRs are common features in the human genome – estimates based in the microsatellite catalogue[1] included in Phase III of the 1000 Genomes Project (described in detail in section 1.2.4.4) indicate that there are more than 670,000 sites. STRs are predominantly multi-allelic markers with new alleles being the product of high mutation rates associated with STRs – those have been estimated to be around 10^{-3} to 10^{-4} (Weber and Wong 1993, Brinkmann et al. 1998, Kayser et al. 2000b, Ellegren 2004, Sun et al. 2012). There are several mechanisms that explain STR mutation, including: i) unequal crossing over in meiosis; ii) retrotransposition mechanism; and iii) strand-slippage replication (Fan and Chu 2007). Such mechanisms are still not well understood but it is thought that the main mutation process giving origin to STRs is strand-slippage replication. In this, DNA polymerase slips in a repeat tract and, when continuing the replication process, it extends a misaligned sequence. If the mismatch repair process fails to fix the mutation, it originates the gain or loss of repeat units. The microsatellite evolution resulting from such mutation process (stepwise mutation model or SMM) was first modelled in 1973. The SMM is a simple model in which the length of the microsatellite varies by single units at a fixed rate independent of repeat length and with the same probability of expansion and contraction (Kimura and Ohta 1973, Ohta and Kimura 1973). But over the years some changes were done to this model: i) while it is true that for tetranucleotide loci most mutations involve the increase or decrease of a single repeat unit, this is not always true for dinucleotide loci; ii) simple repeat loci mutate faster than complex repeat loci, even if they have the same total length; iii) the higher the number of repeats, the higher the mutation rate; and iv) short allele tend to increase in size while longer alleles tend to lose repeat units (Ellegren 2000, Xu et al. 2000, Lai and Sun 2003, Sun et al. 2012). For this reason, the mutation model should be considered to be more dynamic than the over-simplistic SMM: length-dependent slippage leads to equilibrium between expansion and deletion errors of repeat units (Kruglyak et al. 1998, Eckert and Hile 2009).

[1] ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/strs/

1.2.4.2 Single nucleotide polymorphisms

Single nucleotide polymorphisms, or SNPs, are positions in the DNA sequence that can be occupied by different nucleotides in distinct homologous DNA sequences (Griffiths et al. 2005) as shown in **Figure 12**. There are two main processes originating base substitutions: nucleotide misincorporation due to DNA polymerase errors during replication or mutation caused by either endogenous or environmental factors such as UV radiation. Even though SNPs are extremely abundant, there are efficient systems that can detect and repair any error during replication (such as proofreading) or damage to the genome (such as mismatch or base-excision repair). For this reason, base substitutions have an overall low mutation rate: approximately $1.2x10^{-8}$ per nucleotide per generation (Nachman and Crowell 2000, Kong et al. 2012, Sun et al. 2012).

Another mutation process able to generate base substitutions is the methylation of CG dinucleotides. Such dinucleotides are mainly present in CpG islands, that is, regions of DNA, usually in the 5' region of a gene, where high numbers of CG dinucleotides are present and are target to DNA methylation, an important process for the control of gene expression. DNA methylation adds a methyl group ($-CH_3$) to the cytosine ring, originating 5-methylcytosine. Spontaneous or mutagen induced deamination in the CpG islands can have two effects: i) if a cytosine is deaminated, a uracil is produced, which can be efficiently repaired by base-excision; and ii) if a 5-methylcytosine is deaminated, a timine is originated, resulting in a G-T mispair that can either be repaired to a C or A, increasing the probability of mutations in CpG islands.

Base substitutions can be divided into two categories depending on the nature of the bases involved. When a purine base is exchanged by another purine or a pyrimidine by another pyrimidine, the change is called a transition (**Figure 13**). On the other hand, when a pyrimidine is exchanged by a purine, or vice versa, it is called a transversion (**Figure 13**). Interestingly, although each base has the double of possible transversions than transitions (**Figure 13**), the latter are three to ten times more common in the human genome than transversions (Nachman and Crowell 2000, Kong et al. 2012, Jobling et al. 2014). This fact is probably explained by differences in the misincorporation rates or in the detection and repair systems.

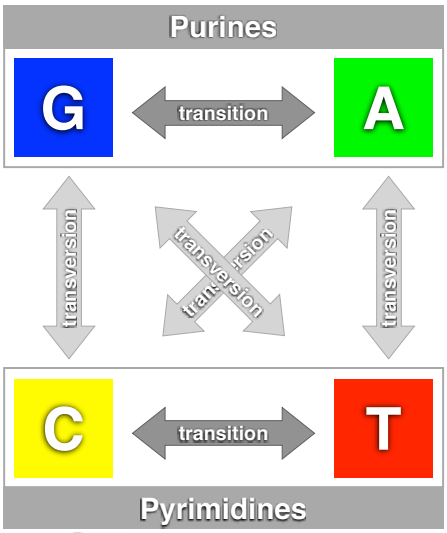


Figure 13: Possible base substitutions. Adapted from Jobling et al. (2014).

SNPs are usually described as bi-allelic since the great majority presents only two variants possible: ancestral and derived, the last being a consequence of a mutation of the prior. The reduced mutation rate implies that there is an extremely low probability that the same genome position mutates on two independent occasions to produce different derived alleles. But there are some known positions in the genome that have three or even four possible nucleotides at the same site (Phillips et al. 2004b). Furthermore, from the reduced mutation rate it is possible to assume that SNPs are more likely to be identical by descent, i.e, the presence of a certain polymorphic base in two individuals indicates that both inherited that base from a common ancestor (Jobling et al. 2014).

Despite the low mutation rate, SNPs are the most abundant polymorphism class in the human genome. In two of the major variation databases (described in detail in section 1.2.4.4) the current estimates are close to 90 millions SNPs. In Phase III of the 1000 Genomes Project, more than 78 million autosomal SNPs were detected[2]. On the other hand, dbSNP build 142 reports approximately 150 million SNPs, of which nearly 98 millions have been validated[3]. Validation status is important to guarantee that the SNP is real and unique, i.e., that it is not the result of a sequencing error and it is mapped only once in the genome (Phillips et al. 2004a).

[2] ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/README_phase3_callset_20150220
[3] http://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi?view+summary=view+summary&build_id=142

1.2.4.3 Insertion/deletion polymorphisms

Insertion/deletion polymorphisms, or Indels, correspond to positions of the genome where nucleotide bases are present or absent of the DNA sequence (**Figure 12**). They can be divided into two categories: the ones with several distinct alleles (multiallelic) or the ones only having two possible alleles (biallelic) (Weber et al. 2002, Edelmann et al. 2009). The latter can present a great length variation between both alleles in a way that the difference can reach tens or hundreds of kilobases (Weber et al. 2002).

When only one base is involved, the Indel can be included in the single nucleotide polymorphisms category (Jobling et al. 2014) but the insertion/deletion events occur less frequently than substitutions (Zhang and Gerstein 2003). Generally SNPs are ten times more frequent than Indels but this difference is relative and it can vary between loci depending of the context sequence (Nachman and Crowell 2000, Ebersberger et al. 2002). For example, it is known that the number of Indels found in regions with known function (in the majority of the coding DNA) or under the influence of evolutionary forces is very small (Clark et al. 2007). Nevertheless, the mutation rate of short Indels (by loci per generation) is very similar to SNPs – $2.3x10^{-9}$ (Nachman and Crowell 2000). As with SNPs, from the low mutation rate it is possible to assume that Indels are more likely to be identical by descent (Jobling et al. 2014).

It is known that Indels are abundant in model organisms' genome so the same is expected in the human genome. The study of Dawson et al. (2001) suggested Indels represent 18% of the human chromosome 22, from where it is possible to infer that those represent 16 to 25% of all polymorphisms in the human genome (Mills et al. 2006). The most common biallelic Indels are the ones with only a few nucleotides of difference between both alleles (Weber et al. 2002). Nachman and Crowell (2000) estimated that short Indels (1-20 bp) correspond to 10% of the changes in non coding sequences of the human genome. In Phase III of the 1000 Genomes Project (described in section 1.2.4.4), more than 3 million autosomal Indels were detected[4]. Other characteristic of Indels is the lower rate of deletion events when compared with insertions and the preference for smaller length differences between both alleles. The study of Zhang and Gerstein (2003) showed that the average length of insertions and deletions is approximately 4 bp with a higher number of 2 bp events observed.

[4] ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/README_phase3_callset_20150220

Indels, as nucleotide substitutions and genomic rearrangements, are one of the main mechanisms generating mutations and genetic defects and, for that reason, are very important for genetic variability and molecular evolution (Britten et al. 2003, Makova et al. 2004, de la Chaux et al. 2007) having an important role in the establishment and maintenance of the human genome diversity (Robledo et al. 2003). This polymorphism class can only be recognised by the detection of length differences when comparing various homologous DNA sequences (Pearce 2006). From the study of approximately 2000 Indels, Weber et al. (2002) concluded that the great majority had their origin after the divergence of the common ancestors of humans, chimpanzees and gorillas, although Indels mutation rate is similar in the primates X and Y-chrs (Weber et al. 2002, Sundstrom et al. 2003). This observation indicates there is no sex-bias and, for that reason, Indels mutation is the result of the combined action of recombination and number of cellular divisions (Sundstrom et al. 2003). Errors introduced during DNA synthesis are correlated with point mutations and their reported biological effects (Garcia-Diaz and Kunkel 2006). The study of Makova et al. (2004) reported evidences of distinct molecular mechanisms originating short and long Indels: the former being related with replication and the latter with recombination. The same study did not replicated those conclusions in rodents being, at the time of the study, the first example of sex-bias in mammals and, according to the authors, an evidence that supports the theory that considers Indels have their origin during the cellular replication phase (Makova et al. 2004). But there are other mutation mechanisms, such as DNA damage, that affect both genders in the same way (Sundstrom et al. 2003).

Another aspect to be considered is the Indel location. The ones in coding regions are subject to stronger selective pressures due to the possible impact in the amino acid sequence or in the gene properties (Taylor et al. 2004, de la Chaux et al. 2007). In fact, there are Indel hotspots, i.e., regions that contain a high number of Indels when compared to the rest of the genome, and that might correspond to recombination hotspots (Mills et al. 2006, Costantini and Bernardi 2009). Several of these regions also present a high number of SNPs and can be considered genetic variation regions (Mills et al. 2006). For this reason, some researchers suggest Indels increase the substitution rates in the surrounding regions (Tian et al. 2008, Chen et al. 2009). Generally, Indel-SNP correlation is considered to be indirect as it would be a response to the composition or structure of the DNA sequence or even a functional restriction of that region (Tian et al. 2008). Another hypothesis suggested and supported by

the study of Tian et al. (2008) is mutation induced by Indels, i.e., when present in heterozygous form, Indels increase the nucleotide substitution rate in adjacent positions. An interesting conclusion of the Costantini and Bernardi (2009) study is the relation between high level of insertion and deletions in isochores – large DNA regions with high GC content. On the other side, SNPs are uniformly distributed across those same regions, which can be explained by the origin of these polymorphisms: errors during DNA replication that are not affected by context sequence (Costantini and Bernardi 2009).

1.2.4.4 Online databases of human genetic variation

Online databases are a major resource for researchers as they include information that can help to locate and select the best markers for a particular application, but also include characterized samples that can be used as a reference.

The National Center for Biotechnology Information (NCBI) site is probably the most important and complete set of genomic databases available. It includes a vast collection of nucleotide, gene, protein, association between genotype-phenotype and inherited disorder databases, among others. Furthermore, complete *H. sapiens* genome sequence and that of over 1000 different organisms are also available at NCBI. Considering the applications described in this thesis, the NCBI dbSNP Short Genetic Variations database (http://www.ncbi.nlm.nih.gov/SNP/) is of great interest – it includes mapping information, molecular context, genotype and population specific frequency data for SNPs and, despite the name, STRs and Indels (Sherry et al. 1999, Sherry et al. 2001). *H. sapiens* dbSNP build 144 (available June 8, 2015) reported 149,735,377 refSNP clusters. A reference SNP (refSNP), represented by an *rs* identification number, is a cluster of identical SNPs (or one of the other polymorphic classes included) submitted to dbSNP by different laboratories. As each submission corresponds to a unique *ss* identification number, an *rs* number can have several associated *ss* numbers. Of those nearly 150 million refSNPs, 97,535,033 have been validated – as mentioned in section 1.2.4.2, validation status is important to guarantee that a SNP is real and unique (Phillips et al. 2004a). To be considered validated, refSNPs have to fulfil at least one of the following conditions: i) multiple independent submissions; ii) frequency or genotype data shows that minor alleles have been observed in at least two chromosomes; iii)

submitter confirmation; iv) all alleles have been observed in at least two chromosomes; v) genotyping by the HapMap project; or vi) sequenced by the 1000 Genomes project.

The HapMap Project (IHMC 2003), originally developed to determine the haplotype structure of the human genome and discover sequence variants associated with common diseases, was later expanded to include the characterization of all common human sequence variation. The project was divided into three phases with different objectives. Phase I analysed approximately 1 million SNPs in 269 samples (IHMC 2005) – a comprehensive review of Phase I data and the conclusions that can be made about the structure of human genetic variation and the molecular and evolutionary processes that shape it was published by McVean et al. (2005). Phase II extended the initial study through the genotyping of further 2.1 million SNPs on the same individual samples as Phase I (IHMC et al. 2007). Phase III explored 1.6 million SNPs in 1,184 individuals belonging to eleven populations from five major groups: African, European, East Asian, South Asian and Native American (IHMC et al. 2010). Data has been made publicly available, including SNP frequencies, genotypes and haplotypes (http://hapmap.ncbi.nlm.nih.gov/). Furthermore, a genome browser is also available, allowing for researchers to explore regions of interest, scan for associations with nearby SNPs and links to dbSNP for more information about each marker.

The 1000 Genomes Project (http://www.1000genomes.org/) goal is "*to discover, genotype and provide accurate haplotype information on all forms of human DNA polymorphism in multiple human populations. Specifically, the goal is to characterize over 95% of variants that are in genomic regions accessible to current high-throughput sequencing technologies and that have allele frequency of 1% or higher (the classical definition of polymorphism) in each of five major population groups (populations in or with ancestry from Europe, East Asia, South Asia, West Africa and the Americas). Because functional alleles are often found in coding regions and have reduced allele frequencies, lower frequency alleles (down towards 0.1%) will also be catalogued in such regions*" (1000GPC 2010). An initial phase of the project included the analysis of 179 individuals from the four original HapMap populations (Yoruba from Ibadan, Nigeria – YRI; European ancestry in Utah – CEU; Han Chinese individuals in Beijing – CHB; and Japanese individuals in Tokyo – JPT) and discovered approximately 9 million novel SNPs, several Indels and identified some large structural variants (1000GPC 2010). Phase I results, which included 1,092 individual genomes from 14 populations, reported 38 million SNPs and 1.4 million

short Indels and showed different populations have distinct profiles of rare and common variants (1000GPC 2012). The last phase of this project, Phase 3, released data for 2,504 individuals from 26 populations from five major groups (31 related individuals were removed from the main release). It includes 78,136,341 SNPs and 3,135,424 Indels, among other polymorphic sites, and all data is available in a public database and accessible through a browser that includes, when available, information about: genomic context, genes and regulation, population genetics, individual genotypes, linkage disequilibrium, phenotype data, phylogenetic context and flanking sequence.

Other databases exist, some of which have more specific applications. For example, ALFRED (Allele FREquency Database, http://alfred.med.yale.edu/) includes frequency data for 663,887 loci in 723 human populations. STRBase (Short Tandem Repeat DNA Internet DataBase, http://www.cstl.nist.gov/strbase/index.htm) is focused in the forensic application of STRs and, among other information, provides details about their sequence and population data. The Mammalian Genotyping Service of the Marshfield Centre for Medical Genetics (http://research.marshfieldclinic.org/genetics/GeneticResearch/compMaps.asp) provides mapping information for several polymorphisms and is very useful to locate STR positions (Broman et al. 1998). Furthermore, it includes genotype data for SNPs, Indels and STRs (http://research.marshfieldclinic.org/genetics/genotypingData_Statistics/humanDiversityPanel.asp). Regarding uniparental markers, EMPOP is the mitochondrial DNA control region database (http://empop.org/) and includes haplotypes from all over the world. On the other hand, YHRD is the Y-chromosome STR haplotype reference database (https://yhrd.org/) and, despite the name, it includes forensically used Y-STRs and also phylogeny information defined by Y-SNPs.

As mentioned in the beginning of this section, characterized samples are sometimes needed as a reference to perform statistical analysis of a profile. Collection of such data usually requires locus-by-locus scrutiny of one or even several databases. The set of web browsers SPS (SNPs for Population Studies) makes this task much more straightforward for any number of SNPs as well as forensically relevant STRs and Indels. SPS allows genotype download (except for STRs) and generates common population genetics indices such as allele frequencies, heterozygosity, F_{ST} or *In*. SPS comprises the following genomic variability browsers:

- *SPSmart home* (Amigo et al. 2008b) http://spsmart.cesga.es/ – simple pre-processing engine that includes five different population based genotype databases: i) 1000 Genomes Phase I May 2011; ii) HapMap Release #28; iii) Perlegen complete data set; iv) HGDP-CEPH Stanford study; and v) HGDP-CEPH NIH-Michigan study. Of all databases included in *SPSmart*, 1000 Genomes and HapMap are of most utility as they have more markers and larger sample sizes (including admixed ancestry populations). This is particularly true for 1000 Genomes data – the ENGINES browser (http://spsmart.cesga.es/engines.php?dataSet=engines) allows scrutiny of SNP variation (down to a minor allele frequency of approximately 1%) of all sites found from 1092 complete genome sequences (Amigo et al. 2011). In contrast, the HGDP-CEPH samples 1,064 individuals from 52 populations and it is widely used in population genetics studies (Cann et al. 2002, Rosenberg 2006). The geographic coverage is patchy in certain regions but all continents and all major genetic ancestry groups defined by studies of variability are represented. Smallest sampling is 28 Oceanians from two populations and just six San from Namibia. *SPSmart* provides genotype data from 650,000 SNPs (obtained with Illumina 650K arrays (Li et al. 2008)) for each HGDP-CEPH sample in the Stanford University study of this panel.
- *SPSmart SNPforID 52-plex and 34-plex variability browser* (Amigo et al. 2008a) http://spsmart.cesga.es/snpforid.php – includes HGDP-CEPH and user submitted data for the SNP*for*ID *52-plex* individual identification (Sánchez et al. 2006) and SNP*for*ID *34-plex* ancestry informative marker sets (Phillips et al. 2007b, Fondevila et al. 2013);
- *pop.STR* (Amigo et al. 2009) http://spsmart.cesga.es/popstr.php - includes HGDP-CEPH data for 60 STRs including commonly used markers in forensic analysis, miniSTRs and ancestry informative STRs;
- *forInDel* (Santos et al. 2015b) http://spsmart.cesga.es/forindel.php includes HGDP-CEPH data for a set of 46 ancestry informative Indels (Pereira et al. 2012b).

1.2.4.5 Typing methods

Nowadays, DNA variation is detected using a range of techniques, the great majority being based in the polymerase chain reaction (PCR). But in the past, before the PCR technique was developed, DNA variation was also studied, mainly through the use of specific restriction enzymes that cleave the DNA molecule when they recognize a particular target sequence. While point mutations can induce the failure or change of a particular cleavage site, length mutations in between two restriction sites will change the size of fragment. Independently of the underlying reason, DNA variants induced changes in the restriction fragment length and were called restriction fragment length polymorphisms (RFLPs). But there were a lot of limitations associated with RFLP detection, the main being the misinterpretation of the obtained band pattern due to possible mutations. With the development of PCR, specific typing techniques appeared over the years.

1.2.4.5.1 Amplifying the signal: the polymerase chain reaction

The impact of PCR on biological and medical research has dramatically increased the progress rate in the study of genomes and gene action. This has created a revolution in the study of infectious diseases, cancer diagnosis, paternity testing and identification of forensic material. It is a fast, simple and low-cost technique that allows the isolation and amplification of virtually any gene of any organism being, at the present time, one essential step of the sequencing genomes projects.

The truth is that PCR is a very complex process with a great power and versatility for the manipulation and analysis of DNA. In the last three decades, since its description by Mullis et al. (1986), the PCR technique has been modified to increase its utility and versatility. In summary, PCR is a reaction that promotes the exponential amplification of specific DNA sequences through the use of specific oligonucleotide primers, a thermal cycling protocol, and a thermally stable DNA polymerase. The use of short oligonucleotide primers (usually with 15-30 nucleotides) allows the amplification of target regions of interest from the genomic DNA template. This amplification occurs through the sequential repeat of three thermal stages: i) heat denaturation of the template DNA at approximately 94-95ºC; ii) primer annealing to the target sites in the single-stranded template – this is possible through reduction

of the reaction temperature to approximately 50-60ºC; and iii) extension of new DNA strands complementary to the template – this step is usually performed at 72ºC and a thermostable DNA polymerase is used to synthesize the new strand through the addition of deoxynucleotides (dNTPs) to the primer 3' end onwards. These three steps are repeated a number of times (usually 25-30 cycles) and, in each of them, both the original genomic DNA and the synthesised amplicons (DNA sequences amplified in a specific PCR reaction) will be used as a template in the following cycles (**Figure 14**). Furthermore, other components such as a buffer and Mg^{2+}, are included in the PCR reaction to guarantee the enzyme activity.

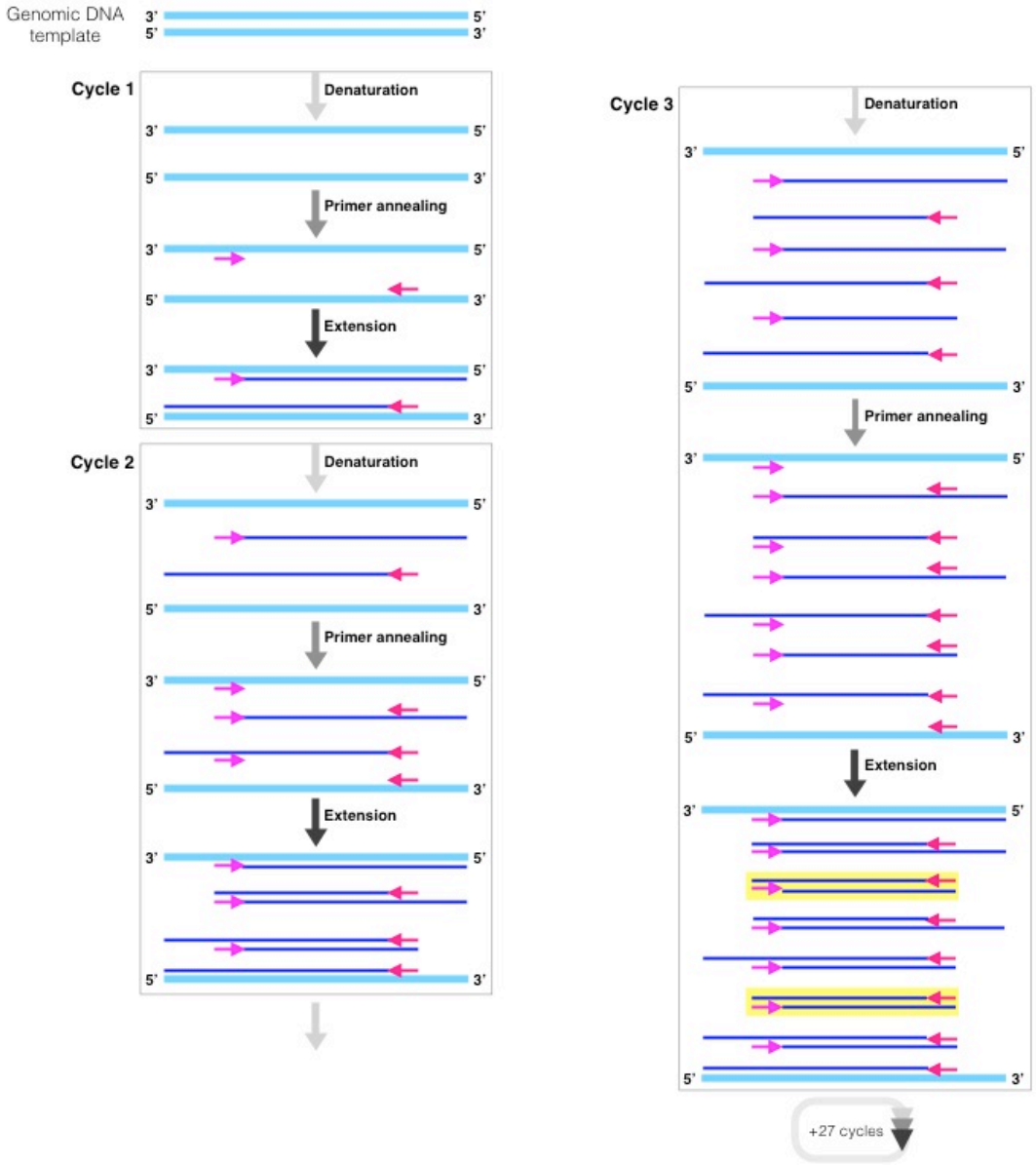


Figure 14: Polymerase chain reaction (PCR). The first three amplification cycles are represented. After denaturation of the double-stranded DNA, primers (pink and magenta arrows) anneal to the target region and a thermostable DNA polymerase extends the sequence. Amplicons delimited at both ends by the primers (highlighted in yellow) only start to accumulate from cycle 3 onwards. Adapted from Jobling et al. (2014).

Theoretically, that is, assuming perfect conditions and 100% of PCR efficiency, the number of target copies would double in each cycle. A 30 cycle PCR as the one shown in **Figure 14** would yield 2^{30} copies of the target sequence. But there are several factors that reduce the PCR yield: i) not all double-stranded DNA molecules denature at each cycle; ii) primer binding efficiency is not 100% so they do not bind to every DNA molecule at each cycle; iii) DNA polymerase loses efficiency at every cycle due to heat denaturation – this affects its ability to complete the extension of the full target sequence; iv) dNTP concentration in the PCR medium decreases at every cycle – as they are the building blocks of the new DNA chains, dNTP concentration is one of the main limiting factors. Nevertheless, PCR is capable of producing enough copies of the target region, allowing its detection and analysis of the sequence or any variation present.

1.2.4.5.2 Targeted typing of short polymorphisms

At the present, typing of short polymorphisms is usually preceded by a multiplex PCR reaction that allows the co-amplification and detection of several target sequences through the use of a set of multiple primer pairs. There are different detection methodologies that can be applied depending on the polymorphism type being analysed but the great majority is based on fragment separation by capillary electrophoresis (CE) followed by fluorescence detection. CE, being an electrophoresis-based technique, separates DNA molecules according to their size by using a voltage gradient: DNA molecules are negatively charged so they migrate towards the positive electrode. In CE the electrophoretic separation occurs inside a fine capillary filled with a separation matrix. Towards the end of the capillary, a laser excites the fluorescent dyes attached to the fragments and a charged coupled device (CCD) detects the fluorescence emitted. Specific software is used to convert the fluorescence signal into digital data and fragment size in base pairs is estimated through comparison with a size standard (fragments with known size labelled with a specific dye). The resulting data is presented in an electropherogram – a plot that contrasts fragment size with signal intensity. The latter is expressed in relative fluorescence units (RFUs) and is associated to the amount of DNA added to the amplification reaction. The existence of a variety of dyes that fluoresce at different wavelengths allows the simultaneous detection and analysis of several loci in one capillary injection, that is, in an electropherogram each peak corresponds to multiple DNA

fragments of equal size in base pairs – same size fragments resulting from the amplification of distinct loci can be discriminated by different fluorescent dyes. Nevertheless, the intensity of the emitted fluorescence is different for each dye and this has implications in the optimization process of any assay, especially for SNP detection (as described later in this section).

When analysing STRs or Indels, both being length polymorphisms, PCR amplicons of different sizes are generated at each loci depending on the number of repeat units or inserted/deleted bases, respectively. When performing the multiplex amplification of the loci of interest, one primer from each pair is labelled with a fluorescent dye at its 5' end (**Figure 15**). This allows the simultaneous analysis of several polymorphic markers at the same time and also a simple workflow: PCR is immediately followed by CE analysis – PCR-to-CE. As the PCR yield is also dependant on the initial template amount, the PCR-to-CE methodology maintains the relationship between initial DNA input and peak height. Together with the fact that, for each locus, alleles are detected using the same dye, this implies that alleles from the same locus in a heterozygous sample will have approximately the same peak height (**Figure 15**). Furthermore, adjusting primer concentration of the weakest markers can easily solve yield differences between loci due to distinct dye intensities.

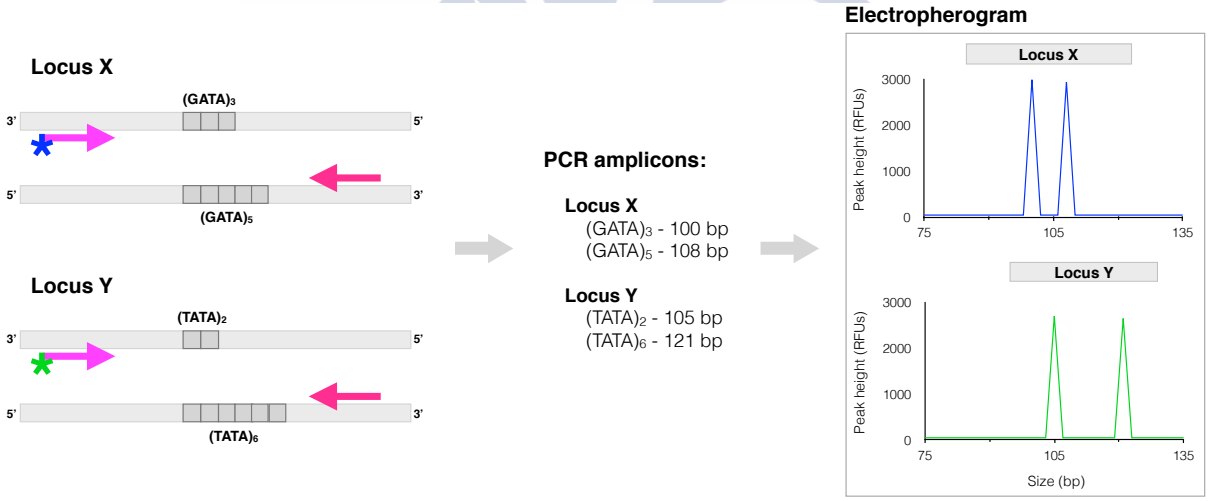


Figure 15: Fragment typing through direct PCR-to-CE. Example electropherogram obtained from the simultaneous PCR amplification of two loci (X and Y). All amplicons are within the same size range but the use of PCR primers labelled with different fluorescent dyes allows their discrimination. The same principle applies to Indel typing.

Contrary to STRs and Indels analysis where different alleles can be distinguished by size selection, SNPs cannot be resolved by PCR-to-CE since possible alleles at a particular locus will generate amplicons with the same size. For that reason, several typing methodologies were developed for SNP typing. As none can be considered completely ideal, SNP selection

should take into account the different methodologies available and the choice of the appropriate technique depends on a set of requirements defined by the user. Nevertheless, the majority of SNP typing reactions can be included in one of the following groups, based on their molecular mechanisms: i) specific allele hybridization; ii) primer extension; iii) oligonucleotide ligation; and iv) invasive cleavage (Gill et al. 2004, Sobrino et al. 2005). The works described in the present thesis used a primer extension technique, specifically minisequencing.

The minisequencing reaction, or single base extension (SBE), allows the determination of the alleles at a polymorphic base through the addition, using DNA polymerase, of one terminator dideoxynucleotide (ddNTP) complementary to the target base. The lack of the hydroxyl group in the 3' carbon of the deoxyribose implies that ddNTPs cannot establish a phosphodiester bond with a new nucleotide. For this reason, the synthesis of a growing DNA chain will be terminated after the incorporation of a ddNTP. In this reaction the primer attaches to the target DNA in a position immediately adjacent to the SNP in study and the method is based on the high fidelity of nucleotide incorporation of the DNA polymerase (Budowle 2004, Sobrino et al. 2005).

One of the most common minisequencing techniques is electrophoresis and fluorescence detection: SNaPshot® (Life Technologies, LT). The SNaPshot® multiplex SBE reaction uses ddNTPs labelled with specific fluorescent dyes as shown in Table 2. A non-marked primer aligns with the 3' end immediately upstream of the polymorphic position and is amplified through the addition of a single ddNTP (**Figure 16**).

Table 2: Dye label and colour of the analysed data of each ddNTP in the SNaPshot® assay.

ddNTP	Dye label	Colour of analysed data
A	dR6G	Green
C	dTAMRA	Yellow/Black
G	dR110	Blue
T	dROX	Red

As with the direct PCR-to-CE methodologies, multiplex PCR reactions allow the detection of several target sequences through the incorporation of a set of multiple primer pairs. This is possible due to the partial separation of the minisequencing products by the use of non-complementary tails with different sizes at the 5' end of the SNaPshot primers (**Figure 16**). According to the manufacturer it is possible to perform 10-plex reactions but bigger

multiplexes have already been designed (Budowle 2004, Yang and Rothman 2004, Sobrino et al. 2005, Sánchez et al. 2006, Phillips et al. 2007b). The products are then separated by electrophoresis in an automatic capillary sequencer (Budowle 2004, Sobrino et al. 2005). With this system genotypes are defined by colour and each locus by the position in the electropherogram. But, as mentioned before, each dye has a different emission intensity – in the case of the SNaPshot® dyes the theoretical proportion is 4G:2A:1C:1T, that is, G peaks are two times higher than A peaks and four times higher than C or T peaks. Contrary to fragment analysis, alleles are interrogated by the same primer but with different dyes, and for this reason the emission intensity cannot be adjusted.

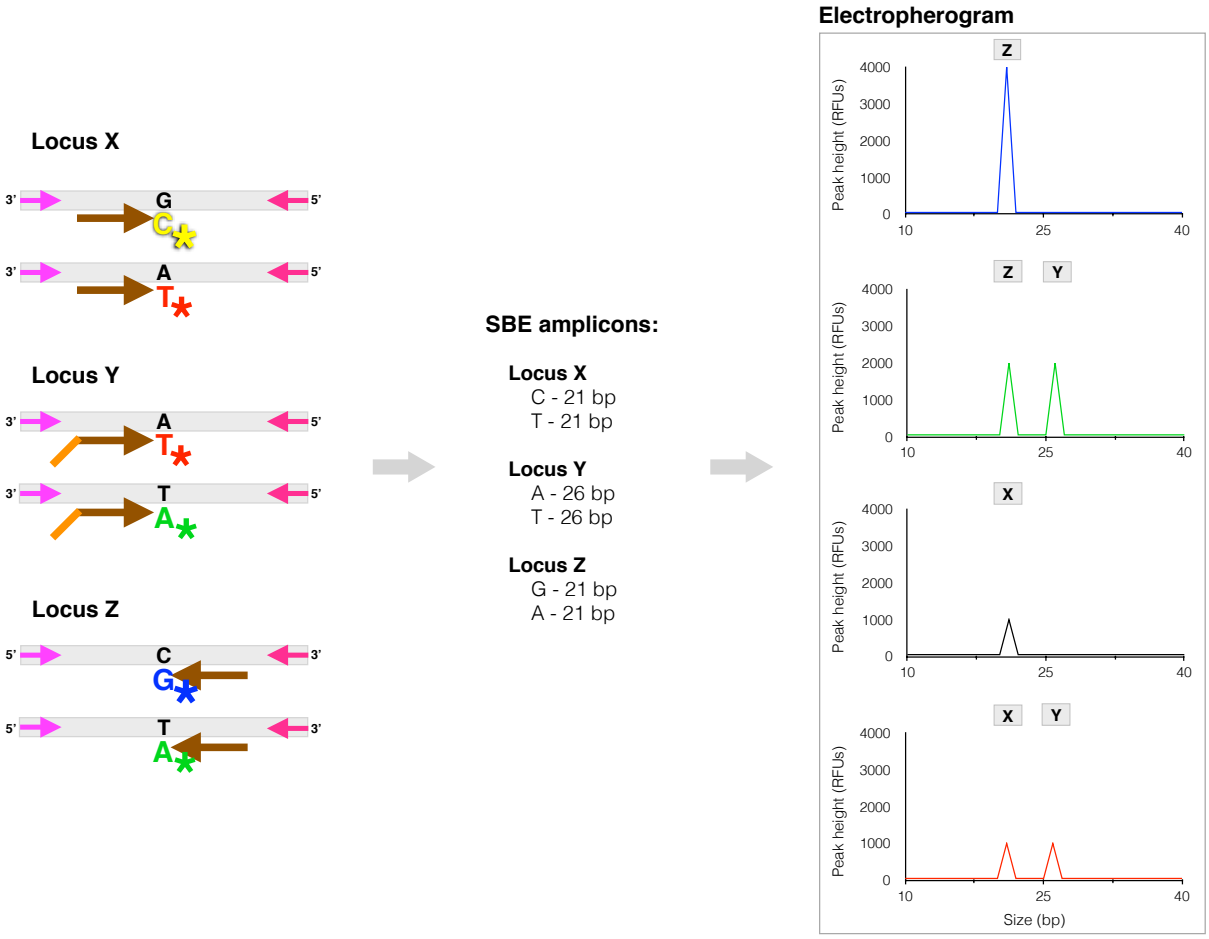


Figure 16: SNaPshot® amplification principle. Example electropherogram obtained from the simultaneous PCR amplification of three loci (X, Y and Z). After the initial PCR to delimit and amplify the target SNPs (using primers represented by the pink and magenta arrows), an unlabelled primer anneals to the 3' end immediately upstream of the SNP position (brown arrows). Size of the SBE amplicon defines which locus is being analysed - different amplicons can be differentiated by use of non-complementary tails (orange segments) at the 5' end of each primer. The use of fluorescently labelled ddNTPs (with different colours) that attach to the polymorphic base permits to distinguish each possible allele.

During the described procedures some treatments with restriction endonucleases need to be performed. This is due to the fact that non-digested DNA residues can interfere with the genotyping fidelity or with the amplification reaction (Ballantyne et al. 2007b). For mutation detection, SNP genotyping reactions pure digestion products are needed, with no traces of undigested DNA. But the digestion does not always occur with 100% efficiency due to the incomplete activity of the restriction enzymes, contamination by the template DNA or sub-optimum reaction conditions.

The multi amplification steps together with different dye intensities disrupt the proportion between initial DNA and final peak heights, which is a disadvantage for some applications (see section 1.2.4.6). Despite that, SNaPshot® is a simple and robust system and since PCR, electrophoresis and fluorescent detection can be achieved using existing equipment in the majority of the laboratories, no new equipment or expertise is required for implementation (Budowle 2004).

1.2.4.5.3 Sequencing

Sequencing methodologies were designed to determine the nucleotide order in a DNA molecule. Initially, such techniques were restricted to regions preceding or following a gene coding for a known protein – from the amino acid sequence, and using the genetic code described in section 1.2.2, the corresponding DNA sequence was inferred and used to design a complementary primer that could then be extended by a polymerase (Padmanabhan and Wu 1972, Wu 1972).

But sequencing methodologies can also be used to interrogate particular positions of the genome, that is, through sequencing of a certain region it is possible to genotype target polymorphisms and also to obtain information about other variability present in the flanking region. This implies that, as the targeted typing methods described in the previous section, an initial PCR amplification (which uses specific oligonucleotide primers) is included in the process. For this reason, targeted sequencing methods require the previous knowledge of the genome sequence – the complete DNA sequence of a genome can be determined by whole genome sequencing methods. Among several techniques available, shotgun sequencing is of particular interest as it was used to determine the initial human reference sequence (IHGSC 2001, Venter et al. 2001). At the time, this technique had been in place for two decades

(Staden 1979) but the developments needed for sequencing large genomes only occurred in the 1990 decade (Edwards et al. 1990, Edwards and Caskey 1991, Roach et al. 1995), being first applied to obtain the *Haemophilus influenzae* bacteria (Fleischmann et al. 1995) and the *Drosophila melanogaster* (Adams et al. 2000) complete genomes. In this technique no specific primers are used. Instead, the complete genome is randomly sheared into short fragments that are then sequenced using a chain termination method (such as the Sanger sequencing). Additionally, each fragment can be inserted into a bacterial artificial chromosome (BAC) where it is amplified. After sequencing, several overlap reads are obtained which allows their *in silico* assembly into a final genomic sequence.

In this work two sequencing techniques are of special interest: Sanger sequencing and next-generation sequencing (NGS).

Sanger sequencing, also known as chain termination sequencing, was first described in 1977 and is based on the incorporation of ddNTPs to a growing DNA chain (Sanger et al. 1977). The automation of Sanger sequencing using labelled ddNTPs and CE (Smith et al. 1986) increased the throughput and reduced the sequencing cost, this being a major factor that made it possible to determine the human reference genome sequence. This technique can be described as an hybrid between the previously described PCR and minisequencing: i) a DNA polymerase is used to synthesize a new strand from a single specific primer – the maximum length of the obtained sequence is defined by the extension time and not by the use of a second primer to delimit the region (this considering that nucleotides on solution are not a limiting factor); ii) as in a minisequencing reaction, ddNTPs are used to terminate the chain synthesis, but in Sanger sequencing unlabelled dNTPs are combined with lower concentrations of labelled ddNTPs, which allows the extension to occur until a ddNTP is added. After a set number of cycles, several sequence fragments are generated – they have different sizes and are labelled with different fluorochromes depending on the base that terminated the extension (**Figure 17**). CE analysis of such fragments maps the position of each base and a sequence can be inferred from the resulting electropherogram (**Figure 17**). Sanger sequencing allows the sequencing of fragments of approximately 1,000 bp with more than 99.99% accuracy (Shendure and Ji 2008). And although the next-generation sequencing methodologies that are about to be described have a much higher throughput, Sanger sequencing is still the standard technology used to validate newly found variants.

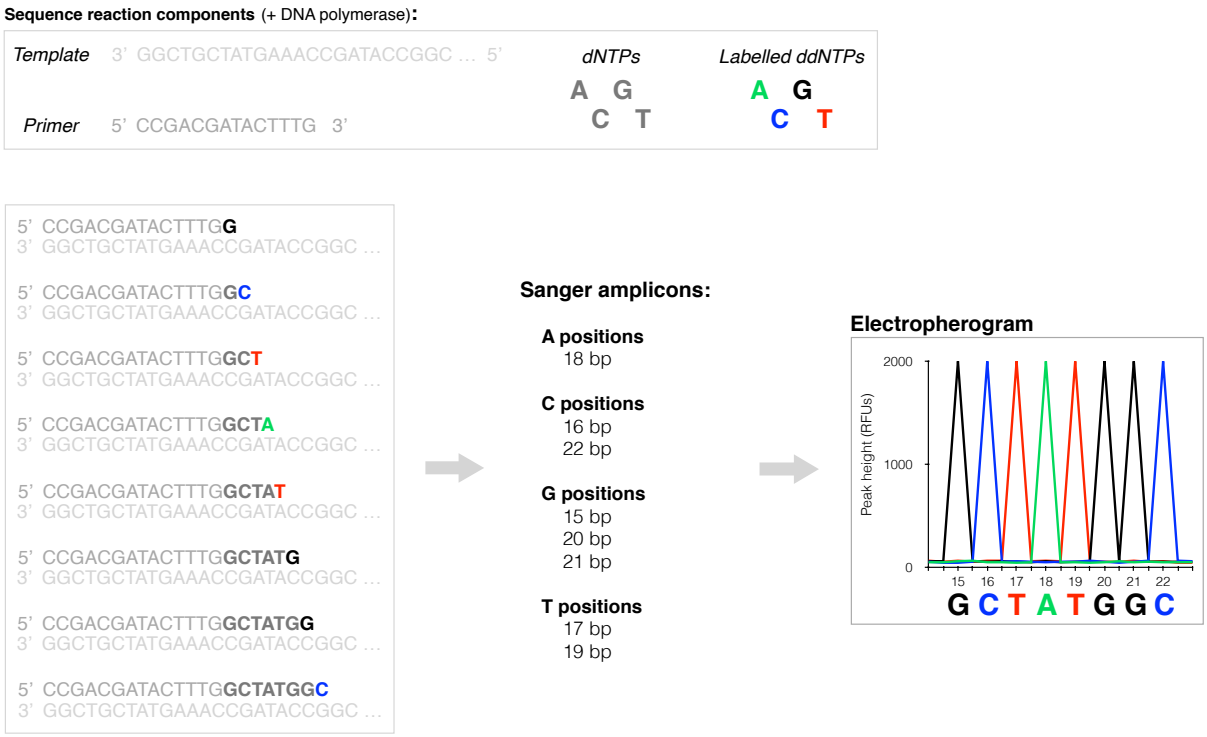


Figure 17: Sanger sequencing principle. This is a PCR based technique where a DNA polymerase is used to extend a single primer. The use of unlabelled dNTPs combined with lower concentrations of labelled ddNTPs allows extension to occur until a ddNTP is added to the growing chain. This generates fragments with different sizes labelled with distinct fluorochromes depending on the base that terminated the extension. CE analysis of such fragments maps the position of each base and the template sequence can be inferred. Adapted from Jobling et al. (2014).

Several NGS methods exist, each platform being based on different principles: i) polymerase-based sequencing by synthesis; ii) ligation-based sequencing; iii) sensing of pyrophosphate release on base incorporation; or iv) semiconductor sensing of H^+ release on base incorporation. Several articles reviewing and contrasting different approaches and techniques available have been published, especially in the last five years (Shendure and Ji 2008, Pettersson et al. 2009, Liu et al. 2012b, Quail et al. 2012, Ross et al. 2013, Buermans and den Dunnen 2014, van Dijk et al. 2014).

All such methods have in common the initial step – the library preparation. In this, DNA fragments are end-repaired so that an adapter sequence, i.e., a known short sequence that will then be used as primer target site, is attached. Libraries can be prepared directly from fragmented genomic DNA but some NGS platforms use a targeted approach for specific applications: a multiplex PCR reaction that can include thousands of primer pairs is performed to enrich the quantity of DNA molecules carrying the target polymorphic sites before adapters are ligated. Furthermore, higher throughput levels can be achieved by the

simultaneous analysis of different samples – this is possible through the use of barcoded adapters (those contain specific small sequences that allow a particular PCR product to be assigned to a specific DNA sample within a multiplexed sequencing run).

The works described in the present thesis used a semiconductor system (Rothberg et al. 2011), specifically the Ion Torrent™ Personal Genome Machine® (PGM). The Ion PGM™ workflow can be divided into five main stages (**Figure 18**): i) DNA extraction; ii) library preparation; iii) template preparation; iv) sequencing; and v) data analysis. DNA extraction is a well-establish procedure in any laboratory that performs genetic analysis. Nevertheless, it is important to mention that it is possible to analyse paraffin embedded samples, which is of great interest for forensic and clinical applications (see section 1.2.4.7). The next step is to prepare a targeted library (**Figure 18**) – using the Ion AmpliSeq™ library preparation kit and a pool of primers that target specific polymorphic sites the regions of interest are amplified. After PCR, partial digestion of the primers[5] prepares both ends of the amplified fragments for adapter ligation. Adapters are called "A" and "P1" – when barcoded adapters are used, the A adapter includes the sample-specific barcode and it is called "X" adapter. After ligation, the oligonucleotide pool comprises amplicons with an X adapter at one end and a P1 adapter at the other. Each sample is then purified, quantified and adjusted in concentration to prepare an equimolar combination of all DNA pools, creating a single library of fragments. One advantage of starting library preparation with PCR enrichment is that established multiplexes can be used to prepare the amplicon pool. This has recently been shown to be possible (Daniel et al. 2015).

In the following stage, the library is amplified by clonal emulsion PCR (emPCR) – the amplification occurs inside small aqueous droplets dispersed in an oil phase in the presence of small particles (Ion Sphere™ particles, ISPs) designed to attach library DNA fragments through their P1 adapters (**Figure 18**). Careful adjustment of the ISP:library proportion ensures that each droplet contains a single ISP and a single library fragment. This guarantees that after emPCR each ISP will have identical amplified fragments in its surface – those are designated as monoclonal. Polyclonal ISPs, that is, when more than one library fragment is amplified in a single ISP, cannot be differentiated from monoclonal ones at this stage but their sequencing results will be considered abnormal and will be filtered out during data analysis.

[5] The primers are designed by specific software (Ion AmpliSeq™ Designer, https://www.ampliseq.com/browse.action) and include proprietary 3'/5' sequence tags that help to optimize the ligation step.

Amplification success can be verified through quantification of specific ISP and A adapter signals. After an enrichment step that isolates ISPs with amplified target, those are placed in a semiconductor-based chip ready for sequencing (**Figure 18**). Each chip contains more than one million wells but only one ISP can enter each one – the resulting sequence from each well is called read. Sequencing proceeds as a series of sequential base flows where incorporation of a new base is registered as a pH change due to H^+ release. The accompanying voltage spikes are detected by a microprocessor at the chip base and are used to create an ionogram of sequential signals. The microprocessor registers multiple base incorporations proportionately, so signal strength corresponds to the number of bases added in each flow step from polybase segments.

Finally, obtained data has to be analysed, this being the most challenging stage. The Torrent Suite™ software is used for data analysis, from base calling (when the ionogram is converted into a base sequence) to the alignment against a reference sequence so that any variants present are typed and reported. Depending on the objectives, Torrent Suite™ includes several plugins, Torrent Variant Caller being the key for variant detection – it reports any variants present in the sequence results. Variant calls are based on posterior probabilities calculated for each possible genotype in a similar way to the widely used variant caller software included in GATK (McKenna et al. 2010). Posterior probabilities are computed using genotype likelihoods, calculated from Phred read quality scores, and prior probabilities, taking into account read depth and minimum allele frequency thresholds. This process reports a quality score with values ranging from 0 to several thousand. For a SNP genotype to be called it must first pass a defined quality score threshold. If the threshold is not met, a 'no-call' genotype is produced for the position (annotated as "NN" or "N"). If the threshold is met, data for that sequencing position undergoes further filtering, based on user-defined parameter thresholds (such as strand bias, maximum common signal shift, etc.). The HID_SNP_Genotyper plugin, which is based in the Torrent Variant Caller, only reports genotypes for target positions (specified in a hotspot file) independently of them being variable in comparison with the reference sequence. Other software programs, such as IGV (Robinson et al. 2011, Thorvaldsdottir et al. 2013), can be used to visualise individual reads, assess atypical sequencing characteristics and examine the context sequence of the target polymorphisms.

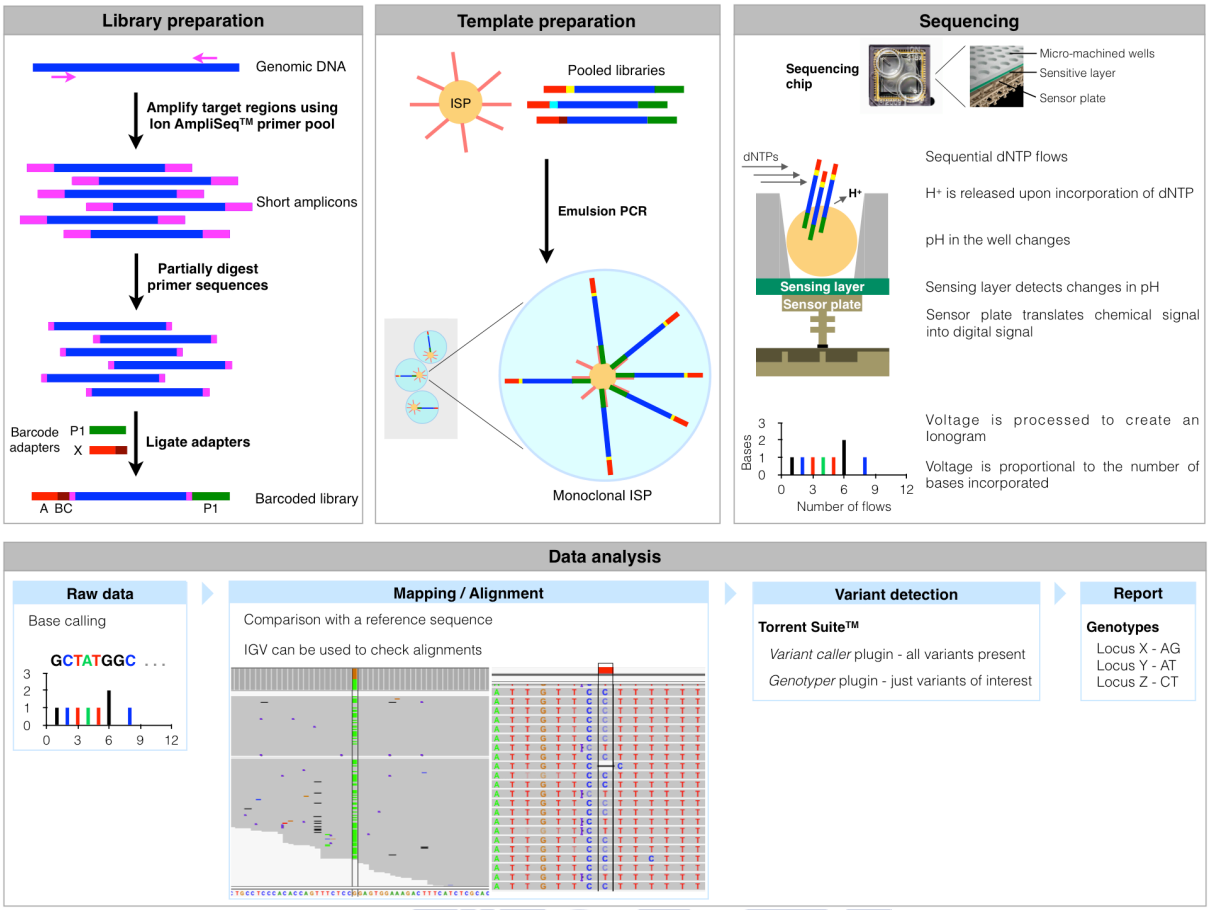


Figure 18: Ion Torrent™ Personal Genome Machine® (PGM) workflow. DNA libraries can be prepared from genomic DNA – in this particular case, when targeting regions of interest, the Ion AmpliSeq™ library preparation kit can be used. If barcoded adapters are used, libraries prepared from different DNA samples can be pooled. An emulsion PCR generates clonal copies of the library on the surface of the Ion Sphere™ particles (ISPs). Sequencing is performed in a chip that contains millions of wells – only a single ISP fits each well. Sequential flows of dNTPs are passed over the chip surface and whenever a nucleotide is incorporated into the growing sequence a H^+ is released. This leads to pH changes that are detected by the sensing layer and converted into a digital signal by the sensor plate. Such voltage changes are proportional to the number of added bases and are processed to generate the raw data (termed ionogram). Nucleotide sequence is inferred from the raw data and is then aligned against a reference sequence. Variants are detected and genotypes reported by the Torrent Suite™ software.

One of the major drawbacks of several NGS techniques is the presence of Indels and homopolymeric tracts in the target sequence. This is particularly true for the Ion PGM™ system – although observed pH changes are proportional to the number of incorporated nucleotides, smaller increases in the pH difference are observed when identical bases are incorporated (Loman et al. 2012, Perkins et al. 2013). The lack of accuracy when assessing the length of the homopolymeric tracts converts them in the major source of Indels and sequence differences between sequencing techniques (Loman et al. 2012, Parson et al. 2013, Ratan et al. 2013). Furthermore, this is highly dependant on the alignment algorithm used and it can benefit from more sensitive base calling software (Parson et al. 2013, Perkins et al.

2013). Indels, due to lack of agreement with the reference for several bases around its position, can produce alignment artefacts that increase the chance of SNP mistyping (Ratan et al. 2013) but their correct identification can be achieved by sequence re-alignment using specific algorithms (Ratan et al. 2015).

Sequence read length is another limitation of NGS techniques, especially when it comes to targeted STR analysis. Nevertheless, STR analysis using NGS technologies is of great interest, especially in the forensic field, as it allows the discrimination of same length alleles that have different sequence or when variation inside a repeat unit is present (Yang et al. 2014). The sequencing by synthesis technology developed by Roche Diagnostics[6] (Margulies et al. 2005) is highly used for *de novo* STR isolation (Delmas et al. 2011, Malausa et al. 2011) and analysis of known markers (Fordyce et al. 2011, van Neste et al. 2012, Dalsgaard et al. 2014, Gelardi et al. 2014, Rockenbauer et al. 2014, Scheible et al. 2014) since it has read lengths of up to 400-1000bp. Nevertheless, recent advances in other platforms have been made and STRs have been analysed in both the Ion PGM™ (Fordyce et al. 2015) and the sequencing by synthesis Illumina®[7] platforms (Bornman et al. 2012, Zeng et al. 2015).

1.2.4.5.4 Amplification artefacts of short polymorphisms

Stutter peaks

As described in section 1.2.4.1, replication slippage is the main *in vivo* mutation process giving origin to STRs. But the DNA polymerase used in PCR amplification can produce the same errors. Strand slippage during PCR results in the amplification of fragments that are typically one or more repeat units shorter than the true allele (**Figure 19**) and are known as stutter products (Walsh et al. 1996). Generally, when analysing good quality samples, stutters do not affect sample genotyping as they can be easily identified as artefacts. However, when analysing challenging samples, for example low template or samples with more than one donor, stutter can lead to genotyping errors. As stutter peaks are the result of random slippage, repetition of the amplification reaction should eliminate those artefacts.

[6] http://my454.com/products/technology.asp
[7] http://www.illumina.com/neoprep-hp-takeover.html

Pull-up peaks

Analysis through CE implies that fluorescent dyes are used to label markers or alleles. As mentioned in section 1.2.4.5.2, each dye emits in different wavelengths. However, although they emit in short ranges, some overlap between each dye spectra is observed (**Figure 19**). When the amplification yield is high, the concentration of dye molecules injected into the capillary is also elevated. In such situation, fluorescence emitted at the extremes of each spectrum is not well resolved by the multicomponent matrix used by the software to differentiate each dye. This will result in an artefact peak in other colours, or pull-ups, that can affect genotyping if their position falls in an expected allele position. A simple solution is to dilute the PCR product prior to re-injection in the capillary sequencer.

Incomplete adenylation

Polymerase frequently adds an extra nucleotide (not present in the template DNA) to the 3' end of the PCR product. It usually performs adenylation, that is, it adds an extra adenine to the end of the product, originating a 1 bp longer fragment. However, excess of DNA template in the PCR reaction can result in incomplete adenylation, which generates a split peak pattern (**Figure 19**). There are two possible solutions to avoid split-peaks from incomplete adenylation: i) increase the final extension time to give the polymerase more time; or ii) a guanine base can be added to the 5'end of the reverse primer (usually unlabelled) to promote the addition of an extra base. This is true for the amplification using dye-labelled primers such as STR and Indel analysis. In SNP analysis, incomplete adenylation can occur during PCR amplification, but it will not be detected because in SNaPshot® analysis fluorescence is associated with the nucleotides. Nevertheless, during SBE amplification, adenylation peaks can be produced: if the PCR yield is high and an excess of product is loaded into the SBE reaction, the polymerase will add a labelled ddATP to the PCR products present in the reaction. For this reason, adenylation peaks in SNP analysis exist but are restricted to the green lane.

Dye-blobs

Dye-blobs are usually associated with amplification using dye-labelled primers, that is, STR and Indel analysis. This is explained by the primer synthesis process – the reaction that attaches the fluorescent dye to the primer sequence is not 100% efficient and if the primer purification does not remove the unattached dye molecules they can be included in the PCR amplification and later injected during CE. The dye-blob peaks are wider and can be easily differentiated from real alleles (**Figure 19**). Nevertheless, if they fall into a true peak position, they can obscure the real allele signal, making genotyping more difficult. As the number of unattached dye molecules will be reduced in comparison with the concentration of the PCR product, re-injecting the sample is usually sufficient to remove the dye-blob artefact. In more serious cases, purification using gel matrix columns can be performed.

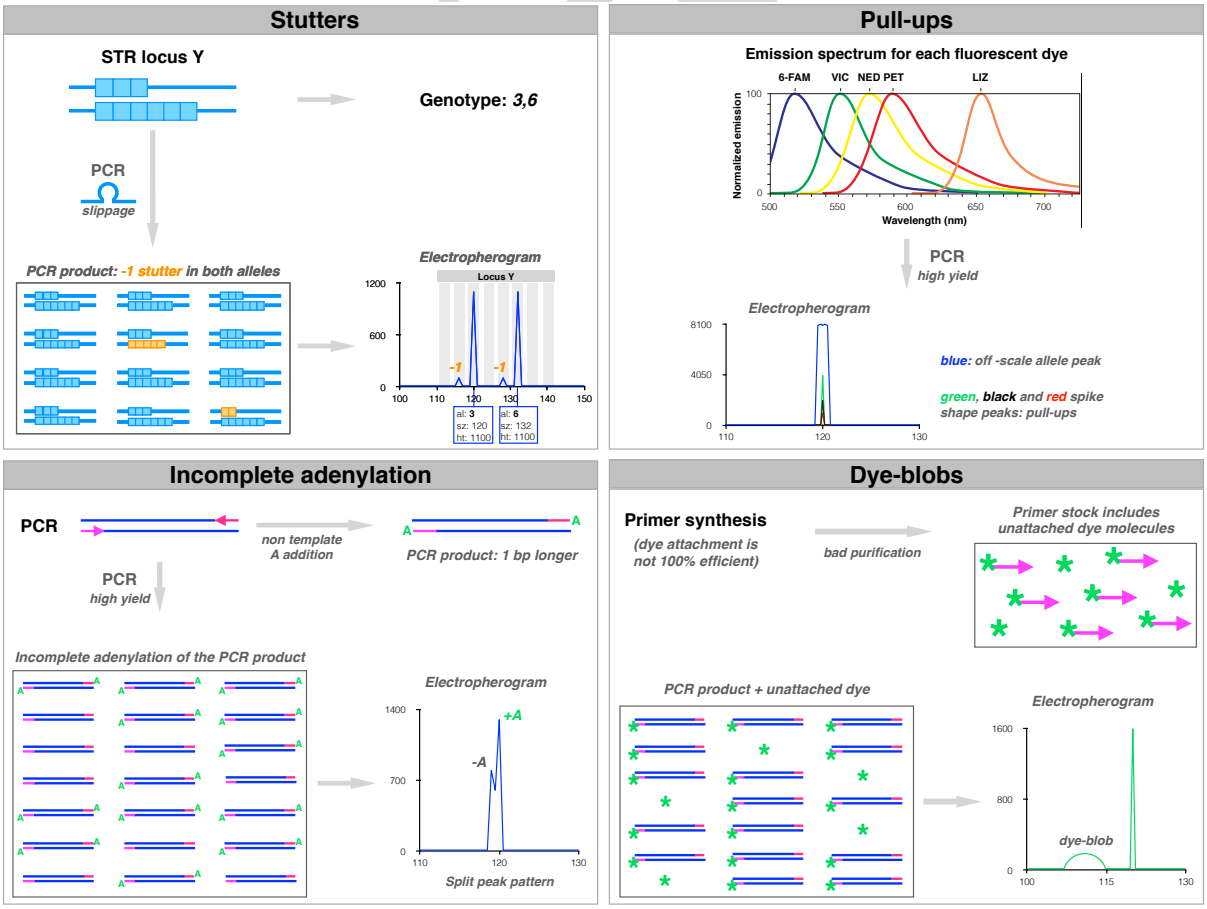


Figure 19: Amplification artefacts usually found during the analysis of short polymorphisms.

1.2.4.6 Advantages and disadvantages of each short polymorphism class

In this section, a discussion on the advantages and disadvantages of each short polymorphism class considered in this thesis is presented. A particular focus on pros and cons for forensic applications (described in section 1.2.4.7.3) is given.

SNPs and Indels share, in comparison with STRs, most of their characteristics and advantages, especially for forensic genetics applications. Particularly, they can be amplified in short amplicons, which is highly advantageous when analysing degraded samples. Furthermore, they are very stable markers so patterns of variation tend not to change greatly between generations. But despite the low mutation rate, and since the world human population is large, changes that transform one allele to another in the total variant pool can occur with a greater frequency than that detected (Jobling and Tyler-Smith 2003). For this reason we can expect to find some mutations that are fixed in one population and that the same mutational event will not occur independently in other populations. So, this common pattern of shared variability can only extend between populations through migration, which is of considerable utility for the study of evolutionary genetics and a population's history.

As previously outlined, SNP amplification offers the possibility to analyse degraded samples with a high rate of success when compared with other techniques, since the polymorphism is constituted by a single base and it only requires a small fragment of intact DNA so that the probability of finding the target sequence is bigger (Gill et al. 2004, Phillips et al. 2004b). In fact, amplicon size is one of the major drawbacks of STRs, especially when analysing degraded DNA samples. The higher the degradation rate, the smaller are the DNA fragments that can be found in the sample which, in extreme conditions, can be smaller than 100 bp. Due to their repetitive nature, the STR region can extend through hundreds of bases, which reduces their amplification success rate – this is particularly true for longer STRs. In order to overcome this problem a new approach to STR analysis, termed miniSTRs, was developed: while trying to keep the most informative markers, new primer designs were made in order to place the PCR primers in the closest position possible to the repeat region. Although this new design reduced amplicon length, this still varies between approximately 50 and 300 bp (Butler et al. 2003, Coble and Butler 2005), with some amplicons reaching 400 bp in a multiplex assay that accommodates 26 STR loci (Hill et al. 2009). This increased sensitivity can lead to greater chances of finding mixed-source DNA patterns, for example

due to contamination, especially when analysing degraded DNA samples that are not as easily amplified as even small amounts of high-quality DNA (Butler 2011). With STRs, this problem can be easily controlled through the detection of more than two alleles per locus. However, it is difficult to detect mixed-source samples when analysing binary markers and so the use of nonbinary SNPs or Indels is recommended. Such markers amplify with the same efficiency as the bi-allelic ones but they can indicate the presence of additional DNA through the detection of a third or a fourth allele (Phillips et al. 2004b).

The binary nature of SNPs and Indels also means that approximately four times more binary markers than STRs have to be used in order to obtain the same discrimination power (Krawczak 1999, Gill 2001b, Amorim and Pereira 2005). Although this is a disadvantage, it can be easily overcome by the construction of bigger assays. In the particular case of SNP analysis, this was possible through advances in SNP typing technologies that allowed the design of multiplex reactions that make the most of the advantages offered by SNP analysis using the least amount of DNA (Gill 2001b, Gill et al. 2004, Phillips et al. 2004a, Sobrino et al. 2005).

One of the major SNP disadvantages is that a complex typing process is required in contrast to direct PCR-to-CE analysis of both STRs and Indels. As described in section 1.2.4.5.2, SNP typing usually includes a post-PCR amplification and several intermediate steps. Despite the complexity inherent to the technique, a second typing reaction means that having amplicons with the same length at the PCR stage is not a problem, which allows several markers to be amplified in short amplicons. But when considering SNaPshot®, the major limitation derives from having both alleles of any SNP reported with different fluorochromes. This generates an inherent allele imbalance that, when analysing good quality profiles, represents a smaller problem that can be easily overcome by expertise. However, as conditions depart from the ideal and in the presence of challenging samples, deviations to the expected balance can occur to the extreme of allele drop-out.

1.2.4.7 General applications of short polymorphisms

SNPs, STRs and Indels have three main applications in the human genetics field: anthropological, clinical and forensic.

1.2.4.7.1 Anthropological applications

Several applications in the field of human genetics need an estimate of the ancestry proportions of the parental population in one population or individual that is the result of the admixture between those parental populations. For this reason, the use of genetic markers to determine the population of origin of one unknown sample is of great interest. Several markers have been identified as relevant to the estimation of ancestry components. In particular, SNPs are very useful since they present a greater variation between populations due to their age and stability of the nucleotide substitutions (Pfaff et al. 2004, Phillips et al. 2004a). Together with SNPs, Indels represent a great proportion of the human biallelic polymorphisms, so their inclusion in genetic studies can allow a higher resolution of the genetic maps and the improvement of the estimation of recombination rates and inferred haplotypes using statistical methods (Bhangale et al. 2005). On the other hand, STRs usually have homogeneous frequencies between populations and so they have limited classification power. Nevertheless, STRs with population specific alleles exist, the best example being D9S1120 (Phillips et al. 2008b), and marker sets have been used to correctly differentiate main population groups (Rosenberg et al. 2002, Kayser and de Knijff 2011, Phillips et al. 2011, Pemberton et al. 2013, Phillips et al. 2013a).

1.2.4.7.2 Clinical applications

The increased knowledge about the human genome and the availability of high throughput technologies at low costs have been used to study genes and markers associated with diseases and normal biological processes. Due to their simplicity, stability, abundance and straightforward analysis, several association studies of SNPs in distinct chromosomes to diseases have been developed. The computation of the linkage disequilibrium measure can help to locate those genes. The genetic analysis of SNPs can also be used in case-control studies, genetic association studies or detection of structural variants causing, for example, cancer (Tsunoda et al. 2004, Strauss et al. 2005, Monzon et al. 2008). Indels can be included in population differentiation estimation analysis (Pearce 2006), which can be beneficial, for example, to detect population substructure in case-control studies (Santos et al. 2010).

Although SNPs represent approximately 67% of known pathological mutations (Stenson et al. 2003), both Indels and STRs have been recognized as possibly disease causing loci. Short insertions and deletions (≤ 20 bp) represent approximately 24% of the mutations causing inherited diseases being associated with, for example, cancer, cystic fibrosis and fragile X syndrome (Collins et al. 1987, Warren et al. 1987, Ball et al. 2005, Ngo et al. 2011, Varela et al. 2011, Kim et al. 2012). Indels are usually responsible for frame shift of the reading frame during DNA transcription, affecting the assembly of completely functional proteins and, ultimately, their function. Several STRs were associated with autoimmune responses because: i) they can lead to the production of new and potentially immunogenic proteins; ii) they can affect splicing and generate aberrant proteins; or iii) affect methylation inducing incorrect transcription initiation (Semsei et al. 2007, Yauk et al. 2008, Ball et al. 2009, Tazi et al. 2009, Ross 2014). Other genetic diseases such as Huntington's, fragile X syndrome, myotonic dystrophy or neurodegenerative disorders, have been associated with STRs (Sutherland and Richards 1995). The majority is associated with trinucleotide repeats, particularly with the CAG triplet – expansion of the number of repeats outside of the normal polymorphic range generates abnormal proteins with longer polyglutamine runs (La Spada et al. 1991, THDCRG 1993, Fan and Chu 2007, Doi et al. 2014).

1.2.4.7.3 Forensic applications

Short polymorphisms can also be applied to forensic casework, for example, to perform individual identification derived from any biological material. STRs are the key markers used in forensic analysis. Currently used loci are located in non-coding regions of DNA (with some markers positioned in gene introns) – they do not affect the individual's fitness and are not submitted to selective pressures. For this reason, a range of alleles can be found, at low frequencies, in any population. This originates a high degree of length polymorphism and a predominantly multiallelic nature, explaining why STRs are highly informative to perform individual identification or affiliation between possibly related individuals. Furthermore, the direct typing methodology (section 1.2.4.5.2) allows quick genotyping of scarce samples.

In the last decade, a great development of the forensic genetics field occurred, increasing analysis resolution and detection sensitivity (Budowle and van Daal 2008). However, STRs sometimes fail to amplify, particularly when analysing challenging DNA samples, i.e., those

that, due to their low quality or quantity, can produce problems during results interpretation. For that reason, it is important to search for other markers suitable for forensic analyses (Budowle and van Daal 2008, Jobling et al. 2014).

The forensic use of SNPs is of great interest since they present a set of characteristics that makes them suitable for forensic studies. First of all, they can provide an additional level of support to analyse highly degraded material, where STRs fail, as they can be amplified through short amplicons, which is critical for the success of the analysis of degraded samples (Gill et al. 2004, Phillips et al. 2004b, Budowle and van Daal 2008). Second, the reduced mutation rate is advantageous for paternity testing, improving the exclusion power for family reconstruction or when there is not a direct reference sample (Budowle and van Daal 2008, Phillips et al. 2008a). SNPs applicability in human identification as been extensively studied (Inagaki et al. 2004, Petkovski et al. 2004, Sánchez et al. 2006, Musgrave-Brown et al. 2007, Fondevila et al. 2008b, Børsting et al. 2009). But more recently, ancestry informative SNPs (Halder et al. 2008, Kersbergen et al. 2009, Nassir et al. 2009, Fondevila et al. 2013, Nievergelt et al. 2013, Phillips et al. 2013b, Kidd et al. 2014, Phillips et al. 2014) and external visible characteristics SNPs (Walsh et al. 2011a, Ruiz et al. 2013, Walsh et al. 2013, Maroñas et al. 2014) were assessed as such markers can provide investigative leads from DNA evidence when no hit to the STR database is obtained or if the sample fails to provide a STR profile.

Along with SNP, Indels have great potential in the forensic genetics field as they combine characteristics from SNPs and STRs (Pereira et al. 2009). Since Indels are binary polymorphisms, it is possible to say they have the same characteristics than SNPs and, for that reason, they have the same advantages in forensic genetics (Edelmann et al. 2009). The most abundant group of biallelic Indels is the one containing polymorphisms with allele differences of 3 to 15 nucleotides (Mills et al. 2006, Pereira et al. 2009, Santos et al. 2010). They can be analysed with techniques established for STRs (which is an advantage in comparison with the multi-step amplification methods developed for SNP analysis) but with smaller amplicon size, improving the amplification of degraded samples and the assembly of multiplex assays (Edelmann et al. 2009, Pereira et al. 2009, Ribeiro-Rodrigues et al. 2009, Pimenta and Pena 2010, Santos et al. 2010). Furthermore, they present distinct frequencies among populations (Hofer et al. 2009, Ribeiro-Rodrigues et al. 2009) and have potential to be ancestry informative markers (Weber et al. 2002, Pereira et al. 2009, Santos et al. 2010). Several

studies were performed using Indels for human identification (Pereira et al. 2009, Li et al. 2011), kinship analysis (Edelmann et al. 2009, Pimenta and Pena 2010) and characterization of human populations and their differentiation (Bastos-Rodrigues et al. 2006, Ribeiro-Rodrigues et al. 2009, Santos et al. 2010, Pereira et al. 2012b, Zaumsegel et al. 2013).

A comparison of short polymorphisms in the human genome, including major advantages and limitations in forensic applications, is presented in Table 3. Their forensic application, particularly ancestry inference, will be discussed in detail in section 1.3.2.

Table 3: Comparison of short polymorphisms classes. Major advantages and limitations of STRs, SNPs and Indels in forensic applications are presented. Adapted from Butler (2011).

		STRs	SNPs	Indels
Number of sites		~1 in every 10 Kb	~1 in every 100 bp	~1 in every 1000 bp
Mutation rate		10^{-3} to 10^{-4}	$1.2x10^{-8}$	$2.3x10^{-9}$
Alleles per marker		5 to 50	Typically 2 (some triallelics exist)	Typically 2
Amplicon size		~75 to 400 bp	Can be less than 100 bp	~50 to 200 bp
Amplification artefacts		Yes	Yes	Yes
Phenotypic information		No	Possible to predict some traits	No
Ancestry information		Limited	Yes	Yes
Specific of forensic applications	**Major advantages**	- Highly informative - Multiallelic markers (easier mixture detection and deconvolution) - Established databases	- Amplified in small amplicons (higher amplification success rates with degraded samples) - Low mutation rates (helpful in kinship analysis) - Phenotype and ancestry prediction	- Typed through direct PCR-to-CE as STRs - Amplified in smaller amplicons (higher amplification success rates with degraded samples) - Low mutation rates (helpful in kinship analysis) - Ancestry prediction
	Limitations	- Amplified in long amplicons (lower amplification success rates with degraded samples)	- Low informativeness (more markers needed to obtain equivalent discrimination power) - Multistep typing methodologies - No widely established core loci - Mixtures are difficult to detect/resolve	- Low informativeness (more markers needed to obtain equivalent discrimination power) - No widely established core loci

1.3 FORENSIC GENETICS STUDIES

1.3.1 A short history of forensic genetics

Forensic sciences are a multidisciplinary specialty that aim to help solving legal questions, both in criminal law but also in civil cases. The first written account of forensic methodologies is included in a medico-legal textbook written by the Chinese jurist Sung Tzhu in the 13th century. The book, named *Xi Yuan Ji Lu* (one possible translation being "The washing away of wrongs"), presented a systematic treatise on forensic medicine and includes selected cases where medicine or, for example, entomology were used (Lu and Needham 1988). But alongside with medicine, the forensic field includes several other areas of knowledge such as chemistry and biology. In particular, DNA analysis has been a central area of forensic sciences, especially in the last 30 years. As described throughout section 1.2, DNA is present in all living organisms and is variable between individuals and populations. For this reason, any biological source associated with a legal case has the potential to give information about its source.

1.3.1.1 Early days of forensic analysis of genetic variation

Two experiments, performed in 1944 and 1952, recognized DNA as the carrier of genetic information (Avery et al. 1944, Hershey and Chase 1952), and its structure was not characterized until 1953 (Watson and Crick 1953). Before that time, proteins were thought to carry genetic information. In fact, the first human genetic polymorphism was defined in 1900 – the human ABO blood group system (Landsteiner 1900). Ten years later, the mendelian inheritance and the transmission as a dominant trait was described (von Dungern and Hirschfeld 1910). This polymorphism can, by itself, be used in forensic investigations to prove exclusion, i.e., that a casework sample has a different profile to that of a suspect, indicating that the suspect is not the individual responsible for the deposition of the biological stain in the crime-scene. It can also show that a sample did actually come from a specific individual, that is, an inclusion or match. But due to the low degree of variation of the ABO system, the estimated inclusion probabilities improved if other markers were added to the analysis. Such markers included a wide range of blood groups defined by antigens such as the

MNS (Landsteiner and Levine 1927), Rh (Levine and Stetson 1939, Landsteiner and Wiener 1940) and the Lewis (Mourant 1946, Andresen 1947) systems. Later on, serum proteins such as gammaglobulines and erythrocitary enzymes (for example, glyoxylase) were described and could be detected using electrophoretic techniques. In 1958 white cell agglutination was discovered and the HLA (human leucocyte antigens) system described (Dausset 1958), being the most polymorphic antigenic systems detected until that moment.

Until 1980's, blood groups and other polymorphic proteins were used as a way to access genetic variation. But they had three major drawbacks: i) they were rapidly degraded after sample deposition; ii) they present relatively low variability and informativeness, that is, two unrelated individuals had high probability of sharing the same profile; and iii) mixed stains (two or more contributors) were difficult to resolve (Jobling and Gill 2004).

In the 1980's two decisive developments occurred. First, hypervariable loci, known as minisatellites, were discovered in 1985 (Jeffreys et al. 1985a), improving the discrimination power between unrelated individuals. Minisatellites were analysed, after DNA digestion with restriction enzymes followed by gel electrophoresis, through hybridization of a radioactively labelled probe, i.e., a short DNA sequence complementary to that present in several highly polymorphic minisatellite loci (Gill et al. 1985, Jeffreys et al. 1985b). The final result was a DNA fingerprint, that is, a band pattern where variable length fragments corresponded to the minisatellites analysed. DNA fingerprints discrimination power was so high that only monozygotic twins shared the same pattern (Jobling and Gill 2004). With time, DNA fingerprinting evolved into the analysis of specific minisatellites through the use of single-locus probe analysis, a technique called DNA profiling. In it, each single-locus probe targeted a single highly polymorphic restriction length polymorphism, simplifying the interpretation and analysis of mixed stains. Approximately at the same time, a DNA extraction technique, particularly useful to co-extract mixed stains as those resulting from sexual assault cases, was developed (Gill et al. 1985). The extraction is based in the differential lysis of distinct cell types (usually epithelial and sperm cells), originating two DNA fractions that can be analysed separately.

1.3.1.2 The PCR revolution

Although high discrimination levels could be achieved through single-locus probe analysis, this technique had some limitations (low sensitivity, particularly when analysing degraded or low DNA quantity samples, and it was both time and sample consuming) that were overcome with the development of PCR in 1986. As described in section 1.2.4.5.1, PCR amplification had a great impact on biological and medical research. The increased sensitivity of the PCR technique allowed the analysis of minimal DNA amounts, as those usually found in casework samples, and its versatility transformed PCR in the basis of all forensic typing technologies (Jobling and Gill 2004). The first PCR-based systems targeted STR markers such as D1S80 (Nakamura et al. 1988, Kasai et al. 1990) and a small number of SNPs in the HLA-DQA1 gene (Helmuth et al. 1990).

The characterization of forensically useful polymorphic STRs, the design of multiplex PCR assays and the automation of DNA (Edwards et al. 1991, Kimpton et al. 1993) were the first steps of a technological revolution that culminated in the powerful systems for human identification available at the present. High discrimination power, sensitivity and ability to resolve mixtures are characteristics that define STRs and explain why they replaced SNPs and single-locus probes analysis. Furthermore, time needed and costs associated with STR analysis are much smaller, which allowed the establishment of national STR DNA databases in a number of countries: they include several hundred thousands to millions of profiles and are usually held by the police or the justice system (Jobling and Gill 2004, Kayser and de Knijff 2011). Such databases can be extremely useful either by linking offenders with previously committed crimes but also by exonerating innocent people who had been wrongly accused. However, DNA databases, and their use, have to reach a balance between individual rights and public safety (Jobling and Gill 2004, Butler 2011). For that reason, a number of issues, mainly regarding privacy and information security, have to be considered: i) the DNA markers included should only be useful for identity testing, being located in non-coding regions of the human genome and not associated with genetic diseases or any predisposition; ii) no names or other individual characterizing data should be stored with the DNA profiles; iii) data should be encrypted and, when shared, it should be done through a secure network; and iv) improper use of criminal DNA samples should be penalized (Butler 2011). Other aspects to be considered are the inclusion parameters and the possible discriminatory nature

of the criminal justice system and of such databases, with some groups probably being overrepresented among suspects (Jobling and Gill 2004, Butler 2011). Some have argued that a global, population-wide, database would eliminate any possible racial inequality regarding sampling (Kaye and Smith 2003) but this approach would raise serious ethical issues (Guillen et al. 2000).

1.3.1.3 New markers and their application in forensic genetics

Gill et al. (2004) stated that "*It is unlikely that SNPs will replace STRs as the preferred method of testing of forensic samples and database samples in the near to medium future*". More than 10 years later this affirmation is still true mainly because STRs have several advantages: i) they are highly informative – approximately four times more binary loci are needed in order to achieve the same discrimination value of STRs; ii) the multiallelic nature of STRs permits an easier interpretation of mixed-source profiles; and iii) extensive national STR databases are already established – the use of SNP-based human identification in routine casework would imply the development of new forensic DNA databases (Gill et al. 2004, Kayser and de Knijff 2011). However, Gill et al. (2004) do recognize that autosomal SNPs are valuable forensic tools that can be used in specific applications. Despite the low discrimination power and inherent difficulty to resolve mixtures, SNPs have characteristics that make them interesting markers to be included in the forensic genetics toolbox. Together with complex kinship analysis (Budowle and van Daal 2008, Phillips et al. 2008a), SNPs are extremely valuable in the analysis of severely degraded samples such as the ones found during disaster victim identification (Prinz et al. 2007), a situation best exemplified by the remains identification effort after the World Trade Center terrorist attacks (Biesecker et al. 2005). Other strategies were also developed to aid the analysis of severely degraded or low quantity DNA samples – low copy number analysis, i.e., increase the number of PCR cycles from 28 to 34 when amplifying STRs (Gill et al. 2000); to perform whole genome amplification of the sample before proceeding to PCR amplification of the loci of interest (Ballantyne et al. 2007a); and post-PCR purification to enhance signal detection sensitivity (Smith and Ballantyne 2007, Forster et al. 2008). However, those techniques can be affected by increased stutters, allelic and locus drop-out, allelic drop-in and complex result interpretation (Gill et al. 2000, Gill 2001a, Jobling and Gill 2004). Additionally, it has been

suggested that the interaction between DNA and histones offers protection to DNA from degradation (Dixon et al. 2005). The hypothesis of nucleosome protection was demonstrated both on STRs (Dong et al. 2015) and in SNPs (Freire-Aradas et al. 2012), increasing the chances of obtaining better profiles with fewer allelic drop-outs. Furthermore, a study from Thanakiatkrai and Welch (2011) showed that most of the forensic STR loci used are somewhat protected from degradation by nucleosomes.

Another tool that adds value to the standard STR analysis is haplotype determination from sets of non-recombining male-specific Y-chromosome STRs (Kayser et al. 1997, Mulero et al. 2006). They can be used for kinship testing or disaster victim identification cases where male relatives are involved, but they are particularly useful when trying to identify the male component in mixed male/female stains in cases of sexual assault (Roewer 2009). However, due to the inheritance model of the Y-chromosome (section 1.2.3.2) and the low mutation rates of the used Y-STRs, in case a match is obtained, closely related males (those that have a patrilineal relationship) cannot be excluded (Roewer 2009, Kayser and de Knijff 2011). In such cases, the use of the recently developed rapidly mutating Y-STRs can help to differentiate related males (Ballantyne et al. 2010, Ballantyne et al. 2012, Ballantyne et al. 2014). Identity can also be established using mtDNA, in this case through comparison with maternally related individuals. This approach is extremely helpful for disaster victim identification, analysis of ancient DNA or genealogical studies, situations where the samples contain little DNA, are old or were severely degraded. Finally, X-chr markers are also useful in forensic analysis, particularly due to their distinct mode of transmission – while males have a single X-chr that is transmitted mostly unchanged to female descendants (low recombination levels are possible in the pseudoautosomal regions as described in section 1.2.3.2.2), females have two copies that recombine in each generation, contributing to the increase of haplotype diversity. In fact, the X-chr has an extensive haplotype structure mainly due to its smaller effective population size, lower SNP density and the restriction of recombination to females (IHMC 2005). Commonly used forensic X-STRs are distributed across the complete chromosome. However, some markers belong to the same linkage group, generating haplotypes that are particularly useful for complex kinship testing involving at least one female, especially when the analysis of autosomal markers is limited or insufficient (Szibor et al. 2003a, Hering et al. 2006, Szibor 2007, Becker et al. 2008, Edelmann et al. 2008, Hundertmark et al. 2008). For example, X-chr markers are useful: i) in paternity cases

(duos or trios with female offspring) where routinely used autosomal STR markers are not sufficiently informative (for example due to incompatibilities); ii) the presumed fathers of a female descendent are close relatives; iii) in incomplete paternity testing where the alleged father, alleged mother or both are absent and it is necessary to test close blood relatives – for example, paternity can be assessed through presumed sisterhood between two females; ii) in the rarer cases of maternity testing (Szibor et al. 2003b, Edelmann et al. 2004, Bobillo et al. 2008, Serra et al. 2008, Toscanini et al. 2009, Trindade-Filho et al. 2013). Such markers are also useful in missing persons or disaster victim identification cases where direct reference samples are not available and it is necessary to test biological relatives (Butler 2011).

Human identification applications, independently of the markers used, are based in profile comparisons. For this reason, and in addition to cases where the sample fails to provide a STR profile, when no hit to reference samples or with the STR database is obtained or if no reliable eyewitness accounts are available, any investigative leads that can be deduced from the DNA evidence are useful. This is known as forensic DNA phenotyping and it includes two main areas: inference of probable origin (biogeographic ancestry) and externally visible characteristics (EVCs) (Kayser and de Knijff 2011).

As described in section 1.1.3, approximately 95% of the human genetic variation is found within populations (Rosenberg et al. 2002), with individuals from distinct populations being more different between themselves than those belonging to the same population. Although a small degree of population differentiation has been shown to exist in commonly used STRs (Phillips et al. 2011) and there are loci with population specific alleles (Phillips et al. 2008b), STR markers are not ideal for ancestry prediction as they are highly variable between individuals and have low inter-population variation. Autosomal binary loci and STRs with large allele frequency differences between populations have been identified (Shriver et al. 1997, Collins-Schramm et al. 2002). Such AIM markers can be used to estimate individual ancestry membership proportions, allowing the prediction of probable biogeographic origin. The application of ancestry estimation in the forensic field will be discussed in detail in section 1.3.2.

It is known that many human phenotypes, such as stature (Lango Allen et al. 2010, Wood et al. 2014), pigmentation (Stokowski et al. 2007, Sulem et al. 2007, Han et al. 2008, Kayser et al. 2008c, Sulem et al. 2008, Liu et al. 2010a) and hair morphology (Fujimoto et al. 2008a, Fujimoto et al. 2008b, Mou et al. 2008, Medland et al. 2009) have a strong genetic

component. Other traits were extensively reviewed in a publication about the forensic application of DNA phenotyping (Kayser 2015). Although environmental factors contribute to EVC variation, increasing the uncertainty of DNA-based EVC prediction, genetic tests that help to predict such characteristics would be very useful in forensic genetics (Kayser and de Knijff 2011).

One of the first EVC traits to be investigated in the forensic field was gender, which can be determined through the detection of variants in the XY-homologous amelogenin gene (Akane et al. 1991) and has been included in several STR profiling kits. However, female false-positive results are possible due to deletions in the Y-chromosome (Roffey et al. 2000, Thangaraj et al. 2002, Lattanzi et al. 2005, Mitchell et al. 2006, Jobling et al. 2007, Oz et al. 2008, Davis et al. 2012). For that reason, alternative gender determination methods were developed (Esteve Codina et al. 2009, Kastelic et al. 2009).

Human pigmentation is the phenotypical trait that underwent more serious investigation considering forensic purposes. The best studied gene, the melanocortin 1 receptor gene (MC1R), has more than 30 known variant alleles that produce amino acid substitutions, several of which are associated with red hair, fair skin and freckling (Valverde et al. 1995, Bastiaens et al. 2001, Beaumont et al. 2008). Red hair can be accurately predicted and the forensic use of such trait was soon considered (Grimes et al. 2001). More recently, a hair colour prediction system has been developed (Branicki et al. 2011) and adapted into a forensic assay that simultaneously examines hair and eye colour (Walsh et al. 2013). Eye colour, defined as a categorical trait, is actually the most accurately predictable pigmentation trait (Liu et al. 2009) and was easily adapted for forensic casework (Walsh et al. 2011a, Walsh et al. 2011b, Ruiz et al. 2013). Skin colour variation is also explained by genetic variation (Lao et al. 2007) and a forensic test was developed (Maroñas et al. 2014).

Genetic variants can also be used for body height prediction but each locus has little effect (Aulchenko et al. 2009, Estrada et al. 2009). For that reason, low accuracy values are obtained, with just one fifth of adult height heritability being explained by approximately 700 variants (Wood et al. 2014). Age estimation can also be of interest in forensic phenotyping – DNA methylation patterns (Lu et al. 2004, Teschendorff et al. 2010, Weidner et al. 2014, Zbiec-Piekarska et al. 2015) and somatic rearrangements in the T cell receptor locus (Zubakov et al. 2010, Ou et al. 2011, Ou et al. 2012) may provide suitable age-predictive biomarkers. DNA methylation can also be used to identify body fluids (Frumkin et al. 2011),

as can mRNA analysis (Liu et al. 2002, Ferri et al. 2004, Juusola and Ballantyne 2005, Nussbaumer et al. 2006, Juusola and Ballantyne 2007, Haas et al. 2009, Zubakov et al. 2009, Fleming and Harbison 2010, Visser et al. 2011). Facial morphology can, to a certain extent, be modelled using DNA markers but this trait has just begun to be explored (Liu et al. 2012a, Paternoster et al. 2012, Claes et al. 2014a, Claes et al. 2014b). Finally, other markers such as hormone concentration in blood or saliva (Ackermann et al. 2010), ratios of different types of RNA (Anderson et al. 2005, Anderson et al. 2011) or mRNA degradation levels (Bauer et al. 2003) can be used to assess two aspects that are relevant in crime scene investigation: time since sample's deposition and the time of the day at which the sample was deposited.

Although forensic DNA phenotyping can be very helpful in certain cases, many countries still limit DNA profiling to STR analysis (Kayser and de Knijff 2011). Physical appearance prediction main ethical concerns are related with privacy and targeting of particular population groups (Cho and Sankar 2004, Ossorio 2006). For this reason, phenotyping is restricted to the analysis of non-coding regions (that do not provide for any functional properties of an organism). However, as discussed by Kayser and Schneider (2009), due to the organization of the genome in linkage disequilibrium blocks, it is possible to find non-coding markers associated with visible traits or that can provide biogeographic ancestry information. Furthermore, if only EVC traits that are visible to everybody since birth and that are not associated to disease are considered, the risk of violating the individual privacy rights is reduced (Kayser and Schneider 2009). Ancestry should be interpreted as geography and not as race or even probable physical appearance, avoiding the discrimination of minority groups while keeping the privacy and data protected (Cho and Sankar 2004, REGWG 2005, Royal et al. 2010).

1.3.2 Application of population genetics to the forensic field

AIM markers can be used to estimate the likely population of origin of a DNA sample where the source individual is not known or is unable to declare its ancestry. Several polymorphisms exist with the potential to be used as AIMs, including Y-chr, mtDNA and autosomal markers (Jorde et al. 2000, Rosenberg et al. 2003, Egeland et al. 2004). As described in section 1.1.3, Y-chr and mtDNA haplogroups are not equally distributed across global human populations. For that reason, such markers can be used to infer the probable

geographic origin of an individual and to study questions of ancestry and migration of peoples throughout history. However, lineage markers have limitations, including: i) estimates of a random match probability are usually done by the counting method (the number of times the haplotype of interest is divided by the total number of haplotypes in the database used), hence the size of the database used can affect the estimated rarity of a profile – very large databases are required to properly gauge population variability (Butler 2011); and ii) there is a risk of finding intact lineages atypical of the population (King et al. 2007). Furthermore, due to their particular inheritance, both Y-chr and mtDNA variation is preserved in both lineages due to the lack of recombination, acting as single loci. In fact, one locus by itself does not provide the amount of information necessary since it is not representative of the entire genome (Lewis 2001, Salas et al. 2006, King et al. 2007). Y-chr and mtDNA are lineage markers that give information about the paternal and maternal lineages, respectively (**Figure 20 A,B**). But this

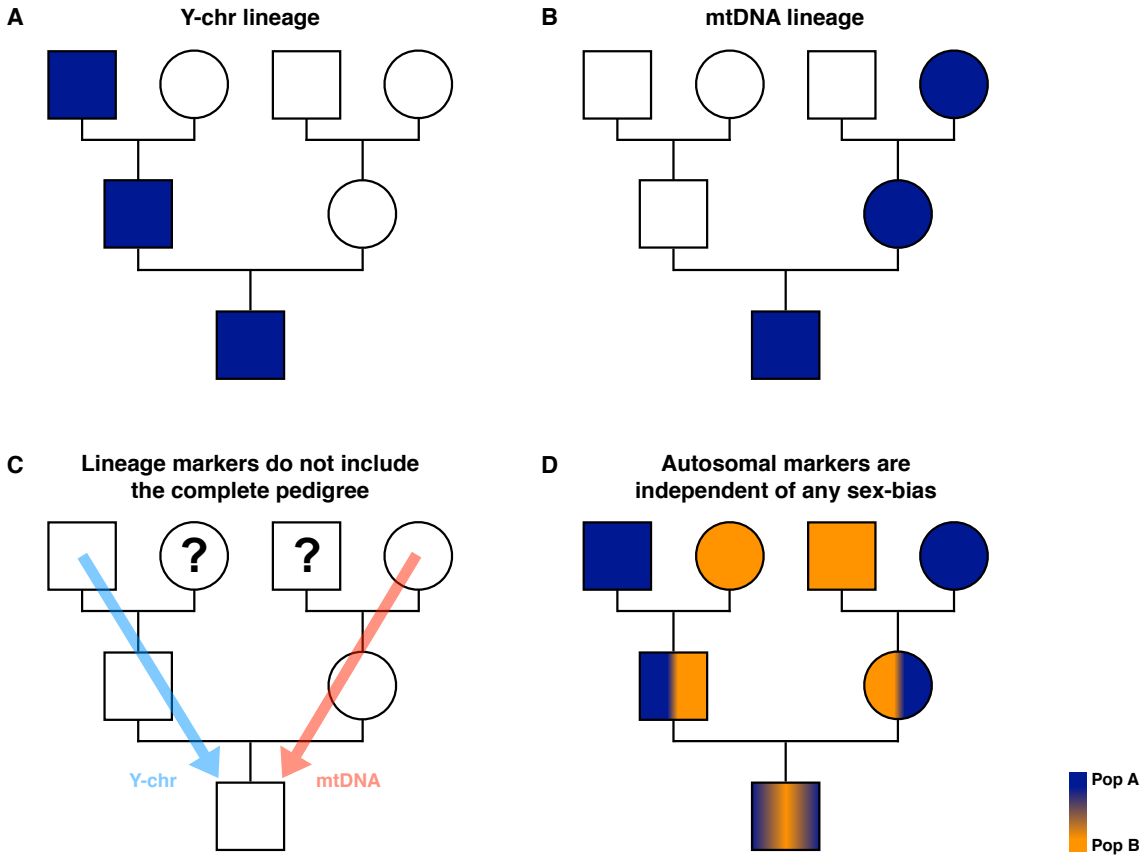


Figure 20: Comparison of uniparental and autosomal markers and possible ancestry inference. Both Y-chr (A) and mtDNA lineages are maintained unaltered between generations. But those markers do not give information about an individual's complete pedigree, with the paternal grandmother and maternal grandfather contribution to the individual genetic background not being studied when using uniparental markers (C). For that reason, the use of autosomal ancestry informative markers, which are independent of any sex-bias effects, is advisable as they give a broader view of an individual's ancestry (D). In this example, blue indicates that the individual belongs to hypothetical population A, orange to hypothetical population B, and individuals with both blue and orange are admixed.

implies that the genetic input from distant male/female lineages (in this case the paternal grandmother and maternal grandfather sides) are not taken into consideration (**Figure 20 C**). And if those have a different ancestral background from that of the detected lineages, a misrepresentation of the overall individual ancestry will be achieved. The risk of finding atypical lineages and that of misinterpreting an individual's overall ancestry pattern is increasing due to population admixture, a characteristic gradually more common in modern human demographics. For that reason, it is convenient to analyse several genetic markers in distinct loci in order to obtain the individual genotype or genetic profile. Autosomal markers have emerged as amongst the best ancestry markers due to their stability, density of distribution and full range of allele frequency patterns amongst populations. Autosomal markers provide valuable complementary data to Y-chr and mtDNA typing for the analysis of the evolutionary history of populations. They have the advantage that small numbers of samples are sufficient to give reliable allele frequencies estimates – with as few as 30-40 samples providing adequate population allele frequency estimates (Phillips 2015) – and also that they are independent of admixture sex bias allowing a more precise inference of the genetic ancestry of an individual (**Figure 20 D**).

But ancestry inference in forensic analysis started as early as 1960s. The recently discovered Duffy polymorphic blood group (Cutbush and Mollison 1950, Cutbush et al. 1950), which was promptly introduced in routine forensic analysis, showed a strong association between its null phenotype and a likely African ancestry component of the individual (Sanger et al. 1955). Several SNPs can be found in the Duffy antigen receptor for chemokines (DARC) gene. SNP rs2814778 in particular, is the most strongly associated with ancestry (Nalls et al. 2008). Later in the 1970s three additional red cell enzyme polymorphisms were used to specifically detect population variation: rs733062 in the peptidase A gene; rs334 in the haemoglobin S variant gene; and rs2228063 in the carbonic anhydrase gene (Phillips 2013). They all have a specific African allele (sometimes at low frequency) that is absent from other major population groups, being, for that reason, highly informative (Phillips 2013).

Section 1.1.3 described how the world's population can be grouped into major discrete ancestral groups with genetic similarity closely corresponding to continental origin. However, this pattern can be highly dependent on the sampling scheme because the majority of human worldwide genetic diversity actually takes the form of geographic clines rather than clades

(Serre and Paabo 2004, Manica et al. 2005, Rosenberg et al. 2005). For that reason, to detect differences between major continental groups, it is essential to find markers that show the most pronounced allele frequency discontinuities between continental regions (Rosenberg et al. 2002). Such markers can be included in small-scale marker sets with population specific genotypes. This implies that a compromise between the amount of information that can be obtained from such small-scale multiplexes and the simplification of the description of the complex human population structure has to be achieved (Phillips 2015).

Geophysical barriers impede free movement of people outside of the regions defined by such barriers, explaining the establishment of the currently observed population structure. When using 377 STRs, five major population groups (sub-Saharan Africa, Eurasia, East Asia, America and Oceania) were defined, with Eurasia separating into three less differentiated groups – Europe, Central South Asia and Middle East (Rosenberg et al. 2002). This study was criticized due to the sampling used – the HGDP-CEPH panel mainly includes mid-continent populations and avoids marginal areas were populations meet, a strategy that does not considers the continuous nature of variation (Serre and Paabo 2004). Further studies, including more markers (933 STRs or 650,000 SNPs) showed that human variation patterns observed when using the HGDP-CEPH panel are robust and not artefacts from uneven sampling along clines – allele frequency differences increase with geographic distance in usually smooth gradients but there are small discontinuities that create the clusters patterns detected (Rosenberg et al. 2005, Li et al. 2008).

Along with geographic barriers, there are other factors that can help to explain the current distribution of human diversity: i) the serial founder effect (described in section 1.1.1.2 and **Figure 2**) explains the reduced diversity of populations such as the Americans and Oceanians (Wang et al. 2007, Friedlaender et al. 2008); ii) different patterns of natural selection depending on bio-geographical such as climate, diseases and diet (Coop et al. 2009); iii) population migration followed by admixture (Pickrell and Reich 2014); and iv) archaic introgression, as exemplified by the presence, in modern humans, of Neanderthal and Denisovan genetic ancestry (Green et al. 2010, Reich et al. 2011, Sankararaman et al. 2012, Pickrell and Reich 2014).

As a summary, for forensic ancestry analysis, a five-group differentiation is a reasonable objective using small-scale marker sets selected to have strong allele frequency differentiation (Phillips 2015). However, for further analysis and differentiation of populations located in the

middle of variability continua, for example Middle East or Central South Asia, which are more likely to show higher misclassification rates, the use of more extensive marker sets that include loci with sharp frequency clines is required (Phillips et al. 2007b).

1.3.2.1 Selection of ancestry informative markers

There are two main points to be considered when selecting markers for forensic ancestry analysis: availability of population data and parameters used to select such markers.

First of all, not all databases have complete coverage of all human populations. The 1000 Genomes project now includes more than 2,500 individuals from 26 populations (Africans, Europeans, South Asians, East Asians, Americans and admixed). Furthermore, this project recently released data for approximately 81 million variants. However, this dataset lacks geographic coverage. For that reason, and despite the much smaller number of loci studied, the HGDP-CEPH panel of samples is a good alternative as it addresses other worldwide regions, including data from Middle Eastern, American and Oceanian populations. As the 1000 Genomes project, the HGDP-CEPH panel of samples also suffers from irregular global sampling (Rosenberg et al. 2002) and the number of samples is very small for some population groups, suggesting that use of the HGDP-CEPH panel to design and predict the performance of a set AIMs may be prone to a degree of bias in marker selection. Although ascertainment bias is a characteristic mainly observed in genome-wide studies (Marth et al. 2004, Clark et al. 2005), it can also affect patterns of genetic variation observed across populations (Romero et al. 2009, Novembre and Ramachandran 2011). Ascertainment bias results from markers being selected from data sets with only a few individuals, decreasing the chance of finding SNPs with low minor allele frequencies, so certain loci are not polymorphic in the populations used during marker selection and are not considered further (Clark et al. 2005, Guillot and Foll 2009, Albrechtsen et al. 2010, Novembre and Ramachandran 2011). Using the HGDP-CEPH panel to select AIMs can also introduce ascertainment bias from reliance on SNP genotypes generated with the Illumina Human Hap650K microarray. Hap650K SNPs were selected from HapMap data, which does not include, for example, Oceanian population samples (Li et al. 2008). A more direct bias might arise from the fact that marker selection is based in the variation recorded in small numbers of individuals in few populations from major groups. Low sample numbers may be sufficient to find a large

proportion of the most differentiated SNPs but cannot be representative of population variability across an entire geographic region.

When selecting markers to build a new assay, several parameters have to be considered. Those include, among others, distance between markers to ensure independence and a suitable context sequence that allows the design of primer sets without known issues. But the first parameter to be considered during the selection process is the amount of information given by a marker. In binary loci, pairwise population differentiation can be easily measured using delta (δ), an allele frequency differential (Shriver et al. 1997). When markers are sorted by their δ value, a ranked list of ancestry markers that maximize the collective divergence amongst the population groups compared is generated. A more refined informativeness value is estimated by the *In* metric – informativeness-for-assignment, which is derived from the Jensen-Shannon's Divergence measure (Rosenberg et al. 2003, Chen et al. 2005). The *In* value can be obtained by the multiplication of Divergence by 0.693 (Phillips 2015). A more common measure of population differentiation is the fixation index, F_{ST} (Wright 1949). All four measures are closely related and they all assess each marker's contribution to ancestry inference (Rosenberg et al. 2003, Phillips et al. 2007b, Winkler et al. 2010). Some studies have suggested that small marker sets targeting the highest possible δ values during SNP selection can be used to assign ancestry or estimate co-ancestry proportions in admixed individuals while genotyping effort, sample usage and costs are kept low (Shriver et al. 1997, Rosenberg et al. 2003). However, when designing a new marker set appropriate for forensic analysis, it is important to balance divergence amongst the target population groups, that is, to differentiate each group with equal power as a way to reduce estimation bias of the ancestry proportions, particularly in admixed individuals (Phillips 2015). This might be difficult to achieve because: i) considering the demographic events that gave origin to modern human diversity, African populations are highly divergent from other groups while East Asians show low divergence values with Americans and Oceanians; ii) due to the lack of fixed markers, each locus adds a varying divergence value to each population comparison, resulting in an overall imbalance of the AIM assay (Phillips 2015). Nevertheless, aiming for balanced divergence allows the setup of the most informative marker sets. This fact, together with the increased multiplexing scales offered by NGS, prompt the development of a set of 128 AIM-SNPs balanced for a five-group comparison (Phillips et al. 2014).

1.3.2.2 Data analysis systems

As described in the previous section, F_{ST} values can be used to estimate a marker's power to differentiate populations of interest. This parameter is derived from the fixation indexes proposed by Wright to measure the deviation of the observed heterozygote frequency with respect to the expected, according to the Hardy-Weinberg theorem (Wright 1949). The Hardy-Weinberg equilibrium (HWE) hypothesis states that, in large populations where random interbreeding occurs and no evolutionary forces are acting, the allele frequencies are maintained constant across generations (Hardy 1908, Weinberg 1908). There are three conditions required to consider that a population is in HWE: i) population size must be large to allow for the maximal transference of genetic variants over generations; ii) all interbreeding events must be random in order to minimize the effect of genetic drift; and iii) selection, migration or mutation forces must be absent in the testing population. Considering a marker with two alleles (A and B) in a population, there are three possible genotypes: AA, AB and BB. HWE can be represented, for biallelic markers, by the following equation:

$$(p+q)^2 = p^2 + 2pq + q^2$$

where p represents allele A frequency in the population and q is allele B frequency in the same population. If a population is in HWE, genotypic frequencies would only be determined by allele frequencies. For multiallelic markers, it is also possible to calculate the equilibrium, but through a more complex way and not through the direct application of a formula.

F_{ST} values can also be used to measure genetic variability between populations as a way to detect population structure, that is, if there are differences between populations or between individuals from a single population. If genetic drift action leads to the differentiation of populations, the F_{ST} value will be high. On the other hand, if there is a lot of gene flow between populations, keeping their similitude, the F_{ST} value will be small. A disadvantage of this parameter is that it is sensitive to changes in any of the populations included in the analysis. Taking this into account, an extension of this parameter was developed – it quantifies the degree of evolution of a particular locus, isolating allele frequency changes and allowing the specification of the degree of evolution produced but also which population suffered that particular change. Such approach is known as locus specific branch length

(Shriver et al. 2004). Other measures used to estimate diversity are usually accompanied by a significance test that demonstrates that the substructure probability of the population is higher than that randomly expected. One example of such methods are the permutation tests known as Monte Carlo, which randomly analyse the empirical data and estimate the measure of interest in each random analysis in a way that the real measure of the data observed is compared with the simulated data to estimate if they are significantly different.

The study that first described a DNA forensic ancestry test (Lowe et al. 2001) suggested Bayes analysis as a way of assigning an STR profile of unknown ancestry to the most likely population of origin. Bayesian population analysis methods use the combined genotype frequencies that define a population and the frequencies found in individuals identified as originating from that population to calculate their likelihood and assign a probability of ancestry. An advantage of such methods is that prior information about the samples can be used to progress the analysis. But the ability to differentiate populations amongst a set of samples is severely restricted from limited sample sizes and small marker numbers (Corander et al. 2003, Corander and Marttinen 2006) – this fact should be taken in consideration when designing any new assay for ancestry inference.

The software *STRUCTURE* analyses differences in the distribution of genetic variants amongst populations with a Bayesian iterative algorithm by placing samples into groups whose members share similar patterns of variation (Pritchard et al. 2000). *STRUCTURE* uses a systematic Bayesian clustering approach applying Markov Chain Monte Carlo (MCMC) estimation. The MCMC process begins by randomly assigning individuals to a pre-determined number of groups. Then, variant frequencies are estimated in each group and individuals re-assigned based on those frequency estimates. This is repeated many times, resulting in a progressive convergence toward reliable allele frequency estimates in each population and membership probabilities of individuals to a population. *STRUCTURE* both identifies populations from the data and assigns individuals to those populations representing the best fit for the variation patterns found.

One disadvantage of *STRUCTURE* is the difficulty of analysing single profiles, typical of forensic testing, since the whole set of parental data plus the unknown profile must be re-analysed in combination each time and this can be both time consuming and complex to perform for a small number of casework samples in turn. For this reason, an alternative online classifier termed *Snipper* (http://mathgene.usc.es/snipper/) was developed to provide a simple

alternative for making ancestry assignments of single profiles in real-time (Phillips et al. 2007b). It is a naïve Bayesian likelihood ratio (LR) based calculator: it analyses single profiles with a near identical Bayesian algorithm and gives likelihoods of membership to ancestry groups inferred from defined training sets acting as reference material. The *Snipper* site holds training sets, providing the reference data from which allele frequencies are calculated. It also allows a crosscheck of novel training set data by one-out cross validation. Custom training set data uploaded to *Snipper*, potentially offers the most useful option since genotypes generated by the laboratory or collected from online/published sources can be assessed and applied to any AIM panel of interest. The likelihood ratio values usually obtained are usually very high but it is important to establish realistic minimum thresholds below which no assignment is made, helping to minimize error when balanced against a reasonable no-classification rate. This approach was applied in the 11-M Madrid bomb attack ancestry analyses presented in section 1.3.2.3.

The difference between both Bayesian methods lies in how the likelihood is computed (more information about these algorithms is detailed in (Pritchard et al. 2000, Phillips et al. 2007b)). Therefore both algorithms require reference data to calculate allele frequencies for comparison to alleles recorded in profiles of unknown origin. Each algorithm makes the same prior assumption, often untested: that the variables, i.e. the component loci, are independent. For this reason, uniparental data (in the form of haplotypes where all markers are linked) is not readily incorporated into either analysis system, though *STRUCTURE* has scope for the analysis of linked loci (Falush et al. 2003).

As the number of populations increases, the number of dimensions needed to visually represent the pairwise genetic distances also increases. The main idea of multivariate analyses is to help to represent, in a comprehensive way, those multiple dimensions. This is done through the reduction of the dimensionality of a data set composed of a large number of interrelated variables maintaining the maximum proportion of the variation present on that data (Jobling et al. 2014). Principal component analysis (PCA) is a commonly used multivariate analysis method, especially as an exploratory tool and to summarize genetic similarities and differences between groups of populations. This is possible through the transformation of those variables into a new set of metrics (principal components: PCs) that are not related and can be ordered in a way that the first PCs retain most of the variation present in the original data – the graphical representation of the first two or three PCs

summarizes as much of the variation as possible in a comprehensive, graphical way (Jolliffe 2002, Jobling et al. 2014). In the graphics that PCA generates, individuals are represented by points distributed according to their coordinates in two-way or three-way PC comparisons (two or three dimensional plots, respectively). PCA can also be used to represent the relation of an unknown study sample with a set of reference population samples, i.e., the study sample will be represented by a point superimposed onto the PCA plot of the reference population samples.

Additionally to the systems already described, several specific software have been developed to analyse multilocus genetic data through allele frequency estimation (either previously or during the inference process) to assign probable populations of origin to individuals (Rosenberg et al. 2003). Some of the most commonly used software in population genetic are:

- *ADMIXTURE* – this software estimates the maximum likelihood of individual ancestries from multilocus SNP genotypes. It allows automated choice of the population number and use of known ancestral populations in a supervised learning mode. As it uses a fast numerical optimization algorithm, it allows the prompt analysis of larger marker sets (Alexander et al. 2009, Alexander and Lange 2011, Zhou et al. 2011);
- *ADMIXMAP* – a program for modelling admixture, using marker genotypes and trait data on a sample of individuals from a probably admixed population. Used markers should be selected to present strongly differentiated allele frequencies between two or more of the ancestral populations contributing to the admixture (McKeigue et al. 2000, Tang et al. 2005, McKeigue 2007)
- *EIGENSOFT* – a program suite that has two main components: *EIGENSTRAT* uses principal component analysis to correct for population stratification in association studies and *SMARTPCA* for the detection and analysis of population structure (Patterson et al. 2006, Price et al. 2006).

1.3.2.3 An example of AIM markers forensic application – the 11-M Madrid bomb attack investigation

This section describes an example of the application of ancestry informative markers in a forensic case, the 11th of March 2004 terrorist attack that occurred in Madrid. The investigation results were summarized in a study published by Phillips et al. (2009).

On the 11th of March 2004, ten improvised explosive devices (IED) were detonated on four commuter trains entering Madrid during the morning rush hour. This coordinated attack killed 191 people and injured 1.755, being one of the major terrorist attacks in Europe. This was the largest forensic case in Spain – DNA analysis included STR profiling of 226 reference samples and more than 600 samples obtained from fragments that included: i) remnants of the detonated IEDs; ii) suspect vehicles; iii) one undetonated IED found at the El Pozo train station; and iv) personal items collected at the bomb assembly locations or in the houses used by the suspects. MtDNA sequence analysis was also performed for the majority of samples.

Seven complete STR profiles obtained from five personal items and a handprint in the undetonated IED failed to match any of the suspects held at the start of the court case in February 2007. Following a court order, the seven unmatched DNA samples were submitted to a specialized analysis in order to estimate the samples ancestry. The court order specified that the ancestry analysis should be confined to the comparison of European with North African variability in order to differentiate these two population groups alone. Although both mtDNA and Y-chr are informative markers for biogeographic ancestry assignment, the populations compared in this analysis were known to show a proportion of shared haplotypes as a result of recent gene flow across the Western Mediterranean. For that reason, the analysis of such markers alone was not adequate and autosomal AIM-SNPs were the selected alternative.

The analysis was performed using a set of 34 AIM-SNPs (Phillips et al. 2007b). This assay was originally designed to differentiate sub-Saharan African, European and East Asian populations. As a consequence, it does not necessarily include the best SNPs to separate Europeans from North Africans. For that reason, prior to the analysis of the case samples, an estimation of the classification success by two training sets (Spanish *versus* Moroccans) was performed using a Bayesian classifier. It allowed the generation of pairwise likelihood plots

(an example plot and its interpretation is shown in **Figure 21**) to gauge patterns of possible admixture and to perform a cross-validation of the used training sets in order in order to assess accuracy (classification error in each population) and performance (the range of values obtained amongst training set samples).

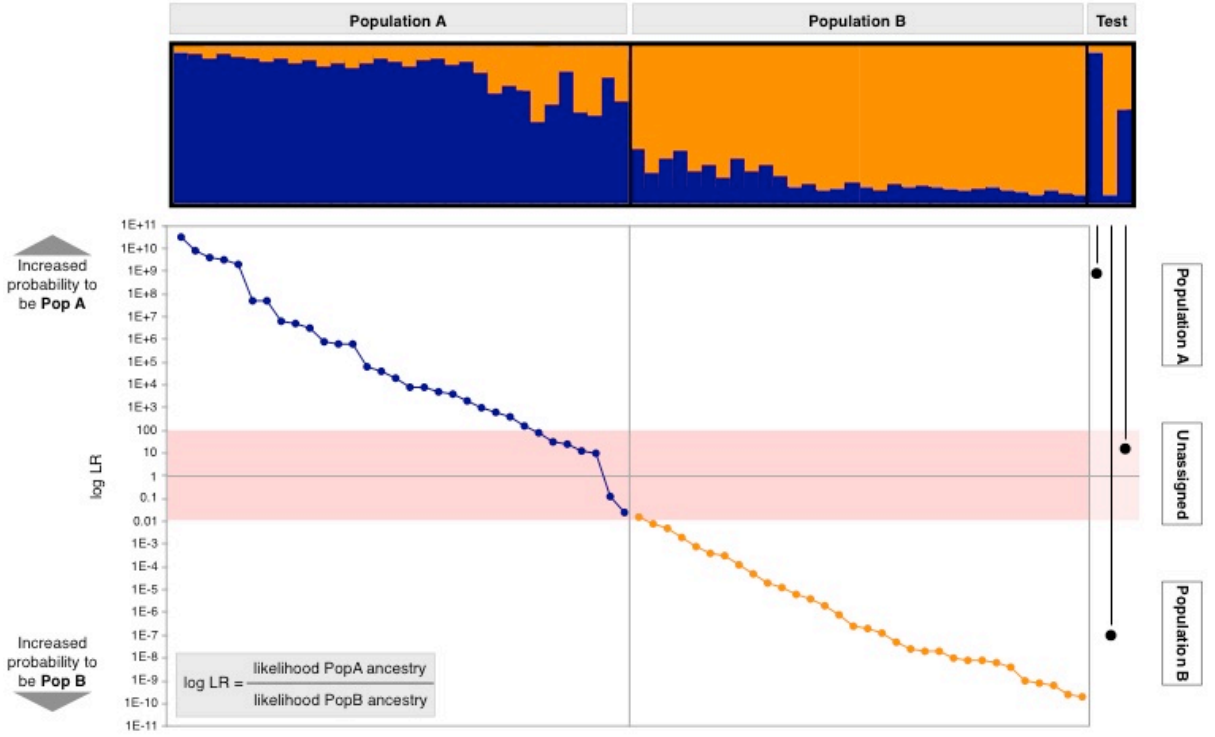


Figure 21: Plot of ranked assignment probabilities obtained from cross validation of two hypothetical training sets (Population A and Population B). Probabilities of example test samples are shown on the right as black circles with probable ancestry assignments shown next to the values. LR is a log scale of the likelihood ratio of (hypothetical Population A classification probability)/(hypothetical Population B classification probability). Values below the midline of 1 (i.e., balanced odds) are reciprocal and the distance from the midline denotes higher probability of ancestry from one of the two populations. Red shaded areas include outlying probability values – samples falling inside that range should be interpreted with caution (such samples might represent admixed individuals, increasing the misclassification risk). A possible *STRUCTURE* analysis plot for *K*=2 is aligned with component samples below. A close match between *STRUCTURE* patterns and assignment probabilities is usually observed.

Results presented by Phillips et al. (2009) showed that three of the samples were more probably North African than European, one was more likely European, and three profiles were not assigned to either populations, that is, the probability values obtained were low to allow a reliable assignment. Those samples probably represented individuals with highly admixed parentage and genomic backgrounds (which is not unusual between Southern Europe and North Africa). One of the assignments obtained using the 34 AIM-SNPs contradicted the Y-chr/mtDNA analysis – the suspect was predicted to be North African by AIM-SNP analysis and European from uniparental data. Although the reported mtDNA and Y-chr haplogroups

are frequently found in Europe, a small proportion was also reported in North Africa. Furthermore, uniparental European and North African databases have highly contrasting sizes, complicating the interpretation of the obtained patterns. For this reason, the results provided by the autosomal AIM-SNPs were determinant for the final conclusions. In fact, the suspect predicted to be North African by AIM-SNP analysis (and European from uniparental data) was identified, later in the investigation, as Algerian in origin.

Such results illustrate the benefits of adding specialized AIM marker sets to the analysis – they can provide enhanced scope and power to an already highly effective system of forensic DNA analysis. But, considering the known populations diversity patterns and the fact that populations from continental margins are expected to show high admixture levels (diminishing populations divergence and increasing misclassification error rates), it is important to continue to develop complementary assays focused on more specific differentiations.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA



OBJECTIVES

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

2 Objectives

2.1 MAIN OBJECTIVES

The research carried out during the present doctoral thesis had three main objectives:

- to understand and apply the most suitable statistical methods to ancestry inference;
- to improve existing forensic ancestry assays and to develop new assays;
- to validate and develop high throughput assays using the Ion PGM™ system.

2.2 SPECIFIC OBJECTIVES

These main objectives included the following specific objectives:

Ancestry informative assays

- to review the design and performance of the SNP*for*ID *34-plex* ancestry informative assay;
- to design a new assay informative for the Pacific region complementary to the SNP*for*ID *34-plex*;
- to design a new assay composed by AIM-Indels and analyse a reference set of samples;
- to design a new assay informative for the Central South Asian region;
- to design a new AIM-STR assay.

Into next generation sequencing

- to verify if already available SNP assay designs could be readily adapted into a next generation sequencing platform;
- to validate the HID-Ion AimpliSeq™ Identity panel using the Ion PGM™ system;

- to select a set of AIM-SNPs with balanced divergence for a five group comparison (Africa, Europe, East Asia, Oceania and America) with the potential to be analysed through next generation sequencing.

Data analysis systems

- to perform an overview of *STRUCTURE* software, its applications and parameter settings;
- to test approaches applicable to the analysis of forensic ancestry marker data: a guide on how to run *STRUCTURE* (and its supporting software), principal component analysis and *Snipper*;
- to assess the possible application of multinomial logistic regression to estimate ancestry proportions.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Author's note:

The main objectives define the organization of the present thesis, this being divided into three thematic blocks. For a better understanding, each block includes a description of the materials and methods used as well as results and discussion of each topic.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

BLOCK 1

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

3 Block 1 – Data analysis systems

Although data analysis is the last step of ancestry inference, this first block describes how to use some of the available informatics tools to estimate biogeographic ancestry and admixture proportions – allowing a better understanding of the results presented in sections 4 and 5. In here, approaches applicable to the analysis of forensic ancestry marker data will be presented, including an overview of one of the most widely used population analysis tools, *STRUCTURE* software (Appendix 1A). Next, a guide on how to run *STRUCTURE* and its supporting software, the use of principal component analysis (including a ready to use *R* script) and of *Snipper* is detailed (Appendix 1B). Finally, results of initial assessments of multinomial logistic regression classification systems to analyse admixed individuals are presented (Appendix 1C).

3.1 Materials and methods

3.1.1 Webpages of interest

- *SPSmart* browser home: http://spsmart.cesga.es/
- *SPSmart* SNP*for*ID *52-plex* and *34-plex* variability browser: http://spsmart.cesga.es/snpforid.php
- Entire genome interface for exploring SNPs (ENGINES) a 1000 Genomes variability browser: http://spsmart.cesga.es/engines.php?dataSet=engines
- pop.STR: http://spsmart.cesga.es/popstr.php
- *forInDel*: http://spsmart.cesga.es/forindel.php
- *Snipper* portal: http://mathgene.usc.es/snipper/
- *STRUCTURE* software: http://pritchardlab.stanford.edu/structure.html
- *Structure harvester*: http://taylor0.biology.ucla.edu/structureHarvester/
- CLUster Matching and Permutation Program (*CLUMPP* software): http://web.stanford.edu/group/rosenberglab/clumpp.html

- *distruct* software: http://web.stanford.edu/group/rosenberglab/distruct.html
- For more information about *R* software: http://www.r-project.org/

3.1.2 Collection of ancestry reference data with the *SPSmart* browser

The statistical analysis of a profile requires reference training sets, i.e., parental populations of interest used to classify casework profiles. Collection of such data previously required locus-by-locus scrutiny of dbSNP or HapMap SNP databases (Phillips 2009), but fortunately *SPS* (SNPs for Population Studies) makes this task much more straightforward for any number of AIM-SNPs as well as up to 52 STRs in routine forensic use.

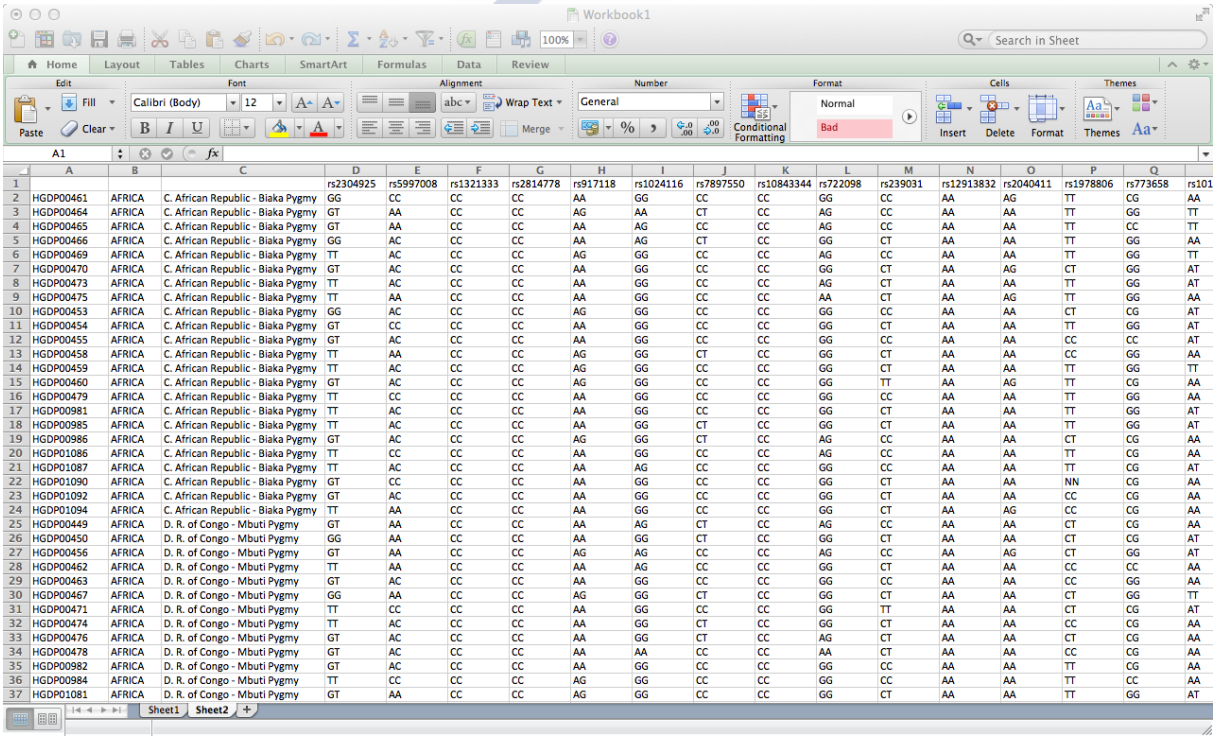
SPSmart (Amigo et al. 2008b) is a simple pre-processing engine that includes five different population based genotype databases: i) 1000 Genomes Phase I May 2011; ii) HapMap Release #28; iii) Perlegen complete data set; iv) HGDP-CEPH Stanford study; and v) HGDP-CEPH NIH-Michigan study (see Note 1). Of all databases included in *SPSmart*, 1000 Genomes and HapMap are of most utility as they have more markers and larger sample sizes (including admixed ancestry populations). This is particularly true for 1000 Genomes – the ENGINES browser allows scrutiny of SNP variation across the whole genome (down to a minor allele frequency of ~1%) from Africans, Europeans and East Asians previously used by HapMap. In contrast, the HGDP-CEPH (Human Genome Diversity Panel) samples 1050 individuals with wide currency in population genetics studies (Cann et al. 2002, Rosenberg 2006). The geographic coverage is patchy in certain regions but all continents and all major genetic ancestry groups defined by studies of variability are represented. Smallest sampling is 28 Oceanians from two populations and just six San from Namibia. *SPSmart* provides freely downloadable genotype data from 650,000 SNPs (obtained with Illumina 650K arrays (Li et al. 2008)) for each HGDP-CEPH sample in the Stanford University study of this panel. *SPSmart* also generates common population genetics indices such as allele frequencies, heterozygosity, F_{ST} or *In* (summarized in the downloads tab of each query).

1. Choose the database(s) to search or choose SEARCH in the SNP*for*ID 'global map' homepage.
2. Choose the populations to merge into groups by ticking selections up to a maximum of five. If opting to review multiple databases, only one population

grouping can be made. Populations are already arranged into sets of genetic diversity based on Rosenberg's original analyses of HGDP-CEPH populations (Rosenberg et al. 2002, Phillips et al. 2007b), though note that Eurasians are subdivided into European, South Asian and Middle East sub-groups.

3. Add the SNP RefSeq (rs-number) identifiers in the search by SNP window. Search by chromosome region or gene name are also available. Click the 'next' button below.
4. Of the filter options presented, MAF is minimum allele frequency and *In* Rosenberg's ancestry informativeness metric (Rosenberg et al. 2003). When reviewing SNP data from multiple databases it is best not to tick option: 'Filter SNPs not genotyped on every compared dataset'.
5. When SNPs are not found from a query, a message tab with the missing rs-numbers appears.
6. In the downloads tab, genotype files are available ready to download, copy and paste into Excel or notepad. The recommended steps being: download, choose all, copy into Excel, transpose the data into rows=samples and columns=SNPs (in edit menu: 'copy'>select new cell>'paste special'>select: 'transpose'). This must be completed for each group while taking care to label each set of sample rows with the appropriate description, e.g. African, South Asian, etc. Another option is to query and download all population groups at once – follow the steps previously described in this point for the genotypes and then download the sample list file. This file includes sample, sub-population and population group information. Open, select all and paste in the genotypes Excel file – confirm that the samples are in the same order and remove the duplicated sample name column (**Figure 22**).
7. The genotypes can be formatted for input to *STRUCTURE*, *Snipper* or PCA data analysis. SNaPshot genotypes may need checking against the reference data if they come from different typing platforms, e.g. a SNaPshot C/T SNP may be an A/G SNP in HapMap requiring base inversions of one dataset. For this reason symmetrical SNPs (C/G and A/T) require particular care as they are a source of error when comparing genotypes generated on different platforms or listed in different databases. The *SPSmart* SNP*for*ID browser makes allowance for most

base inversions by showing the HapMap (or other) allele frequency summary charts with different allele segments if these differ from the strand interrogated by the *34-plex* and *52-plex* extension primers. For example, rs2304925 is listed as a SNP*for*ID GT SNP but a HapMap AC SNP and this applies equally to GC or AT SNPs, e.g. rs10141763. The *SPSmart* help file provides a clear and carefully worded guide in the 'Symmetrical bases' section. There are four symmetrical SNPs in *34-plex*: rs773658, rs10141763, rs1335873 and rs16891982. The latter is the most informative SNP for differentiating populations within Eurasia so it is particularly important to be clear about differences between SNaPshot and database allele calls for this marker.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	
1				rs2304925	rs5997008	rs1321333	rs2814778	rs917118	rs1024116	rs7897550	rs10843344	rs722098	rs239031	rs12913832	rs2040411	rs1978806	rs773658	rs101
2	HGDP00461	AFRICA	C. African Republic - Biaka Pygmy	GG	CC	CC	CC	AA	GG	CC	CC	GG	CC	AA	AG	TT	CG	AA
3	HGDP00464	AFRICA	C. African Republic - Biaka Pygmy	GT	AA	CC	CC	AG	AA	CT	CC	AG	CC	AA	AA	TT	GG	TT
4	HGDP00465	AFRICA	C. African Republic - Biaka Pygmy	GT	AA	CC	CC	AA	AG	CC	CC	AG	CC	AA	AA	TT	CC	TT
5	HGDP00466	AFRICA	C. African Republic - Biaka Pygmy	GG	AC	CC	CC	AA	AG	CT	CC	GG	CT	AA	AA	TT	GG	AA
6	HGDP00469	AFRICA	C. African Republic - Biaka Pygmy	TT	AC	CC	CC	AG	GG	CC	CC	AG	CC	AA	AA	TT	GG	TT
7	HGDP00470	AFRICA	C. African Republic - Biaka Pygmy	GT	AC	CC	CC	AA	GG	CC	CC	GG	CT	AA	AG	CT	GG	AT
8	HGDP00473	AFRICA	C. African Republic - Biaka Pygmy	TT	AC	CC	CC	AA	GG	CC	CC	AG	CT	AA	AA	TT	GG	AT
9	HGDP00475	AFRICA	C. African Republic - Biaka Pygmy	TT	AA	CC	CC	AA	GG	CC	CC	AA	CT	AA	AG	TT	GG	AA
10	HGDP00453	AFRICA	C. African Republic - Biaka Pygmy	GG	AC	CC	CC	AG	GG	CC	CC	GG	CC	AA	AA	CT	CG	AT
11	HGDP00454	AFRICA	C. African Republic - Biaka Pygmy	GT	CC	CC	CC	AA	GG	CC	CC	GG	CT	AA	AA	TT	GG	AT
12	HGDP00455	AFRICA	C. African Republic - Biaka Pygmy	GT	AC	CC	CC	AA	GG	CC	CC	GG	CC	AA	AA	CC	CC	AT
13	HGDP00458	AFRICA	C. African Republic - Biaka Pygmy	TT	AA	CC	CC	AG	GG	CT	CC	GG	CT	AA	AA	CC	GG	AA
14	HGDP00459	AFRICA	C. African Republic - Biaka Pygmy	TT	AC	CC	CC	AG	GG	CC	CC	GG	CT	AA	AA	TT	GG	TT
15	HGDP00460	AFRICA	C. African Republic - Biaka Pygmy	GT	AC	CC	CC	AG	GG	CC	CC	GG	TT	AA	AG	TT	CG	AA
16	HGDP00479	AFRICA	C. African Republic - Biaka Pygmy	TT	CC	CC	CC	AA	GG	CC	CC	GG	CC	AA	AA	TT	GG	AA
17	HGDP00981	AFRICA	C. African Republic - Biaka Pygmy	TT	AC	CC	CC	AA	GG	CC	CC	GG	CT	AA	AA	TT	GG	AT
18	HGDP00985	AFRICA	C. African Republic - Biaka Pygmy	TT	AC	CC	CC	AA	GG	CT	CC	GG	CT	AA	AA	TT	GG	AT
19	HGDP00986	AFRICA	C. African Republic - Biaka Pygmy	GT	AC	CC	CC	AG	GG	CT	CC	AG	CC	AA	AA	CT	CG	AA
20	HGDP01086	AFRICA	C. African Republic - Biaka Pygmy	TT	CC	CC	CC	AA	GG	CC	CC	AG	CC	AA	AA	TT	CG	AA
21	HGDP01087	AFRICA	C. African Republic - Biaka Pygmy	TT	AC	CC	CC	AA	AG	CC	CC	GG	CC	AA	AA	TT	CG	AT
22	HGDP01090	AFRICA	C. African Republic - Biaka Pygmy	GT	CC	CC	CC	AA	GG	CC	CC	GG	CT	AA	AA	NN	CG	AA
23	HGDP01092	AFRICA	C. African Republic - Biaka Pygmy	GT	AC	CC	CC	AA	GG	CC	CC	GG	CT	AA	AA	CC	CG	AA
24	HGDP01094	AFRICA	C. African Republic - Biaka Pygmy	TT	AA	CC	CC	AA	GG	CC	CC	GG	CT	AA	AG	CC	CG	AA
25	HGDP00449	AFRICA	D. R. of Congo - Mbuti Pygmy	GT	AA	CC	CC	AA	AG	CT	CC	AG	CC	AA	AA	CT	CG	AA
26	HGDP00450	AFRICA	D. R. of Congo - Mbuti Pygmy	GG	AA	CC	CC	AA	GG	CT	CC	GG	CT	AA	AA	CT	CG	AT
27	HGDP00456	AFRICA	D. R. of Congo - Mbuti Pygmy	GT	AA	CC	CC	AG	AG	CC	CC	AG	CC	AA	AG	CT	GG	AT
28	HGDP00462	AFRICA	D. R. of Congo - Mbuti Pygmy	TT	AA	CC	CC	AA	AG	CC	CC	GG	CT	AA	AA	CC	CC	AA
29	HGDP00463	AFRICA	D. R. of Congo - Mbuti Pygmy	GT	AC	CC	CC	AA	GG	CC	CC	GG	CC	AA	AA	CC	GG	AA
30	HGDP00467	AFRICA	D. R. of Congo - Mbuti Pygmy	GG	AA	CC	CC	AG	GG	CT	CC	GG	CT	AA	AA	CT	GG	TT
31	HGDP00471	AFRICA	D. R. of Congo - Mbuti Pygmy	TT	CC	CC	CC	AA	GG	CC	CC	GG	TT	AA	AA	CT	CG	AT
32	HGDP00474	AFRICA	D. R. of Congo - Mbuti Pygmy	TT	AC	CC	CC	AA	GG	CT	CC	GG	CT	AA	AA	CC	CG	AA
33	HGDP00476	AFRICA	D. R. of Congo - Mbuti Pygmy	GT	AC	CC	CC	AA	GG	CT	CC	AG	CT	AA	AA	CT	CG	AA
34	HGDP00478	AFRICA	D. R. of Congo - Mbuti Pygmy	GT	AC	CC	CC	AA	AA	CC	CC	AA	CT	AA	AA	CC	CG	AA
35	HGDP00982	AFRICA	D. R. of Congo - Mbuti Pygmy	GT	AC	CC	CC	AA	GG	CC	CC	GG	CC	AA	AA	TT	CG	AA
36	HGDP00984	AFRICA	D. R. of Congo - Mbuti Pygmy	TT	CC	CC	CC	AG	GG	CC	CC	GG	CC	AA	AA	TT	CC	AA
37	HGDP01081	AFRICA	D. R. of Congo - Mbuti Pygmy	GT	AA	CC	CC	AG	GG	CC	CC	GG	CT	AA	AA	TT	GG	AT

Figure 22: Example of a reference ancestry genotype data file obtained from *SPSmart*. The data obtained from *SPSmart* was reorganized (original data was transposed so that samples are now organized in rows and markers in columns) and population information (downloaded from *SPSmart* in a separate file) was added.

3.1.3 *STRUCTURE* software

3.1.3.1 An overview of the *STRUCTURE* program

STRUCTURE is a freely available program for population analysis developed by Pritchard, Stephens and Donnelly (2000). *STRUCTURE* analyses differences in the distribution of genetic variants amongst populations with a Bayesian iterative algorithm by placing samples into groups whose members share similar patterns of variation. *STRUCTURE* both identifies populations from the data and assigns individuals to that population representing the best fit for the variation patterns found. Typically *STRUCTURE* is the first step in examining population structures that emerge from the sample set to provide a preamble to further genetic analysis or to infer the origins of individuals with unknown population characteristics, especially when population admixture has occurred. As *STRUCTURE* uses the core Bayesian principle of comparing likelihoods, prior information about study samples can be supplied to further shape the analysis. For example, information about sampling location can be input – a characteristic that, if shared between individuals, can be associated with their genetic proximity. The definition of populations can be assessed from geographical distribution, but is also often based on alternative criteria, including the phenotype, behaviour and ecology of sampled individuals, while linguistic and cultural characteristics can also define human populations. Therefore it is important to assess whether assignment of individuals to populations with non-genetic criteria is consistent with genetic patterns detected between populations (Pritchard et al. 2000, Waples and Gaggiotti 2006, Jobling et al. 2014).

STRUCTURE uses a systematic Bayesian clustering approach applying Markov Chain Monte Carlo (MCMC) estimation. The MCMC process begins by randomly assigning individuals to a pre-determined number of groups, then variant frequencies are estimated in each group and individuals re-assigned based on those frequency estimates. This is repeated many times, typically comprising 100,000 iterations, in the burnin process that results in a progressive convergence towards reliable allele frequency estimates in each population and membership probabilities of individuals to a population.

Measurement of the assumed number of populations uses the MCMC estimation and is performed separately from the burnin. *STRUCTURE* performs individual analyses for each

assumed population number from one to a reasonably appropriate number for the sampling regime. *STRUCTURE* applies a model to the data of *K* assumed populations or genetic groups, each characterized by a subset of allele frequencies identified in the data. Commonly *K* is not readily defined by the user for the sample set, although this parameter must be pre-selected. Therefore an appropriate first step is to calculate the likelihood of the data for a range of *K* values by creating posterior probabilities of *K*, termed *X* and written: *X*|*K*. Since *K* is not an absolute value, user-defined values should be considered carefully, taking into account characteristics of the sampled populations. Running a range of prescribed *K* settings to obtain their *X* values normally creates probabilities smaller than those for the most appropriate *K* value but beyond this probabilities tend to be very similar for higher *K* values. Therefore plots of *X* values typically progress to a plateau for levels of *K* beyond the most applicable number of detected populations, so the smallest stable *K* value represents the optimum value. Kalinowski (2011) notes that better clusters are created applying the most realistic *K* values, so it is prudent to obtain the smallest value of *K* that maximizes the global likelihood of data – an approach capturing the major underlying population structure in the data without overestimating it.

During each analysis membership coefficients summing to one are assigned to individuals for each group. The membership coefficient matrix, termed the *individual Q-matrix*, is generated with rows for the number of individuals analysed and columns for *K* clusters. The average individual membership coefficients to each population form the *population Q-matrix*. If admixture is not a factor for the population samples analysed, posterior probabilities of belonging to each of *K* groups are calculated for each individual and a sample can be considered a member of the group with the highest probability. If admixture is considered membership coefficients are made across multiple clusters.

Bayesian population analysis methods equate allele frequencies that define the population and the frequencies found in individuals identified as originating from that population. Therefore the ability of Bayesian methods to differentiate populations amongst a set of samples is severely restricted from limited sample sizes and small marker numbers (Corander et al. 2003, Corander and Marttinen 2006). Genetic markers applied to *STRUCTURE* analyses ideally show selective neutrality, low mutation rates and absence of linkage disequilibrium (Pritchard et al. 2000, Corander et al. 2003). In short, they are treated as independent variables and so-called naïve Bayesian approaches assume independence, but without a guarantee this

applies. However, enhancements since *STRUCTURE* 2.3.1 permit inclusion of weakly linked markers with some degree of non-independence (Falush et al. 2003). Additionally, Falush et al. (2007) describe an algorithm that allows application of all available models to dominant markers (loci often characterised by genotype ambiguity).

Note that SNPs, as binary markers, have lower variability than multiple-allele loci, requiring much smaller sample sizes to obtain accurate allele frequency estimates. Shi et al. (2010) suggested population samples as low as four individuals are sufficient to provide reliable data by demonstrating subsamples of four taken from much larger population samples give very similar posterior median population parameters. Microsatellites require considerably larger samples sizes than SNPs to reliably capture patterns of variability in a population. Lastly, recent forensic guidelines for microsatellite population surveys recommend genotyping a minimum 500 samples (Carracedo et al. 2013).

3.1.3.2 Ancestry and allele frequency models

Two terms are relevant to a review of the estimation of ancestry: local ancestry and global ancestry. Local ancestry estimates the extent to which each person's genome is divided into chromosome segments of definite ancestral origin. Global ancestry estimates the proportion of ancestry from each contributing population, considered as an average over the individual's entire genome (Alexander et al. 2009). *STRUCTURE* only estimates global ancestry by implementing different models of population structure to the data. Selection of the most appropriate model depends on the user's data and study objectives. Two ancestry models applied by *STRUCTURE* are the *no admixture* and *admixture* models. If there is no prior knowledge about the origin of the populations under study or if there is reason to consider each population as completely discrete, the *no admixture* model is appropriate. However admixture between populations is a common characteristic such that a large proportion of sampled individuals can have recent ancestors from multiple populations. In these cases knowing the approximate median value of the ancestral population proportions for each individual and/or their populations of origin is a very useful part of the characterisation of the populations under study. In these cases the *admixture* model is more appropriate. Both models can be used with consideration for sampling location information by applying the prior model parameter: *LOCPRIOR* to the population model (Hubisz et al. 2009). This option can be used

when there is additional sample-characteristic data available to the user, including: linguistic, geographical, cultural or phenotypic information. The *LOCPRIOR* parameter is particularly informative when there are weak population structure signals – a situation that can result from using reduced numbers of markers, small sample sizes or due to close relationships between populations.

The third model parameter is *linkage*, based on the *admixture* model and this is designed to deal with admixture linkage disequilibrium (LD): the characteristic of extended LD found in admixed populations and often deliberately sought in association studies. This model was outlined by Falush et al. (2003) and provides more accurate estimates of statistical uncertainty when linked markers are used.

No admixture, *admixture* and *linkage* models can also be analysed as part of the *USEPOPINFO* model. This model uses the population labels to calculate the probability that each individual has of originating from the assumed population – individuals with low probabilities can be considered as hybrids or migrants (Pritchard et al. 2000). This parameter should be used cautiously – applied only when the population labels are well defined beforehand and correspond almost exactly to the groups ultimately defined by the *STRUCTURE* results. The disadvantage of the *USEPOPINFO* model arises with the posterior handling of the results. The *individual Q-matrix* comprises probabilities (and not ancestry membership proportions) that are presented in a format that is incompatible with post-hoc data-processing software such as *CLUMPP* or *distruct*.

All the models considered until now can be used in conjunction with an alternative approach to *USEPOPINFO* – the *POPFLAG* model. *POPFLAG* considers the specified information about the population of origin of a portion of the individuals to help infer the ancestry of other samples with unknown origin. This option should also be used with caution because selected samples will be treated as the "reference" set (pre-assigned *POPFLAG*=1) so allele frequency estimates are based on a reduced sub-set of samples and will directly affect the grouping of the unknown individuals (pre-assigned *POPFLAG*=0). *POPFLAG* is an artificial model that assesses the individual probability of being part of a particular population, but it can be useful if the objective is to efficiently group individuals/populations by comparison with a particularly well-defined and studied reference data set (Pritchard et al. 2000). One such reference set, widely applied to human population genetics studies, is the CEPH human genome diversity panel (HGDP-CEPH) (Cann et al. 2002) with the advantage

that population structure has been identified in this sample set in a wide range of studies using different markers and a variety of data depths, but with consistent findings (Rosenberg et al. 2002, Enoch et al. 2006, HPAC et al. 2009). When the *POPFLAG* model is used in conjunction with the *USEPOPINFO* model the *individual Q-matrix* is composed of two distinct parts: for *POPFLAG*=1 individuals the matrix presents probabilities, while for *POPFLAG*=0 individuals ancestry membership proportions are given according to the admixture model defined (*no admixture*, *admixture* or *linkage*).

Related subjects should be detected then excluded from reference data since shared variation is inflated in frequency causing estimation bias that consequently affects analysis of the ancestry of study samples. Prudent checks of reference data are clearly advisable, for example, an audit of the HGDP-CEPH samples was published by Rosenberg (2006) where two atypical individuals, 13 duplicates and as many as 77 first-degree relative pairs were identified and removed. Related subjects can regularly form part of the study set and this approach is central to linkage studies in genetic epidemiology as these seek loci with correlations between traits of interest and patterns of transmission of DNA sequence over generations in a known pedigree (Astle and Balding 2009).

Two allele frequency models are available. The *correlated allele frequencies* model assumes a level of non-independence, so is more conservative. The *independent allele frequencies* model requires knowledge about the correlation levels across populations – allele frequencies should be reasonably different in distinct populations. The *correlated allele frequencies* model provides greater power to detect distinct populations that are particularly closely related, although this model will give the same results as the *independent allele frequencies* model in the absence of high levels of correlation across populations. Therefore it is prudent to use the *correlated allele frequencies* model since this will guarantee that a previously undetected correlation is identified, but without affecting the results if no such correlation exists.

Summarising the range of models that can be applied to a uniform set of genetic data, **Figure 23** shows bar plots produced by *STRUCTURE* when applying each of the models described above.

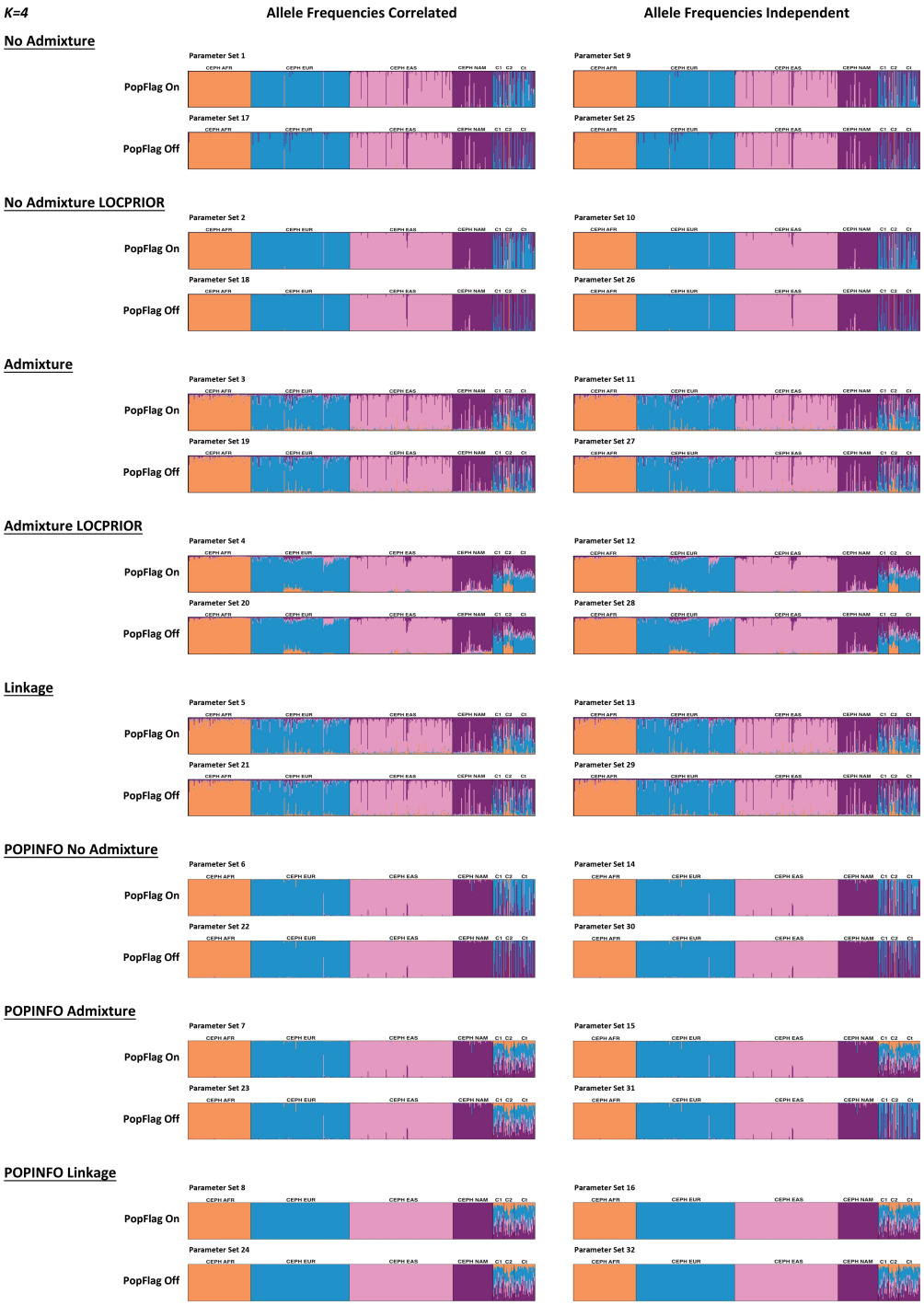


Figure 23: *STRUCTURE* bar plots representing *K*=4 for the principal analysis parameter combinations available to the user. These graphics were obtained with *distruct* and using *CLUMPP* to align the three replicates for *K*=4 (all runs were performed with 100,000 burnin period and 100,000 MCMC repeats after burnin). The exception was the *POPINFO* parameter sets for which direct *STRUCTURE* bar plot outputs were used. Human genetic data comprised genotypes consisting of 100 Africans: CEPH AFR, 158 Europeans: CEPH EUR, 165 East Asians: CEPH EAS, and 64 Native Americans: CEPH NAM from the HGDP-CEPH human diversity panel. An artificial case-control group was created using HapMap Mexican and Puerto Rican samples giving a total 67 sample divided into Cases 1 (C1), Cases 2 (C2), and Controls (Ct). Markers were: 9 AIM-SNPs (two triallelic), 3 phenotype associated SNPs and 5 AIM-SNPs on the X-chromosome. The phenotype and the X-SNPs are linked forming two distinct linkage disequilibrium groups – their genetic distance was used to define linkage disequilibrium groups. Each parameter setting and the results obtained are described in detail in Appendix 2B.

It is noteworthy that a level of finesse exists when implementing analysis models in *STRUCTURE*. All the models described above include specific statistical parameters that can be adjusted to more sensitive values, including *r* (informativeness of the sampling location data), *alpha* (relative admixture levels between populations) and *lambda* (quantifies the independence between markers in terms of their allelic frequency distribution).

Considering the type of analysis required in a forensic context, the *admixture POPFLAG* ancestry model is appropriate. This combines two important features: consideration of admixture between populations (individuals can have recent ancestors from multiple populations so ancestry membership proportions from each ancestral population can be calculated); and some individuals can be used as a reference to help infer the ancestry of the samples under study. Regarding the allele frequencies model, it is advisable to use the correlated allele frequencies model because it will guarantee that an undetected correlation will be identified without affecting the results should it be absent.

3.1.3.3 Preparation of a *STRUCTURE* input file

Data to be analysed with *STRUCTURE* needs to be organized in a single matrix (as a text file) where optional information can be considered to complement the genotypic data. Such information should be included in a pre-defined order and it is important to highlight that only the genotypic data is required for the analysis. We will focus the construction of an input file on the information of greater relevance when analysing a casework profile (Table 4).

- first line: header line. Headers are only included in the markers columns;
- first column: sample name information that can be an alpha-numerical code which can introduce errors when running *CLUMPP*, easily solved in *Structure Harvester*;
- second column: a numerical code representing the population of origin as defined by the researcher. By default this information is not used by the clustering algorithm but can help organize the output file;
- third column: *PopFlag* information. This is a Boolean variable where 1 (TRUE) represents the samples that should be used as reference and 0 (FALSE) the casework/study samples;

- fourth column: *LocPrior* information. A numerical code that denotes sub-population groups, geographical locations or other shared characteristic between individuals inside the population groups defined in the second column. This information is used when considering the *LOCPRIOR* ancestry model;
- any number of extra columns can list useful information for the researcher. For example, as the population and *LocPrior* information is numeric, extra columns with the names can be included as an easy way to cross check data later;
- the following columns include genotypic data for any number of markers (SNPs, Indels or multi-allelic markers such as STRs). Genotypes should be coded as numbers. For SNPs, we routinely use *A=1*; *C=2*; *G=3* and *T=4*. STRs are already numerically coded but in the case of intermediate alleles the "." should be removed, i.e., 19.3=193 – *STRUCTURE* does not assume a particular mutation process so the scale of the number of repeat units in STRs is not considered (only allele frequencies are important). For this reason there is no need to multiply all other alleles by 10 to compensate the transformation of intermediate alleles;
- missing data is usually coded as -9 but any other code not present in the file can be used;
- each allele needs to be represented in a separate cell: both alleles in the same line but in different columns or both alleles in the same column but in different lines (we will focus on the latter as shown in Table 4);
- spaces should not be included.

Table 4: *STRUCTURE* input file format – samples *S1*…*Sn* from populations *1*…*n* analysed with genotypic data from markers *M1*…*Mn*. Samples 1…10 belong to population 1 and can be divided into three locations (1-3). Samples 11…20 belong to population 2 and can be divided into three locations (4-6). Samples 21…*Sn* belong to population *n* and can be divided into three locations (7-9). Samples from populations 1 and 2 are reference (POPFLAG=1) and study samples are from population n (POPFLAG=0).

						M1	M2	M_n
S1	1	1	1	extra	extra	1	2	1
S1	1	1	1	extra	extra	3	4	1
S2	1	1	2	extra	extra	3	2	2
S2	1	1	2	extra	extra	3	2	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
S10	1	1	3	extra	extra	1	4	1
S10	1	1	3	extra	extra	1	4	2
S11	2	1	4	extra	extra	1	2	2
S11	2	1	4	extra	extra	3	4	2
S12	2	1	5	extra	extra	3	2	1
S12	2	1	5	extra	extra	3	2	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
S20	2	1	6	extra	extra	1	4	2
S20	2	1	6	extra	extra	1	4	2
S21	n	0	7	extra	extra	1	2	1
S21	n	0	7	extra	extra	2	4	1
S22	n	0	8	extra	extra	2	2	2
S22	n	0	8	extra	extra	2	2	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
S_n	n	0	9	extra	extra	1	4	**-9**
S_n	n	0	9	extra	extra	1	4	**-9**

3.1.3.4 How to run *STRUCTURE* software

The first stage when running *STRUCTURE* is to create a new project (File > New project) following four established steps:

- step 1 – project information: name of the project, directory where the project will be saved and input file.
- step 2 – information of input data set: number of individuals, ploidy of data, number of markers and missing data value. To confirm the number of markers and individuals select "Show data file format" showing total lines and columns. As an example the input file in Table 4: indicates one line with *m* columns (*m* corresponds to the number of markers) and *n* lines with *m*+6 columns (four columns with prior information and two with extra information) with *n*/2 individuals (two lines per genotype).

- steps 3 and 4 – format of input data set: information contained in rows and columns (for example, row of marker names or individual ID for each individual). When preparing the input file following the example presented in Table 4 there is no need to select the "special format" option because by default *STRUCTURE* assumes genotypes are arranged as two consecutive rows (diploid species) per individual. If both alleles are in the same line but in consecutive columns select "Data file stores data for individuals in a single line".

Before creating the project, *STRUCTURE* presents a summary where it is possible to confirm the selected options. If there are no errors the project opens and the data is visible.

The next stage is to create a new parameter set (Parameter set > New):

- run length – a burnin period of 100,000 is more than sufficient to allow a progressive convergence towards reliable allele frequency estimates in each population and probabilities for membership of individuals to a population. Measurement of the assumed number of populations uses the MCMC estimation and is performed separately from the burnin. About 100,000 MCMC repeats have been shown to provide good ancestry membership proportions estimates. But burnin and MCMC repeat number should be adjusted depending on the study objectives and information contained in the data set.
- select ancestry model – depending on the study objectives and the data to be analysed, different ancestry models can be considered. For the *admixture POPFLAG* model select "Use admixture model" under the "Ancestry model" tab and "Update allele frequencies using only individuals with POPFLAG=1 data" under the "Advanced" tab.
- select allele frequencies model – for the correlated allele frequencies model select the "Allele frequencies correlated" option under the "Allele frequencies model" tab.
- leave the "Compute probability of the data (for estimating *K*)" option under the "Advanced" tab selected so that posterior LnPr(*K*) values are calculated – those will be used to estimate the optimum *K* value.

- save the new parameter set with the desired name and confirm the selected options in the summary window that opens after saving. A tree on the left side of the screen will include all the parameter sets created, indicating the one active at the moment.

There are two ways of starting a simulation:

- run a single *K* value – in the "Parameter set" menu select "Run" and set the assumed number of populations (*K*). This option only allows a single *K* value and replicate at a time.
- schedule multiple runs – in the "Project" menu select "Start a job". A new window opens – select the parameter set(s) to be analysed, the *K* values and the number of iterations for each *K*. *STRUCTURE* performs individual analyses for each assumed population number from one up to a reasonable number for the sampling regime – at least three *K* values more than the number of expected populations. If a *plateau* on the posterior probabilities is not reached, larger *K* values might be needed. Furthermore, clustering algorithms such as the one implemented in *STRUCTURE* can show stochastic variation from the simulations. To diminish their effect, several replicates for each *K* value should be made (at least three to five replicates advised). This option is advantageous for large projects – a new run starts automatically after the previous one has finished so there is no need for constant attention on the progress of the job.

3.1.3.5 Estimate the optimum number of populations

The estimation of the optimum *K* value is performed in an independent way: for each simulation a posterior probability value: Pr(*K*) is calculated. In general, for *K* values under the optimum, Pr(*K*) is low but it tend to stabilize with higher *K* values, so a plateau is commonly observed. When several *K* values have similar LnPr(*K*) estimates, the smallest of them is usually the most appropriate estimate – generally corresponding to the inflection point of the probability plot. However, it is not always possible to know the real *K* value, so it is best to choose the lowest value that captures the maximum structure present in the data (Kalinowski

2011). Evanno et al. (2005) described a method to estimate *K* based in a second order change of probability value taking into account the variability of the probability value between different replicates of each *K* value – delta *K* (Evanno et al. 2005). This method is implemented in *Structure Harvester* (Earl and vonHoldt 2012). After running *STRUCTURE*, pass the zipped folder containing the results files (named *x_run_y_f*, where *x* represents the parameter set name and *y* the run number), to the *Structure Harvester* browser and click Harvest! to start the analysis. Conditions are: a minimum of three sequential *K* values should be included, with more than one replicate for each *K* value (same number of replicates for all *K* values). Posterior probability and *delta K* plots are available to download (**Figure 24**) – the optimum *K* is usually the point with the highest *delta K* value or the one which immediately precedes the LnPr(K) plateau. This software is also useful as it automatically generates input files to run *CLUMPP*.

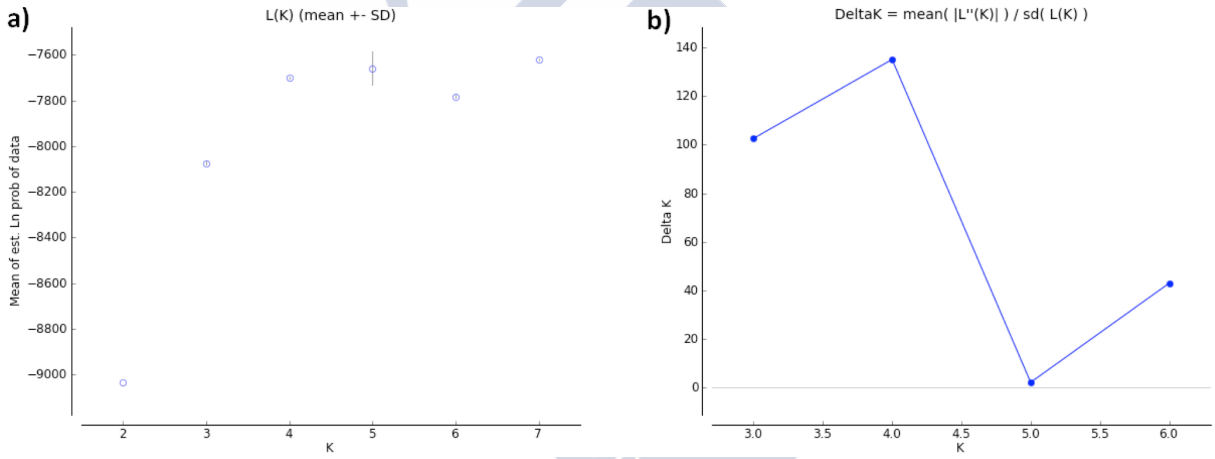


Figure 24: Example of posterior probability and *delta K* plots obtained with *Structure Harvester* for the same analysis. In this case, the optimum *K* value is 4 - the point where the plateau in the posterior probability starts with maximum *delta K* value.

However *K* is not an absolute value, defined values should be carefully considered taking into account any known characteristic of the studied populations. There are several factors that can affect the clustering of the samples: i) number of markers; ii) number of samples; iii) number of clusters; and iv) allele frequency correlation. Endogamy or genotyping errors can have the same effects as true population structure as they can simulate linkage disequilibrium in unlinked markers and deviations from HWE.

3.1.3.6 *STRUCTURE* associated software

Software associated with *STRUCTURE*, for example, *CLUMPP* (Jakobsson and Rosenberg 2007), contains three algorithms for the alignment of multiple replicate analyses of the same data set which allows the transformation of any number of replicate simulations for each *K* in a single set of data. Such data is suitable for analysis with another supporting program: *distruct* (Rosenberg 2004) which allows the visualization of the estimated membership coefficients: populations are represented as colours and individuals as bars portioned into coloured segments that correspond to membership coefficients in the groups.

3.1.3.6.1 CLUMPP

Independently of the origin of differences between clustering results, a method to deal with replicate results is needed. *CLUMPP* uses replicates of the estimated membership proportion matrices for any *K* number – the result is a set of permutated matrices so that all the replicates have the best correspondence possible. It also generates a matrix that corresponds to the median of the permutated matrices. This is done for the population and individual proportions matrices. Currently the easiest way to prepare input files for *CLUMPP* is with *Structure Harvester*. Two files are needed: *.indfile* includes individual ancestry membership proportions tables from all replicates per *K* value and *.popfile* includes population ancestry membership proportions tables for all replicates per *K*. In both a blank line separates each table. If the input files are prepared manually take care with the sample name, which must be numeric; if alpha-numerical an error message appears. Both *.indfile* and *.popfile* files, together with *paramfile* and others, must be saved in the software folder together with the executable file. The paramfile includes important parameters that must be adjusted: *DATATYPE* defines which data is going to be considered for analysis (0=individual; 1=population); *K* is the number of clusters; *C* is the number of individuals or populations (depending on the selected *DATATYPE*); *R* is the number of replicates; *M* is the algorithm used. We recommend *M*=1 so all possible permutations are performed, but with large *K* or *R* values *M*=2 (10,000 random input repeats) is sufficient and for *K* values above 15 *M*=3 is advisable. Metric *S* is the pairwise matrix similarity statistic and we recommend the standardized *G'* (select *S*=2). It is important to note that to obtain a population and individual

mean matrices, two runs are required, adjusting the *.output* file name (no name change overwrites the first run), *DATATYPE* and *C* in between runs. In the Windows version, just execute the *CLUMPP.exe* file – a cmd window opens showing the progress of the simulation. In the MacOS version execute the software through the terminal command line: change the directory to the *CLUMPP* folder location (*type cd > drag the folder into the terminal > ENTER*) and then type *./CLUMPP > ENTER* to run the software.

3.1.3.6.2 distruct

A convenient way of visualizing *STRUCTURE* results (especially for *K*>3) is to show each individual as a straight segment divided into *K* colours that represents the estimated ancestry membership proportions. *STRUCTURE* gives such bar plots but their format cannot be changed and they only present replicate results for one *K* value. *Distruct* offers a great variety of options to generate more informative cluster plots. As with *CLUMPP*, *distruct* uses a set of files stored in the same folder of the executable file. The input files include the *population Q-matrix* (*.popq* file) and the *individual Q-matrix* (*.indivq* file) obtained directly from *STRUCTURE* (in the case of a single *K* replicate) or from *CLUMPP*. Files: *.names* and *.languages* define the labels above and below the plot. Both files have the same format: in each line write the population numeric code, space and preferred name. To define the colour of each cluster open the *.perm* file – with a minimum *K* lines each defining a colour. Colours are assigned to each cluster and not to each predefined population, i.e, if population 1 appears in orange and this population is defined in cluster 3 in *.perm*, define "3 orange" and not "1 orange". The *drawparams* file has several modifiable parameters, notably: *K*, *NUMPOPS* (number of populations) and *NUMINDS* (number of individuals). Remaining parameters adjust graphical aspects of the plot: letter size, distance between text and plot, height of the plot, thickness of the columns representing individuals, thickness of the contour lines, horizontal/vertical orientation, and others (**Figure 25**). The "//" symbol indicates that the following text is a comment describing the parameter function and that it will not be used by the software. When computations are complete a *PostScript* (*.ps*) file containing the plot is created. To visualize the plot with Windows, specific software such as *GhostView* (freely available online), Acrobat Distiller or Illustrator is needed. With MacOS plots can be opened with preview and exported as a pdf.

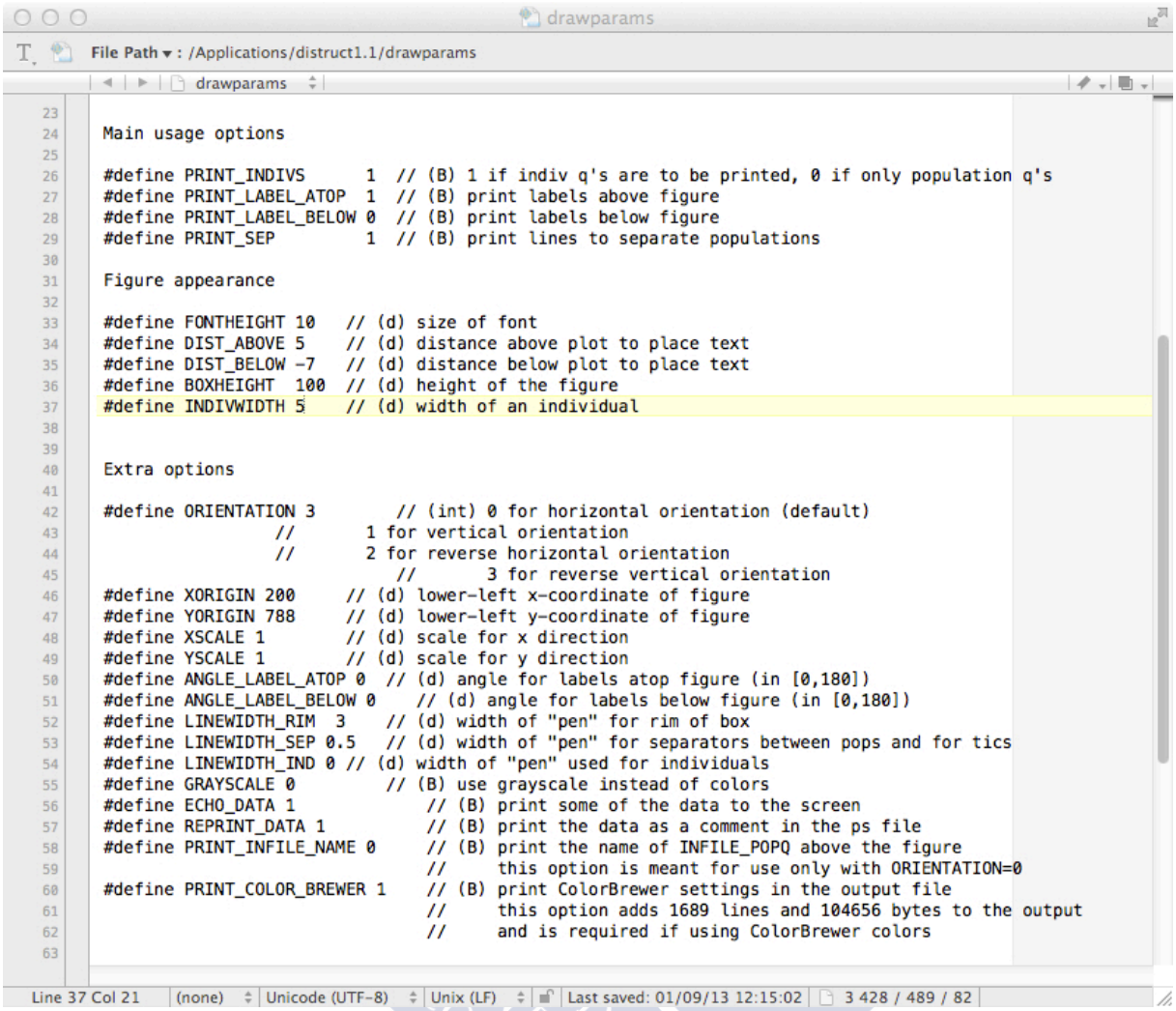


Figure 25: Example of *distruct* parameters. Considering an output file resembling an A4 sheet lying horizontally (longer side down), and depending on the number of samples and the desired effect, it is worth taking advantage of available space. Reverse horizontal orientation uses the longest side of the virtual sheet. Changing the *XORIGIN* and *YORIGIN* values also helps – for example, *XORIGIN*=200 moves the plot away from the margin of the sheet and *YORIGIN*=788 leaves just enough space to separate the plot from the margin without leaving much unused space. The individual bar width (*INDIWITH*) can then be adjusted to an appropriate value that allows all the individuals to be included in the virtual sheet.

3.1.4 The *Snipper* web portal

3.1.4.1 Background on *Snipper* analysis

The *Snipper* portal includes a straightforward Bayesian system for predicting ancestral origin and to estimate the misclassification rate. It uses a set of samples of each population as training sets and assigns individuals to the population that maximizes the posterior probability (maximum likelihood calculation) (Phillips et al. 2007b). The likelihood parameters are

estimated from training set allele frequencies assuming HWE and independence for the used loci. A new option has been recently added to *Snipper* – the ability to apply the HWE principle. Departure from HWE for the markers means less information and classifying power, but an assumption of HWE can inflate the differentiation power of the data that may not be present. As an example the highly informative SNPs: rs26722 and rs16891982 are sometimes included in the same ancestry SNP set, but a separation of 12,177 bases may not be enough to ensure they are free from HWE and can be treated as independent indicators of ancestry.

Snipper was originally designed to provide a real-time ancestry assignment system for *34-plex* profiles with reference to default pre-typed AFR-EUR-E ASN training sets and this still represents the simplest approach for assessment of a single casework profile to obtain an immediate overview of ancestry. The portal has been recently updated to include *34-plex* (Phillips et al. 2007b, Fondevila et al. 2013) and *AIM-Indel* (Pereira et al. 2012b) fixed reference data for five populations groups: AFR-EUR-E ASN-AME-OCE. But the ancestry analyses can be extended beyond the default settings. For example, custom Excel files (including any binary markers that are of interest for the researcher) or frequency based Excel files (helpful when working with STRs or haplotypes) can be used as reference training sets.

A new version of *Snipper* was recently prepared (*Snipper App suite version 2.0*) to include new functionalities including: turn on/off the HWE assumption; prediction of admixture components; batch analysis (multiple profiles); fine-tuning of a training set; classification of single profiles and analysis of training sets through multinomial logistic regression.

3.1.4.2 Preparation of a *Snipper* input file

Careful preparation of the Excel file containing the custom training set profiles and precise matching of unknown profiles to training set data for bases and locus order is important. Therefore it is recommended to sort component SNPs/Indels into ascending rs-number order as an aid to data checking.

For *Snipper* analysis using binary markers, an *.xlsx* Excel file (*.xls* can still be used for certain previous options) with sample, population and genotype information listed (Table 5) –

how that information is organized is also important so the following considerations should be taken into account:

- cell 1A indicates the number of samples; cell 1B the number of markers; and cell 1C the number of populations;
- line 1 (from column D onwards) specifies the marker name (represented by a alpha-numerical code);
- lines 2-5 can be left empty or can be used to include useful notes (for example, one of the lines can be used to store the study/casework sample profile and other line can be used to concatenate that profile – ready for copying and pasting i.e., if the profile is in line 3 type *=D3&E3&F3…* in the desired cell);
- column A (from line 6 onwards) has a numeric value that usually represents a sample;
- column B (from line 6 onwards) has the population names
- column C (from line 6 onwards) has the sample names (which can be represented by an alpha-numerical code);
- column D onwards (from line 6 onwards) include the genotypes (coded as nucleotide bases – ACGT). Missing data should be coded as NN. Other symbols in the file (e.g. ?, spaces) are not recognized. Tri-allelic markers can be included in the analysis;
- a new batch analysis option was recently implemented in *Snipper v2.0* which allows for simultaneous classification of more than one profile. In this case, the input file should be constructed as described in the previous points. An extra column after the last marker (with no headers – start in line 6) needs to be included: training samples are to be marked as 1 and study samples to be classified as 0.

Table 5: *Snipper* input file format. Samples *S1*…*Sn* from populations 1…*Pn* analysed with genotypic data from markers *M1*…*Mn*. Samples 1…10 belong to population 1; samples 11…20 belong to population 2 and samples 21…*Sn* belong to population *Pn*. An extra column after the last marker (in this case column XFD) should be included when trying to classify several study samples simultaneously – samples from populations 1 and 2 are reference (labelled as 1) and samples from population *Pn* are the unknown study (labelled with 0). Lines 2-5 can be used to include useful information – e.g. when a single profile is being classified it can be included (here in line 3) and concatenated (cell D4) – the concatenated profile can then be copy-pasted directly into *Snipper*.

	A	B	C	D	E	…	XFC	XFD
1	# Samples	# Markers	# Populations	M1	M2	…	M_n	
2								
3			Profile	AG	TT	…	AC	
4			Concatenate	=D3&E3&…&XFC3				
5								
6	1	P1	S1	AG	CT	…	AA	1
7	2	P1	S2	GG	CC	…	CG	1
8	3	· ·	· · ·	⋮	⋮	…	⋮	· ·
9	4	P1	S10	AA	TT	…	AG	1
10	5	P2	S11	AG	CT	…	GG	1
11	6	P2	S12	GG	CC	…	AC	1
12	7	⋮	⋮	⋮	⋮	…	⋮	⋮
13	8	P2	S20	AA	TT	…	CC	1
14	9	P_n	S21	AG	CT	…	AA	0
15	10	P_n	S22	GG	CC	…	GG	0
· ·	· ·	⋮	⋮	⋮	⋮	…	⋮	⋮
1048576	n	P_n	S_n	AA	TT	…	**NN**	0

3.1.4.3 How to run *Snipper*

Snipper includes several options to classify individuals and analyse populations. For forensic analysis the two most applicable options are: "*Classification as Europe-East Asia-Africa-America-Oceania (34 SNPs, 46 Indels, or both sets)*" and "*Classification with a custom Excel file of populations*". There is an additional option that works in the same way but allows batch analysis: "*Classification of multiple profiles with a custom Excel file of populations*".

1) The "*Classification as Europe-East Asia-Africa-America-Oceania (34 SNPs, 46 Indels, or both sets)*" option uses fixed training sets and provides a simple system to classify single profiles.
   - Step 1 – go to http://mathgene.usc.es/snipper/popchoosing5groups.html;
   - Step 2 – choose the marker set from three options: *34-plex* SNPs (the original marker set (Phillips et al. 2007b) or the revised set (Fondevila et al. 2013) can be selected), *46-plex* AIM-Indels (Pereira et al. 2012b) or a combination of 80 binary

markers (Indels combined with the revised *34-plex* set). SNPs are listed in rs-number order and AIM-Indels in electrophoretic order – on the left side links give images listing the marker order in each option.

- Step 3 – choose populations. Three to five main population groups are available (Africa, Europe, East Asia plus America plus Oceania).
- Step 4 – choose the classifier. Four options are now available: naïve Bayesian analysis (considering whether the Hardy-Weinberg principle applies or not), multinomial logistic regression and genetic distance algorithm.
- Step 5 – data input. Depending on the option selected in Step 1, a profile including 34, 46 or 80 markers (68, 92 or 160 bases respectively) should be typed. As described before, a profile can be built by concatenating data in Excel (using the "&" operand) allowing individual scrutiny of composite genotypes before direct copy-pasting into the query window left of the "Classify" button.

2) "*Classification with a custom Excel file of populations*" – this option allows extension of ancestry analyses beyond the default five-population group comparisons and 34, 46 or 80 binary markers using *Snipper*.
   - Step 1 – go to http://mathgene.usc.es/snipper/analysispopfile_new.html;
   - Step 2 – data input (population). An Excel file prepared as described above (Table 5 without the final column) is uploaded.
   - Step 3 – choose classifier. Options as described above.
   - Step 4 – data input (individual). A profile string containing the same number of markers in the same order as they appear in the data file uploaded in Step 1 is entered in the query window.

3) "Classification of multiple profiles with a custom Excel file of populations" – go to http://mathgene.usc.es/snipper/analysismultipleprofiles.html. This option works as above but without the need for individual profile submission. Profiles to be classified are indicated as previously described (Table 5). The multinomial logistic regression classifier function is not currently available for this option.

Snipper also includes an option to analyse training sets to gauge characteristics of the component binary markers – "*Thorough analysis of population data of a custom Excel file*" (http://mathgene.usc.es/snipper/analysispopfile2_new.html). This is useful to assess the informativeness of new candidate AIM binary markers for ancestry inference. After uploading the Excel file of custom data and defining Hardy-Weinberg, choose "*Perform a verbose cross-validation analysis of my population data with the best _ SNPs*" adding the relevant number of markers to assess. Cross-validation removes each component sample in turn, recalculates the allele frequencies in the training set, and then assigns ancestry for the removed profile. The other options: "*Try to classify all individuals in the sample*", "*Perform a non verbose cross-validation analysis of my population data*" and "*Compute bootstrap error of my population data*" provides choice of alternative assignment error estimators. Multinomial logistic regression can also be applied to the population data – in this case information given in Step 2 about HWE will be ignored. Once a training set has been assessed for informativeness, users can choose options 2 and 3 described above to compare single or multiple profiles from unknown samples to the custom reference data and assign ancestry in identical fashion to using the fixed training sets.

3.1.5 Principle component analysis (PCA)

3.1.5.1 Background on PCA

Principal component analysis or PCA, is a multivariate data analysis technique allowing the reduction of dimensionality, i.e. it uses fewer variables, while preserving much of information in the data. Usually two or three principal components are made, constructed as linear combinations of the original variables. Working with only two or three variables allows graphical representation of the data in a 2D plane or 3D graphic, providing fast visual recognition of patterns or clusters. Numerous software packages are available to perform PCA analysis when numerical variables are used. When SNP data is considered, an initial transformation (or recodification) is needed to access this existing software. The next section details SNP data preparation using the statistical software *R*.

3.1.5.2 Preparation of PCA input files

SNP analysis with PCA requires two text files with sample, population and genotype information. One of the files should include training set data and the other the study samples to be compared. Both files have the same format (Table 6) – the system for organizing this data is important so the following considerations should be carefully taken into account:

- the first column includes sample name information in the form of an alpha-numerical code. The column header is "Sample";
- the second column has the populations/groups names. The header is "Population";
- the following columns have genotype data, one marker per column. Each column header will have the corresponding marker name, which can be an alpha-numerical code. Genotypes are coded with nucleotide bases (ACGT) and missing data as NN. Note that markers must be in the same order in both input files;
- spaces can be included in the file except as part of the genotype data (they will be considered as a new genotype, i.e., TT ≠ T T);
- tri-allelic markers can be included in the input file but they will not be considered for the principal components calculation.

Table 6: Principal component analysis SNP input file format – samples *S1…Sn* from populations *P1…Pn* analysed with genotypic data from markers *M1…Mn*.

Sample	Population	M_1	M_2	M_n
S1	P1	AG	CT	AA
S2	P1	GG	CC	CG
⋮	⋮	⋮	⋮	⋮
S10	P1	AA	TT	AG
S11	P2	AG	CT	GG
S12	P2	GG	CC	AC
⋮	⋮	⋮	⋮	⋮
S20	P2	AA	TT	CC
S21	P_n	AG	CT	AA
S22	P_n	GG	CC	GG
⋮	⋮	⋮	⋮	⋮
S_n	P_n	AA	TT	NN

3.1.5.3 Creating a PCA plot

In this subheading we include *R* scripts that can be used to generate 2D principal components graphics (only if the number of variables (SNPs) is smaller than the number of samples). The script commands can be copied and pasted into the *R* console. With the main focus on graphics, *R* offers a range of options to manipulate data and generate plots that adjust to user needs. In the case of this script, command lines were added to allow changes in colour, shape and size of the symbols representing individuals. When generating PCA plots it is possible to change graphical parameters to user needs. In the *R* script, command lines are included so the "Population" information can be used to define colour of the symbols in plots. In this case, population names are transformed into colour names, in population alphabetical order and not input order (a complete list of colour available in *R* can be obtained with the command *colours()* (alternatively *colors()*) or with the help of the *Chart of R colours* available at http://research.stowers-institute.org/efg/R/Color/Chart/). The symbol shape can also be changed – for information on available symbols use the *pch* help page by typing *?pch* in *R* console, **Figure 26**). The cex command changes the size of the points. All text after the # symbol represents notes for the user and will not be computed.

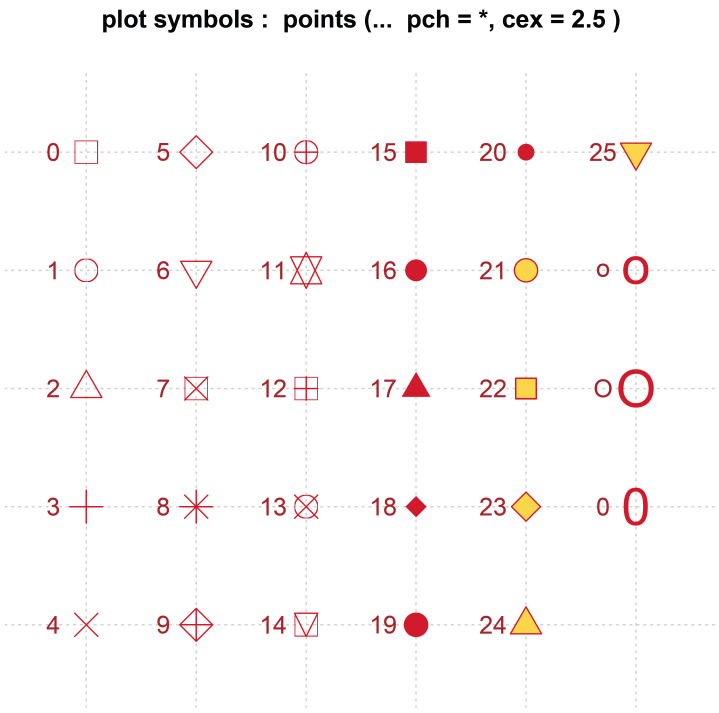


Figure 26: Symbols available in *R* to define the points shown in plots (*pch* command).

```
################################
### Script to make a 2D PCA ###
################################

# Important Note: this script can only be used when the number of samples is equal to, or higher than, the
 number of SNPs.


# First open SNPassoc library - this is an association package that allows you to recode SNP data

library(SNPassoc)

TEST<-function(x){try(snp(x,sep=""),silent=TRUE)} # homemade function that detects if a SNP is bi- or tri-
 allelic

################################
# To read the data from two input files. The computer will prompt for the location of the reference samples
 file and the study samples file, in that order
################################

RefData<-read.table(file.choose(), sep="\t",header=TRUE,na.strings=c("NN")) # function that reads the
 reference data input file and stores it in the dat object

NRefSamples<-dim(RefData)[1] # get the number of reference samples

StudyData<-read.table(file.choose(), sep="\t",header=TRUE,na.strings=c("NN")) # function that reads the file
 with the study samples and stores it in the StudyData object

NStudySamples<-dim(StudyData)[1] # get the number of study samples

### Because tri-allelics are not considered when making the PCA, they are removed from both data sets using
 the function TEST.

apply(RefData,2,TEST)->RT

deletedSNPs<-which(as.numeric(summary(RT)[,1])==1) # defines which SNPs have to be removed

RefData2<-RefData[,-deletedSNPs] # for the reference samples, only the columns with bi-allelic SNP data are
 kept

StudyData2<-StudyData[,-deletedSNPs] # for the study samples, only the columns with bi-allelic SNP data are
 kept
# this removes tri-allelic SNPs if there were any

################################
# To recode and typify the data
################################

ComData<-rbind(RefData2,StudyData2) # combines the reference and study samples in one variable (ComData) -
 this needs to be done prior recoding the SNP data to guarantee the coding uniformity

datSNPT<-apply(ComData,2,function(x) {additive(snp(x,sep=""))}) # the additive function recodes each bi-
 allelic SNP in the data as numeric (0=homozygous for the most frequent allele, 1=heterozygous, 2=
 homozygous for the least frequent allele)

datSNP<-datSNPT[1:NRefSamples,] # after recoding, the reference data set is temporarily isolated into a new
 variable (datSNP) to make some computations

m<-apply(datSNP,2,mean,na.rm=TRUE) # this calculates a vector with the mean value of each "numeric" SNP

s<-apply(datSNP,2,sum,na.rm=TRUE) # this calculates a vector with the number of occurrences of the least
 frequent allele for each SNP

n<-apply(datSNP,2,function(x) sum(!is.na(x))) # this calculates a vector with the number of valid genotypes
 for each SNP (NN genotypes are not considered valid)

p<-s/(2*n) # vector with the frequency of least frequent allele for each SNP

XT<-scale(datSNPT,center=m,scale=sqrt(p*(1-p))) # this typifies the SNPs (to each "numeric" SNP value the
 average is subtracted and then divides by the standard deviation)
```

```
XT[is.na(XT)]<-0 # replaces missing values with 0

X<-XT[1:NRefSamples,] # after recoding and typifying, the reference data set is isolated into a new variable
 (X)

Y<-XT[-c(1:NRefSamples),] # after recoding and typifying, the study data set is isolated into a new variable
 (Y)

if(NStudySamples==1){ # this forces Y to be a matrix if there is only one study sample

              dim(Y)<-c(NStudySamples,length(Y))
             colnames(Y)<-colnames(X)

    }

############################################
# Some computations before preparing the plot
############################################

princomp(X, scale = FALSE)->X.PCA # computes the PC of the reference samples

X.PCA$loadings->M # gets the rotation matrix

(Y%*%M)[,1:2]->StudyCoordinates # computes the new coordinates of the study samples


if(NStudySamples==1){StudyCoordinates<-t(StudyCoordinates)} # forces StudyCoordinates to be a matrix

per<-eigen(cov(X))$values/sum(eigen(cov(X))$values) # this calculates the percentage of explained variance
 for each principal component

per<-round(per*100,2) # values are rounded to two decimal positions

# In this part colours in the plot must be chosen (if you have more than three populations)

mycolours<-c("orange","pink","skyblue2") # choose the colours you want to use for each population
 (considering that populations are in alphabetic order). A complete list of colour names can be obtained
 with the command colours() or with the help of the Chart of R colours available at http://research.stowers-
 institute.org/efg/R/Color/Chart/

colours<-as.character(factor(RefData$Population,labels=mycolours)) # population names are converted to the
 corresponding colour name


######################
# The plot starts here
######################

quartz() # opens a new graphic display window if you use MacOS
# windows() # is the alternative command for windows OS that opens a new graphic display window if you use a
 Windows PC (remove # here and replace # in front of previous line)

plot(X.PCA$scores[,1:2],col=colours,pch=20,main="put here your plot title",xlab=paste("PC1
 ",per[1],"%",sep=""),ylab=paste("PC2  ",per[2],"%",sep=""),cex=1.5)
# this plots the two first principal components. The plot title, pch and cex values can be changed

legend("topleft",legend=levels(factor(RefData$Population)),col=mycolours,pch=20,cex=0.5,y.intersp=1)
# this adds a legend to the plot. Its position can be changed using "topleft", "topright", "bottomleft" or
 "bottomright". Pch should match the one used in the plot.

### With the next set of commands it is possible to include the study individuals superimposed onto the
 previously plotted principal components graphic.

points(StudyCoordinates,col="black",pch=20,cex=2)
# this estimates the coordinates for the study individuals and plots them onto the previously generated
 principal component graphic. colour, pch and cex can be adjusted.

text(StudyCoordinates,as.character(StudyData[,1]),cex=1,pos=1)
# this adds study individual's descriptors to the plot
```

3.1.6 Multinomial logistic regression (MLR) to analyse admixed individuals

3.1.6.1 Background on MLR

Assignment of biogeographical ancestry (BGA) is commonly performed using naïve Bayesian classifiers (e.g. *STRUCTURE* (Pritchard et al. 2000, Hubisz et al. 2009), *Snipper* (Phillips et al. 2007b)) or principal components analysis (PCA). Bayesian classifiers assume that genotypes at all loci are independent of each other (or linked according to a model) and that they are in Hardy-Weinberg equilibrium (HWE) and linkage equilibrium (LE) in assigned "ancestor" populations. This is unlikely to be true for real populations (BGAs) and even less so for other phenotypes like externally visible characteristics (EVCs). While offering a useful method for visualisation of BGA clusters, PCA is less useful for categorical assignment.

Multinomial logistic regression (MLR) is used to predict categorical (non-metric) outcomes from categorical and metric data. It has applications in marketing and polling as well as for the prediction of EVCs like eye and hair colour from single nucleotide (SNP) genotypes at multiple loci (Walsh et al. 2011b, Walsh et al. 2013). A predictor function, f, for a phenotype, p_i, is expressed as:

$$f(p_i) = \beta_{0,i} + \beta_{1,i}g_1 + \beta_{2,i}g_2 + \cdots + \beta_{n,i}g_n$$

where g_j is the genotype at the jth locus and $\beta_{i,j}$ is a coefficient for the ith genotype at the ith locus. For mutually exclusive phenotypes, the probability of the ith phenotype, $\pi(p_i)$, given a genotype, g, is given by:

$$\pi(p_i) = \frac{e^{f(p_i)}}{1 + e^{f(p_1)} + e^{f(p_2)} + \cdots + e^{f(p_n)}}$$

The β coefficients are regressed to maximise probabilities for known phenotypes with known genotypes in a training set. The regressed coefficients are then used to calculate probabilities, $\pi(p_i)$, for unknown phenotypes with known genotypes.

3.1.6.2 Preliminary comparison of MLR and Bayesian methods

A training set of SNP genotypes from 571 individuals was derived from the Human Genome Diversity Project (HGDP) CEPH cell line panel (Cann et al. 2002) and consisted of five BGAs: 98 from Sub- Saharan Africa, 157 from Europe, 225 from East Asia, 27 from Oceania and 64 Indigenous Americans. A subset of 34 autosomal genotypes at each of the loci included in the SNP*for*ID *34-plex* ancestry assay (Fondevila et al. 2013) was used as reference genotypes for the five BGAs. SPSS1 Statistics (IBM, Version 21.0.0.0) was used to regress MLR β coefficients for the training set using each of the BGAs as a pivot reference in turn and Excel 2010 (Microsoft) was used to create predictor functions for BGA assignment. *STRUCTURE* was also used to apportion BGAs to each individual using K=5 with a 100,000 burnin period followed by 100,000 MCMC steps after burnin.

A test set of SNP genotypes from 1092 individuals was derived from the 1000 Genomes Project (1000GPC 2010) (Table 7). These were classified by *STRUCTURE* and MLR using the parameters determined from the HGDP CEPH training set.

Table 7: Training set with 1000 Genomes ID and numbers. Populations in bold are expected to be admixed.

African			European			Asian			American		
US African	**ASW**	**61**	US European	CEU	87	Beijing	CHB	97	**Columbia**	**CLM**	**60**
Kenya	LWK	97	Finland	FIN	93	South China	CHS	100	US Mexican	MXL	66
Nigeria	YRI	88	Britain	GBR	88	Japan	JPT	89	**Puerto Rico**	**PUR**	**55**
			Spain	IBS	14						
			Italy	TSI	98						

3.2 RESULTS AND DISCUSSION

3.2.1 What information can be obtained from *STRUCTURE*?

STRUCTURE output files include information on the estimated clusters, i.e. the population groups generated not the input populations. However, when the populations are defined in a way that they closely match the calculated clusters, the inferences of the population ancestry membership proportions in each of the pre-defined clusters can be considered to be the proportions of the input populations. When attempting to classify a population or individual, the use of reference populations closely matching the inferred clusters is important, especially when analysing admixed samples where it is important to

define the contributing parental populations. The ancestry membership proportions for each individual in each cluster are also calculated by *STRUCTURE*.

Allele frequency divergence among populations, average distances (expected heterozygosity) between individuals in the same cluster, mean F_{ST} values and estimated allele frequencies in each cluster (including estimated ancestral frequencies) are calculated. As a way of quantifying the information given by a particular *STRUCTURE* run and estimating the optimum *K* value it calculates the estimated probability of the data, the mean likelihood value and associated variance. And it calculates the mean value of alpha (α) as a measure of the relative admixture levels between populations – when $\alpha >> 1$ the individuals are highly admixed; for values of $<<1$ each individual has its origin mainly in one population (from our experience with the HGDP-CEPH panel of samples, $\alpha < 0.05$ – this value varies depending on the population groups considered and the differentiation power of the marker sets used).

The population and individual ancestry membership proportions can be represented in two distinct types of plot:

- a bar plot where each individual of the data set is represented by a vertical line divided into *K* coloured segments proportional to the estimated membership into each of the *K* inferred clusters;
- each individual is represented as a coloured point in a triangle. Colours correspond to the population tag in the input file. The estimated ancestry vector for an individual is formed by *K* components that sum up to 1. This type of plot is particularly useful to represent $K=3$ data because the vectors can be represented in one triangular plot. For each point, the distance to the triangle vertices gives each of the three components. Individuals located in one of the vertices are completely assigned to the population represented in it.

Despite the advantages of the triangular plot when visualizing $K=3$ data, bar plots are usually easier to interpret, especially for $K>3$.

In the case of forensic casework analysis, *STRUCTURE* gives information on the training set (allowing the assessment of the used reference data set – the optimum *K* value matches the number of reference populations, which are completely differentiated among them) and it also gives us the individual ancestry membership proportions (such information has considerable potential in guiding investigators to more clearly defined suspect pools, this being particularly true when no eyewitness is available or STR profiles fail to match DNA database records).

3.2.2 Evaluating *Snipper* output

Results from the analysis of a profile comprise: the submitted profile; the assumed classifier; the -log likelihoods (use of -log likelihoods permits easier comparison of the very small likelihood ratio figures normally generated) and percentiles for the training set population groups; the likelihood ratios in verbose format and predicted admixture components and ancestry; a set of plots summarizing the classification; the apparent success of the classification; and a list of the markers in descending order of divergence (calculated based on the number of populations included in the comparison). Missing genotypes are flagged in red in the divergence list to allow some assessment of the potential contribution of gaps in the profile, in other words, assignments made with several red markers at the top of the list will be much less reliable than those with gaps at the bottom, although this will be clear from the probabilities obtained. Apparent success measures the rate of correct assignment of training set samples using the markers of the profile. These values are 100% for a complete set of markers, but drop when significant numbers of gaps occur in the submitted profile (in the case of the three group *34-plex* fixed training set this is particularly true for EUR:E ASN comparisons).

3.2.3 What information can be obtained from PCA?

PCA allows the exploration of data sets and shows proximity between individuals. In fact, it is possible to include a casework sample in the PCA plot generated for the reference populations helping to infer, through visual inspection, the most probable classification of that individual (**Figure 27**).

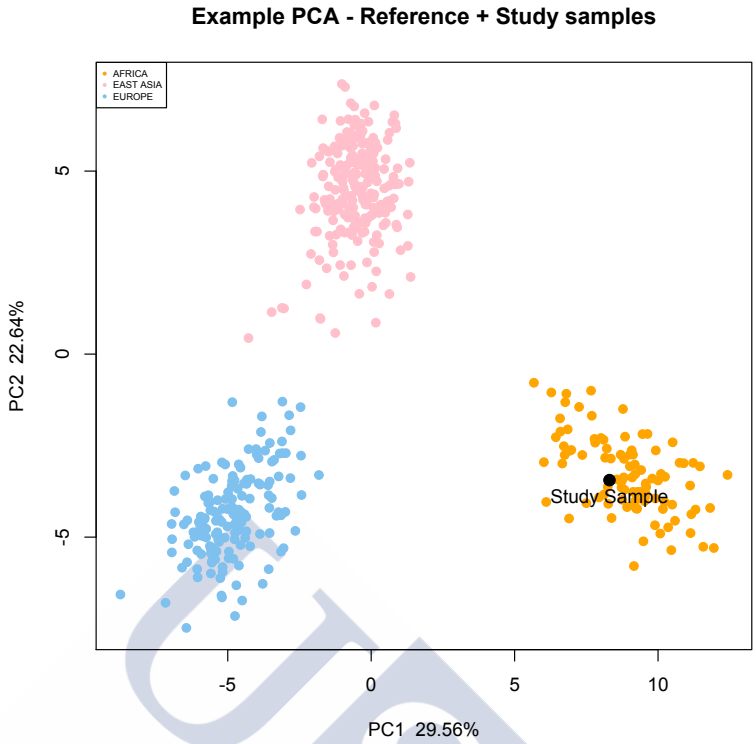


Figure 27: PCA plot generated using the *R* script described in section 3.1.5.2. Three population groups from the HGDP-CEPH panel of samples are used as reference data: Africa (orange), Europe (blue) and East Asia (pink). One study sample was plotted in the reference PCA (black). It is possible to infer that the study sample is likely to be African.

3.2.4 Preliminary comparison of MLR and Bayesian methods

STRUCTURE classification of the HGDP-CEPH training set indicated that individuals within each of the five BGAs (Sub-Saharan African, European, East Asian, Oceania and Indigenous American) had their percentage BGA inferred correctly with mean probabilities ranging from 91% (East Asian) to 96% (Oceania). MLR classification confirmed that there was little admixture between BGAs in the training set.

Individuals from the 1000 Genomes test set were classified by both *STRUCTURE* and MLR (**Figure 28**). They both predict admixture in African Americans (ASW), US Mexicans (MXL), Columbians (CLM) and Puerto Ricans (PUR) and relatively little admixture in other populations. The African Americans have a significant contribution from a European ancestral BGA while the US Mexicans, Columbians and Puerto Ricans are classified as mostly European with a significant contribution from an Indigenous American ancestral BGA. This predicted admixture is consistent with American colonial history.

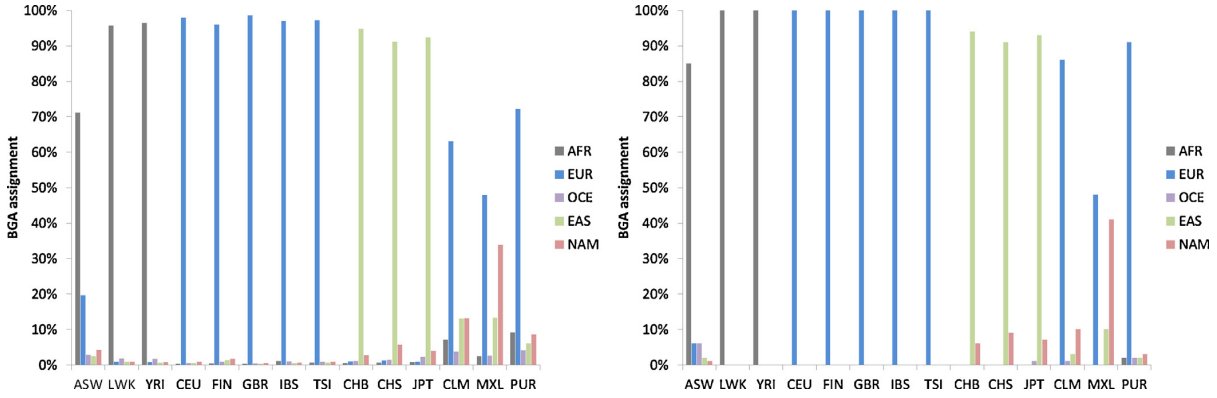


Figure 28: *STRUCTURE* classification (left) and MLR classification (right) of the 1000 Genomes test set.

These preliminary results seem to indicate that MLR offers a viable alternative to *STRUCTURE* as a classifier for BGA. Whereas *STRUCTURE* requires formatted input files and often involves long run times, MLR can be performed in a spreadsheet environment. It also makes no assumptions about genetic equilibrium in ancestral BGAs.

BLOCK 2

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

4 Block 2 – New AIM assays

This second block focus on the development of five forensically relevant ancestry informative assays:

- the revision of the previously published SNP*for*ID *34-plex* forensic ancestry test (Phillips et al. 2007b) to improve marker performance, population differentiation power and to extend population data (Appendix 2A);
- *Pacifiplex*, a set of 29 AIM-SNPs centred on Australia and the Pacific region;
- a set of *46 AIM-Indels* designed to measure population admixture proportions of four major population groups (Appendix 2B) and the completion of a reference database (Appendix 2C);
- *Eurasiaplex*, a set of 23 AIM-SNPs dedicated to the differentiation of European and South Asian populations (Appendix 2D);
- a set of 12 AIM-STRs that provide a good alternative to the detection and analysis of mixed DNA samples (Appendix 2E).

Each assay will be described in the following sections including material and method as well as results and discussion.

4.1 MATERIAL AND METHODS

4.1.1 Revision of the SNP*for*ID *34-plex* forensic ancestry test (Appendix 2A)

4.1.1.1 Population samples

A total of 709 unrelated male samples of self-declared ancestry from the population reference collection of NIST (National Institute of Standards and Technology, Gaithersburg, MD, USA) were genotyped with the enhanced *34-plex* assay reaction conditions described below. This panel is made up of representative samples from the four main U.S. population groups, comprising: 261 Caucasians, 258 African Americans, 140 Hispanics and 50 East Asians. Additionally, the six components of the recently revised NIST PCR-based DNA profiling Standard Reference Material® (SRM) 2391c (Børsting et al. 2011) were typed along with the current standard forensic positive control DNAs: Applied Biosystems (AB)/Promega 9947A; Qiagen XY5, and; Promega 2800M, thereby complementing the reference genotypes for the SNP*for*ID *52-plex* ID-SNP assay published in 2011 (Børsting et al. 2011). NIST SRM 2391c components A, B and C are described as Caucasian, Mexican Hispanic and Melanesian in origin respectively (component D is a 3:1 mixture of A and C), while E and F do not have ancestry descriptions.

Previously generated SNP genotypes from the HGDP-CEPH global diversity panel (Phillips et al. 2007b) were used as reference data for analyses of the NIST populations (CEPH genotype data available from the *SPSmart* SNP*for*ID browser at: http://spsmart.cesga.es/snpforid.php?dataSet=snpforid34), while 1000 Genomes data was collated directly from the ENGINES whole genome SNP browser (Amigo et al. 2011) available from: http://spsmart.cesga.es/engines.php?dataSet=engines. The HGDP-CEPH panel comprises: 105 Africans, 158 Europeans, 232 East Asians, 64 Native Americans and 28 Oceanians (South Asian and Middle East populations were excluded). The 1000 Genomes samples comprise: 246 Africans (including 61 African Americans with mixed ancestry), 380 Europeans, 286 East Asians and 181 Americans with mixed ancestry (Puerto Rican and Mexicans from Central America plus Colombians from South America). Using both panels enabled the most comprehensive geographic survey of SNP variability and ENGINES allows data to be collected for any SNP locus with minor allele frequencies above 1%.

4.1.1.2 *34-plex* component SNP reconfiguration

A single marker substitution was made to replace the consistently under-performing SNP rs727811 (internal code A11) with an East Asian-informative SNP rs3827760 (assigned internal code P28). Representative allele frequency distributions from 1000 Genomes and HapMap (GIH) samples comparing both SNPs are summarised in **Figure 29**. Note that the SNaPshot extension primer of rs727811 interrogated the AC strand but HapMap and 1000 Genomes databases report the GT strand, while for rs3827760 SNaPshot interrogates the AG strand, as listed in the above SNP databases. The revised marker details for the complete assay are listed in Table 8. Components listed with internal codes prefixed with an 'A' are markers overlapping with the SNP*for*ID *52-plex* ID-SNP multiplex (Sánchez et al. 2006), since they show informative allele frequency differences between populations but also serve as housekeeping markers when using both multiplexes together in an analysis. Lastly, it was necessary to design a redundancy in the extension primer of rs3827760, comprising C and T alternate bases corresponding to neighbour SNPs rs144939741 and rs121908454, 10 bp and 15 bp upstream of the target site.

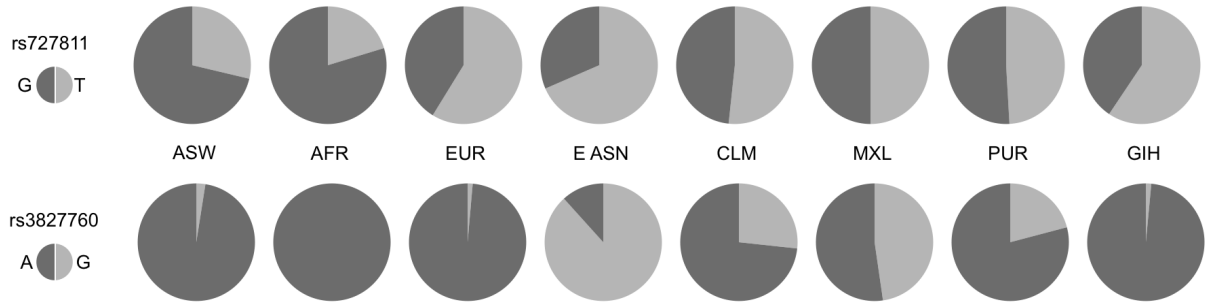


Figure 29: Allele frequency distributions in Africans, Europeans and East Asians from 1000 Genomes and HapMap populations for the replaced rs727811 SNP and new rs3827760. Africans (AFR) comprised YRI: Yoruba in Ibadan, Nigeria and LWK: Luhya in Webuye, Kenya; Europeans (EUR) are CEU: Utah residents with N & W European ancestry from the HGDP-CEPH collection, FIN: Finnish in Finland, GBR: British in England and Scotland, IBS: Iberian populations in Spain and Tuscans in Italy; East Asians (E ASN) are CHB: Han Chinese in Beijing, CHS: China, Han Chinese South and JPT: Japanese in Tokyo, Japan. Certain populations are shown separately due to levels of admixture: ASW, African ancestry in Southwest USA; CLM, Colombian in Medellín, Colombia; MXL, Mexican ancestry in Los Angeles; PUR, Puerto Rican in Puerto Rico; GIH, Gujarati Indians from Houston.

Table 8: *34-plex* composite SNP data.

Markers		PCR size	Strand information			Primers			Primer mix ratios	
Internal code	dbSNP rs-number		1000 G bases	Assay bases	SNaPshot direction	PCR forward primer	PCR reverse primer	SNaPshot single base extension (SBE) primer	PCR (F+R primer stock at 25 μM)	SNaPshot SBE (primer stock at 50 μM)
P02	rs5997008	90	AC	AC	F	GTCAACACTAGAGTATTTGCCCATC	ACAAACCCAAAGACTGTTCTGC	gac[aactaggtgccacgtcgtgaaagtctgac]$_2$ aactctcaCAGGATCGATTGGTTCC	1.9	1.3
P01	rs2304925	98	AC	GT	R	CCCATTAACTCATCAAAGTGGTGAT	CCCCACTCCACCGCTAAT	[aaagtctgacaactaggtgccacgtcgtg]$_2$ aaagtctgacaaCCACTCCACCGCTAAT	2.5	2.5
A07	rs917118	87	CT	AG	R	GCCCTTTAGGGTCGGTTC	GTAAGAGATGACTGAGGTCAACGAG	t[ct]$_2$TGACTGAGGTCAACGAGC	3	2.1
P03	rs1321333	80	AG	CT	R	GTCAGTAAGACGGTAACTCC	CTAACACAAGCCTAAATCCAG	AAGACGGTAACTCCATGGCTG	2.75	1.5
P04	rs2814778	102	CT	CT	F	AACCTGATGGCCCTCATTAGT	ATGGCACCGTTTGGTTCAG	agtctgacaactaggtgccacgtcgtgaaagtc tgacaactaggtgccacgtcgtgaaagtctgacat CTCATTAGTCCTTGGCTCTTA	1.5	2
A29	rs1024116	76	CT	AG	R	CCATGTGTTCTAATAAAAAGGATTGC	TGGGAAGTGAGCAAAAGTAAATACA	cttGTTCTAATAAAAAGGATTGCTCAT	1	2
P05	rs7897550	100	AG	CT	R	CGATGTGTCTTACGGAATACTAGGT	AGAGCTGACAGGCAAAAATGCTAT	t[ct]$_2$TGTGCAGGATTGAAATATAATT	2	1
A21	rs722098	80	AG	AG	F	GGAAGTACACATCTGTTGACAGTAATGA	GGGTAAAGAAATATTCAGCACATCC	agtctgacaaTGACAGTAATGAAATATCCTTG	2.3	4
P06a	rs10843344	87	CT	CT	F	TGTACAATGGTAGATGTGTGCTCAG	GATAGCTCTGGTGTTGCATTATTGT	t[ct]$_5$AGTACTTTGCCAAAGAAACTAAA	4	1.3
P08	rs12913832	99	AG	AG	F	ACGTTGGATGCGAGGCCAGTTTCATTTGAG	ACGTTGGATGAAAACAAAGAGA AGCCTCGG	[ct]$_8$CCAGTTTCATTTGAGCATTAA	3	3
P07	rs239031	70	AG	CT	R	TAGCTGTGAGATAGAAATCCTGGAC	ACTACCCTAATCTCAGCTTCCACTC	t[ct]$_8$cAATCTCAGCTTCCACTC	0.5	1
P09a	rs1978806	94	AG	CT	R	AGAGTTTGACATGATGGTGCTCTA	TCTTGTTTCTAAGCAGGAAAGTTG	t[ct]$_8$cGCAGGAAAGTTGTATTCTGATA	1.8	1
A40	rs2040411	61	AG	AG	F	TCTGGAATGCCAGTTCTTTTGT	CAGAACGCCTATGAAAACCAGT	[ct]$_7$cCTCTGTATTTTCTTACTCTAAGTGC	1	2
P10	rs773658	95	GC	CG	R	ACAAACGGAAAGTAGTATTGGACTG	AGAAGGGGCACAGCAATTTAGTA	[ct]$_{11}$cGGGAAGAATAGAGTCAATCAA	3.2	3
P11	rs10141763	82	TA	AT	R	ACAGACTTGGTTCCCTGAAGTCTA	GTAGATTGTAGGCAAGTCGTAAAGG	t[ct]$_{10}$cGTGTGAGTTGTGTGATAATCTA	3	1.1
P12	rs182549	117	CT	CT	F	AAGTACTGGGACAAAGGTGTGAG	AGAAGTCAGAATACCCCTACCCTAT	[ct]$_{16}$cAGGTGTGAGCCACCG	6.2	6.5
P13	rs1573020	81	AG	AG	R	CTATCTGCCACCTGAGAGAGTATTG	AGGTGTCAGCTTCTTCTGACCAT	[ct]$_{14}$GAGTATTGCCAGCCTGATTC	1	1.3
P14	rs896788	78	CT	CT	R	GTAATGCCTCTGTGGCCCTAT	ATTCCGTCCACATCTTCACTG	[ct]$_{17}$tctACAGTCACCAGCCAC	1.5	1.3
P15	rs2065160	66	AG	AG	F	AAGAATGGCCTCTCGATGAGTA	GATGATACCTACGCATAGTCTGTTT ACTTC	t[ct]$_{14}$GCATAGTCTGTTTACTTCATTTG	1.5	1.3
P16a	rs2572307	63	AG	AG	F	GTGTAGCTATGCCATCATTCAATC	ATCCTTAGAAGGGTGCTAAACTGAG	t[ct]$_{15}$catcATTCAATCAATAGTCATAAAC	2.25	1.1
P17	rs2303798	96	AG	CT	R	CCAGCCTGCACCACTGTC	AGAGATGTGTTCAGGAAGAGGCTA	t[ct]$_{19}$cAGAAGAGGGACGTGGG	4.3	3.5
P18	rs2065982	88	CT	AG	R	CTTGGGGCAGTCTTTAAGTCCT	AGGAAGTGGTCAGTGCCAGTAG	[ct]$_{18}$GGAAAAAAAAGTCCTCTTTGGTAT	2	1.3
P19	rs3785181	156	CT	CT	F	CTCTGTTCAGTTTCAAAGTTCTGG	TTGTGTTCAAAAATTTCAATTAGGTT	t[ct]$_{20}$AGGGCATCTTATCTTGAGC	1.25	0.75
P20	rs881929	93	GT	GT	F	AGCTACCTGGTGTCTAACTC	TTGACCCAGTGGTTCTGAGC	[ct]$_7$aaactaggtgccacgtcgtgaaagtctgacaa GCCTCTTGCCAGCTCTG	3	2.5
P21	rs1498444	106	GT	AC	R	GGCTATTACCACATTAAGAGAAACTGC	CAGCCTCTCAATGCAAATGAT	t[ct]$_{20}$cGGTTGTCAGAATATTTGCTACA	5	3
P22a	rs1426654	74	AG	CT	R	AATTCAGGAGCTGAACTGCC	TGTTCAGCCCTTGGATTGTC	[ct]$_{24}$ttCGCTGCCATGAAAGTTG	3	4
P23	rs2026721	108	CT	AG	R	GAAGACTTTTTGCAAGCACGAG	GGCAAATGCTGTAAGAATCCAT	t[ct]$_{21}$AAATGATTGACATAGTAGGCTATTG	3.2	1.75
P24	rs4540055	76	AGT[a]	ACT	R	TGTGCCTCTGATCACTTTTGAATAC	CCTAGCCAACTCCAGAGTTCAT	[ct]$_{27}$cGAAGCAGTGATCAGCAC	1.2	1.1
A52	rs1335873	110	TA	AT	R	GTGGATGATATGGTTTCTCAAGG	TTCAACAAACGTGTGATGCTCT	t[ct]$_{25}$AGGTACCTAGCTATGTACTCAGTAT	1.2	1.75
P25a	rs16891982	78	CG	GC	R	GAATAAAGTGAGGAAAACACGGAGT	GTTTCTCATCTACGAAAGAGGAGTC	[ct]$_{28}$GGTTGGATGTTGGGGCTT	1.75	2.25
P26	rs730570	92	AG	CT	R	CAGCACCCTGTAAAGTCCAG	CAGCACTCACCTGCATCTCA	t[ct]$_{27}$tCATTAATCACACAAATTTTGCAT	1.3	2.2
A13	rs1886510	86	AG	AG	F	GTCCTTGTCAATCTTTCTACCAGAG	GGATTTTCACAACAACACTTGC	[ct]$_{27}$ACAAGATTTTCACAACAACACTTGC	1.5	1
P27	rs5030240	81	CGT[a]	CGA	R	CCAAAGTGCCAGGATCACAG	TCCCTAGAAATCCTTCAGCC	[ct]$_{32}$cCACAGGAGTGAGCCACTGC	2	2
P28	rs3827760	85	AG	AG	F	GCTCAGCTCCACGTACAAC	CTGTCATGCCCCCAATCTC	tctgacaactaggtgccacgtggtgaaagtctgacaa ctaggtgccacgtcgtgaaagtctgacaactctca GGYGCCAYGTTTTCACA[b]	3	2.75
A11	rs727811	78						H_2O	104.65	5.62
								[200 μM] $(NH_3)_2SO_4$	(SBE mix only)	33.33
								Total	179.45	108.8

[a] dbSNP bases, 1000 Genomes list AC bases only for both of these tri-allelic SNPs.

[b] Y bases denote equimolar primers used with C/T at two positions (rs144939741 and rs121908454).

4.1.1.3 Amplification and single base extension primer modifications

The PCR and single base extension (SBE) primers used in the modified assay are listed in Table 8. As well as novel primers for rs3827760, the SBE primers for component SNPs: rs2304925; rs5997008; rs2814778; rs239031; rs16891982 (internal codes P01, P02, P04, P07 and P25a respectively) have been modified to adjust peak positions for electrophoresis with AB POP-4™ capillary electrophoresis (CE) polymer. Some mobility modifying poly-CT tails were also changed to non-homologous sequences comprising all bases. POP-4™ is now in increasing use and these SBE primer rearrangements anticipate the discontinuation of POP-6™ CE polymer used in the original *34-plex* assay development. In brief, SNPs rs2304925/rs5997008 were moved away from the fastest mobility size range where dye artefacts sometimes interfered with their peak recognition and SNPs rs2814778/ rs239031/rs16891982 were size-adjusted to separate them from the co-incidental mobilities of neighbouring SNPs or closely sited non-specific peaks when using POP-4™. The rearranged SNP positions are listed in Table 9 based on average size estimates of ~750 samples with multiple runs.

Table 9: SNP peak positions for electrophoresis with POP-4™ polymer. SD: standard deviation.

dbSNP rs-number	Internal code	Amplified bases	Theoretical size	G	A	C	T	SD allele 1 size	SD allele 2 size	SD allele 3 size
rs1321333	P03	CT	22			26.89	28.63	0.05	0.06	
rs917118	A07	AG	24	28.17	30.16			0.05	0.04	
rs1024116	A29	AG	28	26.75	28.95			0.02	0.03	
rs7897550	P05	CT	28			31.44	32.87	0.03	0.08	
rs722098	A21	AG	33	33.88	36.28			0.04	0.00	
rs10843344	P06a	CT	35			37.05	38.31	0.04	0.07	
rs239031	P07	CT	36			38.90	40.45	0.08	0.05	
rs12913832	P08	AG	38	40.15	40.90			0.06	0.07	
rs2040411	A40	AG	41	43.18	44.25			0.09	0.07	
rs1978806	P09a	CT	41			44.27	45.52	0.09	0.06	
rs773658	P10	GC	45	46.25		46.73		0.01	0.07	
rs10141763	P11	TA	45		48.74		49.32	0.09	0.05	
rs182549	P12	CT	49			49.11	50.55	0.04	0.04	
rs1573020	P13	AG	49	50.27	50.98			0.06	0.05	
rs896788	P14	CT	53			52.52	53.26	0.06	0.07	
rs2065160	P15	AG	53	54.11	54.82			0.07	0.04	
rs2572307	P16a	AG	57	55.41	56.02			0.06	0.01	
rs2303798	P17	CT	57			57.14	57.73	0.03	0.02	
rs2065982	P18	AG	61	59.17	59.90			0.04	0.03	
rs3785181	P19	CT	61			60.27	61.49	0.05	0.06	
rs881929	P20	GT	64	61.83			63.86	0.05	0.07	
rs1498444	P21	AC	65		64.17	64.01		0.06	0.10	
rs1426654	P22a	CT	68			67.35	68.19	0.04	0.06	
rs2026721	P23	AG	69	68.45	69.37			0.14	0.08	
rs4540055	P24	TCA	73		72.15	71.68	72.48	0.08	0.10	0.06
rs16891982	P25a	GC	75	76.03		76.44		0.13	0.10	
rs1335873	A52	TA	77		78.16		78.31	0.09	0.13	
rs1886510	A13	AG	80	78.71	79.69			0.10	0.07	
rs730570	P26	CT	80			80.11	80.74	0.08	0.11	
rs5030240	P27	GCA	85	83.75	84.91	84.51		0.03	0.04	0.05
rs2304925	P01	GT	87	86.61			87.93	0.09	0.11	
rs5997008	P02	AC	87		88.20	87.86		0.05	0.03	
rs3827760	P28	AG	90	89.64	90.09			0.04	0.04	
rs2814778	P04	CT	90			91.04	91.93	0.07	0.04	

4.1.1.4 Modified PCR and SBE reaction protocols

Both amplification and extension reactions were re-optimised and represent significant modifications to the originally described *34-plex* assay chemistry (Phillips et al. 2007b). We reduced the PCR cycling from the original 35 to 30-32 amplification cycles and from 30 to 28-30 extension cycles – each of the new cycle number ranges was modified according to the peak heights observed on identical 3130*xl* CE detectors in the two study laboratories, so ± two cycles adjusts for instrument sensitivity. Furthermore, cycling conditions were changed from the original protocol by increasing the initial 95ºC denaturation from 10 to 15 min and increasing the final 65ºC extension from 6 to 15 min. Amplification reactions used a total PCR volume of 6.9 µL, comprising: 1× AB AmpliTaq Gold PCR buffer; 25 mM $MgCl_2$ to a final concentration of 5.9 mM; 10 mM dNTP mix to a final concentration of 0.58 mM; 3.2 µg/µL BSA to a final concentration of 0.29 µg/µL, 0.5 units of AB AmpliTaq Gold polymerase, 1 µL of premixed PCR primers, and; 0.75 ng of target DNA. Individual primer pair ratios have also been changed since the original assay description and are detailed in Table 8. PCR cycling comprised: 15 min at 95ºC, 30-32 cycles of 95ºC for 30 s, 60ºC for 50s, 65ºC for 40s, then a final extension at 65ºC for 15min.

PCR products were cleaned up prior to extension using combined Exonuclease I and Shrimp alkaline phosphatase by adding 1.3 µL Exo-SAPit (USB Products, Affymetrix, Santa Clara, CA, U.S.) to a 2.5 µL aliquot of PCR product. Samples were incubated at 37ºC for 45 min then enzymes were heat-inactivated at 85ºC for 15min. SNaPshot primer extension reactions have also been modified from the original protocol to a total SBE volume of 3 µL combining 1 µL of cleaned-up PCR product, 1 µL of SNaPshot ready reaction mix (previously 1.25 µL), 0.5 µL of pre-mixed SBE primers, as outlined in Table 8 (previously 0.75 µL), and 0.5 µL of water, SBE cycling used: 28-30 cycles of 96ºC for 10 s, 55ºC for 5 s, 60ºC for 30 s. SBE products were cleaned-up by adding 1.3 µL of SAP to the whole 3 µL reaction volume and incubating at 37ºC for 80 min and heat-inactivated at 85ºC for 15 min.

For CE detection, SBE products were diluted 1 in 25 and 1 µL of diluted product was added to a mix of 8.9 µL AB HiDi™ formamide plus 0.3 µL AB LIZ-120 size standard. All separations were made with AB 3130*xl* Genetic Analyser, POP-4™ polymer and 36 cm capillary arrays. Standard SNaPshot electrophoresis conditions were used and results analysed with AB GeneMapper ID-X software.

4.1.1.5 Population analysis

Reference genotypes were compiled from the HGDP-CEPH SNP*for*ID listings in *SPSmart* for the five continental population groups of Africa, Europe, East Asia, Oceania and America. This data was used to assess the predictive value of the modified *34-plex* test with emphasis on assessing the enhancement of East Asian and closely related American population classifications that can be expected from the incorporation of rs3827760. Full HGDP-CEPH genotype data for rs3827760 has been added to the *34-plex* tables of the SNP*for*ID browser at: http://spsmart.cesga.es/snpforid.php?-dataSet=snpforid34, while all rs727811 data is retained.

Classification success was estimated using the verbose cross-validation option of the *Snipper* SNP classification portal (at: http://mathgene.usc.es/snipper/ – choosing the 'Thorough analysis of population data with a custom Excel file' method), which removes each sample in turn and classifies it using the modified training set '*n*-1' allele frequency estimates. Population structure analysis was performed using *STRUCTURE* 2.3.3, applying three iterations for each *K* value from 2 to 8, with a 100,000 burn-in period followed by 100,000 MCMC steps after burn-in. We applied the admixture and correlated allele frequencies models and used prior specification of the population of origin of HGDP-CEPH reference samples (i.e. applying POPFLAG = 1) before analysing the population structure of individuals taken from 1000 Genomes or NIST population panels. Analyses of the NIST panel samples combined HGDP-CEPH reference samples with additional SNPforID populations taken from *SPSmart*: 60 each of Mozambican; Ugandan; Danish; NW Spanish; mainland Chinese and; Taiwanese, plus 144 samples from three Colombian populations. Plots were constructed using *CLUMPP* 1.1.2 and *distruct* 1.1 software. After establishing the statistically optimum *K* values for the reference samples using the 34 SNPs (both original and new SNP combinations) we analysed 1000 Genomes individuals combined with the nine control DNAs in simultaneous runs by treating them as population samples of unknown ancestry (i.e. POPFLAG = 0) to test the efficacy of the *34-plex* assay in characterising populations with significant admixture levels. The 709 NIST population reference samples were analysed in identical fashion, pre-empting expected high levels of admixture by also setting POPFLAG = 0. An additional set of *STRUCTURE* analyses examined just the NIST U.S. population panel samples by themselves as an alternative approach to runs combining

reference samples of known un-admixed ancestry (typically HGDP-CEPH panel individuals) in order to assess how closely admixture component ratios matched those obtained with runs combining NIST and reference samples.

4.1.2 *Pacifiplex*: an AIM-SNP panel centred on Australia and the Pacific region

4.1.2.1 Population samples

A total of 507 samples from four population groups of the HGDP-CEPH standardised subset H952 were used as reference samples in this study: 98 from Africa (AFR), 157 from Europe (EUR), 225 from East Asia and 27 from Oceania (OCE). SNP genotypes were previously generated (Li et al. 2008) and collected from the *SPSmart* open access online browser (http://spsmart.cesga.es/). A total of 697 study samples were used to evaluate *Pacifiplex* performance as a test set, comprising: i) 387 individuals from nine different locations in the Pacific region genotyped in this study (21 from Taiwan (Tw); 31 from Philippines (Phl); 31 from Truk Lagoon (TL); 117 from East Timor (ET); 73 Aboriginal Australians from Northern Territory (NT); 16 Aboriginal Australians from Western Australia (WA); 35 from Papua New Guinea (PNG); 33 from Fiji (Fj); and 30 from Easter Island (EI), (sampling locations shown in **Figure 30**); ii) 310 individuals from 1000 Genomes Phase III (108 Yoruba in Ibadan (YRI); 99 Utah residents with European ancestry (CEU); and 103 Han Chinese in Beijing (CHB), selected as representative of African, European and East Asian population groups, respectively.

Sensitivity was assessed using a DNA dilution series comprising DNA extracts from one HGDP-CEPH China-Yizu individual (HGDP01187) and a Coriell cell-line control sample (HG00403). Two highly degraded bone samples (from a tibia (0.0408 ng/µl) and a femur (0.0206 ng/µl)) plus a 50-year old frozen serum sample (0.20 ng/µl) taken from the GWU sample set, were used to test the ability of *Pacifiplex* to successfully genotype degraded and/or low template DNA samples.

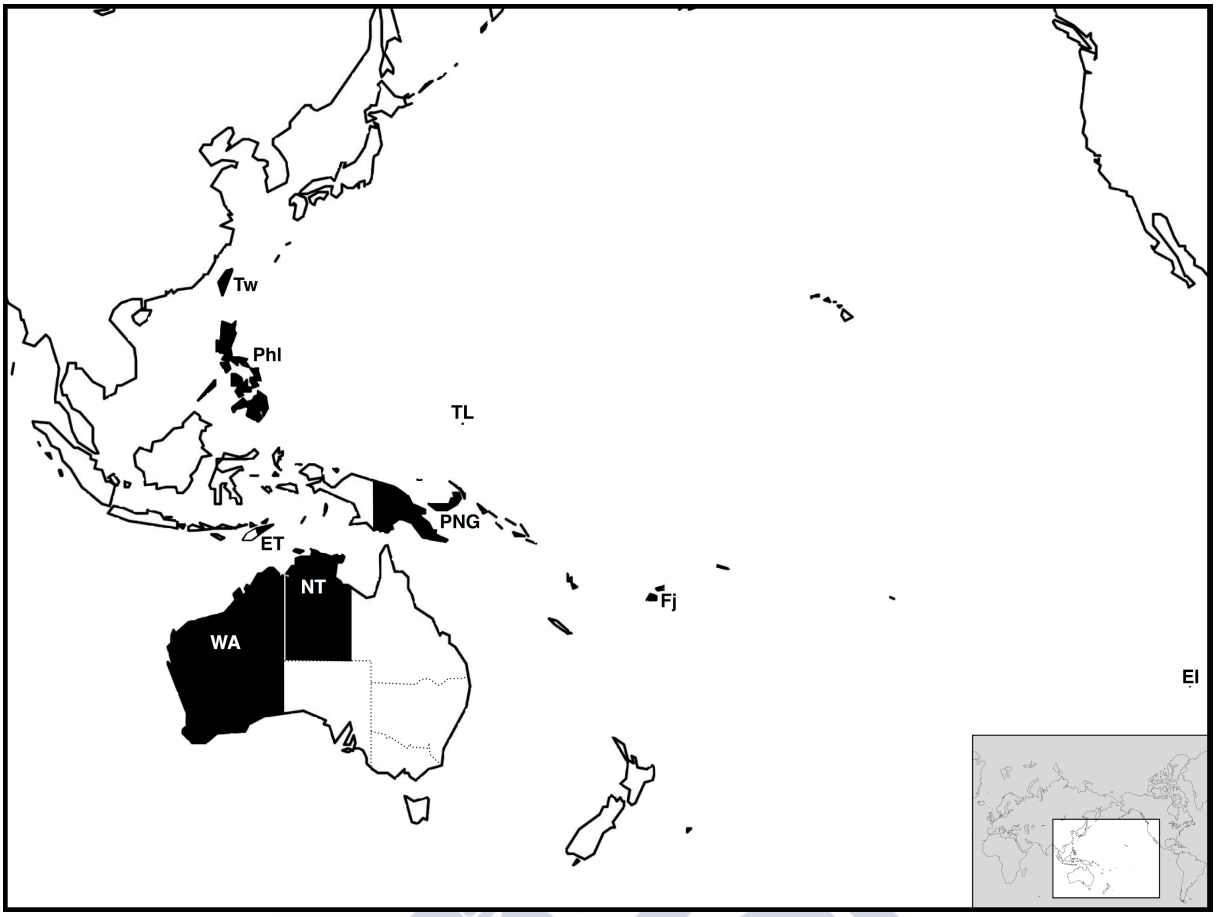


Figure 30: Geographic location of the nine populations genotyped in this study. Philippines (Phl), Aboriginal Taiwanese (Tw), Easter Island (EI), Truk Lagoon (TL), East Timor (ET), Fiji (Fj), Aboriginal Australians from Western Australia (WA), Aboriginal Australians from Northern Territory (NT) and Papua New Guinea (PNG).

4.1.2.2 Marker selection and development of the multiplexed SNP assay

An initial screening of the HGDP-CEPH 650,000 SNP genotype data (from (Li et al. 2008), available online at http://spsmart.cesga.es/) was performed to select the 5% most divergent markers for three pairwise population group *In*/Divergence comparisons: AFR vs. OCE; EUR vs. OCE; and EAS vs. OCE. Furthermore, AFR vs. EAS and EUR vs. EAS comparisons were also assessed in order to improve the differentiation of East Asian populations. A total of 3210 SNPs were selected in this initial screening. From this candidate set, 23 autosomal SNPs were chosen based on the following criteria: i) loci previously described and typed in the HGDP-CEPH panel of samples; ii) delta (δ) allele frequency differentials between OCE-pairwise comparisons higher than 0.5; iii) at least 10 Mb of distance between each selected SNP and SNPs of the SNP*for*ID *34-plex* forensic ancestry test (Phillips et al. 2007b, Fondevila et al. 2013); iv) context sequence devoid of characteristics

which cause primer design issues, i.e., homopolymer stretches, high %GC, etc.; and v) absence of non-specific interactions between primers. In the next stage of SNP selection, four SNPs identified as informative but not genotyped in the HGDP-CEPH panel were included: EAS-informative SNPs rs3827760 and, rs12434466 plus two AFR/OCE-informative tri-allelic SNPs (rs2184030 and, rs2069945). Two X-chromosome SNPs were also added: rs11156577 (OCE-informative) and rs4892491 (EAS-informative). The resulting 29 AIM-SNP panel was highly informative for Oceanian ancestry and complementary to the established *34-plex* forensic ancestry panel (see **Figure 31**, Table 10). Four exceptions to a minimum 10 Mb separation between *Pacifiplex* and *34-plex* SNPs occurred: rs2184030 vs. rs2065160 on chromosome 1; rs12434466 vs. rs730570 (chr. 14); rs2069945 vs. rs1321333 (chr. 20); and rs2274636 vs. rs1978806 (chr. 10). Most selected SNPs located in genes are intronic, except rs3827760 (EDAR), rs4683510 (CLSTN2) and rs2274636 (YME1L1). SNP rs3827760 is associated with hair morphology/thickness (OMIM: 604095.0011) with the C allele found at high frequencies in East Asian populations (Fujimoto et al. 2008a). Furthermore, this SNP was added to a revised *34-plex* assay (Fondevila et al. 2013) during development of *Pacifiplex,* creating a single overlap and a combined set of 62 SNPs. Eleven *Pacifiplex* SNPs were integrated into a 128-SNP next generation sequencing (NGS) ancestry panel developed by the EUROFORGEN-NoE Consortium (Phillips et al. 2014).

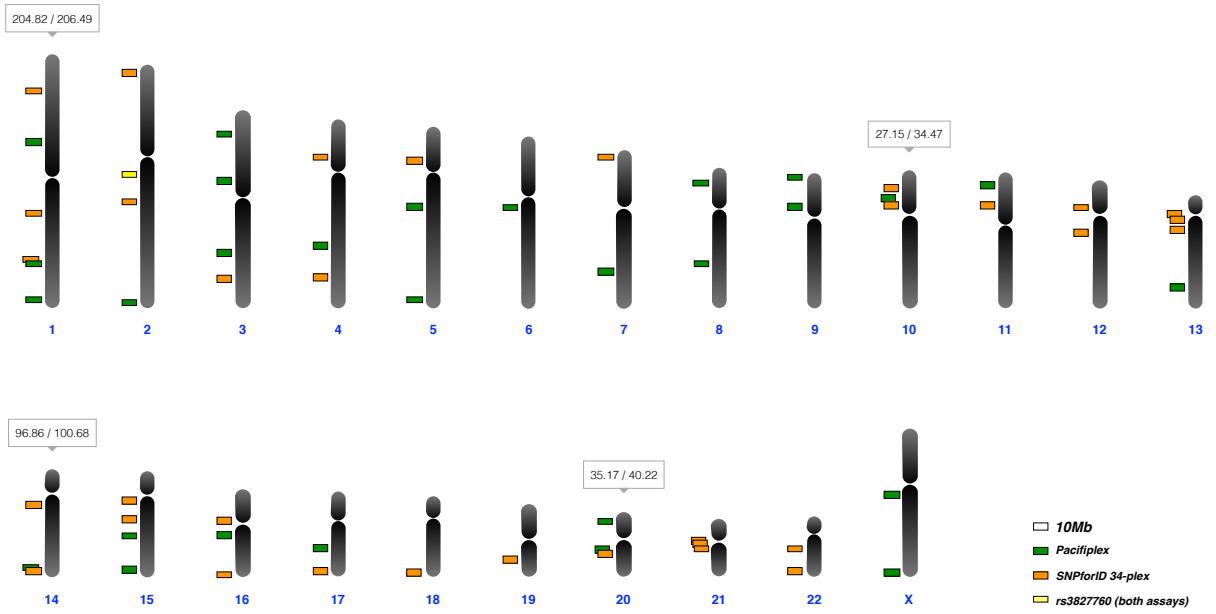


Figure 31: Chromosome positions of *Pacifiplex* (green) and *34-plex* (orange) SNPs (rs3827760 in common in yellow). Positions (Mb) for marker pairs with less than 10Mb separation are indicated. No evidence of linkage disequilibrium was found in the four HGDP-CEPH reference population groups (Africa, Europe, East Asia and Oceania) amongst these pairs.

Table 10: Genomic details of the SNPs included in *Pacifiplex*.

Internal code	rs number	Location	Position (bp) *	Gene	Alleles described in dbSNP
P1	rs9908046	17q12	53563782	-	CT
P2	rs721367	13q32.1	95546650	-	AG
P4	rs3827760	2q12.3	109513601	EDAR	CT
P5	rs4683510	3q23	140285115	CLSTN2	CT
P6	rs2139931	1p31.1	84590527	PRKACB	AG
P7	rs798949	7q31.31	120765954	CPED1	CT
P8	rs7832008	8q22.1	98358246	-	AG
P9	rs10811102	9p24.3	1911291	-	GT
P11	rs2409722	8p23.1	11039816	XKR6	GT
P12	rs9809818	3p13	71480566	FOXP1	AC
P13	rs6886019	5q35.1	170245846	-	CT
P14	rs3751050	11p15.4	9091244	SCUBE2	AG
P15	rs7623065	3p24.3	22385375	LOC100505836	AG
P16	rs1509524	4q28.1	125455038	-	AG
P17	rs16946159	16q12.1	48459558	-	AG
P18	rs2274636	10p12.1	27443012	MASTL, YME1L1	AG
P19	rs4704322	5q13.3	75822474	IQGAP2	CT
P20	rs12405776	1q43	242431557	PLD5	CT
P21	rs6054465	20p12.3	6673018	-	CT
P22	rs3784651	15q26.2	94925273	MCTP2	CT
P23	rs10970986	9p21.1	32453278	-	CT
P24	rs6494411	15q22.31	63835861	USP3	CT
P25	rs10183022	2q37.3	237481969	CXCR7	AG
P26	rs10455681	6q12	69802502	-	AG
P27	rs2184030	1q32.1	206667441	C1orf147	ACG
P28	rs2069945	20q11.22	33761837	PROCR	ACG
P29	rs11156577	Xq28	153660041	ATP6AP1	CT
P30	rs4892491	Xq13.2	73422412	FTX	AG
P31	rs12434466	14q32.2	97324289	VRK1	AG

* Mapping data according to dbSNP (build 138)

4.1.2.3 PCR amplification and genotyping

PCR amplifications were performed in 10 µl total volume containing: 1× GeneAmp PCR Gold buffer II (AB); 5.9 mM of $MgCl_2$ (AB); 3.2 µg of BSA; 0.5 mM of dNTPs (AB); 0.5 units of AmpliTaq Gold DNA Polymerase (AB); 2 µl of premixed PCR primers (detailed in Table 11, PCR primer mix ratios in Table 12) and up to 1 ng of genomic DNA. PCR cycling comprised: initial denaturation of 5 min at 95ºC; followed by 35 cycles of 30 sec at 95ºC, 40

Table 11: PCR and single base extension (SBE) primers; PCR amplicon sizes (bp); SBE amplification direction; observed bases; expected and observed sizes (bp) of the *Pacifiplex* multiplex.

Internal code	rs number	PCR forward primer	PCR reverse primer	PCR amplicon size (bp) [§]	SBE primers [*]	SBE amplification direction	Observed bases in SNaPshot	Expected SNaPshot product size (bp)	Observed SNaPshot size (bp)			
									G	A	C	T
P1	rs9908046	CATGAGGAGAGATGCAGATG	CAGATGAGGTTTACCTTGGC	75	GGAGGTAGGAGCACCCA	rev	A G	18	22.10	24.26		
P2	rs721367	TCTCTCAGACTTTCCTTCCC	CATAGACTGACTGTAATGTTTAG	86	TTCTCAC**Y**CCATAGGTC	rev	C T	18			23.50	26.22
P4	rs3827760	GCTCAGCTCCACGTACAAC	CTGTCATGCCCCCAATCTC	101	tctgacaactaggtgccacgtggtgaaagtctgacaactaggtgccacgtcgtgaaagtctgacaactctcaGG**Y**GCCA**Y**GTTTTCACA	rev	A G	90	90.00	90.30		
P5	rs4683510	GATGACTCCACCCTCCC	GGTCAGGGTTTACAAAAGGG	75	ctctctcgaaCCAGGGGTCTGCTGCC	fwd	C T	27			29.10	31.50
P6	rs2139931	TAGTTATGGGACTACCTGGC	TAGAGCCCTTCGTGTTGATG	82	ctctctcaaAGGGCGTTAG**K**AGATCA	fwd	A G	27	29.00	30.00		
P7	rs798949	TCATGTACCCTGTAGTGCTC	ACTGATCTCTCACCACAAGG	76	tctctctctctttcAGTGCTCCAGGTCAGTA	fwd	C T	32			35.20	36.60
P8	rs7832008	TTGGACACTTGGTGGTTAGC	GGGAGTGTAAGAAAGATGTG	98	ctctctctctctctccAGCCACTCCTGACAC	fwd	A G	32	32.55	33.60		
P9	rs10811102	GGAAAGGCCTTAAAGAGCAT	GATTCCTGTGTCTCTGACAT	99	tctctctctctcctTGCAGGTAACTTACTTTACA	rev	C A	35		37.80	37.00	
P11	rs2409722	TGAAAGAGCACTCTTCTGAC	CAACCCCAGTCTCATTAAGG	79	tctctctctctctctcggaTGACTTAAAATACCACACAG	fwd	G T	41	41.30			42.80
P12	rs9809818	TGTGTGGTTTTCTCAGCGAC	AGCATGGTATGAGCACTGAG	89	tctctctctctctctctctctcTCCTAATAAGAGCTGGC	fwd	C A	41		43.30	42.30	
P13	rs6886019	GGCAAACATGTCTCTGATGG	CATAGATATGATTTGGTTTATGAATG	115	tctctctctctctctctctct

sec at 63.5ºC, 30 sec at 72ºC; then a final extension of 15 min at 72ºC. Prior to single base extension (SBE), PCR products were cleaned with Exonuclease I and Shrimp alkaline phosphatase (SAP) by adding 0.5 µl of 1:2 diluted Illustra™ ExoProStar™ 1-Step (GE Healthcare) to 1.25 µl of PCR product and incubating at 37ºC for 45 min (enzymes then inactivated at 85ºC for 15 min). SBE reactions performed in 3 µl volumes combined 0.625 µl SNaPshot® Multiplex Ready Reaction Mix (AB), 0.75 µl of premixed SBE primers (sequence details in Table 11 and SBE primer mix details Table 12), 1 µl of cleaned PCR product. SBE cycling comprised: 30-35 cycles of 10 sec at 96ºC, 5 sec at 55ºC, 30 sec at 60ºC. SBE products were cleaned by adding 1 µl of SAP (GE Healthcare) to the whole reaction volume and incubating at 37ºC for 80 min (enzymes then inactivated at 85ºC for 15 min).

CE detection combined 3 µl of purified SBE product with a mix of 9.2 µl Hi-Di™ Formamide (AB) and 0.3 µl of GeneScan™ 120 LIZ™ dye Size Standard (AB). CE used standard SNaPshot® electrophoresis conditions in an ABI Prism 3130xl Genetic Analyzer (AB) with 36 cm capillary arrays and POP-4™ polymer (AB). Electropherograms were analysed using GeneMapper® ID v3.2 (AB). SNP*for*ID *34-plex* genotyping was performed as described in (Fondevila et al. 2013) using 32 PCR amplification cycles and 30 SBE cycles.

4.1.2.4 Statistical analysis

Allele frequency estimates, F_{ST} genetic distances, exact tests of Hardy-Weinberg equilibrium (HWE) and linkage disequilibrium (LD) tests were performed using Arlequin v3.5.1.2 (Excoffier and Lischer 2010). The *Snipper* 2.0 online Bayes analysis suite (http://mathgene.usc.es/snipper/) was used to calculate likelihoods of ancestry in all statistical assessments of the SNPs analysed. The *Snipper* site calculates each SNP's Shannon's Divergence value (Phillips et al. 2007b) and these were converted to Rosenberg's informativeness for assignment, *In* (Rosenberg et al. 2003), with *In*=Divergence × 0.69). Together with δ, both F_{ST} and *In* (among other metrics) are comparable measures of each marker's contribution to an ancestry inference (Rosenberg et al. 2003, Phillips et al. 2007b, Winkler et al. 2010).

Table 12: PCR and single base extension (SBE) primer mix ratios.

		PCR primer mix		SBE primer mix	
Internal code	rs number	Volume (µl)[§]	Concentration in mix (µM)	Volume (µl)[*]	Concentration in mix (µM)
P1	rs9908046	0.75	0.125	0.75	0.45
P2	rs721367	9	1.5	8	4.8
P4	rs3827760	0.75	0.125	5	3
P5	rs4683510	1.5	0.25	1	0.6
P6	rs2139931	3	0.5	1	0.6
P7	rs798949	1.5	0.25	1	0.6
P8	rs7832008	3	0.5	1	0.6
P9	rs10811102	3.75	0.625	1.5	0.9
P11	rs2409722	3	0.5	1	0.6
P12	rs9809818	1.5	0.25	2	1.2
P13	rs6886019	12	2	1.2	0.72
P14	rs3751050	0.75	0.125	0.5	0.3
P15	rs7623065	1.5	0.25	1	0.6
P16	rs1509524	7.5	1.25	1.5	0.9
P17	rs16946159	24	4	10	6
P18	rs2274636	0.75	0.125	2	1.2
P19	rs4704322	0.75	0.125	0.5	0.3
P20	rs12405776	1.5	0.25	3	1.8
P21	rs6054465	3	0.5	0.5	0.3
P22	rs3784651	0.75	0.125	2	1.2
P23	rs10970986	1.5	0.25	1.5	0.9
P24	rs6494411	1.5	0.25	1	0.6
P25	rs10183022	1.5	0.25	1.2	0.72
P26	rs10455681	3	0.5	0.5	0.3
P27	rs2184030	0.75	0.125	1.2	0.72
P28	rs2069945	1.5	0.25	1	0.6
P29	rs11156577	12	2	5	3
P30	rs4892491	12	2	10	6
P31	rs12434466	6	1	2	1.2
	H_2O	180		2.55	
	$(NH_4)_2SO_4$ 400mM	-		12.92	
	Total volume	300		83.32	

[§] Forward + Reverse PCR primer stock at 50µM

[*] SBE primer stock at 50µM

Ancestry inferences were made using *STRUCTURE* v 2.3.4 (Pritchard et al. 2000) with 100,000 burnin steps and 100,000 MCMC iterations. Two different ancestry models were considered: i) *admixture* without prior information on sample origin – used to predict clustering of HGDP-CEPH reference individuals as well as those of study population; ii) *admixture POPFLAG* - HGDP-CEPH individuals were defined as the reference data and used by *STRUCTURE* to infer probable ancestry of study populations. Three independent replicates from *K*=1 to *K*=7 genetic clusters were performed for the stand-alone analysis of the HGDP-CEPH individuals and for assessment of study populations using HGDP-CEPH populations as reference data. Ten independent replicates from *K*=1 to *K*=7 genetic clusters were performed for analysis of study populations alone. In all analyses the *correlated allele frequencies* model was applied (Falush et al. 2003). Estimated Ln probability values (-LnP(D)) for each *K* value were plotted with *Structure Harvester* v 0.6.93 (Earl and vonHoldt 2012) that also estimated optimum *K* using the Evanno method (Evanno et al. 2005) and generated *CLUMPP* input files. *CLUMPP* v1.1.2 (Jakobsson and Rosenberg 2007) and *distruct* v 1.1 (Rosenberg 2004) were applied as described later in section 3.1.3, to create the cluster bar plots. Principal component analysis (PCA) was used as a complementary tool for assessment of optimum *K* values and to provide an outline of the genetic relationships between reference and study populations. PCA plots were generated using *R* 3.1.0 (Team 2011) (script presented in section 3.1.5.3). Classification success using the *Pacifiplex* and *34-plex* tests on their own and combined was estimated by cross-validation of the reference HGDP-CEPH data for four population groups using *Snipper* (http://mathgene.usc.es/snipper/analysispopfile2_new.html).

4.1.3 Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing (Appendix 2B)

4.1.3.1 Population samples

A total of 1002 DNA samples were used in this study comprising: i) reference samples from the HGDP-CEPH diversity panel standardized subset H952 (Cann et al. 2002, Rosenberg 2006) with origin in Africa (AFR), Europe (EUR), East Asia (EAS), America (NAM) and also Oceania (OCE), representing a total 584 individuals from 40 populations. Individuals HGDP01219, 1339, 1344 and 1041 were not included in the study since no DNA

was available for analysis; in substitution of 1041 we used 1042 who had been excluded from subset H952 due to a parent/offspring relationship with 1041 (Rosenberg 2006); ii) samples from Angola (48), Portugal (48), Taiwan (48) and Brazilian Amazonas tribes (48) used in a preliminary evaluation of the AIM-Indel assay and as example testing samples; iii) samples from the city of Belém (226), an admixed population in north-eastern Amazonas, Brazil.

4.1.3.2 AIM-Indel selection and development of the multiplex reaction

An initial pool of candidate Indels was assembled by collecting previously available population data on this type of polymorphism included in the Marshfield Diallelic Insertion/ Deletion Polymorphisms database website (http://www.marsh-fieldclinic.org/mgs/; (Weber et al. 2002)) and from later studies that also characterized some candidate Indels in different population groups (Rosenberg et al. 2005, Yang et al. 2005, Pereira et al. 2009, Santos et al. 2010). Considering the allele frequency data compiled from the diverse sources, all markers were sorted according to frequency differentials – δ (Shriver et al. 1997) – comparing four human population groups of Africans, Europeans, East Asians and Native Americans. For this study we selected a set of 46 markers (Table 13) among the most informative Indels for each population group (all with $\delta \geq 0.40$ between at least two groups) and optimized a unique multiplex reaction allowing the simultaneous genotyping of all AIMs in a single PCR and electrophoretic run. The multiplex development followed a similar workflow as in (Pereira et al. 2009, Pereira et al. 2012a) except for the accommodation of certain longer amplicons into a broadened size window up to 230 bp in order to type an extended number of Indels in a single reaction.

4.1.3.3 Amplification and genotyping

PCR amplification of the *46 AIM-Indels* used the QIAGEN Multiplex PCR kit (Qiagen) at 1× Qiagen multiplex PCR master mix, 0.1 μM of all primers (sequence details in Table 14) and 0.3-5 ng of genomic DNA in a 10 μL final reaction volume. Thermocycling conditions were: initial step at 95ºC for 15 min; 30 cycles at 94ºC for 30 sec, 60ºC for 90 sec, and 72ºC for 45 sec; and a final extension at 72ºC for 60 min. The PCR products were then prepared for

Table 13: AIM-INDELs used in the multiplex.

MID*	rs number	Chromosome	Position (bp)**	Alleles described in dbSNP
MID-1470	rs2307666	11	64729920	-/GTTAC
MID-777	rs1610863	16	6551830	-/GAA
MID-196	rs16635	6	99789775	-/CAT
MID-881	rs1610965	5	79746093	-/ACTT
MID-3122	rs35451359	18	45110983	-/ATCT
MID-548	rs140837	6	3708909	-/CT
MID-659	rs1160893	2	224794577	-/CT
MID-2011	rs2308203	2	109401291	-/CTAGA
MID-2929	rs33974167	8	87813725	-/TA
MID-593	rs1160852	6	137345857	-/TT
MID-798	rs1610884	5	56122323	-/GGGAAA
MID-1193	rs2067280	5	89818959	-/AT
MID-1871	rs2308067	7	127291541	-/TT
MID-17	rs4183	3	3192524	-/TAAC
MID-2538	rs3054057	15	86010538	-/AACA
MID-1644	rs2307840	1	36099090	-/GT
MID-3854	rs60612424	6	84017514	-/TCTA
MID-2275	rs3033053	14	42554496	-/TCAGCAG
MID-94	rs16384	22	42045009	-/AAC
MID-3072	rs34611875	18	67623917	-/GCCCCCA
MID-772	rs1610859	5	128317275	-/TAG
MID-2313	rs3045215	1	234740917	-/ATTATAACT
MID-397	rs25621	6	139858158	-/TTCT
MID-1636	rs2307832	1	55590789	-/AA
MID-51	rs16343	4	17635560	-/TTTAT
MID-2431	rs3031979	8	73501951	-/ATTG
MID-2264	rs34122827	13	63778778	-/AAGT
MID-2256	rs133052	22	41042364	-/CAT
MID-128	rs6490	12	108127168	-/ATT
MID-15	rs4181	2	42577803	-/AAATACACAC
MID-2241	rs3030826	6	67176774	-/GTCCAATA
MID-419	rs140708	6	170720016	-/AATGGCA
MID-943	rs1611026	5	82545545	-/TGAT
MID-159	rs16438	20	25278470	-/CCCCA
MID-2005	rs2308161	10	69800909	-/AACAAT
MID-250	rs16687	7	83887882	-/CA
MID-1802	rs2307998	5	7814345	-/GGA
MID-1607	rs2307803	3	108981031	-/TG
MID-1734	rs2307930	6	84476378	-/CCAT
MID-406	rs25630	6	14734341	-/AG
MID-1386	rs2307582	1	247768775	-/AAACTATTCATTTTTCACCCT
MID-1726	rs2307922	1	39896964	-/CAAGAACTATAAT/CACTATCTATTAT
MID-3626	rs11267926	15	45526069	-/AATATAATTTCTCCA
MID-360	rs25584	12	112145217	-/AA
MID-1603	rs2307799	5	70828427	-/TTGT
MID-2719	rs34541393	20	30701405	-/AACT

* Nomenclature according to [7] and Marshfield Diallelic Insertion/Deletion Polymorphisms database;

** Mapping data according to dbSNP (build 132).

Table 14: PCR primer sequences used in the multiplex.

MID	rs number	Labeled primer	Unlabeled primer*
MID-1470	rs2307666	6FAM-GAGTCTGACCCTTCATAAGC	*g*CCATGGTGATATTACGTCCC
MID-777	rs1610863	6FAM-TGGAAGACACGTCCTAAGAG	*g*TATTCCTCCAGGCTCTTTGC
MID-196	rs16635	6FAM-CCAAGTTCTAGCCATATGGA	*gtttctt*GACTATCTTCTCTGACCATC
MID-881	rs1610965	6FAM-TTGGCTCCCTATGATAATCC	*gtttc*TTGTGTTCCCAAAGTTCTCC
MID-3122	rs35451359	6FAM-TCACAAGTCCGGAATACCAG	*g*AGTTATGGGATGGGAAGGAG
MID-548	rs140837	6FAM-AGTCAGGACTGAAGAAACCC	*gttt*CAGTAAAACAAGAGCCCGTG
MID-659	rs1160893	6FAM-CACTGCATCAGACTGACTTC	*gtttctt*GGCTGCTTTGCTTTGAATTG
MID-2011	rs2308203	6FAM-TGAGAAACTAGGAGCTCTGG	*gt*TTCCTAAGAGCCACTGACAT
MID-2929	rs33974167	6FAM-TGTGATGTGGATAGGCAAGG	*g*AGGCTCCATTTGTTAAGAGG
MID-593	rs1160852	6FAM-TGCTCACTTTAGTGGAGACC	*gt*TTGCGTTTAGGTCCCTTCTG
MID-798	rs1610884	6FAM-ACGACAGTGTTCACAAAGAG	*g*CTGTTGTCTGACCTGTGAAG
MID-1193	rs2067280	6FAM-GCTGGGTAGTTTTTTCCTCC	*gt*TCCACCATCTACCTTCTATG
MID-1871	rs2308067	6FAM-TTGTAGTCAGAGAGTGTGCC	*g*AGCCTTTTCCCTAACGTCAC
MID-17	rs4183	6FAM-AGAAACTGCAACCCTCCAAG	GATCCCAGACACTGAAGATG
MID-2538	rs3054057	6FAM-CTCGCAAAGTAGGCAAGTTC	*g*ACACCAACAATCTTGGCACC
MID-1644	rs2307840	6FAM-ACACCACTGAAGATCTGACC	GGTCTAAAGTCAGTGCACAG
MID-3854	rs60612424	VIC-TCACCCTTATTCAGGGTTGC	GCCAGGGATTTAGTGTAGAG
MID-2275	rs3033053	VIC-CTACCTGACTACCACCTATG	*g*ACCCAGCCTATCTGACTTTG
MID-94	rs16384	VIC-TGGTGGCTCATGCACTTTTG	*gttt*ACAGGGTCTCGCTATGATGC
MID-3072	rs34611875	VIC-AGCTTTTTCCGGCAACTCTC	*gtt*TGGATGTGTCTGAGCTCAAC
MID-772	rs1610859	VIC-GTCTCRTTTTCCTGCAGTAG	*gtttctt*ATCCTTCTGCTCACTCTACC
MID-2313	rs3045215	VIC-GCACACATGCAGAAATGCAG	GTTGTAACATCTGTGAGGTC
MID-397	rs25621	VIC-TGGGCTTCTTCTGGGAAAAC	*g*CCACATTCAGGCGTTTTGTC
MID-1636	rs2307832	VIC-TTAGGAAGAGGTGCTATGGG	*g*CCTCCTTTGAAGACACACAG
MID-51	rs16343	VIC-AAGATTGGAGGGAAAAGTGC	*g*CGTCCTCCACCTTCTTTTTC
MID-2431	rs3031979	VIC-AGGAGGAGCTGATAGACTTC	GCAGTGTTGCAACTGATACG
MID-2264	rs34122827	VIC-CTTTGGCTATCCTGTCTCAC	GTAGGAGACCACTCACATTC
MID-2256	rs133052	NED-ATCGAACCGTTCCTAAGGAC	GCAAGAAAAGGAATCCAGGC
MID-128	rs6490	NED-ATCAGGAGACAATCCAGCAG	*g*TCCAGCCATTCAGACAAAGG
MID-15	rs4181	NED-GGGTTATTTGCCTCATCTCC	*gtttct*AGGTATTCTCTGTTCCCACG
MID-2241	rs3030826	NED-ACATACACGTGGAAGACTGC	*gtt*ACTGTCGACTGATCCAATAG
MID-419	rs140708	NED-CAGGAAAGTATGGCCCATTC	GTCCATGTTTTCTTTGAGCATC
MID-943	rs1611026	NED-TCTTCCTACCCCTGTTAGTG	GACAAGATCACTAGCTTGAC
MID-159	rs16438	NED-ACCAGAGCACTACAGCCTTT	*g*CAAGGYAGTAACAATGAGGG
MID-2005	rs2308161	NED-TGTAGCGGCAATATAGGCAG	GAAAGTTGTGGCTTAACTGG
MID-250	rs16687	NED-ATGGAGCAGTAAAGCAGCAC	GTCACTTTGGTTTTTGCAGG
MID-1802	rs2307998	NED-ACGGTCAACTTTGTAGCTCC	*g*CCAGTTGAGAATCACTGCAC
MID-1607	rs2307803	NED-TGTTGCAGAAGAACTCAACC	GATAAGCACCTAACTCCCAG
MID-1734	rs2307930	PET-TTCGTGTTCTCACACTGTCC	GTGCATCCCATACAACTGAC
MID-406	rs25630	PET-TGGCTGCTGTAGATTGTAGG	*g*ACAAATGGACAACGGCCAAG
MID-1386	rs2307582	PET-AGAGGATCATGGAGACCAAC	*gttt*ATGTTCCAAGTCAGCAGCAC
MID-1726	rs2307922	PET-GGTCCAAATGCACCACAATC	GCTCTGCTATTTTGGTTTGC
MID-3626	rs11267926	PET-TGTTGGTTCTCTCCTTTTCC	GGTGACCCCTTCTTTATCTC
MID-360	rs25584	PET-AGATCAACTGCCAATCTGGG	*g*CTCAAGTGACCAACCCACCT
MID-1603	rs2307799	PET-TTACAATTTCAAGCCTCCGC	GGAGCTGTTAGTCTGAGTAG
MID-2719	rs34541393	PET-GTCAGGAGTCTAGAAACTTC	GGGTGATGAAATGTTCCGAA

* Lowercase italic letters represent nucleotide tails added to the primers.

capillary electrophoresis (CE) by adding 1 µL of amplified product to 10 µL Hi-Di™ formamide (AB) and 0.3 µL of GeneScan™ 500 LIZ® size standard (AB). CE was performed using a 3130 Genetic Analyzer prepared with DS-33 matrix standard, POP-7™ polymer and applying virtual filter G5 (AB). The electropherograms were analysed and genotypes were automatically assigned with GeneMapper v4.0 (AB). For practical reasons Indel short alleles were coded as 1 and long alleles as 2.

4.1.3.4 Statistical analysis

Estimation of allele frequencies, exact tests of Hardy-Weinberg equilibrium (HWE), F_{ST} genetic distances and linkage disequilibrium tests were assessed using Arlequin v3.5.1.2 (Excoffier and Lischer 2010). Ancestry inferences were performed using *STRUCTURE* v2.3.3 (Pritchard et al. 2000, Falush et al. 2003) with a burnin length of 100,000 followed by 100,000 MCMC repetitions and a variety of parameter sets were tested depending on the objective of the analysis. Initial runs were made without any prior information on the origin of samples, using the '*admixture* model' and considering either correlated or independent 'allele frequency models'; a minimum of 3 independent runs were performed for each testing *K* value, ranging from *K*=1 to *K*=number of presumed clusters present in the dataset plus three. The estimated ln probability of data (-lnP(D)) values were plotted using *Structure harvester* v0.6.6. (http://taylor0.biology.ucla.edu/structureHarvester/). In a second phase, when using reference samples as training sets to test for 'unknown' individuals or populations, *STRUCTURE* analyses were carried out using the same parameters as before or selecting the 'use population information' option. In these cases, allele frequencies were updated using only the reference individuals with POPFLAG=1 data (option under the Advanced tab). Here, 3 independent runs were performed only for the appropriate number of clusters, as evaluated by the initial analysis. Unless otherwise indicated, results are presented for the default settings considering the '*admixture* model' and correlated allele frequencies. *CLUMPP* v1.1.2 (Jakobsson and Rosenberg 2007) was used to obtain the average permutated individual and population Q-matrices throughout the three replicates for each *K* value. Those matrices were used as input to *distruct* v1.1 (Rosenberg 2004) to obtain bar plots where each individual is represented as a segment divided into *K* colours that represent the estimated membership coefficients from each cluster.

Principal component analysis (PCA) was performed as an additional and independent approach to estimate the number of populations present in the data set. We used *R* 2.11.1 (Team 2011) with *SNPassoc* package (Gonzalez et al. 2007) to obtain two and three-dimensional graphics and the information percentage values associated to each principal component.

The efficiency of the *46 AIM-Indel* set for assigning individuals to population groups was further evaluated by one-out cross-validation based on a flexible single profile analysis system very similar to *STRUCTURE*, calculating likelihood ratio values obtained with a Bayesian classification algorithm implemented in the ''*Snipper* app suite'' website (http://mathgene.usc.es/snipper/; (Phillips et al. 2007b)).

4.1.4 Completion of a world wide reference panel of samples for an ancestry informative Indel assay (Appendix 2C)

4.1.4.1 Population samples

A total of 365 samples from 13 populations of the HGDP-CEPH standardized subset H952 (Cann et al. 2002, Rosenberg 2006) were genotyped: 163 ME and 202 CSA individuals.

4.1.4.2 Amplification reactions

PCR amplification was performed following the protocol in section 4.1.3.3 using 0.75-1 ng/µL of DNA and 28 PCR cycles. Singleplex PCR amplification of component Indel MID-548 used 1× Qiagen multiplex PCR master mix, 0.2 mM primers and 1 ng genomic DNA in 10 µL volumes. Singleplex cycling conditions matched multiplex amplifications. Capillary electrophoresis used an ABI Prism 3130*xl* Genetic Analyzer with 36 cm capillary arrays and POP-4™ polymer (AB). Electropherograms were analysed using GeneMapper® ID v3.2 (AB). MID-548 singleplex PCR products were cleaned up prior to the sequencing reaction using Illustra™ ExoProStar™ 1-step (GE Healthcare) by adding 1 µL of 1:10 ExoProStar to a 2.5 µL aliquot of PCR product. Samples were incubated at 37ºC for 45 min then heat-inactivated at 85ºC for 15 min. Sanger sequencing used BigDye® Terminator v3.1 Cycle Sequencing Kit (AB) and 5µM of reverse PCR primer described in Table 14. Sequence

analysis thermocycling conditions were: denaturation at 96ºC for 1 min; 35 cycles at 96ºC for 30 sec, 50ºC for 10 sec, 60ºC for 1 min. Sequencing products were purified using Centri-Sep™ spin columns (AB) and capillary electrophoresis was performed in an ABI 3730*xl* DNA Analyzer using POP-7™ polymer (AB). Sequence electropherograms were analysed using SeqScape® v2.5.0 (AB).

4.1.4.3 Statistical analyses

Allele frequencies, Hardy Weinberg equilibrium (HWE), linkage disequilibrium (LD) and F_{ST} genetic distances were estimated using Arlequin v 3.5.1.3 (Excoffier and Lischer 2010). Ancestry inferences considering the complete HGDP-CEPH data (African - AFR; European - EUR; East Asian - EAS; Native American - NAM; Oceanian - OCE genotypes available in Supplementary File S1 of (Pereira et al. 2012b)) were made using *STRUCTURE* v 2.3.3 (Pritchard et al. 2000, Falush et al. 2003) with 100,000 burnin steps and 100,000 MCMC iterations. Two different ancestry models were considered: i) *admixture* without prior information on sample origin; ii) *admixture LOCPRIOR* (Hubisz et al. 2009) using additional information, e.g. geographic origin of samples, to help make ancestry estimates (particularly useful when population structure signals are weak). Three independent analyses from K=2 to K=10 genetic clusters were performed for each ancestry model always considering correlated allele frequencies. Estimated ln probability values (-LnP(D)) for each K value were plotted with *Structure harvester* v 0.6.93 (Earl and vonHoldt 2012) also estimating optimum K using the Evanno method (Evanno et al. 2005) and generating *CLUMPP* input files. *CLUMPP* v1.1.2 (Jakobsson and Rosenberg 2007) and *distruct* v 1.1 (Rosenberg 2004) were applied as described in (Pereira et al. 2012b) to create cluster plots.

To evaluate the advantages of joint analysis of the *46 AIM-Indels* with an established forensic set of 34 AIM-SNPs (Fondevila et al. 2013), we performed ancestry inference analyses, as described above, for 34 AIM-SNPs HGDP-CEPH data alone (available at http://spsmart.cesga.es/snpforid.php?dataSet=snpforid34) or combined with *46 AIM-Indel* data (80 loci). Note the joint analysis considered only 937 individuals, as twelve AIM-Indels profiles lacking AIM-SNPs were removed to ensure full 80 marker profiles. PCA was performed in *R 2.13.1* (Team 2011) (script presented in section 3.1.5.3). Finally, we used the *Snipper 2.0 app suite* (http://mathgene.usc.es/snipper/) to estimate the classification success

of AIM-Indels through cross-validation of the full HGDP-CEPH data, then compared this with classification success of the combined 80 marker set.

Formal statistical assessment of *STRUCTURE* genetic cluster patterns was made by applying ANOVA and linear regression (*SPSS v.17.0.0*). To evaluate the differentiation shown by PCA cluster patterns we applied Euclidean distance estimation to measure the degree of centroid separation and Ray-Turi clustering index analysis ((Ray and Turi. 1999); estimated using *R* 3.1.0, package *clusterCrit v1.2.4*).

4.1.5 *Eurasiaplex*: a forensic SNP assay for differentiating European and South Asian ancestries (Appendix 2D)

4.1.5.1 Population samples and SNP genotype databases accesses

SNP variability in Eurasian and East Asian populations was collected from four sources: i) 650,000 SNPs genotyped in 51 HGDP-CEPH populations by Stanford University (Li et al. 2008), ii) HapMap SNP genotypes for 1218 samples from 11 populations (3 Eurasian), iii) 1000 Genomes SNP genotypes for 1093 samples from 14 populations (5 Eurasian), and iv) our study populations comprising: 38 NW Spanish; 31 Germans; 44 Moroccans; 23 Indians from Murcia, Spain; 35 Vietnamese; 42 Greeks; 68 Turkish; 16 Afghans; 38 Iraqis and 8 Iranians. HGDP-CEPH genotypes for eight European, four Middle East/North African, nine Central-South Asian and seventeen East Asian populations, were accessed with *SPSmart* (http://spsmart.cesga.es). HGDP-CEPH Central-South Asians comprise 8 Pakistani populations plus Uygur from extreme western China. *SPSmart* also accessed HapMap Gujarati Indians from Houston (GIH), though seventeen *34-plex* SNPs are not genotyped in GIH.

4.1.5.2 SNP selection and multiplex construction

Sixty candidate SNPs were shortlisted showing highest divergence by estimating Rosenberg's ancestry informativeness metric: *In* (Rosenberg et al. 2003) comparing HGDP-CEPH Central-South Asians *versus* HGDP-CEPH Europeans (excluding Adygei, Sardinian outlier populations). From candidates 23 SNPs were selected (Table 15) with optimum

Table 15: Genomic details and public database coverage of *Eurasiaplex* SNPs. Bold alleles signify 1000 Genomes and SNaPshot assay strands are the same direction and ticks in database coverage denote the presence of SNP data.

Genomic details					1000 Genomes		SNaPshot assay		Reference allele frequencies				SNP database coverage		
No.	SNP	Chromosome	Mb position	Gene	Reference allele	Alternative allele	Reference allele	Alternative allele	HGDP-CEPH EUR[a]	HGDP-CEPH ME	HGDP-CEPH S ASN	HGDP-CEPH E ASN	1000 Genomes	HapMap	Perlegen
1	rs10008492	4	38.77	–	C	T	**C**	**T**	0.61	0.10	0.19	0.01	✓	✓	×
2	rs10131666	14	74.21	C14orf43	A	C	T	G	0.65	0.38	0.33	0.16	✓	✓	✓
3	rs10962599	9	16.80	BNC2	C	T	**C**	**T**	0.25	0.71	0.52	0.98	✓	✓	✓
4	rs11779571	8	18.54	PSD3	T	C	**T**	**C**	0.41	0.76	0.49	0.90	✓	✓	✓
5	rs1363345	19	13.57	CACNA1A	G	A	C	T	0.68	0.38	0.64	0.16	✓	✓	✓
6	rs1519654	2	101.10	NMS	C	T	G	A	0.65	0.24	0.49	0.02	✓	✓	✓
7	rs1544656	12	113.01	–	T	C	A	G	0.89	0.47	0.74	0.15	✓	✓	×
8	rs17625895	16	25.78	HS3ST4	A	C	**A**	**C**	0.47	0.85	0.59	0.99	✓	✓	✓
9	rs1785864	11	104.96	–	A	G	T	C	0.38	0.68	0.71	0.61	✓	✓	✓
10	rs1941411	11	120.13	POU2F3	A	G	**A**	**G**	0.68	0.29	0.55	0.02	✓	✓	✓
11	rs2156208	18	60.13	–	C	T	G	A	0.59	0.21	0.46	0.09	✓	✓	✓
12	rs2196051	8	122.12	–	A	G	T	C	0.64	0.25	0.51	0.02	✓	✓	✓
13	rs2227203	1	172.88	–	T	C	A	G	0.49	0.25	0.18	0.08	✓	✓	✓
14	rs2472304	15	75.04	CYP1A2	G	A	**G**	**A**	0.40	0.76	0.49	0.88	✓	✓	✓
15	rs2835133	21	37.13	–	C	A	G	T	0.98	0.85	0.71	0.73	✓	✓	✓
16	rs39897	5	131.44	–	C	T	**C**	**T**	0.49	0.68	0.75	0.59	✓	✓	✓
17	rs6026972	20	58.15	–	T	G	A	C	0.71	0.30	0.43	0.06	✓	✓	✓
18	rs734482	3	181.59	–	A	G	**A**	**G**	0.39	0.80	0.73	0.96	✓	✓	×
19	rs7354930	1	66.99	–	T	C	A	G	0.40	0.09	0.27	0.00	✓	✓	✓
20	rs756913	7	24.54	–	G	A	**G**	**A**	0.59	0.21	0.45	0.01	✓	✓	✓
21	rs9487258	6	110.26	–	C	T	**C**	**T**	0.80	0.42	0.69	0.13	✓	✓	×
22	rs9522149	13	111.83	ARHGEF7	T	C	A	G	0.20	0.65	0.31	0.97	✓	✓	×
23	rs984038	3	65.76	MAGI1	C	T	G	A	0.28	0.57	0.64	0.39	✓	✓	✓

[a] Excluding Sardinians and Adygei.

multiplex performance. *Eurasiaplex* loci show good genomic separation from *34-plex* SNPs (genomic map in **Figure 32**). We combined *Eurasiaplex* SNPs into single PCR and primer extension multiplexes using SNaPshot (AB) with ~4-5 bp spacing using pigtails. PCR amplification used 2 µL of Qiagen multiplex PCR master mix, 2 µL of PCR primer mix (detailed in Table 16) and 1 ng of genomic DNA. Thermocycling conditions were: initial step at 95ºC for 15 min; 35 cycles at 94ºC for 30 sec, 60ºC for 60 sec, and 72ºC for 50 sec; and a final extension at 72ºC for 10 min. PCR products were cleaned up prior to extension using combined Exonuclease I and Shrimp alkaline phosphatase by adding 1 µL Exo-SAPit (USB Products, Affymetrix, Santa Clara, CA, U.S.) to a 2.5 µL aliquot of PCR product. Samples were incubated at 37ºC for 45 min then enzymes were heat-inactivated at 85ºC for 15min. SBE reactions performed in 3 µl volumes combined 1.25 µl SNaPshot® Multiplex Ready Reaction Mix (AB), 0.75 µl of premixed SBE primers (detailed in Table 16), 1 µl of purified PCR product. SBE cycling comprised: 30 cycles of 10 sec at 96ºC, 5 sec at 55ºC, 30 sec at 60ºC. SBE products were cleaned by adding 1 µl of SAP (GE Healthcare) to the whole reaction volume and incubating at 37ºC for 80 min (enzymes then inactivated at 85ºC for 15 min). Capillary electrophoresis preparation combined 1 µl purified extension products, 10 µl AB HiDi™ formamide and 0.25 µl AB LIZ-120. Separations used AB 3130*xl* detectors, POP-4™, 36 cm capillaries, then AB GeneMapper ID v. 3.2. *34-plex* reactions were as originally described (Phillips et al. 2007b).

To test *Eurasiaplex* sensitivity fifteen typical challenging forensic casework samples and five highly degraded DNAs were analysed in parallel to standard SEFiler and Identifiler STR genotyping.

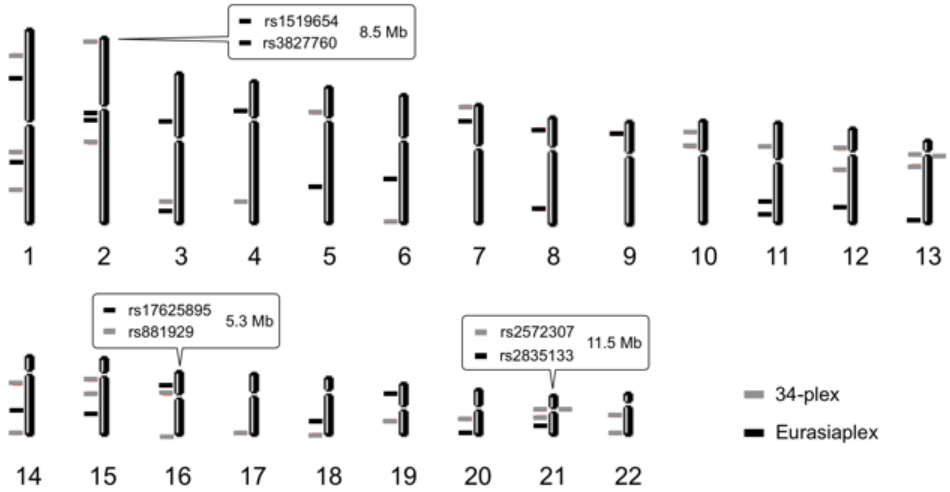


Figure 32: Genomic distribution of 57 SNPs of *Eurasiaplex* and *34-plex* with closest sited pairs indicated.

Table 16: PCR and SNaPshot extension primer details for the *Eurasiaplex* assay.

No.	Internal code	SNP	Forward PCR primer	Reverse PCR primer	Amplicon size	Concentration µM	Extension primer	Extension size	Concentration µM	Detected alleles
1	S6	rs1785864	AGAGTCTTCAGACTTGATTC	GCAGAAAGCCCTGAAAATGG	85	1	tctccTCAGACTTGATTCATAAGACC	26	3	CT
2	S7	rs10962599	GATAGTAAGGCGAACTCCAC	TTCTGTTGTGTGTAGAGGCG	82	0,25	tctctctcccAGGCGAACTCCACATGTGAC	30	0,375	CT
3	S10	rs1519654	TTCTTGGTGAAGTCCACTGC	AGGTCATGATGTGGTTCCTG	86	0,625	tctctctctctcccctACTGCAGCAGTCCCCGA	33	0,5	AG
4	S11	rs2156208	TTGTACACTCCCAGAGCATC	TCCTACCTAGAGTTATAGGC	98	1,55	ctctctctctctctctgTCCTAGGGTTCTCCTATGC	36	3,1	AG
5	S12	rs39897	GGCATCCTCTTCCTGAATTG	ACTCAGCCTCAAAGCCTTAG	78	0,75	tctctctctctctctcggtgATCCTGAGATGACGGTTA	36	3	CT
6	S16	rs2196051	TCCCTCTCAAGTTTGTGAGC	ATTGTTGTAAGCCTGATGGG	115	0,75	ctctctctctctctcccTGAGCAAATATCACTAGAAAGAAGA	42	2,875	CT
7	S15	rs984038	GTAGTGGTATGTCTACCCAG	GTGATTTTCTGGCTCTCCTC	75	0,3125	ctctctctctctctctctcctgctCAGCTGCCTGGAAAACCA	42	0,375	AG
8	S18	rs1363345	TGGGAACTGGAGACTGTTTG	ACCATTGCATGCTAGTCCAC	99	0,5	tctctctctctctctctctcccGGATTGTTTGAAGTTGAAATAGT	45	0,5	CT
9	S17	rs2472304	AACCCTATAGCCAGGAGAAG	ACACAGCAGGCACATAACAG	77	0,375	tctctctctctctctctctctcttcATAGCCAGGAGAAGCCTTGA	45	0,5	AG
10	S20	rs2227203	GAAGCAGAAAGACTCAGTCC	CAGGGCTACAGCTATCATTG	99	0,1875	ctctctctctctctctctctctctctctctcCCACGGCCCCTCTCAGT	48	0,25	AG
11	S22	rs10131666	TCCAGTGGCTTGCAGAAATG	TCTGCAATGCTACATGCACC	78	0,125	ctctctctctctctctctctctctctctctccttccGAAATGGGGTACACTGGCCT	51	0,05	GT
12	S21	rs6026972	TCCAGTACAGTGCAGTACAG	TTGTGTGTGTGGAGAGAAGG	76	1,25	tctctctctctctctctctctctctctctcACAGTCTTTTCTTTTATAATAAAAATC	51	3,1	AC
13	S23	rs734482	AATAGAGCCAGGCACTGTTC	AGCCTCTGCTTCTTCATCTG	118	0,75	ctctctctctctctctctctctctctctctctctctctctctcttcACTGTTCCACGTGCTTT	54	2,5	AG
14	S24	rs9487258	GAACTCATGTTTGAGCAGGG	GGCTCTGGAATTAACAACTG	102	1,0625	tctcctctctctctctctctctctctctctctctctctctctctcCCAAATCAAAGAAGACCTC	54	3,1	CT
15	S25	rs10008492	AGAGCCAGAAAGAGGTTGAG	AATCTGTGCAAGGTTAGCTC	94	0,5	cccaactgactaaactaggtgccacgtcgtgaaagtctgacaaTGAGTCCAGGAATCAAAGA	62	1	CT
16	S26	rs756913	GGCTGCCATGGTTTTCAAAG	TCTTTGACAGAGCTGTCACC	68	0,5	ctctctctctctctctctctctctctctctctctctct

4.1.5.3 Assessment of population divergence and ancestry assignment performance of *Eurasiaplex* and *34-plex* SNPs

Geographically close-sited populations are difficult to assess for clear divergence patterns when they show very close relationships. We used principal component analysis (PCA) to reveal patterns of similarity and divergence amongst 60 populations, progressively decreasing complexity by step-wise reduction of data-points: removing Middle East data and colorizing outlier populations. PCA used *R* 2.11.1 (Team 2011) and the *SNPassoc* package (Gonzalez et al. 2007). HapMap GIH was treated as an un-admixed Indian population as they show no divergence from other Central-South Asian populations. Outlying PCA positions corresponded to geographically or genetically distinct populations. These included: Tuscan, Turkish, Sardinian and Adygei Europeans (the latter two found to be outliers previously (Phillips et al. 2007b)), Hazara and Kalash Central-South Asians (Kalash recognized as atypical in (Rosenberg et al. 2002)), plus Yakut from northeast Siberia.

Ancestry assignment performance of the combined 57 SNPs was measured by constructing differently composed training sets for the *Snipper* online Bayesian classification system (Phillips et al. 2007b). *Snipper* comprises an online training set/profile submission portal generating ancestry likelihoods for single SNP profiles. User-defined SNP sets up to 400 loci plus corresponding reference data (maximum 2000 individuals, 20 populations) can be uploaded as customized training sets to analyse forensic samples of unknown ancestry. This system largely parallels the widely used *STRUCTURE* algorithm, but *Snipper* benefits from an ability to classify single profiles in real time, typifying forensic analyses. We assessed Eurasian ancestry comparisons with two training sets: five group classifications using European, Middle East, Central-South Asian, African and East Asian data, and; four-group classifications excluding Middle East assignment. We chose a central Middle East training set of HGDP-CEPH Druze, Palestinian, Bedouin populations. Training sets and test sets were largely interchangeable, e.g. HGDP-CEPH Pakistani training sets gave similar likelihoods for Indians (Murcia), as Indian training sets for Pakistanis. A North African training set analysing the other Middle East populations gave the greatest contrast in assignment probabilities and success compared to HGDP-CEPH Druze-Palestinian-Bedouin testing North Africans. Therefore all Middle East classifications used the HGDP-CEPH Druze-Palestinian-Bedouin training set. Other training sets selected gave consistently high

assignment probabilities in the widest number of test individuals, consisting of: 1000 Genomes-YRI Africans, 1000 Genomes-CEU Europeans; 1000 Genomes-CHB East Asians and Central-South Asian Indians of Murcia.

A key factor when using *Snipper* is proper assessment of assignment likelihoods. We observed amongst study samples that erroneously assigned individuals rarely gave likelihoods exceeding '100 times more likely'. Therefore, we recorded all assignments below this '100 times more likely' threshold as insufficiently indicative, balancing maximum successful assignments with minimum classification error.

4.1.5.4 AIM-SNP analysis applied to ancestry inference in a challenging identification case

A pelvis and cranium were washed ashore at La Coruña, NW Spain in 2009. We applied routine STR analysis of DNA from a molar to attempt identification but when this failed, cranial morphometric analyses and SNP typing were applied to investigate likely ancestry. Morphometric analysis constructed two-dimensional canonical plots comparing cranial dimensions to data from reference skeletons, mainly museum specimens of US Blacks, US Caucasians, US Hispanics, Native American, Japanese and Chinese – creating a comparative data framework from ~100 individuals (Hardt and Henke 2006, Eiberg et al. 2008). The most reliable morphometric analyses placed the recovered skull dimensions at the far edge of the Japanese female cluster with canonical axes 1 and 5 = 43.3% and 5.3% of variation. Another, less reliable, analysis placed the cranium closest to South Asian clusters.

SNP ancestry analysis used three complementary approaches: joint PCA superimposing the case profile onto full reference population distributions; *STRUCTURE* analysis, and classification using *Snipper* with four- or five-group training sets, as described in section 3.1.4. *STRUCTURE* analyses used v. 2.3.3 with *CLUMPP* and *distruct* software, applying parameters: burnin: 200,000; MCMC steps: 200,000; *admixture*/POPFLAG model with independent allele frequencies; 5 iterations.

4.1.6 Development of a novel forensic STR multiplex for ancestry analysis and extended identity testing (Appendix 2E)

4.1.6.1 Selection of ancestry informative tetranucleotide STRs

The original human population structure study of Rosenberg et al. (2002) used 377 STRs to analyse the CEPH Human Genome Diversity Panel (HGDP-CEPH), sampling 52 populations from seven major global groups of Eurasia (Europe, Middle East, and Central-South Asia), Africa, East Asia, America, and Oceania. We used the published genotype data from Rosenberg extended to 783 loci (Rosenberg et al. 2005), to identify the most informative STRs with tetranucleotide repeats (434/783 STRs). The best ancestry informative markers show the biggest allele frequency differentiation between paired populations and while such genetic distances are straightforward to measure for SNPs using the delta (δ) allele frequency differential (i.e. absolute difference between the allele frequency in one population compared to another) or the Divergence metric: *In* (Rosenberg et al. 2003), estimating equivalent values for STRs is more difficult due to wide ranges of multiple alleles. Therefore, we calculated three genetic distances relevant to STRs for each pairwise population comparison: average squared distance (ASD) and $(\delta\mu)^2$ (Goldstein et al. 1995b, Goldstein et al. 1995a) plus a third value termed "*Aim*" we developed as a supplementary genetic distance providing a cross-check of ASD and $(\delta\mu)^2$. ASD equates genetic distances directly to allele frequency differences across all observed alleles, while $(\delta\mu)^2$ reduces the effect of wide allele ranges by focusing on differences between the average number of alleles. The *Aim* genetic distance we developed calculates the mean of the absolute differences between allele frequencies in the populations compared. In practice all three values correlate very well, but when higher than average differences among distance values were observed, we used $(\delta\mu)^2$ as the defining metric to select the best STRs.

The final selection of STRs to compile the *12-plex* were chosen to balance population differentiation as much as possible, e.g. Americans might show weak differentiation from East Asians so stronger loci for these comparisons were introduced. A second consideration for the final choice of AIM-STRs was genomic distribution, with positions compared to 23 core ID-STRs to give sufficient distance so these new STRs could be applied as

supplementary loci in complex relationship testing. **Figure 33** shows genomic positions of 23 core forensic ID-STRs and the 12 AIM-STRs selected.

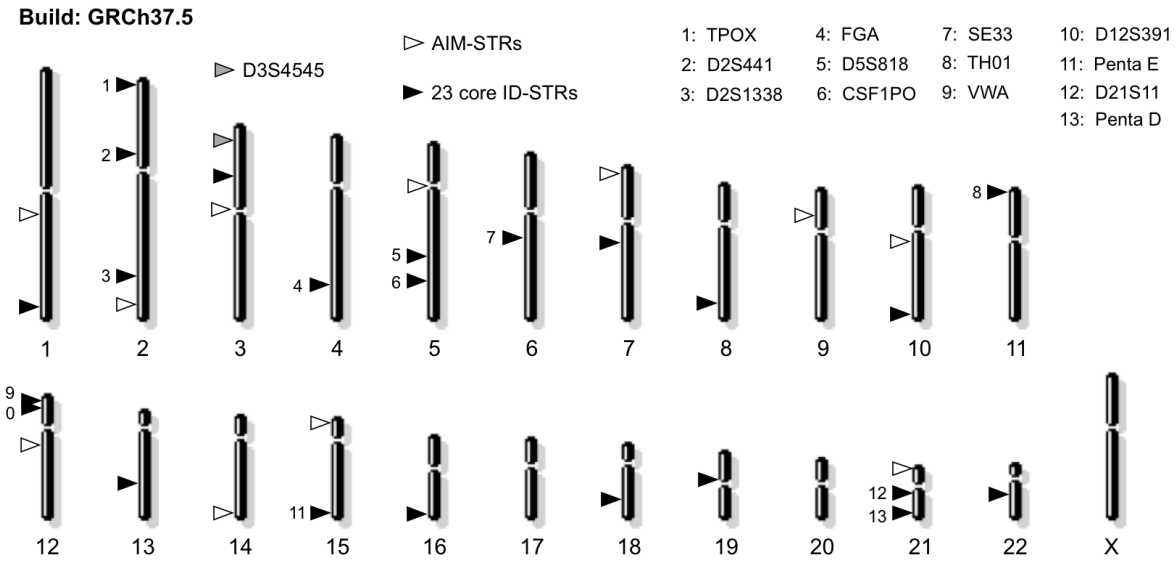


Figure 33: Genomic distribution of 12 AIM-STRs of study and 23 core ID-STRs.

Three STRs performed less well in multiplex development so the final selection comprised 12/15 loci: D1S1679, D2S427, D3S2406, D3S4545, D5S1457, D7S2201, D9S1118, D11S1304, D12S297, D14S1426, D15S822, and D21S1432. Genomic details and primer designs of these loci are outlined in Table 17.

Lastly, since we made use of established ID-STRs in combined-marker ancestry analyses it is instructive to compare the ancestry informativeness of these with the final 12 STRs listed above. Comparing each set's ASD, $(\delta\mu)^2$, *Aim* values and shows an average $(\delta\mu)^2$ value of 0.523 among 20 ID-STRs contrasts with the eightfold greater informativeness of 4.26 among the AIM-STRs. Table 18 lists the individual $(\delta\mu)^2$ values for each population comparison to aid scrutiny of forensic STRs for specific population group comparisons.

Table 17: Genomic and primer details for the 12 component STRs.

AIM-STR	rs-number	GRCh37.p5 location	Range of repeats	Range of sizes	Dye		PCR primers	Tm	Ratio in μM	Sequencing primers
D1S1679	rs112703460	1:162361898–162361981	16–27	171–215	*PET*	F	**GCCATCAAGAAAACTAGTACTGC**	61	6.25	**GCCATCAAGAAAACTAGTACTGC**
						R	CTGACAAAAACTAGGTGCCACGTCGTCAA AGTCTGAACCATGGTACTCAGCAGTGC	60	6.25	GATTAAGCATTCAATGGGGAAG
D2S427	rs111704160	2:232206284–232206436	5–16.3	122–170	*PET*	F	CTGCCTCCCCTTTCAGATTA	58	5.42	GTAGACTGCTAACAGTGTCTG
						R	**GACATTATAAAGCAGTATGAATTTCT**	58	5.42	**GACATTATAAAGCAGTATGAATTTCT**
D3S2406	*rs62242267*[a)]	3:8585036–8585264	27–46	197–269	*NED*	F	GGTACTGCAGTGTTTCTTGAC	60	2.92	CCTATGGGGCTGGATGATC
						R	GGTGCCGGAGAGGCTTTAGAAA AAATTAG	58	2.92	GTCACCATGCAAAAAGTACCAC
D3S4545	rs112283702	3:73258426–73258613	20–37.3	168–244	*VIC*	F	**CTCAGGAGCTTAAGGTTATAGTG**	61	4.17	**CTCAGGAGCTTAAGGTTATAGTG**
						R	ACGTTGGATGACGTTGCCACCCAGGAGAA ACTGTTTTA	60	4.17	TGTCTCTGATGGTCACATCCTC
D5S1457	rs113452942	5:41033858–41033954	8.1–16	82–113	*NED*	F	TCTGTGATGAACAAGGTAGGAG	60	2.08	TGGGTTGGGTATATCTCTCAATG
						R	TTGCTTGGCACACTTCAGGG	60	2.08	GGATTGTTACATGTTGGAAACC
D7S2201	rs111938869	7:5630660–5630751	8–15	80–108	*VIC*	F	GGCAACATAGCAAGATCTCATC	60	2.92	TACAGTGGTTCACACCTGAAATC
						R	CCAAGGCATTTTCTATCGTCTATT	60	2.92	CCAATCTAGAGCAGGACAAAC
D9S1118	rs112557788	9:31925342–31925440	5–16.3	69–115		F	CAGGATATTATGTGATGGAATCC	59	6.25	GAGAATAATACTTACCACAACAGTT
					6-FAM	R	**TCTCTCTCTTTCTCCCTATCTTG**	61	6.25	**TCTCTCTCTTTCTCCCTATCTTG**
D11S1304	rs113308565	11:132586225–132586402	5–19	134–190	*VIC*	F	CAATTGGCATTTGGCTTTTTCAG	59	6.25	TTGCATCTCCTAGAGAAGTGATT
						R	**GCTAAAAGACTGTGGGAACAAA**	58	6.25	**GCTAAAAGACTGTGGGAACAAA**
D12S297	rs113633696	12:52613054–52613283	5–22	202–270	*6-FAM*	F	TTGGTATTGGAGTTTTCAGTGTG	59	6.25	GGAGAAAGGATTTATAAGTATGG
						R	AAATCATCAGTGGAGTTAGCAAAT	58	6.25	GCAACATGGAAATAATGAGACC
D14S1426	rs113351154	14:100619566–100619716	5–17	119–167	*NED*	F	**CTGGGCGACAGTAATGGC**	58	2.5	**CTGGGCGACAGTAATGGC**
						R	**GGGAGAGGCCCTGTATTAGT**	60	2.5	**GGGAGAGGCCCTGTATTAGT**
D15S822	*rs8041848*[a)]	15:27390689–27390838	10–26.3	130–198	*6-FAM*	F	CTCTTTCTCACCTCTCCTTC	58	6.25	TGAGAGTCCATTGAGATTACTTG
						R	**AACAGTCTCAGAGACCTATTAC**	58	6.25	**AACAGTCTCAGAGACCTATTAC**
D21S1432	rs111610289	21:17343484–17343577	7–14	90–118	*PET*	F	CTTAGAGGGACAGAACTAATAG	58	7.5	GTGACATTGTTTTCATGAACATAC
						R	CTCCCTTATCTATCTATCTAATG	58	7.5	AACGTGGAAAGGAGAGATGG

a) rs-number not yet assigned by dbSNP: so the closest SNP site to the STR is given.
Bold sequences denote identical primers for PCR and sequencing analyses.

Table 18: Comparison of $(\delta\mu)^2$ values for nine pairwise population-group differentiations listing 20 standard forensic identification STRs (ID-STRs – upper pane) and the 12 ancestry-informative STRs of this study (AIM-STRs – lower pane) in ascending order of ancestry informativeness (i.e. average of nine pairwise differentiations, weakest STR at the top).

STR	AFR-EUR	AFR-ASN	AFR-AME	AFR-OCE	EUR-ASN	EUR-AME	EUR-OCE	ASN-AME	ASN-OCE	AME-OCE
ID-STRs										
vWA	0.008	0.005	0.091	0.118	0.000	0.153	0.187	0.140	0.173	0.002
D3S1358	0.176	0.040	0.030	0.216	0.049	0.353	0.002	0.139	0.071	0.408
D21S11	0.358	0.056	0.047	0.032	0.131	0.664	0.177	0.205	0.003	0.155
D1S1679	0.217	0.007	0.051	0.143	0.146	0.477	0.712	0.095	0.213	0.023
TPOX	0.033	0.100	0.419	0.712	0.018	0.216	0.438	0.109	0.278	0.039
D19S433	0.690	0.745	0.408	0.487	0.001	0.037	0.018	0.051	0.027	0.004
D12S391	1.111	0.475	0.590	0.401	0.133	0.082	0.177	0.006	0.003	0.018
D2S1338	0.141	0.317	0.194	0.020	0.882	0.666	0.056	0.015	0.494	0.336
CSF1PO	0.321	0.618	1.190	0.621	0.048	0.275	0.049	0.093	0.000	0.092
D22S1045	0.166	0.000	0.653	0.750	0.158	0.161	0.211	0.638	0.734	0.003
D16S539	0.460	0.003	0.022	0.766	0.540	0.684	0.039	0.008	0.869	1.049
TH01	0.003	0.035	1.108	0.149	0.017	1.230	0.195	1.539	0.329	0.445
D8S1179	1.293	0.717	1.243	0.034	0.084	0.000	0.908	0.072	0.440	0.867
D7S820	0.047	0.506	1.970	0.482	0.244	1.407	0.228	0.479	0.000	0.503
D10S1248	0.206	0.020	0.012	1.774	0.097	0.118	0.771	0.001	1.416	1.493
D5S818	0.020	0.332	2.019	0.593	0.190	1.641	0.397	0.714	0.038	0.424
FGA	1.771	0.561	0.002	0.031	0.338	1.891	1.334	0.629	0.329	0.048
D1S1656	0.305	0.470	2.717	0.893	0.018	1.201	0.154	0.927	0.067	0.495
D13S317	0.445	2.749	1.957	2.253	0.982	0.536	0.696	0.067	0.025	0.010
D18S51	3.283	1.547	2.114	0.029	0.323	0.128	2.693	0.044	1.151	1.646
D2S441	0.006	0.532	2.668	3.769	0.425	2.421	3.474	0.817	1.469	0.095
Avg. 20 ID-STRs	0.527	0.468	0.929	0.680	0.230	0.683	0.615	0.323	0.387	0.388
AIM-STRs										
D7S2201	0.247	0.318	0.220	0.203	0.004	0.001	0.002	0.009	0.013	0.000
D21S102	0.082	0.016	0.790	0.126	0.171	0.364	0.005	1.032	0.232	0.285
D5S1457	1.050	0.002	2.196	0.242	1.140	0.209	0.284	2.325	0.286	0.980
D9S1118	0.027	0.003	0.239	**7.518**	0.049	0.426	**8.443**	0.185	**7.201**	**5.076**
D3S2406	1.410	**3.653**	4.748	0.743	0.524	0.983	4.200	0.072	**7.691**	**9.248**
D15S822	**4.221**	0.000	2.283	0.440	**4.301**	**12.714**	1.935	2.225	0.466	**4.729**
D3S4545	0.174	**6.583**	**8.181**	0.720	**4.617**	**5.969**	0.186	0.087	2.948	**4.046**
D14S1426	0.538	0.665	**9.276**	0.875	0.007	**5.346**	2.786	**4.975**	3.065	**15.850**
D11S1304	**2.983**	0.922	**8.248**	**9.425**	**7.220**	1.311	1.804	**14.684**	**16.241**	0.039
D2S427	**14.898**	**16.648**	**11.705**	**20.066**	0.049	0.192	0.384	0.434	0.159	1.120
D12S297	1.100	**9.000**	**31.092**	**41.319**	**3.806**	**20.494**	**28.934**	**6.636**	**11.751**	0.726
Avg. 12 AIM-STRs	2.430	3.437	7.180	7.425	1.990	4.364	4.451	2.970	4.550	3.827

AIM-STR values in bold show above-average AIM-STR values for that group comparison indicating the strongest population-group differentiations. Gray cells show values above the average power for ID-STRs indicating that each AIM-STR is more informative for ancestry than ID-STRs except D7S2201.

4.1.6.2 Multiplex design, amplification, and CE detection of the AIM-STRs

PCR primer designs used Primer3 software (http://bioinfo.ut.ee/primer3/), with criteria: preferred amplicon size range 60–160 bp; optimum Tm 59ºC (minimum 58ºC); optimum GC content 40% (minimum 35%). Primer pairs were checked for nonspecific hybridization elsewhere in the genome using BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) and hairpin/primer–dimer secondary structures checked using AutoDimer (Vallone and Butler 2004). Expected size ranges were balanced for three STRs per fluorochromes: 6-FAM (blue), VIC (green), NED (yellow), and PET (red).

Multiplex amplifications comprised 12 µL reactions consisting of: 2 µL DNA (~0.4 ng/µL) with 10 µL master mix combining: 1 µL 10 × PCR buffer II (AB); 0.8 µL $MgCl_2$ (3.64 mM); 0.25 µL dNTPs (0.23 mM); 0.5 µL BSA (0.07 µg/µL); 0.2 µL AB AmpliTaq Gold; 5 µL Qiagen Multiplex PCR Mix; 1.82 µL PCR primer mix. Primer sequences and ratios are listed in Table 17. Amplifications used AB GeneAmp® 2720 thermocyclers with cycling conditions: 95ºC activation for 15 mins; then 28-30 cycles: 96ºC denaturation for 30 s, 59ºC annealing for 90 s; 72ºC extension for 90 s, and 72ºC final extension for 60 mins.

CE preparation combined 1 µL of amplified product with 9.5 µL of a 33:1 mixture of deionized Hi-Di™ formamide and GS-500 LIZ size standard, respectively (both AB). CE detection used a PRISM 3130 Genetic Analyzer (AB), filter set G5, POP-7™ polymer (AB), and analysis with GeneMapper v4.0 (AB).

4.1.6.3 Sequence analysis of example alleles and construction of allelic ladder

Additional unlabelled primers were designed for sequence analysis of each STR interrogating segments beyond the PCR primer-binding sites to track low frequency SNPs/Indels within these regions while characterizing the repeat structures of common alleles to compile the allelic ladders. Sequencing primer details are shown in Table 17 and allele-specific amplifications used singleplex PCRs (conditions outlined in Table 19). PCR products were fractionated in silver stained polyacrylamide gels (T: 9%, C: 5%) and relevant alleles excised from bands in the gel and re-amplified. Homozygote bands were added directly into sequencing reactions whereas heterozygotes required additional re-amplification due to reduced recovery. Alleles were eluted by incubating gel segments in 50 mL of 20% Chelex™ overnight followed by three freeze-thaw cycles then centrifugation for 4 mins at 13,000 rpm. The resulting purified supernatant was re-amplified and sequenced using standard AB Big-dye™ protocols. All STR repeat numbers were ascribed by recording consensus GATA repeat units or their variants.

Allele reference ladders with four-nucleotide or, where appropriate, intermediate repeat intervals, were constructed combining singleplex amplified alleles, mainly heterozygotes. Prototype ladders were electrophoresed, relative peak heights recorded, and readjusted to reduce imbalance then each ladder was diluted 1000-fold and re-amplified using 35-cycle

PCR. Ladders were combined for each dye with additional rebalancing among dyes before a final combined PCR at 1000-fold dilution was made.

Table 19: Sequencing PCR conditions for each STR.

		D9S1118			D15S822			D12S297			D7S2201			D11S1304			D3S4545		
Mix 1st Amplification	Buffer	1X			1X			1X			1X			1X			1X		
	MgCl2	3.64 mM			3.64 mM			3.64 mM			0.8 mM			3.64 mM			3.81 mM		
	Primer (F & R)	1.82 µM			1.82 µM			1.82 µM			2 µM			1.82 µM			1.9 µM		
	dNTPs	0.23 mM			0.23 mM			0.23 mM			0.25 mM			0.23 mM			0.24 mM		
	BSA	0.07 µg/µL			0.07 µg/µL			0.07 µg/µL			0.8 µg/µL			0.07 µg/µL			0.08 µg/µL		
	Taq Gold	1 U			1 U			1 U			1 U			1 U			1 U		
	H20	4.25 µL			4.25 µL			4.25 µL			4.25 µL			4.25 µL			4.25 µL		
	DNA	1.20 ng/µL			1.20 ng/µL			1.20 ng/µL			0.04 ng/µL			1.20 ng/µL			0.86 ng/µL		
Cycles 1st Amplification	Activation	95ºC	15'		95ºC	15'		95ºC	15'		95ºC	15'		95ºC	15'		95ºC	15'	
	Denaturation	94ºC	30"		94ºC	30"		94ºC	30"		94ºC	30"		94ºC	30"		94ºC	30"	
	Annealing	**59ºC**	1'30"	**35 cycles**	**59ºC**	1'30"	**35 cycles**	**59ºC**	1'30"	**35 cycles**	**60ºC**	1'30"	**33 cycles**	**59ºC**	1'30"	**35 cycles**	**59ºC**	1'30"	**34 cycles**
	Elongation	72ºC	1'30"		72ºC	1'30"		72ºC	1'30"		72ºC	1'30"		72ºC	1'30"		72ºC	1'30"	
	Final extension	72ºC	60'		72ºC	60'		72ºC	60'		72ºC	30'		72ºC	30'		72ºC	60'	
	Soak	4ºC			4ºC			4ºC			4ºC			4ºC			4ºC		
Mix 2nd Amplification	Buffer	1X			1X			1X			1X			1X			1X		
	MgCl2	4 mM			4.44 mM			4 mM			4 mM			4 mM			4.44 mM		
	Primer (F & R)	2 µM			1.11 µM			2 µM			2 µM			2 µM			1.11 µM		
	dNTPs	0.25 mM			0.28 mM			0.25 mM			0.25 mM			0.25 mM			0.28 mM		
	BSA	0.08 µg/µL			0.5 µg/µL			0.08 µg/µL			0.08 µg/µL			0.08 µg/µL			0.5 µg/µL		
	Taq Gold	1 U			1 U			1 U			1 U			1 U			1 U		
	H20	4.25 µL			4.25 µL			4.25 µL			4.25 µL			4.25 µL			4.25 µL		
	PCR Product	1 µL			1 µL			1 µL			1 µL			2µL			0.5µL		
Cycles 2nd Amplification	Activation	95ºC	15'		95ºC	15'		95ºC	15'		95ºC	15'		95ºC	15'		95ºC	15'	
	Denaturation	94ºC	30"		94ºC	30"		94ºC	30"		94ºC	30"		94ºC	30"		94ºC	30"	
	Annealing	**59ºC**	1'30"	**34 cycles**	**60ºC**	1'30"	**34 cycles**	**59ºC**	1'30"	**34 cycles**	**60ºC**	1'30"	**34 cycles**	**59ºC**	1'30"	**34 cycles**	**59ºC**	1'30"	**34 cycles**
	Elongation	72ºC	1'30"		72ºC	1'30"		72ºC	1'30"		72ºC	1'30"		72ºC	1'30"		72ºC	1'30"	
	Final extension	72ºC	30'		72ºC	30'		72ºC	30'		72ºC	30'		72ºC	30'		72ºC	30'	
	Soak	4ºC			4ºC			4ºC			4ºC			4ºC			4ºC		
Mix Sequencing	Big Dye	0.5 µL			0.5 µL			0.5 µL			0.5 µL			0.5 µL			0.5 µL		
	Big Dye Buffer	2.5 µL			2.5 µL			2.5 µL			2.5 µL			2.5 µL			2.5 µL		
	H20	5.5 µL			5.5 µL			5.5 µL			5.5 µL			5.5 µL			5.5 µL		
	Primer	Forward = 1 µL			Forward = 1 µL			Forward = 1 µL			Forward = 1 µL			Reverse = 1 µL			Reverse = 1 µL		
	PCR Product	3 µL			3 µL			3 µL			3 µL			3 µL			3 µL		
Cycles Sequencing	Activation	96ºC	3'		96ºC	3'		96ºC	3'		96ºC	3'		96ºC	3'		96ºC	3'	
	Denaturation	96ºC	30"		96ºC	30"		96ºC	30"		96ºC	30"		96ºC	30"		96ºC	30"	
	Annealing	**50ºC**	15"	**27 cycles**	**50ºC**	15"	**27 cycles**	**50ºC**	15"	**27 cycles**	**50ºC**	15"	**27 cycles**	**50ºC**	15"	**27 cycles**	**50ºC**	15"	**27 cycles**
	Elongation	60ºC	4'		60ºC	4'		60ºC	4'		60ºC	4'		60ºC	4'		60ºC	4'	
	Soak	4ºC			4ºC			4ºC			4ºC			4ºC			4ºC		

		D5S1457			D14S1426			D3S2406			D21S1432			D2S427			D1S1679		
Mix 1st Amplification	Buffer	1X			1X			1X			1X			1X			1X		
	MgCl2	4 mM			3.64 mM			3.81 mM			3.48 mM			3.64 mM			3.48 mM		
	Primer (F & R)	2 µM			1.82 µM			1.9 µM			1.74 µM			1.82 µM			1.74 µM		
	dNTPs	0.25 mM			0.23 mM			0.24 mM			0.22 mM			0.23 mM			0.22 mM		
	BSA	0.5 µg/µL			0.07 µg/µL			0.08 µg/µL			0.07 µg/µL			0.07 µg/µL			0.07 µg/µL		
	Taq Gold	1 U			1 U			1 U			1 U			1 U			1 U		
	H20	4.25 µL			4.25 µL			4.25 µL			4.25 µL			4.25 µL			4.25 µL		
	DNA	0.6 ng/µL			1.20 ng/µL			0.86 ng/µL			0.08 ng/µL			1.20 ng/µL			0.08 ng/µL		
Cycles 1st Amplification	Activation	95ºC	15'		95ºC	15'		95ºC	15'		95ºC	15'		95ºC	15'		95ºC	15'	
	Denaturation	94ºC	30"		94ºC	30"		94ºC	30"		94ºC	30"		94ºC	30"		94ºC	30"	
	Annealing	**60ºC**	1'30"	**35 cycles**	**59ºC**	1'30"	**35 cycles**	**59ºC**	1'30"	**34 cycles**	**60ºC**	1'30"	**35 cycles**	**59ºC**	1'30"	**35 cycles**	**60ºC**	1'30"	**35 cycles**
	Elongation	72ºC	1'30"		72ºC	1'30"		72ºC	1'30"		72ºC	1'30"		72ºC	1'30"		72ºC	1'30"	
	Final extension	72ºC	30'		72ºC	60'		72ºC	60'		72ºC	30'		72ºC	30'		72ºC	30'	
	Soak	4ºC			4ºC			4ºC			4ºC			4ºC			4ºC		
Mix 2nd Amplification	Buffer	1X			1X			1X			1X			1X			1X		
	MgCl2	4 mM			4.55 mM			4 mM			4 mM			4 mM			4 mM		
	Primer (F & R)	2 µM			0.57 µM			2 µM			2 µM			2 µM			2 µM		
	dNTPs	0.25 mM			0.28 mM			0.25 mM			0.25 mM			0.25 mM			0.25 mM		
	BSA	0.08 µg/µL			0.09 µg/µL			0.08 µg/µL			0.08 µg/µL			0.08 µg/µL			0.08 µg/µL		
	Taq Gold	1 U			1 U			1 U			1 U			1 U			1 U		
	H20	4.25 µL			4.5 5 µL			4.25 µL			4.25 µL			4.25 µL			4.25 µL		
	PCR Product	1 µL			0.3µL			1 µL			1 µL			1 µL			1 µL		
Cycles 2nd Amplification	Activation	95ºC	15'		95ºC	15'		95ºC	15'		95ºC	15'		95ºC	15'		95ºC	15'	
	Denaturation	94ºC	30"		94ºC	30"		94ºC	30"		94ºC	30"		94ºC	30"		94ºC	30"	
	Annealing	**60ºC**	1'30"	**34 cycles**	**61ºC**	1'30"	**31 cycles**	**59ºC**	1'30"	**34 cycles**	**60ºC**	1'30"	**34 cycles**	**59ºC**	1'30"	**34 cycles**	**60ºC**	1'30"	**34 cycles**
	Elongation	72ºC	1'30"		72ºC	1'30"		72ºC	1'30"		72ºC	1'30"		72ºC	1'30"		72ºC	1'30"	
	Final extension	72ºC	30'		72ºC	30'		72ºC	30'		72ºC	30'		72ºC	30'		72ºC	30'	
	Soak	4ºC			4ºC			4ºC			4ºC			4ºC			4ºC		
Mix Sequencing	Big Dye	0.5 µL			0.5 µL			0.5 µL			0.5 µL			0.5 µL			0.5 µL		
	Big Dye Buffer	2.5 µL			2.5 µL			2.5 µL			2.5 µL			2.5 µL			2.5 µL		
	H20	5.5 µL			5.5 µL			5.5 µL			5.5 µL			5.5 µL			5.5 µL		
	Primer	Forward = 1 µL			Forward = 1 µL			Reverse = 1 µL			Forward = 1 µL			Forward = 1 µL			Reverse = 1 µL		
	PCR Product	3 µL			3 µL			3 µL			3 µL			3 µL			3 µL		
Cycles Sequencing	Activation	96ºC	3'		96ºC	3'		96ºC	3'		96ºC	3'		96ºC	3'		96ºC	3'	
	Denaturation	96ºC	30"		96ºC	30"		96ºC	30"		96ºC	30"		96ºC	30"		96ºC	30"	
	Annealing	**50ºC**	15"	**27 cycles**	**50ºC**	15"	**27 cycles**	**50ºC**	15"	**27 cycles**	**50ºC**	15"	**27 cycles**	**50ºC**	15"	**27 cycles**	**50ºC**	15"	**27 cycles**
	Elongation	60ºC	4'		60ºC	4'		60ºC	4'		60ºC	4'		60ºC	4'		60ºC	4'	
	Soak	4ºC			4ºC			4ºC			4ºC			4ºC			4ºC		

4.1.6.4 Bayesian analysis of STR data for inference of ancestry

We previously developed a Bayesian ancestry analysis system: *Snipper*, designed for forensic AIM-SNPs. Ancestry assignments of SNP genotype profiles are derived from the

ratio of the two highest likelihoods from multiple population comparisons (Phillips et al. 2007b) with likelihoods calculated from allele frequencies estimated using training sets of HGDP-CEPH reference populations. We adapted *Snipper* to handle the multiple alleles of STRs by accepting training sets of user-input allele frequencies rather than genotypes. An Excel-based data input system was developed using one worksheet per marker (alleles as columns, populations as rows) listing allele frequencies in each cell. This approach simplifies training set preparation, allowing published allele frequencies to be input directly. The *Snipper* portal (http://mathgene.usc.es/snipper/) gives a direct link to the allele frequency analysis option and provides two preset STR frequency data files with HGDP-CEPH data for the 12 AIM-STRs described here plus 20 core ID-STRs in routine use. Both files list data for seven population group analysis (i.e. Eurasians divided into Europe, Middle East, and Central-South Asia), but files are adjustable to five (e.g. retaining Europeans only) by removal of rows in each worksheet. We reclassified the HGDP-CEPH panel using *Snipper* to test the accuracy of this inference system with the 12 AIM-STRs and compared likelihoods with those previously obtained for core ID-STRs always typed in forensic cases and an established *34-plex* AIM-SNP assay (Phillips et al. 2007b).

Ancestry inference of single profiles with *Snipper* closely parallels the Bayesian whole-population analysis system of *STRUCTURE*. *STRUCTURE* assigns individuals to genetic clusters by comparing the extent the allele frequency distributions are shared within a population. Genetic clusters are identified from *K* predefined populations inferred from the data and posterior probabilities of *K* estimated from: InP(X>K) where X = genotypes of analysed individuals. From an optimum number of *K* clusters, each individual has their percentage proportions of membership to each cluster assigned, which are then commonly used to infer ancestry by analysing unknown individuals alongside reference data and assigning them to the cluster with the largest membership proportion in each sample. We used Structure v.2.3.3 (Pritchard et al. 2000) to analyse the HGDP-CEPH panel with 12 AIM-STRs, 12 AIM-STRs plus 20 ID-STRs, then 32 STRs plus 34 AIM-SNPs. Separate analysis of 21 HGDP-CEPH Eurasian populations was made to examine how efficiently 32 STRs and 34 SNPs could differentiate these closely related subpopulations.

4.1.6.5 Expanding *pop.STR*: An open online database of forensic STR population data

Proper compilation of population frequency data relevant to a laboratory's geographic region is an important step toward constructing accurate STR profile frequencies. The open-access STR frequency database *pop.STR* (Amigo et al. 2009) has built a data resource for this purpose with global geographic scope and lists HGDP-CEPH population data for 20 core ID-STRs. The *pop.STR* portal (http://spsmart.cesga.es/popstr.php) allows combination of populations into groups and selection of appropriate ID-STR multiplex sets. Output includes a graphical display of frequencies and downloadable population differentiation values (F_S/F_{ST} metrics) plus allele frequency tables of selected populations. We added the 12 AIM-STR HGDP-CEPH data to *pop.STR* to allow scrutiny of population variability in these new forensic markers and to help users compile training sets for Snipper described in section 4.1.6.4.

4.1.6.6 Forensic statistics

Forensic informativeness parameters including: discrimination power, random match probability (RMP), typical paternity index, and average power of exclusion were calculated using Powerstats (http://www.promega.com/geneticidtools/powerstats/). Forensic metrics were estimated for AIM-STRs then compared to 20 ID-STRs from a previous study of the same HGDP-CEPH populations (Phillips et al. 2011).

4.1.6.7 Genotyping of forensic standard reference material controls

To test the multiplex performance of the *12-plex* assay, we typed four components, A-D, of a Standard Reference Material® control DNA panel: SRM-2391c from NIST. SRM-2391c components A, B, and C are Caucasian, Mexican Hispanic, and Melanesian individuals, respectively, while component D is a 3:1 mixture of A and C. As well as providing reference genotypes to aid use of the 12 STRs, SRM-2391c component D allowed initial assessment of their ability to detect and deconvolute a mixed DNA sample. Forensic sensitivity of the *12-plex* was assessed using doubling dilution series of standard STR kit control DNAs at: 1 ng,

500 picograms (pg), 250 pg, 125 pg, and 62.5 pg, plus genotyping of challenging DNA obtained from degraded skeletal material. Peak balance across the 12 component loci was measured by typing 96 paternity analysis samples from NW Spain.

4.2 RESULTS AND DISCUSSION

4.2.1 Revision of the SNPforID *34-plex* forensic ancestry test (Appendix 2A)

4.2.1.1 Performance of the modified *34-plex* SNaPshot assay: SBE product mobility variation and peak height ratios

An example *34-plex* electropherogram is shown in **Figure 34**. This profile was chosen due to an above-average heterozygosity of 72% with 49 peaks from a total 68, so it illustrates the improved heterozygote balance we observed in comparison to patterns obtained from the previous assay conditions (see Figure 1 of Phillips et al. (2007b)). The migration positions of SBE products not present in this example profile are marked with the SNP and allele code in grey – note three positions are shown for tri-allelic SNPs P24 and P27.

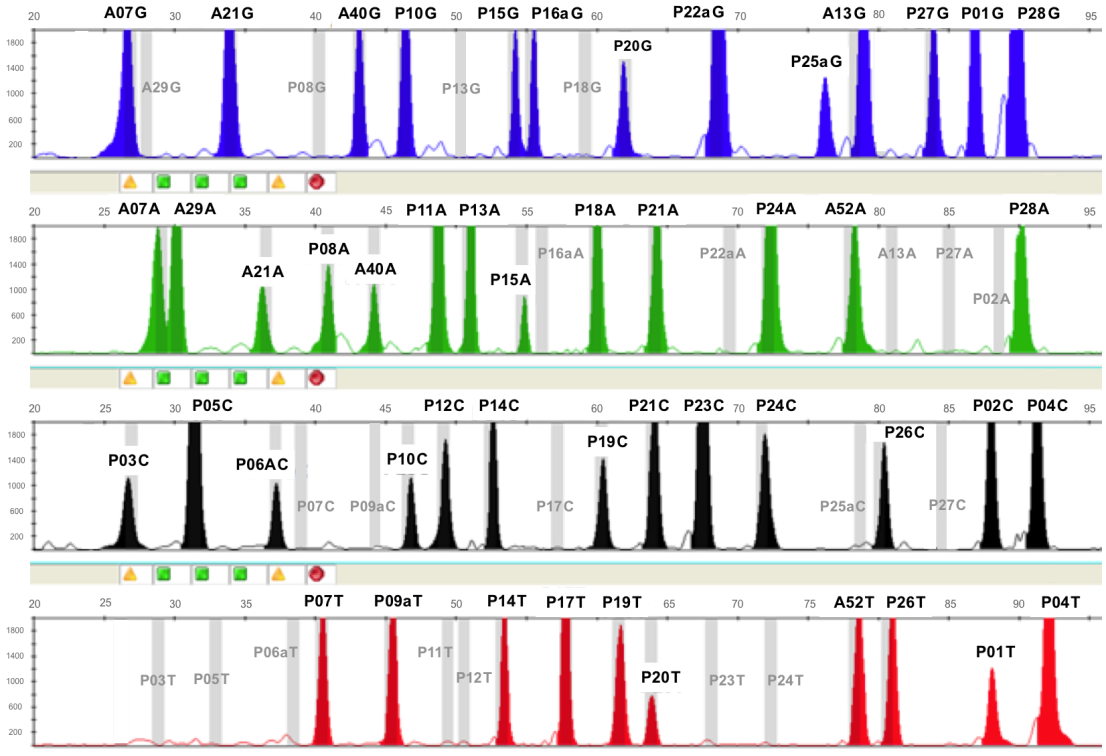


Figure 34: Example electropherogram obtained from the revised *34-plex* assay. SNP allele peaks marked with internal codes listed in Table 8.

Unlike forensic STR typing, profiles generated by SNaPshot tend to give a range of non-specific peaks that are often high enough to mimic product peaks, typically in the green dye channel. Our experiences with the previous *34-plex* assay profiles consistently showed that non-specific signals fall outside the established allelic peak positions and can therefore be confidently excluded as allele extension products. This makes the recording of SBE product mobility variation an important step when validating forensic SNaPshot assays. Therefore we used the genotyping of 712 NIST reference samples to optimise the reconfigured assay and, with reference to multiple runs per sample, plotting all allele mobilities and peak height ratios. Results for these extended electrophoresis analyses are summarised in Table 9 and Table 20. Mean base pair size estimates for each SNP allele were made using standardised conditions of POP-4^{TM} in 36 cm capillary arrays with an AB 3130*xl* Genetic Analyser. The observed allele sizes with the above CE conditions are shown in Table 9. Allele size estimates were found to be very stable across multiple samples, runs, and reactions. Size values for all 34 SNPs in the modified assay gave an average standard deviation across 70 positions of ±0.06 bp and values predominantly equal to or less than 0.1 bp (highest ±0.14 bp), but for most SNPs this was much smaller. Therefore any signal peaks falling outside the highly conservative ±0.5bp size variation window routinely applied to allelic peak positions can be securely designated as artefactual.

Table 20: Mean heterozygote peak height ratios with SNPs in the same order as Fig3.

Marker	Mean	Standard deviation	Marker	Mean	Standard deviation	Marker	Mean	Standard deviation	Marker	Mean	Standard deviation
T/C heterozygotes			G/A heterozygotes			G/T heterozygotes			A/T heterozygotes		
P24-CT	1.2092	0.1692	P27-AG	1.6372	0.1701	P20	2.0731	0.5997	P24-AT	1.7632	0.4127
P04	0.9089	0.1153	P28	5.0806	1.4789	P01	2.3225	0.0398	A52	0.9887	0.1182
P26	1.6778	0.1455	A13	1.4742	0.2727	Total GT	2.1978	0.5183	P11	1.1247	0.1829
P22a	1.2563	0.1349	P23	2.0429	0.5083				Total AT	1.2922	0.2380
P19	1.5359	0.2772	P18	1.7116	0.5266	A/C heterozygotes					
P17	2.1853	0.3817	P16	1.4132	0.3034	P27-AC	2.0240	0.3388	C/G heterozygotes		
P14	1.2118	0.1916	P15	2.3777	0.2960	P24-AC	2.1612	0.5010	P27-CG	3.5695	0.7720
P12	1.0979	0.2036	P13	1.0978	0.2317	P21	1.1858	0.1360	A25a	3.1853	0.6303
P09a	1.2072	0.1850	A40	1.5639	0.3181	P02	1.7830	0.2528	F2	3.0982	0.42424
P07	1.1672	0.1331	P08	4.8865	0.6343	Total AC	1.7885	0.3071	Total CG	3.3215	0.6089
P06a	1.3641	0.1858	A21	3.5275	1.0584						
P05	1.1904	0.3136	A07	2.3412	0.9151						
P03	1.3096	0.2822	A29	1.1750	0.0912						
Total CT	1.3324	0.2091	Total AG	2.3330	0.5234						

Heterozygote imbalance has a more disruptive effect on SNaPshot genotyping consistency than occurrence of non-specific peaks. Large disparities in reporter dye signal strength, varying primer extension efficiencies between the four terminator dideoxynucleotides and stochastic PCR effects can all impact on imbalanced signals from SBE products. While unavoidable, SNaPshot heterozygote imbalance can usually be

accommodated by assigning pre-set peak height ratio (PHR) limits. Previously published guidelines [20,21] suggest that for equal amounts of dye molecules passing the detection window, recorded RFU peak heights for a G-terminating nucleotide (DR110 dye) are generally two times that of a corresponding A (DR6G) and 4 times that of corresponding C (TAMRA) or T (ROX). To explore the widest range of heterozygote peak pairs we collected RFU values from 20 heterozygotes in 80 amplified samples of European, African American, Hispanic and East Asian origin, for each of the 34 SNPs, ensuring many more heterozygotes could be analysed than normally possible in any one population group with these SNPs. The distribution of peak height ratios from these samples is shown in **Figure 35** and the mean values for each PHR range are listed in Table 20.

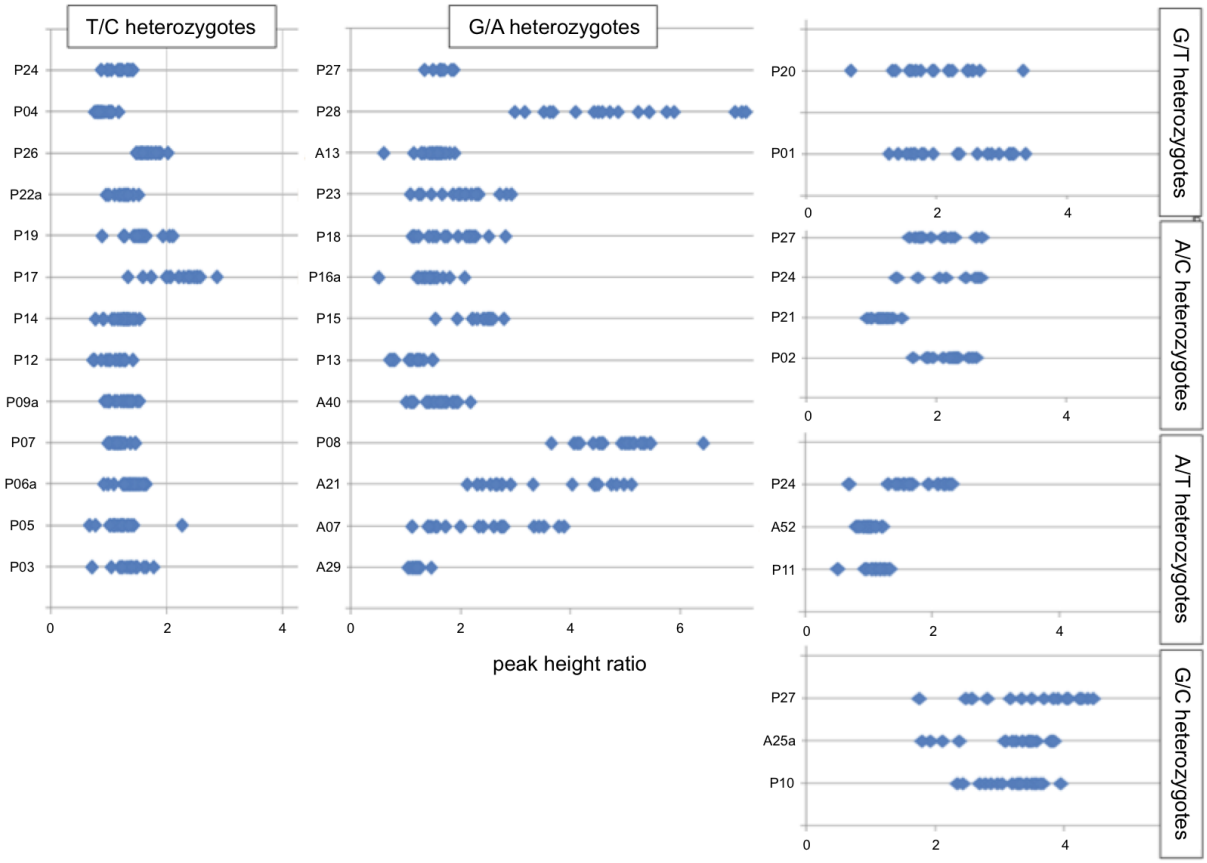


Figure 35: Distribution of peak height ratios recorded in the heterozygotes of 80 NIST population panel samples.

Amplifications were performed with three separate PCR reactions but this was not observed to have a significant effect on peak height ratios (data not shown), indicating stochastic PCR effects represent a minor source of signal variation. CT heterozygotes showed a mean T/C PHR value of 1.33 (i.e. higher T peak) with a standard deviation of 0.21 and SNP

P17 a single outlier with a mean T/C = 2.19. AC and AT heterozygotes showed similar minor variation around mean PHRs of A/C = 1.79 (±0.31) and A/T = 1.29 (±0.23) respectively. The DR110-linked G allele shows the highest deviations from predicted PHR values particularly in AG heterozygotes and it is worth noting the DR110 dye displays the highest energy emission of all the four dyes used in SNaPshot. Some AG SNPs display PHRs close to 1 with little variation, particularly SNPs A07 and A29, but these contrast with the new SNP P28 as well as A21 and P08 that have high between-run variation (up to 0.5) and PHR values greater than predicted. GT peak pairs were the most divergent from the predicted PHR of 4 being closer to a value of 2, likely due to the higher emission energy from the ROX dye of T compared to the TAMRA of C. The standard deviation from average peak height in all G-based heterozygotes was greater than 0.5: more than twice the value of the other heterozygotes and evident in the broader spread of G peak ratios in **Figure 35**. Therefore average PHRs collated in our study for each SNP and summarised in Table 9, should be kept in mind when interpreting *34-plex* SNaPshot profiles, especially those generated from difficult or scant DNA sources where peaks are likely to be much lower than normal and show greater stochastic variation.

4.2.1.2 Population analyses

Structure analysis of HGDP-CEPH reference samples alone using both original and revised *34-plex* SNP sets are presented in **Figure 36A**. The optimum *K* value corresponds to five population clusters in both cases, as shown by the L(*K*) mean value plots in **Figure 36B**. Average ancestry membership proportions (from the HGDP-CEPH analysis) are listed in Table 21 and these show an increase in the values associated with the East Asian ancestry component using the new SNP combination, seen as a median value of 91.16% raised from a previous 87.64%. The revised SNP set creates rather cleaner *STRUCTURE* cluster plots for East Asians at *K*=5 in HGDP-CEPH samples shown in **Figure 36A** and to a lesser extent in the CHB, CHS, JPT samples of the combined HGDP-CEPH-1000 Genomes analyses described below and shown in **Figure 36C**.

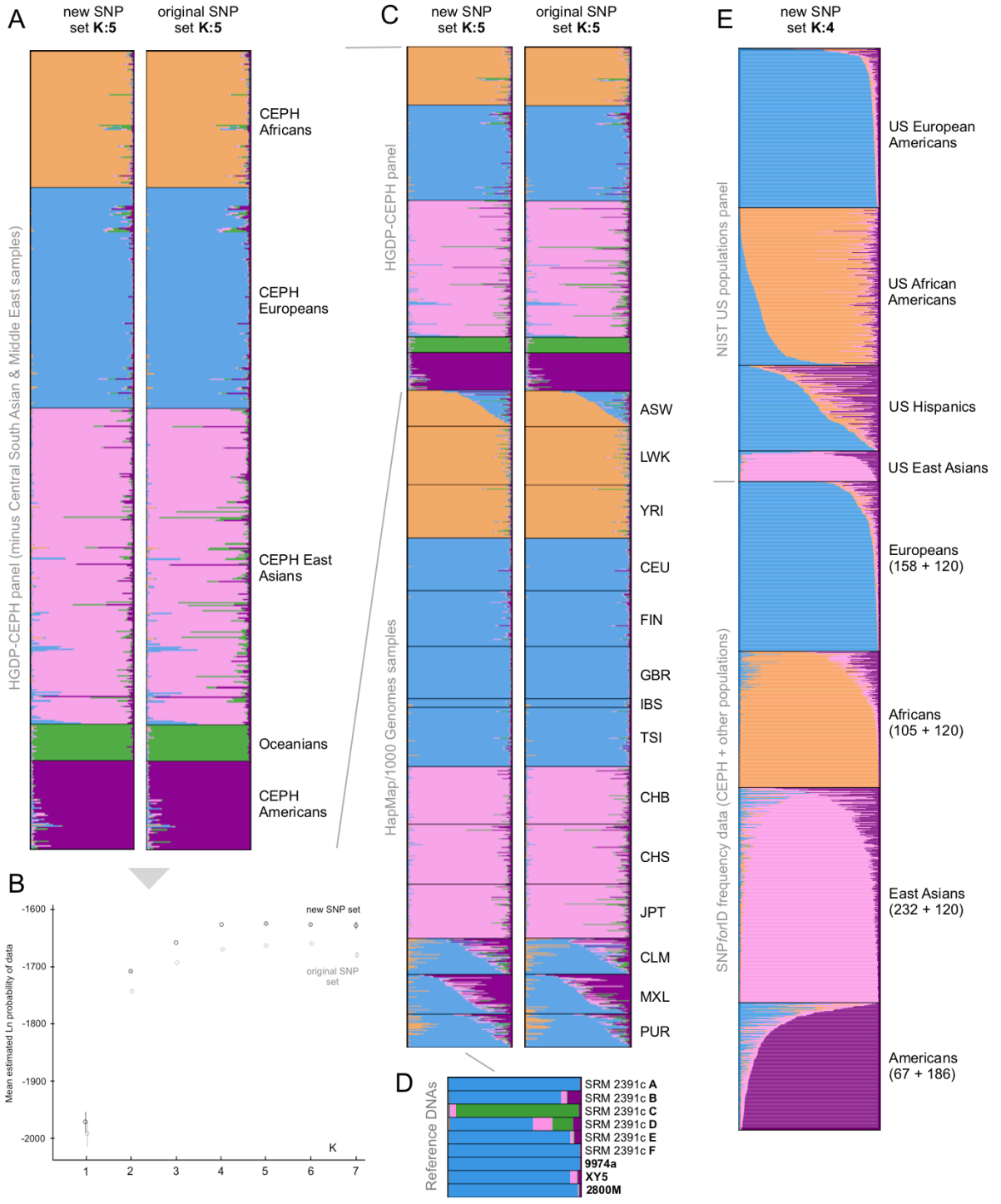


Figure 36: Cluster plots summarising five *STRUCTURE* analyses performed on combinations of reference (HGDP-CEPH plus SNPforID populations) and unknown populations (NIST SRM 2391/control DNAs, 1000 Genomes populations and NIST U.S. panel). *STRUCTURE* runs comprised, (A) HGDP-CEPH alone, old and new SNP sets, with the optimum *K* value of 5 indicated in by the likelihood L(*K*) plot of (B and C) HGDP-CEPH with unknown 1000 Genomes and SRM 2391c/control DNAs (expanded in D), (E) HGDP-CEPH plus SNP*for*ID with unknown NIST U.S. population panel and new SNPs only giving an optimum *K*=4. 1000 Genomes population descriptors as in **Figure 29**.

Analysing 1000 Genomes individuals as study samples of 'unknown' ancestry (lower half of **Figure 36C**) provides results assigning all samples to their expected population clusters. Four populations are arranged in a sequence to indicate rising second admixture component proportions showing the range as a gradient of two-cluster memberships in the following 1000 Genomes populations: ASW (African Americans from Southwest U.S.); CLM (Colombians from Medellín); MXL (Mexicans from LA district, CA); and PUR (Puerto Ricans). Individual sample membership proportions in K=5 clusters for both original and new *34-plex* SNPs (K=4 for NIST populations) estimated from the six *STRUCTURE* analyses of **Figure 36A-E**, indicate ASW samples have a range of 5–45% European co-ancestry with African the major component and PUR; CLM; MXL show increasing levels of American co-ancestry with European the major component plus indications of a significant African third component in many Colombians and Puerto Ricans. A large number of Mexicans (MXL) have comparable proportions of European to Native American ancestry across the population sample, while many PUR, CLM and MXL show their predominant ancestry is European, not Native American.

The *STRUCTURE* analysis of forensic reference/control DNAs is presented in **Figure 36D**. The patterns observed for the ancestries of SRM 2391c A, B and C components match those reported by NIST, while the other reference/control DNAs are consistent with being un-admixed European ancestries.

An initial *STRUCTURE* analysis was made for the NIST population panel alone (i.e., excluding reference populations) to measure the genetic structure discernible in this sample of the U.S. The cluster plots of the optimum K=4 (revised SNP set only, and with cluster colours modified to avoid overlap with in **Figure 36** plots) are shown in **Figure 37**. Four genetic clusters were detected despite the heterogeneity of the sample set and a largely cultural rather than a population genetics-based definition of Hispanics. The analysis therefore detects four clusters in the NIST panel that closely correspond to the population definitions given to the samples. Furthermore most Hispanics are differentiated by showing a distinct genetic component that does not correspond to the clusters defining the other three U.S. populations.

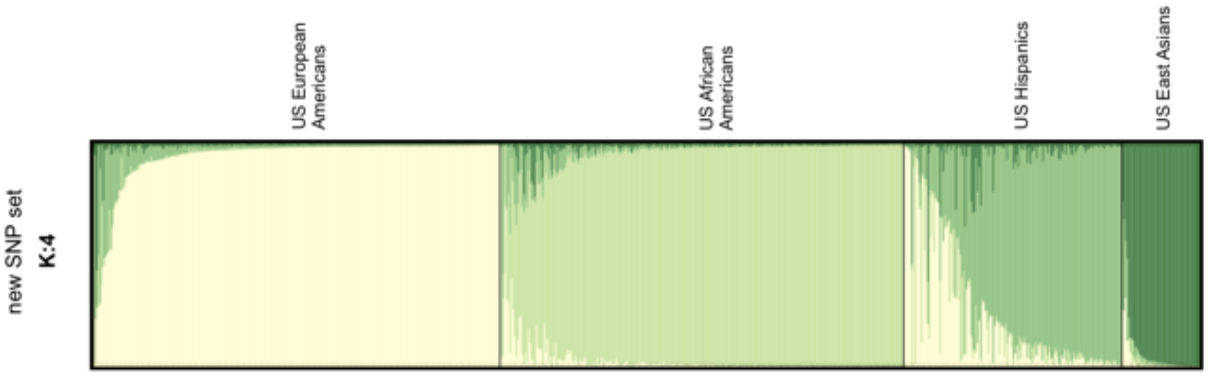


Figure 37: *STRUCTURE* cluster plot of NIST populations at *K*=4.

With the observation that four clusters were detected in the NIST U.S. panel we then performed *STRUCTURE* analysis adding reference populations and the resulting *K*=4 cluster plot is shown in **Figure 36E** (revised SNP set only). In this analysis reference populations were expanded to include several SNP*for*ID populations with admixture. As with the admixed populations in **Figure 36C** (ASW, CLM, MXL, PUR), NIST individuals have been ordered in ascending major co-ancestry component (except descending for U.S. African Americans). It is evident that each NIST group corresponds well with the expected population definitions. U.S. Caucasians and U.S. East Asians show largely un-admixed European and East Asian ancestry. In contrast, a large proportion of African Americans show significant levels of European co-ancestry as a major second component. U.S. Hispanics in particular show considerable heterogeneity and complex patterns of population admixture. The largest proportion of Hispanics show majority European co-ancestry, there is a wide range of Native American co-ancestry proportions and we found detectable African co-ancestry as a third component in most samples. It is evident that U.S. Hispanic individuals have a higher Native American co-ancestry compared to other U.S. population groups, which helps to explain the differentiation pattern observed in **Figure 37**.

As it is difficult to characterise the overall pattern of admixture in any one population when there is a continuous range of detected second or third co-ancestry, we summarised the data from all samples analysed in the *STRUCTURE* run of **Figure 36E** by compiling mean cluster membership proportions with 1st (25th percentile) and 3rd (75th) quartile ranges for the eight population groups analysed using individual *STRUCTURE* membership proportions. These are shown in **Figure 38** divided into four ancestry proportion plots based on the range of membership values to each cluster. The boxes denote 1st-3rd quartile ranges each containing

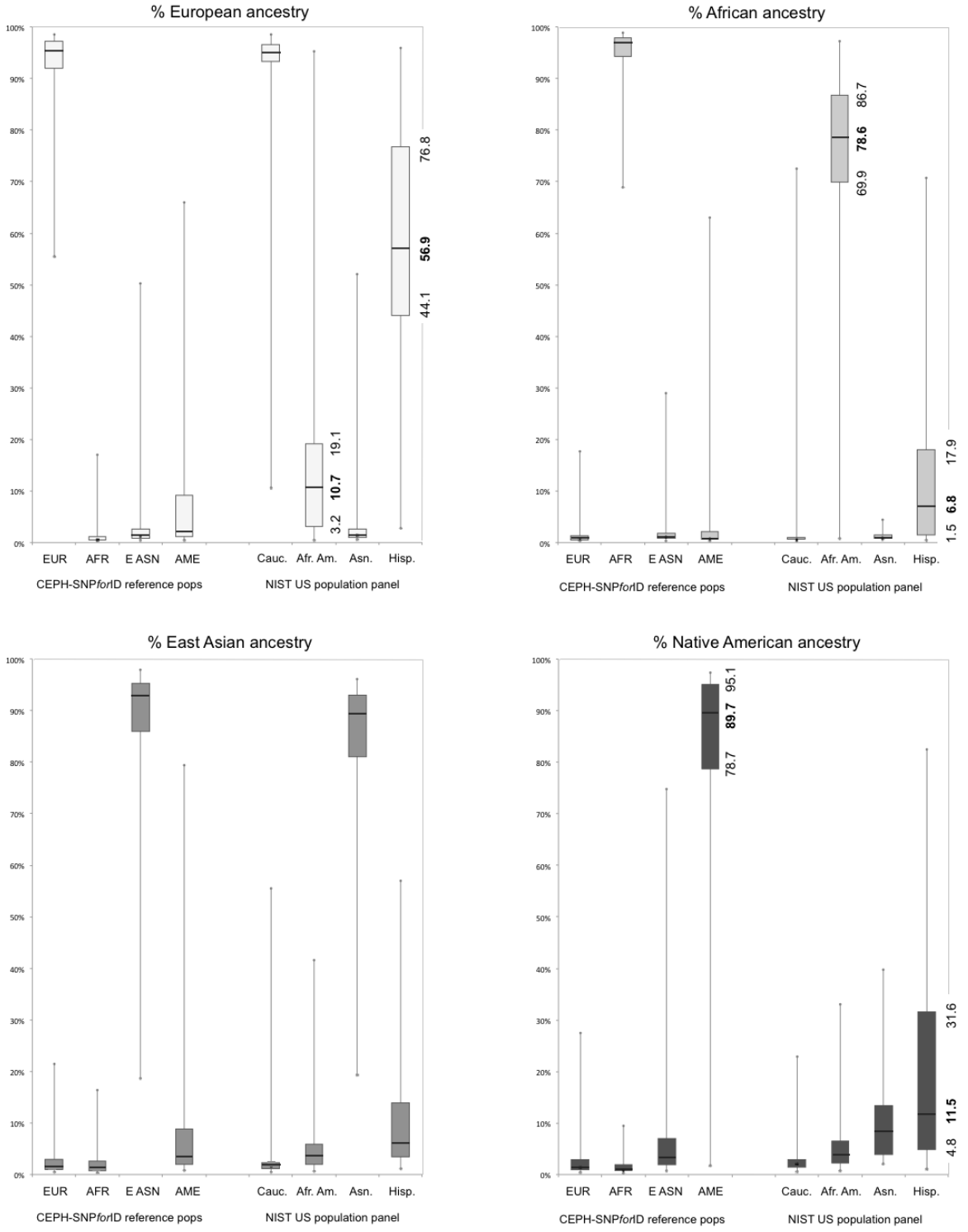


Figure 38: Distribution of cluster membership proportions (% proportions of membership to an optimum *K*=4 clusters) shown by NIST U.S. population panel samples and HGDP-CEPH/SNP*for*ID reference populations. These values were obtained from the *STRUCTURE* run of **Figure 36E**. Boxes indicate 1st and 3rd quartile ranges and contain the mean membership proportion for each population group and *K* ancestry. Whisker lines indicate the full range of values observed. Values placed next to certain selected populations show mean and quartile range values.

the mean value line with whiskers indicating the full range. The lowest means and narrowest quartile ranges are found in the reference European, African and East Asian proportions for their alternative ancestries (e.g. Europeans have minimal African/East Asian ancestry, etc.) and near identical patterns describe the NIST U.S. European and U.S. East Asian samples. NIST Hispanics show the highest mean value for European ancestry of 57% and this has the broadest quartile range, while the mean American ancestry proportion is only 11.5% but with an equally broad quartile range. In the Hispanics the observed African ancestry has a mean of 6.8% so on average this represents a quite minor third component of co-ancestry but the quartile box shows a significant range reaching 18% at the 3rd quartile, so for many NIST Hispanics African ancestry is a detectably larger component. Therefore, as a group, the NIST Hispanics can be described as showing predominantly European ancestry with comparatively minor proportions of Native American co-ancestry plus detectable but generally much lower African co-ancestry. This pattern is in sharp contrast to the HGDP-CEPH and SNP*for*ID Native American reference populations that have a mean 90% American ancestry and minimal European or African co-ancestry. NIST African Americans give a consistent pattern of African major ancestry ranging from 70 to 85% and European co-ancestry of about 5-20% in most samples.

The classification success rates from *Snipper* cross-validation of the HGDP-CEPH panel samples are outlined in Table 22, indicating enhanced performance of European, East Asian and Oceanian classifications, with 0.64%, 2.20% and 3.85% improvements in assignment success respectively. Although the *34-plex* test was not designed to infer Native American or Oceanian ancestry, the enhanced differentiation of East Asians provided by rs3827760 brings improvements to the analysis of all five global groups.

Table 21: Average membership proportions to *K*=5 clusters from the *STRUCTURE* analysis of HGDP-CEPH populations plotted in Figure 36**A.** Values in bold indicate improved performance with the revised SNP combination.

Previous 34-plex (admixture model)							Revised 34-plex (admixture model)						
Given population	Proportion of membership of each pre-defined population in each of the 5 clusters					Number of individuals	Given population	Proportion of membership of each pre-defined population in each of the 5 clusters					Number of individuals
	Inferred clusters							Inferred clusters					
	AFR	EUR	EAS	OCE	NAM			AFR	EUR	E ASN	OCE	AME	
African	**0.9560**	0.0060	0.0110	0.0170	0.0100	98	African	**0.9580**	0.0067	0.0080	0.0190	0.0083	98
European	0.0060	**0.9530**	0.0120	0.0130	0.0160	158	European	0.0057	**0.9573**	0.0100	0.0130	0.0140	158
East Asian	0.0070	0.0250	**0.8764**	0.0546	0.0370	227	East Asian	0.0070	0.0230	**0.9116**	0.0270	0.0313	227
Oceanian	0.0060	0.0050	0.0123	0.9687	0.0080	26	Oceanian	0.0070	0.0060	0.0130	0.9630	0.0110	26
Native American	0.0060	0.0173	0.0190	0.0130	**0.9446**	63	Native American	0.0060	0.0173	0.0190	0.0117	**0.946**	63

Table 22: Classification success (bold % values) from cross validation of new SNP combination.

Previous 34-plex, 5 group training set of supplementary Table S3, worksheet 2						Revised 34-plex, 5 group training set of supplementary Table S3, worksheet 22					
	Estimation with cross-validation to compute the success ratio using the best 34 SNPs:						Estimation with cross-validation to compute the success ratio using the best 34 SNPs:				
	Africa	Europe	East Asia	Oceania	America		Africa	Europe	East Asia	Oceania	America
Population of African origin	**100%**	0.00%	0.00%	0.00%	0.00%	Population of AFRICA origin	**100%**	0.00%	0.00%	0.00%	0.00%
Population of European origin	0.00%	**98.73%**	0.63%	0.00%	0.63%	Population of EUROPE origin	0.00%	**99.37%**	0.00%	0.00%	0.63%
Population of East Asian origin	0.00%	0.00%	**92.51%**	0.88%	6.61%	Population of EAST ASIA origin	0.00%	0.00%	**94.71%**	0.44%	4.85%
Population of Oceanian origin	0.00%	0.00%	3.85%	**96.15%**	0.00%	Population of OCEANIA origin	0.00%	0.00%	0.00%	**100%**	0.00%
Population of American origin	0.00%	0.00%	0.00%	0.00%	100%	Population of AMERICA origin	0.00%	0.00%	0.00%	0.00%	100%

4.2.1.3 Reference/control DNA profiles and AIM-SNP genotypes in seventy populations

SNP profiles for the reference/control DNA samples were compiled through independent, triplicated typing in two laboratories with full consensus in all cases and these genotypes are listed in Table 23. **Figure 36D** shows the *STRUCTURE* analysis of the nine reference DNAs detecting the Oceanian ancestry of SRM 2391c-C, with ~93% membership to the Oceanian cluster identified from the HGDP-CEPH reference samples. The only other SRM 2391c component to show a significant non-European component (excluding the mixed sample D) was sample B with 10.4% membership to the Native American cluster identified from the HGDP-CEPH reference samples, this division of European and American co-ancestry broadly matches values seen in the admixed American populations of CLM, MXL and PUR as well as many NIST Hispanics. The ancestry membership proportions (*K*5=five population clusters) obtained from *STRUCTURE* analysis of the nine reference DNAs are listed in the bottom part of Table 23.

Likelihoods and ancestry assignments were also made for the reference DNAs based on the revised *34-plex* SNP profiles submitted to *Snipper*. Assignments were made using a custom five-group HGDP-CEPH training set (AFR-EUR-E ASN-AME-OCE). The Oceanian classification of SRM 2391c-C matches the *STRUCTURE* inferences described in section 4.2.1.2 and we interpret this as an unequivocal ancestry assignment. Five reference/control DNAs (those except SRM 2391c-B to E) gave very strong European classifications of 1E+15 to 1E+18 times more likely European than the next closest assignment (American or East Asian). Two SRM controls, SRM 2391c-B and E gave reduced European assignments, although the non-European co-ancestry indicated by *Snipper* and measured by *STRUCTURE* as 10% and 5% American membership proportions respectively, is based on SNPs that have less power to differentiate this ancestry compared to EUR-AFR-E ASN.

Table 23: Standard DNA control profiles and ancestry membership proportions in five population cluster estimated from the STRUCTURE analysis shown in Figure 36E. Bold values represent more than 10% proportion in a cluster.

		SRM-A	SRM-B	SRM-C	SRM-D	SRM-E	SRM-F	9947A	XY5	2800M
rs917118	A07	AG	AG	GG	AG	GG	GG	GG	GG	GG
rs1321333	P03	CT	CC	CT	CT	CC	CT	TT	CT	TT
rs1024116	A29	AG	AA	GG	AG	AG	AA	AG	AA	AA
rs7897550	P05	CT	CT	CC	CT	CT	CT	CC	CC	CC
rs722098	A21	AA	AA	AG	AG	AA	AG	AA	AG	AG
rs10843344	P06a	CT	CC	CC	CT	CT	CC	CC	CC	CT
rs239031	P07	TT	TT	CT	TT	TT	TT	TT	TT	TT
rs12913832	P08	GG	AG	AA	AG	AG	AG	GG	AG	AG
rs2040411	A40	AG	AA	AG	AG	AG	AA	GG	AG	AA
rs1978806	P09a	TT	TT	TT	TT	TT	TT	TT	TT	TT
rs773658	P10	CC	CC	CC	CC	CC	CC	CC	CC	CC
rs10141763	P11	AA	AA	AA	AA	AA	AA	AA	AA	AA
rs182549	P12	TT	CT	CC	CT	CT	TT	TT	TT	CT
rs1573020	P13	AA	AA	AA	AA	AA	AA	AA	AA	AA
rs896788	P14	CC	CC	CT	CT	CC	CC	CC	CC	CT
rs2065160	P15	AA	AG	AA	AA	AA	AA	AA	AG	AG
rs2572307	P16a	GG	GG	GG	GG	GG	GG	GG	GG	GG
rs2303798	P17	CC	CC	CT	CT	CC	CC	CC	CC	CC
rs2065982	P18	AA	GG	AA	AA	AG	AA	AG	AA	AA
rs3785181	P19	CC	CC	CC	CC	CC	CC	CC	CC	CT
rs881929	P20	GG	GT	GT	GG	GT	TT	GG	GT	TT
rs1498444	P21	AA	AC	AC	AC	AC	AC	CC	AC	AC
rs1426654	P22a	TT	TT	CC	CT	TT	TT	TT	TT	TT
rs2026721	P23	GG	GG	GG	GG	GG	GG	GG	AG	GG
rs4540055	P24	AT	AT	CC	ACT	AT	AT	AA	AA	AT
rs16891982	P25a	CC	CG	GG	CG	CG	CC	CC	CC	CC
rs1335873	A52	AT	TT	AA	AT	AA	AA	AA	AT	AA
rs1886510	A13	AG	GG	GG	AG	AG	AG	AA	AA	AG
rs730570	P26	CT	TT	CC	CT	CT	TT	TT	CT	TT
rs5030240	P27	CG	CC	AC	AGC	AC	CG	CC	CC	AC
rs2304925	P01	TT	TT	GG	GT	TT	TT	TT	TT	TT
rs5997008	P02	CC	CC	CC	CC	CC	CC	CC	CC	CC
rs3827760	P28	AA	AA	AG	AG	AA	AA	AA	AG	AA
rs2814778	P04	TT	TT	TT	TT	TT	TT	TT	TT	TT
African		0.002	0.005	0.008	0.009	0.003	0.002	0.001	0.004	0.002
European		**0.988**	**0.845**	0.008	0.632	**0.916**	**0.988**	**0.990**	**0.916**	**0.975**
East Asian		0.003	0.042	0.042	0.143	0.024	0.004	0.003	0.049	0.007
Oceanian		0.003	0.004	**0.928**	0.157	0.008	0.002	0.002	0.005	0.005
American		0.004	**0.104**	0.014	0.059	0.049	0.004	0.004	0.026	0.011

4.2.1.4 Tri-allelic AIM-SNPs rs4540055 and rs5030240

Two of the most informative components of the *34-plex* assay are tri-allelic SNPs: rs4540055 and rs5030240, both having three alleles recorded at a single variable nucleotide position (A/C/T and A/C/G respectively) in all population groups, albeit at low frequencies for the third allele in some populations (rs4540055-C and rs5030240-A). We made two noteworthy observations concerning these tri-allelic SNPs: i) their evident success in

detecting three alleles in the two mixture contributors of SRM2391c-D and; ii) the disparity between the allele frequency distributions obtained from our SNaPshot typing compared to the data from the next-generation deep re-sequencing analysis of 1000 Genomes.

The original motivation for identifying and characterising tri-allelic SNPs was to bring mixture detection properties into forensic SNP typing assays (Phillips et al. 2004b) and this approach has been more thoroughly explored in recent studies by Westen et al. (2009). Since tri-allelic SNP allele frequencies commonly show strong population differentiation it is also not surprising to see three alleles in both markers from a mixture of a European and an Oceanian donor. For instance, the rs4540055-C allele is almost eight times more frequent in Oceanians than Europeans and rs5030240-C more than eighteen times more frequent in Europeans. So while tri-allelic SNPs are much weaker than STRs at detecting multiple alleles in mixed donor profiles, the probability of finding more than two alleles is raised if donors are from different populations, potentially helping to identify the ancestry of components in simple mixtures. The three-peak patterns observed in SRM2391c-D are shown in **Figure 39A**.

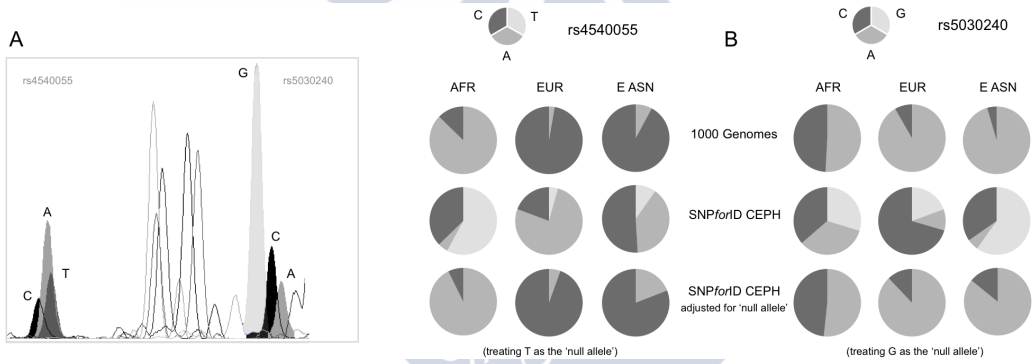


Figure 39: (A) Observed tri-allelic SNP peak patterns in the mixed-source SRM 2391c-D sample. (B) Allele frequency estimates for both tri-allelic SNPs *in 34-plex* obtained from next generation sequencing by 1000 Genomes (top row) and SNaPshot typing of the HGDP-CEPH panel (middle row). The bottom row pie charts indicate the HGDP-CEPH allele frequencies found when the T allele is treated as silent (undetectable) in rs4540055 and the G allele in rs5030240, giving a close match to the estimates from 1000 Genomes and explaining the discrepant values.

The second immediately apparent characteristic of tri-allelic SNPs seen in the allele distribution comparisons in this study is their failure to be characterised as three allele SNPs by both the HapMap and 1000 Genomes projects. Each project has reported the two *34-plex* tri-allelic SNPs as well as those additionally identified as tri-allelic SNPs by Westen et al. (2009) as binary SNPs. There appears to be problems with the genotyping of tri-allelic SNPs with next- generation whole genome sequencing approaches, and it is significant that 1000 Genomes have changed the allele calls of rs4540055 from AT to AC and rs5030240 from AG

to AC between published data revisions of December 2010 and August 2011. Comparisons of the allele frequency distributions of both tri-allelic SNPs from our typing with 1000 Genomes sequencing are summarised in **Figure 39B** and these appear to be quite disparate. We have explained these differences by treating one of the alleles as silent, for example when rs4540055-T is considered silent or undetectable, then AT appears as an AA homozygote and CT as a CC. Therefore the A frequency can be estimated as: (AA + AT + (AC × 0.5)/total genotypes) and C = (CC + CT + (AC × 0.5)/total). When these adjustments are made to the HGDP-CEPH panel genotypes made by SNP*for*ID SNaPshot typing, then the allele frequencies are much better matched. **Figure 39B** shows these HGDP-CEPH panel genotype adjustments made by assuming silent rs4540055-T and rs5030240-G alleles. However, it is discouraging that one of the benefits of such extensive sequencing initiatives: an enhanced ability to locate and identify SNPs with more than two alternative nucleotides, appears to be lost in progressing to next generation sequencing technologies. This is further exacerbated by 1000 Genomes using multiple sequencing centres/platforms but organising analyses to concentrate on samples from one population in any one centre.

4.2.1.5 Concluding remarks

Swapping a single component SNP at the same time as making adjustments to the chemistry and certain mobility positions in the *34-plex* assay provided relatively little disruption to routine forensic SNP typing with this ancestry analysis test. The enhancement to the differentiation of East Asians is evident amongst the seventy populations characterised and the comparisons made in this study. As East Asians had the weakest population differentiation using the original 34 SNP test, the introduction of a near-fixed SNP in this population group represents a major enhancement of the test without upsetting its forensic performance. While SNaPshot primer extension assays remain the only viable multiplexed SNP typing option available to forensic analysis it is crucial to thoroughly gauge the mobility and signal variation entailed with using this chemistry as we have done with the 34 sets of peak pair data observed with the NIST U.S. population panel.

The Hispanics that were studied as part for the NIST panel highlight the difficulties of analysing individuals with complex, potentially three-way, patterns of admixture. Not only is it difficult to make accurate assessments of co-ancestry when these are so varied within the

group as a whole, but also component AIM-SNPs in small forensic multiplexes such as the *34-plex* test cannot provide equally informative differentiations of all co-ancestry contributors. With this in mind we have continued development of additional AIM-SNP based ancestry tests, complementary to the *34-plex*, that improve the differentiation of Native Americans from EUR-AFR-E ASN populations. It is also likely that an optimum strategy will be the addition of ancestry-informative Indels in small multiplexes with similar properties to the *34-plex* SNPs. We recently published details of an AIM-Indel set (Pereira et al. 2012b) that is certain to improve the inferences made for complex admixture patterns when combined with the current *34-plex* SNPs and bring improved differentiation of Native American ancestry. Therefore forensic ancestry inferences based on typing a range of short binary markers will show improved precision while keeping multiplexes small-scale, manageable and with sufficient sensitivity for challenging forensic material. Additionally, we continue to advocate the use of tri-allelic SNPs as highly differentiated ancestry markers. Regrettably, the opportunity to build a well-validated catalogue of tri-allelic SNPs from the output of 1000 Genomes is not available to forensic genetics research at this moment and consequently their number remains small.

Revisions to the *34-plex* primers, chemistry and cycling conditions added since the original development of the assay have evolved gradually and therefore it is difficult to state how each of these changes has influenced the improved peak height balance we report. However the biggest advance in *34-plex* electropherogram quality was obtained with the changes to primer ratios that accompanied changes to the positions of four SNPs. Specifically, the largest improvement resulted from increasing the PCR and EXT primers for rs2814778 (P04) and rs16891982 (P25a), while reducing those of rs2304925 (P01).

Lastly, at the time of writing, a new study of the statistical limitations of familial searching strategies has highlighted the problem of using inappropriate STR allele frequencies when the ancestry of the profile donor is not known (Rohlfs et al. 2012). Using generalised or incorrect allele frequencies can lead to a large number of adventitious matches and the search loses specificity. Therefore it appears that a simple forensic ancestry test with sufficient power to distinguish the major population groups could be applied as an informative adjunct to conventional STR profiling and help enhance the power such tests already provide. This approach would not involve database-wide SNP typing as only the profile would require an inference of ancestry to select the appropriate STR allele frequencies in each search.

4.2.2 *Pacifiplex*: an AIM-SNP panel centred on Australia and the Pacific region

4.2.2.1 Assay performance

An example electropherogram of *Pacifiplex* genotyping 1 ng of 9947A control DNA is shown in **Figure 40**. The dilution series results used to establish minimum DNA input for typical forensic PCR amplifications are summarised in **Figure 41**, indicating SNP drop-out was only observed when 63 pg of DNA was used (applying a 50 RFU peak height threshold). Therefore, a minimum of 125 pg DNA provides suitable sensitivity for all but the lowest levels of DNA encountered in forensic analyses.

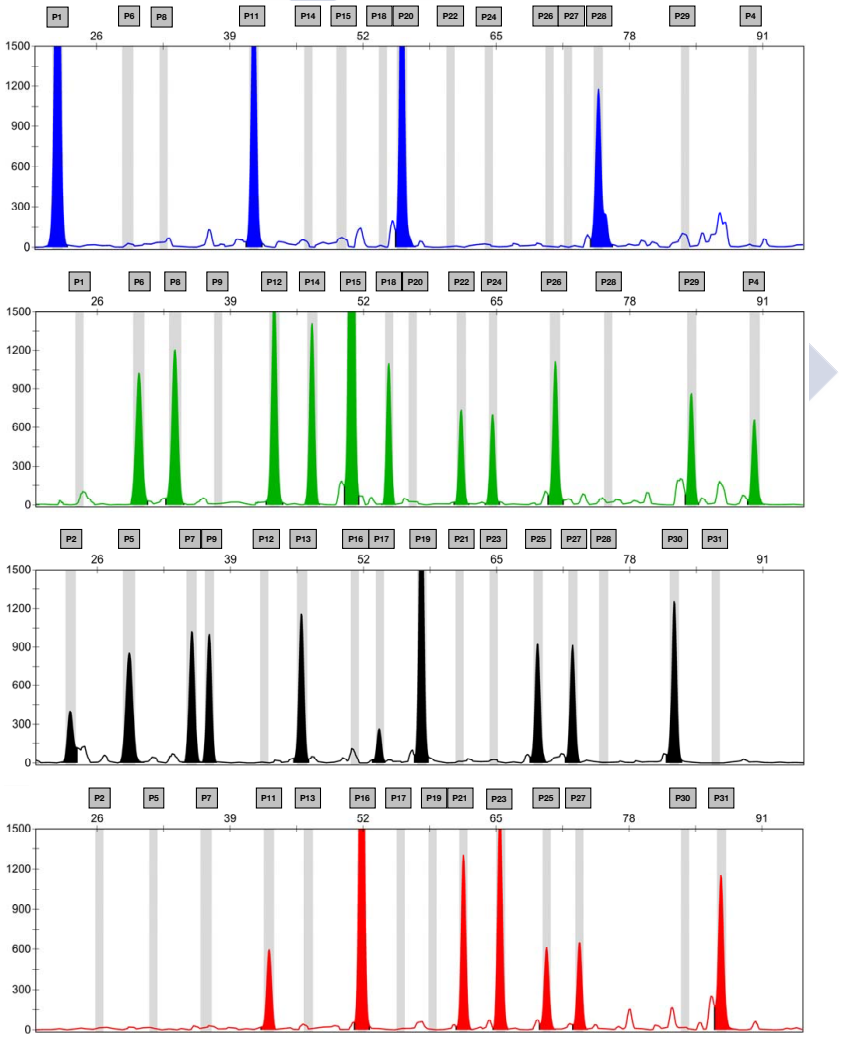


Figure 40: Electropherogram of *Pacifiplex* SNPs obtained from the amplification of 1 ng of 9947A control DNA.

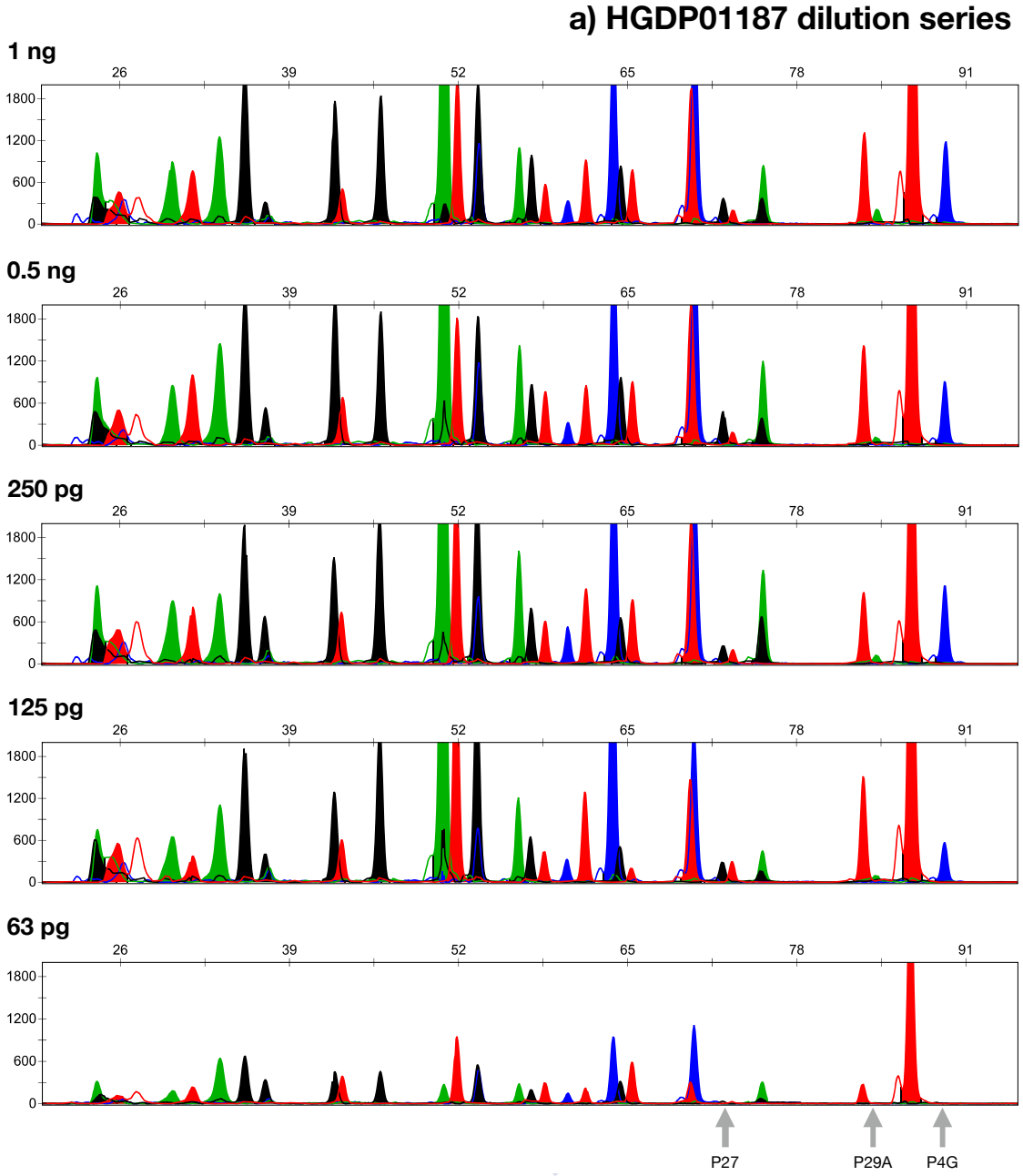


Figure 41: Electropherograms obtained from sensitivity studies of reference DNA samples with five decreasing DNA input of: 1 ng; 0.5 ng; 250 pg; 125 pg and 63 pg. a) HGDP-CEPH panel sample (HGDP01187). Dropout was only observed in the 63 pg sample. b) Coriell sample (HG00403). Dropout was only observed in the 63 pg sample. SNPs failing to amplify are indicated by grey arrows.

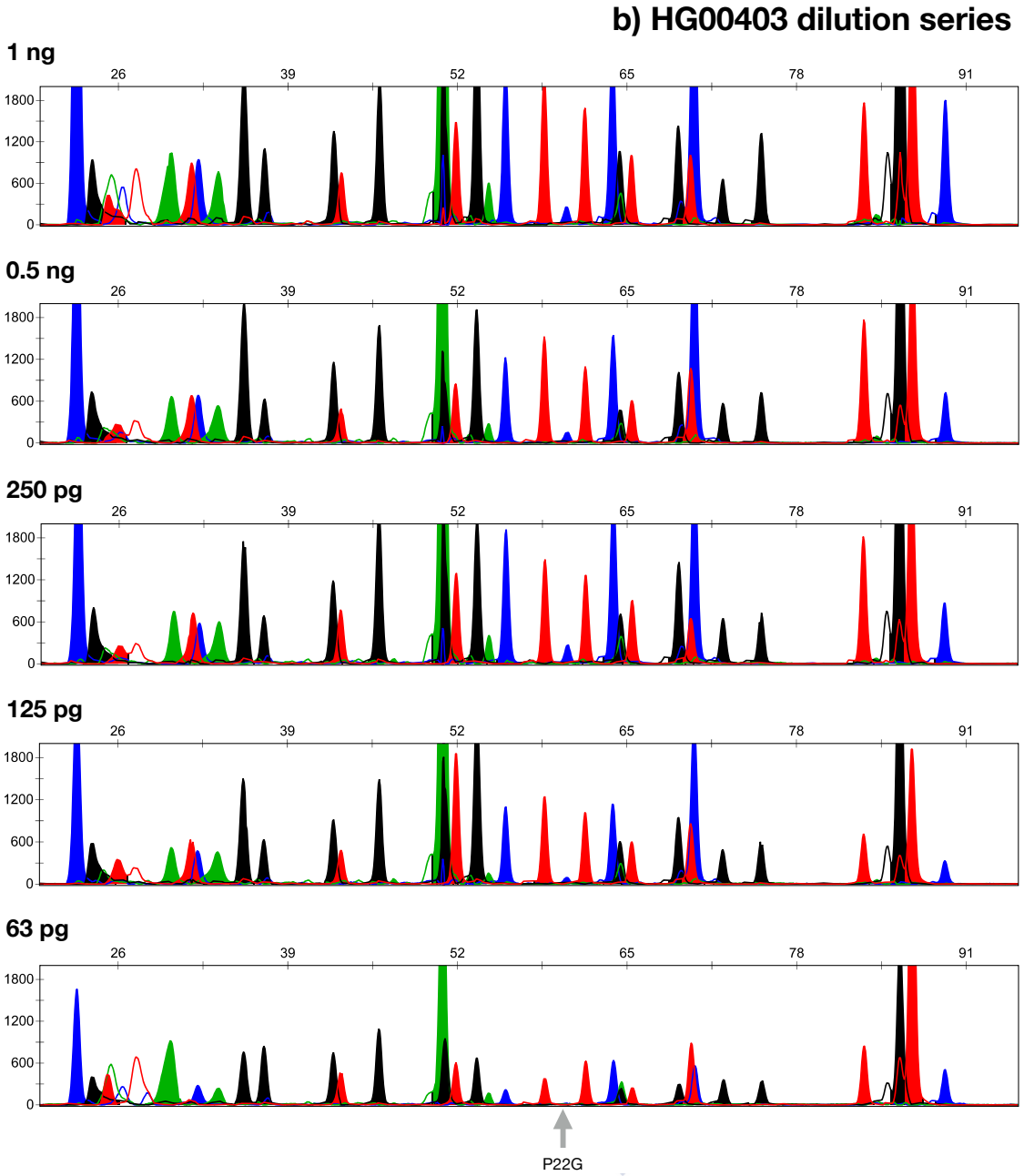


Figure 39 *(continued)*

Design of *Pacifiplex* PCR primers to amplify an average 91 bp fragment size with a maximum 126 bp ensured the assay is applicable to highly degraded DNA. Analysis of two degraded bone samples and serum are summarised in **Figure 42**. Both bone samples had DNA concentrations below the 125 pg minimum outlined above. However, their amplification resulted in 24 of 29 consensus genotypes (two replicates). Although this corresponds to 83% profile completeness, likelihood ratio estimates are still comparable to those of European HGDP-CEPH individuals (see **Figure 43**). Despite low internal PCR

control (IPC) values (femur - Duo IPC 30.67; tibia - Duo IPC 30.02), the values >28 suggest some inhibition was present in both extracts (Fondevila et al. 2008a). Use of BSA in the PCR conditions may have contributed to the observed 83% amplification success rate. In the 50-year-old serum sample, complete and concordant profiles were obtained despite low levels of nucleated cells.

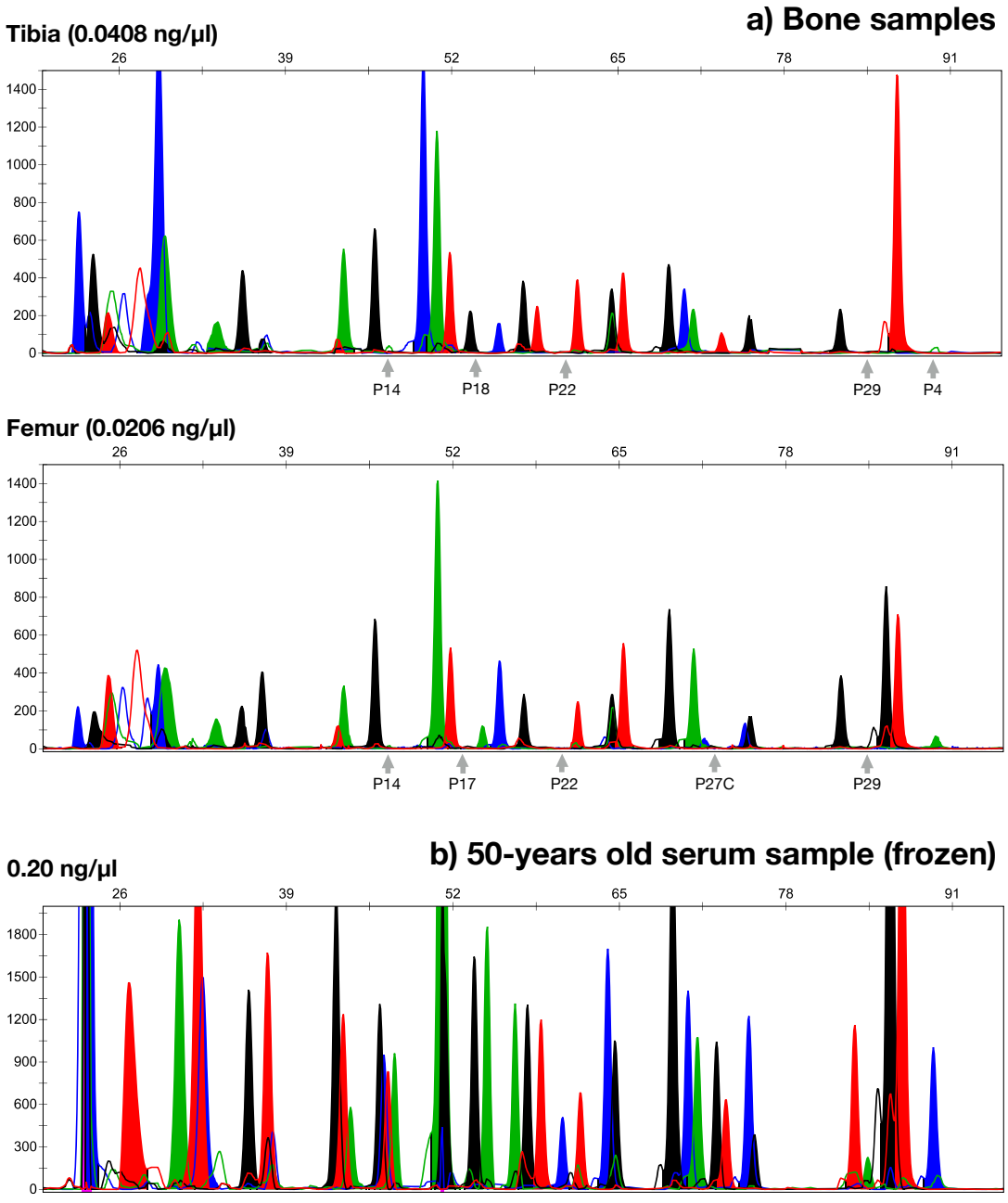


Figure 42: Electropherograms resulting from the amplification of degraded/low template DNA samples. a) Analysis of a tibia (0.0408 ng/µl) and femur (0.0206 ng/µl). Grey arrows indicate SNP dropout. b) 50-year-old serum sample stored in the freezer before DNA extraction (0.20 ng/µl).

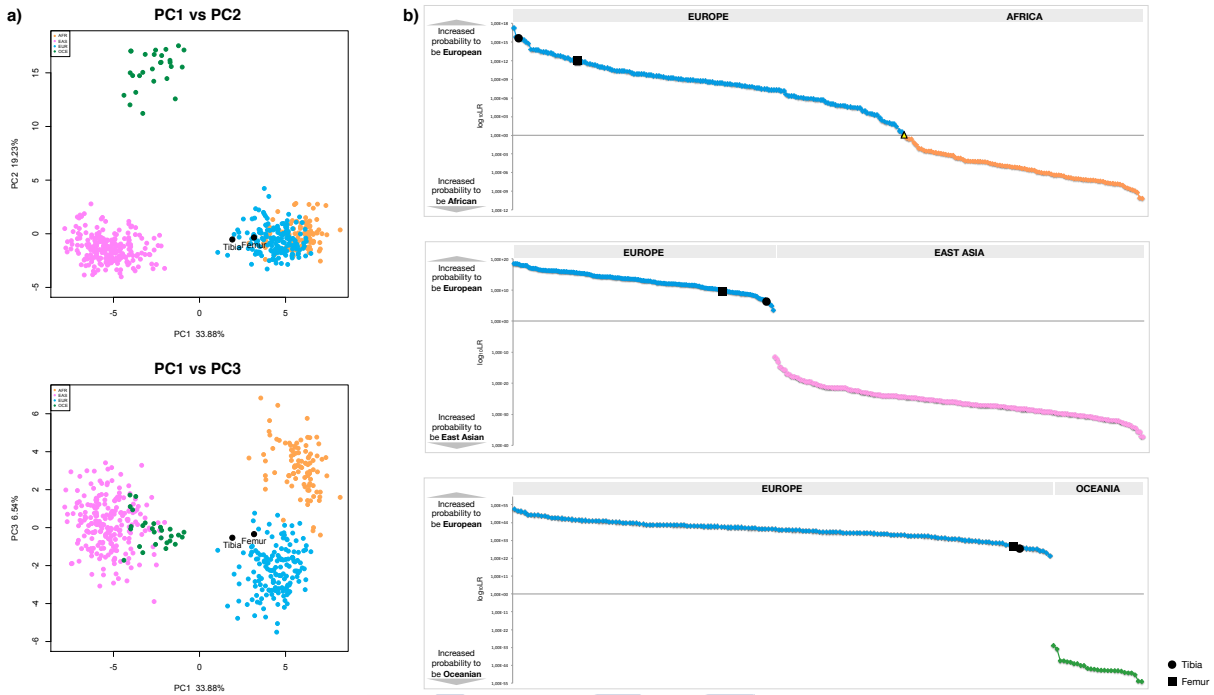


Figure 43: Ancestry assignment of two bone samples analysed with Pacifiplex (see profiles in Figure 42). a) PCA analysis shows that both bone samples cluster with Europeans from the HGDP-CEPH panel of samples. b) Plots of ranked assignment probabilities obtained from cross validation of three training sets based on the HGDP-CEPH population groups considered in this study (top to bottom: European vs. African, European vs. East Asian and European vs. Oceanian). The log10 likelihood ratio (LR) vertical scale of each plot corresponds to the probability of a sample being European over the probability of being from the other population. Increased distances from the mid-line of 1 indicate higher probability of ancestry from one of the two populations considered. A single African individual was misclassified as European (yellow triangle). In all three pairwise comparisons the bone samples are clearly included in the European LR value ranges despite profile completeness of 83% (24/29 SNPs reported).

4.2.2.2 Characterization of HGDP-CEPH reference populations with 62 AIMs

Although the *34-plex* panel was designed for a 3-group comparison differentiating African, European and East Asian ancestries, **Figure 44** indicates *34-plex* SNP data has some capacity to differentiate Oceanian populations comparing four groups. However, a degree of overlap is discernible between East Asians and Oceanians in the PCA analysis and despite 100% classification success in *Snipper* Bayes analyses, *STRUCTURE* tests indicate an optimum *K*=3 genetic clusters (delta *K* analyses and cluster plots of *K*=2 to *K*=7 runs outlined in **Figure 45**). Furthermore, one-against-all average F_{ST} and *In* values for the comparison of HGDP-CEPH Oceanians with the other three groups are reduced by 2x or more (Table 24). For this reason, *Pacifiplex* SNPs were selected to complement those of *34-plex* by targeting Oceanian vs. other group differentiation, but also by enhancing differentiation of the most closely related East Asian group by including several EAS-informative SNPs. Therefore δ

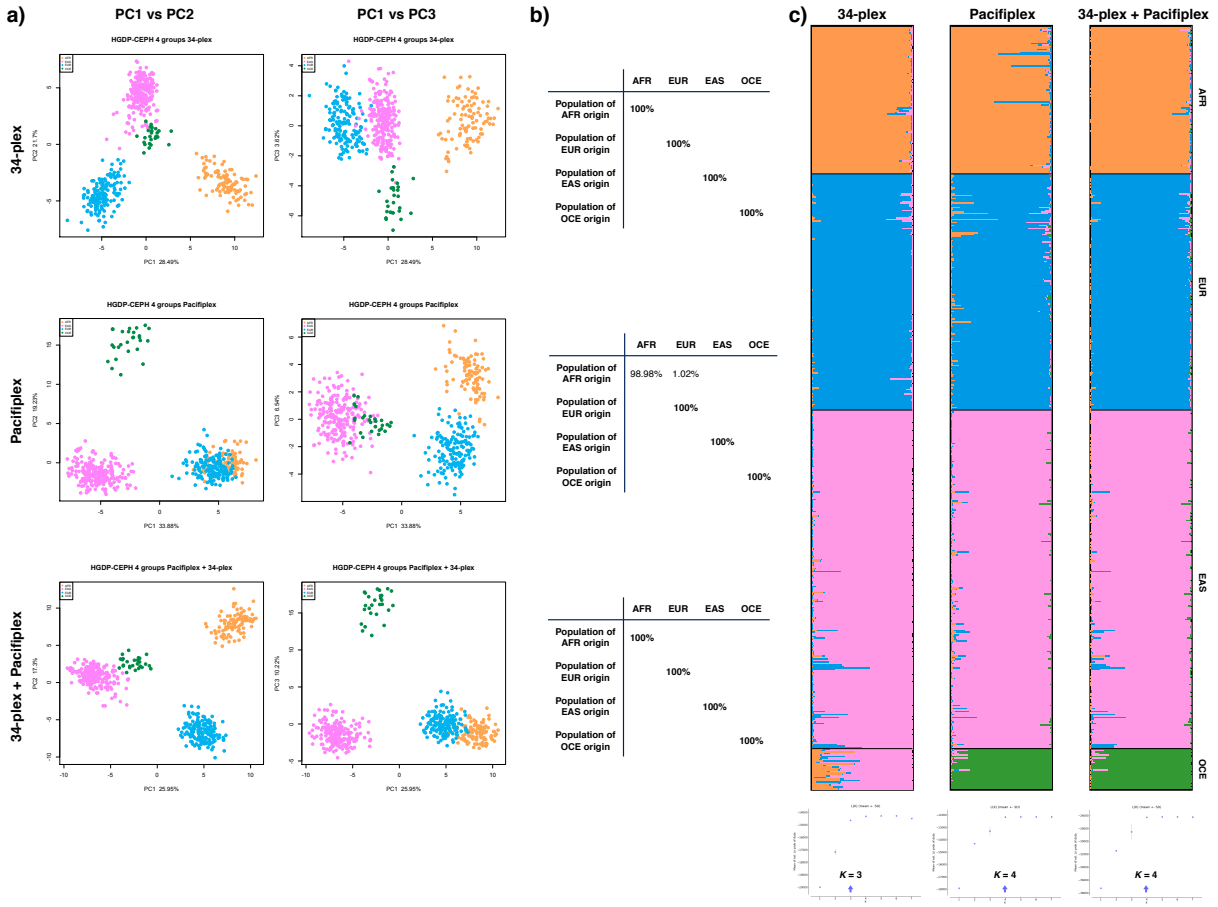


Figure 44: Ancestry analyses of four HGDP-CEPH reference population groups (AFR - Africa; EUR - Europe; EAS - East Asia; OCE - Oceania) using *Pacifiplex*, *34-plex* and both assays combined. a) PCA analysis. Comparisons of the first two principal components (PC1 vs. PC2) plus first and third principal components (PC1 vs. PC3) provide differentiation of the four groups. b) *Snipper* cross-validation classification success values. c) *STRUCTURE* analyses of the three sets of SNPs. Each was analysed with three replicates from *K*=1 to *K*=7 using the admixture ancestry model and correlated allele frequencies (100,000 burnin iterations and 100,000 MCMC repeats). *Structure Harvester* estimated the optimum *K* value (marked by blue arrows in the Ln estimated probabilities plots) and generated input files for *CLUMPP* and *distruct*.

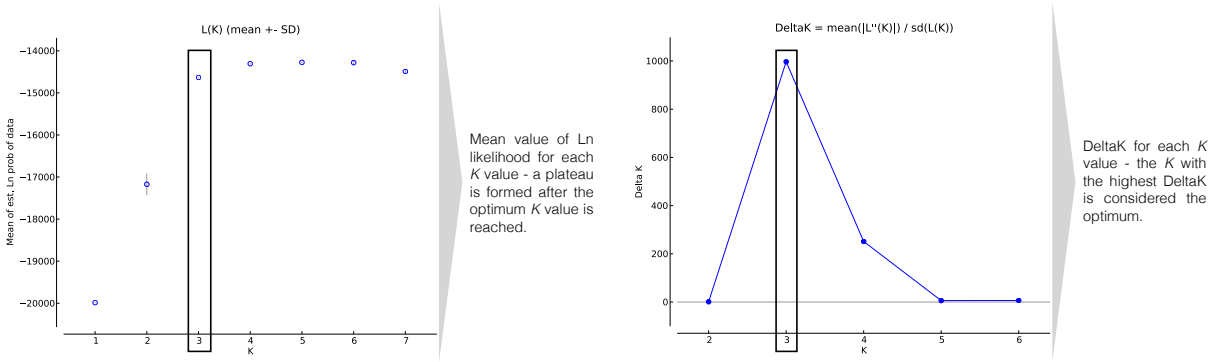


Figure 45: Mean Ln likelihood and *delta K* values for each *K* value in a *STRUCTURE* analysis of four HGDP-CEPH reference population groups using *34-plex*. Runs used the admixture ancestry model, correlated allele frequencies, three replicates from *K*=1 to *K*=7 (100,000 burnin iterations and 100,000 MCMC repeats). *Structure Harvester* estimated the optimum *K* value and generated the plots.

values for the 23 SNP selected from HGDP-CEPH data exceed 0.5 for at least one pairwise comparison of the Oceanian group (Table 24). Component SNPs rs721367, rs3827760 and rs4892491 were selected to enhance East Asian differentiation. The *Pacifiplex* panel results in high one-against-all average F_{ST} values (Table 24), increased separation of the Oceanian and East Asian groups in the PCA analysis and provides an optimum K=4 clusters in *STRUCTURE* analyses (**Figure 44**). As expected, *Pacifiplex* alone has less power to differentiate African and European populations, with some overlap in the PCA and reduction of *Snipper* classification success (**Figure 44**). One highly EUR-informative SNP (rs1592672) was rejected from the panel during assay optimization due to non-specific interactions with other markers. Using *Pacifiplex* and *34-plex* in a combined analysis improves the balance of differentiation power amongst the four population groups (Table 24, average one-against-all F_{ST}: AFR - 0.31024; EUR - 0.26185; EAS - 0.29782; OCE - 0.32662) and this is reflected in the distinct four group differentiation in the PCA and *STRUCTURE* clustering patterns shown in **Figure 44**. Therefore, although small-scale tests, *Pacifiplex* and *34-plex* genotypes can efficiently differentiate four population groups. As other studies suggest, such small marker sets that target the highest possible δ values during SNP selection can be used to assign ancestry or estimate co-ancestry proportions in admixed individuals while genotyping effort, sample usage and costs are kept low (Shriver et al. 1997, Rosenberg et al. 2003).

The HGDP-CEPH panel of samples is a suitable reference training set. Its advantage over HapMap and 1000 Genomes project samples is that it includes Oceanian populations. Nevertheless, this panel does suffer from some irregular global sampling (Rosenberg et al. 2002) and the number of Oceanian samples is very small (N=27) suggesting use of the HGDP-CEPH panel to design and predict the performance of a set of OCE-informative AIMs may be prone to a degree of bias in marker selection. Although ascertainment bias is a characteristic mainly observed in genome-wide studies (Marth et al. 2004, Clark et al. 2005), it can also affect patterns of genetic variation observed across populations (Romero et al. 2009, Novembre and Ramachandran 2011). Ascertainment bias results from markers being selected from data sets with only a few individuals, decreasing the chance of finding SNPs with low minor allele frequencies, so certain loci are not polymorphic in the populations used during marker selection and are not considered further (Clark et al. 2005, Guillot and Foll 2009, Albrechtsen et al. 2010, Novembre and Ramachandran 2011). Using the HGDP-CEPH panel to select AIMs can also introduce ascertainment bias from reliance on SNP genotypes

Table 24: Allele frequencies of the 29 *Pacifiplex* SNPs from HGDP-CEPH African (AFR), European (EUR), East Asian (EAS) and Oceanian (OCE) populations. Absolute allele frequency differentials (δ) and F_{ST} values used to guide marker selection from a candidate pool of 3210 SNPs are shown. Average F_{ST}, *In* values and one-against-all *In* values for each population comparison are shown, plus average values for the one-against-all comparisons.

Internal code	rs number	Alleles described in dbSNP (build 138)		Reference allele frequency				Alternative allele frequency				Absolute allele frequency differentials (δ)					
		Reference	Alternative	AFR	EUR	EAS	OCE	AFR	EUR	EAS	OCE	AFR / EUR	AFR / EAS	AFR / OCE	EUR / EAS	EUR / OCE	EAS / OCE
P1	rs9908046	C	T	0.923	0.920	0.884	0.019	0.077	0.080	0.116	0.981	0.003	0.039	0.904	0.036	0.901	0.865
P2	rs721367	G	A	0.923	0.965	0.283	0.926	0.077	0.035	0.717	0.074	0.042	0.640	0.003	0.682	0.039	0.643
P4	rs3827760	T	C	1.000	0.993	0.165	0.944	0.000	0.007	0.835	0.056	0.007	0.835	0.056	0.828	0.049	0.779
P5	rs4683510	T	C	0.148	0.127	0.802	0.241	0.852	0.873	0.198	0.759	0.021	0.654	0.093	0.675	0.114	0.561
P6	rs2139931	A	G	0.944	0.771	0.860	0.019	0.056	0.229	0.140	0.981	0.173	0.084	0.925	0.089	0.752	0.841
P7	rs798949	T	C	0.724	0.510	0.120	0.944	0.276	0.490	0.880	0.056	0.214	0.604	0.220	0.390	0.434	0.824
P8	rs7832008	G	A	0.541	0.099	0.522	1.000	0.459	0.901	0.478	0.000	0.442	0.019	0.459	0.423	0.901	0.478
P9	rs10811102	G	T	0.985	0.732	0.300	0.019	0.015	0.268	0.700	0.981	0.253	0.685	0.966	0.432	0.713	0.281
P11	rs2409722	T	G	0.015	0.475	0.862	1.000	0.985	0.525	0.138	0.000	0.460	0.847	0.985	0.387	0.525	0.138
P12	rs9809818	C	A	0.010	0.067	0.833	0.981	0.990	0.933	0.167	0.019	0.057	0.823	0.971	0.766	0.914	0.148
P13	rs6886019	C	T	0.959	0.946	0.984	0.167	0.041	0.054	0.016	0.833	0.013	0.025	0.792	0.038	0.779	0.817
P14	rs3751050	A	G	0.974	0.895	0.938	0.093	0.026	0.105	0.062	0.907	0.079	0.036	0.881	0.043	0.802	0.845
P15	rs7623065	A	G	0.133	0.654	0.847	0.000	0.867	0.346	0.153	1.000	0.521	0.714	0.133	0.193	0.654	0.847
P16	rs1509524	A	G	0.883	0.780	0.856	0.093	0.117	0.220	0.144	0.907	0.103	0.027	0.790	0.076	0.687	0.763
P17	rs16946159	G	A	0.908	0.997	0.976	0.222	0.092	0.003	0.024	0.778	0.089	0.068	0.686	0.021	0.775	0.754
P18	rs2274636	A	G	0.959	0.911	0.736	0.111	0.041	0.089	0.264	0.889	0.048	0.223	0.848	0.175	0.800	0.625
P19	rs4704322	C	T	0.964	0.787	0.140	0.556	0.036	0.213	0.860	0.444	0.177	0.824	0.408	0.647	0.231	0.416
P20	rs12405776	C	T	0.633	0.965	0.458	0.093	0.367	0.035	0.542	0.907	0.332	0.175	0.540	0.507	0.872	0.365
P21	rs6054465	T	C	0.954	0.889	0.780	0.037	0.046	0.111	0.220	0.963	0.065	0.174	0.917	0.109	0.852	0.743
P22	rs3784651	T	C	0.372	0.876	0.251	0.000	0.628	0.124	0.749	1.000	0.504	0.121	0.372	0.625	0.876	0.251
P23	rs10970986	T	C	0.990	0.723	0.680	0.019	0.010	0.277	0.320	0.981	0.267	0.310	0.971	0.043	0.704	0.661
P24	rs6494411	C	T	0.066	0.115	0.896	0.346	0.934	0.885	0.104	0.654	0.049	0.830	0.280	0.781	0.231	0.550
P25	rs10183022	G	A	0.000	0.602	0.638	0.981	1.000	0.398	0.362	0.019	0.602	0.638	0.981	0.036	0.379	0.343
P26	rs10455681	A	G	0.995	0.844	0.178	0.019	0.005	0.156	0.822	0.981	0.151	0.817	0.976	0.666	0.825	0.159
P27 [§]	rs2184030	G	A,C	0.571	0.480	0.578	0.750	0.194	0.510	0.415	0.019	0.091	0.007	0.179	0.098	0.270	0.172
P28 [§]	rs2069945	C	A,G	0.112	0.459	0.652	0.037	0.184	0.061	0.075	0.444	0.347	0.540	0.075	0.193	0.422	0.615
P29	rs11156577	C	T	0.393	0.105	0.011	0.907	0.607	0.895	0.989	0.093	0.288	0.382	0.514	0.094	0.802	0.896
P30	rs4892491	G	A	0.684	0.815	0.071	0.722	0.316	0.185	0.929	0.278	0.131	0.613	0.038	0.744	0.093	0.651
P31 [*]	rs12434466	A	G	0.923	0.927	0.227	0.426	0.077	0.073	0.773	0.574	0.004	0.696	0.497	0.700	0.501	0.199
											Average:	**0.191**	**0.429**	**0.568**	**0.362**	**0.583**	**0.560**

[§] For these markers, the third allele was observed with frequency = 1-(reference allele frequency + alternative allele frequency). Third alleles are C and G for P27 and P28, respectively.

[*] This marker is not genotyped in HGDP-CEPH panel of samples. For this reason we used rs2224442 as a proxy to estimate allele frequencies and perform statistical analysis. P31 G allele corresponds to proxy's C allele and P31 A allele to the proxy's A allele.

Table 24 *(continued)*

Internal code	rs number	Pairwise F_{ST}						One-against-all F_{ST}				One-against-all *In*				
		AFR / EUR	AFR / EAS	AFR / OCE	EUR / EAS	EUR / OCE	EAS / OCE	AFR *vs.* ALL	EUR *vs.* ALL	EAS *vs.* ALL	OCE *vs.* ALL	AFR *vs.* ALL	EUR *vs.* ALL	EAS *vs.* ALL	OCE *vs.* ALL	
P1	rs9908046	0.06457	0.02652	0.88886	0.00542	0.86442	0.79758	0.34248	0.34789	0.34217	0.82915	0.0120	0.0141	0.0037	0.7013	
P2	rs721367	0.01858	0.56047	-0.00195	0.63933	0.01520	0.52461	0.55391	0.58906	0.65502	0.51001	0.1182	0.2191	0.3814	0.094	
P4	rs3827760	0.00019	0.59589	0.04546	0.62328	0.04516	0.50109	0.52027	0.54539	0.61314	0.42519	0.2710	0.3210	0.6216	0.1334	
P5	rs4683510	0.09408	0.09727	0.90475	0.05693	0.64312	0.78076	0.30344	0.26404	0.27640	0.75025	0.1096	0.1686	0.3402	0.0347	
P6	rs2139931	0.08686	0.57482	0.11016	0.31478	0.29556	0.77288	0.43296	0.32382	0.45476	0.55445	0.0472	0.0025	0.0121	0.6131	
P7	rs798949	0.42931	0.03585	0.39796	0.34041	0.84381	0.34534	0.27684	0.41085	0.28654	0.55062	0.1194	0.0181	0.2113	0.2966	
P8	rs7832008	0.19633	0.60760	0.96701	0.31289	0.59786	0.16653	0.54116	0.40459	0.47096	0.57383	0.0163	0.1900	0.0249	0.3521	
P9	rs10811102	0.40315	0.80644	0.97584	0.31075	0.39732	0.08562	0.68146	0.46941	0.56857	0.56657	0.3077	0.0526	0.1532	0.3242	
P11	rs2409722	0.06397	0.77305	0.97682	0.73440	0.88476	0.07382	0.75700	0.74882	0.77567	0.76652	0.4667	0.0191	0.1997	0.2143	
P12	rs9809818	-0.00408	0.01674	0.83479	0.01991	0.82319	0.92020	0.45347	0.45615	0.48046	0.89377	0.3172	0.2673	0.3891	0.3277	
P13	rs6886019	0.04001	0.02145	0.91322	0.01559	0.77735	0.85611	0.37991	0.35953	0.37675	0.84619	0.0059	0.0028	0.0369	0.5567	
P14	rs3751050	0.43404	0.67706	0.12544	0.10376	0.52594	0.75731	0.58295	0.38071	0.48533	0.63378	0.0317	0.0003	0.0163	0.5964	
P15	rs7623065	0.03942	0.06385	0.76542	0.04976	0.59439	0.72625	0.21912	0.20476	0.22061	0.68564	0.2715	0.0040	0.1551	0.3750	
P16	rs1509524	0.10799	0.05841	0.69770	0.03745	0.91651	0.87933	0.44005	0.46654	0.46261	0.88481	0.0146	0.0006	0.0128	0.4525	
P17	rs16946159	0.03248	0.15778	0.88053	0.09730	0.78853	0.51392	0.28426	0.26165	0.24610	0.68770	0.0017	0.0439	0.0219	0.4911	
P18	rs2274636	0.10710	0.79366	0.52950	0.60754	0.12758	0.42075	0.65083	0.57870	0.66880	0.44252	0.0640	0.0343	0.0148	0.4276	
P19	rs4704322	0.31973	0.08954	0.40208	0.47905	0.90216	0.25806	0.28682	0.45345	0.36157	0.54877	0.2924	0.1137	0.3739	0.0009	
P20	rs12405776	0.02263	0.10985	0.91102	0.04385	0.79990	0.63907	0.29425	0.26475	0.25230	0.74152	0.0000	0.2442	0.0737	0.2702	
P21	rs6054465	0.46376	0.04909	0.31015	0.55351	0.80269	0.14754	0.35258	0.54696	0.43040	0.51716	0.0543	0.0178	0.0029	0.5731	
P22	rs3784651	0.20872	0.23446	0.97527	-0.00076	0.58333	0.52424	0.32438	0.19455	0.20212	0.63372	0.0077	0.2972	0.1008	0.2450	
P23	rs10970986	0.03921	0.80942	0.26884	0.76850	0.19599	0.59121	0.70955	0.70264	0.76469	0.56702	0.1704	0.0000	0.0041	0.4972	
P24	rs6494411	0.55342	0.55077	0.99188	-0.00364	0.25939	0.20514	0.54462	0.29007	0.31298	0.47403	0.2224	0.2088	0.4954	0.0106	
P25	rs10183022	0.11252	0.76911	0.98332	0.61771	0.74548	0.09872	0.70259	0.65111	0.69815	0.69690	0.4192	0.0098	0.0318	0.2623	
P26	rs10455681	0.13721	0.10137	0.08759	0.03111	0.28763	0.21604	0.13215	0.10446	0.08749	0.20663	0.3556	0.1625	0.3195	0.3085	
P27	rs2184030	0.18269	0.33908	0.12354	0.07884	0.24349	0.35750	0.27885	0.15744	0.22841	0.30186	0.1058	0.0452	0.0422	0.2069	
P28	rs2069945	0.29760	0.55681	0.47761	0.08152	0.77467	0.95538	0.52438	0.43554	0.50028	0.81098	0.1587	0.0119	0.0960	0.2540	
P29	rs11156577	0.09211	0.64421	0.07049	0.73705	0.05323	0.73376	0.54640	0.62144	0.66616	0.51653	0.0829	0.0098	0.1280	0.5126	
P30	rs4892491	0.02204	0.64301	0.52910	0.66922	0.58388	0.12398	0.59440	0.61007	0.65470	0.46647	0.0594	0.2102	0.3975	0.0601	
P31	rs12434466	-0.00579	0.79468	0.16726	0.81147	0.10670	0.71620	0.79370	0.79991	0.84478	0.74574	0.1660	0.2190	0.3380	0.0215	
		0.21997	**0.51669**	**0.69409**	**0.43031**	**0.62661**	**0.56983**	**0.27583**	**0.22636**	**0.39007**	**0.47012**	**0.10158**	**0.06920**	**0.11894**	**0.21922**	**Average F_{ST} / *In* over all 29 *Pacifiplex* loci**
								0.34150	**0.30007**	**0.23460**	**0.11924**	**0.12175**	**0.09886**	**0.07322**	**0.04300**	**Average F_{ST} / *In* over all *34-plex* loci**
								0.31024	**0.26185**	**0.29782**	**0.32662**	**0.11127**	**0.08301**	**0.08887**	**0.12463**	**Average F_{ST} / *In* over all 62 *unique* loci**

generated with the Illumina Human Hap650K microarray. Hap650K SNPs were selected from HapMap data, which does not include Oceanian population samples (Li et al. 2008). More direct bias might arise from the fact that the selected SNPs were chosen from variation recorded in only 27 individuals in two Melanesian populations (Bougainville and New Guinea). Low sample numbers may be sufficient to find a large proportion of the most differentiated SNPs but cannot be representative of population variability across the entire Pacific region. To address this issue it was important to test *Pacifiplex* genotype patterns across a range of Pacific region locations and assess the effectiveness of *Pacifiplex* beyond Melanesia.

4.2.2.3 Study population analyses

As well as study populations, genotypes were obtained from 1000 Genomes Phase III published data to assess the classification efficiency of 62 AIMs for non-reference populations outside of the Pacific region. Note that Phase III version 4 data was used so no X-chromosome SNP genotypes were available. Furthermore, *34-plex* rs1321333 data is not listed in Phase III, so Phase I genotypes were used. Estimated allele frequencies for the study populations are detailed in Table 25. Genetic distances between study and HGDP-CEPH populations shown in **Figure 46A** confirm that 1000 Genomes YRI are very closely related to HGDP-CEPH AFR, CEU to EUR and CHB to EAS, with low numbers of pairwise differences amongst these population comparisons. Pairwise difference analyses of Oceanian study populations are more heterogeneous but indicate patterns of divergence that, with the exception of Easter Islanders, largely correspond with distance from mainland East Asia. First, Philippine Islanders and Taiwanese show higher divergence with HGDP-CEPH Bougainvillean (Melanesian) and PNG samples and this is reflected in the absence of divergence with East Asian reference populations in these two study populations plus Easter Island, Truk Lagoon and East Timor populations. Second, a high number of within-population pairwise differences are seen in Philippine, Easter Island, Turk Lagoon, East Timor and Fiji samples, which could signal a higher level of population admixture in these populations, particularly as divergence levels contrast with the very low within-population divergence of both Papuan samples. Third, Fiji, WA and NT Aboriginal Australian samples reveal relatively low divergence with the reference Oceanian populations used, indicating the AIM-SNPs

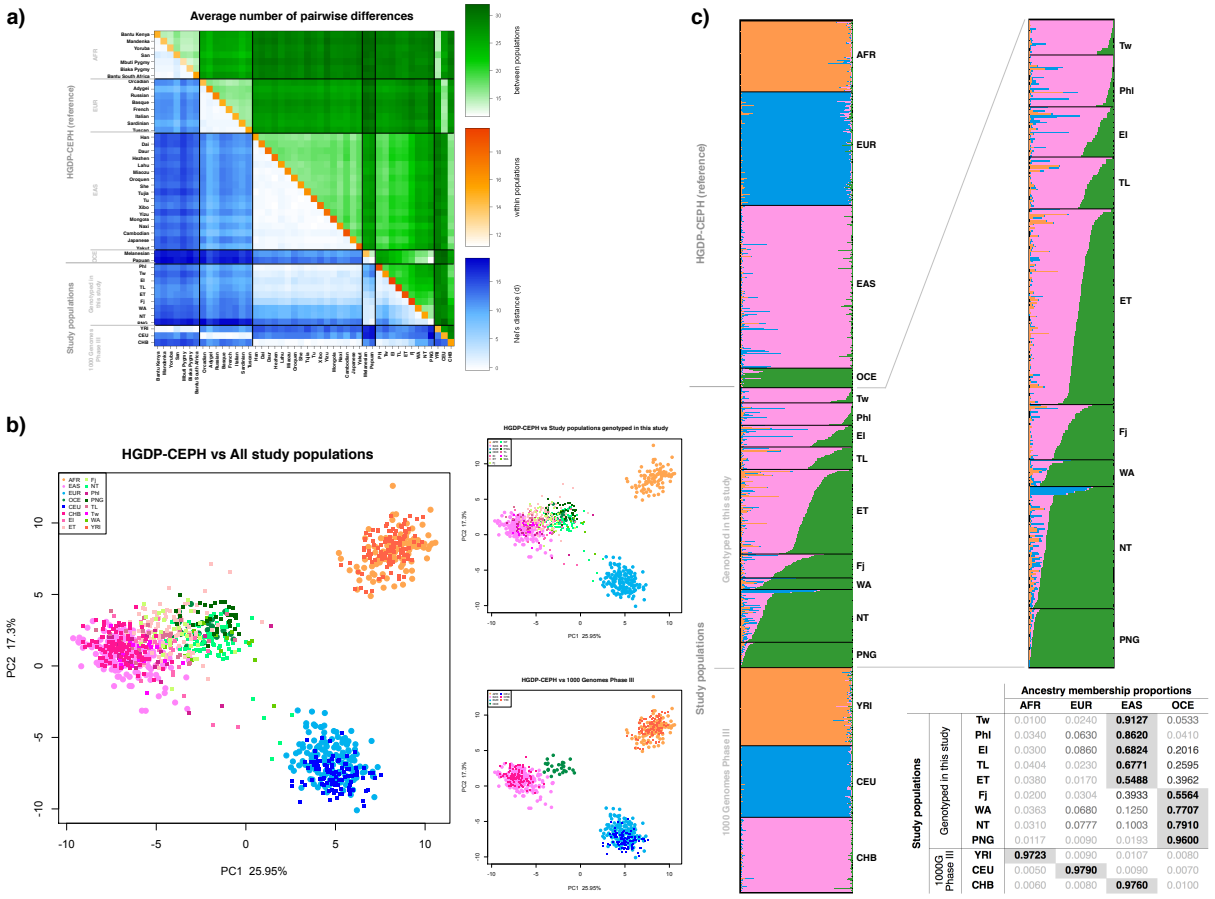
chosen for *Pacifiplex* are informative for Oceanian populations with the closest relationship to the HGDP-CEPH Melanesians used for marker selection.

Table 25: Allele frequencies of *34-plex* and *Pacifiplex* SNPs in nine study populations: Philippines (Phl), Aboriginal Taiwanese (Tw), Easter Island (EI), Truk Lagoon (TL), East Timor (ET), Fiji (Fj), Aboriginal Australians from Western Australia (WA), Aboriginal Australians from Northern Territory (NT) and Papua New Guinea (PNG). SNP rs3827760 common to both assays uses internal code: P4 in *Pacifiplex*, P28 in *34-plex*.

			Alleles described in dbSNP (build 138)		Reference allele frequency									Alternative allele frequency								
	Internal code	rs number	Reference	Alternative	Phl	Tw	EI	TL	ET	Fj	WA	NT	PNG	Phl	Tw	EI	TL	ET	Fj	WA	NT	PNG
SNPforID *34-plex* ancestry forensic test	P11	rs10141763	T	A	0.661	0.595	0.633	0.468	0.513	0.409	0.625	0.678	0.371	0.339	0.405	0.367	0.532	0.487	0.591	0.375	0.322	0.629
	A29	rs1024116	G	A	0.935	0.952	0.867	0.806	0.600	0.545	0.531	0.473	0.200	0.065	0.048	0.133	0.194	0.400	0.455	0.469	0.527	0.800
	P06a	rs10843344	C	T	0.887	0.857	0.817	0.935	0.902	0.844	1.000	0.993	0.914	0.113	0.143	0.183	0.065	0.098	0.156	0.000	0.007	0.086
	P08	rs12913832	A	G	0.952	1.000	0.950	1.000	0.978	0.985	0.969	0.952	1.000	0.048	0.000	0.050	0.000	0.022	0.015	0.031	0.048	0.000
	P03	rs1321333	A	G	0.016	0.000	0.167	0.016	0.091	0.250	0.281	0.178	0.186	0.984	1.000	0.833	0.984	0.909	0.750	0.719	0.822	0.814
	A52	rs1335873	A	T	0.306	0.476	0.017	0.267	0.165	0.333	0.000	0.014	0.000	0.694	0.524	0.983	0.733	0.835	0.667	1.000	0.986	1.000
	P22a	rs1426654	A	G	0.113	0.000	0.200	0.048	0.009	0.061	0.031	0.048	0.000	0.887	1.000	0.800	0.952	0.991	0.939	0.969	0.952	1.000
	P21	rs1498444	A	C	0.710	0.833	0.850	0.710	0.658	0.667	0.656	0.500	0.700	0.290	0.167	0.150	0.290	0.342	0.333	0.344	0.500	0.300
	P13	rs1573020	G	A	0.129	0.071	0.167	0.194	0.097	0.061	0.031	0.021	0.000	0.871	0.929	0.833	0.806	0.903	0.939	0.969	0.979	1.000
	P25a	rs16891982	C	G	0.919	1.000	1.000	1.000	0.968	0.970	1.000	0.966	1.000	0.081	0.000	0.000	0.000	0.032	0.030	0.000	0.034	0.000
	P12	rs182549	C	T	0.984	1.000	0.917	0.984	1.000	1.000	0.969	0.958	1.000	0.016	0.000	0.083	0.016	0.000	0.000	0.031	0.042	0.000
	A13	rs1886510	C	T	0.855	0.905	0.883	0.887	0.884	0.894	1.000	0.945	0.914	0.145	0.095	0.117	0.113	0.116	0.106	0.000	0.055	0.086
	P09a	rs1978806	A	G	1.000	1.000	1.000	0.968	0.991	1.000	1.000	1.000	1.000	0.000	0.000	0.000	0.032	0.009	0.000	0.000	0.000	0.000
	P23	rs2026721	A	G	0.016	0.000	0.000	0.000	0.004	0.000	0.000	0.007	0.000	0.984	1.000	1.000	1.000	0.996	1.000	1.000	0.993	1.000
	A40	rs2040411	G	A	0.565	0.725	0.450	0.484	0.509	0.182	0.094	0.151	0.286	0.435	0.275	0.550	0.516	0.491	0.818	0.906	0.849	0.714
	P15	rs2065160	T	C	0.339	0.190	0.483	0.452	0.538	0.500	0.813	0.726	0.643	0.661	0.810	0.517	0.548	0.462	0.500	0.188	0.274	0.357
	P18	rs2065982	T	C	0.355	0.452	0.400	0.371	0.615	0.727	1.000	0.945	1.000	0.645	0.548	0.600	0.629	0.385	0.273	0.000	0.055	0.000
	P17	rs2303798	G	A	0.806	0.905	0.583	0.661	0.513	0.455	0.531	0.589	0.200	0.194	0.095	0.417	0.339	0.487	0.545	0.469	0.411	0.800
	P01	rs2304925	T	G	0.403	0.286	0.200	0.371	0.227	0.197	0.438	0.658	0.143	0.597	0.714	0.800	0.629	0.774	0.803	0.563	0.342	0.857
	P07	rs239031	C	T	0.161	0.048	0.000	0.129	0.114	0.094	0.094	0.089	0.000	0.839	0.952	1.000	0.871	0.886	0.906	0.906	0.911	1.000
	P16a	rs2572307	T	C	0.016	0.000	0.000	0.000	0.013	0.000	0.000	0.000	0.000	0.984	1.000	1.000	1.000	0.987	1.000	1.000	1.000	1.000
	P04	rs2814778	A	G	0.984	1.000	1.000	0.984	0.911	1.000	1.000	1.000	1.000	0.016	0.000	0.000	0.016	0.089	0.000	0.000	0.000	0.000
	P19	rs3785181	G	A	0.532	0.238	0.400	0.613	0.530	0.621	0.875	0.904	0.557	0.468	0.762	0.600	0.387	0.470	0.379	0.125	0.096	0.443
	P24 *	rs4540055	T	A,G	0.452	0.500	0.633	0.258	0.355	0.197	0.250	0.316	0.143	0.419	0.429	0.250	0.484	0.290	0.212	0.188	0.205	0.100
	P27 *	rs5030240	G	C,T	0.387	0.357	0.250	0.242	0.185	0.106	0.031	0.034	0.000	0.532	0.571	0.400	0.548	0.668	0.697	0.688	0.719	0.900
	P02	rs5997008	C	A	0.984	0.976	0.967	1.000	0.918	0.955	0.875	1.000	1.000	0.016	0.024	0.033	0.000	0.082	0.045	0.125	0.000	0.000
	A21	rs722098	A	G	0.581	0.595	0.583	0.435	0.162	0.381	0.625	0.630	0.414	0.419	0.405	0.417	0.565	0.838	0.619	0.375	0.370	0.586
	P26	rs730570	G	A	0.581	0.619	0.450	0.645	0.823	0.682	0.969	0.966	1.000	0.419	0.381	0.550	0.355	0.177	0.318	0.031	0.034	0.000
	P10	rs773658	C	G	0.065	0.000	0.450	0.113	0.194	0.379	0.156	0.178	0.800	0.935	1.000	0.550	0.887	0.806	0.621	0.844	0.822	0.200
	P05	rs7897550	G	A	0.516	0.619	0.817	0.645	0.814	0.788	0.438	0.664	1.000	0.484	0.381	0.183	0.355	0.186	0.212	0.563	0.336	0.000
	P20	rs881929	G	T	0.210	0.095	0.483	0.581	0.536	0.548	0.781	0.877	0.838	0.790	0.905	0.517	0.419	0.464	0.452	0.219	0.123	0.162
	P14	rs896788	G	A	0.274	0.429	0.567	0.371	0.339	0.364	0.313	0.212	0.329	0.726	0.571	0.433	0.629	0.661	0.636	0.688	0.788	0.671
	A07	rs917118	C	T	0.800	0.750	0.817	0.694	0.688	0.833	0.813	0.788	0.643	0.200	0.250	0.183	0.306	0.312	0.167	0.188	0.212	0.357
Pacifiplex	P1	rs9908046	C	T	0.839	0.976	0.767	0.937	0.612	0.333	0.000	0.123	0.143	0.161	0.024	0.233	0.063	0.388	0.667	1.000	0.877	0.857
	P2	rs721367	G	A	0.306	0.950	0.267	0.589	0.608	0.609	0.844	0.877	1.000	0.694	0.050	0.733	0.411	0.392	0.391	0.156	0.123	0.000
	P5	rs4683510	T	C	0.806	0.810	0.883	0.677	0.668	0.697	0.313	0.301	0.157	0.194	0.190	0.117	0.323	0.332	0.303	0.688	0.699	0.843
	P6	rs2139931	A	G	0.855	0.667	0.800	0.776	0.522	0.591	0.500	0.452	0.114	0.145	0.333	0.200	0.224	0.478	0.410	0.500	0.548	0.886
	P7	rs798949	T	C	0.177	0.119	0.317	0.371	0.466	0.530	0.938	0.973	1.000	0.823	0.881	0.683	0.629	0.534	0.470	0.063	0.027	0.000
	P8	rs7832008	G	A	0.677	0.053	0.738	0.519	0.804	0.924	0.688	0.958	0.985	0.323	0.947	0.262	0.481	0.196	0.076	0.313	0.042	0.015
	P9	rs10811102	G	T	0.484	0.524	0.731	0.423	0.355	0.177	0.125	0.089	0.000	0.516	0.476	0.269	0.577	0.645	0.823	0.875	0.911	1.000
	P11	rs2409722	T	G	0.855	0.905	0.967	0.919	0.927	0.939	1.000	0.979	1.000	0.145	0.095	0.033	0.081	0.073	0.061	0.000	0.021	0.000
	P12	rs9809818	C	A	0.532	0.690	0.700	0.806	0.886	0.833	0.781	0.692	1.000	0.468	0.310	0.300	0.194	0.114	0.167	0.219	0.308	0.000
	P13	rs6886019	C	T	1.000	1.000	0.767	0.919	0.705	0.533	0.469	0.479	0.343	0.000	0.000	0.233	0.081	0.295	0.467	0.531	0.521	0.657
	P14	rs3751050	A	G	0.839	0.357	0.633	0.339	0.645	0.515	0.094	0.089	0.000	0.161	0.643	0.367	0.661	0.355	0.485	0.906	0.911	1.000
	P15	rs7623065	A	G	0.790	0.905	0.683	0.613	0.479	0.394	0.125	0.068	0.014	0.210	0.095	0.317	0.387	0.521	0.606	0.875	0.932	0.986
	P16	rs1509524	A	G	0.661	0.762	0.417	0.769	0.547	0.328	0.406	0.534	0.045	0.339	0.238	0.583	0.231	0.453	0.672	0.594	0.466	0.955
	P17	rs16946159	G	A	0.887	0.833	1.000	0.556	0.564	0.500	0.719	0.564	0.348	0.113	0.167	0.000	0.444	0.436	0.500	0.281	0.436	0.652
	P18	rs2274636	A	G	0.581	0.571	0.467	0.452	0.496	0.500	0.219	0.322	0.157	0.419	0.429	0.533	0.548	0.504	0.500	0.781	0.678	0.843
	P19	rs4704322	C	T	0.145	0.024	0.400	0.242	0.383	0.318	0.750	0.541	0.671	0.855	0.976	0.600	0.758	0.617	0.682	0.250	0.459	0.329
	P20	rs12405776	C	T	0.274	0.143	0.207	0.133	0.177	0.288	0.563	0.563	0.103	0.726	0.857	0.793	0.867	0.823	0.712	0.438	0.438	0.897
	P21	rs6054465	T	C	0.710	0.900	0.850	0.828	0.650	0.391	0.125	0.240	0.000	0.290	0.100	0.150	0.172	0.350	0.609	0.875	0.760	1.000
	P22	rs3784651	T	C	0.306	0.238	0.466	0.145	0.115	0.197	0.000	0.027	0.000	0.694	0.762	0.534	0.855	0.885	0.803	1.000	0.973	1.000
	P23	rs10970986	T	C	0.694	0.619	0.300	0.517	0.397	0.214	0.281	0.062	0.088	0.306	0.381	0.700	0.483	0.603	0.786	0.719	0.938	0.912
	P24	rs6494411	C	T	0.855	0.675	0.717	0.758	0.709	0.788	0.313	0.432	0.544	0.145	0.325	0.283	0.242	0.291	0.212	0.688	0.568	0.456
	P25	rs10183022	G	A	0.613	0.548	0.783	0.710	0.862	0.803	0.969	0.979	1.000	0.387	0.452	0.217	0.290	0.138	0.197	0.031	0.021	0.000
	P26	rs10455681	A	G	0.129	0.167	0.138	0.161	0.098	0.015	0.000	0.042	0.029	0.871	0.833	0.862	0.839	0.902	0.985	1.000	0.958	0.971
	P27 ‡	rs2184030	G	A,C	0.661	0.690	0.667	0.677	0.795	0.682	0.594	0.685	0.757	0.306	0.310	0.317	0.194	0.107	0.061	0.125	0.055	0.114
	P28 ‡	rs2069945	C	A,G	0.435	0.381	0.317	0.383	0.288	0.182	0.125	0.185	0.136	0.161	0.167	0.217	0.167	0.275	0.167	0.219	0.192	0.500
	P29	rs11156577	C	T	0.032	0.048	0.036	0.194	0.327	0.682	1.000	0.986	0.986	0.968	0.952	0.964	0.806	0.673	0.318	0.000	0.014	0.014
	P30	rs4892491	G	A	0.355	0.095	0.196	0.300	0.483	0.485	0.625	0.575	0.971	0.645	0.905	0.804	0.700	0.517	0.515	0.375	0.425	0.029
	P31	rs12434466	A	G	0.274	0.333	0.200	0.242	0.368	0.318	0.563	0.500	0.443	0.726	0.667	0.800	0.758	0.632	0.682	0.438	0.500	0.557
Both	P4	rs3827760	T	C	0.210	0.000	0.500	0.516	0.315	0.697	0.969	0.993	1.000	0.790	1.000	0.500	0.484	0.685	0.303	0.031	0.007	0.000

* For these markers, the third allele was observed with frequency = 1-(reference allele frequency + alternative allele frequency). Third alleles are G and T for P24 and P27, respectively.

‡ For these markers, the third allele was observed with frequency = 1-(reference allele frequency + alternative allele frequency). Third alleles are C and G for P27 and P28, respectively.



Ancestry membership proportions

Study populations			AFR	EUR	EAS	OCE
Genotyped in this study		Tw	0.0100	0.0240	**0.9127**	0.0533
		Phl	0.0340	0.0630	**0.8620**	0.0410
		EI	0.0300	0.0860	**0.6824**	0.2016
		TL	0.0404	0.0230	**0.6771**	0.2595
		ET	0.0380	0.0170	**0.5488**	0.3962
		Fj	0.0200	0.0304	0.3933	**0.5564**
		WA	0.0363	0.0680	0.1250	**0.7707**
		NT	0.0310	0.0777	0.1003	**0.7910**
		PNG	0.0117	0.0090	0.0193	**0.9600**
1000G Phase III		YRI	**0.9723**	0.0090	0.0107	0.0080
		CEU	0.0050	**0.9790**	0.0090	0.0070
		CHB	0.0060	0.0080	**0.9760**	0.0100

Figure 46: Ancestry analyses of the study populations (for this study and from 1000 Genomes Phase III) compared with the four HGDP-CEPH reference populations. a) Genetic distances and pairwise differences within and between populations. b) PCA plot of the first two components (PC1 vs. PC2). Study individual coordinates were calculated from the principal components of reference samples. c) *STRUCTURE* analysis results. Three replicates from *K*=1 to *K*=7 using the admixture POPFLAG ancestry model and correlated allele frequencies (100,000 burnin iterations and 100,000 MCMC repeats). Optimum cluster number was *K*=4. Study populations are ordered from highest to lowest EAS ancestry membership proportions. *STRUCTURE* cluster membership proportions of all study populations are summarised in the table.

Pairwise difference patterns match those observed in the PCA (**Figure 46B**) with PNG and WA/NT Australian samples displaying greater separation from the East Asian cluster than the other study populations. The *R* script used to generate PCA plots allows estimation of individual coordinates of study samples from the principal components calculated for reference samples. The coordinates allow the assessment of relationships between study individuals and reference populations in a similar way to *STRUCTURE* runs using *admixture POPFLAG* parameters. Equivalent distributions are obtained when principal components are estimated for reference and study individuals together and when *STRUCTURE* analysis is performed without prior information on sample origin (outlined in **Figure 47**). Results indicate that the study populations form an ordered distribution of points between EAS and

OCE, while PCA distributions of study populations are not affected whether HGDP-CEPH data is used as a reference or not. Regardless of the regime applied in both genetic clustering approaches, *STRUCTURE* analyses match well with the PCA patterns in all cases, with study populations ranging from high EAS cluster membership proportions (Taiwanese and Philippines) to complete OCE cluster membership in the PNG (individual PCA plots for each of the populations genotyped in this study are outlined in **Figure 48**). The division of OCE study populations revealed in the pairwise difference grids extends to *STRUCTURE* cluster plot patterns, with **Figure 46C** indicating Fijian, WA/NT Australian and PNG have majority OCE cluster membership while the other five have a majority or complete EAS cluster membership.

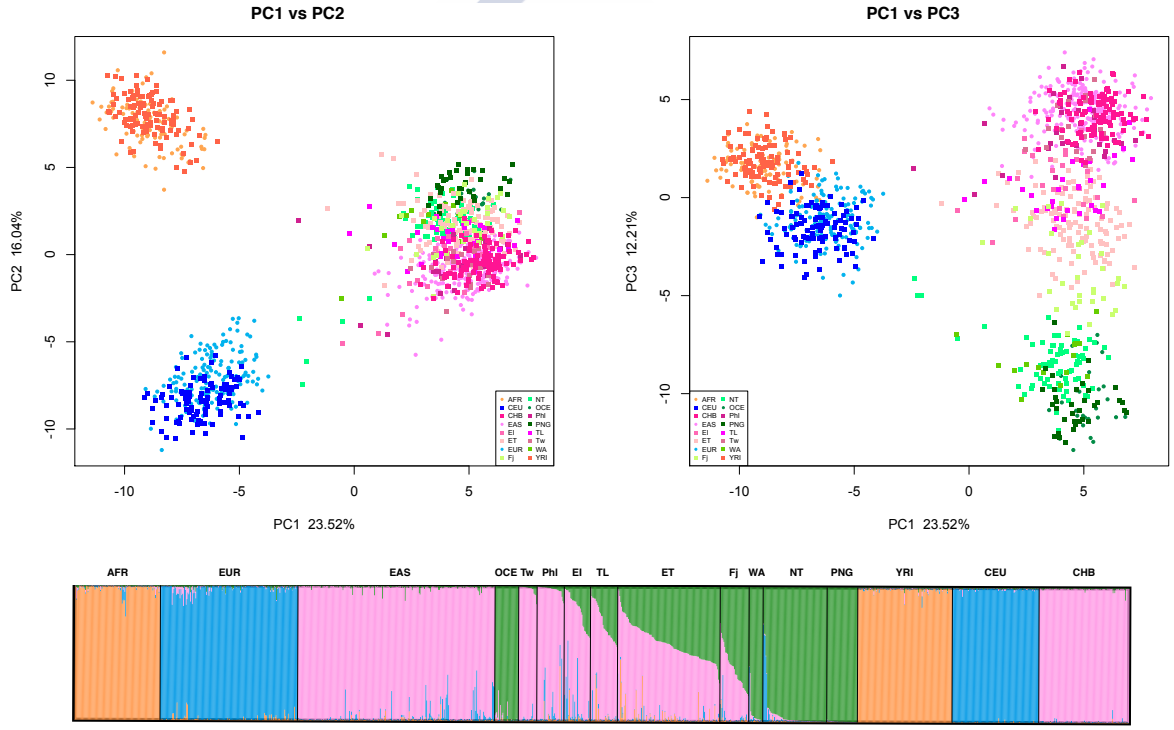


Figure 47: PCA and *STRUCTURE* analysis of HGDP-CEPH and study populations with no prior information assumed. Principal components were estimated for reference and study individuals together with 62 SNP data. Observed distributions are similar to those obtained in **Figure 46** where study individuals coordinates were estimated from HGDP-CEPH principal components. The optimum *STRUCTURE* *K*=4 cluster plot was obtained with runs using analysis parameters as detailed in **Figure 45**.Study populations from 1000 Genomes Phase III: Yoruba in Ibadan (YRI), Utah residents with European ancestry (CEU) and Han Chinese in Beijing (CHB). Study populations genotyped in this study: Philippines (Phl), Aboriginal Taiwanese (Tw), Easter Island (EI), Truk Lagoon (TL), East Timor (ET), Fiji (Fj), Aboriginal Australians from Western Australia (WA), Aboriginal Australians from Northern Territory (NT) and Papua New Guinea (PNG).

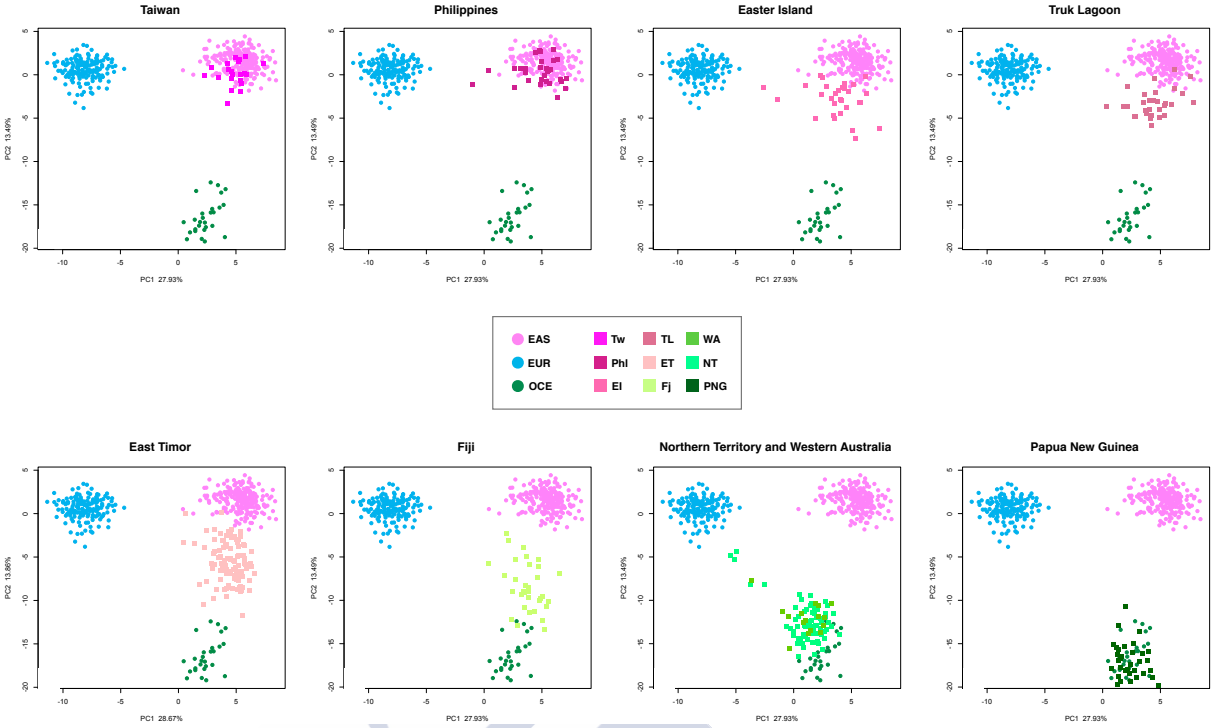


Figure 48: Individual PCA plots for each of the populations genotyped in this study in comparison to three HGDP-CEPH reference groups.

Only four populations have more than 5% European cluster membership overall, where certain individuals have above-average European co-ancestry resulting from admixture. In the case of the NT samples, it was possible to compare the five indicating higher-than-average European co-ancestry (>30% EUR cluster membership, discernible as blue segments in the cluster plots of **Figure 46C**) with mtDNA/Y-marker analyses (personal communication: Nano Nagle, Kaye Ballantyne and Manfred Kayser). Table 26 indicates these five samples have Oceanian mtDNA and Eurasian Y-marker haplogroups and illustrates the risk of misinterpretation of an individual's overall ancestry when analyses are based on uniparental data from atypical lineages.

Table 26: Autosomal SNP European cluster membership proportions (A-EUR) of five NT samples and comparison with the geographic distributions of their mtDNA and Y-marker haplogroups.

A-EUR	mtDNA haplogroup	Y-marker haplogroup
0.7012	Oceanian	European
0.6292	Australian Aboriginal	Eurasian*
0.6270	Australian Aboriginal	European
0.4477	Oceanian	Eurasian*
0.3413	Australian Aboriginal	European

* Eurasian distribution equates to European and Asian regions, but in this Y-haplogroup includes North Africa.

When analysing study populations alone (**Figure 49**) by excluding reference populations, an optimum *K*=2 is obtained in *STRUCTURE* that again correlates with the PCA patterns. PCA plots position Taiwanese and Philippine samples in one extreme and WA/NT Australian plus PNG at the other. Aboriginal Taiwanese individuals can be expected to share much of their genetic ancestry with other East Asian populations, and a previous study showed they cluster together with East Asians and have patterns that are very similar to those reported here (Friedlaender et al. 2008). Furthermore, the same study reported Polynesian and Micronesian populations have an intermediate position between East Asian and Melanesian (Friedlaender et al. 2008) population variation. A higher East Asian component in Polynesians, but also a considerable Melanesian component has been detected in other large-scale studies (Kayser et al. 2008b, Wollstein et al. 2010). Lastly, Fijians have been reported elsewhere to share 58% of their ancestry with New Guinea Highlanders (Wollstein et al. 2010).

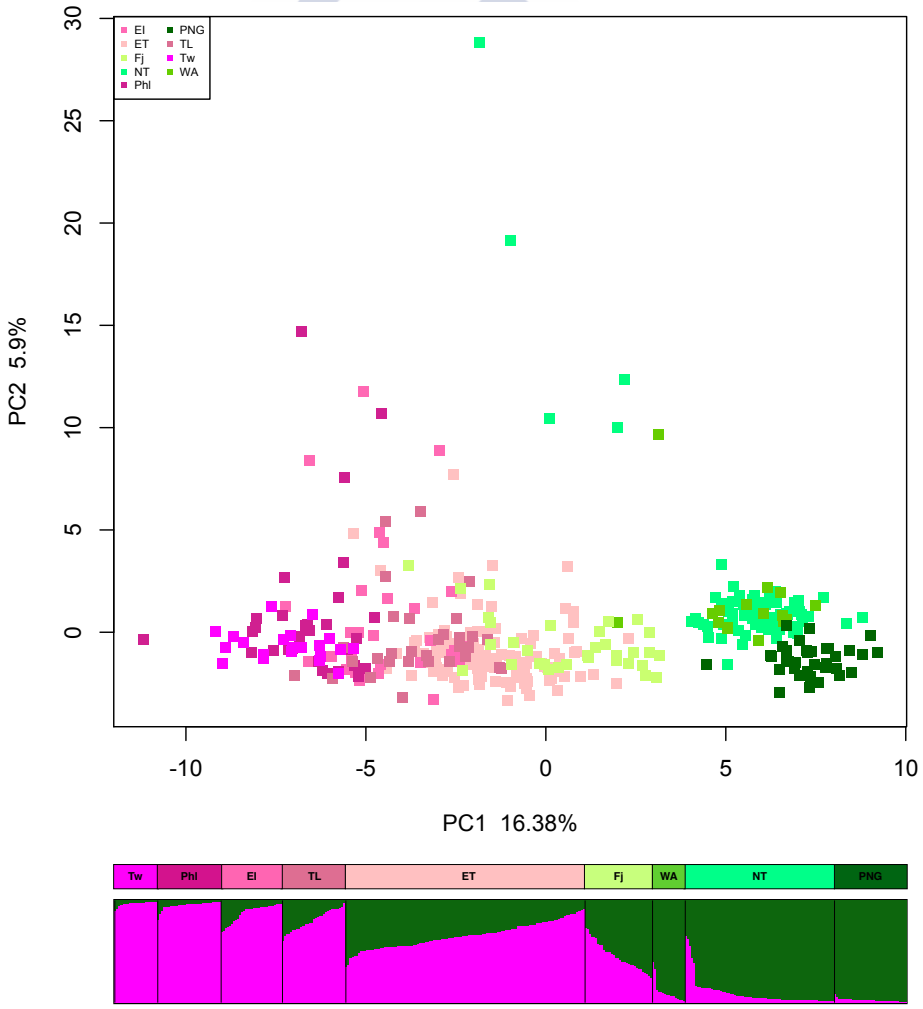


Figure 49: Additional ancestry analyses of study samples with 62 SNPs. Comparison of PCA plot of study populations alone with the optimum *K*=2 *STRUCTURE* cluster plot. Individuals in cluster plot re-ordered to indicate gradient between the two cluster membership proportion distributions across study populations.

Although in comparison to other studies, our population analyses are based on a limited set of SNPs and populations, the strategy to use two Melanesian populations to select AIM-SNPs that address Oceanian variation as a whole, provides an informative AIM-SNP panel. The *Pacifiplex* panel is more effective for Papua New Guinea and Australian Aboriginal samples than populations more divergent from Melanesia, reflecting the Oceanian-wide heterogeneity seen in Micronesia (Truk Lagoon) or Polynesia (Easter Island). In addition, the Aboriginal Taiwanese and Philippine samples have very high East Asian cluster memberships as would be expected from their inferred demographic history and positions between mainland East Asia and the western edge of Oceania.

An important component of this study was to assess the extent to which the HGDP-CEPH Oceanian populations could be an efficient proxy for other Pacific region populations. Results in **Figure 46**, **Figure 47** and **Figure 48** indicate that PNG and NT/WA Aboriginal Australians cluster together and have very similar ancestry membership proportions to the HGDP-CEPH Melanesian populations. For this reason, we consider that the 62 AIM-SNPs examined in this study will be sufficiently informative to differentiate European and Aboriginal Australian ancestries relevant to forensic investigations in Australia. Since population admixture has been a significant feature of the recent history of this continent, reliance on uniparental genetic data alone is prone to the risk of misinterpretation and therefore the *Pacifiplex* and *34-plex* AIM-SNPs assessed here make suitably informative and sensitive marker panels to use alongside mtDNA and Y-marker analyses.

4.2.2.4 Concluding remarks

The *Pacifiplex* assay developed in this study fulfils the three goals that were set for the 29 SNP panel: it successfully enhances an existing 34 AIM-SNP panel by enabling the differentiation of East Asian and Oceanian populations; it provides an informative complementary test to Y-marker and mtDNA analyses when admixture may be a dominant characteristic of the populations studied; and it has proven sensitivity when analysing challenging DNA samples. *Pacifiplex* was able to detect as little as 125 pg of DNA with a high level of genotype completeness, providing a key characteristic for forensic analysis of trace quantities of DNA. The high sensitivity and robustness of *Pacifiplex* and *34-plex* assays was equally important for the population studies as it allowed us to expand the sample

coverage of Oceania by including the GWU population collection of 50-year-old serum samples. In all cases except a proportion of WA Aboriginal Australian samples excluded from this study, *Pacifiplex* and *34-plex* assays were able to successfully genotype the scant DNA we extracted from the serum samples containing very low levels of nucleated cells. Further studies of *Pacifiplex* will focus on the ability of *Pacifiplex* to detect DNA mixtures and provide assessment of their genotype patterns.

The combined 62 AIM-SNPs are able to differentiate ancestry patterns observed in Oceanian populations (such as Aboriginal Australians and Papua New Guineans) from those of populations more closely related to East Asia (Taiwan and Philippines). Enhanced differentiation of Micronesia and Polynesia will be gained when the 62 AIM-SNPs are combined with uniparental data. The benefits of combining different marker types extend to all populations where admixture is a factor. Although, numerous AIM sets designed to differentiate global populations have been described, the development of *Pacifiplex* addresses a gap in forensically optimised AIM sets relevant for casework needs in the Australasian and Pacific regions.

4.2.3 Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing (Appendix 2B)

A simple and informative multiplex was developed for the simultaneous analysis of *46 AIM-Indels* reported to have high δ values between the AFR, EUR, EAS or NAM population groups. All markers were analysed in short fragments (<230 bp) through a single PCR followed by capillary electrophoresis (**Figure 50**). The workflow of the Indel assay is straightforward, reducing considerably the steps and resources needed to genotype a large set of biallelic AIMs.

After optimization of the method we created a database including HGDP-CEPH diversity panel genetic data, commonly used by the research community as reference populations for the four groups AFR, EUR, EAS, NAM and also from Oceania.

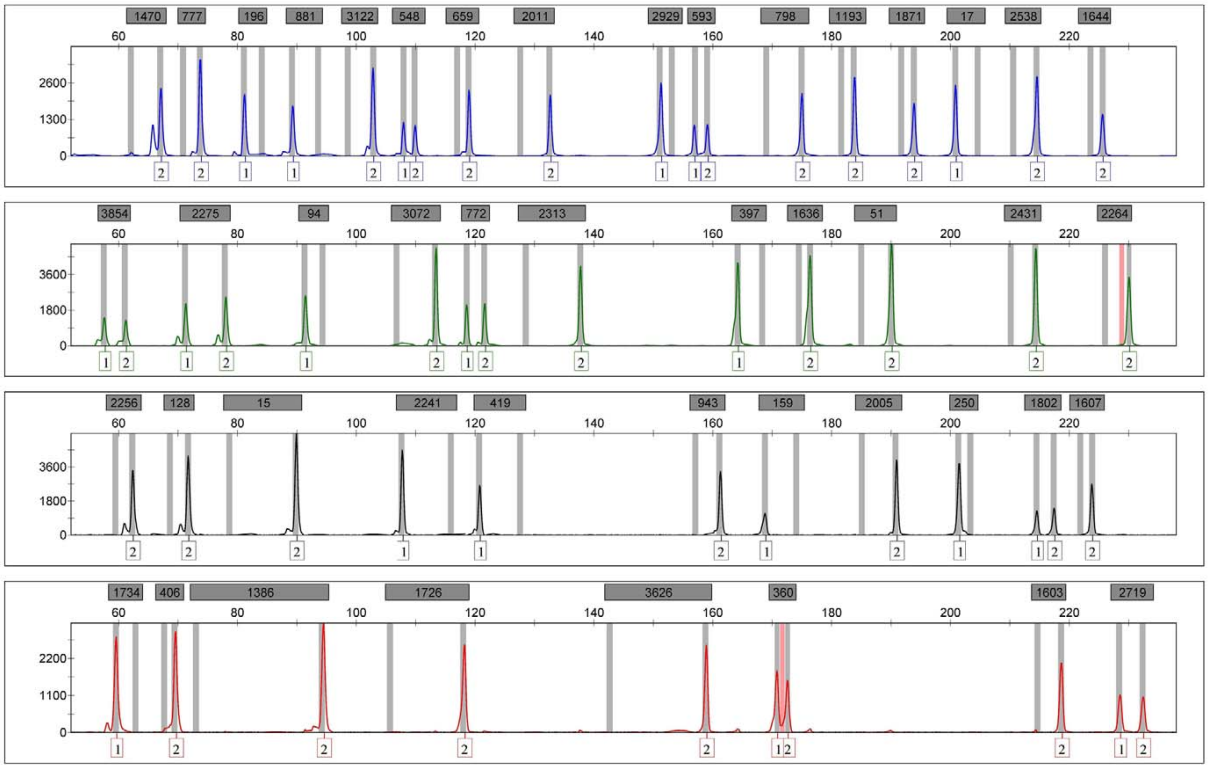


Figure 50: Example of an electropherogram obtained for the HGDP-CEPH 0452 sample with the *46 AIM-Indel* multiplex (markers are identified by MID number).

4.2.3.1 Genetic characterization of reference populations

Patterns of Indel variability observed in the HGDP-CEPH samples from the population groups AFR, EUR, EAS and NAM are detailed in Table 27 as well as δ and pairwise F_{ST} for each marker. With few exceptions, the vast majority of the Indels show high allele frequency differentials and genetic distances between at least two groups (39 with $\delta \geq 0.4$ and 44 with $\delta \geq 0.3$). No significant departures from Hardy-Weinberg equilibrium were found in the studied populations and pairwise linkage disequilibrium exact tests did not detect significant associations within the marker set.

Table 27: Allele frequencies, δ and F_{ST} values for the *46 AIM-Indels* in HGDP-CEPH diversity panel population samples from Africa (AFR), Europe (EUR), East Asia (EAS) and Native America (NAM).

MID	rs number	frequency of allele 1				frequency of allele 2				allele frequency differentials (δ)					
		AFR	EUR	EAS	NAM	AFR	EUR	EAS	NAM	AFR/ EUR	AFR/ EAS	AFR/ NAM	EUR/ ASN	EUR/ NAM	EAS/ NAM
MID-1470	rs2307666	0.162	0.563	0.234	0.031	0.838	0.437	0.766	0.969	0.401	0.072	0.131	0.330	0.532	0.202
MID-777	rs1610863	0.390	0.288	0.550	0.844	0.610	0.712	0.450	0.156	0.103	0.160	0.453	0.262	0.556	0.294
MID-196	rs16635	0.538	0.472	0.550	0.063	0.462	0.528	0.450	0.938	0.067	0.012	0.476	0.079	0.409	0.488
MID-881	rs1610965	0.324	0.880	0.932	0.992	0.676	0.120	0.068	0.008	0.556	0.609	0.668	0.053	0.112	0.060
MID-3122	rs35451359	0.552	0.991	1.000	1.000	0.448	0.009	0.000	0.000	0.438	0.448	0.448	0.009	0.009	0.000
MID-548	rs140837	0.338	0.196	0.478	0.172	0.662	0.804	0.522	0.828	0.142	0.140	0.166	0.282	0.024	0.306
MID-659	rs1160893	0.276	0.104	0.463	0.258	0.724	0.896	0.537	0.742	0.172	0.187	0.018	0.358	0.153	0.205
MID-2011	rs2308203	0.162	0.791	0.876	0.906	0.838	0.209	0.124	0.094	0.629	0.714	0.744	0.084	0.115	0.031
MID-2929	rs33974167	0.790	0.655	0.974	0.977	0.210	0.345	0.026	0.023	0.135	0.183	0.186	0.319	0.322	0.003
MID-593	rs1160852	0.695	0.006	0.002	0.008	0.305	0.994	0.998	0.992	0.689	0.693	0.687	0.004	0.001	0.006
MID-798	rs1610884	0.295	0.642	0.157	0.117	0.705	0.358	0.843	0.883	0.347	0.138	0.178	0.485	0.525	0.040
MID-1193	rs2067280	0.178	0.098	0.341	0.141	0.822	0.902	0.659	0.859	0.080	0.163	0.037	0.243	0.043	0.200
MID-1871	rs2308067	0.081	0.326	0.666	0.289	0.919	0.674	0.334	0.711	0.245	0.585	0.208	0.340	0.037	0.377
MID-17	rs4183	0.795	0.338	0.797	0.898	0.205	0.662	0.203	0.102	0.458	0.002	0.103	0.459	0.561	0.102
MID-2538	rs3054057	0.000	0.494	0.024	0.063	1.000	0.506	0.976	0.938	0.494	0.024	0.063	0.470	0.431	0.038
MID-1644	rs2307840	0.229	0.953	0.251	0.195	0.771	0.047	0.749	0.805	0.724	0.023	0.033	0.701	0.757	0.056
MID-3854	rs60612424	0.748	0.013	0.111	0.008	0.252	0.987	0.889	0.992	0.735	0.636	0.740	0.099	0.005	0.104
MID-2275	rs3033053	0.486	0.095	0.775	0.523	0.514	0.905	0.225	0.477	0.391	0.289	0.038	0.680	0.429	0.252
MID-94	rs16384	0.071	0.155	0.059	0.648	0.929	0.845	0.941	0.352	0.084	0.012	0.577	0.096	0.493	0.589
MID-3072	rs34611875	0.100	0.940	0.996	0.977	0.900	0.060	0.004	0.023	0.840	0.896	0.877	0.056	0.037	0.019
MID-772	rs1610859	0.943	0.968	0.465	0.781	0.057	0.032	0.535	0.219	0.025	0.478	0.162	0.503	0.187	0.316
MID-2313	rs3045215	0.110	0.247	0.587	0.730	0.890	0.753	0.413	0.270	0.137	0.478	0.621	0.341	0.483	0.143
MID-397	rs25621	0.771	0.766	0.655	0.344	0.229	0.234	0.345	0.656	0.006	0.116	0.428	0.111	0.422	0.311
MID-1636	rs2307832	0.148	0.785	0.897	0.992	0.852	0.215	0.103	0.008	0.637	0.750	0.845	0.113	0.207	0.095
MID-51	rs16343	0.057	0.642	0.878	0.867	0.943	0.358	0.122	0.133	0.585	0.821	0.810	0.235	0.225	0.011
MID-2431	rs3031979	0.052	0.111	0.373	0.234	0.948	0.889	0.627	0.766	0.058	0.321	0.182	0.263	0.124	0.139
MID-2264*	rs34122827*	0.162	0.465	0.059	0.133	0.838	0.449	0.941	0.867	0.303	0.103	0.029	0.406	0.332	0.074
MID-2256	rs133052	0.000	0.222	0.039	0.836	1.000	0.778	0.961	0.164	0.222	0.039	0.836	0.182	0.614	0.797
MID-128	rs6490	0.024	0.481	0.079	0.047	0.976	0.519	0.921	0.953	0.457	0.055	0.023	0.402	0.434	0.032
MID-15	rs4181	0.195	0.453	0.290	0.836	0.805	0.547	0.710	0.164	0.257	0.095	0.641	0.162	0.383	0.546
MID-2241	rs3030826	0.814	0.293	0.081	0.070	0.186	0.707	0.919	0.930	0.521	0.734	0.744	0.212	0.223	0.010
MID-419	rs140708	0.919	0.810	0.153	0.563	0.081	0.190	0.847	0.438	0.109	0.766	0.357	0.657	0.248	0.410
MID-943	rs1611026	0.176	0.791	0.223	0.398	0.824	0.209	0.777	0.602	0.615	0.047	0.222	0.568	0.393	0.176
MID-159	rs16438	0.752	0.570	0.128	0.734	0.248	0.430	0.872	0.266	0.183	0.624	0.018	0.441	0.165	0.606
MID-2005	rs2308161	0.029	0.671	0.142	0.325	0.971	0.329	0.858	0.675	0.642	0.113	0.297	0.529	0.345	0.183
MID-250	rs16687	0.824	0.715	0.795	0.328	0.176	0.285	0.205	0.672	0.109	0.029	0.496	0.080	0.387	0.467
MID-1802	rs2307998	0.543	0.003	0.007	0.000	0.457	0.997	0.993	1.000	0.540	0.536	0.543	0.003	0.003	0.007
MID-1607	rs2307803	0.424	0.161	0.260	0.555	0.576	0.839	0.740	0.445	0.262	0.164	0.131	0.098	0.393	0.295
MID-1734	rs2307930	0.852	0.835	0.539	0.375	0.148	0.165	0.461	0.625	0.017	0.313	0.477	0.296	0.460	0.164
MID-406	rs25630	0.048	0.823	0.919	0.703	0.952	0.177	0.081	0.297	0.775	0.871	0.656	0.096	0.120	0.216
MID-1386	rs2307582	0.138	0.218	0.336	0.766	0.862	0.782	0.664	0.234	0.080	0.198	0.628	0.118	0.547	0.429
MID-1726	rs2307922	0.176	0.700	0.643	0.758	0.824	0.300	0.357	0.242	0.524	0.466	0.582	0.057	0.058	0.115
MID-3626	rs11267926	0.124	0.699	0.186	0.039	0.876	0.301	0.814	0.961	0.575	0.062	0.085	0.514	0.660	0.147
MID-360*	rs25584*	0.591	0.864	0.116	0.523	0.317	0.136	0.884	0.477	0.273	0.476	0.068	0.748	0.340	0.408
MID-1603	rs2307799	0.081	0.389	0.546	0.820	0.919	0.611	0.454	0.180	0.308	0.465	0.739	0.157	0.431	0.274
MID-2719	rs34541393	0.448	0.345	0.635	0.469	0.552	0.655	0.365	0.531	0.103	0.188	0.021	0.290	0.124	0.167
									mean:	0.349	0.332	0.380	0.282	0.301	0.215

* For these markers an allele 3 was observed with frequency = 1 – (allele 1 frequency + allele 2 frequency)

Table 27 ***(continued)***

MID	rs number	pairwise F_{ST} (single locus and overall at the bottom)						F_{ST} (single locus and mean at the bottom)			
		AFR/ EUR	AFR/ EAS	AFR/ NAM	EUR/ ASN	EUR/ NAM	EAS/ NAM	AFR *vs* Others	EUR *vs* Others	ASN *vs* Others	NAM *vs* Others
MID-1470	rs2307666	0.282	0.012	0.076	0.206	0.432	0.118	0.148	0.301	0.150	0.244
MID-777	rs1610863	0.020	0.046	0.337	0.128	0.454	0.158	0.091	0.170	0.114	0.290
MID-196	rs16635	0.004	-0.003	0.386	0.010	0.297	0.363	0.066	0.062	0.096	0.316
MID-881	rs1610965	0.504	0.622	0.615	0.014	0.070	0.029	0.645	0.300	0.379	0.336
MID-3122	rs35451359	0.477	0.562	0.391	0.009	0.001	0.000	0.665	0.337	0.393	0.299
MID-548	rs140837	0.048	0.035	0.061	0.155	-0.004	0.168	0.044	0.107	0.130	0.103
MID-659	rs1160893	0.094	0.066	-0.005	0.255	0.083	0.077	0.065	0.182	0.175	0.062
MID-2011	rs2308203	0.562	0.683	0.702	0.024	0.039	-0.001	0.650	0.301	0.394	0.322
MID-2929	rs33974167	0.039	0.192	0.129	0.319	0.231	-0.005	0.112	0.266	0.213	0.153
MID-593	rs1160852	0.728	0.778	0.635	0.000	-0.005	0.000	0.845	0.588	0.638	0.534
MID-798	rs1610884	0.211	0.054	0.080	0.404	0.413	0.001	0.138	0.380	0.257	0.215
MID-1193	rs2067280	0.024	0.058	-0.002	0.146	0.004	0.085	0.044	0.098	0.120	0.053
MID-1871	rs2308067	0.153	0.491	0.143	0.205	-0.002	0.241	0.341	0.174	0.333	0.161
MID-17	rs4183	0.342	-0.004	0.031	0.359	0.456	0.029	0.175	0.397	0.230	0.235
MID-2538	rs3054057	0.446	0.014	0.075	0.486	0.319	0.019	0.328	0.557	0.378	0.256
MID-1644	rs2307840	0.727	-0.002	-0.004	0.654	0.786	0.004	0.391	0.643	0.455	0.393
MID-3854	rs60612424	0.765	0.613	0.694	0.068	-0.004	0.058	0.744	0.476	0.456	0.427
MID-2275	rs3033053	0.332	0.171	-0.005	0.625	0.405	0.138	0.215	0.492	0.431	0.201
MID-94	rs16384	0.028	-0.003	0.563	0.048	0.431	0.649	0.198	0.187	0.266	0.610
MID-3072	rs34611875	0.834	0.927	0.855	0.055	0.009	0.016	0.910	0.714	0.775	0.686
MID-772	rs1610859	0.004	0.376	0.111	0.439	0.191	0.173	0.296	0.346	0.430	0.179
MID-2313	rs3045215	0.055	0.365	0.586	0.207	0.379	0.037	0.311	0.217	0.254	0.297
MID-397	rs25621	-0.004	0.028	0.315	0.026	0.312	0.172	0.071	0.078	0.064	0.254
MID-1636	rs2307832	0.571	0.732	0.813	0.046	0.145	0.053	0.704	0.352	0.458	0.432
MID-51	rs16343	0.516	0.787	0.809	0.147	0.109	-0.005	0.675	0.327	0.476	0.367
MID-2431	rs3031979	0.017	0.224	0.138	0.160	0.054	0.037	0.155	0.119	0.182	0.059
MID-2264	rs34122827	0.213	0.058	-0.002	0.407	0.231	0.034	0.143	0.378	0.265	0.143
MID-2256	rs133052	0.188	0.025	0.870	0.149	0.535	0.841	0.419	0.364	0.465	0.778
MID-128	rs6490	0.394	0.022	0.002	0.357	0.327	0.002	0.250	0.445	0.262	0.213
MID-15	rs4181	0.132	0.020	0.576	0.053	0.250	0.436	0.173	0.120	0.152	0.402
MID-2241	rs3030826	0.422	0.731	0.698	0.147	0.126	-0.004	0.639	0.294	0.435	0.346
MID-419	rs140708	0.042	0.723	0.309	0.607	0.141	0.348	0.533	0.482	0.605	0.306
MID-943	rs1611026	0.544	0.003	0.114	0.486	0.288	0.071	0.286	0.471	0.317	0.173
MID-159	rs16438	0.066	0.589	-0.006	0.367	0.050	0.583	0.372	0.262	0.454	0.337
MID-2005	rs2308161	0.588	0.062	0.302	0.462	0.209	0.101	0.361	0.501	0.336	0.199
MID-250	rs16687	0.028	-0.001	0.411	0.015	0.260	0.380	0.096	0.077	0.111	0.359
MID-1802	rs2307998	0.589	0.635	0.486	-0.002	-0.003	-0.001	0.741	0.430	0.476	0.381
MID-1607	rs2307803	0.158	0.057	0.028	0.025	0.312	0.172	0.091	0.109	0.070	0.192
MID-1734	rs2307930	-0.003	0.187	0.402	0.175	0.387	0.048	0.173	0.178	0.168	0.248
MID-406	rs25630	0.739	0.853	0.665	0.040	0.036	0.179	0.757	0.441	0.550	0.357
MID-1386	rs2307582	0.017	0.090	0.579	0.031	0.461	0.298	0.172	0.133	0.128	0.417
MID-1726	rs2307922	0.425	0.347	0.511	0.005	0.002	0.024	0.383	0.153	0.153	0.161
MID-3626	rs11267926	0.495	0.010	0.037	0.430	0.579	0.073	0.275	0.514	0.301	0.326
MID-360	rs25584	0.136	0.471	0.029	0.720	0.267	0.379	0.343	0.533	0.557	0.273
MID-1603	rs2307799	0.213	0.356	0.727	0.045	0.302	0.139	0.352	0.144	0.190	0.342
MID-2719	rs34541393	0.018	0.066	-0.005	0.153	0.027	0.051	0.040	0.100	0.106	0.032
		0.365	0.393	0.443	0.284	0.298	0.220	0.340	0.311	0.312	0.293

One interesting finding was the occurrence of an unexpected third allelic state (coded as allele 3) for MID360 and MID2264. Sequencing analysis confirmed our observations as a result of additional sequence length variants within the amplicon fragments. For MID360 the third allelic state observed is due to a T insertion associated with the short allele, 8 bases downstream of the targeted polymorphism (allele 1D8Tins). Conversely for MID2264, allele 3 corresponds to a T deletion occurring in the long allele background (allele 2D68Tdel). Interestingly, the MID360 variant alleles were only found in AFR samples whereas the MID2264 variants seemed specific of EUR, further contributing to the differentiation of the two groups.

4.2.3.2 Inferring genetic ancestry

Before implementing the HGDP- CEPH diversity panel reference database, a preliminary evaluation on the performance of a panel comprising 44 Indels at the time (without MID94 and MID1734) had been performed using 48 samples with origin in each of the four groups under study (detailed results in **Figure 51**). In brief, analyses with *STRUCTURE*, PCA and one-out cross validation clearly supported the efficiency of the panel in clustering individuals into four population groups.

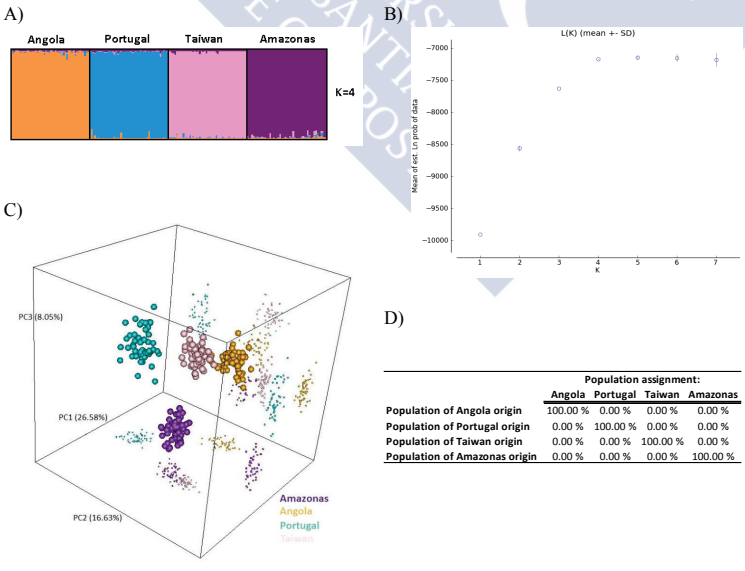
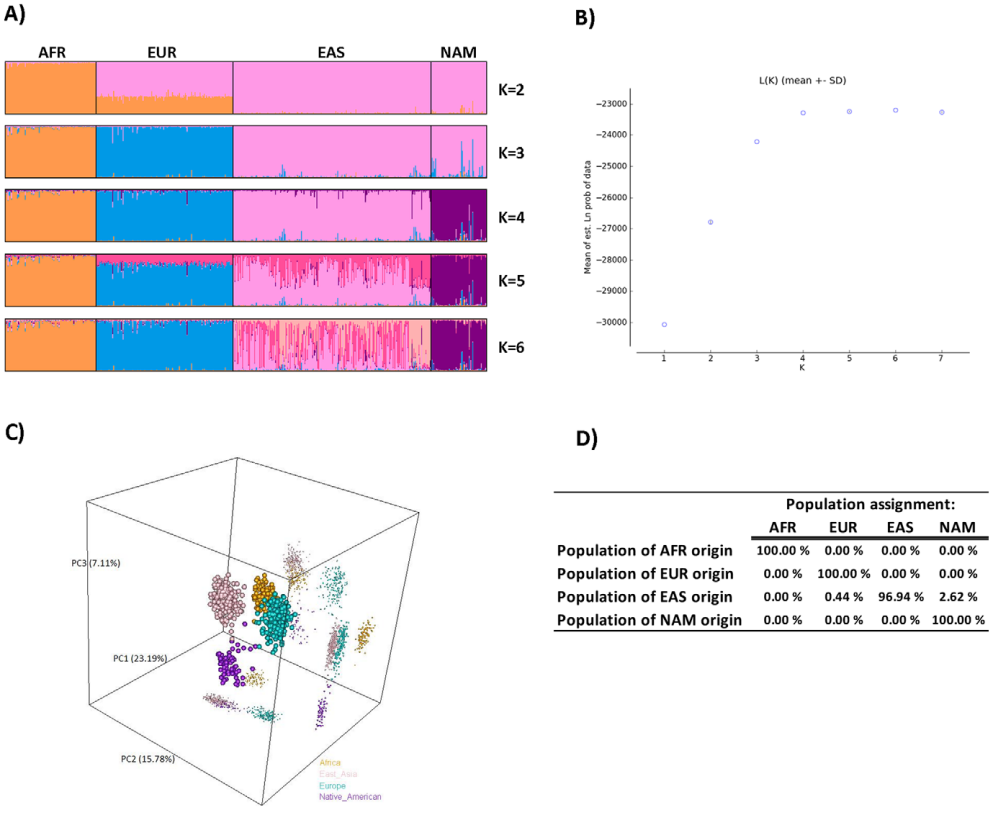


Figure 51: Analysis of population samples from four different continental origins using a preliminary set of 44 AIM-Indels (without MID94 and MID1734): A) ancestral membership proportions (based on *STRUCTURE* results from 3 independent runs treated in *CLUMPP* and plotted with *distruct*); B) estimated ln probability of the data (-lnP(D) obtained with *STRUCTURE* and plotted using *Structure harvester*); C) principal component analysis 3D plots; D) estimation on population assignment success (results from one-out cross validation studies using the *Snipper* app suite; see methods for details on the analyses). Angola (Africa); Portugal (Europe); Taiwan (East Asia); Brazilian Amazonas tribes (Native America).

The results obtained for the complete AIM-Indel panel with HGDP-CEPH AFR, EUR, EAS and NAM populations strongly corroborate these preliminary findings (**Figure 52**). *STRUCTURE* ancestry estimates considering *K*=4 still produce an enhancement in -lnP(D) values while a plateau is reached thereafter, which points to 4 as the smallest *K* number capturing the major population structure in the data and supports the inference that a four group clustering better fits the genetic data (**Figure 52A** and **B**).

PCA for the same dataset allows an independent non-model based view of the individual clustering. The first three PCs define approximately half of the variance in the dataset (46.1%) yet allow a clear spatial separation of four different groups (**Figure 52C**). Likewise, cross-validation studies (**Figure 52D**) revealed the Indel panel to show a high accuracy of population assignment, with a global classification error of 1.26% (specifically 7 of 556). All AFR, EUR and NAM were correctly assigned whereas misclassified individuals were all from the Yakut population in Siberia except for one individual from Oroquen, China.



	Population assignment:			
	AFR	EUR	EAS	NAM
Population of AFR origin	100.00 %	0.00 %	0.00 %	0.00 %
Population of EUR origin	0.00 %	100.00 %	0.00 %	0.00 %
Population of EAS origin	0.00 %	0.44 %	96.94 %	2.62 %
Population of NAM origin	0.00 %	0.00 %	0.00 %	100.00 %

Figure 52: Analysis of HGDP-CEPH diversity panel samples from four continental origins using a set of *46 AIM-Indels*. A) ancestral membership proportions (based on *STRUCTURE* results from 3 independent runs treated in *CLUMPP* and plotted with *distruct*; individuals were first sorted by geographic origin of population, and within those by ascending population code and HGDP individual number); B) estimated ln probability of the data (-lnP(D) obtained with *STRUCTURE* and plotted using *Structure harvester*); C) principal component analysis 3D plots; D) estimation of population assignment success (results from one-out cross validation studies using the *Snipper* app suite; see methods for details of the analyses). AFR: Africa; EUR: Europe; EAS: East Asia; NAM: Native America.

Reference HGDP-CEPH diversity panel genetic data from the four population groups (AFR, EUR, EAS and NAM) was used to estimate ancestry proportions of individuals/populations from different geographic locations. We tested samples from Angola, Portugal, Taiwan and Brazil (Amazonas Amerindian tribes and Belém, a northeastern Amazonas city). The individual and global admixture estimates obtained with genetic data only (no prior population information) correspond well with expected patterns, knowing the origin of the subjects (**Figure 53**; Table 28). In general, individuals from the non-admixed populations show high membership proportion in the same cluster as HGDP-CEPH representatives of the same population group. In contrast individuals from Belém show highly variable admixture patterns mainly of European, Native American and African origin (**Figure 53**; Table 28), resulting in average ancestry proportions of 53.5% EUR, 22.9% NAM, 14.8% AFR and 8.8% EAS. Considering the historical formation and peopling of Brazil in which there were three main contributing ancestral populations (NAM, EUR and AFR) we performed a three-group analysis for the particular case of Belém – specifically, excluding EAS and using only NAM, EUR and AFR ancestral groups with K=3. In particular the Native American proportion increased (53.7% EUR, 29.5% NAM and 16.8% AFR), having captured most of the previous East Asian component.

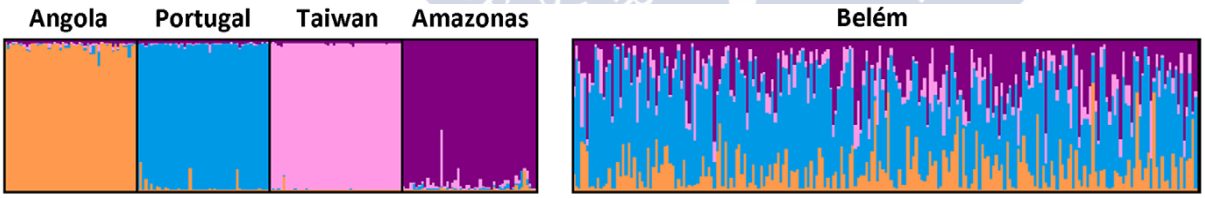


Figure 53: Ancestral membership proportions for testing population samples from different continental origins using the HGDP-CEPH diversity panel genetic data as training sets. Angola (Africa); Portugal (Europe); Taiwan (East Asia); Brazilian Amazonas tribes (Native America); Belém is an example of a highly admixed Brazilian city in northeastern Amazonas.

Table 28: Ancestral membership proportions for HGDP-CEPH diversity panel samples and testing populations from four continental origins (AFR: Africa; EUR: Europe; EAS: East Asia; NAM: Native America).

	46 AIM-INDELs (this study)				210 INDELs [32]				48 I_n4 AIM-SNP set [44]				
	AFR	EUR	EAS	NAM	AFR	EUR	EAS	NAM		AFR	EURA	EAS	AMI
HGDP-CEPH AFR	**0.969**	0.011	0.012	0.008	**0.977**	0.009	0.009	0.005	AFR	**0.97**	0.02	0.01	0.01
HGDP-CEPH EUR	0.008	**0.963**	0.014	0.014	0.007	**0.967**	0.013	0.013	EURA	0.01	**0.96**	0.02	0.01
HGDP-CEPH EAS	0.006	0.018	**0.952**	0.024	0.007	0.021	**0.955**	0.017	EAS	0.01	0.04	**0.91**	0.03
HGDP-CEPH NAM	0.008	0.041	0.027	**0.924**	0.011	0.028	0.015	**0.946**	AMI	0.01	0.03	0.04	**0.92**
Testing populations:	AFR	EUR	EAS	NAM									
Angola	**0.970**	0.011	0.011	0.008									
Portugal	0.018	**0.966**	0.008	0.008									
Taiwan	0.004	0.003	**0.984**	0.009									
Br. Amazonas tribes	0.010	0.013	0.032	**0.945**									
Belém (4G-analysis)	0.148	0.535	0.088	0.229									
Belém (3G-analysis)	0.168	0.537	-	0.295									

The AIM-Indel panel was primarily designed as a tool for ascertaining ancestry from four major population groups. Nonetheless, as there is general interest in AIM panels able to distinguish populations at the broader continental level, we extended our study to HGDP-CEPH Oceanian samples and assessed the ability of the panel to differentiate populations with origin in all five continent regions. Following the same evaluation strategy as before, the assay proved to consistently recognize a fifth cluster corresponding to Oceanians and that *K*=5 captures most of the structure in the dataset (**Figure 54**, see **Figure 55** for details). PCA plots (**Figure 55C**) show most HGDP-CEPH Oceanians form a distinguishable cloud lying between EUR and EAS even though the separation is not perfectly achieved. In a five-group classification, the one-out cross validation error rate increased slightly to 1.54% (9/584). The assignment of Oceanians was accurately made but two EAS (from Cambodia) were now misclassified as OCE.

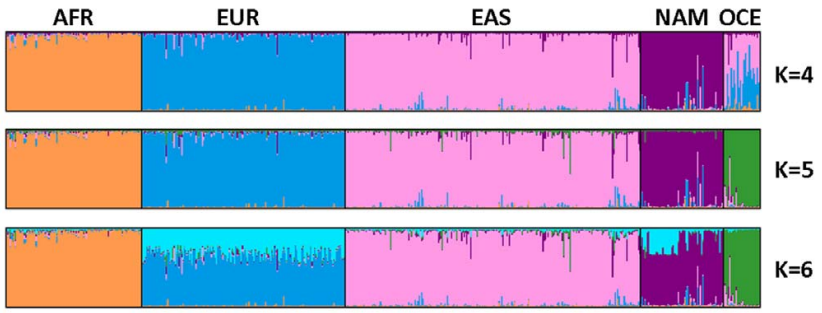
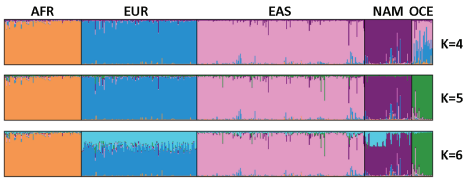
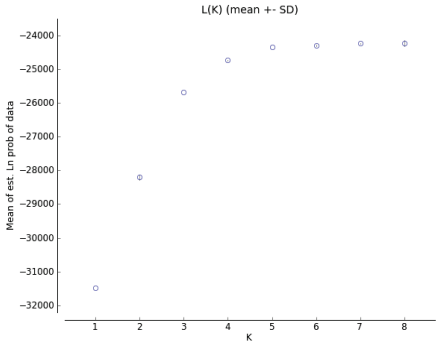


Figure 54: Ancestral membership proportions for HGDP-CEPH diversity panel samples from five continental origins using a set of *46 AIM-Indels* (based on *STRUCTURE* results from 3 independent runs treated in *CLUMPP* and plotted with *distruct*; individuals were first sorted by geographic origin of population, and within those by ascending population code and HGDP individual number). AFR: Africa; EUR: Europe; EAS: East Asia; NAM: Native America; OCE: Oceania.

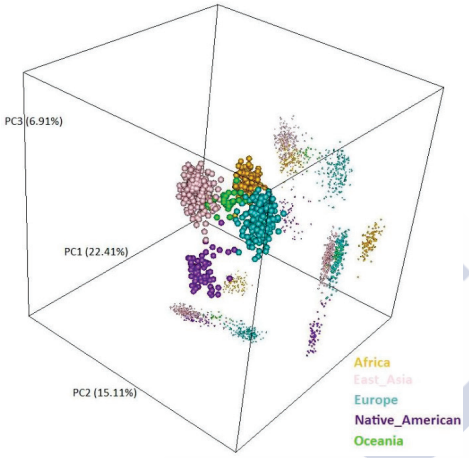
A)



B)



C)



D)

	Population assignment:				
	AFR	EUR	EAS	NAM	OCE
Population of AFR origin	100.00 %	0.00 %	0.00 %	0.00 %	0.00 %
Population of EUR origin	0.00 %	100.00 %	0.00 %	0.00 %	0.00 %
Population of EAS origin	0.00 %	0.44 %	96.07 %	2.62 %	0.87 %
Population of NAM origin	0.00 %	0.00 %	0.00 %	100.00 %	0.00 %
Population of OCE origin	0.00 %	0.00 %	0.00 %	0.00 %	100.00 %

Figure 55: Analysis of HGDP-CEPH diversity panel samples from five continental origins using a set of *46 AIM-Indels*: A) ancestry membership proportions (estimated based on *STRUCTURE* results from 3 independent runs treated in *CLUMPP* and plotted with *distruct*; individuals were first sorted by geographic origin of population. and within those by ascending population code and HGDP individual number); B) estimated ln probability of the data (-lnP(D) obtained with *STRUCTURE* and plotted using *Structure harvester*); C) principal component analysis 3D plots. D) estimation on population assignment success (results from one-out cross validation studies using the *Snipper* app suite; see methods for details on the analyses). AFR: Africa; EUR: Europe; EAS: East Asia; NAM: Native America; OCE: Oceania.

4.2.3.3 Discussion

The main objective of this study was to provide a simple tool for inferring ancestry and estimate admixture proportions from four different population origins that can be widely applied to genetic studies. We describe a new AIM assay comprising 46 Indels that are simply analysed in a multiplex PCR followed by CE detection. With this approach we were able to combine the ancestry informative power of biallelic markers (exemplified by AIM-SNP panels) with the simplified analysis based in fragment size separation (as in STR typing). The methodology of the assay is straightforward and can be readily and inexpensively implemented in any molecular genetics laboratory. In contrast, the majority of AIM sets published in recent years involve more complex genotyping protocols or are limited to specific platforms not available to all laboratories and therefore requiring additional resources (Lao et al. 2006, Phillips et al. 2007b, Halder et al. 2008, Kersbergen et al. 2009, Kosoy et al.

2009, Nassir et al. 2009, Lao et al. 2010, Paschou et al. 2010, Santos et al. 2010, Kidd et al. 2011). Another important aspect is that some AIM sets are directed to differentiate specific population groups depending on the main ancestral contributors to the individuals or populations under investigation (Phillips et al. 2007b, Santos et al. 2010, Tandon et al. 2011). We aimed to develop a generic panel, designed to target the four major population groups of AFR, EUR, EAS and NAM, similarly to Halder et al. (2008). Our objective was to balance combining the highest number of AIMs possible into a single reaction with use of amplicon lengths suitable for the analysis of low quality DNA. The limitation of large multiplex reliability restricts the maximum number of markers in a single reaction. On the other hand, AIMs have an important application in forensic investigations where the quantity and quality of the samples are often limiting factors. We were eventually able to multiplex 46 highly informative Indels, with a scope of markers comparable to other AIM sets reported (Phillips et al. 2007b, Kersbergen et al. 2009, Lao et al. 2010, Paschou et al. 2010, Santos et al. 2010). Kosoy et al. (2009) have shown that small AIM sets can distinguish major population groups and correct for false positive results in association studies. Other studies have addressed ancestry prediction of the HGDP-CEPH samples using large-scale SNP datasets obtained with high-throughput microarrays, and have also evaluated the performance of small subsets of markers ascertained following different strategies such as F_{ST}, allele differentials δ, informativeness of assignment index *In* (Rosenberg et al. 2003) or PCA (Lao et al. 2006, Biswas et al. 2009, Paschou et al. 2010). These studies have shown that inference of continental ancestry for the HGDP-CEPH panel is quite clear, and can be performed with a relatively small number of SNPs (10 to 50). They also showed that, when using SNPs, it is possible to predict individual ancestry down to the population level, although such approaches require an increased number of markers ranging from several hundred to thousands (Biswas et al. 2009, Paschou et al. 2010). Due to the multiplexing limitation associated with the number of markers that it is possible to analyse in a single PCR, and in the same respect for other small-scale AIM sets, the *46 AIM-Indel* assay we outline is only going to be particularly useful when broad assignment to continental ancestry is desired, or when estimating admixture proportions in individuals/populations that received ancestral contributions of different continental origins. Assessing within-continent population structure requires much larger arrays of markers, well beyond the number included in our set and in most of the alternative AIM sets, and therefore it will have limited application for that purpose.

The AIM-Indel assay allowed a rapid and cost-effective genotyping of a large number of samples including HGDP-CEPH individuals from five continental groups (AFR, EUR, EAS, NAM and OCE) and representative testing samples with different origins and admixture levels. From the genetic characterization of the reference ancestral samples we observed a high level of differentiation from the chosen Indels, as expected from the selection criteria. Although some markers revealed lower differences than those expected from previous data, this is possibly due to the samples representing each group and allele frequency estimation strategies being different (Weber et al. 2002). The pairwise F_{ST} values obtained with the *46 AIM-Indels* (Table 27) are clearly above the usually found at the continental level with random markers (Rosenberg et al. 2002, Nassir et al. 2009) and similar to those obtained with other AIM panels for the same population groups (Kersbergen et al. 2009, Kosoy et al. 2009, Nassir et al. 2009).

The results from the HGDP-CEPH diversity panel and other representative populations underlined the capacity of the panel to distinguishing four continental population groups. Furthermore, the ancestry estimates obtained in a four-group analysis are very similar to those obtained in Kosoy et al. (2009) with a 48 *In4* AIM- SNP set for equivalent population groups (Table 28), as well as using a much larger number of INDELs (210) for the same HGDP-CEPH individuals (Table 28; (Rosenberg et al. 2005)). This concordance in the ancestry estimates highlights the accuracy of the AIM-Indel panel in inferring ancestry proportions from African, European, East Asian and Native American origin. Furthermore, in spite of the assay being primarily designed for studies considering only four major population groups, extension to five groups revealed the capacity to reliably distinguish Oceanians.

The population assignment cross validation studies based on Bayesian likelihood ratios provided additional evidence of the utility of the assay, particularly for forensic applications where single profiles are often analysed one at a time. Here the error rates in classifications considering either four or five population groups were low (1.26% and 1.54% respectively). The AIM-Indel panel achieves very high accuracy for population assignment in the five broad continental regions, similar to results observed by Paschou et al. when using subsets of 50 SNPs ascertained by PCA and estimation of *In* metrics (Paschou et al. 2010). In our study, the great majority of misclassified individuals were from a single population (Yakut of eastern Siberia) localized near the northeastern fringe of the Asian continent. This intermediate position between East Asia and the American continent can explain differences in patterns of

divergence between individuals and their misclassification as American. Likewise, the cross validation studies with five groups revealed two misclassified Cambodians as Oceanians. Together, these results suggest a weaker performance of the panel with differentiation of East Asians. In fact, the accumulated divergence assessed for EAS vs. others is slightly smaller than for the other ancestral groups, and the fact that the HGDP-CEPH EAS group analysed is so diverse (229 individuals from 18 subpopulations) may contribute to this reduced differentiation for East Asians. Another important aspect is the proximity of the East Asian and Native American gene pools. Considering the history of modern humans these groups have diverged over the shortest time, and furthermore, the original peopling of Americas from Beringia involved a significant bottleneck effect that is still reflected in Native American variability. Despite this slightly reduced level of differentiation in the AIM-Indels selected, *STRUCTURE*, PCA and cross validation studies together support the capacity of the panel to properly distinguish both groups.

AIM panels are regularly applied in population genetics studies to analyse admixed populations by estimating admixture proportions both at the individual and population level. Depending on the historical context of populations under study, there are different principal ancestral contributors to the formation of the current ancestry characteristics of the region. For example, Brazil and the majority of south-American countries underwent admixture between the pre-existent Native Americans, colonizing Europeans and later African influences resulting from the slave trade to create essentially tri-hybrid populations. In such cases, it is appropriate for genetic studies to perform three-group analyses of ancestry estimates. Our study analysed ancestry proportions in Belém. We first considered the possibility of a fourth EAS minor ancestral contributor in initial analyses and K=4 resulted in a low level but detectable fraction of membership of this cluster at 8.8% (Table 28). However, although not statistically significant (exact test of differentiation p-value=0.136), the three-group membership proportion estimates at K=3 showed a noticeable increase in the Native American component to 29.5% (Table 28) which is in very close agreement with the admixture proportions previously reported for the same population but using a different set of AIM-Indels (average NAM estimate: 28.4%; (Santos et al. 2010)). Nevertheless, a preliminary four-group analysis has persuasive arguments for considering all four potential contributors to admixture in these regions. In particular, some locations in Brazil (e.g. São Paulo, Campinas; IBGE – Instituto Brasileiro de Geografia e Estatística, www.ibge.gov.br)

include significant East Asian communities, despite having joined these populations rather recently. When using Brazilian samples from such geographic areas, particularly as case and control samples for association studies, a preliminary four-group analysis is recommended to detect the presence of East Asian ancestry amongst individuals in the study. Otherwise there is considerable risk that the a priori rejection of this hypothesis based on three-group analyses could lead to an over-estimation of the Native American proportion in the global admixture estimates (data not shown) due to a strong bias caused by the presence of East Asian individuals in the population under study. Conversely, when "forcing" a four-group analysis in south-American tri-hybrid populations, it is possible that the fourth East Asian component can produce a spurious fraction of membership arising from the Native American component, due to the close relationship of the East Asian and Native American population groups. In summary, we advocate adopting an approach taking due regard for the particular population under study. Consideration of the known recent population history and demographics helps make appropriate adjustment for the different principal ancestral contributors. In the special case of south-American populations, we recommend a preliminary study taking advantage of the full potential of the AIM-Indel assay to identify and possibly exclude East Asian study subjects, and subsequently perform a comprehensive three- group analysis. The AIM-Indel assay can be efficiently used in three-group analyses AFR/EUR/NAM, similarly to Santos et al. (2010) and also AFR/EUR/EAS, as in Phillips et al. (2007b). Nonetheless, the reliability of the four-way analysis we repeatedly achieve with this multiplex allows a clear distinction of all groups.

4.2.4 Completion of a worldwide reference panel of samples for an ancestry informative Indel assay (Appendix 2C)

4.2.4.1 Completing the HGDP-CEPH reference database

Genotypes of 365 ME and CSA HGDP-CEPH individuals for the *46 AIM-Indels* were generated. This new data, together with the existing data in Pereira et al. (2012b), completes the HGDP-CEPH H952 subset (note three individuals absent in the original study: 949 samples listed here). ME and CSA population allele frequency estimates are presented in Table 29. No deviations from HWE were observed (all *p*-values ≥0.01471). Two significant

Table 29: Allele frequency estimates of the *46 AIM-Indels* for the HGDP-CEPH ME and CSA populations. Short allele = 1; long allele = 2.

MID	rs number	Allele 1 frequency		Allele 2 frequency	
		ME	CSA	ME	CSA
MID-1470	rs2307666	0.546	0.411	0.454	0.589
MID-777	rs1610863	0.322	0.500	0.678	0.500
MID-196	rs16635	0.494	0.411	0.506	0.589
MID-881	rs1610965	0.804	0.760	0.196	0.240
MID-3122	rs35451359	0.948	0.993	0.052	0.007
MID-548 *	rs140837	0.163	0.134	0.837	0.854
MID-659	rs1160893	0.092	0.153	0.908	0.847
MID-2011	rs2308203	0.696	0.733	0.304	0.267
MID-2929	rs33974167	0.571	0.752	0.429	0.248
MID-593	rs1160852	0.104	0.017	0.896	0.983
MID-798	rs1610884	0.552	0.446	0.448	0.554
MID-1193	rs2067280	0.156	0.252	0.844	0.748
MID-1871	rs2308067	0.248	0.446	0.752	0.554
MID-17	rs4183	0.313	0.537	0.687	0.463
MID-2538	rs3054057	0.377	0.228	0.623	0.772
MID-1644	rs2307840	0.896	0.800	0.104	0.200
MID-3854	rs60612424	0.126	0.087	0.874	0.913
MID-2275	rs3033053	0.218	0.250	0.782	0.750
MID-94	rs16384	0.236	0.252	0.764	0.748
MID-3072	rs34611875	0.831	0.936	0.169	0.064
MID-772	rs1610859	0.923	0.874	0.077	0.126
MID-2313	rs3045215	0.258	0.473	0.742	0.527
MID-397	rs25621	0.767	0.700	0.233	0.300
MID-1636	rs2307832	0.791	0.772	0.209	0.228
MID-51	rs16343	0.571	0.577	0.429	0.423
MID-2431	rs3031979	0.071	0.139	0.929	0.861
MID-2264 *	rs34122827	0.426	0.431	0.512	0.535
MID-2256	rs133052	0.083	0.149	0.917	0.851
MID-128	rs6490	0.414	0.374	0.586	0.626
MID-15	rs4181	0.534	0.376	0.466	0.624
MID-2241	rs3030826	0.310	0.277	0.690	0.723
MID-419	rs140708	0.831	0.693	0.169	0.307
MID-943	rs1611026	0.687	0.639	0.313	0.361
MID-159	rs16438	0.652	0.595	0.348	0.405
MID-2005	rs2308161	0.613	0.399	0.387	0.601
MID-250	rs16687	0.730	0.599	0.270	0.401
MID-1802	rs2307998	0.055	0.032	0.945	0.968
MID-1607	rs2307803	0.175	0.198	0.825	0.802
MID-1734	rs2307930	0.831	0.765	0.169	0.235
MID-406	rs25630	0.660	0.750	0.340	0.250
MID-1386	rs2307582	0.199	0.292	0.801	0.708
MID-1726	rs2307922	0.549	0.765	0.451	0.235
MID-3626	rs11267926	0.583	0.631	0.417	0.369
MID-360 *	rs25584	0.650	0.691	0.340	0.309
MID-1603	rs2307799	0.347	0.334	0.653	0.666
MID-2719	rs34541393	0.316	0.408	0.684	0.592

* For these markers, allele 3 was observed with frequency = 1-(allele 1 frequency + allele 2 frequency).

results for pairwise LD tests were observed applying a Bonferroni corrected α=0.00022. Pairs showing significant association were: MID-3854:MID-1802 in ME, and MID-3854:MID-1734 in CSA. Notably, MID-3854 and MID-1734 is the closest same-chromosome pair in the set, separated by 459 kilobases, but neither pair showed significant LD results in the other six population groups studied (Pereira et al. 2012b). Such results collectively point towards an absence of real associations between the Indel markers at the population level, permitting a conventional *STRUCTURE* cluster analysis of our data without the need for a linkage model (Falush et al. 2003).

Pairwise F_{ST} genetic distances are presented in Table 30. As previously observed with AIM-SNPs chosen for forensic analysis (Phillips et al. 2007b, Lao et al. 2010), these Indels selected for the same purpose display much lower within population-group diversity levels than between group diversity (see **Figure 56**) indicating the major proportion of variation detected can efficiently differentiate individuals with different populations of origin. However, it is noticeable that the pairwise difference heat maps for both ME and CSA show levels of within population group diversity higher than other groups, relating, in part, to higher levels of population admixture in both regions.

Table 30: Pairwise F_{ST} values comparing seven HGDP-CEPH population groups (African, European, East Asian, Native American and Oceanian genotype data previously published in (Pereira et al. 2012b) used to perform comparisons with ME and CSA). Diagonal grey cells represent zero value output from Arlequin (within-population analyses not made). Plus signs (+) above the diagonal indicate distances between population pairs are significant in each case (p-value < 1×10^{-5}).

	AFR	EUR	ME	CSA	EAS	NAM	OCE
AFR		+	+	+	+	+	+
EUR	0.36515		+	+	+	+	+
ME	0.28955	0.01485		+	+	+	+
CSA	0.28922	0.03533	0.02735		+	+	+
EAS	0.39289	0.28406	0.25193	0.18104		+	+
NAM	0.44273	0.29768	0.26519	0.19707	0.21990		+
OCE	0.37447	0.22520	0.17420	0.16655	0.22480	0.30981	

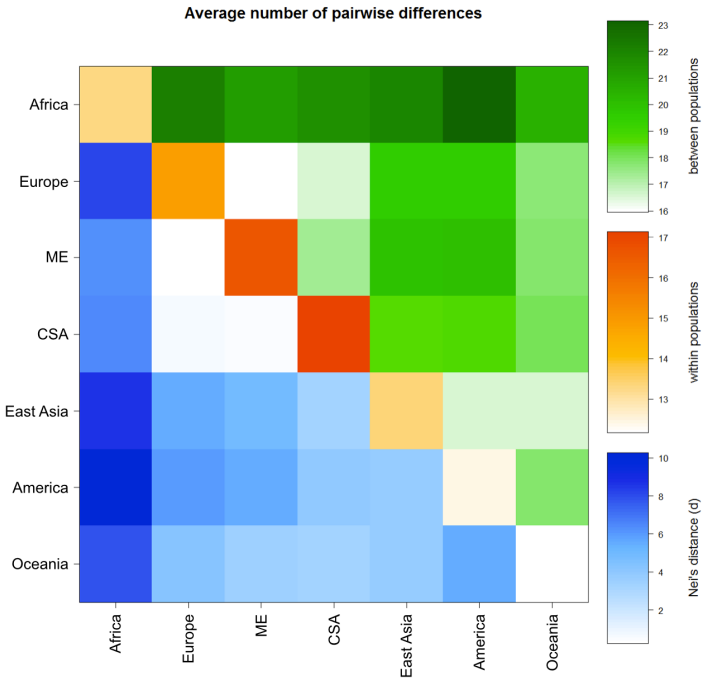


Figure 56: Pairwise difference heat maps for between and within population F_{ST} values plus Nei's distance: d, for the HGDP-CEPH 7 population groups.

4.2.4.2 Characterization of third alleles in component Indel MID-548

Five CSA individuals (HGDP00131, 137, 139, 165, and 187) showed an atypical heterozygous genotype for MID-548 (confirmed in singleplex amplification; **Figure 57**), with a novel short allele 3 basepairs (bp) smaller than the expected short allele. Scrutiny of MID-548 context sequence in dbSNP (the NCBI database of short genetic variations), revealed a neighbor insertion deletion polymorphism (rs375726490) that lacked frequency data and validation information. Comparing the dbSNP context sequences of rs375726490 and MID-548 indicated a small portion of discordant bases (**Figure 58**), but no discordances between our sequencing reads and dbSNP MID-548 were found. For this reason, using another neighbor SNP (rs74656636) as a sequence anchor, we cross-checked the chromosomal region in dbSNP and 1000 Genomes. In both cases the sequence matched that of MID-548, and the discordant bases in the dbSNP sequence of rs375726490 were absent (**Figure 58**). Due to the close proximity of each allele, it was not possible to achieve adequate band separation in T_9C_5 or $T_{14}C_{10}$ polyacrylamide gels before sequencing. Therefore, sequences show the characteristic heterozygous pattern downstream of the point where the novel deletion occurs (**Figure 58**). All five samples with this allele were sequenced and gave identical sequencing

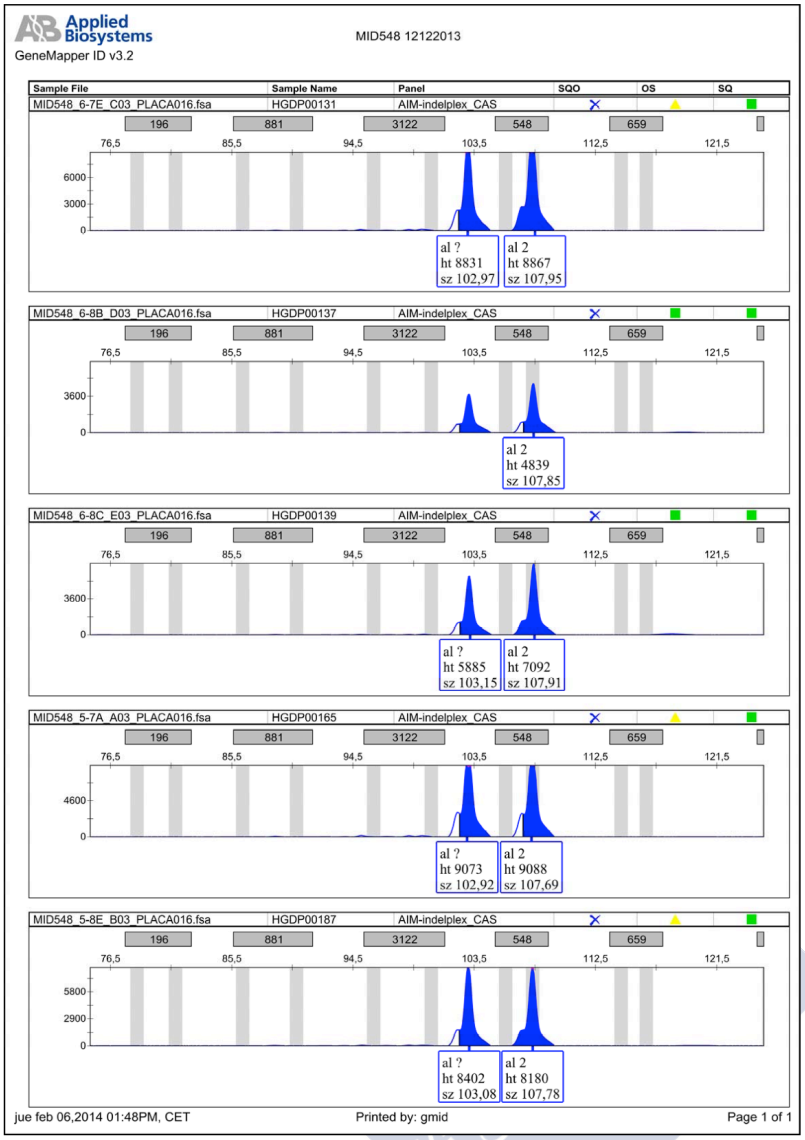


Figure 57: MID-548 third allele characterization. Singleplex results confirmed that a third allele in MID-548 was present in five Central South Asian HGDP-CEPH samples.

patterns matching those expected. This confirms the existence of a 5bp deletion in the context sequence of the MID-548 long allele and is likely to be the rs375726490 polymorphism listed in dbSNP. In the HGDP-CEPH H952 panel, the MID-548 third allele was only observed in CSA populations. Despite a 1.2% frequency (Table 29), marker discrimination power is increased for differentiating CSA populations.

MID-548 dbSNP Fwd sequence
Validated by cluster and sequenced in 1000Genomes
AGTCAGGACTGAAGAAACCCCTCTGCCCAGGCATGCATTAAATTAAATTGGGCGGAGCAGCAGGTCCTGTCTCTGCCAAA(-/CT)CTGTCCACACGGGCTCTTGTTTTACTG

rs375726490 dbSNP Fwd sequence
GGGAAGTCAAGGGTAAGTCAGGACTGAAGAAACCCCTCTGCCCAGGCATGC (-/ATTAA)TTAAATTAAATTGGGCGGAGCAGCAGGTCCTGTCTCTGCCAAACTCTGTC
Note: grey C is missing and light blue TTAA are extra in rs375726490 (in comparison with MID-548)

Using rs74656636 as anchor to inspect the area in 1000Genomes and dbSNP.

1000Genomes TAGAGAAGTCACTCCTGAAACCCACAGAGTGTAGCCGGGAAGTCAAGGGTA**AGTCAGGACTGAAGAAACCCCTCTGCCCAGGCATGCATTAAATTAAATTGGGCGGAGCAGCA**
GGTCCTGTCTCTGCCAAACTCTGTCCACACGGGCTCTTGTTTTACTGGTGAACCCACATGCTGCACCCCTTGATTCATCATAATTAATCCATCCATA

dbSNP - validated by cluster TAGAGAAGTCACTCCTGAAACCCACAGAGTGTAGCCGGGAAGTCAAGGGTR**AGTCAGGACTGAAGAAACCCCTCTGCCCAGGCATGCATTAAATTAAATTGGGCGGAGCAGCA**
GGTCCTGTCTCTGCCAAACTCTGTCCACACGGGCTCTTGTTTTACTGGTGAACCCACATGCTGCACCCCTTGATTCATCATAATTAATCCATCCATA

Both MID-548 flanking sequence and rs74656636 sequence in 1000Genomes and dbSNP match.
rs375726490 flanking sequence does not match for some bases around the indel motif.
Despite the flanking region discordances between MID-548 and rs375726490, as the latter is not validated, we consider that MID-548 flanking region is correct and rs375726490 insertion is included in it.

Sanger sequence - MID-548 Rev primer was used.

MID-548 dbSNP Rev sequence CAGTAAAACAAGAGCCCGTGTGGACAG(-/GA)TTTGGCAGAGACAGGACCTGCTGCTCCGCCCAATTTAATTTAATGCATGCCTGGGCAGAGGGGTTTCTTCAGTCCTGACT

Allele 2 C A G G A T T T G G C A G A G A C A G G A C C T G C T G C T C C G C C C A A T T T A A T T T A A T G C A T G C C T G G
Allele 3 C A G G A T T T G G C A G A G A C A G G A C C T G C T G C T C C G C C C A A T T T A A T G C A T G C C T G G G C A G A

Expected sequence C A G G A T T T G G C A G A G A C A G G A C C T G C T G C T C C G C C C A A T T T A A T K Y A W K S C W K G S C W G R

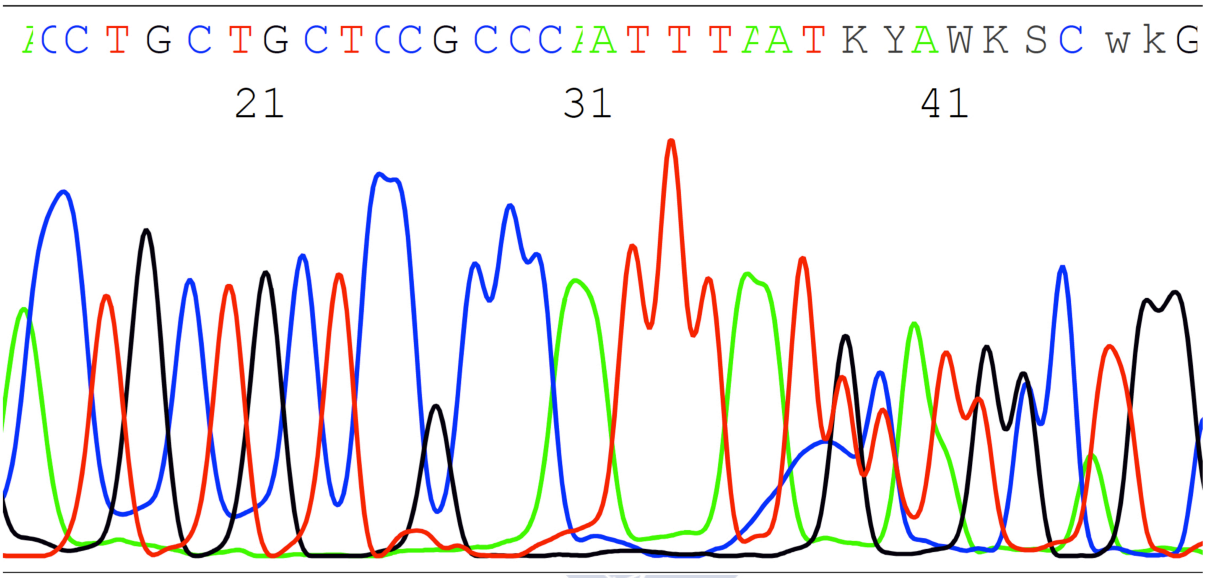


Figure 58: MID-548 third allele characterization. The five Central South Asian HGDP-CEPH samples were analysed with Sanger sequencing using the available reverse PCR primer (an example of the sequences obtained is presented). Results confirmed that the third allele is due the presence of the 5bp deletion allele of rs375726490, upstream of the MID-548 location.

4.2.4.3 Genetic ancestry analysis

Although the *46 AIM-Indels* assay was originally designed to differentiate African, European, East Asian and Native American populations, it has already been shown to be capable of distinguishing Oceanians (Pereira et al. 2012b). Once genotyping of the HGDP-CEPH panel was completed, we extended the assessment of this multiplex to differentiate ME and CSA population groups. Furthermore, forensic ancestry assessment using HGDP-CEPH

samples as reference data is now improved through the combined analysis of these AIM-Indels with several available forensic AIM-SNP or STR sets (Fondevila et al. 2013, Phillips et al. 2013a, Phillips et al. 2013b).

When analysing the seven population groups of HGDP-CEPH with *STRUCTURE*, both *admixture* and *admixture LOCPRIOR* ancestry models gave an optimum *K*=6 genetic clusters (see **Figure 59**). The sixth cluster defines an ancestry component mainly present in CSA populations (also characterized by a second European ancestry component). Exceptions are seen in Hazara (Pakistan) and Uygur (extreme West China), with a clear East Asian ancestry component; explained by the geographic location of Hazara near the western border of China and from historical demographic factors described for the origins of the Uygur (Li et al. 2009). The Middle East populations cannot be separated using the *46 AIM-Indels* – they can be described as mainly European with minor African components. As expected, North African Mozabites show noticeably higher proportions of sub-Saharan African components, as shown by the ANOVA results presented in **Figure 60** (p-value = 3.61×10^{-25}).

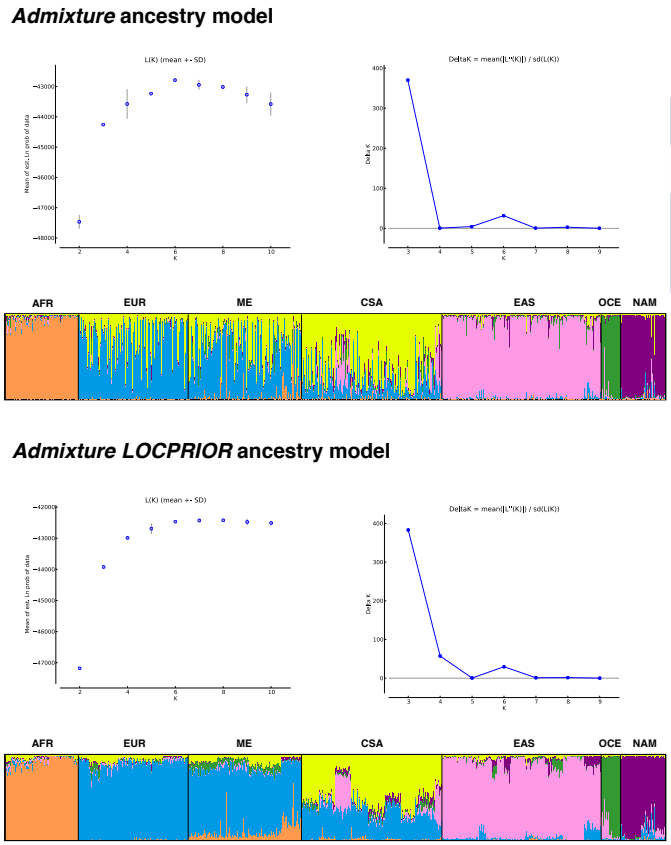
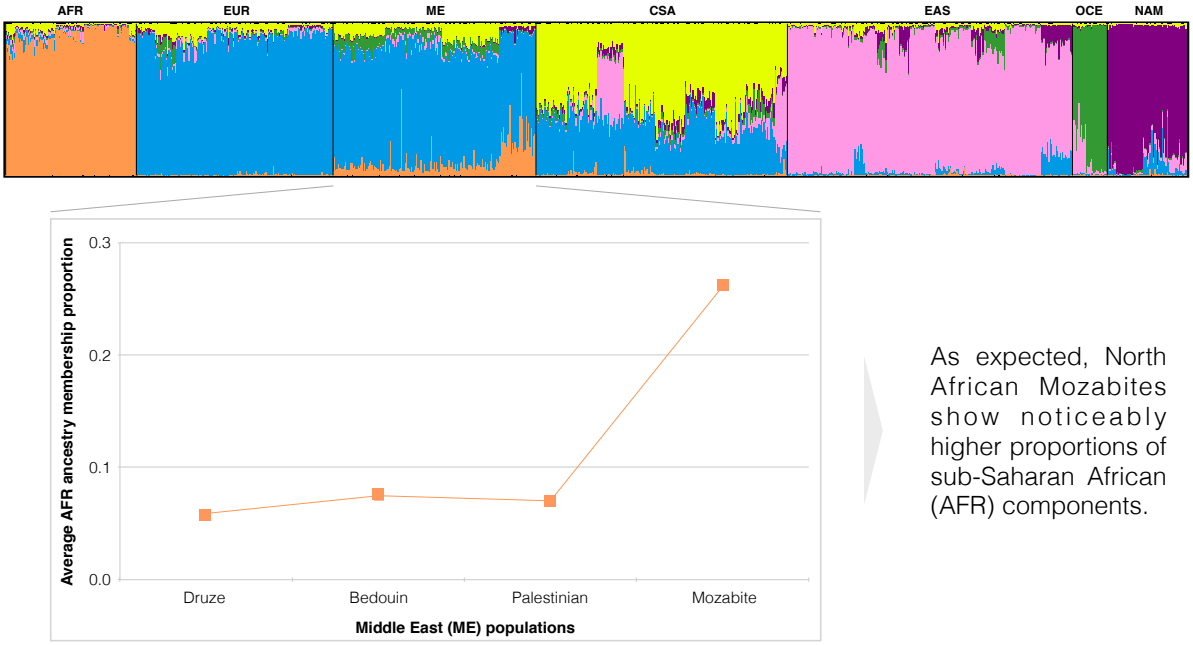


Figure 59: Comparison of STRUCTURE results. Mean LnP(D) and *delta K* plots were generated with *Structure Harvester* for both ancestry models considered: *admixture* and *admixture LOCPRIOR*. In both cases, optimum *K*=6 bar plots were generated using *CLUMPP* and *distruct*.

The Middle East populations cannot be separated using the 46 AIM-Indels – they can be described as mainly European with minor African components.



As expected, North African Mozabites show noticeably higher proportions of sub-Saharan African (AFR) components.

The ANOVA test shows there are significant differences between the four groups compared (α<0.05).

AFR component	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	0.898	3	299	57.310	3.61E-25
Within Groups	0.830	159	5		
Total	1.729	162			

Both Bonferroni and Tukey B post hoc tests indicate that the Mozabite group has a significant higher African ancestry membership component than that observed for the Druze, Bedouin and Palestinian groups.

Multiple comparisons: Bonferroni
Dependent variable: AFR component

(I) Pop	(J) Pop	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Druze	Bedouin	-0.015756	0.015424	1.00E+00	-0.05696	0.02545
	Palestinian	-0.011060	0.015424	1.00E+00	-0.05227	0.03015
	Mozabite	-0.202673*	0.017449	3.41E-22	-0.24929	-0.15605
Bedouin	Druze	0.015756	0.015424	1.00E+00	-0.02545	0.05696
	Palestinian	0.004696	0.015070	1.00E+00	-0.03557	0.04496
	Mozabite	-0.186918*	0.017137	2.97E-20	-0.23270	-0.14113
Palestinian	Druze	0.011060	0.015424	1.00E+00	-0.03015	0.05227
	Bedouin	-0.004696	0.015070	1.00E+00	-0.04496	0.03557
	Mozabite	-0.191613*	0.017137	5.28E-21	-0.23740	-0.14583
Mozabite	Druze	0.202673*	0.017449	3.41E-22	0.15605	0.24929
	Bedouin	0.186918*	0.017137	2.97E-20	0.14113	0.23270
	Palestinian	0.191613*	0.017137	5.28E-21	0.14583	0.23740

* The mean difference is significant at the 0.05 level.

Homogeneous subsets: AFR component
Tukey B[a,b]

Pop	N	Subset for alpha = 0.05	
		1	2
Druze	42	0.05881	
Palestinian	46	0.06987	
Bedouin	46	0.07457	
Mozabite	29		0.26148

Means for groups in homogeneous subsets are displayed.
a. Uses Harmonic Mean Sample Size = 39.304.
b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Figure 60: Analysis of *STRUCTURE* membership proportions in ME populations using ANOVA. Bonferroni and Tukey B tests show Mozabites have a significantly higher African ancestry membership component than Druze, Bedouin and Palestinian groups.

The *Snipper* cross-validation success rates obtained by classifying the full HGDP-CEPH panel of seven population groups are shown in Table 31. Success rates are listed for 46 Indels and using 80 AIMs combined. Increased success rates are evident with 80 AIMs, in line with the findings of other studies (Fig. 3A-B in (Phillips et al. 2013a), Fig. 3 in (Phillips et al. 2011) and Fig. 2 in (Lao et al. 2010)). Nevertheless, Table 31 shows that considerable error rates persist when dealing with the Eurasian group, in particular when classifying EUR (37.58%) and ME (42.24%) individuals. To less extent, levels of classification error are also found for CSA, with 7.46% incorrectly assigned to other population groups. As the differentiation of populations within Eurasia was not the intended purpose of both marker sets, markedly lower classification success of ME in a seven-group differentiation is not unexpected. This is a widely recorded result with a range of different ancestry marker sets having low divergence between EUR or CSA and ME populations (Rosenberg et al. 2002, Li et al. 2008, Phillips et al. 2013b). The described error rates derive from the fact that *Snipper* always assigns individuals to the group with the lowest likelihood (-log likelihoods: smaller values denote higher probabilities), not taking into account the fact that there can be very similar values for other population groups. When considering average likelihoods obtained for each population group (Table 32), it is evident that correctly classified individuals have much lower likelihoods compared with the values obtained for other possible assignments (e.g. AFR individuals classified as Africans have average likelihoods of ~60, but when classified as Europeans this value increases to ~176). Conversely, likelihood values for EUR classified as Europeans or Middle Eastern are near identical, highlighting the lack of divergence between these two groups. This lack of divergence extends to ME compared to CSA. Overall, the Snipper likelihood ratios from a large proportion of Eurasian comparisons, particularly those involving EUR and ME, are too low for reliable assignments. Therefore we recommend that these markers are combined with other AIMs to improve the classification of more closely related populations and likelihood ratio (LR) thresholds are defined in order to avoid misclassifications (Phillips et al. 2009).

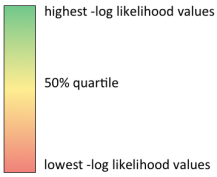
Table 31: Comparison of classification success (bold diagonal values) estimated from cross-validation of the 7 group HGDP-CEPH training set in Snipper using *46 AIM-Indels* or 80 AIMs.

	***46 AIM-Indels* HGDP-CEPH 7 group training set**						
	Estimation of the success ratio with cross-validation using *46 AIM-Indels*						
	AFR	**EUR**	**ME**	**CSA**	**EAS**	**OCE**	**NAM**
Population of AFR origin	**100 %**	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %
Population of EUR origin	0.00 %	**40.51 %**	39.87 %	19.62 %	0.00 %	0.00 %	0.00 %
Population of ME origin	1.23 %	0.61 %	**46.01 %**	52.15 %	0.00 %	0.00 %	0.00 %
Population of CSA origin	0.00 %	0.00 %	0.50 %	**90.10 %**	7.43 %	1.49 %	0.50 %
Population of EAS origin	0.00 %	0.00 %	0.00 %	0.00 %	**96.94 %**	0.87 %	2.18 %
Population of OCE origin	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	**100 %**	0.00 %
Population of NAM origin	0.00 %	0.00 %	0.00 %	1.56 %	0.00 %	0.00 %	**98.44 %**

	80 AIMs: *46 AIM-Indels* + 34 AIM-SNPs HGDP-CEPH 7 group training set						
	Estimation of the success ratio with cross-validation using 80 binary AIM markers						
	AFR	**EUR**	**ME**	**CSA**	**EAS**	**OCE**	**NAM**
Population of AFR origin	**100 %**	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %
Population of EUR origin	0.00 %	**62.42 %**	33.76 %	3.82 %	0.00 %	0.00 %	0.00 %
Population of ME origin	1.24 %	0.00 %	**57.76 %**	40.99 %	0.00 %	0.00 %	0.00 %
Population of CSA origin	0.00 %	0.00 %	0.00 %	**92.54 %**	6.97 %	0.00 %	0.50 %
Population of EAS origin	0.00 %	0.00 %	0.00 %	0.00 %	**97.79 %**	0.00 %	2.21 %
Population of OCE origin	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	**100 %**	0.00 %
Population of NAM origin	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	**100 %**

Table 32: Average -log likelihoods are summarized on the table on the right. Diagonal values show the likelihoods observed for correct classification of population group members, but the middle box of Eurasian population groups have very similar values, particularly ME comparisons. This highlights the raised risk of classification error for individuals from these three groups.

	AFR	EUR	ME	CSA	EAS	OCE	NAM
Africa	60.932	176.775	124.051	137.140	168.001	143.578	187.098
Europe	191.840	62.020	62.729	65.480	119.718	122.938	123.916
Middle East	169.553	95.901	69.644	69.854	116.415	115.978	124.203
Central South Asia	173.603	101.307	89.583	69.968	101.188	111.778	107.352
East Asia	186.219	167.157	139.627	115.864	57.967	91.848	83.252
Oceania	150.845	140.410	114.287	106.461	98.835	57.320	109.594
America	202.052	159.638	136.674	111.539	100.871	139.034	58.205



Furthermore, *Snipper* should not be considered a stand-alone classification tool. PCA and *STRUCTURE* analysis (**Figure 61**) reflect the better differentiation of CSA compared to ME for these AIM-Indels. When combining the 34 AIM-SNPs with AIM-Indels, the PCA analyses indicate increased separation between African and Eurasian clusters, as well as between European and East Asian clusters (**Figure 61A**). We estimated Euclidean distances between cluster centroids in order to assess cluster separations. Results are outlined in **Figure 62**.

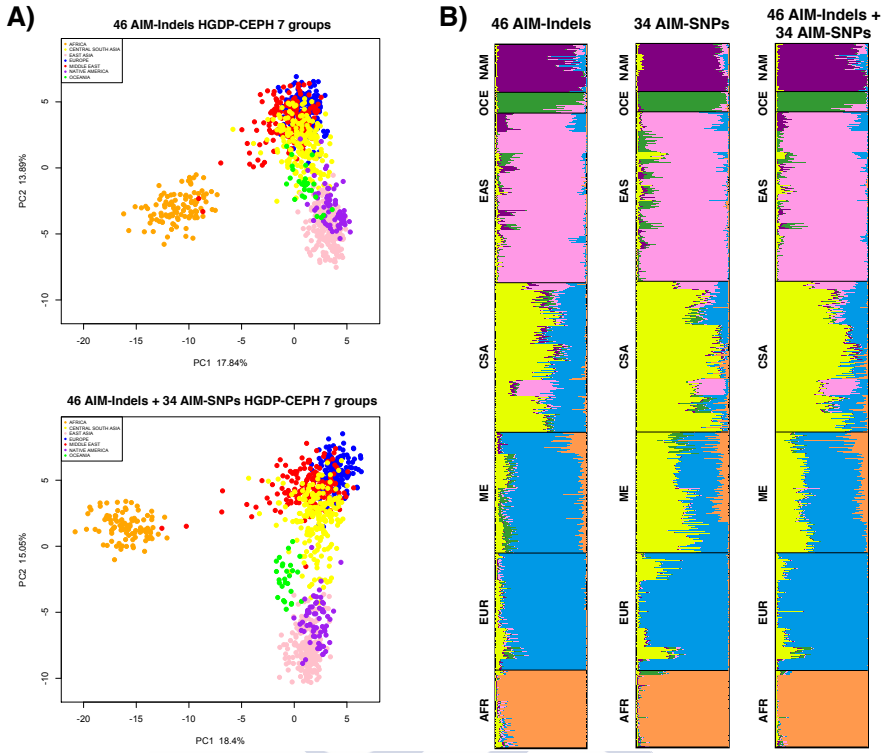
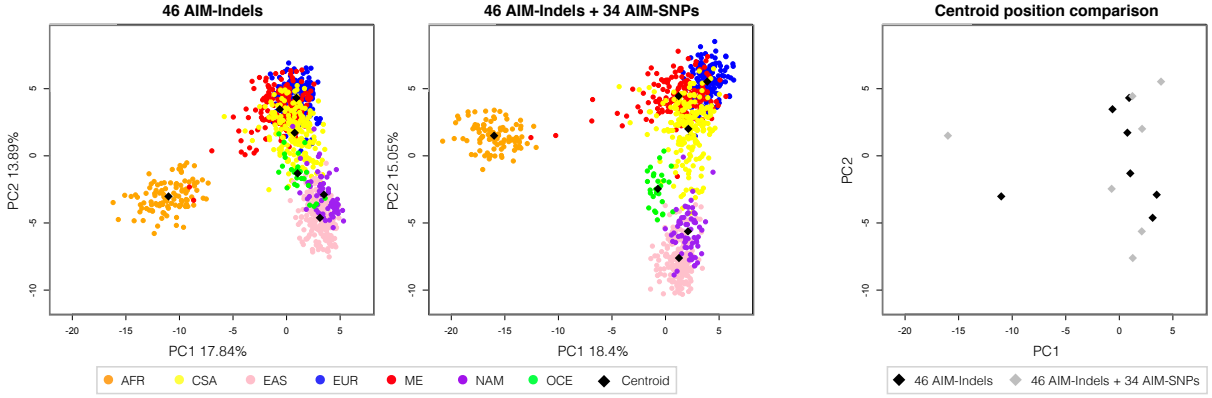


Figure 61: Principal component analysis and *STRUCTURE*. (A) PCA for the complete HGDP-CEPH panel using the *46 AIM-Indels* and combined *46 AIM-Indels* plus 34 AIM-SNPs. (B) *STRUCTURE* ancestry analysis of the complete HGDP-CEPH panel based on *46 AIM-Indels*, 34 AIM-SNPs and both sets together. Results are based on 3 independent *STRUCTURE* runs for K=6 considering *admixture LOCPRIOR* ancestry model and correlated allele frequencies model. Replicates were merged using *CLUMPP* and plotted with *distruct*.

In the *STRUCTURE* analysis (**Figure 61B**) patterns of ancestry proportions match the expected cluster patterns and on the basis of their ancestry membership proportions, CSA individuals can be well differentiated from Europeans when adding 34 SNPs. Considering only *46 AIM-Indels* or 34 AIM-SNPs, ME populations cannot be clearly differentiated from EUR and CSA populations, in line with the above *Snipper* analyses. However, the increased differentiation power observed for 80 binary AIMs generates a gradient in the ancestry membership proportions estimated by *STRUCTURE*, with ME populations lying between EUR and CSA populations. This gradient in the European ancestry membership proportions is better correlated with populations' longitude, especially if populations more dispersed from a W-E clinal axis, such as the Mozabite, Adygei and Russians are removed from the analysis (**Figure 63**). The same gradient is found from F_{ST} analysis of the *46 AIM-Indels* alone (see Table 30). It is noteworthy that the distance is only 0.035 between EUR and CSA, with ME in an intermediate position (0.027 for ME-CSA, but closer still between EUR and ME with a value of 0.015). These contrast with F_{ST} values seen outside the Eurasian population groups with 0.167 being the smallest distance (CSA:OCE).

Considering the PCA analysis presented in the main Fig. 1A, centroid positions were estimated for each population cluster (average PC1 and PC2 scores based on the individuals belonging to each cluster). When comparing centroid positions estimated with 46 AIM-Indels alone or in combination with 34 AIM-SNPs, it is apparent that distances increase when using the 80 binary AIM markers.



This increased separation is of special interest when considering the African and the Eurasian groups (the latter including Europe, Middle East and Central South Asia) but also between Europe and East Asia.
To confirm these two assumptions, simple Euclidean distances between centroids were estimated using the following formula:

$$\sqrt{(POP1_{PC1} - POP2_{PC1})^2 + (POP1_{PC2} - POP2_{PC2})^2}$$

We also estimated the Ray-Tury clustering index using the *clusterCrit R* package. This index compares the tightness of clusters to the minimum distance between clusters. Smaller index values indicates clusters are tighter and/or that clusters are more separated, with zero the lowest value possible.

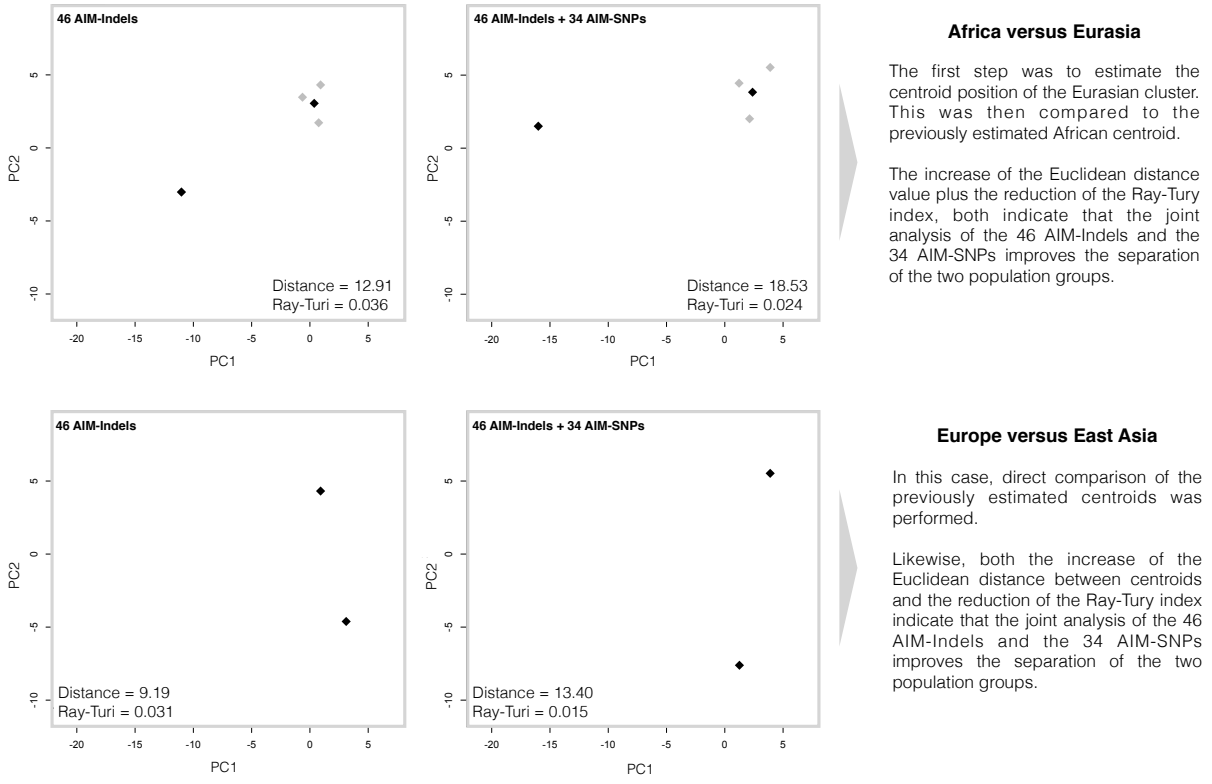
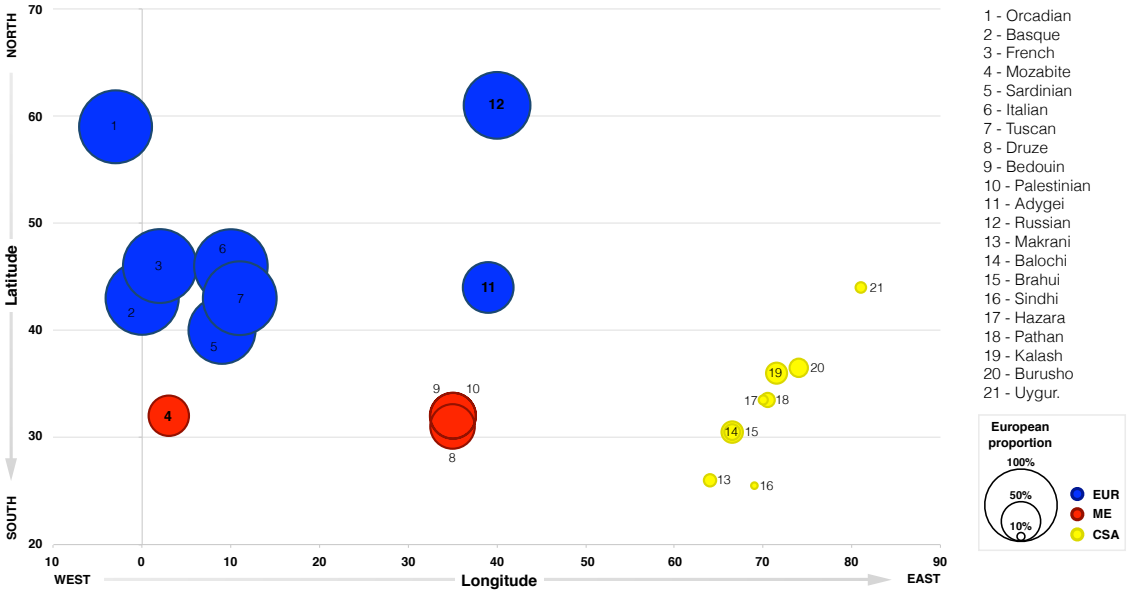
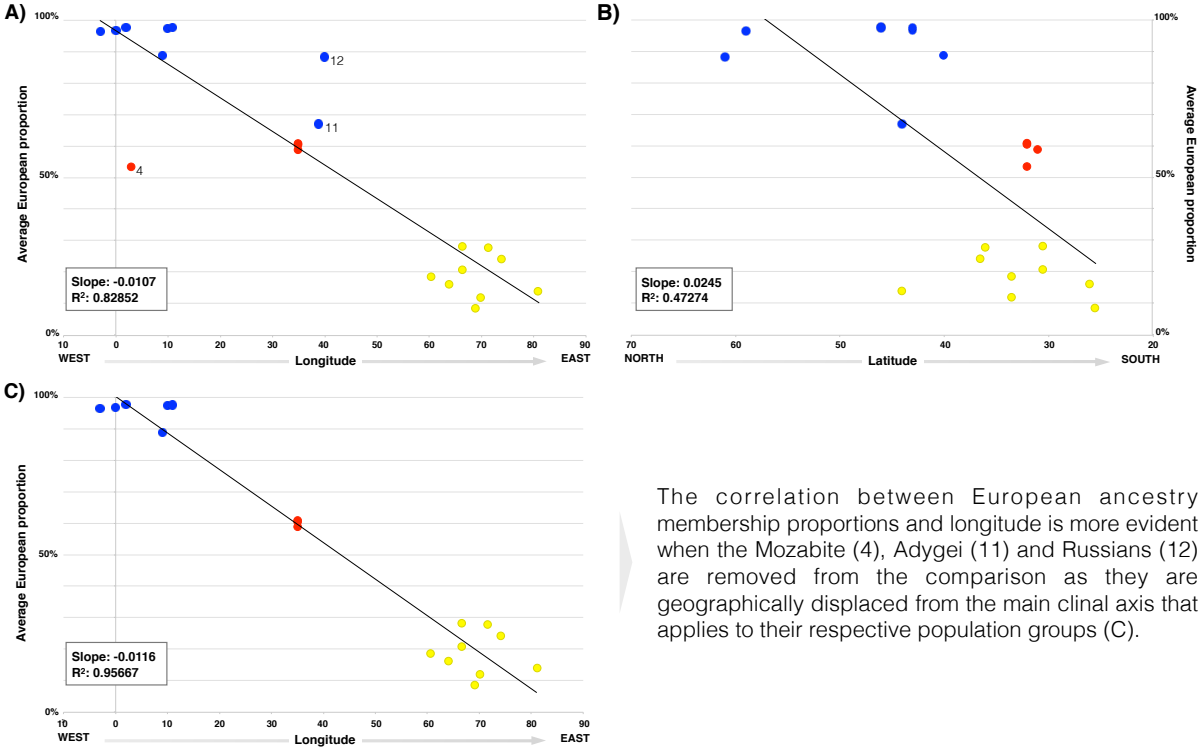


Figure 62: PCA differentiation tests measuring centroid separations and tightness of clustering. Results indicate better separation of African and Eurasian groups using both marker sets together. The same improved separation is evident in PCA comparisons of Europe and East Asia.

Geographic distribution of the European (EUR), Middle Eastern (ME) and Central South Asian (CSA) populations from the HGDP-CEPH diversity panel with reference to their latitude and longitude co-ordinates (in degrees). Grey arrows indicate direction of movement. The size of each point represent the average European ancestry membership proportions estimated by *STRUCTURE* using 80 binary AIMs (see main Fig. 1B).



Average European ancestry membership proportions distribution in comparison with longitude (A) and latitude (B) of the EUR, ME and CSA populations. Linear regression results are presented: the slope of both lines shows there is a gradual decrease of the European ancestry membership proportions from EUR to ME to CSA; R^2 values give the correlation between membership proportions and the geographic location. In this case the gradient appears to be stronger when considering latitude, but European ancestry membership proportions are more correlated with longitude.



The correlation between European ancestry membership proportions and longitude is more evident when the Mozabite (4), Adygei (11) and Russians (12) are removed from the comparison as they are geographically displaced from the main clinal axis that applies to their respective population groups (C).

Figure 63: Distribution of European membership proportions in EUR, ME and CSA groups with relation to longitude and latitude. Membership proportions show a clear correlation to longitude in an easterly direction.

4.2.4.4 *forInDel*: an open access forensic Indel frequency browser

In forensic analysis the availability of marker genotypes or allele frequency estimates for sets of relevant reference populations provides key data. Following the example of the SNP*for*ID online forensic SNP databases (Amigo et al. 2008a) and with the same framework, we created an open access allele frequency browser listing *46 AIM-Indel* genotypes for the complete HGDP-CEPH H952 subset, named *forInDel* – Forensic Indel browser (http://spsmart.cesga.es/forindel.php). Different populations can be combined into groups following a geographic basis or according to a user's own grouping regime (maximum five groups compared), making it easy to combine allele frequency estimates and download genotypes. Population parameters: observed and expected heterozygosity, F_S, F_{ST} estimates and *In* are calculated. As we made use of a SNP based framework, alleles in *forInDel* are recoded similarly using conventional nucleotide letters: A – short allele; C – long allele; and G – third allele (applies to MID-548, MID-360, MID-2264) and identical ACG genotype codes are employed in *Snipper*. Allele frequencies are represented by pie charts color-coded as: A – blue; C – red; and G – green. Note that colour assignments do not match equivalent SNP data in *SPSmart*, which uses the convention of blue segments for reference alleles. In the future we aim to add full HGDP-CEPH data for 38 human identification Indels (Pereira et al. 2009) to *forInDel* along with additional populations that contributors are invited to submit, increasing the browser's scope and coverage.

4.2.4.5 Concluding remarks

With this study the genotyping of the HGDP-CEPH diversity panel for a set of *46 AIM-Indels* has been completed, allowing a broader geographic scope for combined analyses with equivalent marker sets such as the *34-plex* AIM-SNPs (Fondevila et al. 2013). The differentiation of European and South Asian ancestries applying two compact AIM sets is now feasible by analysing DNA of unknown origin with standard approaches that compare genetic data with the expanded reference population genotypes reported here. All genotypes and frequency estimates are publicly available and can be used as reference data in forensic applications such as inference of ancestry of casework samples. The genotype data is now compiled in a dedicated web browser for forensic Indels named *forInDel* and these genotypes

or major population indices can be downloaded. Ancestry analysis of the complete HGDP-CEPH H952 panel indicates the *46 AIM-Indels* can differentiate six major population groups but with noticeable classification error in EUR and ME. It is important to stress that the differentiation of EUR from CSA and ME populations was not a target for the original marker selection.

4.2.5 *Eurasiaplex*: a forensic SNP assay for differentiating European and South Asian ancestries (Appendix 2D)

4.2.5.1 Characteristics of the *Eurasiaplex* assay and component SNPs

Figure 64 shows the *Eurasiaplex* SNP profile for standard forensic control DNA 9947A. In casework samples SEFiler showed no detectable genotypes in ten DNAs with remaining profiles showing average 9% profile completeness, while five degraded DNA samples gave 60% profile completeness with Identifiler. Eurasiaplex profiles gave an average 83% completeness with a median 19/23 SNPs detected in the twenty samples genotyped.

The genomic map of 57 SNPs in **Figure 32** shows the closest *Eurasiaplex* SNP pair is rs1785864–rs1941411 (11:104955489-120130512 ≈ 15.17 Mb) and closest *Eurasiaplex*:*34-plex* pair is rs17625895-rs881929 (16:25775102-31079371 = 5.3 Mb). Therefore marker independence can be assumed combining both SNP sets in Bayesian analyses. One extra AIM-SNP rs3827760, added to *34-plex* late in the study improves differentiation of East Asians from European Eurasians (Fondevila et al. 2013) and full HGDP-CEPH data is in *SPSmart* (http://spsmart.cesga.es/search.php?dataSet=snpforid34).

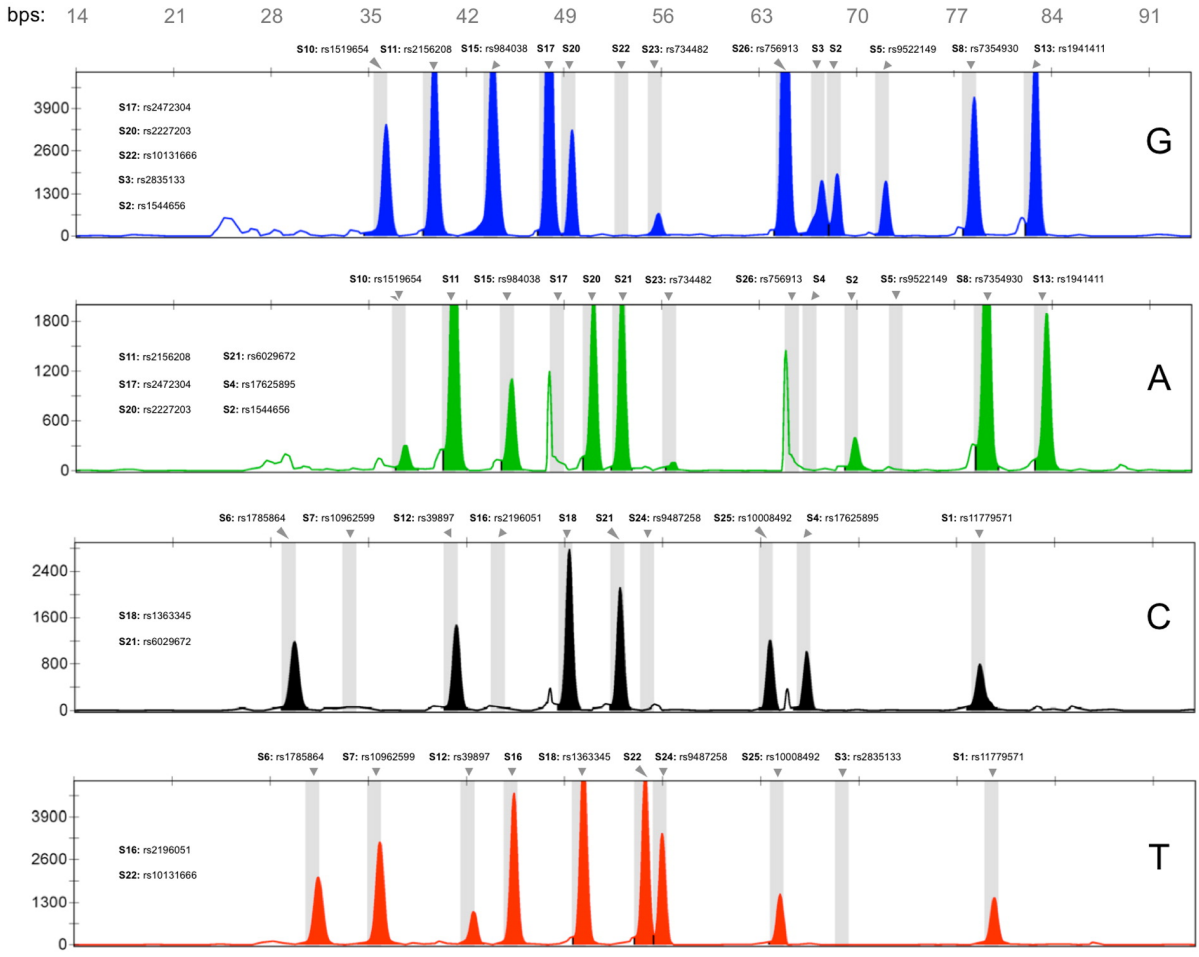


Figure 64: SNaPshot electropherogram of *Eurasiaplex* SNPs for standard forensic control DNA 9947A. SNP peaks labelled with the rs-numbers and internal codes listed in Table 15.

Genomic details and summary allele frequencies for the *Eurasiaplex* SNPs in Europeans, South Asians, East Asians and Middle Eastern populations of HGDP-CEPH is outlined in Table 15. Derived allele frequencies for sixty populations studied indicate a clear clinal distribution in *Eurasiaplex* SNPs with strong contrasts between Europeans and East Asians with intermediate frequencies observed in intervening populations located on the NW-SE longitudinal axis (certain positions displaced to north e.g. Adygei, or south e.g. Indians). **Figure 65** plots individual SNP allele frequencies for 23 Eurasian populations positioned west to east along the cline described above. SNP rs10008492 gave the strongest European to East Asian contrast (top plot) and rs2835133 the least contrast. Although some SNP clines are less pronounced, the average allele frequency plot indicates a strong geographic correlation to longitudinal position, although Iranians, Indians and Uygur deviate slightly from this pattern.

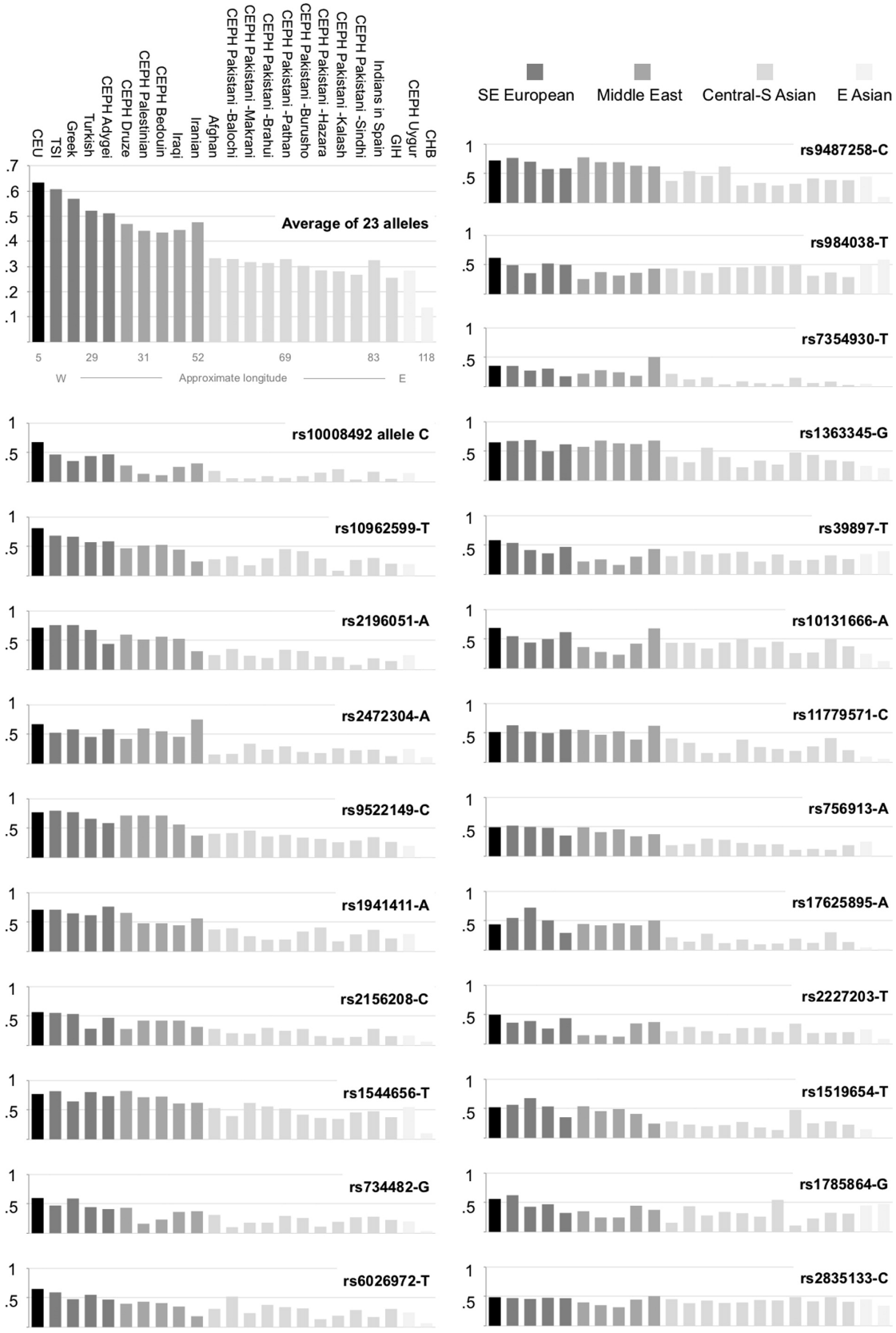


Figure 65: Individual and average *Eurasiaplex* SNP allele frequency clines defined by the geographic points of twenty-three populations positioned west to east to match their longitudes. Note that HGDP-CEPH Uygur are labelled as E Asian here to reflect geography not genetic affiliation.

To gauge individual SNP differentiation power we calculated Rosenberg's *In* divergence values (Rosenberg et al. 2003) comparing European to South Asian population groups. Each *Eurasiaplex* value is plotted in **Figure 66** in descending order of differentiation power alongside an equivalent number of the most divergent *34-plex* SNPs. The most informative SNP is rs16891982 (60% of *34-plex* cumulative divergence) while *34-plex* rs12913832 and rs722098 are comparable to the best *Eurasiaplex* SNPs. Interestingly, rs12913832 is strongly associated with blue eye colour (Eiberg et al. 2008) with a widely recorded cline from NW to S Europe, while rs722098 was chosen because of high European-South Asian divergence. Remaining *34-plex* SNPs drop markedly in informativeness below *Eurasiaplex* SNPs, reflected in the final *Eurasiaplex* cumulative divergence twice as high. Therefore, combining multiplexes provides optimum differentiation of Europeans and South Asians, but just three SNPs provide the bulk of *34-plex* differentiation power, having a bearing on interpretation of partial *34-plex* profiles.

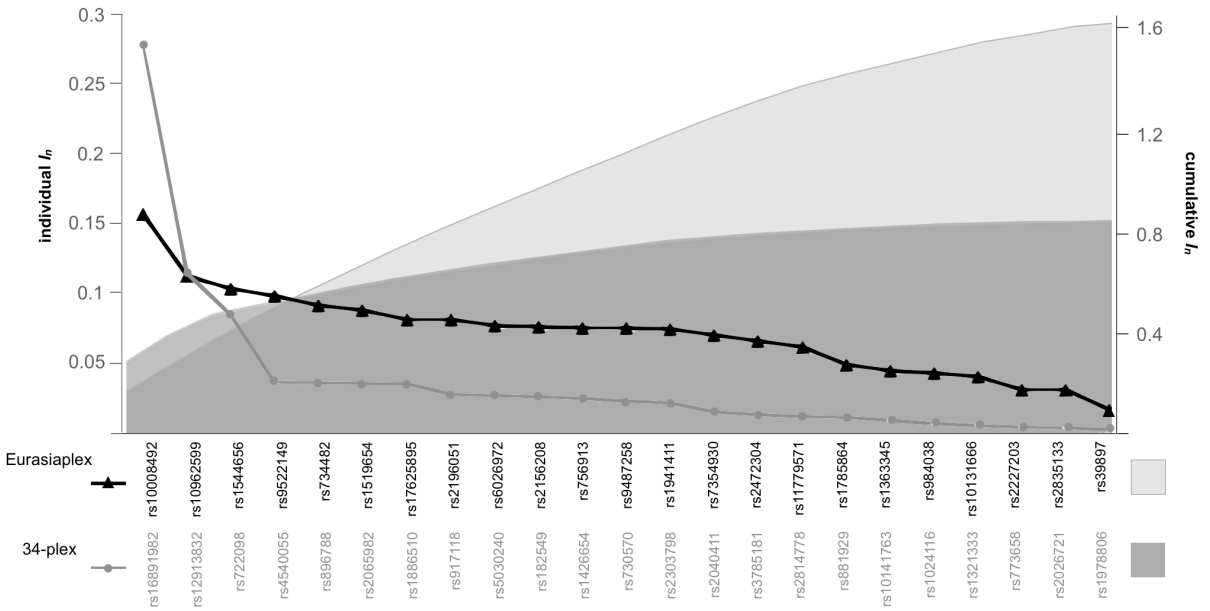


Figure 66: Individual (linked points, left axis) and cumulative (solid curves, right axis) Rosenberg Divergence values: *In* for all *Eurasiaplex* and twenty-three most differentiating *34-plex* SNPs. Loci are ordered in decreasing *In*, left-to-right, *Eurasiaplex* SNPs shown as black triangles and lighter grey shade.

4.2.5.2 Population divergences assessed with principal component analysis

PCA results analysing sixty populations from all five groups, are plotted for the two principal components in **Figure 67A**. As we compared a wide range of closely related populations, plots were simplified to clarify the spatial separations amongst Eurasian samples. **Figure 67** includes eleven Middle Eastern populations, excluded in **Figure 67B** to better display degrees of overlap amongst Europeans and Central-South Asians. The atypical Yakut and HapMap ASW (African Americans of SW America, with high levels of European admixture) were also removed. Outlying populations were highlighted by assigning grey colours to most populations in cluster centres, shown in **Figure 67C**. These indicate ten intermediate position or outlier populations (plus skewed position GIH due to incomplete data). The clearest trend in the full PCA distributions is extensive overlap of Middle East populations with European and South Asian clusters, although they are well separated from each other. Outliers comprise: Yakut occupying space between South Asia and East Asia consistent with their geographic and genetic differentiation from East Asia; Moroccan and HGDP-CEPH Algerian distributions below the Middle East cluster, extending toward the African cluster, consistent with North Africa's geographic proximity to sub-Saharan Africa. Positions of Tuscans (HGDP-CEPH and 1000 genomes), HGDP-CEPH Adygei and HGDP-CEPH Sardinians are highlighted in **Figure 67C** as they gave more erroneous ancestry assignments with *Snipper* (section 4.2.5.3), but in PCA analysis these populations position well within the European cluster. In contrast, Greek and Turkish occupy intermediate positions between European and South Asian clusters with overlap into the South Asian cluster. Amongst South Asians, Afghans show the most spread distribution and encroach the space separating Europeans, consistent with their geographic position on the western edge of Central-South Asia. The HGDP-CEPH Pakistani Kalash and Hazara encroach toward East Asia contrasting with HGDP-CEPH Uygur in the centre of the South Asian cluster, lacking affinity with East Asia despite occupying the extreme west of China. In summary, comparing Europeans and South Asians using PCA indicates very good separation. Greek and Turkish from east Mediterranean give the least separation and closer proximity to South Asia, which was unexpectedly less pronounced in the HGDP-CEPH Adygei of the Caucasus despite their more intermediate geographic position and showing high misclassification rates in analyses detailed below.

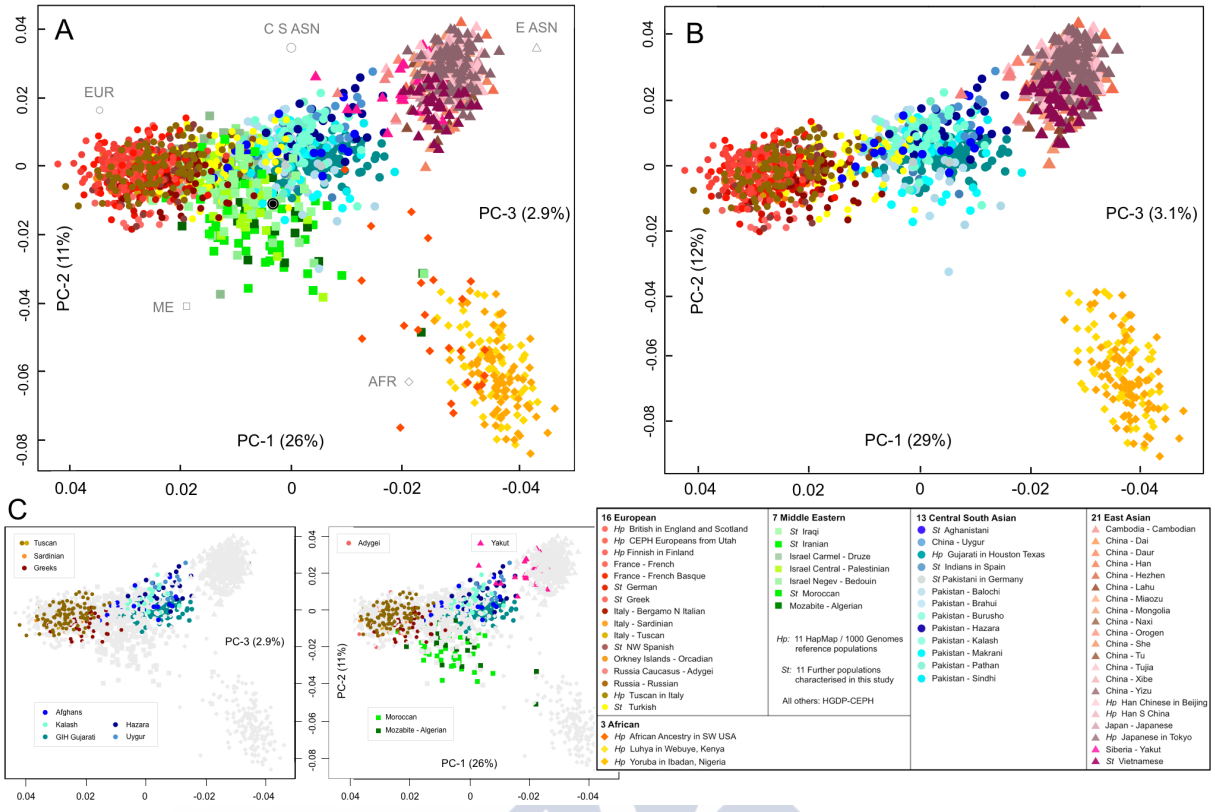


Figure 67: (A) 2-Dimensional PCA plot of 60-population analyses with colours and symbols used to differentiate the five groups and with certain population outliers with more contrasting colour labels. The black circle mid-plot shows the PCA position of the 47-SNP profile obtained from the unidentified cranium case using joint PCA with the same population data. Eurasians comprised EUR: European, ME: Middle Eastern, C S ASN: Central-South Asians. Other groups: AFR sub-Saharan Africans, E ASN: East Asians. (B) PCA plot from a similar analysis to that of (A) excluding seven Middle East populations. (C) Two PCA plots from the 60-population analysis of (A) simplified by colour reduction to light grey of the well-separated populations to highlight populations with outlying or with intermediately clustered points.

4.2.5.3 Bayesian classification analysis

We assessed the ability of *34-plex* and *Eurasiaplex* AIM-SNPs to differentiate Eurasians by classifying all study and reference populations from Eurasia and East Asia with *Snipper*. Training set individuals were analysed by reclassification (i.e. assigning ancestry to individuals using the training set to which they belonged) and all other profiles were systematically classified using four and/or five-group training sets. The overall classification performance of the 57 SNPs is summarized by charting likelihood ratios as: proportions correct and greater than the '100 times more likely' threshold (>100×), those correct <100× more likely, those incorrect <100× and those incorrect >100×. Therefore applying a threshold measures: non-classification rate from total probabilities <100×; classification success from total correct >100×, and classification error from total incorrect >100×. **Figure 68A** and **B**

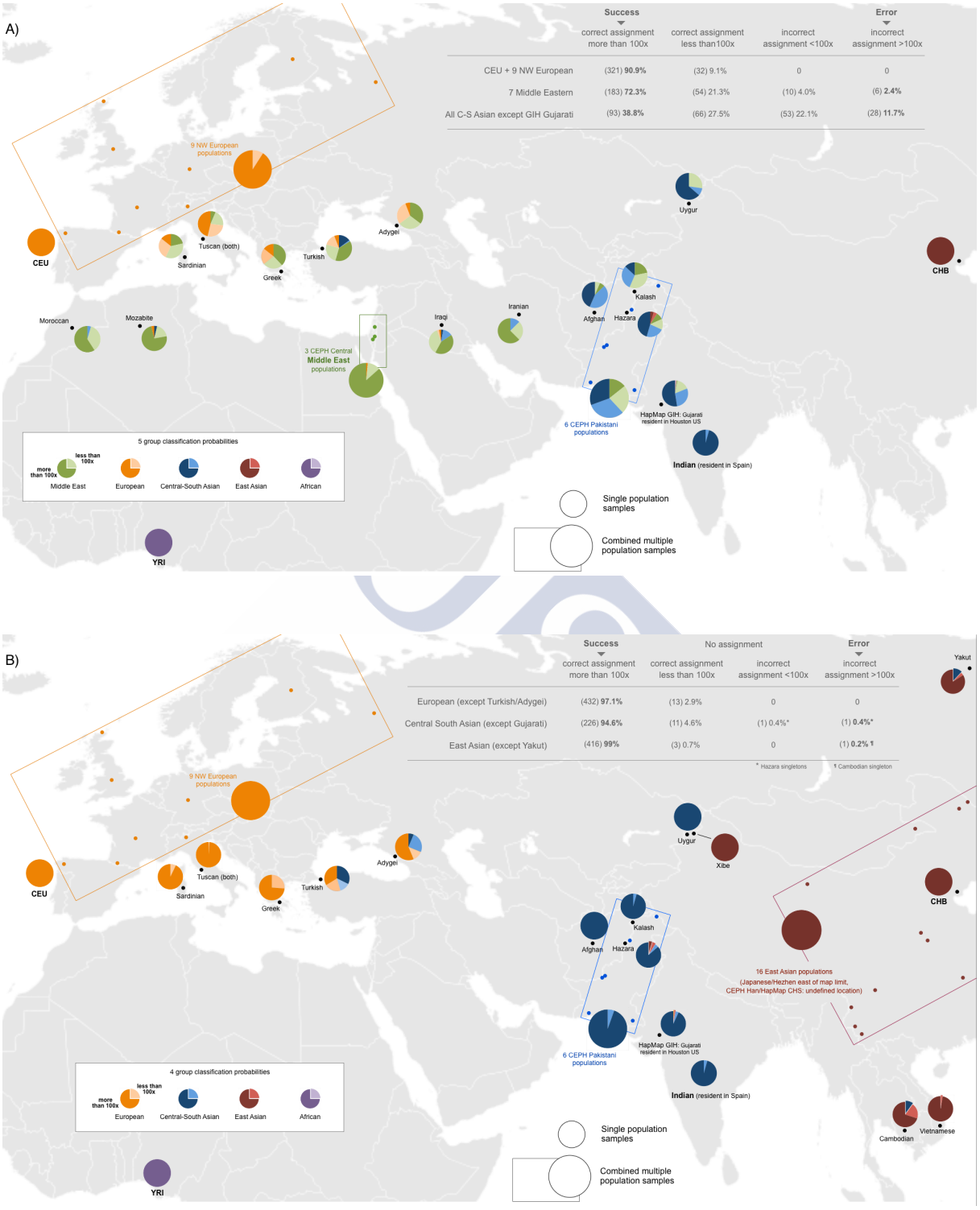


Figure 68: (A) Pie charts summarizing the proportion of individuals in a range of populations classified with *Snipper* using a five-group training set as pie charts displaying ratios of successful ancestry assignments, erroneous assignments and not assigned (defined by a probability threshold of less than 100 times more likely). Populations with near identical classification proportions were grouped together as defined by coloured points and surrounding boxes. The summary top-right table of success/error/non-assignment percentages excludes several outlying or intermediate position populations mapped in the charts below. Training set populations labelled in bold. (B) Assignment success, error and non-assignment proportions with *Snipper* using a four-group training set, with assignments extended to analyse East Asian populations.

chart distributions of assignment success/error. Where populations from the same group gave very similar assignment performance we combined them into single pie charts as follows: 10 NW European; 3 Middle East (comprising the training set populations); 6 HGDP-CEPH Pakistani and 16 East Asians. Classification performance with the Middle East five-group training set indicates high levels of error, notably amongst SE European and Central-South Asian populations to the NW of India. High error rates extend to Italian Tuscans and Turkish. Therefore, although NW Europeans have high classification success of 91% and only 9% were unclassified, success drops to 72% for the Middle East and just 39% for Central-South Asians, where classification error reaches 12% and non-classification 28%. In contrast, applying a four-group training set raises classification success to 97% in Europeans, 94.6% in Central-South Asians and 99% in East Asians. Erroneous assignments in Central-South Asia were confined to a single Hazara individual classified as East Asian. Certain omissions to the above success rate estimates should be highlighted: we excluded outlier Turkish and Adygei from Europe plus Siberian Yakut from East Asia. HapMap Gujarati were excluded due to incomplete SNP data.

Therefore applying a four-group training set provides very good classification success rates ranging from 95% to 99%, differentiating East Asians, Central-South Asians and nearly all Europeans except those on the far eastern edge of the continent. The non-classification rate reaches an acceptable maximum 5%, while East Asian differentiation is particularly strong. Furthermore, there is scope for more systematic analysis of probability thresholds to further improve balance between classification success and non-classification rates. By contrast, achieving adequate differentiation of Middle East populations from neighbouring south and southeast European plus Central-South Asian populations outside India remains work in progress, likely to require much larger SNP sets to reach acceptable likelihoods.

4.2.5.4 AIM-SNP and morphometric analysis of unidentified cranium

Formal canonical plot analyses placed the cranium dimensions closest to Japanese female reference data. Replicated SNP typing gave 47 duplicate genotypes. Genetic analysis produced strong statistical indications against East Asian ancestry with reduced likelihoods against South Asian ancestry. Taking each analysis in turn, combined PCA of reference and cranium SNP data in the 60-population analysis of **Figure 67A** positioned the unknown

individual between Middle East and European clusters, slightly closer to the Middle East group. The cranium's profile position indicated by the black circle in **Figure 67A**, lies just to the right of most European points. Next, *STRUCTURE* analyses, using five reference population groups labelled with POPFLAG and treating the cranium data as unknown, are summarized in **Figure 69**. Plots show the optimum cluster number analysing *34-plex*, *Eurasiaplex*, then all SNPs combined. The *34-plex* analysis has an optimum *K*=3 (i.e. three genetic clusters detected from data) and shows clustering of the cranium profile with the Eurasian cluster. The *34-plex*, *K*=3 membership proportions were: Africa: 0.0624; Eurasia: 0.9152; East Asia: 0.0224. Analysis with *Eurasiaplex* SNPs and combined sets both indicated an optimum *K*=5 with some lack of differentiation between the three Eurasian clusters for Eurasiaplex SNPs, more clearly differentiated adding *34-plex* data. In both analyses the cranium profile shows joint membership to European and Middle East clusters but much less to the Central-South Asian cluster, with majority Middle East membership increasing with combined SNP data. The *Eurasiaplex*, *K*=5 membership proportions were: Africa: 0.1008; Europe: 0.3380; Middle East: 0.4336; Central-South Asia: 0.1090; East Asia: 0.0186. Combined SNP analysis *K*=5 membership proportions were: Africa: 0.0883; Europe: 0.2145; Middle East: 0.5920; Central-South Asia: 0.0949; East Asia: 0.0102. These *STRUCTURE* results were interpreted to indicate Eurasian ancestry with Middle East or European origin more likely than Central-South Asian. Lastly, the *Snipper* classifications indicated that probabilities obtained for the five-group training set likelihood ratio were 205 times more likely Middle Eastern than European and 593,238 times more likely Middle Eastern than Central-South Asian. Using the four-group training set gave: 3000 times more likely European than Central-South Asian and 9E+27 more likely European than East Asian, strongly excluding the possibility of Central-South Asian or East Asian origin suggested by morphometric analysis. Taken as a whole, the genetic analysis of the cranium suggests Middle Eastern or European ancestry more strongly than Central-South Asian. Despite *Snipper* results being more indicative of Middle Eastern than European ancestry we ignored this distinction due to the high classification error rate of Europe-Middle East differentiations reported in section 4.2.5.3.

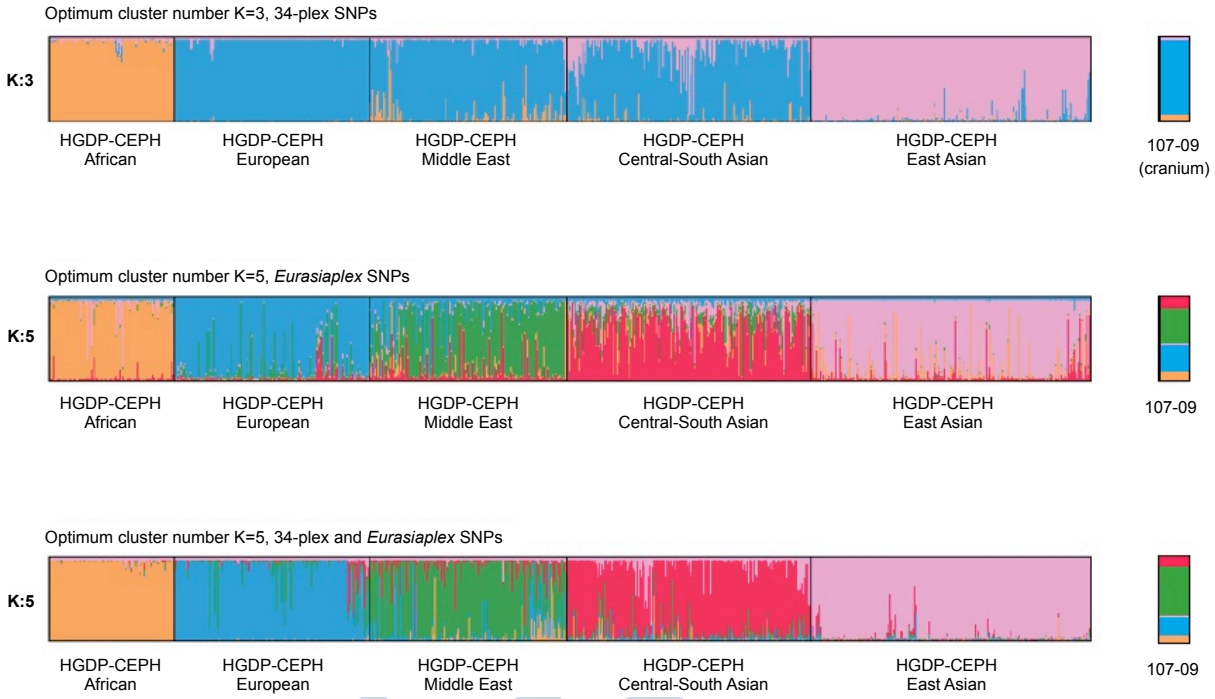


Figure 69: *STRUCTURE* analysis of *34-plex*, *Eurasiaplex* and combined SNP profiles from genotypes of unidentified skeletal remains washed ashore at La Coruna, 2009.

4.2.5.5 Discussion

The studies of Eurasian population variability we report indicate the SNPs of *Eurasiaplex* and *34-plex* differentiate the bulk of European and South Asian populations to high levels of statistical likelihood. Classification success remains high when extending analyses west from India to Afghanistan and Pakistan but moving east from Europe toward Turkey and South Caucasus regions bordering Iran produces considerable classification error, while differentiation of Middle Eastern populations shows even higher error rates. These observations highlight the difficulty of using forensic-scale SNP tests to analyse populations in closely sited regions within a continuous genetic variation gradient – as with the cline characterizing SNP frequencies across Eurasia. Since most European and South Asian populations are sited at the cline's extremes they are reliably separated using *Snipper*, whereas populations closer together and occupying middle regions reduce the overall reliability of this Bayesian classifier. Since *Snipper* reports the lowest likelihood ratio amongst all pairwise population comparisons and assigns likely ancestry to the numerator of the ratio, users can have difficulties gauging the relative magnitude of each likelihood. Therefore *Snipper* ancestry assignments can appear qualitative not quantitative, despite the

likelihood value itself forming the key output and availability of density plots comparing profile likelihoods to training set distributions. Two approaches can help interpret *Snipper* results from closely related populations in an allele frequency continuum: i) adopting threshold values balancing classification success with minimal error rates; and ii) combining *Snipper* with complementary analyses including *STRUCTURE* and PCA that assess similarity patterns amongst populations – helping to highlight the dangers of very close relationships. In this study we advocated a discrimination threshold of '100 times more likely' to balance *Snipper*'s performance differentiating Europeans and South Asians. The table in **Figure 68B** shows this produced relatively high success rates of 97% and 95% respectively and a single misclassification in the outlier population Hazara, just 0.4% error. More precise threshold settings can be achieved by receiver operating characteristic (ROC) curve analysis where the ROC space can indicate an optimum trade off between true positives (success) and false positives (error) by finding the point closest to the upper left corner where success=1, error=0. ROC curve analysis or setting more aggressive thresholds altogether can provide the securest framework for *Snipper* to classify closely related populations. However, we also advocate expanding markers used to numbers appropriate for the levels of differentiation that exist. Clearly, differentiation of Middle East populations from Europeans and South Asians either side of the cline requires more SNPs than described for the *Eurasiaplex* test, moreover, *Eurasiaplex* SNPs were not chosen to make this differentiation. When selecting additional SNPs *STRUCTURE* and PCA can gauge the extent population groups have improved separation adding markers specific for the purpose. The 'cost' being reduced forensic sensitivity as more tests are added, although uniparental markers improve differentiation markedly and already have necessary sensitivity. We are now working on dedicated SNP sets to analyse Middle East populations and early indications suggest more than 100 SNPs will be needed. Despite this, *Eurasiaplex* in its current format provides a successful system for ancestry inference of West/Central Europeans and South Asians, with manageable error rates applying a probability threshold that allows the great majority of samples from these groups to be classified.

4.2.6 Development of a novel forensic STR multiplex for ancestry analysis and extended identity testing (Appendix 2E)

4.2.6.1 AIM-STR allele frequency distributions

The *pop.STR* database now includes HGDP-CEPH data for 12 AIM-STRs and population combinations are selectable to generate summary allele frequencies and statistics. To simplify this complex variation data, we outline allele frequencies in the four main population groups in **Figure 70**. These reveal an absence of common distribution patterns, e.g. D15S822 gives differentiations of East Asian/Native American alleles compared with Europeans: comprising common allele ranges in bimodal patterns of 14-17.3 and 19.3-23.3 repeats, respectively. However, only D14S1426 and D12S297 give similar bimodal patterns but with less well-defined differences. D3S4545, despite a wide range of observed alleles, is confined to sharp contrasts in alleles: 27, 30.3, 31.3, and 32.3.

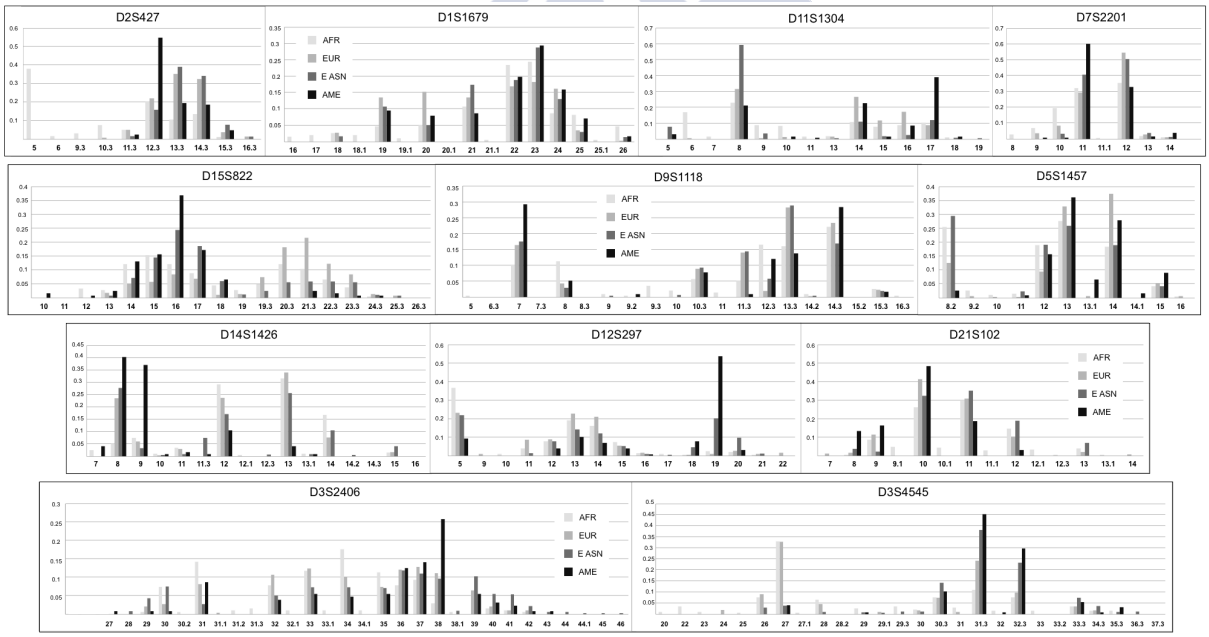


Figure 70: Summary allele frequency charts for European, African, East Asian and Native American population groups.

4.2.6.2 Multiplex performance and analysis of standard control DNAs

Figure 71 shows typical profiles obtained from three SRM-2391c DNAs (D = 3:1 mixture of A and C, respectively) and profile data indicates a degree of signal strength variation between STRs, notably D21S1432 and D7S2201 lower than ideal, but reasonably balanced peak height ratios within each STR. **Figure 72** shows an optimized set of AIM-STR allelic ladders. The peak patterns shown suggest that all alleles, including single nucleotide repeat intervals (e.g. D14S1426 11.3-12), give generally comparable signals despite a complex mix of representative alleles. Initial studies of dilution series of 9947A control DNA indicates the AIM-STR multiplex produces complete profiles at 125 picograms of DNA, comparable to ID-STR multiplex sensitivity. Furthermore, initial genotyping of challenging DNA obtained from skeletal extracts produced complete profiles with some peak imbalance but free from artefact signals and allele/locus dropout. An example profile from an extract of a femur (6 years internment, damp environment) is shown in **Figure 73**.

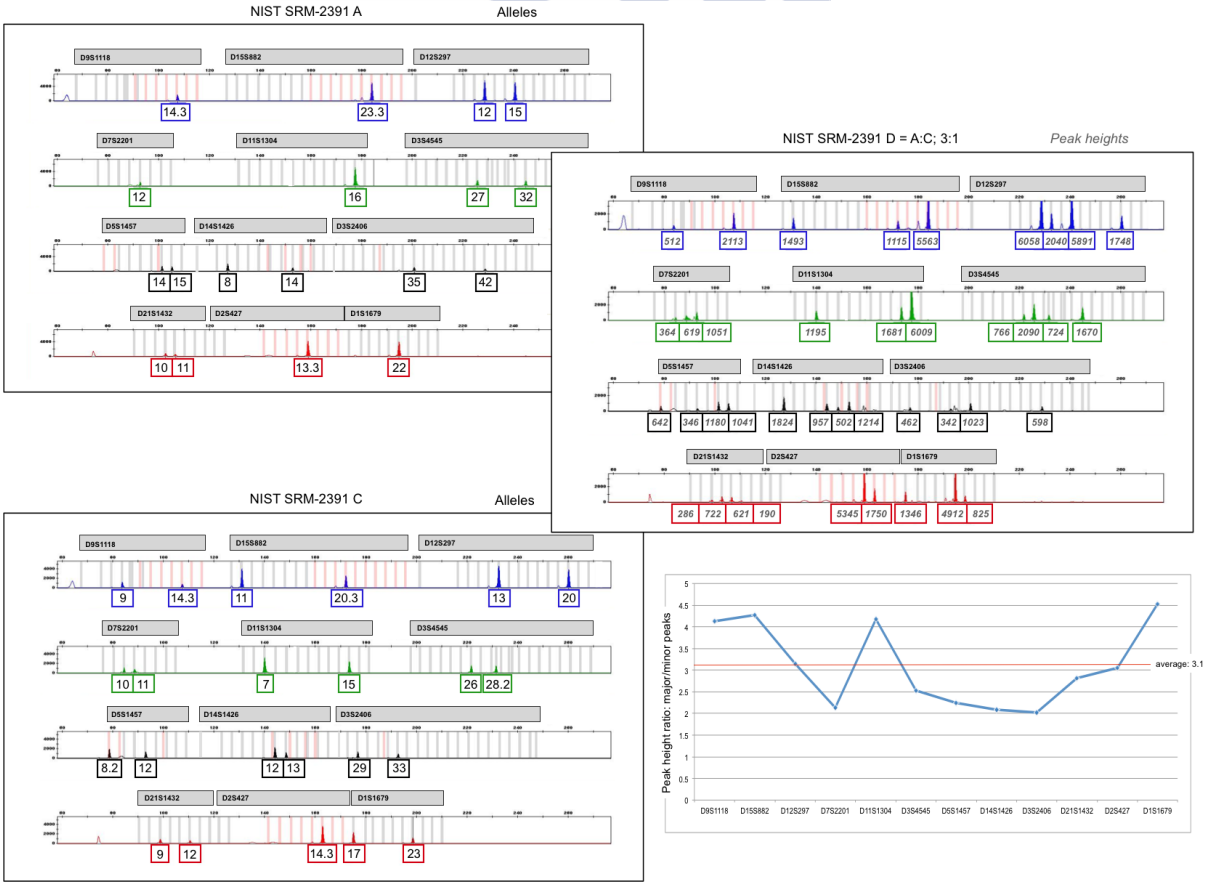


Figure 71: Three SRM-2391c reference DNA profile electropherograms for components A, C, and D. Profile of component D is a 3:1 mixture of A and C with peak heights shown rather than allele calls. Chart shows peak height ratios of profile D and table shows genotypes for unmixed components A, B, and C.

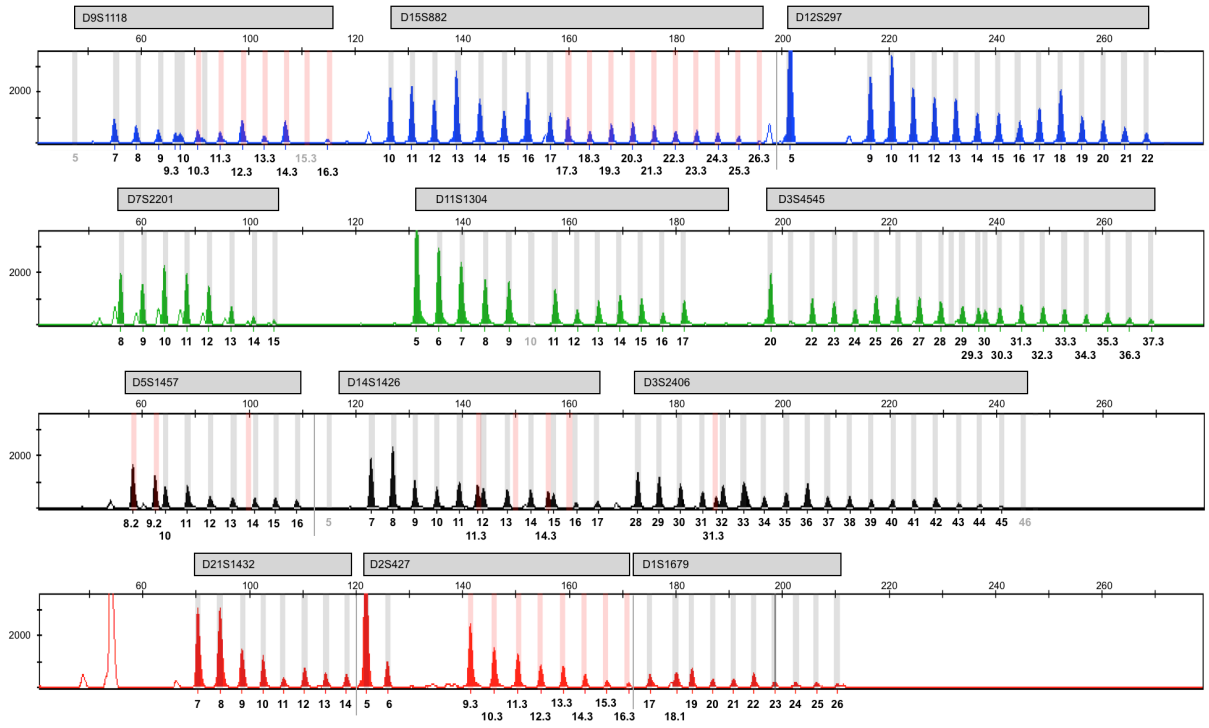


Figure 72: Allelic ladders for the *12-plex* assay. Grey numbers refer to recorded alleles not included in ladder.

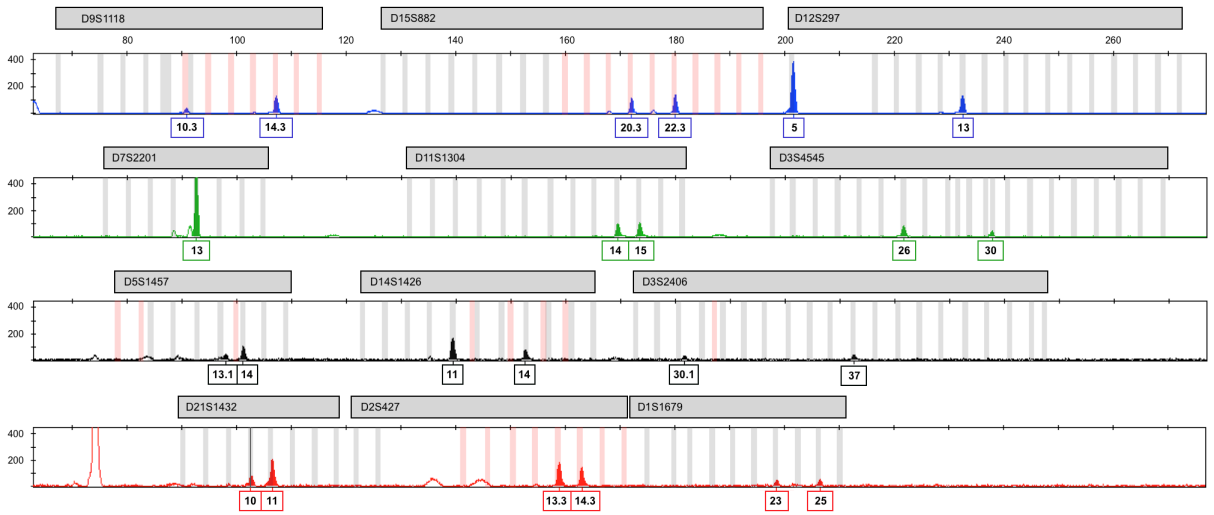


Figure 73: Electropherogram of DNA from a decomposed femur with six years internment in a damp environment.

A formal analysis of profile balance with 96 DNA samples indicate good peak height ratios (PHR) for all STRs, with the bulk of loci giving PHRs ranging from 0.75 to 0.95. However, inter-locus balance was considerably more variable, with D3S2406 and D1S1679 giving median relative fluorescence units of 600 and 900: some four to six-fold lower signal strength than the strongest STRs (all blue and first two green-labelled STRs). This suggests

some further optimization of the multiplex ratios of PCR primers is required to reach the balanced mixture interpretation system this set is designed to provide in addition to STR-based ancestry analysis. Stutter ratios mirror the shift in repeat structure seen in many of the *12-plex* components, notably the 18.3-19-19.3 alleles of D15S882. Overall, the stutter ratios, though slightly higher than those of kit STRs, show stutter peaks will not interfere markedly with mixture interpretation. Therefore, both the stutter and PHR values suggest good signal balance within any one STR, but imbalance exists between several STRs requiring further multiplex adjustment.

Genotyping of SRM-2391c D mixed source DNA represents a single experiment ahead of more extensive studies, but illustrates how mixture components with differing ancestries and therefore allele distributions can maximize peak pattern contrasts (here, European-Oceanian). **Figure 71** lists peak heights obtained in SRM-2391 D and their ratios are plotted below. Despite some variation, STRs with four peak patterns, half the peak sets observed, were closer to the mean ratio of 3.1:1, very closely matching the actual 3:1 mixture ratio of SRM-2391c D. As one critical shortcoming of forensic SNP analysis is inability to detect mixtures, extending available STRs represents the best way to improve mixture interpretation.

4.2.6.3 Sequence analysis of STR repeat structures

Repeat unit structures found from sequence analysis of a full range of alleles are summarized in **Figure 74**. Wherever possible common intermediate alleles (.2 and .3 = 2 and 3 nucleotide repeat units) were sequenced alongside normal four-nucleotide repeats. With the exception of D1S1679 all STRs showed a core GATA tandem repeat unit with nine STRs showing sequence variations where a nucleotide substitution appeared fixed in a tandem repeated series (commonly GACA) or single incidences of two or three nucleotide repeats were imbedded within runs of GATA tandem repeats. GATA-related repeats were counted to obtain the final allele designations, but only D7S2201 and D21S1432 gave simple (GATA)n repeats and it is noteworthy that both show the smallest allele size ranges. All two or three nucleotide repeats were single instances and with .3 repeat units these usually produced a series of common three nucleotide intermediate alleles exemplified by D15S882 and D3S4545. By far the greatest sequence variation was found in D3S4545 and D3S2406, showing ten and six separate repeat motifs, respectively. In D3S4545, three GACA and one

GAGA motif intersperse the core GATA repeats creating a complex and highly variable repeat region. However, D3S4545 does not show the most complex allele frequency distributions since much variability is concentrated in allele 27 in Europeans/Africans or alleles 30.3, 31.3, and 32.3 in East Asians/Native Americans. Compared to D3S4545, D3S2406 shows less motif variation but also comprises interspersed motifs, where GATG, GACG, and GACA units are located ahead of a single GATA tandem repeat series. However, D3S2406 shows much more allelic variability and finding a novel D3S2406 30.1 allele in **Figure 73** suggests this marker has a larger potential number of alleles than those we already detected – promising much higher population differentiation than the other STRs.

D9S1118:	$(GATA)_n$ $(GTA)_{0-1}$ $(GATA)_3$	Blue dye
D15S882:	$(G/AATA)_1$ $(GATA)_n$ $(GAT)_{0-1}$ $(GATA)_p$	
D12S297:	$(GATA)_n$ $(CATA)_p$ CAGA	
D7S2201:	$(GATA)_n$	Green dye
D11S1304:	$(GATA)_n$ $(GACA)_2$ $(GATA)_p$	
D3S4545:	$(GATA)_n$ $(GAT)_{0-1}$ $(GATA)_p$ $(GACA)_q$ GAGA GATA $(GACA)_r$ $(GATA)_{0-1}$ $(GACA)_s$ $(GATA)_3$	
D5S1457:	G/AAAT $(GATA)_n$ $(TA)_{0-1}$ $(GATA)_p$	Yellow dye
D14S1426:	$(GATA)_n$ $(GAT)_{0-1}$ $(GATA)_p$	
D3S2406:	$(GATG)_n$ $(GACG)_p$ $(GACA)_{0-1}$ $(GACG)_q$ $(GACA)_r$ $(GATA)_s$	
D21S1432:	$(GATA)_n$	Red dye
D2S427:	$(GATA)_n$ $(GAT)_{0-1}$ $(GATA)_p$	
D1S1679:	$(GAGA)_2$ $(GAAA)_2$ $(AAGA)_{0-1}$ $(AGGA)_n$	

Figure 74: Repeat unit sequence structures.

4.2.6.4 Population analysis and inference of ancestry of HGDP-CEPH samples

Ancestries of SRM-2391c A-C genotypes were estimated with *Snipper* by uploading European, African, East Asian, Native American, and Oceanian training sets. Likelihoods from *12-plex* genotypes were A: 160 times more likely European; B: 92 times more likely European; C: 2433 times more likely Oceanian. Combining *12-plex* and 20 ID-STR data gave likelihoods: 10 028 times more likely European; 274 times more likely European and 18 million times more likely Oceanian. Therefore, likelihoods reach sufficiently high probability levels when 32 STRs are combined, but since it is not envisaged that 12 AIM STRs would be used alone in any one forensic analysis, use of 27-32 STR data with the modified *Snipper* tool

represents the most secure forensic ancestry inference approach when mixed profiles are encountered.

Two principal approaches were used to gauge the efficiency of the 12 AIM STRs to infer the ancestry of a forensic sample: i) using *STRUCTURE* analysis to assess how closely the genetic cluster patterns of HGDP-CEPH samples matched those previously found (Rosenberg et al. 2002, Rosenberg et al. 2003) for the same individuals, and ii) using *Snipper* to reclassify HGDP-CEPH samples and measure success rates from correct assignments. With each approach the 12 AIM-STRs were tested alone, then combined with 15 or 20 core ID-STRs and finally with 34 AIM-SNPs. *STRUCTURE* analyses were: 15 ID-STRs; 12 AIM-STRs; 12+15 STRs; 12+20 STRs; 12 AIM-STRs+34 AIM-SNPs; 32 STRs+34 AIM-SNPs.

Results from four Structure analyses are summarized in **Figure 75**. In cluster plot **Figure 75A**, analysis with 12 AIM-STRs differentiates HGDP-CEPH samples into five genetic clusters at an optimum grouping of K=5, (inferred from likelihood of K: L(K) plots shown, where optimum K is suggested from the point before subsequent probabilities of mean L(K) stabilize). Cluster patterns for Africans suggest the *12-plex* markers give reduced differentiation of Africans than alternative AIM sets since this cluster is usually the most well defined. Among the five population group clusters discernible levels of co-ancestry exist as mixed membership proportions, notably certain African and American samples but these patterns diminish in Africans when ancestry informative SNPs and STRs are combined or using all 66 markers, as shown in **Figure 75B** and C. Cluster plots of **Figure 75C** also indicate a degree of differentiation is possible between European and Central South Asian populations at each K value. The best differentiation of the three Eurasian subgroups is shown in the middle K=7 plot. Interestingly, the Kalash population forms an eighth cluster (K=8) in an identical way to the K=6 differentiation of Rosenberg's original study (Rosenberg et al. 2002). A third *STRUCTURE* analysis examined 21 Eurasian populations with all 66 loci and results are shown in **Figure 75D** including the cluster plot, triangle plot, and percentage membership proportion distributions. Firstly, the cluster plot indicates all populations except the Middle East Druze can be differentiated using majority cluster memberships (i.e. >50%). These patterns are reflected in the triangle plot (where close proximity to vertices equate to high cluster memberships) with Druze and Adygei individuals forming the discernable subclusters closer to the middle of the plot. The range of percentage membership proportions estimated for Eurasians are summarized in the bar plots of **Figure 75D**, showing mean values,

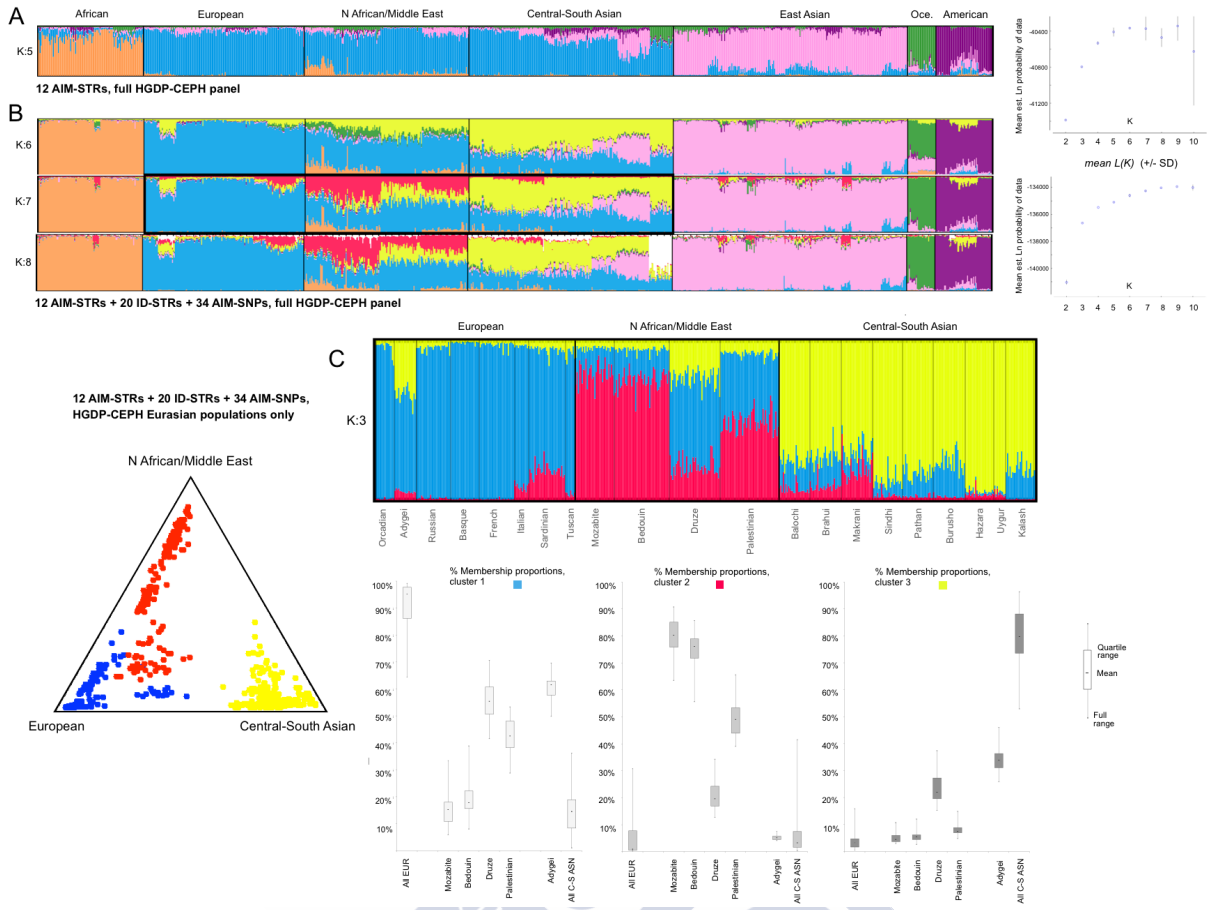


Figure 75: Four *STRUCTURE* population analyses using *12-plex* STRs. (A) *K*=5 cluster plot of full HGDP-CEPH panel analysed with *12-plex* AIM-STRs alone. (B) *K*=6 cluster plot of same populations analysed with 12 AIM-STRs and 34 AIM-SNPs. (C) Cluster plots of analysis with 32 STRs and 34 AIM-SNPs, optimum *K*=7 inferred from charts on right indicating mean estimated probability of data. (D) Cluster and triangle plots from analysis of 21 HGDP-CEPH Eurasian populations analysed separately with 32 STRs and 34 AIM-SNPs with an optimum *K*=3. Bar charts show mean, quartile, and full ranges of percentage membership proportions estimated from the same analysis.

quartile ranges, and full ranges for membership proportions. Europeans (except Adygei) and Central South Asians are combined and range between 90-100% and 80-90%, respectively, indicating European and South Asian ancestries can be differentiated using 32 STRs and 34 AIM-SNPs whereas previously this has not been possible with small-scale marker sets. The four Middle East subpopulations are plotted individually indicating Druze have low cluster-2, high cluster-1 membership proportions, making this population poorly differentiated. All Adygei and most Palestinians have membership proportions above 50% allowing successful assignment to European and Middle East ancestries, respectively. This fourth *STRUCTURE* analysis represents a successful analysis of European and Central South Asian ancestries by combining 12 AIM-STR data with 34 AIM-SNPs from a single assay plus ID-STRs always typed in forensic cases. Therefore current forensic markers can be enhanced by adding the *12-*

plex AIM-STR assay and progress can be made toward successfully analysing ancestries of closely related populations that hitherto could not be well differentiated (Phillips et al. 2009).

Bayesian classifications with *Snipper* provided a set of likelihoods of assignment to the five major population groups, analysing the HGDP-CEPH samples. The highest likelihood ratio obtained in each sample provided the assignment. The training set comprised HGDP-CEPH panel samples excluding Middle Eastern and Central South Asian populations. The percentage success and error rates obtained with each marker set combination are summarized in Table 33A-E (success rates in bold). Although applying 12 AIM-STRs (Table 33A) improves classification of East Asians and Europeans compared to 20 ID-STRs (Table 33B), Africans are poorly classified, affecting success rates when *12-plex* STRs are combined with ID-STRs (Table 33C). However, Table 33E shows the best classification performance (100% success in five groups) from forensic markers obtained in any study to date: reached when we combine 66 markers in a mixed marker approach exploiting the added flexibility of the frequency-based *Snipper* tool. This represents an improvement of the performance using the 34 AIM-SNPs alone (Table 33D) requiring a single extra assay that enhances the mixture interpretation characteristics of ID-STRs, lacking in forensic SNP genotyping tests.

Table 33: Percentage reclassification success (bold) and error of HGDP-CEPH individuals using expanding marker sets. (A-C) STRs, (D) AIM-SNPs, (E) all 66 markers combined.

			Classified as:				
			AFR	**AME**	**E ASN**	**EUR**	**OCE**
A	**12 AIM-STRs**	HGDP-CEPH Africans	**48,54**	1,94	9,71	36,89	2,91
		HGDP-CEPH Americans		**93,75**	3,13	3,13	
		HGDP-CEPH East Asians		3,08	**92,95**	3,08	0,88
		HGDP-CEPH Europeans			3,18	**95,54**	1,27
		HGDP-CEPH Oceanians			3,7		**96,3**
B	**20 ID-STRs**	HGDP-CEPH Africans	**95,15**			4,85	
		HGDP-CEPH Americans	1,56	**96,88**	1,56		
		HGDP-CEPH East Asians	6,17	0,44	**85,02**	8,37	
		HGDP-CEPH Europeans	12,10		3,18	**84,08**	0,64
		HGDP-CEPH Oceanians					**100,00**
C	**12 AIM-STRs + 20 ID-STRs**	HGDP-CEPH Africans	**81,55**		0,97	17,48	
		HGDP-CEPH Americans		**98,44**		1,56	
		HGDP-CEPH East Asians	3,96	0,44	**92,07**	3,52	
		HGDP-CEPH Europeans	5,73		1,27	**92,99**	
		HGDP-CEPH Oceanians					**100,00**
D	**34 AIM-SNPs**	HGDP-CEPH Africans	**99,03**		0,97		
		HGDP-CEPH Americans	1,56	**98,44**			
		HGDP-CEPH East Asians		6,11	**91,70**		2,18
		HGDP-CEPH Europeans				**100,00**	0,00
		HGDP-CEPH Oceanians			8,00		**92,00**
E	**32 STRs + 34 AIM-SNPs**	HGDP-CEPH Africans	**100**				
		HGDP-CEPH Americans		**100**			
		HGDP-CEPH East Asians			**100**		
		HGDP-CEPH Europeans				**100**	
		HGDP-CEPH Oceanians					**100**

4.2.6.5 Forensic informativeness of the *12-plex* assay

Although the AIM-STRs were chosen to differentiate populations, their allele frequency distributions show a full range of alleles but with frequencies that do not necessarily coincide – maintaining population differentiation while allowing their use as additional identification markers. Individual STR forensic informativeness metrics are listed in Supporting Information Table 5. To gauge forensic informativeness of the *12-plex* assay, we calculated expected cumulative RMP (RMP: probability of an identical profile in a random individual from the same population) to predict a typical result using these STRs for identification. Supporting Information Fig. 5 provides comparisons of cumulative RMPs from two widely used ID-STR sets (ESS and Identifiler) and AIM-STRs in four population groups. In all cases, ESS STRs provide the highest levels of discrimination, reaching two orders of magnitude greater RMPs than Identifiler but the AIM-STR multiplex matches or exceeds the 12 best Identifiler loci in Europeans, Africans, and East Asians. All values exceed 1 in 9 billion level of unique identity (i.e. RMPs above a global population size). Therefore the AIM-STR assay provides a straightforward but powerful forensic identification system capable of extending current STR sets when applications such as complex relationship testing require additional power.

4.2.6.6 Discussion

The studies of Rosenberg et al. showed that several hundred STRs differentiate a global sample panel into groups closely matching their geographic distribution (Rosenberg et al. 2002). Rosenberg went on to demonstrate STR numbers could be brought down considerably by choosing the most ancestry informative loci (Rosenberg et al. 2003). Such reduced-scale marker sets should be ideal for forensic application but they consist of dinucleotide repeat STRs wholly unsuitable for forensic analysis due to very high stutter peaks. Therefore, it is appropriate to assess the informativeness of the best tetranucleotide STRs for the same purpose and gain more balanced profile signals by building a manageable multiplex. Our results indicate the reduction in population informativeness in going to tetranucleotide loci creates less powerful microsatellite-based sets for differentiating populations. However, as binary SNPs developed in parallel to STRs for forensic ancestry inference cannot detect and

analyse mixed DNA samples our tetranucleotide AIM-STRs provide the best alternative in such commonly encountered situations. Furthermore, although existing ID-STRs would not be expected to differentiate populations well, they do have this characteristic to some degree. A recent study of CODIS STRs (core US loci) analysing global populations (Silva et al. 2012) revealed that differentiations can be made, albeit with half the power possible from dedicated markers such as Rosenberg's dinucleotide STR subsets. Since core forensic ID-STRs are always typed in routine profiling, our strategy to find the most ancestry informative tetranucleotide STRs then combine these with core STRs with the option to add AIM-SNPs from a single assay creates a powerful final set of markers showing complementary characteristics. At the time of writing, core ID-STR sets are being expanded to 21-23 loci (Powerplex 21 and Promega Fusion/AB GlobalFiler, respectively) further increasing the likely population informativeness of standard STRs. Analysis of the resulting mixed marker data is possible with both *STRUCTURE* and the enhanced *Snipper* tool, adapted to handle multiple allele loci, with *Snipper* offering the advantage of assessing single profiles in real time and providing HGDP-CEPH-based training set data for all the STRs assessed in this study.

Although we have reservations about use of cluster membership proportions to differentiate closely related populations, our studies reveal that Europeans and Central South Asians can be analysed in this way. Combining 66 markers of different characteristics improves differentiation of these Eurasian population groups compared to previous marker sets, although Middle East populations remain very difficult to differentiate without recourse to much larger scale, mainly SNP-based multiplexes – as our on-going efforts to develop dedicated Middle East-informative AIM-SNP sets suggest.

Mixed profile analysis depends on good signal balance but equally important is securing the most points of reference possible (e.g. four alleles in any one STR are easier to interpret than an imbalanced peak pair). Therefore the STRs outlined offer considerable additional scope to improve mixture analysis. As the AIM-STRs also give comparable levels of discrimination to existing forensic markers for routine identification purposes, the novel *12-plex* assay developed could be suitable for a range of forensic analyses in addition to ancestry inference.

Beyond forensic applications, ancestry analysis with small-scale tests is a different field with distinct needs. The forensic ID-STRs we included, that bring a degree of extra

population differentiation, may not be available, although these kits are by design, efficient, and easy to run. As already noted, forensic ID-STR multiplexes are expanding – certain to increase their population differentiation power significantly. However, we would suggest the optimum strategy for simple ancestry analysis outside of forensic applications is to bring together the AIM-STRs described here with small-scale AIM-SNP tests (Phillips et al. 2007b) plus the best dinucleotide AIM-STRs, free from the need to reduce stutter to controllable levels. The 36 STRs identified by Londin et al. (Londin et al. 2010) are very well suited for this purpose and in an identical approach to our study, Londin developed three 12-plex tests labelled with the same CE dyes.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

BLOCK 3

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

5 Block 3 – Into NGS

This third block focus on the possible use of next generation sequencing, as it allows the analysis of hundreds of markers in a single assay. Three different questions were made: can we still use the existing assays? How good is the technology? Which are the best markers? To answer these questions, three studies were performed and will be described in the following sections including material and method as well as results and discussion.

5.1 MATERIAL AND METHODS

5.1.1 A SNaPshot of next generation sequencing for forensic SNP analysis (Appendix 3A)

5.1.1.1 Samples

Three DNA templates were used in this study; female human cell line control DNA 9947A, male human cell line control DNA 007 (both Life Technologies: LT) and one male donor sample (S1). S1 was extracted by a standard phenol/chloroform method (Liu and Harada 2013). All templates were quantitated using Quantifiler® Human DNA Quantification Kit[8] according to the manufacturer's protocol. Template DNA of 0.1, 0.2 and 0.3 ng was used for amplification of all samples and a non-template negative control was included with each PCR.

5.1.1.2 Preparation of PCR amplicons

Published PCR primer sequences and protocols were used for amplifications of each of the following assays: SNP*for*ID *52plex* (Sánchez et al. 2006), SNP*for*ID *34-plex* (Fondevila et

8 Applied Biosystems, Quantifiler® Kits: User's Manual, Life Technologies Corporation, Carlsbad, CA, 2012

al. 2013), *Eurasiaplex* (Phillips et al. 2013b), *Pacifiplex* (see sections 4.1.2 and 4.2.2) and *IrisPlex* (Walsh et al. 2011a). The *Pacifiplex* PCR amplification reaction was performed in a 10 µL total volume containing: 1× GeneAmp PCR Gold buffer II (LT), 8 mM $MgCl_2$, 1.6 µg of BSA, 1.875 mM dNTPs, 0.5 units of AmpliTaq Gold DNA Polymerase (LT), 0.2 µM PCR primers (Table 12) and 0.1, 0.2 or 0.3 ng DNA. PCR cycling comprised denaturation for 5 min at 95ºC followed by 35 cycles of 30 sec at 95ºC, 40 sec at 61ºC, 60 sec at 72ºC, and then a final extension of 15 min at 72ºC.

5.1.1.3 SNaPshot analysis of PCR amplicons

PCR efficiency was evaluated through SBE analysis of the amplicons. PCR products were purified using shrimp alkaline phosphatase with exonuclease 1 (ExoSAP-IT™: GE Healthcare), followed by SNaPshot primer extension. SBE products were then purified with shrimp alkaline phosphatase (SAP: GE Healthcare). SBE primer sequences, purification and protocols were as previously described: SNP*for*ID *52plex* – in two separate SBE multiplexes (Sánchez et al. 2006), SNP*for*ID *34-plex* (Fondevila et al. 2013), *Eurasiaplex* (Phillips et al. 2013b) and *IrisPlex* (Walsh et al. 2011a). *Pacifiplex* PCR products were purified with 0.65 µL of ExoSAP-IT™ added to 1.25 µL PCR product. Samples were incubated at 37ºC for 45 min then enzymes were heat-inactivated at 85ºC for 15 min. SBE reactions were performed in 3 µL volumes combining 1 µL SNaPshot Multiplex Ready Reaction Mix with 1 µL premixed SNaPshot primers (Table 12) and 1 µL purified PCR products. SBE thermocycling conditions were: 30 cycles of 10 sec at 96ºC, 5 sec at 55ºC, 30 sec at 60ºC. SBE products were purified by adding 1.3 µL of SAP to the whole reaction volume and incubating at 37ºC for 80 min followed by heat-inactivation of the enzyme at 85ºC for 15 min.

SBE products were analysed by capillary electrophoresis (CE) using aliquots of 0.25–3 µL of purified SBE product added to a mix of Hi-Di™ Formamide (LT) and GeneScan™ 120 LIZTM Size Standard (LT). CE was performed using standard SNaPshot electrophoresis conditions in an ABI Prism 3130*xl* Genetic Analyzer (LT) with 36cm capillary arrays and POP-4™ polymer (LT). Electropherograms were analysed using GeneMapper1 ID-X (LT).

5.1.1.4 Sanger sequencing of PCR amplicons

Sanger sequencing of the positions and context sequences of twelve SNPs with discordant SNaPshot and PGM genotypes was performed. Two SNPs (rs1493232 and rs1029047) required specific primer designs binding outside the previous amplicon spans (Sánchez et al. 2006): rs1029047-F GGAAATTTGTAGATCATTCTGGGTA; rs1029047-R CACTGAAGTTTCAGGGCTCA; rs1493232-F GGAACCTTCTAAATCTCCTTGATG; rs1493232-R TGAGAAAACTTGAAAATGTTGGTT. Established PCR primer designs were used for the remaining ten SNPs (Sánchez et al. 2006, Phillips et al. 2007b, Fondevila et al. 2013, Phillips et al. 2013b). Primers were diluted to their multiplex concentrations in singleplex PCRs. After PCR, amplified products were treated with EXO I and SAP (Illustra™ ExoStar™ 1-Step, GE Healthcare), using: 2.5 µL amplified product and 1 µL enzyme mix. The PCR products were incubated for 15 min at 37ºC followed by enzyme inactivation at 85ºC for 15 min in an ABI GeneAmp® PCR system 2700 thermal cycler (LT). Sanger sequencing was performed in each direction using BigDye® Terminator v3.1 Cycle Sequencing Kit (LT), using: 0.5 µL BigDye kit, 2 µL BigDye buffer, 2 µL water, 1 µL 10 µM sequencing primer and 1.5 µL purified PCR product. Cycling comprised: 96ºC for 1min, 32 cycles of 96ºC for 15 sec, 50ºC for 10 sec and 60ºC for 1 min, 60ºC for 4 min. Sequencing products were then purified with Montage™ SEQ96 clean up kits (Millipore) following the manufacturer's protocol. Sequencing products were electrophoresed on an ABI Prism 3730*xl* DNA analyser (LT) using POP-7™ polymer (LT) and standard electrophoretic conditions and analysed with Sequence Analysis version 5.2 (LT).

5.1.1.5 NGS library preparation

The same aliquot of PCR products from each assay used for SNaPshot analysis were purified individually using the QIAquick® PCR purification kit (Qiagen) following the manufacturer's protocol and eluted in 30 µL of elution buffer. Equal 10 µL volumes of purified PCR products from the five assays were pooled together for each template and input amount (i.e. containing PCR product from five assays per input amount). The PCR negative controls for each assay were also pooled to form one combined negative control. Each pooled sample contained a total of 145 amplicons with nine SNPs in multiple assays (rs1024116,

rs1335873, rs1886510, rs204041, rs3827760, rs722098, rs917118, rs12913832 and rs16891982), two with differing primer designs (rs12913832 and rs16891982). The amplicons for the 136 unique SNPs had sizes ranging from 51 to 156 bp. The 10 samples prepared from three DNA templates are outlined in Table 34.

Table 34: Total sequence coverage obtained from 136 unique SNPs for each of ten samples analysed using the Ion 314™ v1 chip.

Sample no.	Sample	DNA template amount (ng)	Coverage
1	9947A	0.1	94,997
2	9947A	0.2	107,311
3	9947A	0.3	71,227
4	007	0.1	93,983
5	007	0.2	70,068
6	007	0.3	79,270
7	S1	0.1	54,437
8	S1	0.2	63,024
9	S1	0.3	94,351
10	Neg control	-	1365
	Total		730,033

As amplicons were relatively short, DNA fragments were not sheared prior to library preparation. The Ion P1- and A-barcoded adaptors were ligated to the 5' and 3' fragment ends and nick translated using the Ion Xpress Plus Fragment Library Kit (LT) and Ion Xpress Barcode Adaptors 1-16 Kits (LT) following the manufacturer's protocols[9]. Adaptor-ligated fragment libraries were purified using Agencourt® AMPure® XP beads (Beckman Coulter) following the manufacturer's protocols. Barcoded libraries were then amplified and quantitated using the Agilent 2100 Bioanalyzer (Agilent Technologies) prior to pooling in equimolar 13 pM quantities to balance the representation of each barcoded library.

[9] Life Technologies, Ion Xpress™ Plus gDNA and Amplicon Library Preparation User Guide, Life Technologies, Foster City, CA, 2011.

5.1.1.6 Template preparation

A volume of 25 µL of the diluted, barcoded, pooled library was used for Ion Sphere™ Particle (ISP) preparation. Emulsion PCR was performed using the Ion OneTouch™ 200 Template Kit v2. Template-positive Ion Sphere particles were then isolated on the Ion One Touch™ ES (enrichment system) according to the manufacturer's protocol[10]. Template-positive ISPs were loaded directly onto two Ion 314v1 chips (samples 1-6 on one chip and samples 7-10 on a second chip) and sequenced using the Ion PGM™ Sequencing 200 kit according to the manufacturer's recommended protocol[11].

5.1.1.7 Ion Torrent data analysis

Signal Processing, basecalling and barcode deconvolution were performed using Torrent Suite software v 3.0 (LT). Alignment to the Hg19 human reference genome (Genome Reference Consortium Human Build 37: GRCh37[12]) was performed using TMAP (LT) (Homer et al. 2014) and the alignment output to BAM format. The Torrent Suite software was also employed to generate relevant run metrics such as the total number of sequences per sample. Individual amplicon coverage metrics were calculated using the Torrent Suite Coverage analysis plug-in v 3.0.1 (LT). Variants were identified with Torrent Variant Caller software v 3.0 (LT) using default *Germline* parameters and the results output to VCF format. Amplicon genomic positions and known variant coordinates were defined in BED format.

5.1.1.8 Allele call thresholds

A simple allele designation was applied, based on allele frequency thresholds. SNP amplicons with allele frequencies of 90% or greater were deemed to be homozygous for that allele, and those with allele frequencies between 10% and 90% heterozygous. Allele frequencies of less than 10% were ignored (as the alternate allele was designated a

10 Life Technologies, Ion OneTouch™ 200 Template Kit v2 User Guide, Life Technologies Corporation, Carlsbad, CA, 2012
11 Life Technologies, Ion PGM™ 200 Sequencing Kit User Guide, Life Technologies Corporation, Carlsbad, CA, 2012
12 Genome Reference Consortium, Human Genome Overview: GRCh37, National Institutes of Health, 2014 Available from: http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/ (accessed 27.06.14)

homozygote). Ion Torrent PGM and SNaPshot genotypes were compared and discordant genotypes were analysed by Sanger sequencing.

5.1.1.9 Statistical analysis

Statistical analysis was performed using IBM SPSS Statistics (v21). Non-parametric statistical tests were applied as a result of the skewed amplicon coverage distribution. The distributions of GC content (%) and amplicon length (bp) of the amplicons with the highest and lowest 10% of coverage were compared with a Mann-Whitney U test for the (null) hypothesis that there was no difference between them. The distributions of combined coverage (across all three template amounts) for amplicons in each of the assays (*52-plex*, *34-plex*, *Eurasiaplex*, *Pacifiplex* and *IrisPlex*) were compared with a Kruskal-Wallis test for the (null) hypothesis that there was no difference between them.

5.1.2 Inter-laboratory evaluation of SNP-based forensic identification by massively parallel sequencing using Ion PGM™ (Appendix 3B)

All Ion PGM™ protocols followed published laboratory guidelines [13,14,15,16,17]. The term sample is used here for DNA extracts that were amplified then prepared for Ion PGM™ in different ways (i.e. several samples may be used from one donor). The term run refers to sequencing tests made using one Ion PGM™ chip, combining multiple samples. The term analysis is used to describe sequencing of a specific DNA sample forming part of a run. *Somatic* and *Germline* analysis parameter settings are distinguished from the biological terms using capitals. The term allele frequency is used in Ion PGM™ analysis software, describing how many sequence reads carry each allele per SNP. To avoid confusion with the population genetics term we use *allele read frequency* (ARF).

---

13 Thermo Fisher Scientific, Life Technologies: Sequencing technology solutions, 2014. Available from: https://www.lifetechnologies.com/au/en/home/life-science/sequencin/sequencing-technology-solutions.html (accessed 06.14).

14Thermo Fisher Scientific, Life Technologies: Ion PGM™ system for next-generation sequencing, 2014. Available from: https://www.lifetechnologies.com/au/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-run-sequence/ion-PGM-system-for-next-generation-sequencing.html (accessed 06.14).

15 Thermo Fisher Scientific, Life Technologies: Ion AmpliSeq™ library preparation user guide, July 2013.

16 Thermo Fisher Scientific, Life Technologies: Ion OneTouch™ 200 Template Kit v2 user guide, 2012.

17 Thermo Fisher Scientific, Life Technologies: Ion PGM™ 200 Sequencing Kit user guide, 2012.

5.1.2.1 DNA samples, extraction of DNA and preparation of artificial mixtures

Common DNAs were used to measure genotyping concordance or assess consistency of sequence quality across three laboratories. These DNAs comprised: i) six voluntary staff donors (S1-S6) that could be repeatedly analysed and exchanged between laboratories; ii) standard 9947A and 007 forensic controls; iii) Coriell cell-line control DNAs that allowed checks against online genotype data published by 1000 Genomes and Complete Genomics (CG) projects (comprising: NA06994; NA07000; NA07029; NA18498; HG00403; NA10540; NA11200). These DNAs provide comparisons of three independent SNP genotyping systems using NGS sequencing: 1000 Genomes mainly used Illumina HiSeq (1000GPC 2012) and CG a proprietary DNA nanoarray method (Drmanac et al. 2010).

Dilutions of 9947A and 007 DNAs assessed the forensic sensitivity of Ion PGM™, using 10 ng, 1 ng, 100 pg, 50 pg and 25pg of DNA amplified with varying PCR cycle numbers, as outlined in Table 35. Two runs used eight picomolar (pM) library pools (i.e. following standard Ion AmpliSeq™ library preparation guidelines). Another three runs used libraries pooled at ~26 pM dilution to determine if increasing library concentrations enhanced genotyping of low-level DNA. Input DNA <1 ng was either amplified in 25 cycles alone, or with 5 extra amplification cycles after library preparation. Two approaches assessed re-amplification: i) re-amplify half the prepared library per sample and compare to no re-amplification; ii) prepare separate libraries with and without re-amplification for each sample. Samples were quantified for pooling with LT Ion Library Quantitation Kit.

The ability of Ion PGM™ to detect mixed DNA was evaluated with mixtures of male-female DNAs S5 and S6 at ratios 1:9, 1:3, 1:1, 3:1, 9:1. Each mixture ratio was prepared once, then two libraries constructed for each. The two differently-barcoded libraries of each ratio were combined in one template preparation step and sequenced on a single Ion 316™ chip.

One ancient male DNA sample extracted from 12th century archaeological remains (S7 or aDNA) was analysed. The skeletal preservation conditions from the site in Volders, Tyrol, Austria are detailed in (Bauer et al. 2013). Sample S7 was analysed in two separate PCRs with maximum input DNA (450 pg quantified with LT Quantifiler Duo), using 25 PCR cycles and 25 PCR + 5 library re-amplification cycles. Although this sample lacked reference genotypes, consistency of SNP genotyping was checked between analyses.

Table 35: Sensitivity study DNA dilutions added to five sequencing runs, their pooling concentration, input quantities and PCR cycling regimes. Five additional cycles of amplification after library preparation, applied to the lowest level DNA, is denoted by '+5'.

Run	Cycles	8 pM		26 pM		
		lab1-A	lab1-B	lab1-E	lab1-F	lab1-C
9947A 10 ng	18			●		
9947A 1 ng	21	●		●		
9947A 100 pg	21			×		
9947A 100 pg	25	●				
9947A 50 pg	25	●		×		
9947A 25 pg	25			×		
9947A 100 pg	21 + 5				×	
9947A 100 pg	25 + 5	●				
9947A 50 pg	25 + 5	●			×	
9947A 25 pg	25 + 5				×	
007 10 ng	18			●		
007 1 ng	21	●		●		
007 100 pg	21			×		
007 100 pg	25	●	Δ			Δ
007 50 pg	25	●	Δ	×		Δ
007 25 pg	25			×		
007 100 pg	21 + 5				×	
007 100 pg	25 + 5	●	Δ			Δ
007 50 pg	25 + 5	●	Δ		×	Δ
007 25 pg	25 + 5				×	

×: Same sample re-amplified; Δ: library replicates.

5.1.2.2 Ion PGM™ library and template preparation, enrichment and sequencing

HID-Ion Ampliseq™ Identity Panel v2.2 libraries were constructed with Ion AmpliSeq™ Library Kit 2.0 following manufacturer's protocols[15]. Prior quantification of DNAs used Qubit® dsDNA HS Assay Kit diluting samples (not all) to guidance inputs of 10 ng in ≤ 6 µL. Targets were amplified as recommended for 196 primer pairs with 18-21 cycles of PCR. After partial digestion of primer sequences, Ion Xpress™ Barcode Adapters were ligated for tagging and resulting ligation products purified with Agencourt AMPure XP magnetic beads. Library quality was checked with either Qubit® dsDNA HS Assay Kit, Agilent® High Sensitivity DNA Kit or Ion Library Quantitation Kit to equalise a final library of 100 pM in ≥ 20 µL[15].

Template preparation used Ion OneTouch™ 200 Template Kit v2, following manufacturer's protocols[16]. After recovering template-positive Ion Sphere particles (ISPs), Ion Sphere™ Quality Control Kit was used to ensure 10-30% templated ISPs before enrichment with Ion PGM™ Enrichment Beads, following manufacturer's protocols. Sequencing was performed using Ion PGM™ Sequencing 200 Kit v2 and Ion 314™ or 316™ chips (both types either v1 or v2) following manufacturer's protocols[17].

5.1.2.3 Data analysis

Data analysis used Torrent Suite™ 4.0.2 (herein TS) and HID_SNP_Genotyper 4.0.1 plugin (herein Genotyper) with low stringency parameter settings[18]. We applied *HID_SNP_v2.2.2_hotspots.bed* plus *HID_SNP_v2.2.2_targets.bed* files, identifying SNPs with genome build Hg19. Genotyper makes variant calls using posterior probabilities calculated for each possible genotype in similar fashion to GATK (McKenna et al. 2010). Posterior probabilities are computed from genotype likelihoods (using Phred quality scores and prior probabilities), accounting for read depth and minimum allele frequency thresholds to report quality scores (QUAL values of 0 to several thousand). SNP genotypes are called when they pass a quality score plus user-defined sequence filter thresholds, or are given as "NN"/"N" no-calls.

Genotyper output comprises a web-based graphical overview and two report files: a custom-format text file plus a variant call format (vcf) file with SNP details. The text file lists genotype calls with corresponding quality *p*-values, total sequence coverage from forward and reverse sequence reads, number of calls for all four bases and number of no-calls at each SNP position. For this study all SNP data processing of both Genotyper files was made using *R* (v3.0.3, 2014-03-06) (Gentleman et al. 2004, Team 2011).

18 Thermo Fisher Scientific, Life Technologies: Torrent Suite™ software 4.0.2. user guide. November 2013

5.1.3 Building a forensic ancestry panel from the ground up: the EUROFORGEN Global AIM-SNP set (Appendix 3F)

5.1.3.1 Sources of AIM-SNPs and allele frequencies in the five main global population groups

Candidate AIM-SNPs were compiled from three sources: i) SNP sets previously developed for a range of forensic ancestry test initiatives at Santiago (USC); ii) allele frequency screens of the Stanford HGDP-CEPH 650K SNP dataset (Li et al. 2008) – identifying SNPs with the highest divergence between targeted population comparisons by finding the top 5% most differentiated in each case; and iii) AIM-SNP lists published both before and after availability of whole genome scan (WGS) high-density SNP arrays. This third approach collected a large number of SNPs highly informative for ancestry but previously systematically excluded from WGS SNP sets due to their lack of association power (e.g. rs16891982 shows a highly differentiated G allele, close to fixation in Europeans and therefore uninformative for correlated variation in SLC45A2 where it is sited). Mainly, SNPs identified as good AIMs prior to use of WGS arrays are commonly part of forensic ancestry sets and include many of the best population differentiators such as rs2814778, rs16891982, rs1426654 and rs3827760, all absent from WGS sets. As Stanford HGDP-CEPH SNP data uses Illumina 650K WGS array genotyping, the second strategy appears contradictory. However, we identified Oceanian and Native American informative SNPs from Stanford HGDP-CEPH data, as Illumina selected the 650,000 loci based solely on European, African and East Asian variability. Sources of SNP variation we scrutinized are outlined in **Figure 76**. Published SNP sets were: 178 of the DNAprint Ancestry-by-DNA forensic set of Halder et al. (2008); 128 of a American population analysis panel (Kosoy et al. 2009, Kidd et al. 2011); 445 of the Latin American Cancer Epidemiology (LACE) group's American population analysis panel of Galanter et al. (2012); 47 of the forensic ancestry panel of Kersbergen et al. (2009), and; 55 listed in FROGkb, forming several Kiddlab forensic ancestry panels (Nievergelt et al. 2013). AIM-SNPs from previous USC developments comprised: 34 of an established forensic ancestry panel (Phillips et al. 2007b, Fondevila et al. 2013); 28 of a European-African admixture detection panel (unpublished); 25 of a CT-SNP set developed for enhanced SNaPshot peak balance (unpublished); 46 of a USC-Applied

Biosystems Genplex forensic AIM-SNP set, developed in 2007/8 but not released as a kit (Phillips et al. 2007a), and 48 of a Sequenom iPLEX® case-control association study (CCAS) stratification adjustment panel (Borel et al. 2012). A further ~260 American and Oceanian-informative SNPs were identified by screening Stanford HGDP-CEPH data, including 28 SNPs already forming a dedicated Oceanian-informative forensic SNaPshot multiplex, termed *Pacifiplex* (see sections 4.1.2 and 4.2.2). Nine Oceanian and six East Asian-informative SNPs were novel selections for this study. The above ten sets provided 1031 SNPs for scrutiny. There was some SNP commonality between USC and independently published forensic ancestry sets, comprising: rs16891982, rs2814778 in Kiddlab's FROGkb set and DNAprint; rs12913832, rs1426654, rs2471552, rs2715883 in the FROGkb set (only rs12498138 in the published Kiddlab set (Nievergelt et al. 2013)) and rs3827760 in Kersbergen's set.

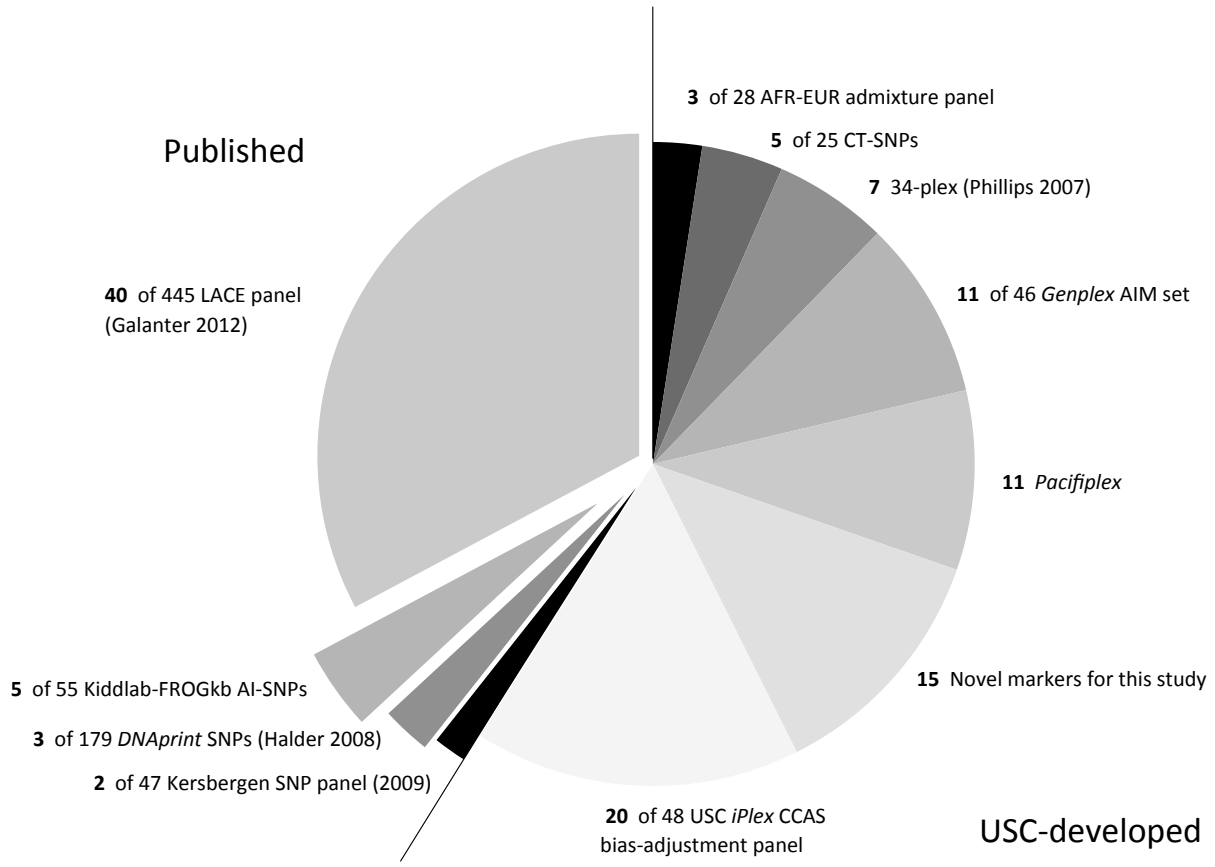


Figure 76: Sources of AIM-SNP candidates and numbers selected from each source.

Candidate SNP allele frequencies were collected for African (herein AFR), European (EUR) and East Asian (E ASN) reference populations in 1000 Genomes using the USC ENGINES online browser (Amigo et al. 2011). The ENGINES portal accesses ~28 million variant sites in 14 populations but we chose the ten unadmixed populations combined in three groups to assess candidate's allele frequency distributions. Frequencies were compiled from: 97 Luhya in Webuye, Kenya (LWK) and 88 Yoruba in Ibadan, Nigeria (YRI) combined as 185 AFR; 87 Utah residents with north and west European ancestry from the CEPH collection (CEU), 93 Finnish in Finland (FIN), 88 British in England and Scotland (GBR), 98 Tuscans in Italy (TSI), 14 Iberians from Spain (IBS) = 380 EUR; 89 Japanese in Tokyo, Japan (JPT), 97 Han Chinese in Beijing, China (CHB), 100 Han Southern Chinese (CHS) = 286 E ASN. 1000 Genomes data of four admixed populations: 61 individuals of African ancestry in Southwest USA (ASW), 66 individuals of Mexican ancestry in Los Angeles, California (MXL), 55 Puerto Ricans in Puerto Rico (PUR) and 60 Colombians in Medellín (CLM) was accessed later to assess the ability of selected AIM-SNPs to gauge admixture. American (herein AME) and Oceanian (OCE) allele frequencies were collected from Stanford HGDP-CEPH SNP data using the USC *SPSmart* online browser (Amigo et al. 2008b) to access 14 Brazilian Karitiana, 8 Brazilian Surui, 7 Colombian Piapoco, 21 Mexican Maya, 14 Mexican Pima = 64 AME plus 11 Bougainville Melanesian, 17 Papua New Guinean = 28 OCE.

5.1.3.2 Selection criteria for the final AIM-SNP set

Relative population differentiation power for each candidate was estimated by ranking SNPs by allele frequency differential (or delta: δ), given by the absolute value of the frequency of an allele in one population minus the frequency in another. Pairwise population comparison deltas were summed for the three main groups, e.g. AFR = (AFR-EUR δ) + (AFR-E ASN δ). Delta correlates closely to the other widely used population differentiation metrics: F_{ST} and Rosenberg's *In* Divergence (Rosenberg et al. 2003), e.g. Fst≈δ/(2-δ). F_{ST} and *In* metrics were applied once markers had been ordered by their population-specific differentiation values, but note F_{ST} and *In* normally measure overall population divergence, whereas summed δ estimates provide the optimum way to identify best markers for any one pairwise population differentiation. This is particularly important when selecting SNPs to differentiate closely related populations that share recent demographic history and therefore

show reduced divergence, as demonstrated by most E ASN-AME allele frequency differences. OCE-informative SNPs had not been identified in any previously compiled AIM-SNP sets so we accessed USC lists of the top 5% most divergent markers compiled to construct *Pacifiplex*.

A minimum one megabase (Mb) inter-marker spacing filter was applied to prevent multiple, highly divergent SNPs from the same genomic segment (i.e. in linkage disequilibrium: LD) occupying the top of each population differentiation ranking. Choosing multiple markers with limited recombination between them can bias co-ancestry proportion estimates in admixed individuals. Furthermore, LD can sometimes extend beyond 1 Mb – notably in the very high extended haplotype homozygosities seen in the LCT region due to rapid recent evolution in North Europe (Bersaglieri et al. 2004). However, such instances are well documented, so we checked LD using HapMap D' estimates between only the most closely sited SNP pairs. The same phenomenon of allele frequency correlations in closely sited SNPs was exploited to broaden the choice of alternative candidates when first-strike markers failed in multiplex. We previously observed that a series of SNPs with identical allele frequencies occurred in succession in a large number of locations in the human genome (Costas et al. 2005). Therefore alternative markers are often available to select with no change in their differentiation power. We used the *SPSmart* chromosome segment query option when alternative SNPs were required and one example in the VRK1 gene is shown in **Figure 77**.

A key goal was balancing the degree of differentiation as equally as possible amongst all five-population groups. To achieve this we calculated the population-specific Divergence (PSD) *In* value in each population group for each SNP candidate. PSD is equivalent to the locus specific branch length (LBSL) *In* measurement described by Galanter et al. (2012) and used to balance AFR-EUR-AME population divergence in the LACE SNP set. PSD values were calculated by uploading 1000 Genomes and Stanford HGDP-CEPH SNP genotypes into the *Snipper* forensic ancestry analysis portal (Phillips et al. 2007b) – deriving each SNP's PSD by pairwise labelling of uploaded populations: i.e. AFR:non-AFR (all other groups), EUR:non-EUR, etc. *Snipper* calculates Divergences from uploaded SNP data, but values are higher than Rosenberg's *In* (*In* = *Snipper* Divergence × 0.69), with cumulative values obtained by addition. We first identified the most powerful OCE, then AME-informative SNPs followed by those of the other three groups, carefully gauging the cumulative values until reaching a point of PSD convergence while aiming for ~120-140 SNPs in combination.

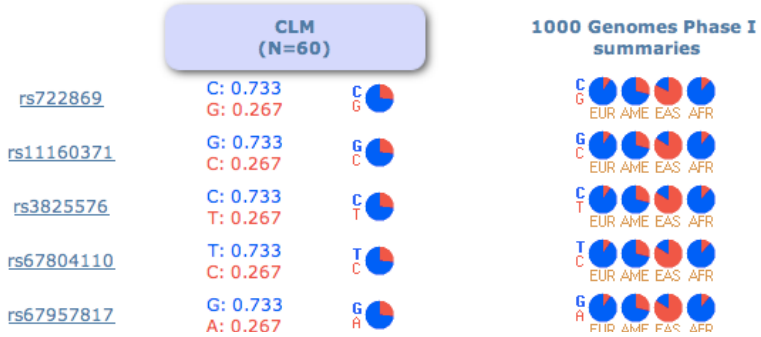


Figure 77: Examples of SNP substitutes obtained using the *SPSmart* chromosome segment query – providing alternative SNPs available to swap in when first-strike candidates fail. AIM-SNPs in the E ASN-divergent VRK1 chromosome segment: rs722869 was the Global SNP set selection, but the other four show identical allele frequencies.

5.1.3.3 Marker subsets for smaller-scale multiplexed forensic tests

To aid development of ancestry panels for established forensic technologies such as SNaPshot, the completed set of balanced divergence AIM-SNPs (herein the Global AIM-SNP set) was systematically reduced in size to a minimum number of markers capable of successfully classifying AFR-EUR-E ASN, these plus AME, then these plus OCE, while keeping the PSD values balanced throughout each stage of reduction. Additionally, since forensic NGS multiplex designs may opt to combine core STRs, identification SNPs and pigmentation-predictive SNPs alongside AIM-SNPs, we also compiled SNP sets suitable for such purposes into sets aiming for ~90 and 50–60 components.

5.1.3.4 Tri-allelic SNPs included in the selection process

We previously identified several tri-allelic SNPs exhibiting highly skewed allele frequency distributions amongst the five groups. We included SNPs: rs4540055 and rs5030240, already part of the *34-plex* test that successfully identified mixed source DNA from donors with different ancestries (notably when third alleles show contrasting common and rare frequencies in contributor populations (Fondevila et al. 2013)). Unlike WGS techniques, sequencing or iPLEX and SNaPshot SBE tests reliably detect three alleles at substitution sites, so another four tri-allelic SNPs were added and their HGDP- CEPH allele frequencies are summarized in **Figure 78**.

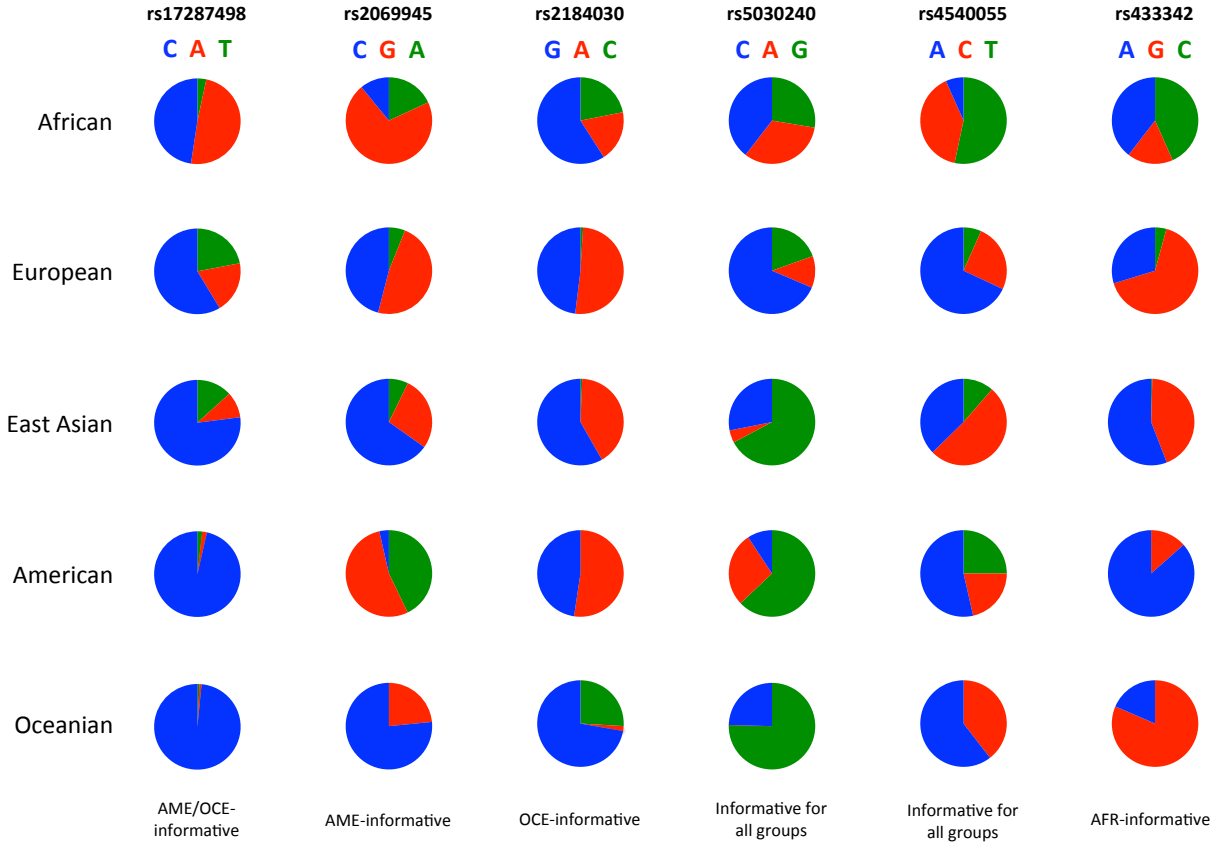


Figure 78: HGDP-CEPH allele frequency distributions in five population groups of the six tri-allelic SNPs included in the Global AIM-SNP set. Blue and red segments correspond to the reference and alternative alleles recorded in 1000 Genomes respectively.

5.1.3.5 Population analyses

The Global AIM-SNP set was assessed for ability to infer co-ancestry proportions in 1000 Genomes ASW, MXL, PUR and CLM admixed populations (described in section 5.1.3.1). We compared co-ancestry components detected using 128 Global AIM-SNPs and 350 of 445 LACE SNPs (AME data not available for 95): comprising a large-scale CCAS analysis panel (Galanter et al. 2012) designed to differentiate AFR-EUR-AME admixture components, with E ASN and OCE components not expected to be detectable parts of the above four 1000 Genomes admixed population's ancestry profiles.

We applied standard population analysis approaches of *STRUCTURE* and principal component analysis (PCA). *STRUCTURE* (v.2.3.3) was used with *CLUMPP* and *distruct* software (Porras-Hurtado et al. 2013), applying parameters: burnin = 100,000; MCMC steps = 100,000; *admixture/POPFLAG* model with *correlated allele frequencies* and 5 iterations. PCA used *R* v.2.11.1 and the *SNPassoc* package (Gonzalez et al. 2007). Values of *K* clusters

from 2 to 8 were explored but *K*=5 was identified as the optimum number of populations from comparison of posterior probabilities given by *X*|*K*.

A proportion of populations in any one group can be divergent from those we used for AIM-SNP selection and this may reduce the ability of the marker set to be optimally informative for all regions of a population group's geographic distribution. We assessed within-group variation in SNP allele frequencies by analysing 62 Greenlanders, 32 Fijians and 77 Somalis in 116/128 SNPs with these additional samples exemplifying geographically distinct populations in America, Oceania and Africa. We also typed 22 Pathan Pakistanis from the HGDP-CEPH panel to compare to European allele frequency distributions. In addition, within-group variation of SNPs reported for HGDP-CEPH by Stanford were analysed, although this data set comprises 75/128 marker coverage. SNP data from the 45 extra populations was analysed using Arlequin (v.3.5.1.3) and *STRUCTURE*.

5.2 RESULTS AND DISCUSSION

5.2.1 A SNaPshot of next generation sequencing for forensic SNP analysis (Appendix 3A)

Total sequence coverage in excess of 730,000× was obtained from two Ion 314 v1 chips (Table 34). An average of 500× coverage was obtained per amplicon in each sample excluding the negative control. Assuming an average amplicon size of ~100bp, the amount of data obtained from each run corresponded to ~36 Mb from each 314 v1 chip, nominally rated at 10 Mb, or over three times more sequence data than expected.

5.2.1.1 Amplicon coverage bias

The coverage recorded at the three template input amounts per sample is shown in **Figure 79**. Uneven/inconsistent coverage was observed between different SNP amplicons in all three templates. High and low coverage was repeatedly observed for the same amplicons in all samples and across three template amounts, indicating that the coverage bias was not random. **Figure 80** shows the ordered SNP amplicons from low to high coverage per sample and

template amount indicating a highly skewed distribution across all samples with the majority of amplicons having less than 1000× coverage.

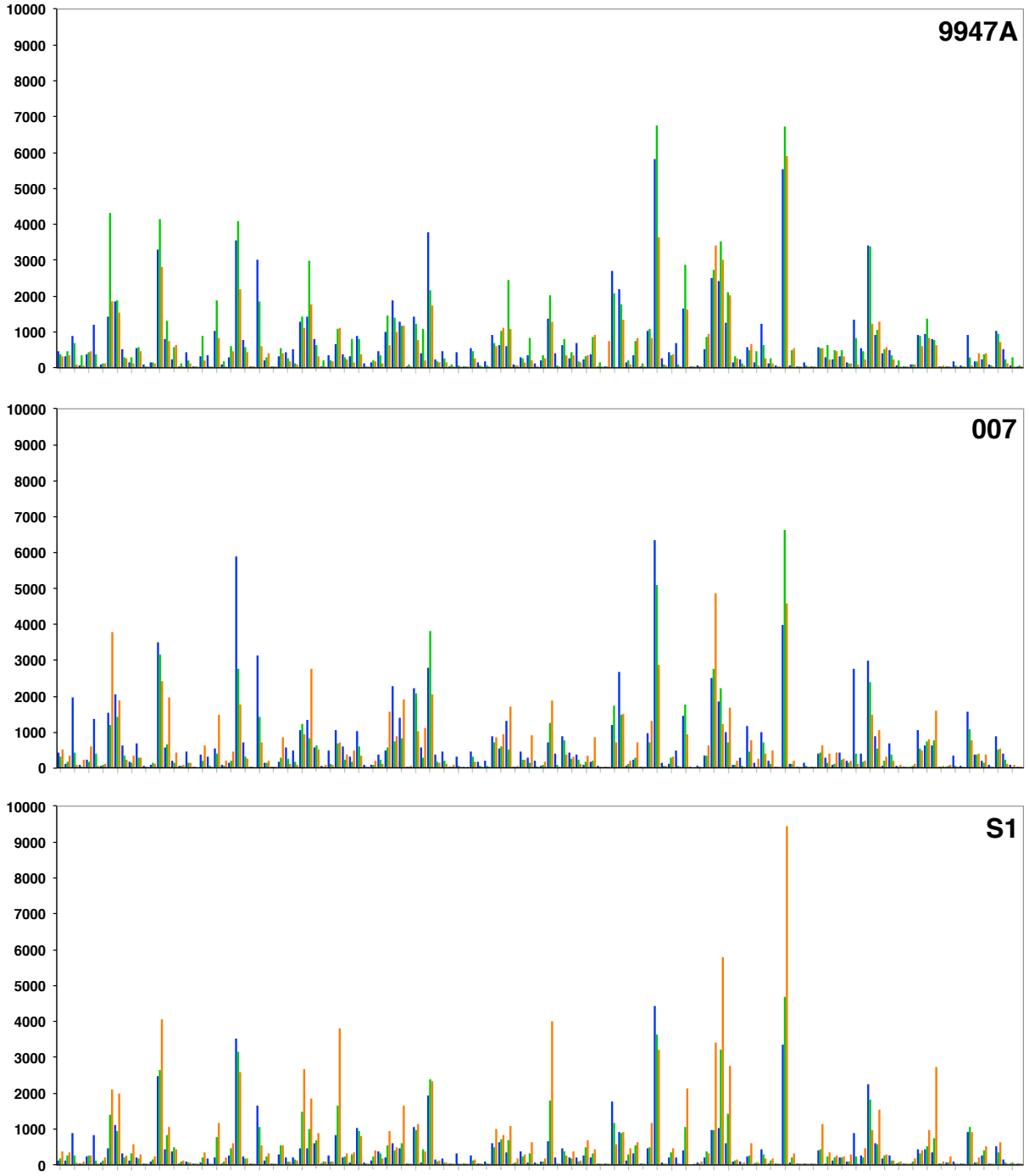


Figure 79: Combined sequence coverage (vertical axis) of SNP amplicons in numerical order of SNP rs-number identifiers (horizontal axis). Template input amounts were 0.1ng (blue), 0.2 (green) and 0.3ng (orange) for each DNA sample of 9947A, 007 and S1.

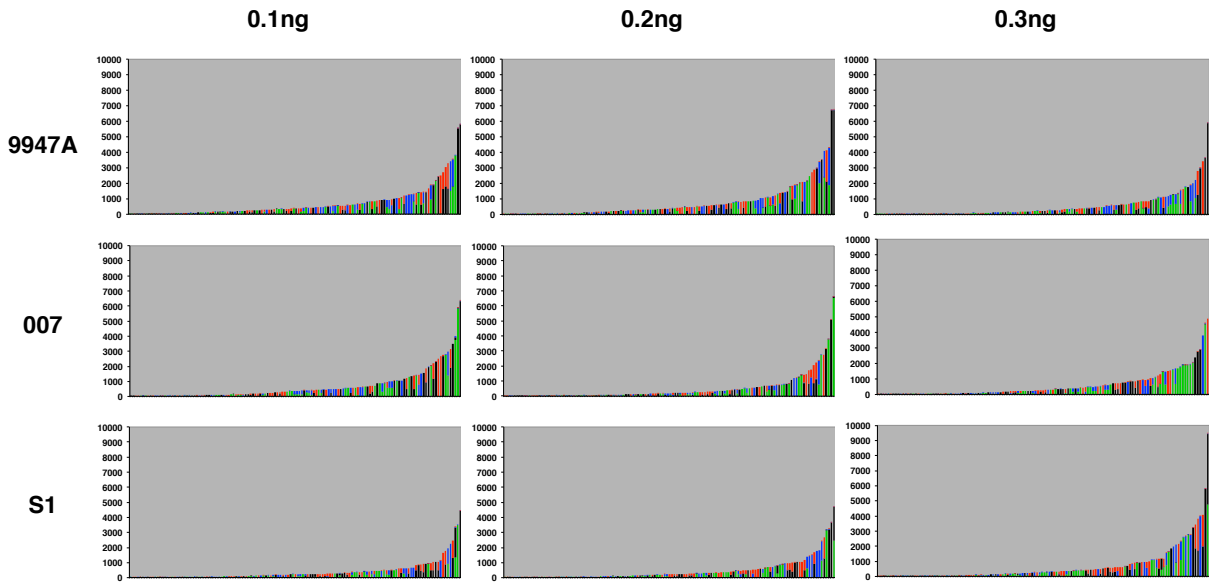


Figure 80: SNP amplicons (horizontal axis) ordered from low (left) to high (right) sequence coverage (vertical axis). The alleles present at each locus are represented by green (A), black (C), blue (G) and red (T).

Of the 10% of amplicons that obtained the highest coverage for all template input amounts, 64% of these also obtained the highest coverage in all three templates. A similar observation occurred with the 10% of amplicons with the lowest coverage in all three template input amounts where 64% of these amplicons also had the lowest coverage in all three templates. This indicates that specific characteristics of the amplicons, or their ratios in the original SNaPshot assays, are likely to influence their sequencing success in relation to depth of coverage obtained. The influences of GC content, amplicon length and PCR assay were systematically examined to determine the origin of the coverage bias observed.

5.2.1.1.1 Effect of GC content on coverage

The GC content of all SNP amplicons ranged from 28 to 65% with a mean of 44% (**Figure 81**). The GC content of the amplicons with the highest and lowest 10% of coverage ranged from 34 to 58% and 27-58%, respectively, with means of 46% and 39%, respectively (Table 36). A Mann-Whitney U test supported the hypothesis that there was no difference in GC content between amplicons with the highest and lowest 10% coverage with 95% confidence (*p*-value=0.19). Therefore, coverage bias of the SNP amplicons did not appear to be associated with GC content.

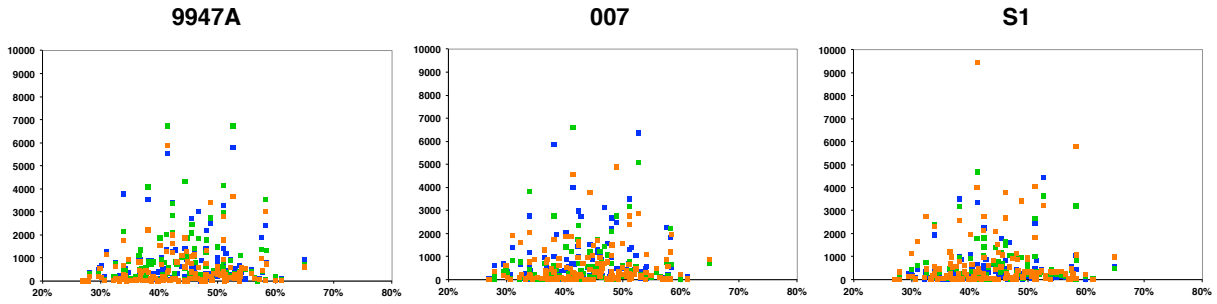


Figure 81: Combined sequence coverage (vertical axis) of SNP amplicons by GC content (horizontal axis). Template input amounts were 0.1 ng (blue), 0.2 ng (green) and 0.3 ng (orange) in each DNA sample of 9947A, 007 and S1.

Table 36: SNP amplicons with the highest and lowest 10% of sequence coverage.

SNP rsID	GC content (%)	Amplicon length (bp)	PCR assay	Coverage 007	Coverage 9947A	Coverage S1
A: Highest 10%						
rs17625895	41	99	Eurasiaplex	5058	6058	5826
rs9522149	53	93	Eurasiaplex	4763	5397	3765
rs734482	38	118	Eurasiaplex	3460	3278	3087
rs2016276	49	90	52plex Auto 1	3376	2879	1778
rs1519654	51	86	Eurasiaplex	3023	3412	3066
rs2196051	34	115	Eurasiaplex	2876	2558	2216
rs1363345	42	99	Eurasiaplex	2272	2678	1674
rs2184030	44	99	Pacifiplex	2171	2536	1321
rs1800407	58	127	Irisplex	1758	2986	3341
Mean	46	103				
B: Lowest 10%						
rs11156577	58	108	Pacifiplex	15	43	22
rs729172	45	60	52plex Auto 1	15	25	13
rs1509524	33	105	Pacifiplex	13	13	11
rs16946159	27	115	Pacifiplex	9	13	7
rs10970986	39	98	Pacifiplex	4	9	6
rs2046361	28	79	52plex Auto 1	4	8	4
rs3785181	47	156	34plex	3	8	8
rs12434466	33	51	Pacifiplex	2	6	1
rs1357617	44	90	52plex Auto 1	1	2	5
Mean	39	96				

5.2.1.1.2 Effect of amplicon length on coverage

The sequence coverage obtained in each sample in relation to the amplicon lengths is shown in **Figure 82**. Amplicons with the highest and lowest 10% of coverage ranged from 86 to 127 bp and 51-156 bp, respectively, with means of 103 bp and 96 bp, respectively (Table 36). A Mann-Whitney U test supported the hypothesis that there was no difference in length between amplicons with the highest and lowest 10% coverage with 95% confidence (*p*-value=0.49). Therefore, coverage bias of the SNP amplicons did not appear to be associated with amplicon length. However, **Figure 82** indicates a sharp reduction in coverage of amplicons less than ~70 bp in length for all three samples and template input amounts. This finding may be attributed to shorter amplicons not being retained during the purification

process using QIAquick columns that have a 100 bp size cut-off, resulting in the loss of smaller amplicons.

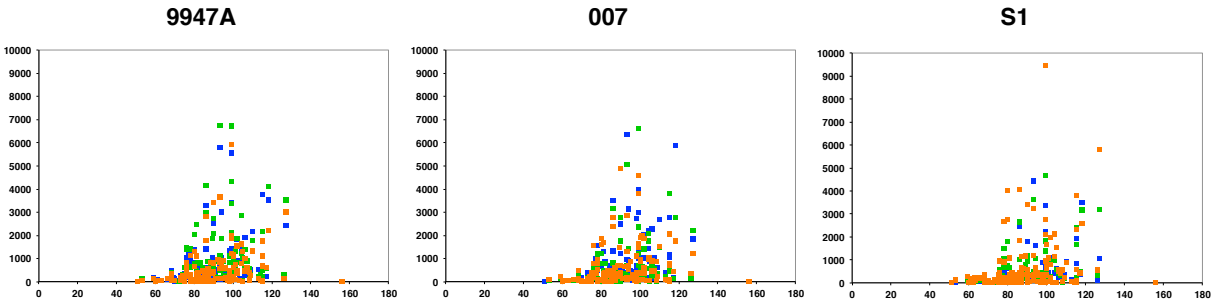


Figure 82: Combined sequence coverage (vertical axis) of SNP amplicons by amplicon length (bp: horizontal axis). Template input amounts were 0.1 ng (blue), 0.2 ng (green) and 0.3 ng (orange) for each DNA sample of 9947A, 007 and S1.

5.2.1.1.3 Effect of PCR assay on coverage

The coverage of amplicons within each of the five multiplex PCR assays also indicated a skewed distribution from low to high coverage (**Figure 83**). *Eurasiaplex* amplicons were overrepresented relative to the other four assays. Table 36 indicates that six of the nine amplicons in the top 10% by coverage (rs17625895; rs9522149, rs734482, rs1519654, rs2196051 and rs1363345) were from *Eurasiaplex* with a range of 1755× to 6741× coverage. Conversely, all but one of the amplicons in the bottom 10% by coverage were from *Pacifiplex* and *52-plex* with 1× to 43× coverage. A Kruskal-Wallis test rejected the hypothesis that there was no difference in combined coverage (across all three template amounts) between amplicons from different assays with 95% confidence (*p*-value=0.000). Hence, there was a significant difference between amplicon coverage between assays, with amplicons from *Eurasiaplex* assays over-represented. SNaPshot electropherograms indicated that higher allele peaks were generally observed from *Eurasiaplex* amplifications compared to the other SNP assays (data not shown).

A Kruskal-Wallis test rejected the hypothesis that there was no difference in combined coverage (across all three template amounts) between amplicons with unique SNPs (present in only one PCR assay) and amplicons with non-unique SNPs (present in more than one assay) with 95% confidence (*p*-value=0.014). Thus, SNPs appearing in more than one assay had significantly greater coverage than unique SNPs.

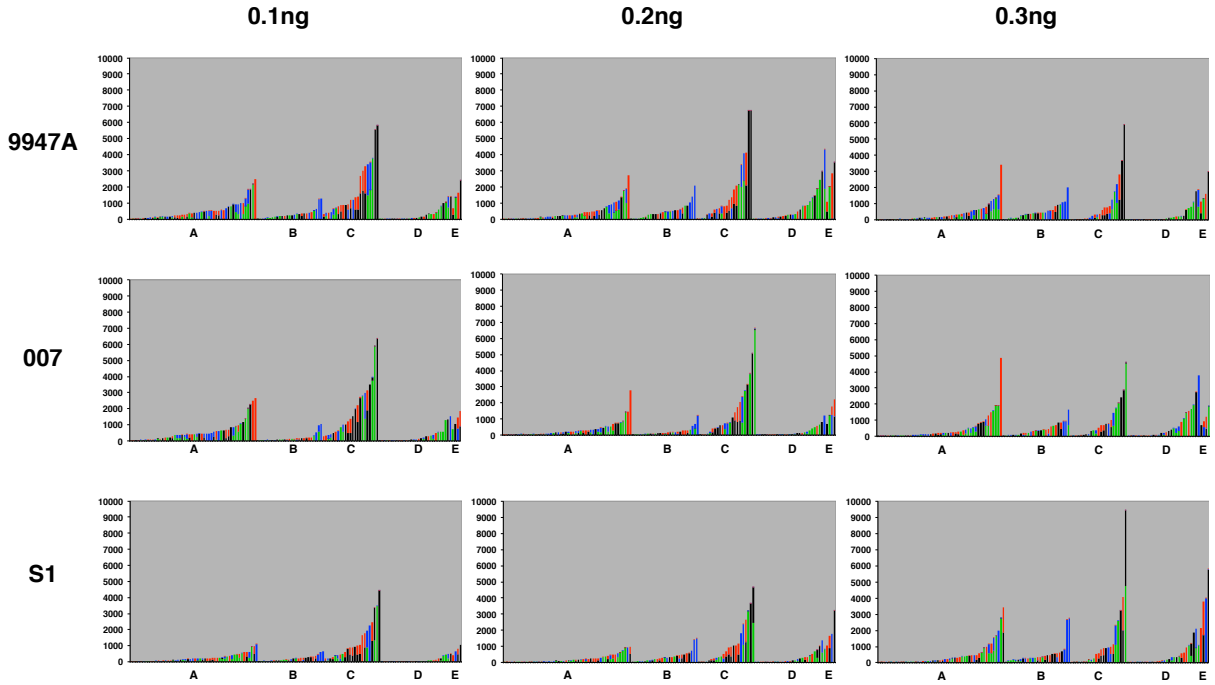


Figure 83: SNP amplicons (horizontal axis) ordered from low to high coverage (vertical axis) within each multiplex PCR assay (A: *52-plex*, B: *34-plex*, C: *Eurasiaplex*, D: *Pacifiplex* and E: *Irisplex*). The alleles present at each locus are represented by green (A), black (C), blue (G) and red (T).

5.2.1.2 Negative controls

The negative control contained eleven amplicons listed in Table 37. Nine of these were *52-plex* amplicons, five had coverage of 6× or less and the remainder had coverage ranging from 23× to 709× With the exception of three amplicons (rs740910 at 2×, rs719366 at 5× and rs722098 at 86×), the genotypes associated with the remainder indicate a possible origin of one or more of the three templates used in this study (9947A, 007 and S1). This may indicate PCR contamination during amplicon preparation. Alternatively, it may represent 'background' or 'noise' and highlights the need to establish a threshold for background levels of sequence output from non-target amplicons.

Table 37: Amplicons detected in the pooled PCR negative control.

SNP ID	Assay	Genotype	Possible template	Coverage (×) A	Coverage (×) C	Coverage (×) G	Coverage (×) T	Total coverage (×)
rs10495407	52plex Auto2	AA	007	195	0	0	1	196
rs12405776	Pacifiplex	CC	9947A, 007	0	1	0	0	1
rs1519654	Eurasiaplex	CC	007	0	701	4	4	709
rs2056277	52plex Auto2	CT	9947A	0	25	0	42	67
rs735155	52plex Auto1	CC	9947A, 007	2	265	0	0	267
rs2107612	52plex Auto1	AG	007	9	0	13	1	23
rs873196	52plex Auto2	CT	S1, 007	0	2	0	4	6
rs1382387	52plex Auto2	AA	S1, 9947A, 007	3	0	0	0	3
rs740910	52plex Auto1	AA	–	2	0	0	0	2
rs719366	52plex Auto1	GG	–	0	0	5	0	5
rs722098	52plex Auto1/34plex	AG	–	61	0	24	1	86

5.2.1.3 Allele read frequencies

Allele read frequency variation (termed 'allele frequency' in the Ion Torrent data analysis software and herein) is shown with respect to the number of alleles in **Figure 84A** and compared to depth of coverage in **Figure 84B**. In each template, a cluster of high numbers of alleles with high coverage is centred around 50% allele frequency corresponding with heterozygous genotypes. Likewise, two clusters around 0% and 100% allele frequencies denote homozygous genotypes. Low numbers of alleles with correspondingly low coverage are present in allele frequency ranges 10-30% and 70-90%. Values in these ranges represent ambiguous genotype designations where it is uncertain whether they are homozygote (with a significant number of spurious alternate alleles) or heterozygote (with pronounced allelic imbalance). As such, a threshold was applied to aid identification, and subsequent discounting, of ambiguous genotypes.

To determine genotype designations, two options were considered based on the allele frequency distributions shown in **Figure 84**. The first strategy was to apply a coverage threshold below which the corresponding amplicons would be excluded. However, this would also result in rejection of a high number of amplicons in the unambiguous frequency ranges of 50% and 100%, since coverage distribution is highly skewed. Setting a minimum threshold of 100× coverage per amplicon, for example, while only rejecting 1 or 2% of coverage, would reject over 20% of amplicons in 9947A, as shown in the cumulative coverage plot for this template in **Figure 85**. This represents a loss of useful genotype data from the sequence analysis.

A second strategy applies an allele frequency threshold to define homozygotes and heterozygotes. For an allele frequency above a 90% threshold, genotypes were designated homozygous for the dominant allele. For the range 10–90%, genotypes were designated

heterozygous for both alleles and SNPs with allele frequency below a 10% threshold were not included (as the alternate allele was designated a homozygote). Application of this threshold regime resulted in high concordance between genotypes obtained for each amplicon at all template input amounts for each sample as well as high concordance between Ion Torrent and SNaPshot genotypes.

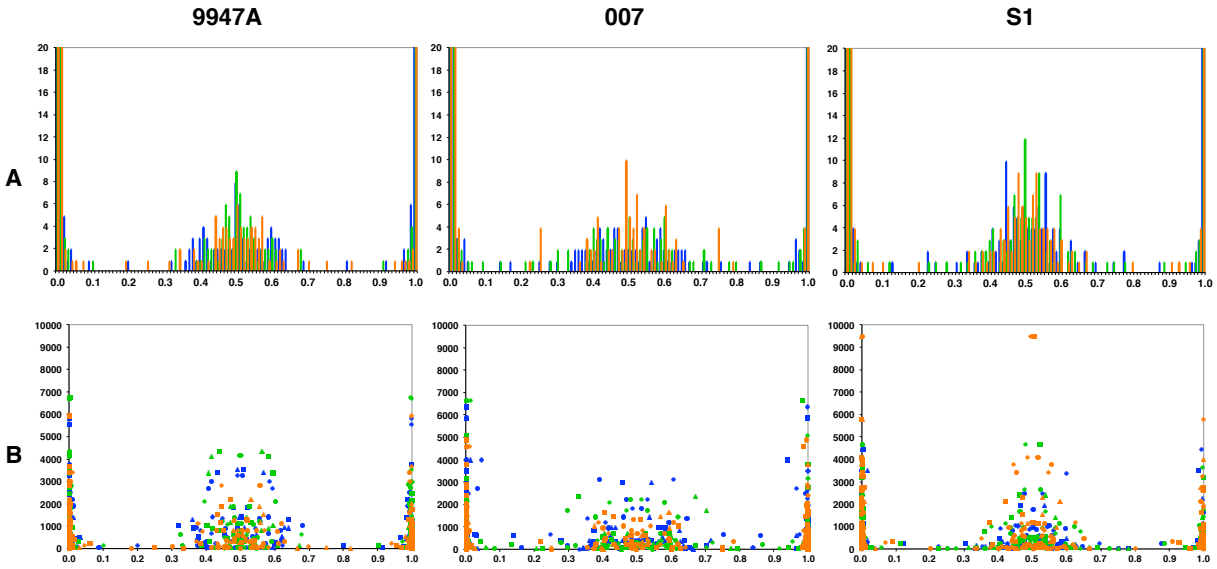


Figure 84: Allele frequency variation (horizontal axis) with respect to number of alleles (A) and sequence coverage (B). Template input amounts were 0.1 ng (blue), 0.2 ng (green) and 0.3 ng (orange) for each DNA sample of 9947A, 007 and S1.

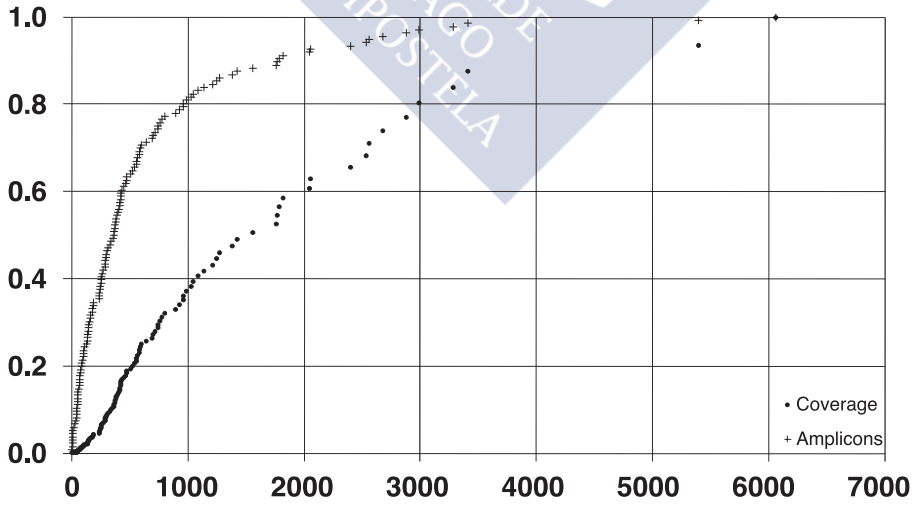


Figure 85: Cumulative fraction (vertical axis) of sequence coverage and amplicons as a function of coverage per amplicon (horizontal axis) for 9947A.

5.2.1.4 Genotype concordance

Table 38 details the discordant genotypes suggested by within-sample comparisons of different template inputs. Of the 136 unique SNPs, all but five provided the same genotype across three template amounts in each of the three samples applying a 90% allele frequency threshold. The five SNPs originated from the *52-plex* or *Pacifiplex* assays. Four SNPs had very low coverage across all three template amounts (<13×), while rs1493232 in 9947A was only discordant with Sanger sequencing at 0.1 ng input. Overall, genotype concordance between the template input amounts was 99%, 98% and 99% for 9947A, 007 and S1, respectively. Amongst 131 Ion Torrent genotypes that were concordant between template input amounts, further discordances between Ion Torrent and SNaPshot genotypes were detected as shown in Table 39. Genotypes for two SNPs (rs2227203 and rs938283 in S1) were not observed in SNaPshot (missing alleles), however, the Ion Torrent genotypes were concordant with Sanger sequencing. For the six genotypes discordant with SNaPshot, Ion Torrent coverage ranged from 90× to 9457× Sanger sequencing confirmed four of the Ion Torrent genotypes and only one of the SNaPshot genotypes. One of the discordant genotypes (rs717302 in S1) produced a different genotype for Ion Torrent, SNaPshot and Sanger sequencing analysis. Overall, of the 408 genotypes from 3 samples and 136 unique SNPs, 395 (97%) were fully concordant with SNaPshot and across three template amounts. Of the discordant genotypes, Ion Torrent data for six SNP/sample combinations (an additional 1.5%) were fully concordant with Sanger sequencing and across three template amounts.

Table 38: Ion Torrent genotypes (and coverage) that were discordant between the DNA template amounts. Missing genotype data is denoted by '-'.

Template	SNPID	0.1 ng	0.2 ng	0.3 ng	SNaPshot genotype	Sanger genotype	Assay
007	rs10970986	CC 1×	– 0×	CT 12×	CT	CT	Pacifiplex
007	rs12434466	AA 7×	– 0×	– 0×	TT	AG/CT	Pacifiplex
9947A	rs1493232	AC 151×	AA 122×	AA 132×	AC	AA	52plex-Auto1
S1	rs2046361	AT 9×	AT 2×	TT 2×	–	AA/TT	52plex auto 1
S1	rs12434466	– 0×	– 0×	AA 3×	AA	AG	Pacifiplex

Table 39: Ion Torrent, SNaPshot and Sanger genotypes and coverage for samples that were discordant between Ion Torrent and SNaPshot. Ion Torrent and SNaPshot genotypes that were discordant with Sanger genotypes are shaded grey. Missing genotype data is denoted by '-'.

Template	SNPID	0.1 ng	0.2 ng	0.3 ng	SNaPshot genotype	Sanger genotype	Assay
007	rs2304925	AA 90×	AA 107×	AA 411×	GT	AA	34plex
9947A	rs1029047	AT 374×	AT 293×	AT 224×	AA	AA/TT	52plex Auto 1
S1	rs12405776	CT 120×	CT 312×	CT 565×	CC	CT	Pacifiplex
S1	rs1544656	CT 1769×	CT 1161×	CT 569×	CC	CT	Eurasiaplex
S1	rs17625895	AC 3342×	AC 4680×	AC 9457×	CC	AC	Eurasiaplex
S1	rs717302	AG 211×	AG 146×	AG 122×	GG	AA/TT	52plex Auto 1
S1	rs2227203	CC 826×	CC 131×	CC 15×	–	CC	Eurasiaplex
S1	rs938283	TT 200×	TT 195×	TT 240×	–	AA/TT	52plex Auto 2

The two Ion Torrent genotypes discordant with Sanger sequencing were rs1029047 in 9947A and rs717302 in S1. Both Ion Torrent genotypes were heterozygotes with balanced allele frequencies (no more than 62% of either allele) where Sanger sequencing predicted homozygotes. In addition, the coverage of each SNP was more than 100× for each template input amount. Therefore, the Ion Torrent genotypes were not ambiguous with respect to coverage or allele frequencies. The rs717302 S1 genotype was discordant across all three genotyping platforms, as highlighted above, leaving the most likely genotype unknown. The rs1029047 9947A AT genotype is assumed to be incorrect, as both SNaPshot and Sanger sequencing gave a consistent AA genotype. Seo et al. (2013) also found the same heterozygous Ion Torrent genotype for this SNP to be inconsistent with that determined by the IlluminaGAIIx NGS system.

5.2.1.5 Sensitivity

As 97% of Ion Torrent genotypes were concordant with SNaPshot and a further 1.5% were concordant with Sanger sequencing, it can be assumed that the Ion Torrent PGM system can reliably genotype PCR amplicons from multiplex assays with at least 0.1 ng of DNA template. Because the Ion Torrent sequencing run included amplicons pooled from five multiplex assays, a total of 0.5 ng appeared to be sufficient to accurately genotype 97% of the

136 unique SNPs. This is comparable to recommended input amounts for established STR assays, for example, 0.5ng for PowerPlex® 21[19] and AB GlobalFiler™ [20].

Four of the five amplicons for which Ion Torrent genotypes were discordant across the three template amounts had very low coverage (<13×). This suggests that accurate SNP genotyping can be achieved at amplicon coverage as low as 20× in agreement with the studies of Bentley et al. (2008) and Quail et al. (2012) who were able to accurately call heterozygous SNPs with 15× coverage.

5.2.1.6 Discussion

Based on this preliminary study, the Ion Torrent system has demonstrated considerable potential for use in forensic DNA analysis as a low to medium throughput NGS platform. Using Ion 314 v1 chips, three times more data was obtained than expected and without the need to optimise or further adapt PCR multiplexes from previously established SNaPshot assays to produce the amplicons for sequencing. The average coverage of 500× per SNP amplicon was sufficient to generate a genotyping accuracy of 97% for the SNPs assessed in this study.

A skewed distribution of coverage of the SNP amplicons generated from the combined PCRs indicated a noticeable sequencing coverage bias. Patterns of high and low coverage were repeated for the same amplicons over the three samples and template amounts, indicating non-random bias. Aside from the expected bias between individual multiplex PCR assays (with *Eurasiaplex* amplicons over-represented), the source of bias within assays was not determined. However, the data suggests that it may arise from varying amplification efficiencies of the SNPs in these assays prior to library preparation and/or emulsion PCR after library preparation. The SNaPshot PCR protocols used to generate SNP amplicons were used without further adjustment or attempt to balance output between assays prior to pooling. In addition, equal volumes of PCR amplicons were pooled prior to library preparation. Adjusting the number of PCR cycles for individual assays and careful equimolar pooling of PCR

19 Promega, Technical manual: PowerPlex® 21 System, Promega Corporation, Madison, WI, 2012

20 Applied Biosystems, GlobalFiler™ Express PCR Amplification Kit User Guide, Life Technologies Corporation, Carlsbad, CA, 2012

products may assist in reducing bias between PCR assays while bias within assays may be addressed by adjusting primer concentrations.

GC bias has often been associated with coverage bias in next generation sequencing and library preparation conditions can also result in sequencing bias (Ross et al. 2013). While Quail et al. (2012) observed that Ion Torrent PGM sequences had strong coverage bias against extremely GC-poor amplicons and Ross et al. (2013) observed Ion Torrent bias for both GC-rich and GC-poor amplicons, we observed no such bias. This may be due to higher GC in our amplicons (28-65%) compared to the genomic DNA in their study (19% GC in *Plasmodium falciparum*). Furthermore, the amplicons sequenced in this study were derived from optimised PCR assays where optimal GC content of both primer and amplicon sequence had been an important consideration during marker selection and PCR primer design.

Semiconductor technology is employed in Ion Torrent sequencing, therefore, homopolymeric stretches increase sequencing errors, in common with the luciferase-mediated signals employed in 454 sequencing (Shendure and Ji 2008). This is may be related to GC bias, as G and C homopolymers are more likely in GC-rich sequence (and A and T homopolymers more likely in GC-poor sequence). While homopolymeric stretches are avoided in SNaPshot primer design, there are some homopolymeric stretches between primer sequence and target SNP in some of the PCR amplicons of this study. SNaPshot genotyping is largely unaffected by homopolymers in this region, as shown by rs1029047 in *52-plex*, the only amplicon discordant between Ion Torrent and SNaPshot but concordant between SNaPshot and Sanger sequencing. The rs1029047 amplicon has TTT and AAAAAAAAA sequences immediately flanking each side of the SNP that can produce sequencing errors for the Ion Torrent. This same SNP was mistyped with the same erroneous genotype (AT) by Seo et al. (2013), using Ion 314 Chips and the Ion AmpliSeq™ HID SNP panel. In common with our concordant SNaPshot and Sanger sequencing genotypes for rs1029047 in 9947A, we note that Seo et al. also obtained the AA genotype when using the Illumina GAIIx NGS system. Homopolymeric sequences not immediately flanking the SNP do not seem to affect Ion Torrent sequencing. The rs89170 amplicon in the *52-plex* has an AAAAAAAAAA stretch within 8 bp of the SNP and a further two amplicons (rs2065982 in *34-plex* and rs2196051 in *Eurasiaplex*) have a TTTTTTTT stretch within the amplicon.

As with GC content, no obvious influence of amplicon length on coverage for amplicons between 70 and 110 bp in length was observed (**Figure 82**). The purification columns used

during library preparation may have produced a partial loss of amplicons below 70 bp although this did not result in lower genotyping accuracy. However, this highlighted the importance of selecting purification methods appropriate for the range of PCR primers lengths and amplicon sizes analysed.

Coverage of the nine SNPs represented in more than one assay was not amongst the top 10% of the highest coverage. Therefore, unequal representation of SNPs in multiple assays resulted in higher coverage of those SNPs but this was not associated with the overall coverage bias observed. Amplicon coverage bias is most likely to result from imbalance within and between each of the five PCR assays. The SNaPshot PCR protocols used to generate SNP amplicons were used without further adjustment or attempt to balance output between assays prior to pooling. PCR products were combined based on equal volumes rather than equimolar amounts resulting in a degree of coverage bias, with *Eurasiaplex* noticeably over-represented, readily corrected by reducing this assay's PCR cycling appropriately.

Although not true replicates, consistent genotypes amongst three template input amounts indicate high reproducibility from the Ion Torrent. In addition, 97% of 136 unique SNPs were concordant with SNaPshot for all input amounts (with a further 1.5% concordant with Sanger sequencing), indicating that Ion Torrent sensitivity can be as low as 0.1 ng per PCR assay or 0.5 ng in total, comparable to current STR analysis requirements. The accuracy of some SNaPshot genotypes has been called into question by comparison with Ion Torrent genotypes in this study (Table 5). Of the six genotypes discordant between SNaPshot and Ion Torrent, four were concordant between Ion Torrent and Sanger sequencing. The remaining two SNPs indicated homopolymer stretches or discordancy between all three genotyping platforms. Four of the five discordant SNaPshot genotypes were homozygotes with heterozygous Ion Torrent and/or Sanger genotypes, suggesting low-level allelic drop out could be an inherent characteristic of SNaPshot electropherograms. However, the recommended range of template input amounts for the five SNaPshot assays used is between 0.5 and 1ng. Therefore, the SNaPshot genotyping accuracies observed in this study were based on less than the recommended template input amount for each assay.

The advantage of NGS for SNP genotyping compared to multiple SNaPshot assays is that amplicons are pooled in the library preparation step and are therefore processed as a single reaction sample. This reduces the potential for cross contamination and post-PCR contamination that is more likely to occur when multiple reactions are run in parallel to

generate SNP genotypes. In addition, nucleotide barcoding allows individual samples to be pooled together and analysed on a single chip, reducing inter-run variability while reducing the time, effort and costs associated with sequencing.

In addition to sensitivity and accuracy, a complete forensic validation of SNP genotyping with NGS would require assessment of additional factors. Although arbitrary allele frequency thresholds of 90% for homozygotes and 10-90% for heterozygotes applied in this study resulted in high genotype accuracy, higher stringency thresholds, such as 95% (homozygotes) and 40-60% (heterozygotes), will increase accuracy but reduce the number of useable genotypes. This would have the added benefit of increasing the ability to detect mixed source genotypes from their unbalanced allele frequencies. Establishing allele frequency thresholds was the preferred approach to determine SNP genotypes in this study rather than applying coverage thresholds that lead to significant loss of data.

However, some level of coverage threshold is necessary to accommodate background sequence coverage in negative controls. Five of the SNP amplicons detected in the pooled PCR negative control had less than 10× coverage, but six had higher coverage of 23× to 709×. These latter values could be considered above an acceptable level of background sequence output. Although some negative control genotypes could be attributed to PCR contamination, others with relatively low coverage (2×, 5× and 86×) did not match genotypes in the samples used. Detecting low-level sequence did not impact genotyping accuracy, however, appropriate thresholds will need to be established in the long term. Although the negative control was a pool of amplicons analysed in the same manner as the DNA samples, individual amplification reaction negative controls could be more informative when sequenced separately. Therefore, our approach, while pragmatic, is not the most appropriate for evaluating baseline thresholds in detail. However, this preliminary study could be adapted to improve future studies of negative control baselines to assess low level base miscalls.

The resolution of mixtures will present certain interpretative challenges when NGS is applied to forensic analysis. As most SNPs are biallelic, SNaPshot suffers from severe limitations when attempting to resolve mixtures. In addition, the SNaPshot fluorochromes do not generate equal signal strength with 4:2:1:1 average peak height ratios for the G:A:C:T nucleotides, further increasing the difficulty in comparing peak heights. Mixture resolution of biallelic markers in NGS analysis will not be without challenges if the coverage biases observed in this study indicate the levels of variation in sequence output users can expect

from the Ion Torrent system. However, by sequencing every base in an amplicon, NGS has the ability to detect mutations up- and down-stream of SNPs that may provide some capacity to distinguish between contributors to a mixture.

5.2.1.7 Conclusions

The Ion Torrent has successfully genotyped at least 97% of 136 unique SNPs for three samples at three input amounts down to 0.5 ng of template DNA. The only SNP confirmed to be mistyped by the Ion Torrent was rs1029047 with known homopolymeric stretches immediately flanking either side of the SNP. Notwithstanding this, the Ion Torrent and other NGS platforms have the potential to genotype hundreds of SNPs in large numbers of samples analysed in parallel at a high level of genotyping accuracy. The majority of the costs associated with Ion Torrent genotyping in particular, and NGS in general, are for library preparation rather than sequencing.

Although SNaPshot has served the forensic community well as a method to genotype up to 30–40 SNPs per assay from picogram quantities of DNA, the ability to provide accurate BGA and EVC intelligence to forensic investigators requires the analysis of more SNPs than can be genotyped from a single SNaPshot assay. A single NGS assay containing hundreds of SNPs can maximise use of scant evidential material as well as reducing both sample processing times and potential contamination risks. However, the recently published assessment of the Ion AmpliSeq HID SNP panel by Seo et al. (2013) indicated that 10 ng of template DNA was currently required to detect all loci reliably, whereas using 1 ng of DNA resulted in one of the 136 SNPs showing extreme heterozygote imbalance. Using five established SNaPshot assays in five PCR multiplexes, reproducible genotypes were obtained from 0.1 to 0.3 ng of DNA for each, equating to 0.5-1.5 ng of DNA, making it a viable alternative strategy. This approach enables users with significant investment in SNaPshot assays to make immediate use of NGS technology for SNP genotyping and allows flexibility with respect to the selection of specific assays to cater to the unique needs of an investigation.

This study represents the very low end of the scale with respect to NGS throughput. Sequencing capacity increases with the Ion 314^{TM} v2 Chip (up to 100 Mb), the Ion 316 v2 Chip (1 Gb) and Ion 318 v2 Chip (2 Gb) on the Ion Torrent and then to 10 Gb on the Ion Proton (LT), 15 Gb on the MiSeq (Illumina) and 900 Gb on the HiSeq (Illumina). Hence,

there is opportunity to scale up 1000-fold from the levels of this study, in the number of samples, the number of SNPs, or both. NGS systems can utilise combinations of current forensic assays and accommodate new phenotypically predictive SNPs that are continually being discovered.

5.2.2 Inter-laboratory evaluation of SNP-based forensic identification by massively parallel sequencing using Ion PGM™ (Appendix 3B)

Before presenting the results obtained in this study, it is important to mention that, although this study is not based in ancestry inference markers, it has allowed us to start to understand the viability of the Ion PGM™ in forensic analysis and to define which are the parameters and thresholds of interest.

5.2.2.1 Sequence coverage from Ion PGM™

Sequencing depth (depth of coverage or simply 'coverage') has a direct bearing on the sensitivity and genotyping accuracy of NGS systems applied to forensic SNP typing. Its value specifies the number of times each base has been read in the sequencing run. For whole genome applications it is usually stated as an average value per base. However, for SNP detection applications such as HID SNP, actual depth of coverage at the targeted SNP site is more relevant and is given in number of reads targeting the site (herein SNP Target Reads). This final number will depend on sequencing technology, raw read filtering methods and how variant calls are processed. In Ion PGM™ sequencing runs, the number of wells per chip that can be filled with ISPs defines the number of possible reads. Sample pooling, template preparation (influencing the number of non-templated and polyclonal ISPs) and loading efficiency (influencing the number of empty wells) determine the final number of successfully read ISPs (monoclonal reads). During the base calling steps of TS data processing, monoclonal reads are further filtered for low quality and adapter dimer reads. When sequencing multiple barcoded samples, equimolar pooling ahead of template preparation aims for a homogenous distribution of reads between samples of the same run.

In this study, all 12 runs reached overall sequencing throughput, measured in Mb per run, in compliance with TS guidelines for each chip version used (**Figure 86**). It is noticeable that

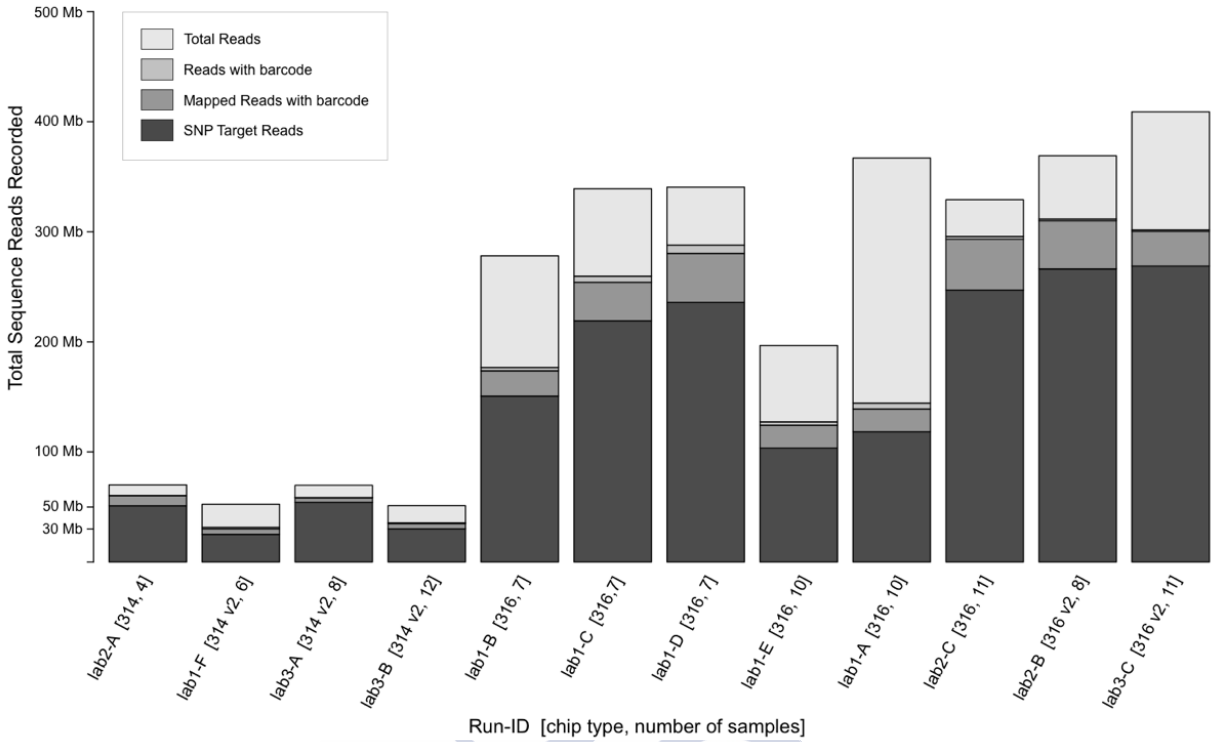


Figure 86: Proportions of four types of sequence reads from 12 Ion PGM™ runs using the full range of available sequencing chips.

for runs pooling low-level and optimum input DNA samples (lab1), more reads are filtered during the base calling process. A more comprehensive description of primer sequence and primer dimer issues in low-level DNA samples as well as sequencing results of negative controls is summarised in Appendix 3C. While the amount of filtered low quality reads per run is similar for all runs, the percentage of filtered primer dimer reads is slightly higher (*p*-value=0.029, alpha=0.05) in lab1 runs with low-level DNA and optimum input DNA samples combined on the same chip. This is indicated by the SNP Target Read distributions for all 101 analyses in **Figure 87A**. The distribution of quartiles reveals variation both within and between runs, but **Figure 87C** indicates that runs combining low-level DNA alongside optimum input DNA samples have higher variation between samples. **Figure 87B** shows the deviation from maximum achievable SNP Target Reads (see figure legend for this metric's definition). In comparison to low-level DNA samples the analysis of optimum input DNA samples (68 high quantity/quality DNAs of 1-10 ng) gave less deviation from expected SNP Target Reads. Furthermore, Ion PGM™ coverage analysis shows significantly higher off-target reads (*p*-value=0.00045, alpha=0.05) in low-level DNA samples. We detected an increased number of sequenced multiplex primers from target amplification in low-level DNA

samples. These primer sequences are aligned to the reference genome and account for the total number of monoclonal reads in TS, but are not considered part of the amplicon, thus increasing the amount of off-target reads (Appendix 3C).

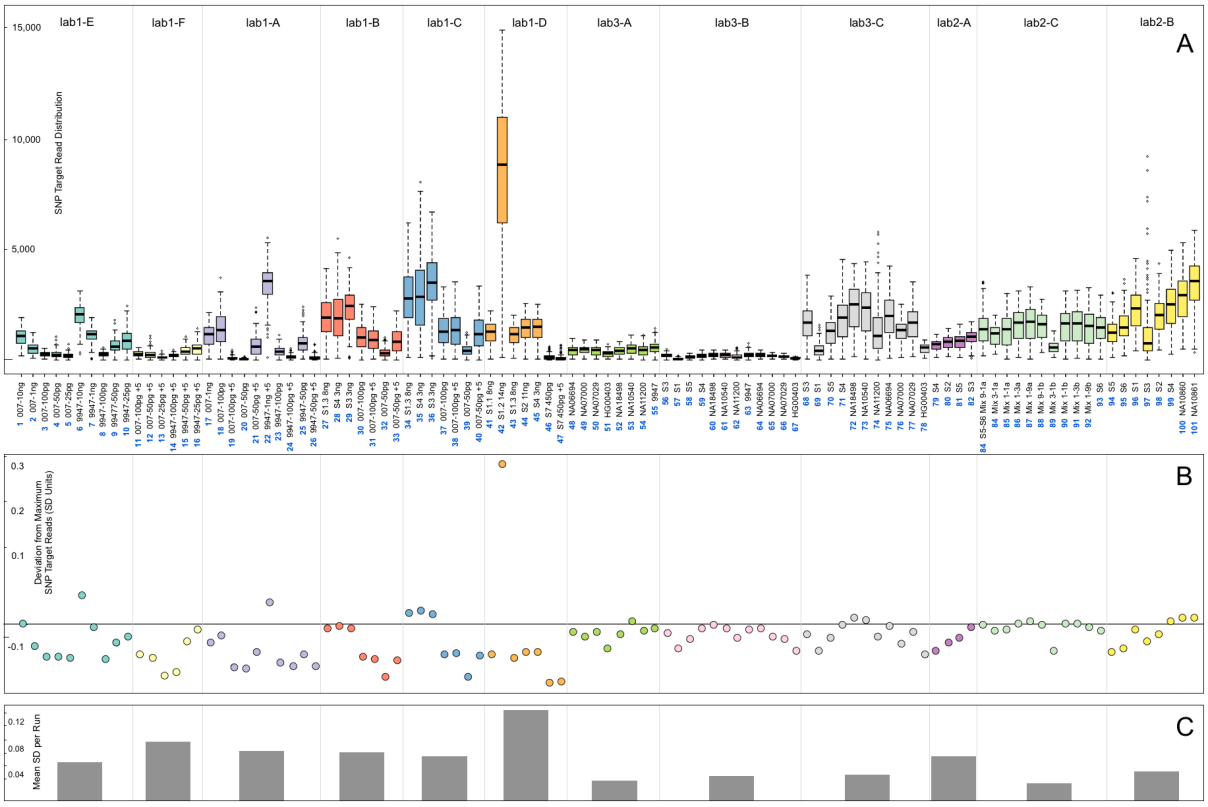


Figure 87: (A) Box plots of recorded SNP Target Reads from 101 samples in 12 runs (interquartile range: IQR boxes, IQR 1.5 whiskers, means shown as mid-plot bars). Blue numbers are used to identify samples listed in **Figure 88**. (B) Deviation from expected maximum SNP Target Reads adjusted for a wide range of chip types, sample numbers per chip and DNA quality amongst the runs made. The deviation metric is calculated as: [(SNP Target Reads – Expected SNP Target Reads)/Sum of Clonal Reads per Chip], where: Clonal Reads = number of reads passing the polyclonal filter; Expected SNP Target Reads = Clonal Reads/number of samples. (C) Summary bar plots of mean standard deviation of SNP Target Reads to expected SNP Target Reads per run. Generally, runs combining optimum input and low-level DNA samples show higher variation from expected SNP Target Reads than runs with optimum input DNA only.

There are two main considerations for multiplex SNP typing in massively parallel sequencing analyses: minimum coverage thresholds for reliable genotyping and number of samples that can be sequenced in parallel to meet those thresholds. LT guidelines suggest minimum coverage thresholds for germline and somatic SNP detection of 30 and 500, respectively. The threshold for somatic SNP calling is close to the values cited in whole genome and enrichment variant detection studies (Bentley et al. 2008, Koboldt et al. 2010, Nielsen et al. 2011, Elsharawy et al. 2012, Quail et al. 2012, Sims et al. 2014). However, minimum coverage thresholds generally depend on the sequencing application, the SNP

variant-calling algorithms used and analysis parameter settings. For forensic applications, a threshold of ~20× could be sufficient coverage to reliably detect variants in high quality single source DNA samples, whereas mixture detection and low-level DNA samples will require much higher coverage. In this study, the lowest coverage values with concordant genotypes in autosomal and Y-chromosome SNPs (herein A- and Y-SNPs) were 13× and 41×, respectively, discounting outlier SNPs. This largely matches results of a recent study by Daniel et al. finding a similar minimum coverage estimate of 20× for reliable SNP genotyping (Daniel et al. 2015). In mixtures, however, minimum coverage should be set higher to reliably identify minor alleles in heterozygous markers. For A-SNPs, concordance between the expected genotypes in the mixture and those of the components was obtained with an average 269× coverage or higher. Y-SNPs gave concordant genotypes with an average of 63× coverage in the 1:9 male-female mixture whereas this value increased to 274× in the 9:1 male-female mixture.

To gauge samples loaded per run, LT provides guidelines for pooling samples to reach the estimated minimum coverage for 95% of bases. In this study samples were pooled in a run to aim for a minimum coverage between 42× and 286× for 95% of bases (Table 40). Information on minimum coverage per sample for 95% of bases is not included in TS output. In the HID SNP panel, the targeted 95% base minimum coverage thresholds were only reached for all SNPs in 8 samples (all optimum input DNA). When accounting for outlier SNPs, 31 optimum input DNA samples reach the desired minimum coverage threshold. From the general coverage assessments made, we infer that a targeted minimum coverage of at least 62× for 95% of bases is necessary to accomplish a minimum coverage of 13× for all SNPs in the panel, which is in agreement with minimum coverage threshold values for concordance samples. For this reason, Run lab3-B was omitted from further concordance studies since none of the optimum input DNA samples reached this threshold. The heatmaps in **Figure 88** outline differences between analyses by ranking cells with increasing coverage per analysis (top to bottom, topmost analyses comprising mainly low-level DNA), and per SNP (left to right). Although a similar SNP coverage pattern across samples is discernible, the leftmost columns show more heterogeneity than average. In fact, further analysis shows that per sample coverage distribution of all SNPs in the panel is not uniform across samples (Figure S5 in Appendix 3C). In conclusion, LT guidelines are useful for initial estimation of sample numbers per chip and minimum coverage. However, the guidelines do not function well when

estimating minimum coverage for all HID SNPs in the panel, as well as when considering low-level DNA samples. For this reason, it is important to adjust numbers of samples loaded on each chip to a particular SNP set and to carefully gauge the quality and quantity of DNA samples to be sequenced.

Table 40: Expected sequence throughput of Ion PGM™ based on chosen sample numbers and 3-series chip type used.

RunID	Chip Type	Number of samples	Minimum 95% Coverage Set
lab1_F	314	6	83
lab2_A	314	4	125
lab3_A	314v2	8	62.5
lab3_B	314v2	12	42
lab1_E	316	10	200
lab1_A	316	10	200
lab1_B	316	7	286
lab1_C	316	7	286
lab1_D	316	7	286
lab2_B	316	11	181.82
lab3_C	316v2	11	181.82
lab2_B	316v2	8	250

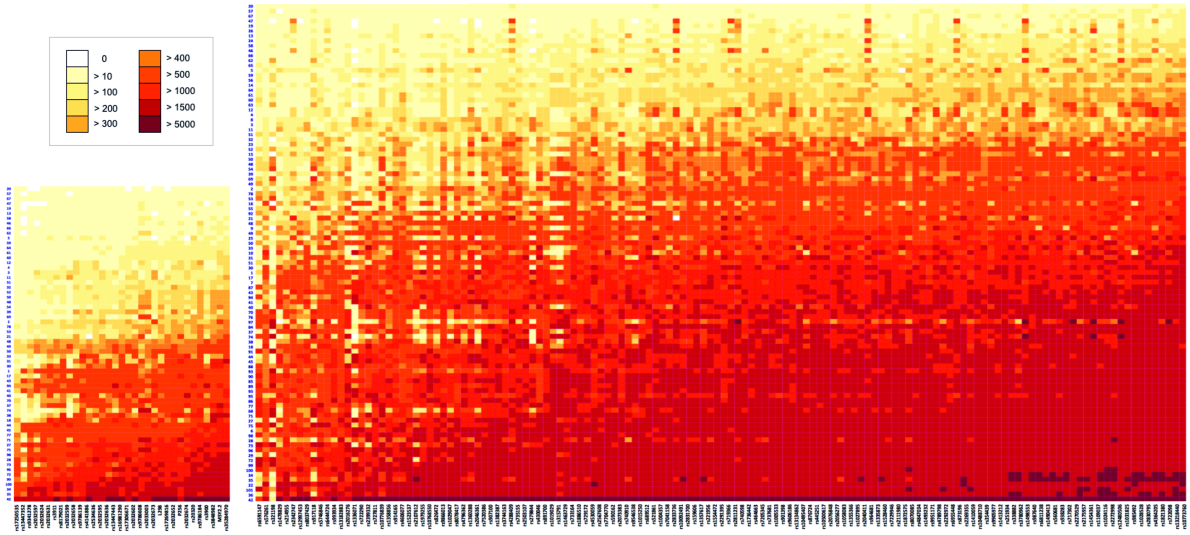


Figure 88: Analysis vs. SNP heatmap arranged as: increasing mean simple coverage levels top to bottom, increasing mean SNP coverage levels left to right. Left map shows Y-SNPs and for brevity, blue run identifiers are as detailed in **Figure 87A**.

5.2.2.2 Sequencing characteristics Ion PGM™ that impact forensic SNP genotyping

Considering the sequence data in Genotyper output or obtained from this study's comparisons amongst runs and laboratories, we focused on sequence coverage, base misincorporation, allele read frequency balance and strand bias, as factors impacting the reliable differentiation of SNP heterozygotes from homozygotes. While artificial mixtures can help assess how mixed DNA changes standard Ion PGM™ sequence data and creates atypical patterns, it is important to assess the range of values observed in HID SNP sequences with unmixed DNA. From the value ranges recorded, outlier SNPs were identified which either should be removed from the HID SNP set or excluded from the data analysis applied to more complex forensic analyses, including genotyping low-level and extremely degraded DNA or detecting mixtures.

5.2.2.2.1 Base misincorporation rates

To gauge the overall rate of base misincorporation of Ion PGM™ (incorrect bases detected at the SNP site in small proportions of sequence reads), the incidence of non-specific 3rd/4th base incorporation (e.g. G and T in an A/C SNP) was compared to incidences of incorrect alleles in homozygotes (e.g. very low occurrence of A bases in C homozygotes). If such rates are comparable then a simple baseline rate of misincorporation can be established. If different, then levels of extraneous target DNA detected by Ion PGM™, akin to allele drop-in, can be estimated by how much more allelic misincorporation is seen. In either case, any outlying SNPs with above-average misincorporation can be identified and appropriate safeguards applied when detecting mixed DNA with minor components below 10%. **Figure 89** records frequencies of misincorporated bases in ranked order and shows allelic and non-specific misincorporation rates were similar in nearly all SNPs and hardly rose above 0.2% in all but 12/169 SNPs. Amongst the twelve SNPs with higher misincorporation (on the right-hand side), only rs8078417, rs2399332, Y-rs2032597, rs9866013 and rs1523537 reach 1% or more. These SNP's data should be discounted from assessments of imbalanced homozygote patterns in mixtures, particularly rs2399332 and rs1523537 with disproportionately high allelic misincorporation.

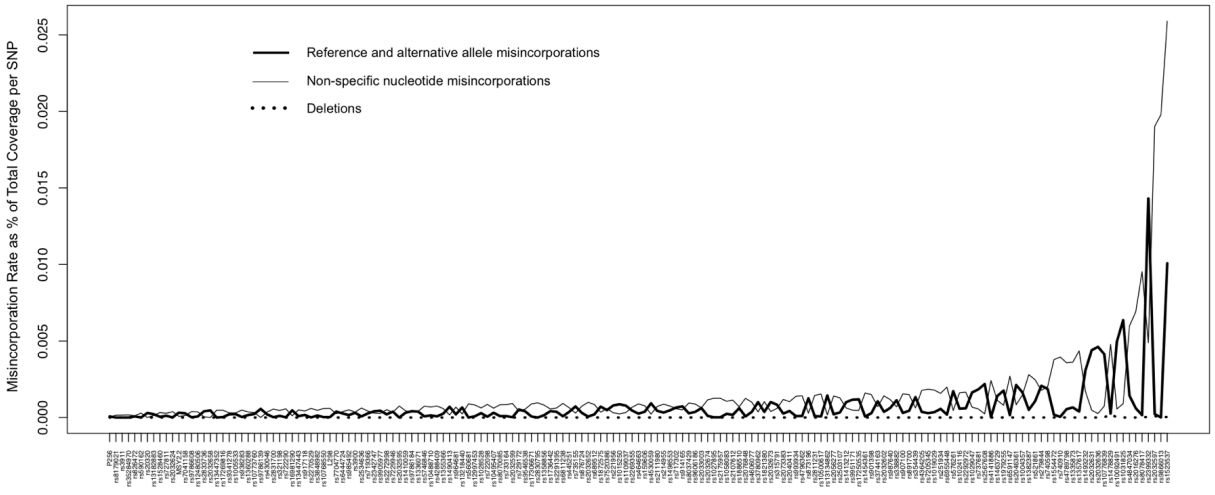


Figure 89: Base misincorporation rates recorded as the presence of non-allelic reference or alternative bases (e.g. low levels of A in G homozygotes plus G in A homozygotes); non-specific base incorporation (e.g. C or T in an A/G SNP) and deletions.

Although allelic and non-specific misincorporation rates are similar enough to largely discount drop-in, Y-chromosome sequences were observed in female DNA. **Figure 90** shows 34 Y-SNP nucleotide reads made in six analyses of two female samples. This data represents male SNP target sequence in processed Genotyper output, but only 34 sequences amongst more than 2 million female-specific sequences indicates extremely low levels of drop-in genotypes from extraneous DNA for the Ion PGM™ system.

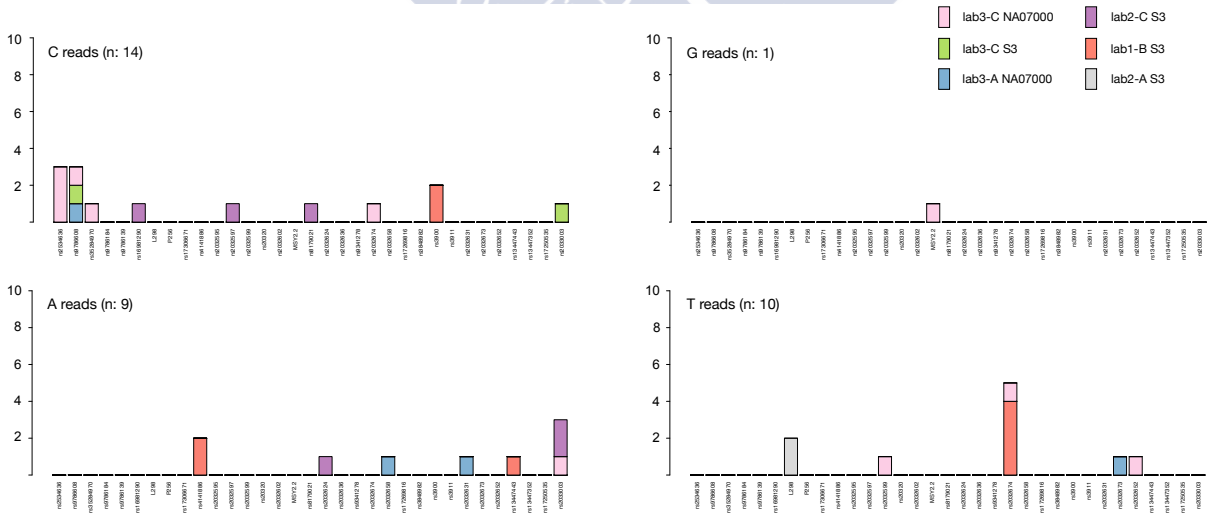


Figure 90: Y-SNP nucleotide reads recorded in analyses of female DNA samples. Numbers of reads indicate very low levels of extraneous male sequences amongst much higher quantities of autosomal SNP target sequence obtained (34 sequences in 6 samples).

5.2.2.2.2 Allele read frequency balance

All forensic genotyping approaches must reliably differentiate imbalanced heterozygote signals, created by stochastic effects in PCR, from the combined allele signals of mixed DNA. This is particularly important for the 136 binary A-SNPs of the HID SNP set, as mixtures can only be detected by measuring the signal of one allele against its alternative. Furthermore, the Y-SNPs, chosen to help infer population divergent male phylogenies, are much more restricted in detecting multiple genotypes (i.e. males from the same population are minimally differentiated). We defined ARF settings that could equate to signal ratios commonly observed in forensic markers and then assessed their effect on genotype calls. Allele reads were reviewed from 38 analyses, comprising 169 SNPs in 28 male DNAs, 136 A-SNPs in 10 female DNAs. **Figure 91A** shows the distributions obtained from the ratio of reference and total ARF values. A-SNP heterozygotes mostly showed good levels of clustering around the 0.5 'perfect balance' midline. Homozygote data at the top and bottom is even more regular in distribution, indicating ratios do not cross 0.1/0.9 thresholds.

Applying an 'aggressive' 45% allele balance thresholds, (i.e. a maximum 55:45 heterozygote ratio) was assessed, but marked too many SNPs as imbalanced when in all other respects their genotypes were concordant and reliably detected (see Sections 5.2.2.3 and 5.2.2.4). A 40% threshold (60:40 heterozygote ratio), indicated by the middle grey box over A-SNPs in **Figure 91A**, gave better equilibrium between gaining the highest proportion of reliable genotypes and balanced signals in optimum input DNA samples. Several SNPs with atypical ARF distributions are evident in **Figure 91A** and were identified from divergent average heterozygote ARF values. SNPs rs2399332, rs1029047, rs8037428, rs430046 and rs1523537 were identified as poorly balanced ARF markers, in common with the analysis of HID SNP performance by Børsting et al. (2014). Additionally, rs2107612 was poorly balanced in our study, but not singled out by Børsting et al. Interestingly, SNPs rs10776839, rs4530059 and rs1031825 found to be problematic by Børsting et al., gave reasonably balanced ARFs here, although **Figure 91A** indicates rs4530059 and rs1031825 have small proportions of genotypes lying outside the threshold range.

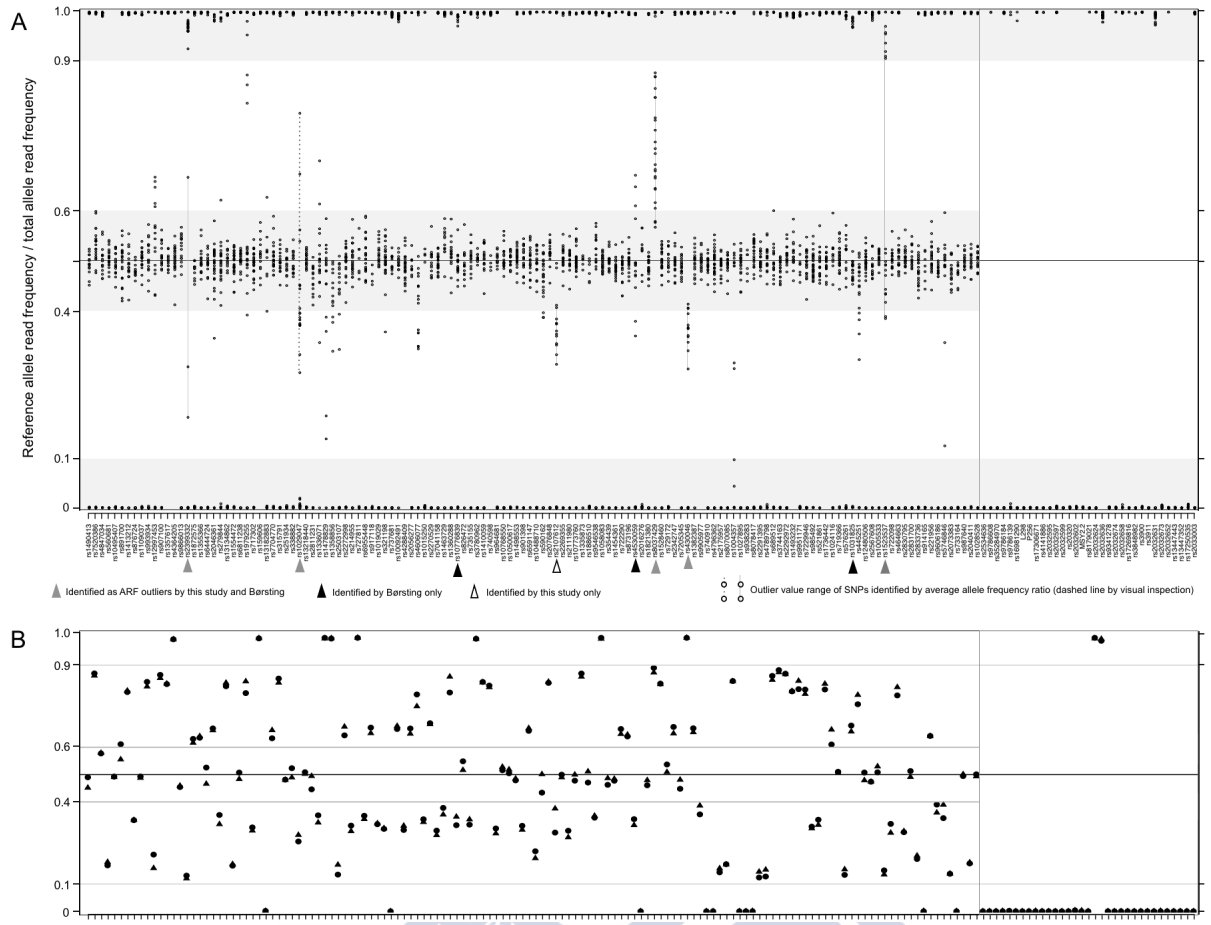


Figure 91: (A) ARF balance in 169 SNPs (listed in Genotyper order, Y-SNPs rightmost) with this study's analysis parameter thresholds marked with grey boxes denoting reference/total ARF ratios of: 0–0.1 and 0.9–1 for A-SNP homozygotes/Y-SNP hemizygotes and 0.4–0.6 for A-SNP heterozygotes. The marked outlier SNPs were identified by recording average ARF ratios (solid lines) or for rs10129047 by visual inspection, as values positioned each side of midline affect the average. Outlier SNPs identified from the study of the same HID SNPs by Børsting et al. (2014) are marked for comparison. (B) ARF balance observed in the 1:1 mixture (S5 and S6 male–female donors), SNPs listed in the same order as (A). Circle and triangle points show replicate values from two independent library runs.

Allele read frequency ratios also apply to homozygotes but in a different way. The presence of other bases at a low proportion in the Ion PGM™ data arise from non-specific incorporation, but the proportion of a second allele must exceed 10% for Genotyper to call the genotype. For this reason, when ARFs reach ≥90% samples cannot be mistyped as heterozygotes.

5.2.2.2.3 Strand bias

Ion PGM™ measures strand bias from forward strand SNP Target Reads divided by total SNP Target Reads, indicating the ratio of sequencing in each direction. Arguably, sequence

output heavily biased towards one strand direction is less reliable, but we observed a large range of strand bias from 0.5 (no discernible bias, equal sequencing of both strands) to values occasionally close to one or zero (output exclusively from forward or reverse strand respectively). We set strand bias to 25-75%: equating to three-fold differences in output from each direction. The range of strand bias values observed is summarised in **Figure 92**. Nine SNPs are marked at the plot ends with average strand bias values outside the threshold set, three of these SNPs gave a small proportion of genotype no-calls and this is discussed in more detail in section 5.2.2.4.2.

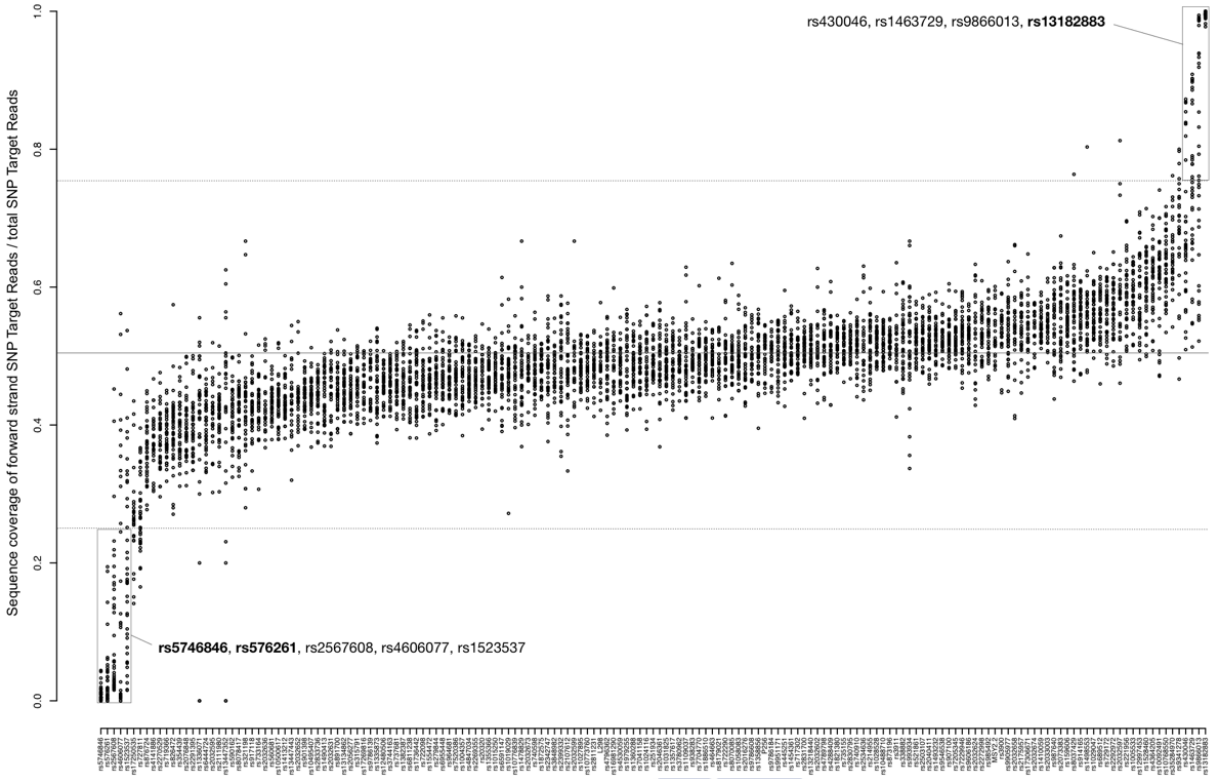


Figure 92: Distribution of strand bias (forward strand SNP Target Reads / total SNP Target Reads) for 136 autosomal HID SNPs. The midline represents no discernible strand bias and dotted lines the 25%-75% value range used to identify nine SNP outliers with mean strand bias values outside this range (extreme values marked by boxes). SNPs in bold gave several no-calls and are discussed in section 5.2.2.4.2 and the IGV overview of rs13182883 is given in Appendix 3D – SNP2.

5.2.2.3 Genotype concordance

Genotype concordance was assessed in three ways: i) between replicate runs of the same sample in each laboratory (inter-run concordance, 13 samples, 38 analyses); ii) between laboratories running identical samples (inter-lab concordance, 6 samples, 24 analyses) and iii) by comparing Ion PGM™ genotypes of Coriell cell-line control DNAs to those listed for HID

SNPs in 1000 Genomes and Complete Genomics public databases. The individual concordance rate for each sample is based on the number of called genotypes, to account for varying numbers of replicates for different samples and varying numbers of no-call results (one or more runs with NN calls for a SNP or ambiguous genotypes in project data). In the following section the total values for no-call, concordance and discordance rates are given.

5.2.2.3.1 Inter-run and inter-lab concordance

The no-call rate for inter-run samples was as low as 1.2% (70/6092) from eleven SNPs, while 99.8% of called genotypes were concordant in between runs of the same sample, with only 0.2% discordant genotypes (13/6022). Discordances were observed in rs2399332, rs1004357, rs938283, rs1979255 and rs2032597 in six different samples. Possible explanations for the discordances and no-calls are discussed in section 5.2.2.4.1.

In addition to the 38 analyses for inter-run concordance we observed a complete absence of discordances and no-calls between library replicate analyses lab1_B and lab1_C. These replicates correspond to Ion PGM™ libraries, prepared from the same original sample, but processed separately in two distinct template preparations and sequencing runs.

Inter-lab concordance of called genotypes was 99.7% (3751/ 3763), with a no-call rate of 0.8% (29/3792). The same five SNPs as those from inter-run comparisons accounted for the inter-lab discordances of 0.3% (12/3763) in five samples. No discordances were seen in 9947A analyses.

5.2.2.3.2 Coriell cell-line control DNA concordance between Ion PGM™ genotypes and online data

Genotypes are available from 1000 Genomes for four of the seven Coriell cell-line control DNAs used (NA06994, NA07000, HG00403, NA18498), while Y-SNPs data is not yet compiled from this project and four A-SNPs are not listed. Therefore, Ion PGM™ vs. 1000 Genomes concordance comparisons assessed 1056 genotypes from 132 SNPs, with a no-call rate of 2.4% (25/1056) from rs1029047, rs13182883, rs13447352, rs2399332 and rs5746846. Genotyping concordance was 99.5% (1026/1031) with 0.5% discordances in rs8078417, rs10768550 and rs2399332, as shown in Table 41. However, during completion of

this study, 1000 Genomes Phase III data was released and two genotyping discordances are now resolved by Phase III revisions, leaving rs2399332 the single discordant genotype (Ion PGM™: TT *versus* 1000 Genomes: GT) amongst 1031 comparisons, giving a revised genotype concordance of 99.9%.

CG online data lists five of the Coriell cell-line control DNAs used (the above DNAs plus NA07029) and includes all HID SNPs, giving 1624 genotype comparisons. CG SNP genotypes for the Coriell cell-line controls DNAs were based on CG assembly software v2.2.0.26, except for the genotypes of sample NA06994 where v2.2.0.19 was used. In addition to 30 no-call genotypes from Ion PGM™ results, 8 no-call genotypes resulted from ambiguous CG genotype calls (Table 41); therefore, the combined no-call rate was 2.3% (38/1624). However, 99.7% (1583/1586) of called genotypes were concordant between Ion PGM™ and CG data. The three discordant genotypes occurred in rs2032597 and rs2399332, as shown in Table 41. SNP rs2399332 also showed a discordance for the same sample between Ion PGM™ and 1000 Genomes, whereas 1000 Genomes and CG gave identical genotypes.

Overall, comparisons of Coriell cell-line control DNA genotype data generated from different SNP genotyping systems indicate a very high concordance rate of 99.8%.

Table 41: Concordance details for comparisons made between Ion PGM™ genotype calls and online data for Coriell cell-line control DNAs. Italic-bold genotypes denote suggested discordances on the basis of consensus.

SNP ID	Coriell cell-line control DNA no.	Ion PGM™ genotype	CG genotype	1000 Genomes-Phase I genotype	1000 Genomes-Phase III genotype	Comments on discordance
Y-rs2032597	NA06994	T	C	(No Y data)	(No Y data)	See Sections 3.3.2 and 3.4.1
Y-rs2032597	NA07029	T	C	(No Y data)	(No Y data)	See Sections 3.3.2 and 3.4.1
rs2399332	NA18498	***TT***	GT	GT	GT	See Sections 3.3.2 and 3.4.1
rs2342747	NA07000	AG	***NN***	AG	AG	No call on either allele in CG
rs4288409	NA18498	AC	***NN***	AC	AC	No call on either allele in CG
rs4847034	NA07000	GG	***NG***	GG	GG	No call for 1st allele in CG
rs4847034	NA07029	GG	***GN***	GG	GG	No call for 2nd allele in CG
rs8078417	HG00403	TT	TT	***CT***	TT	Likely 1000 Genomes-Phase I error
rs10768550	NA18498	CT	CT	***CC***	CT	Identified by CG, but annotated as two base substitution instead of a SNP; 1000 Genomes-Phase I error from neighbouring SNP 2 bp distant

5.2.2.4 Outlier SNPs: HID SNP markers showing discordances or requiring data scrutiny

Outlier SNPs were identified by collating performance data from coverage, analysis parameter thresholds and genotyping concordance. SNPs were ranked according to their risk of mistyping by comparing: i) SNPs with discordant genotypes; ii) SNPs with no-calls; and

iii) SNPs with mean analysis parameter values deviating from thresholds defined for our data set (38 analyses of 13 samples); iv) SNPs without problems. **Figure 93** summarises these four categories and indicates 85.2% of HID SNPs showed no deviation from defined thresholds and were fully concordant. Five SNPs showed discordances, nine had no-calls and eleven gave outlying mean analysis parameter values. SNPs with atypical sequencing characteristics were then analysed in detail by examining their VCF files and using IGV sequence visualisation software (Robinson et al. 2011, Thorvaldsdottir et al. 2013).

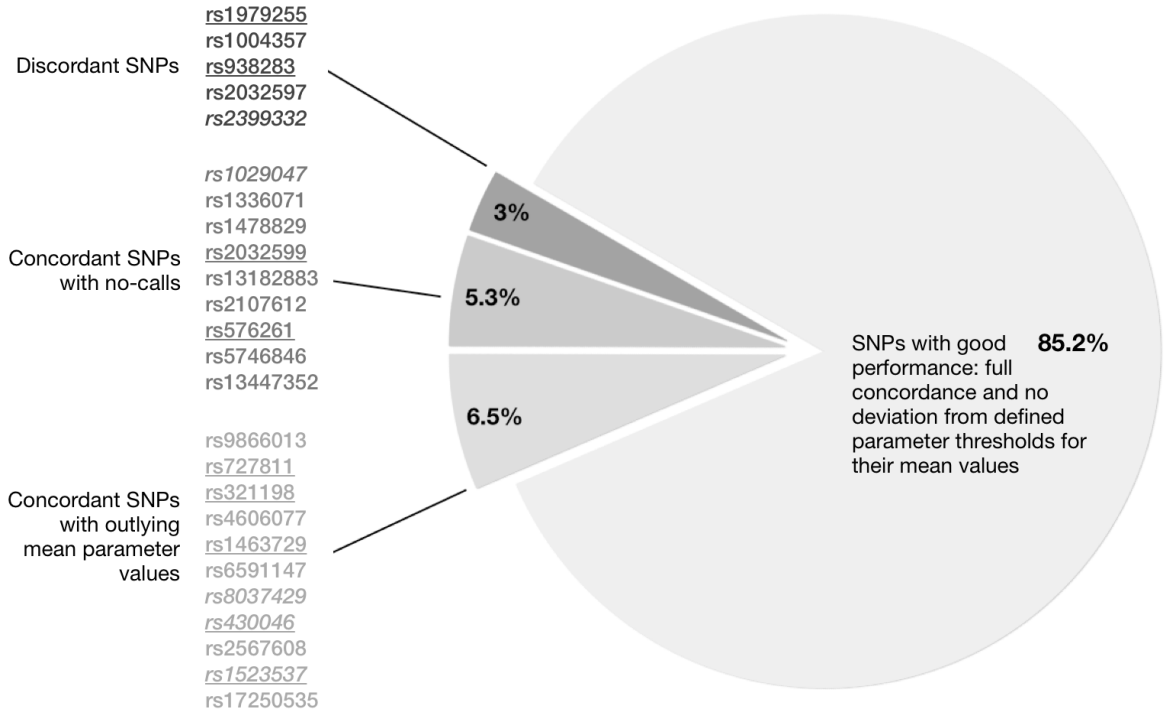


Figure 93: Schematic representation of the proportion of HID SNPs with good performance, por performance or outlier characteristics. Markers listed left were identified as: five SNPs with genotype discordances; nine concordant SNPs with no-calls; eleven concordant SNPs showing deviation from analysis parameter thresholds defined in this study. Underlined SNPs are still retained in the HID SNP set, to the best of the authors' knowledge. Italic SNPs show 5/8 markers recommended for removal by Børsting et al. (2014) study of the same SNP panel. Another three SNPs identified by Børsting: rs10776839; rs4530059 and rs1031825 did not show problematic characteristics in our study.

5.2.2.4.1 Discordant SNPs

Five SNPs were identified as showing consistent patterns of discordant genotyping. Section 5.2.2.3 listed SNPs rs2032597 and rs2399332 as showing genotype differences between replicates in more than one sample and it is notable that they share the characteristic of having closely sited polymeric tracts around the target SNP site.

The A/C Y-SNP rs2032597 gave a non-allelic T base in ~20% of male analyses. As shown in the IGV graphics (Appendix 3D – SNP1), the base immediately upstream of the SNP position is C (the anchor base). An rs2032597-C genotype leads to a large proportion of misaligned reads in both sequencing directions, with a false C insertion being generated and the SNP's C allele becoming the anchor base. This displaces the downstream poly-T tract one base into the SNP position and as it is hemizygous, when the number of T reads exceeds the minimum ARF, the genotype is called T instead of C.

The G/T A-SNP rs2399332 is sited within a poly-T tract. Examination in IGV showed that many G reads had an extra T in the poly-T tract downstream of the SNP position. This caused misalignments and the G allele was considered an insertion, incorrectly placing a T in the SNP position. As this usually happens at a frequency <10%, Genotyper correctly reports GG for most samples, but discordant genotypes can occur when the T frequency exceeds the 10% threshold. This phenomenon explains the above-average allelic misincorporation rate of rs2399332 (**Figure 89**) as well as the single discordant genotype observed (Table 41). Furthermore, rs2399332 shows a clear deviation from expected ARF ratios in heterozygotes (**Figure 91A**). Those samples can reach ARF imbalances of 0.2:0.8 (20% of sequences G), however, IGV shows these are not caused by misalignment from the poly-T tract. For this reason, context sequence was scrutinised for possible primer binding site polymorphisms that could hinder production of sequences carrying the G allele. Several SNPs were found in the region encompassing the amplicon plus 30 bp upstream/downstream of the amplicon ends. In particular SNP rs2399333 is very likely to be in the forward primer-binding site as it is located ~10 bp within the inferred 5'-amplicon end. Furthermore, if the reverse primer is long enough, rs9866331 could also interfere with balanced PCR of each allele as it is ~25 bp within the inferred 3'-amplicon end. Depending on the PCR efficiency and the degree to which neighbour SNPs affect primer binding, the rs9866331-G ARF may drop to 10%, causing heterozygotes to be reported as homozygous T genotypes, as seen in discordant S5 replicates.

The remaining three SNPs had discordant genotypes in 1-2 analyses of single samples. In rs1979255 and rs1004357, heterozygotes had balanced ARFs in all but the single discordant sample. The third SNP rs938283 showed balanced heterozygote allele distributions including the discordant sample. IGV context sequence analysis failed to indicate distinct features that could create misalignments and produce mistyping in the samples analysed.

5.2.2.4.2 SNPs with no-calls

Genotyper reports no-calls when SNPs fail to fulfil *Germline* analysis parameter settings, but additionally dropouts were observed, defined here as SNPs with nil sequence output (QUAL=0). **Figure 94** summarises total SNPs with no-calls or dropouts in 74 analyses (mixtures and lab3-B run excluded). In the 38 concordance analyses, no-calls were recorded in nine SNPs. First, rs5746846, rs576261 and rs13182883 had insufficient coverage in one strand (**Figure 92**). In these SNPs sequencing is initiated on both strands but one fails to reach the SNP position, illustrated by the IGV overview of rs13182883 (Appendix 3D – SNP2) with 0.994 strand bias. This phenomenon produces the very strong strand bias deviations shown at the ends of the distribution plot of **Figure 92** and remains unexplained from all analyses made in IGV.

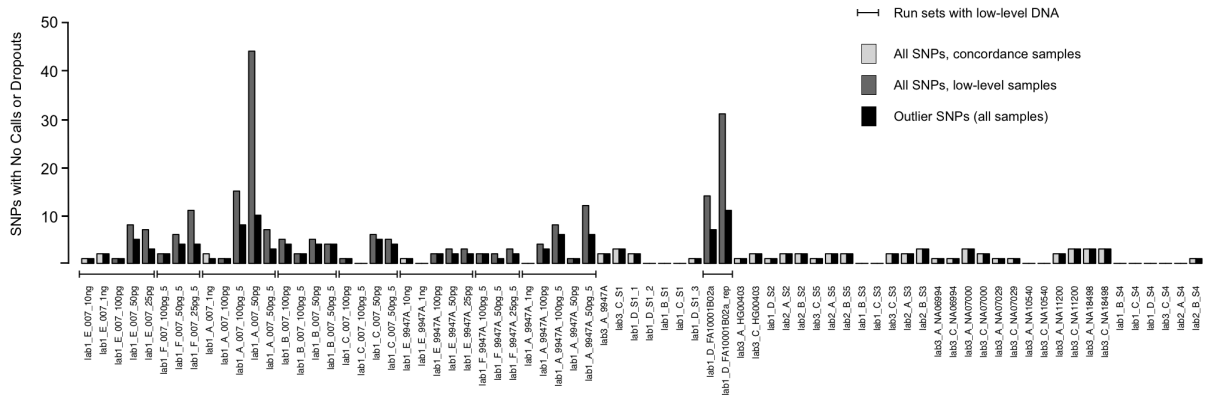


Figure 94: Numbers of SNPs showing no-calls (sequence quality outlying analysis parameter thresholds) or dropouts (QUAL = 0) in concordance study or low-level DNA analyses (marked by horizontal bars for each dilution series or for aDNA S7).

Second, rs13447352 and rs1336071 consistently showed low numbers of sequence reads; failing to reach minimum values for both strands and total coverage. The same observation was made for SNPs rs2032599, rs2107612 and rs1478829, but only in single analyses. Notably, rs1478829 had zero reads in one analysis.

Lastly, as well as the coverage-related analysis parameters and sequence quality thresholds detailed in sections 5.2.2.1 and 5.2.2.2, analysis parameter settings: *VCF minimum quality* (min_variant_score=10) plus *maximum common signal shift* (filter_unusual_predictions=0.3) affected genotype reporting and occasionally caused no-calls, the latter most strongly in rs1029047. Comprehensive review of rs1029047 data in IGV revealed uncertainty about Genotyper heterozygote calls, even when all replicates were

concordant (Appendix 3D – SNP3). This A/T SNP lies between poly-T tract and long poly-A tracts plus several Indels, highly likely to produce systematic misalignments. This same SNP was identified as poorly performing by Børsting et al. (2014), while Seo et al. (2013) also reported discordant genotypes.

5.2.2.4.3 SNPs with mean analysis parameter values deviating from defined thresholds

Despite an absence of genotyping problems affecting the eleven SNPs of this third category (**Figure 93**), examination of their mean values showed consistent atypical behaviour with respect to the analysis parameter thresholds we defined, particularly sequence coverage and strand bias. IGV files from all analyses of the eleven SNPs were scrutinised, but failed to indicate sequence problems. An example is rs430046 that, despite strong strand bias and a high frequency of base deletion calls at the target site, gave consistent genotypes across all replicates (typical IGV data in Appendix 3D – SNP4). There is no strong reason to doubt SNP genotype calls predominantly based on sequences in one direction, despite an increased rate of no-calls observed in such markers.

5.2.2.5 Assessments of Ion PGM™ sensitivity

Assessing sequence data from input DNA well below recommended quantities, the Ion PGM™ system is evidently a very sensitive SNP detection system. Levels of SNP data completeness in low-level DNA analyses are indicated by dark grey columns in **Figure 94**, counting SNPs with no-calls and dropouts. At 100-50-25 pg inputs, SNPs generally show more no-calls/dropouts than optimum input DNA, although runs lab1-E and F maintain good genotyping performance at these lowest inputs. Only rs2016276 appeared disproportionately amongst failing markers in 100-50-25 pg dilutions, giving 6/23 male and 6/13 female no-calls. Although concordance study DNAs mainly had missing genotypes in only 1-3 SNPs, low-level DNA rarely exceeded 8-12 SNPs with missing genotypes. Furthermore, this has little impact on random match probability (RMP) values. **Figure 95** indicates approximately 40-50% of missing data (including outlier SNPs) is needed to decrease the cumulative RMP to a value similar to GlobalFiler. Half or less of outlier SNPs (using each category defined in

Section 5.2.2.4) had missing genotypes in aDNA and lab1-A runs. Five extra library amplification cycles did not increased sensitivity.

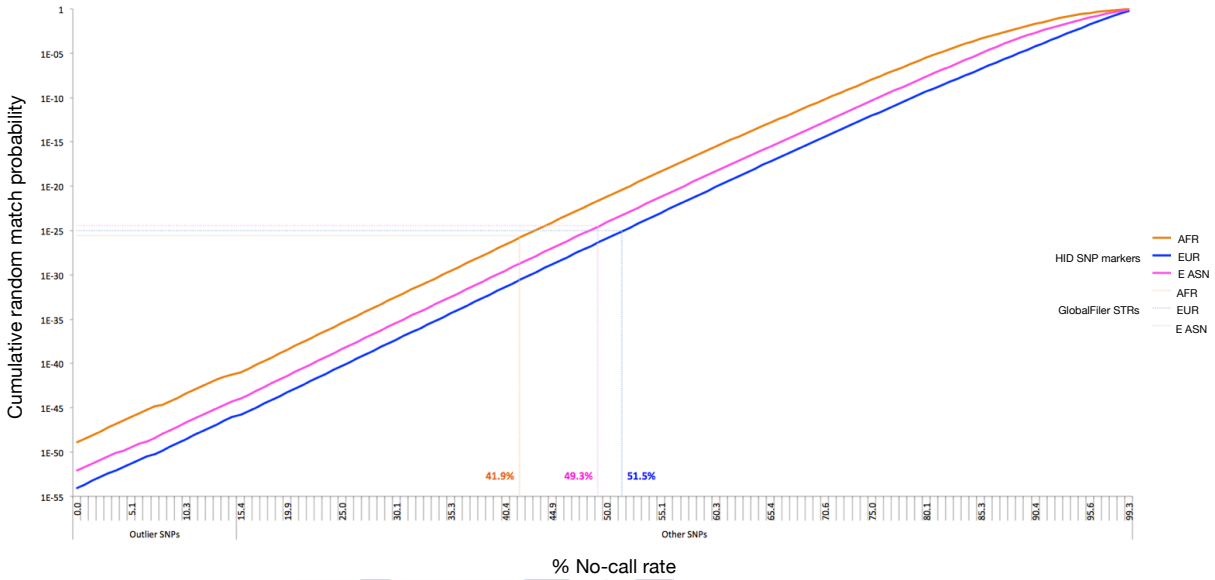


Figure 95: Reduction in cumulative RMP with increasing no-call rate.

The highly degraded aDNA sample gave more SNP failures than most dilution series analyses. Although this is limited initial NGS data, these results indicate very high sensitivity for Ion PGMTM when target sequences are highly degraded or inhibited. Therefore, although good sensitivity to low-level DNA has been recognised in this and other studies (Seo et al. 2013, Børsting et al. 2014), specific effects of aggressive degradation need to be comprehensively assessed to properly test the effectiveness of NGS analysing skeletal remains typical of missing person identification.

Although the two analyses of aDNA gave relatively low levels of SNP Target Reads and the lowest mean read lengths of any samples (data not shown), genotypes had very good levels of agreement. In all, 128/169 genotype pairs were called identically (75.7%) and a further 23 genotypes called from one analysis (totalling 89.3% genotypes). More no-calls and dropouts (QUAL=0) were recorded applying library re-amplification. The 25-cycle PCR gave 10 no-calls, 4 dropouts, whereas 25+5 cycles gave 18 no-calls, 13 dropouts (6 no-calls, 4 dropouts in common). Unmodified PCR also achieved higher average sequence coverage and quality scores: 128 sequences and QUAL=422.7, compared to 72 sequences and QUAL=286.5 in 25+5 analysis, plus just 1/23 singleton genotypes.

The slight rise in numbers of common genotypes to ‘common results’ (same SNPs giving genotypes or no-calls/dropouts in both analyses) from 75.7% to 81.7%, suggests some locus

dropout in Ion PGM™ SNP genotyping may be systematic rather than random, but many more highly degraded DNA samples must be assessed to test this assumption. Despite lacking reference genotypes, the aDNA heterozygosity of 51% compares to an expected 46% heterozygosity for these SNPs (1000 Genomes CEU data), suggesting very little allele dropout.

5.2.2.6 Mixture analysis

Detection of mixed source DNA and possible identification of components in simple mixtures is challenging when genotyping binary SNPs with the commonly used SNaPshot® system. In contrast, NGS data from this study of Ion PGM™ and AmpliSeq™ technology gave balanced heterozygous genotypes, providing a more secure basis for analysing mixtures. It is important to reliably recognise SNP data as originating from a mixture and not a single profile. Furthermore, development of enhanced statistical analyses, prompted by our results from Ion PGM™ runs, will allow likelihood ratio calculations when one of the component DNAs is known. For these reasons, our assessment of NGS data from artificial mixtures was more comprehensive than for the other DNAs. Detailed descriptions of these mixed sequence data analyses are given in Appendix 3E.

Scrutiny of the single 1:1 mixture ARF plot of **Figure 91B** and all other ARF plots in Figure S9 in Appendix 3E, shows mixtures generally have patterns quite distinct from unmixed samples, with more heterozygous SNPs outside the 40-60% ARF range. Additionally, increased heterozygosity and reduced Y-SNP coverage provide clear indications of the presence of mixed DNA in HID SNP data (Table S5 in Appendix 3E). Our initial analyses of limited numbers of mixtures indicate *Germline* analysis parameter settings should be used for forensic samples of unknown origin. If any of the described mixture indicators is found, data should then be re-analysed with *Somatic* settings to improve accuracy of A-SNP genotyping. Even with this two-tier approach, care is needed with more extreme mixture ratios (herein 1:9 and 9:1), as there is increased probability minor alleles escape detection. Y-SNPs should be analysed independently with *Germline* analysis parameter settings as this guarantees higher genotyping rates while maintaining allele call quality.

5.2.2.7 Context sequence examinations with IGV

To further assess HID SNPs for forensic analysis, the context sequence of each marker was scrutinised using IGV. This provided checks on characteristics that could influence alignment, including Indels or polymeric tracts, but also screened for extra polymorphisms close to target sites. In staff donors, we detected clustering polymorphisms associated with target SNPs. Table 42 summarises data for these additional polymorphisms. In SNP rs430046 there are three well-characterised and closely sited SNPs adding discrimination power (all variant allele homozygotes in Appendix 3D – SNP4). Variants at sequence extremes and next to polymeric tracts tended to produce unreliable reads (see rs1109037 in Appendix 3D – SNP5).

Table 42: Details of clustering variants identified from IGV analysis of HID SNP sequences.

HID SNP	Clustering Variant	Type	Alleles (Ref/Alt)	C	Position	Minor allele frequency range	Comments
rs891700	rs12047255	SNP	A/G	1	239881878	0.125–0.174	
rs1413212	rs10926803	SNP	T/C	1	242806748	0.085–0.342	
rs1413212	rs6669024	SNP	C/A	1	242806743	0.283–0.517	
rs876724	rs35414538	Indel	Del/In	2	114976	Not reported	Poly-A tract
rs1109037	No reported	SNP	G/A	2	10085636	Not reported	
rs1109037	rs34861500	Indel	In/Del	2	10085764	Not reported	Poly-C tract, at end of sequence
rs12997453	rs72883670	SNP	C/T	2	182413238	0.142–0.225	
rs9866013	rs9883594	SNP	A/T	3	59488282	0.329–0.368	
rs279844	rs279845	SNP	T/A	4	46329723	0.456–0.556	
rs338882	rs42875	SNP	A/G	5	178690776	0.075–0.233	
rs7704770	rs35593173	Indel	In/Del	5	159487969	Not reported	Poly-C tract
rs1029047	rs201933068	Indel	In/Del	6	1135938	Not reported	Same position as HID SNP
rs1336071	rs7760004	SNP	C/T	6	94537144	0.194–0.476	
rs727811	rs1390470	SNP	C/T	6	165045290	0.022–?	
rs1478829	rs7751035	SNP	C/T	6	120560627	0.246–0.483	
rs6955448	rs6950322	SNP	G/A	7	4310317	0.288–0.3	
rs6955448	rs6955464	SNP	C/T	7	4310397	0.244–0.347	
rs4288409	rs35574091	Indel	In/Del	8	136839227	Not reported	
rs4606077	rs58774517	SNP	C/T	8	144656763	0.075–0.167	
rs4606077	rs1869434	SNP	G/A	8	144656764	0.192–0.432	
rs10776839	rs7037930	SNP	A/G	9	137417305	0.103–0.325	
rs2270529	rs2270530	SNP	A/C	9	14747156	0.261–0.3	
rs6591147	rs72975101	SNP	C/T	11	105912913	0.033–0.153	
rs2076848	rs5795898	Indel	Del/In	11	134667482	0.325–?	
rs954538	rs60940032	Indel	Del/In	13	84456695	Not reported	Poly-A tract
rs1058083	rs701537	SNP	A/T	13	100038271	0.326–0.417	
rs1058083	rs75653253	SNP	G/A	13	100038285	Not reported	
rs2016276	rs72705536	SNP	C/G	15	24571814	0.008–0.117	
rs2342747	rs2342748	SNP	G/C	16	5868729	0.222–0.450	
rs430046	rs381840	SNP	A/T	16	78017077	0.008–0.034	See Supplementary file S2 – SNP 4
rs430046	rs430044	SNP	C/T	16	78017045	0.263–0.467	See Supplementary file S2 – SNP 4
rs430046	rs409820	SNP	C/A	16	78017034	0.242–0.482	See Supplementary file S2 – SNP 4
rs7205345	rs34743902	SNP	T/C	16	7520277	0.034–0.225	
rs9905977	rs73298992	SNP	C/T	17	2919461	0.042–0.133	
rs740910	rs60810599	SNP	A/G	17	5706584	0.076–0.083	
rs985492	Unassigned	SNP	C/T	18	29311074	Not reported	
rs985492	Unassigned	Indel	In/Del	18	29311062	Not reported	
rs1736442	rs371957125	Indel	In/Del	18	55225736	Not reported	Poly-G tract
rs445251	rs376918760	SNP	T/C	20	15124994	Not reported	
rs2567608	rs3746728	SNP	C/T	20	23017044	0.244–0.407	
rs2567608	rs2567609	SNP	T/C	20	23017017	0.378–0.617	
rs12480506	rs6034433	SNP	T/C	20	16181362	0.267–0.542	
rs914165	rs755095	SNP	C/G	21	42415976	0.006–0.212	
rs722098	rs55916325	SNP	G/A	21	166588530	0.042–0.083	
rs2830795	rs12626695	SNP	T/C	21	28608125	0.033–0.167	
rs2073383	rs2073384	SNP	C/T	22	23802242	0.307–?	
rs20320	rs13305774	SNP	A/G	Y	14898094	0.149–?	
rs9786139	rs9785971	SNP	G/A	Y	6753511	0.253–?	

In contrast to SNPs close to target sites, Indel discovery and genotyping with Ion PGM™ sequence data remains more restricted. Small sequencing errors, usually linked to short polymeric tracts of four or more bases, tend to produce artefact Indels at high frequency. Mostly deletions were observed in such cases, but insertions occasionally occur in misaligned polymeric tracts. Two other observations made from IGV sequences are worth noting. First, Indel artefacts are affected by sequence directionality and tend to occur exclusively on one strand, aiding the differentiation of true from artefact Indels (rs430046 in Appendix 3D – SNP4, shows 12 direction-dependent Indel calls). Second, false Indels can be generated from misaligned sequences containing repetitive motifs, although handling of short tandem repeat alignments is being refined and such artefacts will be better controlled as sequence analysis software improves.

5.2.2.8 Concluding remarks

The evaluation of Ion PGM™ sensitivity and genotyping accuracy made here, give strong support for the application of NGS technology to forensic DNA analysis. Sequence data obtained in all three laboratories had sufficiently high coverage and gave reliable SNP genotyping for most loci in HID SNP. We discovered five SNPs with discordances that should be excluded from the panel. We note rs1004357 and rs2032597 are already removed from the revised version of HID SNP, while rs2399332 was identified in Børsting et al. (2014) study as a problematic SNP. We also found discordant genotypes in rs1979255 and rs938283, and their continued inclusion in HID SNP needs critical review. Furthermore, rs2107612 showed imbalanced heterozygote reads and should also be removed from the panel in addition to the eight problematic markers identified by Børsting et al. Lastly, mention should be made of rs1029047, which gives genotyping inconsistencies in all NGS studies of this SNP made so far (Seo et al. 2013, Børsting et al. 2014). There are clearly characterised context sequence factors affecting the alignment and therefore the reliability of allele calls for rs1029047 (Appendix 3D – SNP3), which have not affected SNaPshot genotyping of this SNP (Sánchez et al. 2006). Therefore, careful scrutiny of sequence characteristics is required of any SNP chosen for forensic use. This is particularly important for coding SNPs in forensic phenotype predictive tests, since these must work well for the SNP analyses to be sufficiently informative.

The estimation of optimum sample numbers for each of the six Ion PGM™ chip versions presented this study with the biggest challenge, both in harmonising NGS runs across three laboratories and ensuring the coverage obtained was appropriate for assessing forensic sensitivity. Since low-level DNA appears to accentuate coverage variability in HID SNP markers, this will be a major problem when initially optimising NGS for routine forensic use. As Ion PGM™ chip capacities have now reached very reasonable levels of sequence output, users can be cautious by loading fewer samples than coverage estimation guidelines suggest. Furthermore, there is some consensus that ~15-20 minimum coverage thresholds can safeguard the reliability of allele calls made with NGS (Bentley et al. 2008, Quail et al. 2012, Daniel et al. 2015).

Although the Torrent Suite™ software provides several sequence quality parameters in the data output, we found there was little or no scope for changing the default analysis parameters settings to more aggressive thresholds. Setting such thresholds would provide a way to exclude miscalled genotypes from underperforming SNPs or mixed DNA. This finding has consequences for the average forensic scientist's capacity to properly scrutinise the extensive data that Ion PGM™ produces. Since mixture detection with binary markers is severely restricted compared to multi-allele STRs, it is all the more important to properly assess deviations from balanced heterozygote patterns. We largely agree with the conclusions of Børsting et al. (2014) that the Ion PGM™ analysis software needs further optimisation to be fully suitable for forensic application, although it is being constantly revised to this end. In particular, there is an evident need to apply *Somatic* analysis parameter settings to properly analyse mixtures, even though *Germline* analysis parameter settings are set in place for forensic SNP analysis with Ion PGM™. This reduces the capacity of the system to alert the analyst to mixtures and represents a critical shortfall when the very high sensitivity of Ion PGM™ is borne in mind.

5.2.3 Building a forensic ancestry panel from the ground up: the EUROFORGEN Global AIM-SNP set (Appendix 3F)

5.2.3.1 Characteristics of the ancestry informative SNPs selected

A final set of 122 bi-allelic and 6 tri-allelic SNPs were selected from a total candidate pool of 189 loci (and 12 tri-allelic) and are detailed in Table 43. Global AIM-SNP allele frequency distributions in five population groups are summarized in **Figure 96**. The cumulative PSD values in each group required a smaller number of AFR-informative SNPs and for 28 candidates, Oceanian and American allele frequency data was absent from the Stanford HGDP-CEPH study so these loci were kept in reserve. This pool of alternative SNPs able to substitute loci failing in multiplex comprised: 6 tri-allelics; 12 AFR; 17 EUR; 7 AME; 2 E ASN and one OCE informative SNP.

To measure the overall population differentiation power of the 128 Global AIM-SNPs, we compared their individual F_{ST} values for three, four and five group comparisons with SNPs in the Kosoy et al. (2009) ancestry set. The Kosoy set also has 128 AIM-SNPs and only rs9522149 in common. Observed F_{ST} values are shown in ascending order in **Figure 97**. Average F_{ST} values for AFR-EUR-E ASN comparisons are Kosoy=0.260, Global=0.378, AFR-EUR-E ASN-AME comparisons: 0.321 *versus* 0.438 and AFR-EUR-E ASN-AME-OCE: 0.309 *versus* 0.487. Overall, only 38 of 384 SNP comparisons indicate higher differentiation power for Kosoy SNPs, none for F_{ST} values for AFR-EUR-E ASN-AME-OCE comparisons. As with the LACE panel, Kosoy's SNP set aims to primarily differentiate AFR, EUR and AME co-ancestry components as these are found in a large proportion of US and Latin American admixed individuals used in association studies. Therefore many individual F_{ST} values should be higher for AFR-EUR-E ASN-AME comparisons. However, only 21 Kosoy SNPs have higher F_{ST} values for this comparison. The largest contrast of average F_{ST} is observed in AFR-EUR-E ASN-AME-OCE comparisons, with 63% higher differentiation in Global AIM-SNPs. Lastly, a noticeable jump in F_{ST} values is seen in the four and five-group charts for about 10 Global AIM-SNPs at the far right. These could be considered the core forensic AIM-SNP set, showing the highest ancestry informativeness. Top components are: rs2814778; rs16891982;

Table 43: Details of 128 markers selected for the Global AIM-SNP set.

AIM details							Reference allele frequencies					Individual population-specific divergence (PSD) and cumulative PSD (CD) values									
PG	SNP ID	C	Source	Position	Gene	RA	Afr	Eur	Asn	Ame	Oce	Afr PSD	Afr CD	Eur PSD	EurCD	AsnPSD	AsnCD	Ame PSD	Ame CD	OCE PSD	OCE CD
1	rs9908046	17	USC CCAS panel	53563782	–	C	0.965	0.949	0.879	0.992	0.018	0.010	0.01	0.008	0.01	0.002	0.00	0.024	0.02	0.522	0.52
2	rs2139931	1	USC CCAS panel	84590527	PRKACB	A	0.897	0.774	0.904	0.898	0.018	0.008	0.02	0.006	0.01	0.013	0.01	0.007	0.03	0.424	0.95
3	rs715605	22	Novel	30640308	–	T	0.851	0.913	0.991	1	0.089	0.006	0.02	0.000	0.01	0.034	0.05	0.025	0.06	0.415	1.36
4	rs3751050	11	Pacifiplex	9091244	SCUBE2	A	0.949	0.888	0.96	0.961	0.089	0.006	0.03	0.001	0.01	0.012	0.06	0.008	0.06	0.411	1.77
5	rs6054465	20	USC CCAS panel	6673018	–	T	0.962	0.87	0.748	0.859	0.036	0.036	0.07	0.005	0.02	0.010	0.07	0.001	0.06	0.396	2.17
6	rs26951	5	Novel	59759657	PDE4D	G	0.997	0.907	0.629	0.82	0.036	0.083	0.15	0.024	0.04	0.047	0.12	0.000	0.06	0.380	2.55
7	rs6886019	5	USC CCAS panel	170245846	–	C	0.941	0.942	1	0.883	0.161	0.000	0.15	0.001	0.04	0.031	0.15	0.004	0.07	0.374	2.92
8	rs10970986	9	USC CCAS panel	32453278	–	T	0.981	0.726	0.734	0.711	0.018	0.085	0.24	0.002	0.05	0.001	0.15	0.002	0.07	0.362	3.28
9	rs16830500	2	Novel	152814129	CACNB4	T	0.938	0.939	0.428	0.875	0	0.049	0.28	0.079	0.12	0.128	0.28	0.014	0.08	0.359	3.64
10	rs3804030	21	Novel	45629565	–	A	0.811	0.903	0.865	0.992	0.089	0.003	0.29	0.001	0.13	0.005	0.28	0.043	0.13	0.356	4.00
11	rs2274636	10	USC CCAS panel	27443012	MASTL	A	0.995	0.882	0.719	0.906	0.107	0.066	0.35	0.006	0.13	0.022	0.30	0.007	0.13	0.314	4.31
12	rs4806654	19	Novel	55768276	PPP6R1	T	0.719	0.924	0.886	0.813	0.125	0.018	0.37	0.020	0.15	0.004	0.31	0.001	0.14	0.303	4.62
13	rs4391951	13	Novel	44755071	–	T	0.892	0.841	0.477	0.883	0.036	0.035	0.41	0.027	0.18	0.069	0.38	0.024	0.16	0.303	4.92
14	rs10149275	14	Novel	43268640	–	G	0.881	0.903	0.26	0.805	0.018	0.033	0.44	0.097	0.28	0.188	0.56	0.013	0.17	0.295	5.21
15	rs1509524	4	USC CCAS panel	125455038	–	A	0.784	0.795	0.82	0.844	0.089	0.000	0.44	0.000	0.28	0.002	0.57	0.004	0.18	0.288	5.50
16	rs7623065	3	Pacifiplex	22385375	–	A	0.251	0.684	0.871	0.969	0	0.131	0.57	0.001	0.28	0.061	0.63	0.101	0.28	0.284	5.79
17	rs10811102	9	Pacifiplex	1911291	–	G	0.995	0.774	0.32	0.336	0.018	0.181	0.75	0.033	0.31	0.101	0.73	0.049	0.33	0.265	6.05
18	rs10455681	6	Pacifiplex	69802502	BAI3	A	0.989	0.846	0.178	0.445	0.018	0.175	0.93	0.083	0.39	0.216	0.94	0.018	0.34	0.260	6.31
19	rs7832008	8	Pacifiplex	98358246	–	G	0.53	0.137	0.612	0.156	1	0.016	0.94	0.100	0.49	0.054	1.00	0.038	0.38	0.257	6.57
20	rs9809818	3	Pacifiplex	71480566	FOXP1	C	0.019	0.088	0.871	0.82	0.982	0.167	1.11	0.152	0.65	0.264	1.26	0.116	0.50	0.254	6.82
21	rs9934011	16	Novel	13915807	–	T	0.814	0.816	0.217	0.68	0.018	0.041	1.15	0.072	0.72	0.160	1.42	0.004	0.50	0.247	7.07
22	rs3784651	15	Pacifiplex	94925273	MCTP2	T	0.516	0.879	0.248	0.836	0	0.004	1.16	0.137	0.86	0.123	1.55	0.043	0.55	0.240	7.31
23	rs2282107	21	Novel	37707581	MORC3	A	0.881	0.932	0.579	0.867	0.161	0.012	1.17	0.030	0.89	0.040	1.59	0.006	0.55	0.226	7.53
24	rs12405776	1	Pacifiplex	242431557	PLD5	C	0.497	0.979	0.413	0.672	0.089	0.023	1.19	0.200	1.09	0.070	1.66	0.000	0.55	0.203	7.74
25	rs1592672	12	Pacifiplex	80128593	–	T	0.065	0.891	0.248	0.602	0	0.174	1.36	0.253	1.34	0.062	1.72	0.007	0.56	0.182	7.92
26	rs10183022	2	Pacifiplex	237481969	CXCR7	G	0.027	0.611	0.631	0.813	0.982	0.252	1.62	0.010	1.35	0.011	1.73	0.053	0.61	0.173	8.09
27	rs1877751	20	LACE	57967906	–	A	0.849	0.692	0.166	0.031	0.036	0.106	1.72	0.053	1.40	0.124	1.85	0.179	0.79	0.163	8.26
28	rs798949	7	Pacifiplex	120765954	C7orf58	T	0.778	0.525	0.15	0.07	0.929	0.092	1.81	0.010	1.41	0.099	1.95	0.110	0.90	0.158	8.41
1	rs1557553	22	LACE	44760984	–	C	0.962	0.908	0.717	0.086	0.786	0.046	1.86	0.028	1.44	0.011	1.96	0.473	1.37	0.000	8.41
2	rs2080161	7	LACE	13331150	–	T	0.968	0.789	0.713	0.099	0.82	0.075	1.93	0.004	1.44	0.002	1.97	0.415	1.79	0.023	8.44
3	rs10483251	14	LACE	21671277	–	G	0.889	0.779	0.923	0.099	0.712	0.013	1.95	0.000	1.44	0.033	2.00	0.413	2.20	0.005	8.44
4	rs12498138	3	CT-plex	121459589	GOLGB1	G	0.995	0.916	0.914	0.094	0.911	0.046	1.99	0.006	1.45	0.004	2.00	0.411	2.61	0.002	8.44
5	rs8137373	22	USC CCAS panel	41729216	ZC3H7B	G	0.827	0.742	0.923	0.023	0.982	0.004	2.00	0.002	1.45	0.041	2.04	0.394	3.01	0.062	8.51
6	rs1452501	16	LACE	80623262	–	C	0.978	0.959	0.874	0.214	0.635	0.030	2.03	0.028	1.48	0.000	2.04	0.376	3.38	0.042	8.55
7	rs17130385	10	USC CCAS panel	115196019	–	G	0.976	0.934	0.769	0.086	0.518	0.049	2.08	0.034	1.51	0.005	2.05	0.357	3.74	0.058	8.60
8	rs2471552	7	CT-plex	45977173	–	C	0.081	0.211	0.217	0.945	0.107	0.032	2.11	0.001	1.51	0.000	2.05	0.340	4.08	0.015	8.62
9	rs5757362	22	LACE	39306080	–	T	0.014	0.375	0.219	0.935	0.212	0.118	2.23	0.012	1.53	0.007	2.06	0.322	4.40	0.004	8.62
10	rs17359176	13	LACE	23667334	–	G	0.997	0.911	0.822	0.204	0.982	0.062	2.29	0.010	1.54	0.002	2.06	0.320	4.72	0.033	8.66
11	rs174570	11	LACE	61597212	FADS2	C	0.992	0.841	0.593	0.031	0.482	0.115	2.40	0.023	1.56	0.023	2.08	0.319	5.04	0.035	8.69
12	rs10012227	4	USC CCAS panel	18637315	–	G	0.905	0.867	0.528	0.047	0.768	0.044	2.45	0.044	1.60	0.040	2.12	0.304	5.34	0.002	8.69
13	rs11960137	5	LACE	155338081	–	C	0.938	0.859	0.785	0.122	0.964	0.029	2.48	0.007	1.61	0.001	2.12	0.304	5.65	0.035	8.73
14	rs2051827	12	USC CCAS panel	47956031	–	G	0.949	0.932	0.937	0.195	0.518	0.013	2.49	0.011	1.62	0.011	2.13	0.303	5.95	0.085	8.81
15	rs12402499	1	CT-plex	101528954	–	G	1	0.909	1	0.258	1	0.033	2.52	0.000	1.62	0.041	2.17	0.300	6.25	0.014	8.83
16	rs11625446	14	LACE	48244558	–	C	0.951	0.739	0.614	0.043	0.982	0.081	2.60	0.002	1.62	0.009	2.18	0.296	6.55	0.089	8.92
17	rs4979274	9	USC CCAS panel	116444269	–	C	0.995	0.88	0.57	0.086	0.946	0.105	2.71	0.032	1.66	0.044	2.23	0.293	6.84	0.040	8.96
18	rs7151991	14	LACE	32635572	–	A	0.154	0.199	0.15	0.935	0	0.005	2.71	0.001	1.66	0.007	2.23	0.277	7.12	0.057	9.01
19	rs4780476	16	USC CCAS panel	12862007	CPPED1	A	0.227	0.308	0.247	0.945	0.232	0.007	2.72	0.000	1.66	0.006	2.24	0.272	7.39	0.004	9.02
20	rs6088466	20	LACE	32913534	–	G	0.978	0.679	0.572	0.034	0.75	0.131	2.85	0.000	1.66	0.010	2.25	0.266	7.65	0.005	9.02
21	rs2302013	2	USC CCAS panel	242042331	FARP2	T	1	0.997	0.845	0.344	0.339	0.045	2.90	0.057	1.71	0.002	2.25	0.219	7.87	0.168	9.19
22	rs4792928	17	USC CCAS panel	42105174	–	T	1	0.955	0.369	0.195	0.804	0.126	3.02	0.111	1.82	0.154	2.40	0.175	8.05	0.004	9.20
1	rs2814778	1	34plex	159174683	DARC	A	0.003	0.997	1	0.992	1	0.674	3.70	0.123	1.95	0.105	2.51	0.065	8.11	0.051	9.25
2	rs2789823	9	CT-plex	136769888	VAV2	G	0.908	0.003	0	0	0	0.528	4.23	0.109	2.06	0.093	2.60	0.057	8.17	0.044	9.29
3	rs1871534	8	Genplex	145639681	SLC39A4	C	0.886	0.003	0	0	0	0.503	4.73	0.106	2.16	0.091	2.69	0.056	8.23	0.042	9.33
4	rs6875659	5	LACE	175158653	–	G	0.046	0.916	0.956	0.977	1	0.483	5.21	0.051	2.21	0.071	2.76	0.062	8.29	0.065	9.40
5	rs1369290	18	Kersbergen	67691520	RTTN	A	0.881	0.032	0	0	0	0.463	5.67	0.069	2.28	0.097	2.86	0.060	8.35	0.046	9.44

Table 43 *(continued)*

AIM details							Reference allele frequencies					Individual population-specific divergence (PSD) and cumulative PSD (CD) values									
PG	SNP ID	C	Source	Position	Gene	RA	Afr	Eur	Asn	Ame	Oce	Afr PSD	Afr CD	Eur PSD	EurCD	AsnPSD	AsnCD	Ame PSD	Ame CD	OCE PSD	OCE CD
6	rs310644	20	Kiddlab	62159504	PTK6	A	0.043	0.928	0.965	0.953	0.107	0.454	6.13	0.073	2.36	0.091	2.95	0.051	8.40	0.239	9.68
7	rs1197062	17	LACE	58641118	–	A	0.111	0.953	0.974	1	0.857	0.432	6.56	0.065	2.42	0.072	3.02	0.068	8.47	0.004	9.69
8	rs6034866	20	USC CCAS panel	17603728	RRBP1	A	0.924	0.058	0.082	0.054	0.179	0.430	6.99	0.075	2.50	0.042	3.07	0.045	8.51	0.003	9.69
1	rs1426654	15	34plex	48426484	SLC24A5	A	0.043	0.996	0.014	0.039	0	0.152	7.14	0.617	3.11	0.237	3.30	0.126	8.64	0.145	9.83
2	rs16891982	5	34plex	33951693	SLC45A2	C	0.992	0.029	0.983	0.984	1	0.115	7.26	0.547	3.66	0.229	3.53	0.154	8.79	0.154	9.99
3	rs8072587	17	USC CCAS panel	19211073	EPN2	C	0.978	0.192	1	0.796	1	0.135	7.39	0.370	4.03	0.207	3.74	0.019	8.81	0.122	10.11
4	rs12142199	1	AFR-EUR admix SNPs	1249187	CPSF3L	G	0.954	0.2	0.983	1	1	0.104	7.50	0.369	4.40	0.168	3.91	0.128	8.94	0.107	10.22
5	rs9522149	13	CT-plex	111827167	ARHGEF7	T	0.954	0.258	0.997	0.945	1	0.090	7.59	0.334	4.73	0.175	4.08	0.121	9.06	0.107	10.32
6	rs7531501	1	USC CCAS panel	234338303	–	G	0.065	0.886	0.128	0.258	0.089	0.136	7.72	0.330	5.06	0.109	4.19	0.018	9.08	0.084	10.41
7	rs12913832	15	34plex	28365618	HERC2	A	1	0.289	0.998	0.883	1	0.140	7.86	0.315	5.38	0.169	4.36	0.028	9.11	0.089	10.50
8	rs7084970	10	AFR-EUR admix SNPs	119750413	–	C	0.854	0.063	0.769	0.833	0.214	0.117	7.98	0.294	5.67	0.086	4.45	0.068	9.18	0.044	10.54
9	rs11778591	8	LACE	12720349	–	C	0.624	0.088	0.895	0.786	0.857	0.010	7.99	0.282	5.95	0.175	4.62	0.059	9.23	0.076	10.62
10	rs4749305	10	LACE	28391596	MPP7	A	0.332	0.849	0.051	0.018	0.036	0.007	8.00	0.279	6.23	0.186	4.81	0.173	9.41	0.126	10.74
11	rs820371	3	LACE	123404711	MYLK	A	0.965	0.237	0.949	0.734	0.981	0.123	8.12	0.278	6.51	0.132	4.94	0.003	9.41	0.113	10.86
12	rs1924381	13	LACE	72321856	DACH1	A	0.924	0.13	0.811	0.661	0.786	0.138	8.26	0.272	6.78	0.077	5.02	0.012	9.42	0.034	10.89
13	rs2715883	11	Genplex	120133494	POU2F3	A	0.122	0.708	0.005	0	0	0.039	8.30	0.270	7.05	0.172	5.19	0.115	9.54	0.096	10.99
14	rs595961	1	DNAprint	36367780	EIF2C1	A	0.084	0.859	0.189	0.172	0.464	0.128	8.42	0.264	7.32	0.075	5.26	0.051	9.59	0.000	10.99
15	rs3759171	12	LACE	72307616	TBC1D15	A	0.032	0.743	0.103	0.048	0.018	0.121	8.55	0.264	7.58	0.077	5.34	0.084	9.67	0.109	11.09
16	rs917115	7	Kiddlab	28172586	JAZF1	T	0.178	0.779	0.024	0.156	0.179	0.035	8.58	0.262	7.84	0.184	5.52	0.037	9.71	0.029	11.12
17	rs1486341	12	LACE	39042063	–	A	0.957	0.126	0.733	0.679	0.911	0.183	8.76	0.259	8.10	0.042	5.57	0.019	9.73	0.098	11.22
18	rs11074130	15	LACE	93583742	–	T	0.776	0.096	0.773	0.780	0.804	0.063	8.83	0.257	8.36	0.081	5.65	0.036	9.76	0.055	11.28
19	rs7937598	11	LACE	44745048	–	A	0.778	0.151	0.827	0.982	0.929	0.041	8.87	0.255	8.61	0.083	5.73	0.169	9.93	0.102	11.38
20	rs10186877	2	LACE	216618818	–	G	0.941	0.279	0.902	0.982	0.911	0.090	8.96	0.242	8.86	0.078	5.81	0.127	10.06	0.050	11.43
21	rs1567803	2	LACE	101343018	–	C	0.965	0.322	0.916	0.969	1	0.100	9.06	0.232	9.09	0.074	5.88	0.084	10.14	0.095	11.52
22	rs2889678	20	LACE	31189993	–	C	0.949	0.339	0.948	0.961	0.946	0.078	9.14	0.229	9.32	0.096	5.98	0.073	10.22	0.058	11.58
23	rs4791868	17	LACE	9698244	–	G	0.157	0.895	0.341	0.125	0.429	0.111	9.25	0.225	9.54	0.032	6.01	0.104	10.32	0.004	11.58
24	rs7630522	3	LACE	107153088	–	T	0.703	0.199	0.914	0.938	0.679	0.013	9.26	0.219	9.76	0.143	6.15	0.106	10.43	0.006	11.59
25	rs10735825	12	LACE	50768339	FAM186A	C	0.895	0.08	0.58	0.537	0.732	0.183	9.44	0.210	9.97	0.021	6.17	0.001	10.43	0.048	11.64
26	rs730570	14	34plex	101142890	–	G	0.765	0.161	0.802	0.555	0.982	0.048	9.49	0.201	10.17	0.085	6.26	0.000	10.43	0.174	11.81
27	rs930072	5	AFR-EUR admix SNPs	36666071	SLC1A3	C	0.962	0.136	0.642	0.438	0.911	0.219	9.71	0.198	10.37	0.022	6.28	0.002	10.43	0.117	11.93
28	rs794672	6	LACE	95458317	–	G	0.057	0.917	0.432	0.741	0.357	0.246	9.96	0.197	10.57	0.020	6.30	0.017	10.45	0.025	11.95
29	rs634392	18	Genplex	70216045	–	T	0.797	0.066	0.696	0.102	0.089	0.127	10.08	0.195	10.76	0.090	6.39	0.071	10.52	0.074	12.03
30	rs4787040	16	LACE	7560980	RBFOX1	A	0.957	0.289	0.914	0.519	0.732	0.119	10.20	0.193	10.95	0.102	6.49	0.005	10.52	0.005	12.03
31	rs7307862	12	Genplex	112437514	–	C	0.576	0.899	0.142	0.426	0.839	0.000	10.20	0.172	11.13	0.209	6.70	0.010	10.53	0.046	12.08
32	rs862500	3	LACE	64272649	–	T	0.862	0.196	0.836	0.944	0.232	0.099	10.30	0.166	11.29	0.106	6.81	0.159	10.69	0.033	12.11
33	rs2503770	6	LACE	110266415	–	T	0.911	0.266	0.904	0.929	0.786	0.109	10.41	0.146	11.44	0.131	6.94	0.059	10.75	0.170	12.28
1	rs3827760	2	Genplex	109513601	EDAR	T	1	0.986	0.117	0.111	0.944	0.164	10.57	0.214	11.65	0.351	7.29	0.199	10.95	0.072	12.35
2	rs6437783	3	Genplex	108172817	MYH15	C	0.254	0.159	0.997	0.891	0.589	0.047	10.62	0.170	11.82	0.348	7.64	0.110	11.06	0.005	12.36
3	rs17822931	16	Genplex	48258198	ABCC11	C	0.997	0.854	0.096	0.615	0.875	0.180	10.80	0.077	11.90	0.346	7.99	0.000	11.06	0.060	12.42
4	rs1229984	4	Kiddlab	100239319	ADH1B	A	0	0.021	0.733	0	0.071	0.105	10.91	0.116	12.01	0.330	8.32	0.080	11.14	0.028	12.45
5	rs12594144	15	USC CCAS panel	64161351	–	C	1	0.855	0.101	0.778	0.393	0.207	11.11	0.102	12.12	0.290	8.61	0.112	11.25	0.000	12.45
6	rs4935501	10	USC CCAS panel	55935850	PCDH15	C	0.995	0.878	0.129	0.536	0.286	0.179	11.29	0.102	12.22	0.287	8.89	0.009	11.26	0.065	12.51
7	rs2180052	6	Novel	170589989	–	C	0.868	0.834	0.11	0.742	0.232	0.048	11.34	0.088	12.31	0.271	9.16	0.014	11.27	0.074	12.58
8	rs4918664	10	Kiddlab	94921065	–	A	0.989	0.874	0.131	0.141	0.821	0.175	11.52	0.106	12.41	0.271	9.44	0.146	11.42	0.027	12.61
9	rs366178	8	LACE	8771154	–	G	0.854	0.725	0.142	0.016	0.357	0.102	11.62	0.064	12.48	0.157	9.59	0.211	11.63	0.013	12.62
10	rs434504	1	DNAprint	4815477	AJAP1	G	0.168	0.2	0.902	0.481	0.268	0.063	11.68	0.081	12.56	0.256	9.85	0.007	11.64	0.016	12.64
11	rs8104441	19	Novel	51441807	–	T	0.83	0.833	0.129	0.875	0.179	0.048	11.73	0.084	12.64	0.253	10.10	0.056	11.69	0.104	12.74
12	rs1371048	2	Kersbergen	145753166	–	C	0.83	0.641	0.019	0	0.321	0.129	11.86	0.060	12.70	0.245	10.35	0.180	11.87	0.007	12.75
13	rs4657449	1	Novel	165465281	–	G	0.905	0.887	0.117	0.117	0	0.106	11.96	0.153	12.85	0.242	10.59	0.141	12.01	0.234	12.99
14	rs881929	16	34plex	31079371	ZNF668	G	0.954	0.6	0.082	0.603	0.732	0.187	12.15	0.008	12.86	0.232	10.82	0.025	12.04	0.024	13.01
15	rs10079352	5	LACE	117494640	–	G	0.038	0.451	0.972	0.984	0.429	0.266	12.42	0.018	12.88	0.229	11.05	0.166	12.20	0.010	13.02
16	rs203150	18	Novel	38037221	–	A	0.595	0.796	0.044	0.156	0	0.014	12.43	0.166	13.05	0.227	11.28	0.063	12.27	0.168	13.19
17	rs722869	14	Genplex	97277005	VRK1	C	0.895	0.901	0.168	0.259	0.429	0.078	12.51	0.137	13.18	0.224	11.50	0.102	12.37	0.019	13.21

Table 43 *(continued)*

18	rs1366220	5	LACE	153497780	–	C	0.941	0.678	0.077	0.055	0.589	0.187	12.70	0.044	13.23	0.218	11.72	0.154	12.52	0.004	13.21
19	rs499827	20	Genplex	3604447	ATRN	G	0.081	0.242	0.816	0.696	0.982	0.096	12.79	0.031	13.26	0.207	11.93	0.002	12.53	0.128	13.34
20	rs2065982	13	34plex	34864240	–	T	0.943	0.939	0.245	0.103	0.929	0.089	12.88	0.136	13.39	0.205	12.13	0.211	12.74	0.057	13.40
21	rs7246968	19	Novel	12651854	ZNF564	A	0.265	0.232	0.879	0.641	0.056	0.031	12.91	0.075	13.47	0.203	12.33	0.019	12.76	0.122	13.52
22	rs12435594	14	LACE	69555866	DCAF5	G	0.957	0.672	0.089	0.107	0.250	0.207	13.12	0.041	13.51	0.202	12.54	0.097	12.85	0.036	13.55
23	rs2851060	4	LACE	101953322	PPP3CA	C	0.097	0.224	0.871	0.946	0.411	0.119	13.24	0.076	13.59	0.200	12.74	0.181	13.04	0.001	13.55
24	rs17544484	13	Novel	88369434	–	C	0.968	0.775	0.145	0.25	0	0.179	13.42	0.067	13.65	0.195	12.93	0.059	13.09	0.225	13.78
25	rs3782973	13	Genplex	95102697	DCT	C	0.035	0.184	0.762	0.75	0.321	0.134	13.55	0.056	13.71	0.167	13.10	0.071	13.17	0.001	13.78
26	rs67302	16	DNAprint	65821799	–	T	0.859	0.511	0.066	0.219	0.411	0.163	13.71	0.011	13.72	0.164	13.26	0.027	13.19	0.000	13.78
27	rs868767	3	LACE	141377330	–	A	0.981	0.891	0.269	0.164	0.446	0.143	13.86	0.099	13.82	0.160	13.42	0.221	13.41	0.023	13.80
28	rs2833250	21	Genplex	32439270	–	T	0.392	0.046	0.699	0	0.589	0.004	13.86	0.152	13.97	0.156	13.58	0.122	13.54	0.038	13.84
29	rs1834619	2	Kiddlab	17901485	SMC6	G	1	0.929	0.304	0.047	0.286	0.158	14.02	0.122	14.09	0.152	13.73	0.274	13.81	0.082	13.92
30	rs7911953	10	LACE	61791039	ANK3	G	0.135	0.089	0.713	0.875	0.196	0.045	14.06	0.116	14.21	0.149	13.88	0.180	13.99	0.015	13.94
31	rs2585897	13	LACE	21398979	XPO4	G	0.87	0.829	0.206	0.047	0.089	0.083	14.15	0.101	14.31	0.146	14.03	0.204	14.19	0.150	14.09
1	rs433342	8	USC Tri-allic SNP	17747876	FGL1	AG	0.40/0.17	0.30/0.66	0.56/0.44	0.87/0.13	0.19/0.81	0.163	14.31	0.049	14.36	0.048	14.07	0.169	14.36	0.102	14.19
2	rs2069945	20	USC Tri-allic SNP	33761837	PROCR	CG	0.11/0.71	0.46/0.48	0.65/0.27	0.04/0.54	0.77/0.23	0.126	14.44	0.006	14.36	0.039	14.11	0.104	14.47	0.191	14.38
3	rs4540055	4	34plex	38803255	TLR1	AC	0.07/0.40	0.68/0.25	0.37/0.51	0.54/0.21	0.60/0.40	0.183	14.62	0.062	14.43	0.021	14.13	0.104	14.57	0.022	14.40
4	rs5030240	11	34plex	32424389	WT1	CA	0.39/.0.32	0.69/0.12	0.28/0.05	0.09/0.28	0.25/0	0.059	14.68	0.089	14.52	0.059	14.19	0.067	14.64	0.078	14.48
5	rs17287498	10	USC Tri-allic SNP	54530788	MBL2	GA	0.48/0.49	0.59/0.19	0.77/0.10	0.96/0.02	0.98/0.01	0.096	14.77	0.021	14.54	0.017	14.21	0.145	14.78	0.075	14.55
6	rs2184030	1	USC Tri-allic SNP	206667441	–	CG	0.59/0.19	0.48/0.51	0.58/0.41	0.48/0.52	0.72/0.02	0.072	14.85	0.024	14.56	0.022	14.23	0.042	14.83	0.156	14.71

Component SNPs are ranked in descending order of informativeness (defined by individual population-specific divergence = PSD) in population groups (PG): Oceanian (28); American (22); African (8); European (33) and East Asian (31). Six tri-allelic SNPs are listed at the end with allele frequencies of 1000 Genomes listed alleles. C: chromosome; RA: reference allele; CD: cumulative PSD (derived by addition of individual PSDs). SNP positions are from genome build 37.1 (GRCh37).

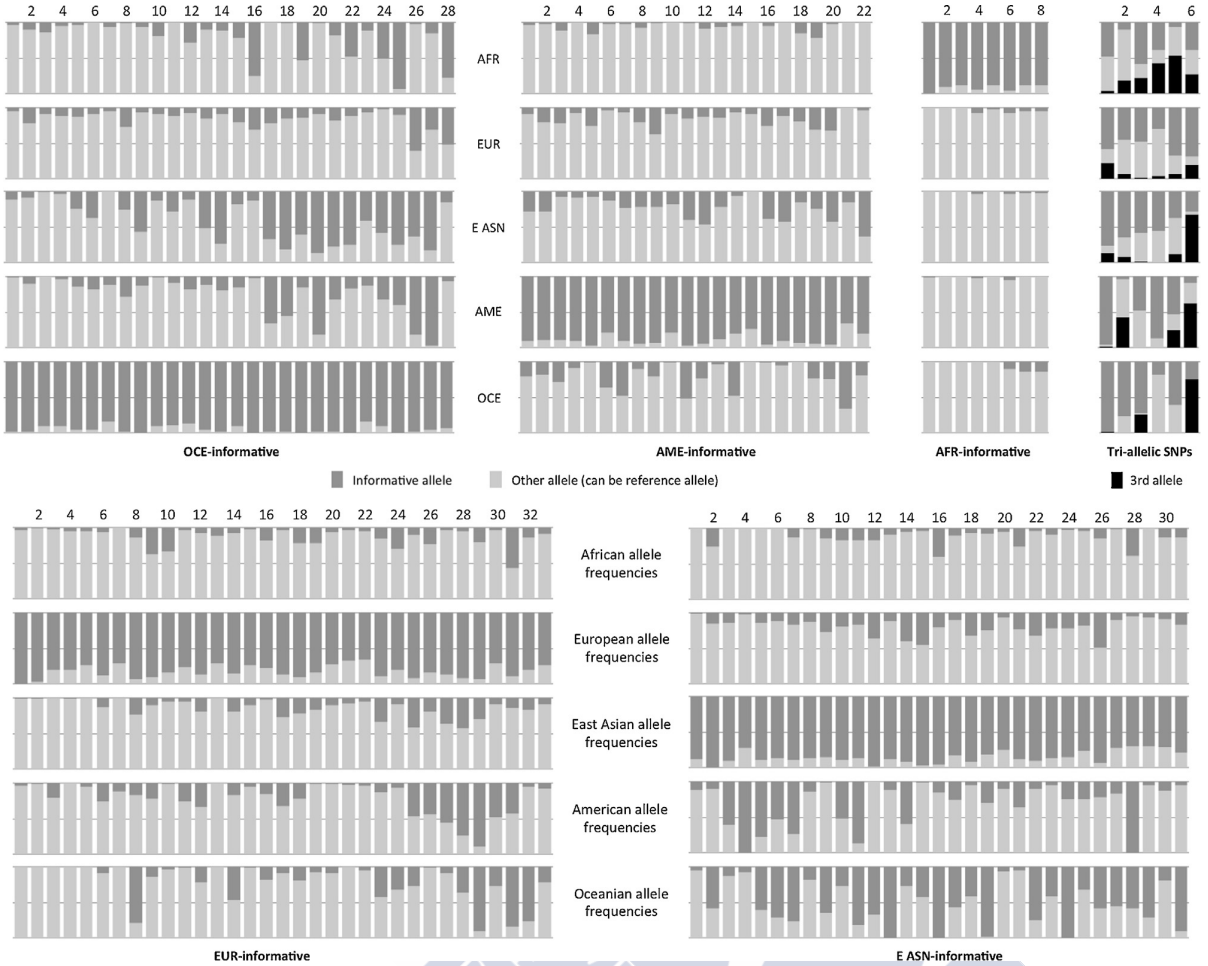


Figure 96: Allele frequency distributions of the Global AIM-SNP set arranged into the five groups they best differentiate plus tri-allelic SNPs. Column numbers identify individual SNPs ranked by population-group informativeness in Table 1. Mid-grey bars in tri-allelic SNPs denote reference allele, light grey and black bars 2nd and 3rd alleles respectively.

rs1426654; rs3827760; rs2789823; rs1871534; rs1369290; rs6875659; rs310644; rs9908046. Only the last three loci are part of the Illumina 650 K WGS panel often used for AIM-SNP selection.

Median inter-marker distance amongst syntenic Global AIM-SNPs was 14.8 Mb, with four pairs below average close physical linkage (including two transgressions of the 1 Mb filter). HapMap data indicated absence of correlated allele frequencies across 500 Kb distances (maximum possible in Haploview) within each span. Therefore, these SNP pairs lack allelic correlation and unless very recently admixed (e.g. parentally), correlations between ancestry-informative alleles in individuals with co-ancestry will erode rapidly (Gill et al. 2012).

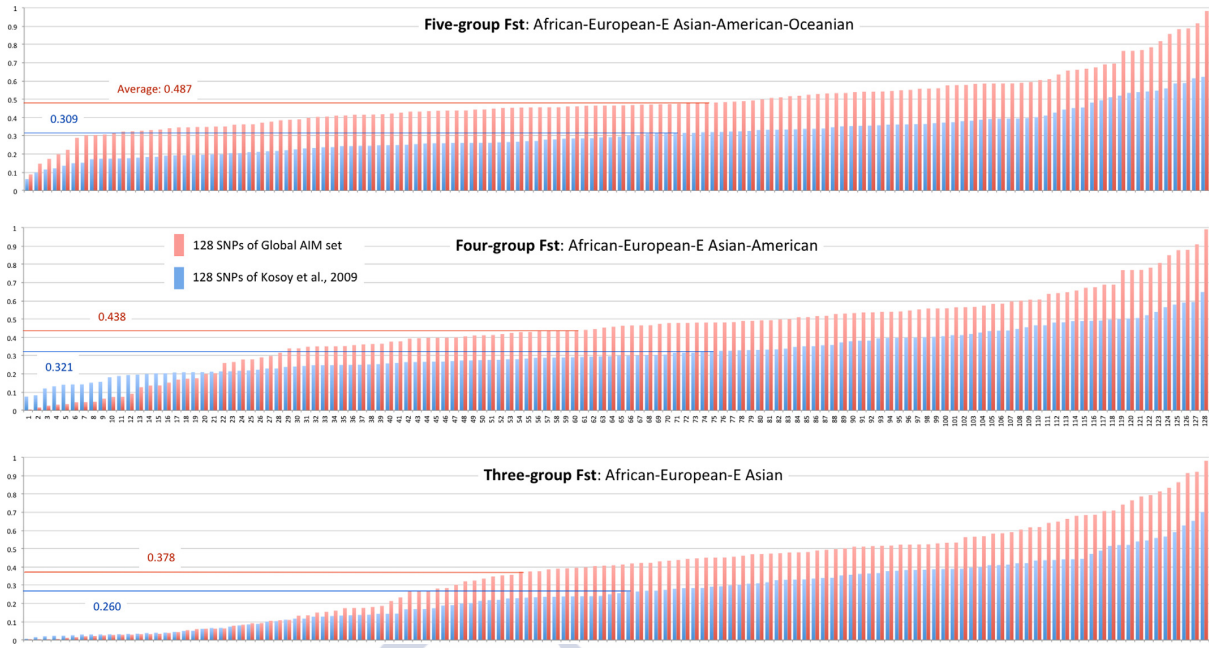


Figure 97: Ranked F_{ST} values for three, four and five group comparisons analysed with 128 Global AIM-SNPs or with the closest equivalent panel: 128 SNPs of the AFR-EUR-AME-informative ancestry panel of Kosoy et al. (2009).

5.2.3.2 Balancing population-specific Divergence levels and subsets of the full Global AIM-SNP set

The five sets of SNPs listed in Table 43 that focus on each global population group comprise: 33 EUR-informative; 31 E ASN-informative; 28 OCE-informative; 22 AME-informative and 8 AFR-informative markers, with a further six tri-allelic SNPs added. The 122-SNP set combined in these proportions reached cumulative PSD values of EUR=14.31, E ASN=14.03, OCE=14.09, AME=14.19, AFR=14.15 (a standard deviation of ±~0.1). This very balanced distribution of PSD values in each of the five groups is maintained adding the six tri-allelic markers, giving cumulative PSD values of EUR=14.56, E ASN=14.23, OCE=14.71, AME=14.82, AFR=14.84 ±~0.22). Therefore this SNP set has near-identical power to differentiate all five population groups, allows balanced analysis of each of the ten pairwise group comparisons and can be used to assess patterns of admixture without exaggerating the contribution of any one group due to an excess of informative markers for that group compared to the others. **Figure 98** charts the progression toward a point of PSD convergence, combining 122+6 component SNPs by carefully adjusted stepwise accumulation of divergence across five groups.

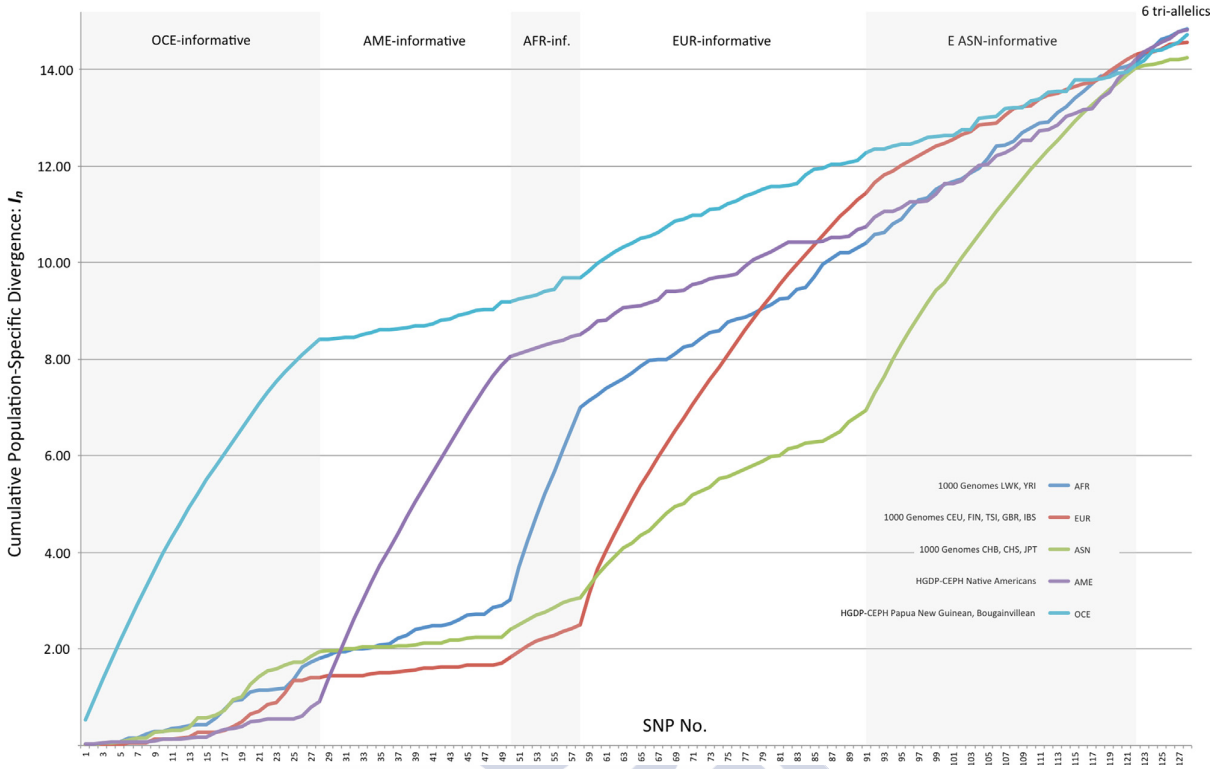


Figure 98: Convergence of five different cumulative population-specific Divergence (PSD) values in the 128 Global AIM-SNP set with components ordered by OCE-informative, best to worst differentiators, then: AME; AFR; EUR and E ASN-informative. 122 SNPs reach a PSD convergence value of ~14 then tri-allelic SNPs introduce some dispersion of PSD balance, notably in E ASNs.

Reaching a point of PSD convergence for 128 SNPs, we next sought to reduce component numbers to subsets of approximately 90, 60 and 30 SNPs or less, while preserving PSD balance. The medium-scale subsets consisted of 88 SNPs with PSD values of EUR=10.58, E ASN=10.52, OCE=10.47, AME=10.36, AFR=10.35 (±~0.09) and 55 SNPs with PSD values of EUR=6.99, E ASN=6.98, OCE=6.81, AME=7.07, AFR=7.12 (±~0.1). Construction of smaller scale subsets from the most divergent SNPs, examined how low the number of AIMs can be reduced while preserving both balance and classification success (using *Snipper* cross-validation). The three smallest subsets are summarized in **Figure 99**, which shows 12 SNPs reach a three group PSD of 3 and 100% classification success, extending this subset to 20 SNPs adds 100% classification success for Americans and extension to 28 provides reasonable PSD balance across four groups, converging on 4.7-5.0.

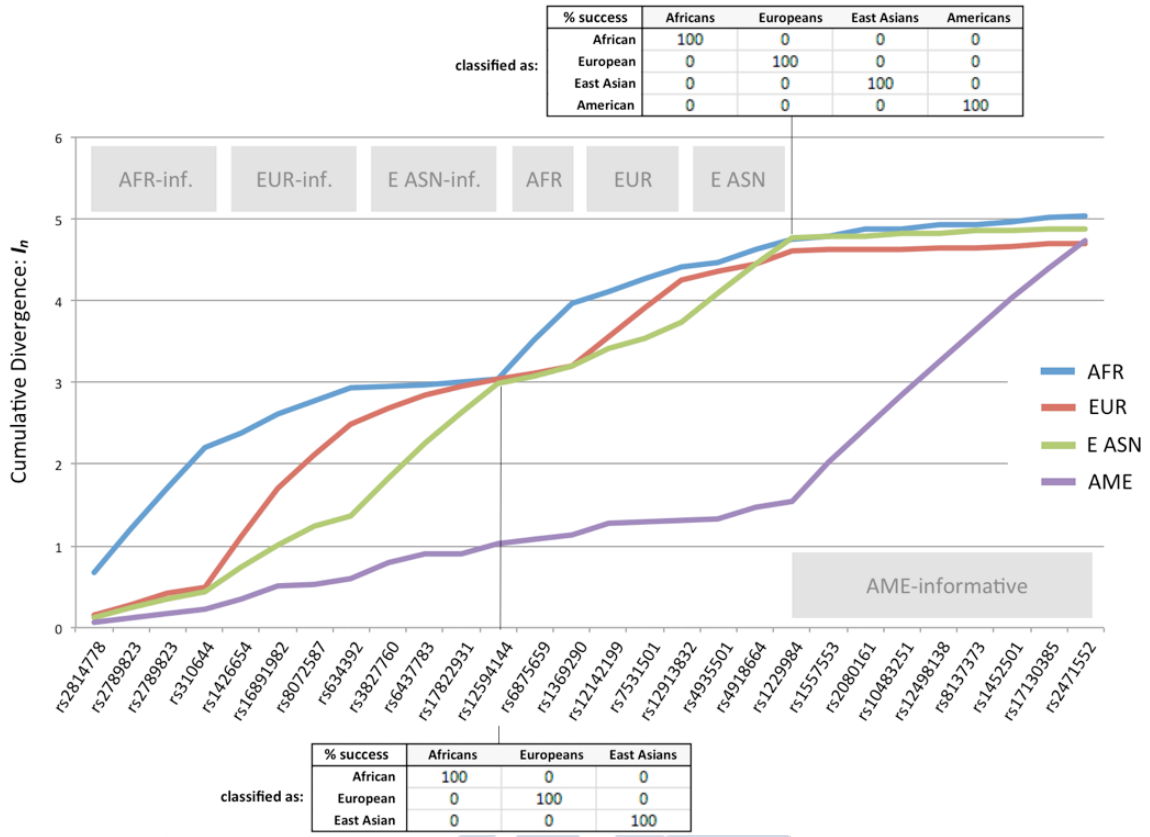


Figure 99: Convergence of AFR-EUR-E ASN cumulative population specific divergence (PSD) in reduced subsets of 12 and 20 best AIM-SNPs, and four group PSD convergence with a set of 28. Tables show classification success of 1st and 2nd set.

5.2.3.3 Analysis of admixed populations with the Global AIM-SNP set

PCA and STRUCTURE analyses of reference and admixed populations are summarized in **Figure 100**. The uppermost plot of **Figure 100A** shows *STRUCTURE* results for an optimum *K*=5 genetic clusters (indicated on the likelihood-of-*K* chart right): differentiating five reference groups and four admixed populations. Almost identical patterns are seen in the paired plots below for an optimum *K*=3 genetic clusters comparing the same admixed samples (but with three reference groups) using 128 Global SNPs and 350 LACE SNPs. While visual inspection of cluster plots suggests a close match between the co-ancestry patterns seen in ASW, CLM, MXL and PUR individuals for 128 vs. 350 AIM-SNPs, individual co-ancestry proportion estimates that underlie the cluster plots provide a more precise comparison. Global SNPs gave average proportions for ASW of: AFR= 3.7%-EUR=21.6%-AME=4.7%; CLM=9%-67.3%-23.7%; MXL=8.1%-46.7%-45.2%; PUR=12.1%-74.7%-13.2%; while LACE gave: ASW of: AFR=75.3%-EUR=20.8%-AME=3.9%; CLM=9.3%-66.8%-23.9%;

MXL=4.5%-49.0%-46.5%; PUR=12.9%-75.0%-12.1%. These comparisons show minimal differences in co-ancestry proportion estimates between each SNP set of about 1%, reaching a maximum 3.6% difference for the AFR component estimates in MXL. The very close match between SNP sets is also shown by the mean and quartile differences in individual co-ancestry proportion estimates listed and charted in Supplementary Table S4. The great majority of estimates differ by much less tan 5% between each SNP set, reaching a maximum 8% difference for the 75% quartile in MXL, AFR component estimates. Therefore the 128 SNPs we have collected provide a very close match, in terms of admixture detection capabilities, to a powerful AIM-SNP set almost three-fold larger and centred in its selection on three of the five population groups the Global set differentiates.

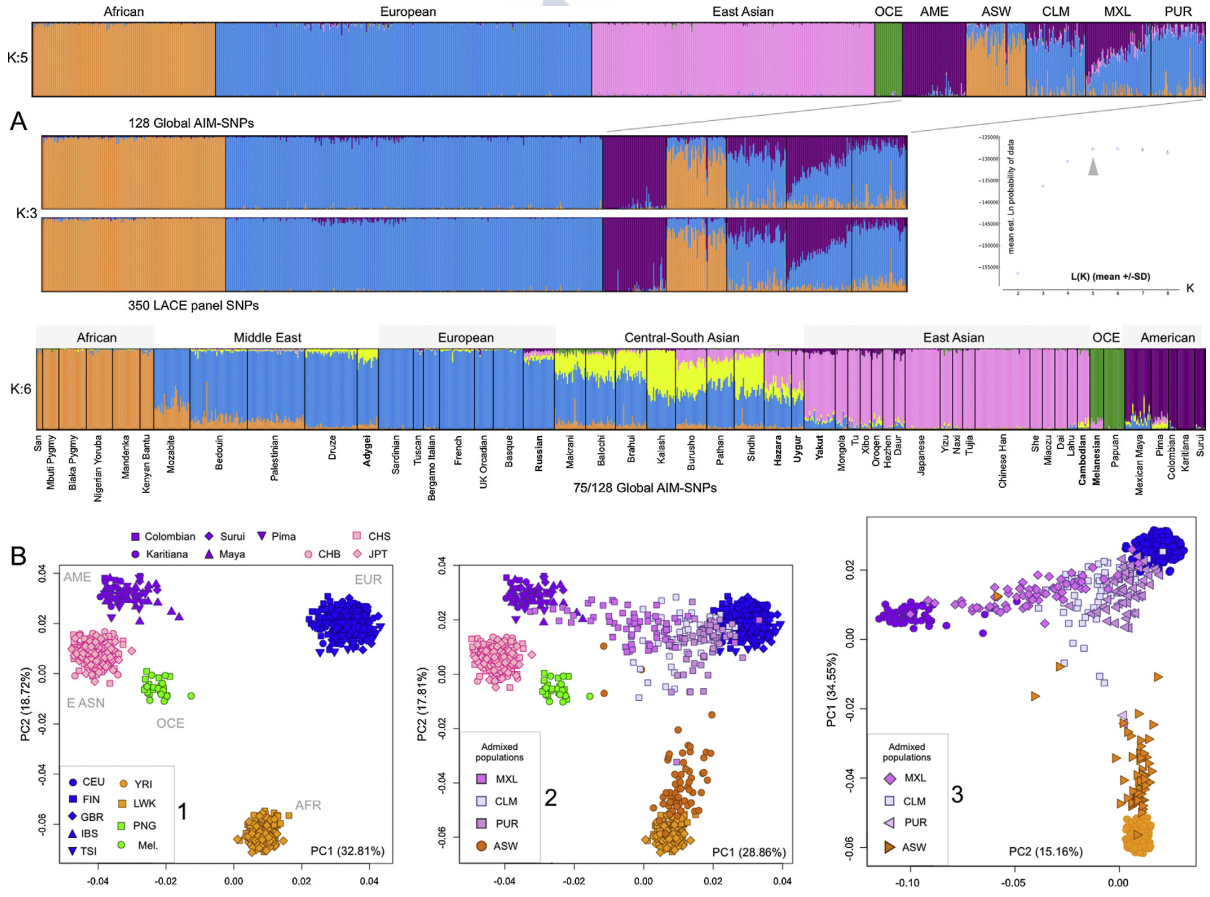


Figure 100: Population analyses of 1000 Genomes and HGDP-CEPH SNP data. (A) *STRUCTURE* cluster plots analysing, top to bottom: five reference and four admixed populations, optimum *K*=5 (indicated in L(*K*) chart right); three reference (AFR-EUR-AME) and admixed populations, optimum *K*=3 with 128 Global *versus* 350 LACE SNPs; all HGDP-CEPH populations, optimum *K*=6 with a subset of 75/128 Global SNPs. Bold population labels below indicate seven outlier populations that show strong similarities to cluster plot patterns obtained by Li et al. (Figure 1 of Li et al. (2008)) analysing the same samples at *K*=7 with 650,000 SNPs. The order of populations in each group is the same as Li's cluster plot to allow easy comparison. (B1) PCA plot with 128 SNPs analysing the same population combinations as the top *STRUCTURE* results. Colours matched to each reference group cluster. (B2) PCA adding four 1000 Genomes admixed populations. (B3) PCA of the second *STRUCTURE* plot population combinations, comprising 128 SNPs and three reference groups.

Figure 100B shows three PCA plots that link to the *STRUCTURE* cluster plots above. PCA with 128 Global SNPs of reference populations alone (**Figure 100B1**) estimated the first three principal components (PC) accounted for PC1=32.8%, PC2=18.7% and PC3=6.4% of the total variation. In the case of the middle and right-hand PCA plots (**Figure 100B2-3**), visual inspection is the only means to assess how well the co-ancestry analysis capabilities of both SNP sets match. Cluster patterns generally appear well matched with very similar outlier positions for three mid-cluster ASW and one PUR. These plots suggest the smaller 128 SNP Global set would give consistent positioning of admixed individuals on PC1-PC2 plots compared with the larger LACE set, in agreement with the findings from *STRUCTURE* analyses.

5.2.3.4 Assessment of within-population group heterogeneity of selected AIM-SNPs

Analysis of within-population, within-group and between-group SNP allele frequency heterogeneity was assessed with pairwise F_{ST} values and average pairwise genotype differences generated by Arlequin. Analyses compared the 1000 Genomes-CEPH populations used for AIM selection with additional CEPH AFR, EUR and E ASN populations or with four novel divergent populations, and are summarized in **Figure 101**. Additionally, *STRUCTURE* analyses were made of the complete HGDP-CEPH panel, restricted to the 75/128 SNPs in the Stanford data. The resulting plot of the optimum K=6 clusters is shown in the lower graphic of **Figure 100A**. Genetic cluster patterns for novel AFR, EUR and E ASN populations are mainly the same as their 1000 Genomes counterparts. The sixth, yellow cluster is confined to Central-South Asians and Caucasian Adygei (the most closely sited European population). However, Central-South Asian and Middle East populations do not produce distinctive clusters since Global AIM selection did not target differentiation of Europeans from these Eurasian sub-groups. A comparison of the *STRUCTURE* cluster plot with the comprehensive analysis of the same samples by Li et al. that detected seven clusters (Figure 1 in Li et al. (2008)), indicates some closely matching patterns for many populations, notably: Adygei, Russian, Hazara, Uygur, Yakut, Cambodian and Melanesian. Overall, despite the constraint of analysing only 60% of SNP data, *STRUCTURE* analysis of 30 new AFR, EUR, E ASN populations does not reveal significant allele frequency heterogeneity.

Similarly, Arlequin analysis shows low within-group divergences for the same groups (first, third and fourth large diagonal boxes, upper plot of **Figure 101**).

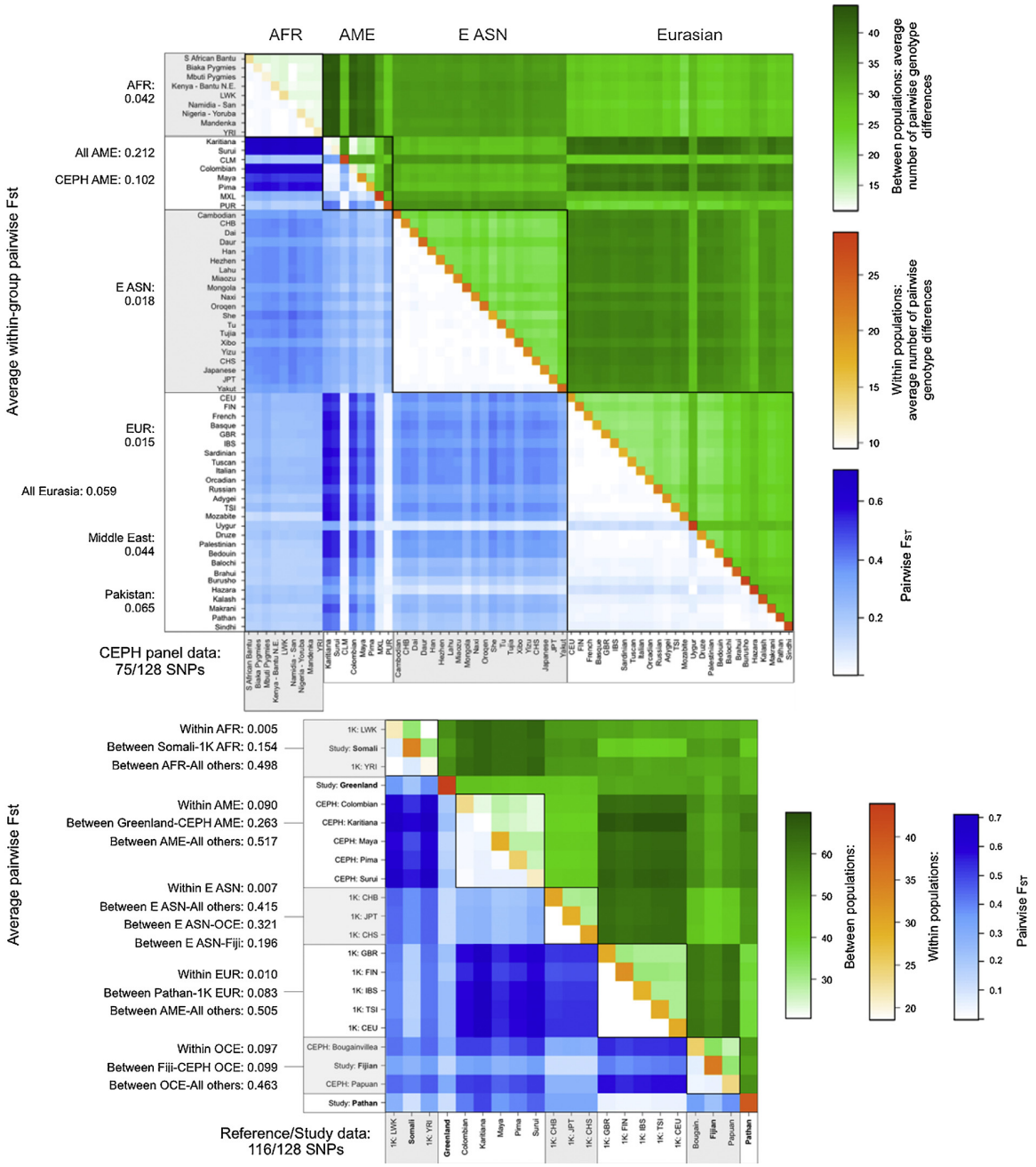


Figure 101: Arlequin population divergence analyses of pairwise F_{ST} (blue squares), pairwise between-population genotype differences (green) and pairwise within-population differences (orange). Upper chart plots comparisons of HGDP-CEPH populations with equivalent 1000 Genomes populations, restricted to 75/128 SNPs. Lower chart shows comparisons of four additional study populations likely to be divergent from 1000 Genomes/CEPH populations used for AIM selection. Upper chart values, listed left, show within-group F_{ST} averages (lower half of large diagonal boxes) indicating very low within-group divergence is maintained adding HGDP-CEPH populations to those used for selection. Raised F_{ST} is evident within the American group, particularly amongst admixed 1000 Genomes CLM, MXL, PUR. Lower chart F_{ST} values on the left indicate a uniformly high ratio of between-group to within-group divergence. However, divergence between Greenland and other AME populations is very high, while there is markedly reduced divergence between East Asians and Fijians – values likely reflecting different demographic histories between Arctic *versus* Meso-S American populations and Near *versus* Remote Oceanian populations respectively.

A lack of heterogeneity amongst the selected AIMs can be attributed to the loci as much as the populations we added to extended analyses, i.e. in SNPs at or near to fixation there is limited scope for drift to strongly influence very low frequency alleles. In contrast, a review of certain Arlequin pairwise comparisons in **Figure 101** reveals admixture or divergent population origins can create higher levels of heterogeneity that may affect how well the AIMs we selected differentiate all populations worldwide. 1000 Genomes CLM, MXL and PUR patterns indicate substantial divergence can occur from admixture. All three show less divergence from EUR populations than AME, as well as raised within-population heterogeneity.

Lastly, differences in the origins and demographics of populations placed in the same group due to geographic proximity can create the largest levels of within-group heterogeneity. Greenlanders produced distinctive patterns of pairwise F_{ST}, shown in the lower plot of **Figure 101**, suggesting they are different to all the populations we used for AIM selection. While EUR admixture may contribute some divergence, most of the heterogeneity seen is more likely to be due to the unique peopling of Greenland in two very recent migratory waves: from Siberia by Paleo-Eskimos 4500 years ago and from Beringia by Inuit 1000 years ago (Colonna et al. 2011) – in stark contrast to the separate, earlier colonization of the rest of America. Some heterogeneity is also seen in Fijians, not in increased divergence within OCE, but in much reduced divergence with E ASN. The observation of divergence between Near and Far Oceania is, again, likely to reflect differences in the origins of the migrant populations that first colonized two distinct parts of the same continent as well as recent migration from e.g. South Asians. While it can be argued that strong divergence within a group will compromise the ability of the AIMs selected to fully differentiate populations, when we construct *Snipper* training sets of both Greenlanders and Fijians, in place of AME and OCE data or in addition to the five groups, all samples are assigned correctly to their population of origin.

5.2.3.5 Discussion

This study shows the current extensive human genome variation catalogues can be easily accessed and their allele frequency data used to select highly differentiating ancestry informative SNPs. We were able to build sets with a range of sizes that meet the statistical

power demands of forensic analysis, while focusing on the key characteristic of population differentiation balance. Although prompted by the previous study of Galanter et al. (2012), that addressed AFR-EUR-AME populations, for all but the smallest subsets, we successfully kept five population group comparisons equally well differentiated and did not confine SNP informativeness to three or four global groups. Up until now, many AIM-SNP selections have ignored balanced differentiation of East Asians in favour of the closely related Native American group. Such approaches preclude unbiased and globally applicable analysis of a forensic sample of unknown geographic origin, despite East Asian admixture being rare in The Americas, and despite Oceanian ancestry representing a minor proportion of the worldwide demographic landscape outside of Australia, New Zealand and Hawaii.

The *SPSmart* genome browsers (Amigo et al. 2008b, Amigo et al. 2011) proved to be particularly valuable aids to building comprehensive forensic AIM-SNP sets in three respects. Firstly, they provide the only viable means to interrogate hundreds of markers with specific population properties in the same query, whereas 1000 Genomes, HapMap and HGDP-CEPH data gateways only allow SNPs to be screened locus-by-locus. Secondly, the option to identify groups of SNPs with identical allele frequency patterns, and therefore divergence (Costas et al. 2005), is straightforward in *SPSmart* by defining the chromosome segment to be queried, thus providing alternative markers for many first-strike SNPs that fail in multiplex or NGS analysis due to poor context sequence. Thirdly, *SPSmart* enables a very simple system for downloading genotype data from selected population groupings for porting into *Snipper* to assess patterns of divergence amongst markers and population pairings chosen.

With our study's primary focus on equilibrating the differentiation power evenly amongst five population groups, it is important to assess how well the Global AIM-SNP set can gauge admixture between them and whether forensic analysis has the right tools for the task. Our assessments of both of these aspects remain somewhat limited by the availability of data, being confined to four populations from The Americas. The imminent data release of 1000 Genomes Peruvians from Lima (PEL) may reveal East Asian admixture components, while the recent detailed study of American populations by Reich et al. also provides useful SNP data (Reich et al. 2012). Analysis of Fijians with 116/128 SNPs suggests that divergence between Near and Remote Oceania is detectable but small, although recent migration and integration of South Asians to Fiji is an additional factor that should be considered in any interpretation of ancestry patterns. Furthermore, the distinction between populations with

admixture and those on continental margins where absence of geographic barriers has led to increased gene-flow, is a complicating factor that requires caution and makes population analyses challenging when moving beyond unadmixed 'continental' populations such as the reference data used in this study. For example, it is not straightforward to distinguish allele frequency patterns found in North Africans from those found in admixed US African-Americans (see Figure 4 of Phillips (2013)). *STRUCTURE* provides the most commonly applied system for estimating co-ancestry components in admixed individuals by accurately defining the optimum number of genetic clusters in the data (Rosenberg et al. 2002, Borel et al. 2012, Johansen et al. 2013). However, there are widely held doubts about the validity of using mid-continental reference populations and not taking sufficient account of the clinal variation characteristic of so much of worldwide demographic structure (Serre and Paabo 2004, Rosenberg et al. 2005). A simpler and more easily interpreted approach to the complex task of distinguishing patterns of admixed SNP variation from those of unadmixed ancestral populations, is to superimpose individuals with unknown ancestry alongside reference population data on distance matrices generated by multi-dimensional scaling systems such as PCA. In forensic analysis it is beneficial to have a flexible system of ancestry inference that can easily handle single SNP profiles without using lengthy and PC-intensive calculations (typified by *STRUCTURE*). Therefore we have improved forensic ancestry analysis using real-time tools in two ways. Firstly, by adding a co-ancestry component estimator in *Snipper* based on the ratio of each likelihood to the sum of all likelihoods (the number of ancestries defined by the training set data uploaded). Secondly, by developing scripts that position de novo AIM-SNP profiles on top of reference data PCA plots (such as **Figure 100B**) using the first and second principal component percentages to define the profile's plot axes (http://mathgene.usc.es/snipper/analysismultipleprofiles.html). In this way, visualization of complex genetic data in simple 2D space can bring a more intuitive approach to ancestry inference, aided by this study's advances in assembling some of the most informative AIM-SNPs for forensic use.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA



FINAL DISCUSSION

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

6 Final discussion

As described in the Introduction (section 1), approximately 95% of the human genetic variation is found within populations (Rosenberg et al. 2002), with individuals from each population being more different between them than populations between themselves. STR markers, commonly used in human identification applications, are not ideal for ancestry prediction as they are highly variable between individuals and have low inter-population variation. Nevertheless, there are loci with population specific alleles (Phillips et al. 2008b) that can offer some degree of population differentiation. On the other hand, binary loci with large allele frequency differences between populations have been proposed as an alternative to estimate individual ancestry membership proportions, allowing the prediction of probable biogeographic origin. Autosomal markers have emerged as amongst the best ancestry markers due to their stability, density of distribution and full range of allele frequency patterns amongst populations. Autosomal markers provide valuable complementary data to Y-chr and mtDNA typing for the analysis of the evolutionary history of populations. They have the advantage that small numbers of samples are sufficient to give reliable allele frequencies estimates – with as few as 30-40 samples providing adequate population allele frequency estimates (Phillips 2015) – and also that they are independent of admixture sex bias allowing a more precise inference of the genetic ancestry of an individual.

The development of assays using such markers is of particular interest in forensic genetics as they can provide investigative leads from DNA evidences in cases where the source individual is not known or is unable to declare its ancestry. However, it is important to understand the current distribution of human diversity. Studies using several hundred markers suggested that the world's population can be grouped into seven major discrete ancestral groups (Rosenberg et al. 2002, Enoch et al. 2006, Allocco et al. 2007, Li et al. 2008), with groups based on genetic similarity closely corresponding to continental origin. However, this pattern can be highly dependent on the sampling scheme because the majority of human worldwide genetic diversity actually takes the form of geographic clines rather than clades (Serre and Paabo 2004, Manica et al. 2005, Rosenberg et al. 2005). For that reason, to detect

differences between major continental groups, it is essential to find markers that show the most pronounced allele frequency discontinuities between continental regions (Rosenberg et al. 2002). Such markers can be included in small-scale marker sets with population specific genotypes. This implies that a compromise between the amount of information that can be obtained from such small-scale multiplexes and the simplification of the description of the complex human population structure has to be achieved (Phillips 2015). For forensic ancestry analysis, a five-group differentiation (Africa *versus* Europe *versus* East Asia *versus* Oceania *versus* Native America) is a reasonable objective using small-scale marker sets selected to have strong allele frequency differentiation (Phillips 2015). However, for further analysis and differentiation of populations located in the middle of variability continua, for example Middle East or Central South Asia, which are more likely to show higher misclassification rates, the use of more extensive marker sets that include loci with sharp frequency clines is required (Phillips et al. 2007b).

The five AIM assays described in this thesis (section 4) were developed at different times with complementary objectives. First of all, the *34-plex* assay was originally developed to differentiate three population groups: Africa, Europe and East Asia (Phillips et al. 2007b). The revised version presented in this thesis targets the same population groups but the swapping of a component SNP (rs727811 was replaced by rs3827760) improved the differentiation of East Asian populations. However, this test lacks sufficient power to differentiate Oceanian and Native American populations – this prompt the development of additional ancestry assays complementary to the *34-plex*. One of such assays is *Pacifiplex* – it successfully enhances *34-plex* by enabling the differentiation of East Asian and Oceanian populations. Their joint analysis improves the differentiation of the ancestry patterns observed in Oceanian populations (such as Aboriginal Australians and Papua New Guineans) from those of populations more closely related to East Asia (Taiwan and Philippines). Other of such assays is the set of *46 AIM-Indels* – it was developed to provide a simple tool for inferring ancestry from four different population origins that can be widely applied to genetic studies (African European, East Asian and Native American). Furthermore, although this assay was primarily designed for studies considering only four major population groups, extension to six groups revealed the capacity to reliably distinguish Oceanians and, to a certain extent, to detect a Central South Asian component. Although this assay was not designed to complement *34-plex*, the joint analysis of both assays reduces the classification

rate when classifying Eurasian populations (Europe, Middle East and Central South Asia). Nevertheless, other AIMs should be combined in order to improve the classification of more closely related populations. *Eurasiaplex* is an AIM-SNP assay that, when analysed together with *34-plex*, differentiates the bulk of European and South Asian populations with high levels of statistical likelihood. However, differentiation of Middle Eastern populations shows high classification error, highlighting the difficulty of using forensic-scale AIM tests to analyse populations in closely sited regions within a continuous genetic variation gradient – as with the cline characterizing marker frequencies across Eurasia. Since most European and South Asian populations are sited at the cline's extremes they are reliably separated whereas populations closer together and occupying middle regions reduce the overall reliability of classification. Clearly, differentiation of Middle East populations from Europeans and South Asians either side of the cline requires more SNPs than described for the *Eurasiaplex* test. In fact, to predict individual ancestry of closely related populations (belonging to the same major ancestral group) an increased number of markers is required, ranging from several hundred to thousands (Biswas et al. 2009, Paschou et al. 2010). Moreover, marker selection highly influences an assay's ability to differentiate certain population groups (especially those more closely related) with populations not considered during marker selection being more prone to classification errors.

All of the described AIM assays were designed and optimized to be applied in forensic casework – the markers are amplified in short amplicons (ensuring the assays are applicable to highly degraded DNA) and have suitable sensitivity amplifying less than 0.5 ng (an advantage when analysing scant samples usually found in forensic cases). However, SNP analysis major disadvantage is the typing methodology usually used – SNaPshot is a multi-step process where the samples are subject to multiple tube tests, increasing the risk of contamination and transfer error. Furthermore, SNaPshot fluorochromes do not generate equal signal strength with 4:2:1:1 average peak height ratios for the G:A:C:T nucleotides, further increasing the difficulty in comparing peak heights. This problem is exacerbated when in presence of a mixed-source sample – mixture detection and deconvolution presents interpretative challenges when analysing biallelic markers. Although the inclusion of tri-allelic markers increases the probability of mixture detection (Phillips et al. 2004b), this problem can be better addressed by the described AIM-Indel assay as it brings together the benefits of biallelic markers to the straightforward PCR-to-CE workflow, reducing the steps

and resources needed to genotype a large set of biallelic AIMs and allowing the comparison of allele peak height ratios. However, mixture detection and deconvolution can be achieved when using STR markers. The tetranucleotide *12-plex* AIM-STR forensic ancestry assay described in section 4, although less informative than binary marker assays, provides the best alternative to detect and analyse mixed DNA samples. The combined analysis of these AIM-STRs with core identification STRs and *34-plex* AIM-SNPs creates a powerful set of markers – combining 66 markers of different characteristics improves differentiation of Eurasian population groups, although Middle East populations remain very difficult to differentiate without recourse to much larger marker sets. Furthermore, the *12-plex* AIM-STRs also give comparable levels of discrimination to existing forensic markers for routine identification purposes, being suitable for a range of forensic analyses in addition to ancestry inference.

Once AIM genotypes are obtained from a DNA sample, it is necessary to estimate the probable origin and ancestry membership proportions. Several methods exist, each with a set of advantages and disadvantages, being applicable in different situations. A possible approach starts with a set of predefined populations and then classifies individuals of unknown origin into these populations. Allele frequencies estimates in each source population at a series of unlinked loci can be used to compute a set of likelihoods that a given profile of genotypes originates in each population. These likelihoods allow the assignment of individuals of unknown origin to populations based on the highest likelihood ratio (Pritchard et al. 2000).

In the present thesis, a set of tools applicable to the inference of ancestry in forensic genetics was described (section 3). Principal component analysis is a commonly used exploratory tool that summarizes the genetic similarities and differences between groups of populations or individuals. The described *R* script can be used to represent the relation of a casework sample with a set of reference population samples helping to infer, through visual inspection, the most probable classification of that individual. Considering the computational speed of principal component analysis, it can be used as a preamble to one or both of the two Bayesian methods considered in this thesis: *STRUCTURE* and *Snipper*.

STRUCTURE output files include information on the estimated clusters, i.e. the population groups generated not the input populations – when the populations are defined in a way that they closely match the calculated clusters, the inferences of the population ancestry membership proportions in each of the pre-defined clusters can be considered to be the proportions of the input populations. When attempting to classify a population or individual,

the use of reference populations closely matching the inferred clusters is important, especially when analysing admixed samples where it is important to define the contributing parental populations. *STRUCTURE* gives information on the training set (allowing the assessment of the used reference data set – the optimum *K* value matches the number of reference populations, which are completely differentiated among them) and it also gives us the individual ancestry membership proportions. The population and individual ancestry membership proportions can be represented in two distinct types of plot – a triangle plot where each individual is represented as a coloured point and a bar plot where each individual is represented by a vertical line divided into *K* coloured segments proportional to the estimated membership into each of the *K* inferred cluster. The latter is particularly useful when visualizing data for $K>3$.

Snipper output includes a variety of information that can be useful both during the assay design stage (such as the apparent success of the classification and a list of the markers in descending order of divergence, calculated based on the number of populations included in the comparison) and the classification of a casework sample (likelihood ratios in verbose format and predicted admixture components and ancestry). Pairwise likelihood ratio plots such as the one generated for the 11-M Madrid bomb attack (Phillips et al. 2009) (section 1.3.2.3, exemplified in **Figure 21**) can be used to gauge patterns of possible admixture and to perform a cross-validation of the used training sets in order to assess accuracy and performance. In this analysis it is important to define outlying probability values, that is, likelihood values under which a sample cannot be reliably assigned to a certain population group. *STRUCTURE* analysis of the same samples can be used to detect admixed individuals (that increase the misclassification risk), helping to define such thresholds.

Two recent studies (Phillips 2015, Santos et al. 2015a) have shown that the three data analysis systems described here provide closely matched results between *STRUCTURE* patterns, *Snipper* assignment probabilities and the positioning of unknown samples in comparison with reference populations in PCA plots.

Two approaches to ancestry inference or to estimate co-ancestry proportions in admixed individuals have been previously proposed: i) small marker sets targeting the highest possible divergence (specifically, δ) values during SNP selection (Shriver et al. 1997, Rosenberg et al. 2003); ii) balance divergence amongst the target population groups to differentiate each group with equal power as a way to reduce estimation bias of the ancestry proportions, particularly

in admixed individuals (Phillips 2015). Although this is not always easy to achieve, aiming for balanced divergence allows the setup of the most informative marker sets.

This fact, together with the increased multiplexing scales offered by next generation sequencing, prompt the development of the Global AIM-SNP set that includes 128 markers balanced for a five-group comparison. Although a limited number of population samples were analysed (four admixed populations from 1000 Genomes), it was possible to assess how well the Global AIM-SNPs can gauge admixture between the five target population groups. Furthermore, analysis of Fijians with 116/128 SNPs suggests that divergence between Near and Remote Oceania is detectable but small, although recent migration and integration of South Asians to Fiji is an additional factor that should be considered in any interpretation of ancestry patterns. Another objective of the Global AIM-SNP study was to assess whether forensic analysis has the right tools for the task. In forensic analysis it is beneficial to have an alternative and flexible system of ancestry inference that can easily handle single profiles without using lengthy and intensive calculations such as the ones performed by *STRUCTURE*. Therefore we have improved forensic ancestry analysis using real-time tools in two ways. Firstly, by adding the co-ancestry component estimator in *Snipper* based on the ratio of each likelihood to the sum of all likelihoods (the number of ancestries defined by the training set data uploaded). Secondly, by developing the *R* script that places *de novo* binary AIMs profiles on top of reference data PCA plots using the first and second principal component percentages to define the profile's plot axes. This script has been implemented in *Snipper*.

However, when considering the application of next generation sequencing in forensic analysis, two questions have to be considered – can we still use the existing forensic AIM assays? How good is the technology?

The first question was addressed in the study presented in section 5 (Appendix 3A). Based on this preliminary study, the Ion PGM™ system has demonstrated considerable potential for use in forensic DNA analysis as a low to medium throughput NGS platform. Using Ion 314 v1 chips, three times more data was obtained than expected and without the need to optimise or further adapt PCR multiplexes from previously established SNaPshot assays to produce the amplicons for sequencing. The average coverage of 500× per SNP amplicon was sufficient to generate a genotyping accuracy of 97% of 136 unique SNPs assessed in this study for three samples at three input amounts down to 0.5 ng of template DNA. The only SNP confirmed to be mistyped by the Ion Torrent was rs1029047 with known

homopolymeric stretches immediately flanking either side of the SNP (Seo et al. 2013, Børsting et al. 2014). The ability to provide accurate ancestry and EVC intelligence to forensic investigators requires the analysis of more SNPs than can be genotyped from a single SNaPshot assay. A single NGS assay containing hundreds of SNPs can maximise use of scant evidential material as well as reducing both sample processing times and potential contamination risks.

A more formal assessment of the parameters and thresholds was performed in the study presented in section 5 (Appendixes 3B-E). The evaluation of Ion PGM™ sensitivity and genotyping accuracy made in this study gives strong support for the application of NGS technology to forensic DNA analysis. Sequence data had sufficiently high coverage and gave reliable SNP genotyping for most loci in the HID SNP panel. Five SNPs with discordances should be excluded from the panel – two (rs1004357 and rs2032597) were already removed from the revised version of HID SNP panel, while rs2399332 was identified in Børsting et al. (2014) study as a problematic SNP. Lastly, mention should be made of rs1029047 that, as in the Appendix 3A study, gives genotyping inconsistencies in all NGS studies of this SNP made so far (Seo et al. 2013, Børsting et al. 2014). Careful scrutiny of sequence characteristics is required of any SNP chosen for forensic use, particularly coding SNPs in forensic phenotype predictive tests since these must work well for the SNP analyses to be sufficiently informative. Furthermore, there is some consensus that ~15-20× minimum coverage thresholds can safeguard the reliability of allele calls made with NGS (Bentley et al. 2008, Quail et al. 2012, Daniel et al. 2015). The conclusions of this study agree with the ones of Børsting et al. (2014) – the Ion PGM™ analysis software needs further optimisation to be fully suitable for forensic application, although it is being constantly revised to this end.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

CONCLUSIONS

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

7 Conclusions

7.1 ABOUT THE DATA ANALYSIS SYSTEMS

1. The use of Bayesian methods (such as *STRUCTURE* and *Snipper*) offers several advantages, especially assigning admixed individuals to population clusters, since it is possible to use prior information to assist the calculation of ancestry proportions for these individuals. Therefore, information on reference data, the markers applied and the type of analysis desired is relevant before the selection of the analysis parameters.

2. Principal component analysis (PCA) is an alternative to Bayesian methods – it can be used as a preamble to both *STRUCTURE* and *Snipper*, allowing a fast recognition of clustering patterns between individuals. The described *R* script can be used to represent a casework sample in a PCA plot generated for the reference population helping to infer the most probable classification of that individual.

3. The best approach in forensic ancestry analysis includes the integration, in a single analysis, of the different techniques described in this thesis – closely matched results between *STRUCTURE* patterns, *Snipper* assignment probabilities and the positioning of unknown samples in comparison with reference populations in PCA plots.

7.2 ABOUT THE NEW ANCESTRY INFORMATIVE ASSAYS

7.2.1 Revision of the SNP*for*ID *34-plex* forensic ancestry test

1. Swapping a single component SNP at the same time as making adjustments to the chemistry and certain mobility positions in the *34-plex* assay provided relatively little disruption to routine forensic SNP typing with this ancestry analysis test. As East

Asians had the weakest population differentiation using the original *34-plex* test, the introduction of a near-fixed SNP in this population group represented a major enhancement of the test without upsetting its forensic performance.

2. The Hispanics that were studied as part for the NIST panel highlight the difficulties of analysing individuals with complex, potentially three-way, patterns of admixture. Not only it is difficult to make accurate assessments of co-ancestry when these are so varied within the group as a whole, but also component AIM-SNPs in small forensic multiplexes such as the *34-plex* test cannot provide equally informative differentiations of all co-ancestry contributors. With this in mind we have continued development of additional AIM-SNP based ancestry tests, complimentary to the *34-plex*.

7.2.2 Pacifiplex: a 29 AIM-SNPs assay dedicated to the Pacific region

1. The *Pacifiplex* assay developed in this study fulfils the three goals that were set for the 29 SNP panel: it successfully enhances an existing *34-plex* AIM-SNP panel by enabling the differentiation of East Asian and Oceanian populations; it provides an informative complementary test to Y and mtDNA analyses when admixture may be a dominant characteristic of the populations studied; and it has proven sensitivity when analysing challenging DNA samples.

2. *Pacifiplex* was able to detect as little as 125 pg of DNA with a high level of genotype completeness, providing a key characteristic for forensic analysis of trace quantities of DNA. The high sensitivity and robustness of *Pacifiplex* and *34-plex* assays was equally important for the population studies as it allowed us to expand the sample coverage of Oceania by including the GWU population collection of 50-year-old serum samples. In all cases except a proportion of WA Aboriginal Australian samples excluded from this study, *Pacifiplex* and *34-plex* assays were able to successfully genotype the scant DNA we extracted from the serum samples containing very low levels of nucleated cells.

3. The combined 62 AIM-SNPs are able to differentiate ancestry patterns observed in Oceanian populations (such as Aboriginal Australians and Papua New Guineans) from those of populations more closely related to East Asia (Taiwan and Philippines). Enhanced differentiation of Micronesia and Polynesia will be gained when the 62 AIM-SNPs are combined with uniparental data.

4. Although numerous AIM sets designed to differentiate global populations have been described, the development of *Pacifiplex* addresses a gap in forensically optimised AIM sets relevant for casework needs in the Australasian and Pacific regions.

7.2.3 A *46-plex* AIM-Indel assay and the completion of the HGDP-CEPH panel of samples

1. A set of 46 AIM-Indels were multiplexed in a simple and informative assay, enabling a more straightforward alternative to the commonly available AIM-SNP typing methods dependent on multi-step protocols and/or implementation of dedicated genotyping technologies.

2. The AIM-Indel assay produces accurate individual ancestry estimates of four different origins – Africa, Europe, East Asia and Native America. However, analysis can be extended to five continental groups while reliably distinguishing Oceanians. Given the relatively high efficiency in population assignment of individuals from all five continental origins, the multiplex represents a tool of considerable potential in forensic applications.

3. The completion of the genotyping of the HGDP-CEPH diversity panel for a set of 46 AIM-Indels allows a broader geographic scope for combined analyses with equivalent marker sets such as the *34-plex* AIM-SNPs.

4. Ancestry analysis of the complete HGDP-CEPH H952 panel indicates the 46 AIM-Indels can differentiate six major population groups but with noticeable classification

error in EUR and ME. It is important to stress that the differentiation of EUR from CSA and ME populations was not a target for the original marker selection.

5. All genotypes and frequency estimates are publicly available and can be used as reference data in forensic applications such as inference of ancestry of casework samples. The genotype data is now compiled in a dedicated web browser for forensic Indels named *forInDel* (http://spsmart.cesga.es/search.php?dataSet=forindel46) and these genotypes or major population indices can be downloaded.

7.2.4 Eurasiaplex: a 23 AIM-SNP assay for Europe *versus* South Asia differentiation

1. The studies of Eurasian population variability indicate that the SNPs of *Eurasiaplex* and *34-plex* differentiate the bulk of European and South Asian populations to high levels of statistical likelihood. Classification success remains high when extending analyses west from India to Afghanistan and Pakistan but moving east from Europe toward Turkey and South Caucasus regions bordering Iran produces considerable classification error, while differentiation of Middle Eastern populations shows even higher error rates. These observations highlight the difficulty of using forensic-scale SNP tests to analyse populations in closely sited regions within a continuous genetic variation gradient – as with the cline characterizing SNP frequencies across Eurasia.

2. Since most European and South Asian populations are sited at the cline's extremes they are reliably separated using *Snipper*, whereas populations closer together and occupying middle regions reduce the overall reliability of this Bayesian classifier.

3. Two approaches can help interpret *Snipper* results from closely related populations in an allele frequency continuum: adopting threshold values balancing classification success with minimal error rates and combining *Snipper* with complimentary analyses including *STRUCTURE* and PCA that assess similarity patterns amongst populations – helping to highlight the dangers of very close relationships. In this study we advocated a discrimination threshold of '100 times more likely' to balance *Snipper*'s performance differentiating Europeans and South Asians.

7.2.5 A 12-plex AIM-STR assay

1. The results obtained indicate the reduction in population informativeness when using tetranucleotide loci creates less powerful microsatellite-based sets for differentiating populations. However, as binary SNPs developed in parallel to STRs for forensic ancestry inference cannot detect and analyse mixed DNA samples, tetranucleotide AIM-STRs provide the best alternative in such commonly encountered situations.

2. Although existing ID-STRs would not be expected to differentiate populations well, they do have this characteristic to some degree. Since core forensic ID-STRs are always typed in routine profiling, a possible strategy is to find the most ancestry informative tetranucleotide STRs then combine these with core STRs with the option to add AIM-SNPs from a single assay creates a powerful final set of markers showing complimentary characteristics.

3. Analysis of the resulting mixed marker data is possible with both *STRUCTURE* and the enhanced *Snipper* tool, adapted to handle multiple allele loci. Although some reservations about use of cluster membership proportions to differentiate closely related populations exist, this studies reveals that Europeans and Central South Asians can be analysed in this way. Combining 66 markers of different characteristics improves differentiation of these Eurasian population groups compared to previous marker sets, although Middle East populations remain very difficult to differentiate without recourse to much larger scale, mainly SNP-based multiplexes.

4. The STRs outlined here offer considerable additional scope to improve mixture analysis. As the AIM-STRs also give comparable levels of discrimination to existing forensic markers for routine identification purposes, the novel *12-plex* assay developed could be suitable for a range of forensic analyses in addition to ancestry inference.

7.3 ABOUT ANCESTRY INFERENCE AND NEXT GENERATION SEQUENCING

7.3.1 Use of already existing AIM-SNP assays in NGS

1. The Ion Torrent™ has successfully genotyped at least 97% of 136 unique SNPs for three samples at three input amounts down to 0.5 ng of template DNA. The only SNP confirmed to be mistyped by the Ion Torrent was rs1029047 with known homopolymeric stretches immediately flanking either side of the SNP.

2. A single NGS assay containing hundreds of SNPs can maximise use of scant evidential material as well as reducing both sample processing times and potential contamination risks. Using five established SNaPshot assays in five PCR multiplexes, reproducible genotypes were obtained from 0.1 to 0.3 ng of DNA for each, equating to 0.5-1.5 ng of DNA, making it a viable alternative strategy. This approach enables users with significant investment in SNaPshot assays to make immediate use of NGS technology for SNP genotyping and allows flexibility with respect to the selection of specific assays to cater to the unique needs of an investigation.

3. This study represents the very low end of the scale with respect to NGS throughput. Sequencing capacity increases with other Ion PGM™ chips and platforms with the opportunity to scale up 1000-fold from the levels of this study, in the number of samples, the number of SNPs, or both. NGS systems can utilise combinations of current forensic assays and accommodate new phenotypically predictive SNPs that are continually being discovered.

7.3.2 Evaluation of the HID-Ion AmpliSeq™ Identity panel with the Ion PGM™

1. The evaluation of Ion PGM™ sensitivity and genotyping accuracy made here, give strong support for the application of NGS technology to forensic DNA analysis. Sequence data obtained in all three laboratories had sufficiently high coverage and

gave reliable SNP genotyping for most loci in HID SNP. We discovered five SNPs with discordances that should be excluded from the panel.

2. The estimation of optimum sample numbers for each of the six Ion PGM™ chip versions, presented this study with the biggest challenge, both in harmonising NGS runs across three laboratories and ensuring the coverage obtained was appropriate for assessing forensic sensitivity. Since low-level DNA appears to accentuate coverage variability in HID SNP markers, this will be a major problem when initially optimising NGS for routine forensic use. As Ion PGM™ chip capacities have now reached very reasonable levels of sequence output, users can be cautious by loading fewer samples than coverage estimation guidelines suggest. Furthermore, there is some consensus that ~15-20× minimum coverage thresholds can safeguard the reliability of allele calls made with NGS.

3. Although the Torrent Suite™ software provides several sequence quality parameters in the data output, we found there was little or no scope for changing the default analysis parameters settings to more aggressive thresholds. Setting such thresholds would provide a way to exclude miscalled genotypes from underperforming SNPs or mixed DNA. This finding has consequences for the average forensic scientist's capacity to properly scrutinise the extensive data that Ion PGM™ produces. Since mixture detection with binary markers is severely restricted compared to multi-allele STRs, it is all the more important to properly assess deviations from balanced heterozygote patterns.

4. We largely agree with the conclusions of Børsting et al. (2014) that the Ion PGM™ analysis software needs further optimisation to be fully suitable for forensic application, although it is being constantly revised to this end. In particular, there is an evident need to apply *Somatic* analysis parameter settings to properly analyse mixtures, even though *Germline* analysis parameter settings are set in place for forensic SNP analysis with Ion PGM™. This reduces the capacity of the system to alert the analyst to mixtures and represents a critical shortfall when the very high sensitivity of Ion PGM™ is borne in mind.

7.3.3 The EUROFORGEN Global AIM-SNP set

1. This study showed that the current extensive human genome variation catalogues can be easily accessed and their allele frequency data used to select highly differentiating ancestry informative SNPs. Marker sets with a range of sizes that meet the statistical power demands of forensic analysis, while focusing on the key characteristic of population differentiation balance, were developed.

2. Five population group comparisons were kept equally well differentiated and did not confine SNP informativeness to three or four global groups. Up until now, many AIM-SNP selections have ignored balanced differentiation of East Asians in favour of the closely related Native American group. Such approaches preclude unbiased and globally applicable analysis of a forensic sample of unknown geographic origin.

3. The *SPSmart* genome browsers proved to be particularly valuable aids to building comprehensive forensic AIM-SNP sets in three respects: i) they provide the only viable means to interrogate hundreds of markers with specific population properties in a single query; ii) the option to identify groups of SNPs with identical allele frequency patterns, and therefore divergence, is straightforward in *SPSmart* by defining the chromosome segment to be queried; and iii) *SPSmart* enables a very simple system for downloading genotype data from selected population groupings for porting into *Snipper* to assess patterns of divergence amongst markers and population pairings chosen.

4. Although *STRUCTURE* provides the most commonly applied system for estimating co-ancestry components in admixed individuals, there are widely held doubts about the validity of using mid-continental reference populations and not taking sufficient account of the clinal variation characteristic of so much of worldwide demographic structure. A simpler and more easily interpreted approach to the complex task of distinguishing patterns of admixed SNP variation from those of unadmixed ancestral populations, is to superimpose individuals with unknown ancestry alongside

reference population data on distance matrices generated by multi-dimensional scaling systems such as PCA.

5. Forensic ancestry analysis using real-time tools was improved in two ways: i) a co-ancestry component estimator based on the ratio of each likelihood to the sum of all likelihoods (the number of ancestries defined by the training set data uploaded) was added to *Snipper*; and ii) a PCA *R* script to position *de novo* AIM-SNP profiles on top of reference PCA plots using the first and second principal component percentages to define the profile's plot axes was implemented in *Snipper*.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

RESUMEN

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

8 Resumen

Aproximadamente el 95% de la variabilidad genética humana es intra-poblacional (Rosenberg et al. 2002), de tal manera que los individuos de cada población presentan más diferencias entre ellos que las propias poblaciones entre sí. Los microsatélites (STRs), comúnmente utilizados en aplicaciones de identificación individual, no son los más adecuados para la predicción de ancestralidad, ya que son altamente variables entre individuos y tienen baja variabilidad inter-poblacional. Sin embargo, existen STRs con alelos específicos de poblaciones (Phillips et al. 2008b), que pueden añadir un cierto grado de diferenciación poblacional. Por otra parte, se han propuesto marcadores binarios con grandes diferencias de frecuencias alélicas entre las poblaciones como alternativa para estimar las proporciones individuales de ancestralidad, permitiendo una predicción probable de origen biogeográfico. Los polimorfismos autosómicos, en particular, se encuentran entre los mejores marcadores informativos de ancestralidad (AIMs) debido a su estabilidad, densidad de distribución en el genoma y a un amplio rango de diferentes patrones de frecuencias alélicas entre poblaciones. Los polimorfismos autosómicos son un complemento apropiado al genotipado del cromosoma-Y y ADN mitocondrial en el análisis de la historia evolutiva de las poblaciones. Éstos tienen la ventaja de que se necesita un pequeño numero de muestras para obtener estimas fiables de sus frecuencias alélicas – alrededor de 30-40 muestras (Phillips 2015) – y presentan independencia de los fenómenos de mezcla sesgada por género, permitiendo una estima más precisa de la ancestralidad de un individuo.

En el campo de la genética forense, existe un interés en el desarrollo de ensayos utilizando estos marcadores ya que pueden ofrecer pistas, a partir de evidencias de ADN, en los casos en los que no se conoce la persona fuente de la muestra o ésta no puede declarar su ascendencia. Sin embargo, es importante comprender la distribución de la diversidad poblacional actual. Estudios realizados utilizando cientos de marcadores sugieren que la población mundial puede ser dividida en siete grupos ancestrales discretos (Rosenberg et al. 2002, Enoch et al. 2006, Allocco et al. 2007, Li et al. 2008), basados en similitud genética y que tienden a representar orígenes continentales. No obstante, este patrón puede ser muy

dependiente del muestreo ya que la mayoría de la diversidad genética humana está, en realidad, organizada en clinas y no en grupos (Serre and Paabo 2004, Manica et al. 2005, Rosenberg et al. 2005). Por este motivo, para detectar las diferencias entre los principales grupos continentales, es esencial encontrar marcadores que presenten diferencias en las frecuencias alélicas muy marcadas entre las regiones continentales (Rosenberg et al. 2002). Estos marcadores pueden ser incluidos en pequeños multiplexes para generar genotipos específicos de poblaciones. Para ello, es necesario un compromiso con la cantidad de información que se puede obtener con estos marcadores, simplificando el grado de complejidad en la estructura poblacional humana (Phillips 2015). En genética forense, la separación de cinco grupos ancestrales (África versus Europa versus Asia Oriental versus Oceanía versus población Nativo Americana) constituye un objetivo abordable con pequeños multiplexes que comprendan marcadores con frecuencias muy sesgadas (Phillips 2015). Para la diferenciación de poblaciones localizadas en el medio del continuo de variabilidad, como Oriente Medio y el Centro/Sur de Asia, que conllevan una mayor probabilidad de error de clasificación, se necesitan conjuntos más de extensos de marcadores que incluyan algunos con fuertes clinas de frecuencias (Phillips et al. 2007b).

Las cinco reacciones multiplex que se describen en esta tesis (sección 4) han sido desarrolladas en distintos momentos pero con objetivos complementarios. En primer lugar, la reacción *34-plex* fue desarrollada originariamente para diferenciar tres grupos poblacionales: África, Europa y Asia Oriental (Phillips et al. 2007b). La versión revisada presentada en esta tesis intenta diferenciar los mismos grupos poblacionales pero el cambio de un SNP (rs727811 remplazado por rs3827760) incrementa la diferenciación de las poblaciones de Asia Oriental. Pese a esto, este multiplex no tiene poder suficiente para diferenciar poblaciones de Oceanía y Nativo Americanas, por lo que se planteó el diseño de reacciones complementarias a *34-plex*. Uno de estos ensayos es *Pacifiplex* –que presenta una mejora en comparación con *34-plex* al permitir la diferenciación de las poblaciones de Asia Oriental con las de Oceanía. El análisis combinado de las dos reacciones acentúa la diferenciación de los patrones de ancestralidad observados en poblaciones de Oceanía (como Aborígenes Australianos y de Papúa Nueva Guinea) contra los de las poblaciones de Asia Oriental más cercanas (como Taiwán y Filipinas). Otra de las reacciones es un conjunto de *46 AIM-Indels* – diseñada como una herramienta simple para inferir ancestralidad en cuatro grupos poblacionales distintos (África, Europa, Asia oriental y población Nativo Americana).

Aunque esta reacción ha sido inicialmente diseñada para considerar cuatro grandes grupos poblacionales, la extensión a seis grupos reveló la capacidad de distinguir con seguridad Oceanía y, hasta cierto punto, detectar un componente de Centro/Sur de Asia. Asimismo, y aunque inicialmente no fue planteado, el análisis conjunto de esta reacción y *34-plex* reduce el error de clasificación cuando se analizan poblaciones Euroasiáticas (Europa, Oriente Medio y Centro/Sur de Asia). No obstante, otros AIMs deben ser añadidos para mejorar la clasificación de poblaciones cercanas. *Eurasiaplex* es una reacción compuesta por AIM-SNPs que, analizada en conjunto con *34-plex*, es capaz de diferenciar a la mayoría de poblaciones Europeas y Centro/Sur Asiáticas con elevados niveles de probabilidad. Con todo, la diferenciación de poblaciones de Oriente Medio tiene altos niveles de error de clasificación, evidenciando la dificultad de utilizar reacciones forenses de AIMs para distinguir poblaciones en regiones cercanas con gradiente de variabilidad genética. La mayoría de poblaciones Europeas y Sur Asiáticas pueden ser separadas de una forma fiable ya que estas están localizadas en los extremos de la clina. Por otra parte, poblaciones que ocupan las regiones centrales reducen la fiabilidad general de la clasificación – la diferenciación de las poblaciones de Oriente Medio de las Europeas o Centro/Sur Asiáticas requiere más marcadores que los incluidos en Eurasiaplex. De hecho, para predecir ancestralidad de poblaciones relacionadas (que pertenecen al mismo grupo ancestral) se necesitan muchos más marcadores: entre cientos y miles de polimorfismos (Biswas et al. 2009, Paschou et al. 2010). Además, la selección de los polimorfismos tiene mucha influencia en la capacidad para diferenciar determinados grupos poblacionales (particularmente los más próximos), siendo las poblaciones no consideradas en el proceso de selección las más propensas a errores de clasificación.

Todas las reacciones descritas en esta tesis han sido diseñadas y optimizadas para su aplicación en casos forenses – los marcadores son amplificados en fragmentos cortos (asegurando su aplicabilidad en muestras de ADN muy degradado) y tienen la sensibilidad necesaria para amplificar menos de 0,5 ng (una ventaja cuando se analizan muestras escasas como las que se suelen encontrar en casos forenses). No obstante, la mayor desventaja del análisis de SNPs es el método de tipado utilizado normalmente – SNaPshot® es una técnica en varios pasos que conlleva el cambio de tubos de reacción, aumentando el riesgo de contaminación y los errores de transferencia. Además, SNaPshot® utiliza distintos fluorocromos para detectar cada alelo, originando un desbalance en la intensidad, la cual

tiene, en media, un cociente de 4:2:1:1 para los nucleótidos G:A:C:T; esto aumenta la dificultad al intentar comparar la altura de los picos. Este problema se ve acentuado cuando la muestra de estudio es una mezcla – la detección y deconvolución de las muestras presenta desafíos para la interpretación cuando se analizan marcadores binarios. A pesar de que la inclusión de marcadores tri-alélicos puede aumentar la probabilidad de detección de mezclas (Phillips et al. 2004b), el problema puede ser abordado preferiblemente mediante los AIM-Indels, que tienen los mismos beneficios de los marcadores bialélicos pero un método de tipado directo, reduciéndose el numero de pasos y recursos necesarios para el genotipado y permitiendo la comparación del cociente entre las alturas de los picos para cada par de alelos. Sin embargo, el análisis de STRs permite la detección de mezclas y la separación de los genotipos de cada componente de dichas mezclas. El análisis del *12-plex* de AIM-STRs tetranucleotídicos descrito en la sección 4, a pesar de proporcionar menos información que las reacciones compuestas por marcadores binarios, es la mejor alternativa para detectar y analizar mezclas. El análisis conjunto de los AIM-STRs, los STRs rutinarios de identificación y los AIM-SNPs de *34-plex* constituye un conjunto potente de marcadores – la combinación de los 66 marcadores con distintas características mejora la diferenciación de las poblaciones Eurasiáticas, aunque las poblaciones de Oriente Medio siguen siendo muy difíciles de diferenciar sin analizar conjuntos mucho más grandes de marcadores. Además, el *12-plex* tiene niveles de discriminación individual comparables a los marcadores utilizados en rutina, siendo por eso aplicable para otros fines forenses además de la inferencia de ancestralidad.

Una vez generados los genotipos de los AIMs a partir de una muestra de ADN, es necesario inferir el origen más probable y estimar las proporciones de ancestralidad. Existen varios métodos, cada uno con sus ventajas y desventajas, aplicables en distintas situaciones, Un posible enfoque implica un conjunto de poblaciones predefinidas contra las cuales se clasifican individuos de ancestralidad desconocida. Las frecuencias alélicas estimadas para un conjunto de marcadores independientes en cada población parental son utilizadas para calcular la probabilidad que un determinado perfil genotípico se haya originado en cada una de las poblaciones posibles. Éstas probabilidades permiten asignar a los individuos desconocidos a poblaciones, teniendo en cuenta el cociente de verisimilitudes más alto (Pritchard et al. 2000).

En la presente tesis se describe un conjunto de herramientas aplicables a la inferencia de ancestralidad en genética forense (sección 3). El análisis de componentes principales (PCA)

es una herramienta exploratoria comúnmente utilizada para sintetizar las similitudes y diferencias genéticas entre grupos de poblaciones o individuos. El script de *R* incluido puede ser utilizado para representar una muestra de un caso forense con un conjunto de muestras de referencia ayudando a inferir, a través de inspección visual, la clasificación más probable del individuo. Teniendo en cuenta la alta velocidad de cálculo de las componentes principales, este análisis puede ser utilizado como preámbulo de los métodos Bayesianos considerados en esta tesis: *STRUCTURE* y *Snipper*.

Los ficheros de salida de *STRUCTURE* incluyen información acerca de los grupos estimados, es decir, de los grupos poblacionales generados y no los grupos proporcionados. La utilización de poblaciones de referencia muy similares a los grupos inferidos es importante cuando se intenta clasificar una población o individuo, en particular cuando se analizan individuos mestizos y se necesita saber cuáles son las poblaciones parentales que contribuyen. Además, *STRUCTURE* aporta información sobre el conjunto de entrenamiento utilizado (permitiendo la evaluación de los datos de referencia utilizados – el valor optimo de *K* es igual al numero de poblaciones de referencia, las cuales son completamente distintas entre si) y las proporciones de ancestralidad de cada muestra. Estas proporciones, tanto para cada población como para cada individuo, pueden ser representadas en dos tipos de gráficas: una gráfica triangular donde cada individuo está representado por un punto coloreado y una gráfica de barras donde cada individuo está representado por una línea vertical dividida en *K* segmentos de colores, cada uno correspondiente a la proporción estimada en cada uno de los *K* grupos inferidos. Esta última es particularmente útil cuando se intentan visualizar datos para $K>3$.

La salida de *Snipper* incluye una variedad de información que puede ser útil tanto en la etapa de diseño del ensayo (por ejemplo, el éxito de clasificación y una lista de los marcadores en orden descendente de divergencia calculada en base al número de poblaciones incluidas en la comparación) como en la clasificación de una muestra de un caso forense (cocientes de probabilidades en formato detallado y proporciones de ancestralidad). Gráficas de cocientes de verosimilitud por pares, tales como la generada para el análisis de ancestralidad realizado para el caso del atentado del 11-M en Madrid (Phillips et al. 2009) (sección 1.3.2.3, ejemplificada en la **Figure 21**), se pueden utilizar para medir los posibles patrones de mestizaje poblacional y para realizar una validación cruzada del conjunto de entrenamiento para evaluar la precisión y el rendimiento de los marcadores utilizados. En este

análisis es importante definir umbrales de valores atípicos de probabilidad, es decir, aquellos por los que una muestra no se puede asignar de forma fiable a un determinado grupo poblacional. El análisis de las mismas muestras utilizando *STRUCTURE* ayuda a definir dichos umbrales para detectar individuos mestizos (que aumentan el riesgo de errores de clasificación).

Dos estudios recientes (Phillips 2015, Santos et al. 2015a) han demostrado que los tres sistemas de análisis de datos aquí descritos proporcionan resultados muy similares en patrones obtenidos con *STRUCTURE*, probabilidades de asignación de *Snipper* y la posición en las gráficas de PCA de las muestras desconocidas en comparación con poblaciones de referencia.

Dos enfoques diferentes para la inferencia de ancestralidad o la estima de proporciones de co-ancestralidad en individuos mestizos han sido propuestas anteriormente: i) pequeños conjuntos de marcadores cuyo objetivo es alcanzar la mayor divergencia posible (en concreto, la medida δ) durante la selección de SNPs (Shriver et al. 1997, Rosenberg et al. 2003); ii) balancear la divergencia entre los grupos de poblaciones de manera que se diferencie a cada grupo con igual poder, reduciendo el sesgo en la estima de las proporciones de ancestralidad principalmente cuando se analizan individuos mestizos (Phillips 2015). Aunque no es un objetivo fácil de alcanzar, equilibrar la divergencia permite la construcción de conjuntos de marcadores más aplicables y apropiados.

Este hecho, junto con el aumento de la capacidad de multiplex que ofrece la secuenciación de nueva generación, condujo al desarrollo del conjunto *Global AIM-SNPs*, que incluye 128 marcadores balanceados para la comparación de cinco grupos poblacionales. Aunque se analizaron un número limitado de muestras (cuatro poblaciones mestizas del proyecto 1000 Genomas), fue posible observar como el conjunto *Global AIM-SNPs* puede medir mestizaje entre los cinco grupos poblacionales considerados correctamente. Además, el análisis de los Fiyianos con 116/128 SNPs sugiere que, a pesar de la reciente inmigración e integración de los Sur Asiáticos en Fiyi, la divergencia entre Oceanía Lejana y Cercana es detectable, siendo un factor adicional a considerar en la interpretación de los patrones de ancestralidad. Otro de los objetivos del estudio de los *Global AIM-SNPs* fue evaluar si las herramientas de análisis son las adecuadas. En el análisis forense es beneficioso tener sistemas alternativos y flexibles para la inferencia de ancestralidad y que estos puedan manejar fácilmente los perfiles individuales sin realizar cálculos intensos y de larga duración como los realizados por *STRUCTURE*. Por esto, hemos mejorado el análisis de ancestralidad

utilizando herramientas en tiempo real. En primer lugar, mediante la adición de un estimador de proporciones de co-ancestralidad en *Snipper* basado en el coeficiente de cada probabilidad con la suma de todas las probabilidades (dependiendo del número de poblaciones definidas por los datos del conjunto de entrenamiento). En segundo lugar, mediante el desarrollo de un script de *R*, que coloca los perfiles de los casos en las gráficas PCA generadas con los datos de referencia utilizando los porcentajes de la primera y segunda componentes principales para definir los ejes de la gráfica sobre la que se representa el nuevo perfil. Este script ha sido implementado en *Snipper*.

Sin embargo, al considerar la aplicación de secuenciación de nueva generación en el análisis forense, dos preguntas deben ser planteadas – ¿Podemos seguir utilizando los ensayos de AIMs ya desarrollados para genética forense? ¿Cómo de buena es la tecnología?

La primera cuestión se abordó en el estudio presentado en la sección 5 (Anexo 3A). Teniendo en cuenta los resultados de este estudio preliminar, el sistema Ion PGM™ ha demostrado un gran potencial para su uso en el análisis forense de ADN como plataforma de NGS de rendimiento bajo a medio. Los chips Ion 314 v1 utilizados generaron tres veces más datos de lo esperado, sin necesidad de optimizar o adaptar las reacciones de PCR multiplex que se utilizaron (a partir de reacciones SNaPshot® desarrolladas previamente) para producir los amplicones para secuenciación. La cobertura promedio de 500 × por amplicón (SNP) fue suficiente para alcanzar una precisión de genotipado del 97% para los 136 SNPs evaluados en este estudio, en tres muestras con tres cantidades diferentes de ADN amplificadas (hasta un mínimo de 0,5 ng). El único SNP que ha sido genotipado de forma errónea en Ion Torrent fue rs1029047, el cual tiene secuencias homopoliméricas conocidas en las zonas que flanquean inmediatamente ambos lados del SNP (Seo et al. 2013, Børsting et al. 2014). La capacidad de estimar de forma precisa las proporciones de ancestralidad y información acerca posibles características visibles externas (ECVs) requiere el análisis de más SNPs que aquellos que puede ser amplificados en una única reacción SNaPshot®. Un único ensayo de NGS puede contener cientos de SNPs, maximizando el uso de una escasa cantidad de muestra y reduciendo el tiempo de procesamiento de la muestra y los posibles riesgos de contaminación.

Una evaluación más formal de los parámetros y umbrales ha sido realizada en el estudio presentado en la sección 5 (Anexos 3B-E). La evaluación de la sensibilidad y precisión de genotipado del Ion PGM™ realizada en este estudio proporciona un fuerte soporte a aplicación de las tecnologías NGS al análisis forense de ADN. Los datos de secuenciación

generaron valores suficientemente altos de cobertura y originaron genotipos fiables para la mayoría de los marcadores incluidos en el panel HID SNP. Cinco SNPs con discordancias deben excluirse del panel – dos de ellos (rs1004357 y rs2032597) ya se han retirado de la versión revisada del panel de HID SNP, mientras que rs2399332 fue identificado en el estudio de Børsting et al. (2014) como problemático. Por último, hay que mencionar rs1029047 que, como en el estudio presentado en el Anexo 3A, genera inconsistencias de genotipado en todos los estudios NGS que lo incluyen realizados hasta la fecha (Seo et al. 2013, Børsting et al. 2014). Es importante realizar un escrutinio cuidadoso de las secuencias flanqueantes de cualquier SNP elegido para uso forense, en particular de SNPs codificantes de los ensayos ECVs, ya que estos deben funcionar bien para que los análisis sean suficientemente informativos. Además, existe cierto consenso en que un umbral mínimo de aproximadamente 15-20 × de cobertura puede garantizar la fiabilidad de los alelos genotipados con NGS (Bentley et al. 2008, Quail et al. 2012, Daniel et al. 2015). Finalmente, las conclusiones de este estudio coinciden con las de Børsting et al. (2014) – el software de análisis incluido en el sistema Ion PGM™ necesita una mayor optimización para su total adecuación a las aplicaciones forenses, aunque está siendo revisado continuamente para este fin.

APPENDIX

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

9 Appendixes

As a supplement to the introduction:

1. *Appendix I* – "Nuevos biomarcadores en genética forense" in Santiago Delgado Bueno, Fernando Bandrés Moya y Joaquín Lucena Romero (orgs.), Tratado de Medicina Legal y Ciencias Forenses - Patología y Biología Forense. Barcelona, Editorial Bosch, S.A., 1151- 1169;
2. *Appendix II* – "New turns from old STaRs: enhancing the capabilities of forensic short tandem repeat analysis" (http://www.ncbi.nlm.nih.gov/pubmed/24888494);
3. *Appendix III* – "Inference of ancestry in forensic analysis I: autosomal ancestry-informative marker sets" (Book chapter to be included in "Forensic DNA Profiling Protocols" – expected online in 2015).

From Block 1:

4. *Appendix 1A* – "An overview of *STRUCTURE*: applications, parameter settings, and supporting software" (http://www.ncbi.nlm.nih.gov/pubmed/23755071);
5. *Appendix 1B* – "Inference of ancestry in forensic analysis II: analysis of genetic data" (Book chapter to be included in "Forensic DNA Profiling Protocols" – expected online in 2015);
6. *Appendix 1C* – "An assessment of Bayesian and multinomial logistic regression classification systems to analyse admixed individuals" (http://www.sciencedirect.com/science/article/pii/S1875176813000334).

From Block 2:

7. *Appendix 2A* – "Revision of the SNP*for*ID *34-plex* forensic ancestry test: assay enhancements, standard reference sample genotypes and extended population studies" (http://www.ncbi.nlm.nih.gov/pubmed/22749789);

8. *Appendix 2B* – "Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing" (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3260179/);
9. *Appendix 2C* – "Completion of a worldwide reference panel of samples for an ancestry informative Indel assay" (http://www.ncbi.nlm.nih.gov/pubmed/25840342);
10. *Appendix 2D* – "*Eurasiaplex*: a forensic SNP assay for differentiating European and South Asian ancestries" (http://www.ncbi.nlm.nih.gov/pubmed/23537756);
11. *Appendix 2E* – "Development of a novel forensic STR multiplex for ancestry analysis and extended identity testing" (http://www.ncbi.nlm.nih.gov/pubmed/23400880).

From Block 3:

12. *Appendix 3A* – "A SNaPshot of next generation sequencing for forensic analysis" (http://www.ncbi.nlm.nih.gov/pubmed/25282603);
13. *Appendix 3B* – "Inter-laboratory evaluation of SNP-based forensic identification by massively parallel sequencing using Ion PGM™" (http://www.ncbi.nlm.nih.gov/pubmed/25955683);
14. *Appendix 3C* – "Assessments of sequence coverage obtained with HID SNP markers and the Ion PGM™" (Supplementary Material associated with Appendix 3B publication);
15. *Appendix 3D* – IGV overview of outlier SNPs (Supplementary Material associated with Appendix 3B publication);
16. *Appendix 3E* – "Mixture analysis with the Ion PGM™" (Supplementary Material associated with Appendix 3B publication);
17. *Appendix 3F* – "Building a forensic ancestry panel from the ground up: the EUROFORGEN Global AIM-SNP set" (http://www.ncbi.nlm.nih.gov/pubmed/24631693).

REFERENCES

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

10 References

1000GPC (2010). "*A map of human genome variation from population-scale sequencing*." Nature **467**(7319): 1061-1073.

1000GPC (2012). "*An integrated map of genetic variation from 1,092 human genomes*." Nature **491**(7422): 56-65.

Abecasis, G. R., D. Ghosh and T. E. Nichols (2005). "*Linkage disequilibrium: ancient history drives the new genetics*." Hum Hered **59**(2): 118-124.

Abi-Rached, L., M. J. Jobin, S. Kulkarni, A. McWhinnie, K. Dalva, L. Gragert, F. Babrzadeh, B. Gharizadeh, M. Luo, F. A. Plummer, et al. (2011). "*The shaping of modern human immune systems by multiregional admixture with archaic humans*." Science **334**(6052): 89-94.

Ackermann, K., K. N. Ballantyne and M. Kayser (2010). "*Estimating trace deposition time with circadian biomarkers: a prospective and versatile tool for crime scene reconstruction*." Int J Legal Med **124**(5): 387-395.

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle, et al. (2000). "*The Genome Sequence of Drosophila melanogaster*." Science **287**(5461): 2185-2195.

Akane, A., H. Shiono, K. Matsubara, Y. Nakahori, S. Seki, S. Nagafuchi, M. Yamada and Y. Nakagome (1991). "*Sex identification of forensic specimens by polymerase chain reaction (PCR): two alternative methods*." Forensic Sci Int **49**(1): 81-88.

Albrechtsen, A., F. C. Nielsen and R. Nielsen (2010). "*Ascertainment biases in SNP chips affect measures of population divergence*." Mol Biol Evol **27**(11): 2534-2547.

Alexander, D. H. and K. Lange (2011). "*Enhancements to the ADMIXTURE algorithm for individual ancestry estimation*." BMC Bioinformatics **12**: 246.

Alexander, D. H., J. Novembre and K. Lange (2009). "*Fast model-based estimation of ancestry in unrelated individuals*." Genome Res **19**(9): 1655-1664.

Allocco, D. J., Q. Song, G. H. Gibbons, M. F. Ramoni and I. S. Kohane (2007). "*Geography and genography: prediction of continental origin using randomly selected single nucleotide polymorphisms*." BMC Genomics **8**: 68.

Amigo, J., C. Phillips, M. Lareu and A. Carracedo (2008a). "*The SNPforID browser: an online tool for query and display of frequency data from the SNPforID project*." Int J Legal Med **122**(5): 435-440.

Amigo, J., C. Phillips, A. Salas, L. Fernandez-Formoso, Á. Carracedo and M. V. Lareu (2009). "*pop.STR—An online population frequency browser for established and new forensic STRs*." Forensic Sci Int Genet Suppl Ser **2**(1): 361-362.

Amigo, J., A. Salas and C. Phillips (2011). "*ENGINES: exploring single nucleotide variation in entire human genomes*." BMC Bioinformatics **12**: 105.

Amigo, J., A. Salas, C. Phillips and Á. Carracedo (2008b). "*SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access*." BMC Bioinformatics **9**: -.

Amorim, A. and L. Pereira (2005). "*Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs*." Forensic Sci Int **150**(1): 17-21.

Anderson, S., B. Howard, G. R. Hobbs and C. P. Bishop (2005). "*A method for determining the age of a bloodstain*." Forensic Sci Int **148**(1): 37-45.

Anderson, S. E., G. R. Hobbs and C. P. Bishop (2011). "*Multivariate analysis for estimating the age of a bloodstain*." J Forensic Sci **56**(1): 186-193.

Andresen, P. H. (1947). "*Blood group with characteristic phenotypical aspects*." Acta Pathologica Microbiologica Scandinavica **24**(5-6): 616-618.

Ankel-Simons, F. and J. M. Cummins (1996). "*Misconceptions about mitochondria and mammalian fertilization: Implications for theories on human evolution*." Proc Natl Acad Sci U S A **93**(24): 13859-13863.

Armitage, S. J., S. A. Jasim, A. E. Marks, A. G. Parker, V. I. Usik and H. P. Uerpmann (2011). "*The southern route "out of Africa": evidence for an early expansion of modern humans into Arabia*." Science **331**(6016): 453-456.

Astle, W. and D. J. Balding (2009). "*Population structure and cryptic relatedness in genetic association studies*." Statist. Sci. **24**(4): 451-471.

Atkinson, Q. D. (2011). "*Phonemic diversity supports a serial founder effect model of language expansion from Africa*." Science **332**(6027): 346-349.

Aulchenko, Y. S., M. V. Struchalin, N. M. Belonogova, T. I. Axenovich, M. N. Weedon, A. Hofman, A. G. Uitterlinden, M. Kayser, B. A. Oostra, C. M. van Duijn, et al. (2009). "*Predicting human height by Victorian and genomic methods*." Eur J Hum Genet **17**(8): 1070-1075.

Avery, O. T., C. M. Macleod and M. McCarty (1944). "*Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III*." J Exp Med **79**(2): 137-158.

Ball, E. V., P. D. Stenson, S. S. Abeysinghe, M. Krawczak, D. N. Cooper and N. A. Chuzhanova (2005). "*Microdeletions and microinsertions causing human genetic disease: common mechanisms of mutagenesis and the role of local DNA sequence complexity*." Human Mutation **26**(3): 205-213.

Ball, M. P., J. B. Li, Y. Gao, J. H. Lee, E. M. LeProust, I. H. Park, B. Xie, G. Q. Daley and G. M. Church (2009). "*Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells*." Nat Biotechnol **27**(4): 361-368.

Ballantyne, K. N., M. Goedbloed, R. Fang, O. Schaap, O. Lao, A. Wollstein, Y. Choi, K. van Duijn, M. Vermeulen, S. Brauer, et al. (2010). "*Mutability of Y-chromosomal*

microsatellites: rates, characteristics, molecular bases, and forensic implications." Am J Hum Genet **87**(3): 341-353.

Ballantyne, K. N., V. Keerl, A. Wollstein, Y. Choi, S. B. Zuniga, A. Ralf, M. Vermeulen, P. de Knijff and M. Kayser (2012). "*A new future of forensic Y-chromosome analysis: rapidly mutating Y-STRs for differentiating male relatives and paternal lineages*." Forensic Sci Int Genet **6**(2): 208-218.

Ballantyne, K. N., A. Ralf, R. Aboukhalid, N. M. Achakzai, M. J. Anjos, Q. Ayub, J. Balazic, J. Ballantyne, D. J. Ballard, B. Berger, et al. (2014). "*Toward male individualization with rapidly mutating y-chromosomal short tandem repeats*." Hum Mutat **35**(8): 1021-1032.

Ballantyne, K. N., R. A. van Oorschot, I. Muharam, A. van Daal and R. John Mitchell (2007a). "*Decreasing amplification bias associated with multiple displacement amplification and short tandem repeat genotyping*." Anal Biochem **368**(2): 222-229.

Ballantyne, K. N., R. A. H. van Oorschot, M. Kayser and R. J. Mitchell (2007b). "*Determining restriction digest efficiency using the SNaPshot single-base extension method and CE*." Electrophoresis **28**(10): 1514-1517.

Balme, J. (2013). "*Of boats and string: The maritime colonisation of Australia*." Quat Int **285**(0): 68-75.

Bamshad, M., S. Wooding, B. A. Salisbury and J. C. Stephens (2004). "*Deconstructing the relationship between genetics and race*." Nat Rev Genet **5**(8): 598-609.

Barker, G., H. Barton, M. Bird, P. Daly, I. Datan, A. Dykes, L. Farr, D. Gilbertson, B. Harrisson, C. Hunt, et al. (2007). "*The 'human revolution' in lowland tropical Southeast Asia: the antiquity and behavior of anatomically modern humans at Niah Cave (Sarawak, Borneo)*." J Hum Evol **52**(3): 243-261.

Bastiaens, M., J. ter Huurne, N. Gruis, W. Bergman, R. Westendorp, B. J. Vermeer and J. N. Bouwes Bavinck (2001). "*The melanocortin-1-receptor gene is the major freckle gene*." Hum Mol Genet **10**(16): 1701-1708.

Bastos-Rodrigues, L., J. R. Pimenta and S. D. Pena (2006). "*The genetic structure of human populations studied through short insertion-deletion polymorphisms*." Ann Hum Genet **70**(Pt 5): 658-665.

Bauer, C. M., H. Niederstatter, G. McGlynn, H. Stadler and W. Parson (2013). "*Comparison of morphological and molecular genetic sex-typing on mediaeval human skeletal remains*." Forensic Sci Int Genet **7**(6): 581-586.

Bauer, M., S. Polzin and D. Patzelt (2003). "*Quantification of RNA degradation by semi-quantitative duplex and competitive RT-PCR: a possible indicator of the age of bloodstains?*" Forensic Sci Int **138**(1-3): 94-103.

Beaumont, K. A., S. N. Shekar, A. L. Cook, D. L. Duffy and R. A. Sturm (2008). "*Red hair is the null phenotype of MC1R*." Hum Mutat **29**(8): E88-94.

Becker, D., H. Rodig, C. Augustin, J. Edelmann, F. Gotz, S. Hering, R. Szibor and W. Brabetz (2008). "*Population genetic evaluation of eight X-chromosomal short tandem*

repeat loci using Mentype Argus X-8 PCR amplification kit." Forensic Sci Int Genet **2**(1): 69-74.

Bellwood, P. (2001). "*Early agriculturalist population diasporas? Farming, languages, and genes*." Annu Rev Anthropol **30**(1): 181-207.

Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, et al. (2008). "*Accurate whole human genome sequencing using reversible terminator chemistry*." Nature **456**(7218): 53-59.

Bersaglieri, T., P. C. Sabeti, N. Patterson, T. Vanderploeg, S. F. Schaffner, J. A. Drake, M. Rhodes, D. E. Reich and J. N. Hirschhorn (2004). "*Genetic signatures of strong recent positive selection at the lactase gene*." Am J Hum Genet **74**(6): 1111-1120.

Beyin, A. (2011). "*Upper Pleistocene Human Dispersals out of Africa: A Review of the Current State of the Debate*." Int J Evol Biol **2011**: 615094.

Bhangale, T. R., M. J. Rieder, R. J. Livingston and D. A. Nickerson (2005). "*Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes*." Hum Mol Genet **14**(1): 59-69.

Biesecker, L. G., J. E. Bailey-Wilson, J. Ballantyne, H. Baum, F. R. Bieber, C. Brenner, B. Budowle, J. M. Butler, G. Carmody, P. M. Conneally, et al. (2005). "*Epidemiology. DNA identifications after the 9/11 World Trade Center attack*." Science **310**(5751): 1122-1123.

Biswas, S., L. B. Scheinfeldt and J. M. Akey (2009). "*Genome-wide insights into the patterns and determinants of fine-scale population structure in humans*." Am J Hum Genet **84**(5): 641-650.

Blust, R. (1999). *Selected Papers from the Eighth International Conference on Austronesian Linguistics*. Symp. Ser. Inst. Linguist., Acad. Sinica. E. Zeitoun and P. J. K. Li. Taipei, Taiwan.

Bobillo, C., M. Marino, A. Sala, L. Gusmão and D. Corach (2008). "*X-STRs: Relevance in complex kinship cases*." Forensic Sci Int Genet Suppl Ser **1**(1): 496-498.

Borel, C., F. Cheung, H. Stewart, D. A. Koolen, C. Phillips, N. S. Thomas, P. A. Jacobs, S. Eliez and A. J. Sharp (2012). "*Evaluation of PRDM9 variation as a risk factor for recurrent genomic disorders and chromosomal non-disjunction*." Hum Genet **131**(9): 1519-1524.

Bornman, D. M., M. E. Hester, J. M. Schuetter, M. D. Kasoji, A. Minard-Smith, C. A. Barden, S. C. Nelson, G. D. Godbold, C. H. Baker, B. Yang, et al. (2012). "*Short-read, high-throughput sequencing technology for STR genotyping*." Biotech Rapid dispatches **2012**: 1-6.

Børsting, C., S. L. Fordyce, J. Olofsson, H. S. Mogensen and N. Morling (2014). "*Evaluation of the Ion Torrent HID SNP 169-plex: A SNP typing assay developed for human identification by second generation sequencing*." Forensic Sci Int Genet **12**: 144-154.

Børsting, C., E. Rockenbauer and N. Morling (2009). "*Validation of a single nucleotide polymorphism (SNP) typing assay with 49 SNPs for forensic genetic testing in a laboratory accredited according to the ISO 17025 standard*." Forensic Sci Int Genet **4**(1): 34-42.

Børsting, C., C. Tomas and N. Morling (2011). "*SNP typing of the reference materials SRM 2391b 1-10, K562, XY1, XX74, and 007 with the SNPforID multiplex*." Forensic Sci Int Genet **5**(3): e81-82.

Bowler, J. M., H. Johnston, J. M. Olley, J. R. Prescott, R. G. Roberts, W. Shawcross and N. A. Spooner (2003). "*New ages for human occupation and climatic change at Lake Mungo, Australia*." Nature **421**(6925): 837-840.

Branicki, W., F. Liu, K. van Duijn, J. Draus-Barini, E. Pospiech, S. Walsh, T. Kupiec, A. Wojas-Pelc and M. Kayser (2011). "*Model-based prediction of human hair color using DNA variants*." Hum Genet **129**(4): 443-454.

Brinkmann, B., M. Klintschar, F. Neuhuber, J. Huhne and B. Rolf (1998). "*Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat*." Am J Hum Genet **62**(6): 1408-1415.

Britten, R. J., L. Rowen, J. Williams and R. A. Cameron (2003). "*Majority of divergence between closely related DNA samples is due to indels*." Proc Natl Acad Sci U S A **100**(8): 4661-4665.

Broman, K. W., J. C. Murray, V. C. Sheffield, R. L. White and J. L. Weber (1998). "*Comprehensive human genetic maps: individual and sex-specific variation in recombination*." Am J Hum Genet **63**(3): 861-869.

Brown, P., T. Sutikna, M. J. Morwood, R. P. Soejono, Jatmiko, E. W. Saptomo and R. A. Due (2004). "*A new small-bodied hominin from the Late Pleistocene of Flores, Indonesia*." Nature **431**(7012): 1055-1061.

Brunet, M., F. Guy, D. Pilbeam, H. T. Mackaye, A. Likius, D. Ahounta, A. Beauvilain, C. Blondel, H. Bocherens, J. R. Boisserie, et al. (2002). "*A new hominid from the Upper Miocene of Chad, Central Africa*." Nature **418**(6894): 145-151.

Budowle, B. (2004). "*SNP typing strategies*." Forensic Sci Int **146 Suppl**: S139-142.

Budowle, B. and A. van Daal (2008). "*Forensically relevant SNP classes*." Biotechniques **44**(5): 603-608, 610.

Buermans, H. P. J. and J. T. den Dunnen (2014). "*Next generation sequencing technology: Advances and applications*." Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease **1842**(10): 1932-1941.

Butler, J. M. (2011). *Advanced topics in forensic DNA typing: methodology*, Academic Press.

Butler, J. M., Y. Shen and B. R. McCord (2003). "*The development of reduced size STR amplicons as tools for analysis of degraded DNA*." J Forensic Sci **48**(5): 1054-1064.

Cann, H. M., C. de Toma, L. Cazes, M.-F. Legrand, V. Morel, L. Piouffre, J. Bodmer, W. F. Bodmer, B. Bonne-Tamir, A. Cambon-Thomsen, et al. (2002). "*A Human Genome Diversity Cell Line Panel*." Science **296**(5566): 261-262.

Cann, R. L. (2000). "*Linguistics. Talking trees tell tales*." Nature **405**(6790): 1008-1009.

Carracedo, Á., J. M. Butler, L. Gusmao, A. Linacre, W. Parson, L. Roewer and P. M. Schneider (2013). "*New guidelines for the publication of genetic population data*." Forensic Sci Int Genet **7**(2): 217-220.

Chen, C. H., T. J. Chuang, B. Y. Liao and F. C. Chen (2009). "*Scanning for the signatures of positive selection for human-specific insertions and deletions*." Genome Biol Evol **1**: 415-419.

Chen, F. C. and W. H. Li (2001). "*Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees*." Am J Hum Genet **68**(2): 444-456.

Chen, H. D., C. H. Chang, L. C. Hsieh and H. C. Lee (2005). "*Divergence and Shannon information in genomes*." Phys Rev Lett **94**(17): 178103.

Cho, M. K. and P. Sankar (2004). "*Forensic genetics and ethical, legal and social implications beyond the clinic*." Nat Genet **36**(11 Suppl): S8-12.

Claes, P., H. Hill and M. D. Shriver (2014a). "*Toward DNA-based facial composites: preliminary results and validation*." Forensic Sci Int Genet **13**: 208-216.

Claes, P., D. K. Liberton, K. Daniels, K. M. Rosana, E. E. Quillen, L. N. Pearson, B. McEvoy, M. Bauchet, A. A. Zaidi, W. Yao, et al. (2014b). "*Modeling 3D Facial Shape from DNA*." PLoS Genet **10**(3): e1004224.

Clark, A. G., M. J. Hubisz, C. D. Bustamante, S. H. Williamson and R. Nielsen (2005). "*Ascertainment bias in studies of human genome-wide polymorphism*." Genome Res **15**(11): 1496-1502.

Clark, T. G., T. Andrew, G. M. Cooper, E. H. Margulies, J. C. Mullikin and D. J. Balding (2007). "*Functional constraint and small insertions and deletions in the ENCODE regions of the human genome*." Genome Biol **8**(9): R180.

Coble, M. D. and J. M. Butler (2005). "*Characterization of new miniSTR loci to aid analysis of degraded DNA*." J Forensic Sci **50**(1): 43-53.

Collins, A. (2009). "*Allelic association: linkage disequilibrium structure and gene mapping*." Mol Biotechnol **41**(1): 83-89.

Collins, F. S., M. L. Drumm, J. L. Cole, W. K. Lockwood, G. F. Vande Woude and M. C. Iannuzzi (1987). "*Construction of a general human chromosome jumping library, with application to cystic fibrosis*." Science **235**(4792): 1046-1049.

Collins-Schramm, H. E., C. M. Phillips, D. J. Operario, J. S. Lee, J. L. Weber, R. L. Hanson, W. C. Knowler, R. Cooper, H. Li and M. F. Seldin (2002). "*Ethnic-Difference Markers for Use in Mapping by Admixture Linkage Disequilibrium*." Am J Hum Genet **70**(3): 737-750.

Colonna, V., L. Pagani, Y. Xue and C. Tyler-Smith (2011). "*A world in a grain of sand: human history from genetic data*." Genome Biol **12**(11): 234.

Conrad, D. F., M. Jakobsson, G. Coop, X. Wen, J. D. Wall, N. A. Rosenberg and J. K. Pritchard (2006). "*A worldwide survey of haplotype variation and linkage disequilibrium in the human genome*." Nat Genet **38**(11): 1251-1260.

Coop, G., J. K. Pickrell, J. Novembre, S. Kudaravalli, J. Li, D. Absher, R. M. Myers, L. L. Cavalli-Sforza, M. W. Feldman and J. K. Pritchard (2009). "*The role of geography in human adaptation*." PLoS Genet **5**(6): e1000500.

Cooper, A. and C. B. Stringer (2013). "*Paleontology. Did the Denisovans cross Wallace's Line?*" Science **342**(6156): 321-323.

Corander, J. and P. Marttinen (2006). "*Bayesian identification of admixture events using multilocus molecular markers*." Mol Ecol **15**(10): 2833-2843.

Corander, J., P. Waldmann and M. J. Sillanpaa (2003). "*Bayesian Analysis of Genetic Differentiation Between Populations*." Genetics **163**(1): 367-374.

Corser, C. A., P. A. McLenachan, M. J. Pierson, G. L. Harrison and D. Penny (2012). "*The Q2 mitochondrial haplogroup in Oceania*." PLoS One **7**(12): e52022.

Costantini, M. and G. Bernardi (2009). "*Mapping insertions, deletions and SNPs on Venter's chromosomes*." PLoS One **4**(6): e5972.

Costas, J., A. Salas, C. Phillips and A. Carracedo (2005). "*Human genome-wide screen of haplotype-like blocks of reduced diversity*." Gene **349**: 219-225.

Cutbush, M. and P. L. Mollison (1950). "*The Duffy blood group system*." Heredity **4**(3): 383-389.

Cutbush, M., P. L. Mollison and D. M. Parkin (1950). "*A New Human Blood Group*." Nature **165**(4188): 188-189.

Dalsgaard, S., E. Rockenbauer, A. Buchard, H. S. Mogensen, R. Frank-Hansen, C. Borsting and N. Morling (2014). "*Non-uniform phenotyping of D12S391 resolved by second generation sequencing*." Forensic Sci Int Genet **8**(1): 195-199.

Daniel, R., C. Santos, C. Phillips, M. Fondevila, R. A. van Oorschot, A. Carracedo, M. V. Lareu and D. McNevin (2015). "*A SNaPshot of next generation sequencing for forensic SNP analysis*." Forensic Sci Int Genet **14**: 50-60.

Dausset, J. (1958). "*Iso-leuco-anticorps*." Acta Haematol **20**(1-4): 156-166.

Davis, C., M. Illescas, C. Tirado, R. Lopez, B. Budowle and T. D. Cruz (2012). "*A case of Amelogenin Y-null: a simple primer binding site mutation or unusual genetic anomaly?*" Leg Med (Tokyo) **14**(6): 320-323.

Dawson, E., Y. Chen, S. Hunt, L. J. Smink, A. Hunt, K. Rice, S. Livingston, S. Bumpstead, R. Bruskiewich, P. Sham, et al. (2001). "*A SNP Resource for Human Chromosome 22: Extracting Dense Clusters of SNPs From the Genomic Sequence*." Genome Res **11**(1): 170-178.

de la Chaux, N., P. W. Messer and P. F. Arndt (2007). "*DNA indels in coding regions reveal selective constraints on protein evolution in the human lineage*." BMC Evol Biol **7**: 191.

Delfin, F., S. Myles, Y. Choi, D. Hughes, R. Illek, M. van Oven, B. Pakendorf, M. Kayser and M. Stoneking (2012). "*Bridging near and remote Oceania: mtDNA and NRY variation in the Solomon Islands*." Mol Biol Evol **29**(2): 545-564.

Delmas, C. E. L., E. Lhuillier, A. Pornon and N. Escaravage (2011). "*Isolation and characterization of microsatellite loci in Rhododendron ferrugineum (Ericaceae) using pyrosequencing technology*." Am J Bot **98**(5): e120-e122.

Dennell, R. and M. D. Petraglia (2012). "*The dispersal of Homo sapiens across southern Asia: how early, how often, how complex?*" Quat Sci Rev **47**(0): 15-22.

Derricourt, R. (2005). "*Getting "Out of Africa": sea crossings, land crossings and culture in the hominin migrations*." J World Prehist **19**: 119-132.

Diamond, J. (1988). "*Express train to Polynesia*." Nature **336**: 307-308.

Diamond, J. (2001). "*Reply to Oppenheimer and Richards*." Nature **410**: 167.

Diamond, J. and P. Bellwood (2003). "*Farmers and their languages: the first expansions*." Science **300**(5619): 597-603.

Dixon, L. A., C. M. Murray, E. J. Archer, A. E. Dobbins, P. Koumi and P. Gill (2005). "*Validation of a 21-locus autosomal SNP multiplex for forensic identification purposes*." Forensic Sci Int **154**(1): 62-77.

Doi, K., T. Monjo, P. H. Hoang, J. Yoshimura, H. Yurino, J. Mitsui, H. Ishiura, Y. Takahashi, Y. Ichikawa, J. Goto, et al. (2014). "*Rapid detection of expanded short tandem repeats in personal genomics using hybrid sequencing*." Bioinformatics **30**(6): 815-822.

Dong, C., Y. Yang, J. Yan, L. Fu, X. Zhang, B. Cong and S. Li (2015). "*Evaluation of the protective capabilities of nucleosome STRs obtained by large-scale sequencing*." Electrophoresis.

Drmanac, R., A. B. Sparks, M. J. Callow, A. L. Halpern, N. L. Burns, B. G. Kermani, P. Carnevali, I. Nazarenko, G. B. Nilsen, G. Yeung, et al. (2010). "*Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays*." Science **327**(5961): 78-81.

Duggan, A. T., B. Evans, F. R. Friedlaender, J. S. Friedlaender, G. Koki, D. A. Merriwether, M. Kayser and M. Stoneking (2014). "*Maternal history of Oceania from complete mtDNA genomes: contrasting ancient diversity with recent homogenization due to the Austronesian expansion*." Am J Hum Genet **94**(5): 721-733.

Earl, D. A. and B. M. vonHoldt (2012). "*STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method*." Conserv Genet Resour **4 (2)**(2): 359-361.

Ebersberger, I., D. Metzler, C. Schwarz and S. Pääbo (2002). "*Genomewide Comparison of DNA Sequences between Humans and Chimpanzees*." Am J Hum Genet **70**(6): 1490-1497.

Eckert, K. A. and S. E. Hile (2009). "*Every microsatellite is different: Intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome*." Mol Carcinog **48**(4): 379-388.

Edelmann, J., S. Hering, C. Augustin and R. Szibor (2008). "*Characterisation of the STR markers DXS10146, DXS10134 and DXS10147 located within a 79.1 kb region at Xq28*." Forensic Sci Int Genet **2**(1): 41-46.

Edelmann, J., S. Hering, C. Augustin and R. Szibor (2009). "*Indel polymorphisms - An additional set of markers on the X-chromosome*." Forensic Sci Int Genet Suppl Ser **2**(1): 510-512.

Edelmann, J., R. Lessig, M. Klintschar and R. Szibor (2004). "*Advantages of X-chromosomal microsatellites in deficiency paternity testing: presentation of cases*." Int Congr Ser **1261**(0): 257-259.

Edwards, A. and C. T. Caskey (1991). "*Closure strategies for random DNA sequencing*." Methods **3**(1): 41-47.

Edwards, A., A. Civitello, H. A. Hammond and C. T. Caskey (1991). "*DNA typing and genetic mapping with trimeric and tetrameric tandem repeats*." Am J Hum Genet **49**(4): 746-756.

Edwards, A., H. Voss, P. Rice, A. Civitello, J. Stegemann, C. Schwager, J. Zimmermann, H. Erfle, C. T. Caskey and W. Ansorge (1990). "*Automated DNA sequencing of the human HPRT locus*." Genomics **6**(4): 593-608.

Egeland, T., H. M. Bovelstad, G. O. Storvik and A. Salas (2004). "*Inferring the most likely geographical origin of mtDNA sequence profiles*." Ann Hum Genet **68**(Pt 5): 461-471.

Eiberg, H., J. Troelsen, M. Nielsen, A. Mikkelsen, J. Mengel-From, K. W. Kjaer and L. Hansen (2008). "*Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression*." Hum Genet **123**(2): 177-187.

Ellegren, H. (2000). "*Heterogeneous mutation processes in human microsatellite DNA sequences*." Nat Genet **24**(4): 400-402.

Ellegren, H. (2004). "*Microsatellites: simple sequences with complex evolution*." Nat Rev Genet **5**(6): 435-445.

Elsharawy, A., M. Forster, N. Schracke, A. Keller, I. Thomsen, B. S. Petersen, B. Stade, P. Stahler, S. Schreiber, P. Rosenstiel, et al. (2012). "*Improving mapping and SNP-calling performance in multiplexed targeted next-generation sequencing*." BMC Genomics **13**: 417.

Enoch, M. A., P. H. Shen, K. Xu, C. Hodgkinson and D. Goldman (2006). "*Using ancestry-informative markers to define populations and detect population stratification*." J Psychopharmacol **20**(4 Suppl): 19-26.

EPC (2012). "*An integrated encyclopedia of DNA elements in the human genome*." Nature **489**(7414): 57-74.

Esteve Codina, A., H. Niederstatter and W. Parson (2009). "*"GenderPlex" a PCR multiplex for reliable gender determination of degraded human DNA samples and complex gender constellations*." Int J Legal Med **123**(6): 459-464.

Estrada, K., M. Krawczak, S. Schreiber, K. van Duijn, L. Stolk, J. B. van Meurs, F. Liu, B. W. Penninx, J. H. Smit, N. Vogelzangs, et al. (2009). "*A genome-wide association study of northwestern Europeans involves the C-type natriuretic peptide signaling*

pathway in the etiology of human height variation." Hum Mol Genet **18**(18): 3516-3524.

Evanno, G., S. Regnaut and J. Goudet (2005). "*Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study*." Mol Ecol **14**(8): 2611-2620.

Excoffier, L. and H. E. Lischer (2010). "*Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows*." Mol Ecol Resour **10**(3): 564-567.

Falush, D., M. Stephens and J. K. Pritchard (2003). "*Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies*." Genetics **164**(4): 1567-1587.

Falush, D., M. Stephens and J. K. Pritchard (2007). "*Inference of population structure using multilocus genotype data: dominant markers and null alleles*." Mol Ecol Notes **7**(4): 574-578.

Fan, H. and J. Y. Chu (2007). "*A brief review of short tandem repeat mutation*." Genomics Proteomics Bioinformatics **5**(1): 7-14.

Feathers, J., R. Kipnis, L. Piló, M. Arroyo-Kalin and D. Coblentz (2010). "*How old is Luzia? Luminescence dating and stratigraphic integrity at Lapa Vermelha, Lagoa Santa, Brazil*." Geoarchaeology **25**(4): 395-436.

Ferri, G., C. Bini, S. Ceccardi and S. Pelotti (2004). "*Successful identification of two years old menstrual bloodstain by using MMP-11 shorter amplicons*." J Forensic Sci **49**(6): 1387.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, et al. (1995). "*Whole-genome random sequencing and assembly of Haemophilus influenzae Rd*." Science **269**(5223): 496-512.

Fleming, R. I. and S. Harbison (2010). "*The development of a mRNA multiplex RT-PCR assay for the definitive identification of body fluids*." Forensic Sci Int Genet **4**(4): 244-256.

Fondevila, M., C. Phillips, N. Naverán, M. Cerezo, A. Rodríguez, R. Calvo, L. M. Fernández, Á. Carracedo and M. V. Lareu (2008a). "*Challenging DNA: Assessment of a range of genotyping approaches for highly degraded forensic samples*." Forensic Sci Int Genet Suppl Ser **1**(1): 26-28.

Fondevila, M., C. Phillips, N. Naveran, L. Fernandez, M. Cerezo, A. Salas, A. Carracedo and M. V. Lareu (2008b). "*Case report: identification of skeletal remains using short-amplicon marker analysis of severely degraded DNA extracted from a decomposed and charred femur*." Forensic Sci Int Genet **2**(3): 212-218.

Fondevila, M., C. Phillips, C. Santos, A. Freire-Aradas, P. M. Vallone, J. M. Butler, M. V. Lareu and Á. Carracedo (2013). "*Revision of the SNPforID 34-plex forensic ancestry test: Assay enhancements, standard reference sample genotypes and extended population studies*." Forensic Sci Int Genet **7**(1): 63-74.

Fordyce, S. L., M. C. Avila-Arcos, E. Rockenbauer, C. Borsting, R. Frank-Hansen, F. T. Petersen, E. Willerslev, A. J. Hansen, N. Morling and M. T. Gilbert (2011). "*High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform*." Biotechniques **51**(2): 127-133.

Fordyce, S. L., H. S. Mogensen, C. Borsting, R. E. Lagace, C. W. Chang, N. Rajagopalan and N. Morling (2015). "*Second-generation sequencing of forensic STRs using the Ion Torrent HID STR 10-plex and the Ion PGM*." Forensic Sci Int Genet **14**: 132-140.

Forster, L., J. Thomson and S. Kutranov (2008). "*Direct comparison of post-28-cycle PCR purification and modified capillary electrophoresis methods with the 34-cycle "low copy number" (LCN) method for analysis of trace forensic DNA samples*." Forensic Sci Int Genet **2**(4): 318-328.

Freire-Aradas, A., M. Fondevila, A. K. Kriegel, C. Phillips, P. Gill, L. Prieto, P. M. Schneider, Á. Carracedo and M. V. Lareu (2012). "*A new SNP assay for identification of highly degraded human DNA*." Forensic Sci Int Genet **6**(3): 341-349.

Friedlaender, J. S., F. R. Friedlaender, F. A. Reed, K. K. Kidd, J. R. Kidd, G. K. Chambers, R. A. Lea, J. H. Loo, G. Koki, J. A. Hodgson, et al. (2008). "*The genetic structure of Pacific Islanders*." PLoS Genet **4**(1): e19.

Frumkin, D., A. Wasserstrom, B. Budowle and A. Davidson (2011). "*DNA methylation-based forensic tissue identification*." Forensic Sci Int Genet **5**(5): 517-524.

Fu, Q., M. Meyer, X. Gao, U. Stenzel, H. A. Burbano, J. Kelso and S. Pääbo (2013). "*DNA analysis of an early modern human from Tianyuan Cave, China*." Proc Natl Acad Sci U S A **110**(6): 2223-2227.

Fujimoto, A., R. Kimura, J. Ohashi, K. Omi, R. Yuliwulandari, L. Batubara, M. S. Mustofa, U. Samakkarn, W. Settheetham-Ishida, T. Ishida, et al. (2008a). "*A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness*." Hum Mol Genet **17**(6): 835-843.

Fujimoto, A., J. Ohashi, N. Nishida, T. Miyagawa, Y. Morishita, T. Tsunoda, R. Kimura and K. Tokunaga (2008b). "*A replication study confirmed the EDAR gene to be a major contributor to population differentiation regarding head hair thickness in Asia*." Hum Genet **124**(2): 179-185.

Galanter, J. M., J. C. Fernandez-Lopez, C. R. Gignoux, J. Barnholtz-Sloan, C. Fernandez-Rozadilla, M. Via, A. Hidalgo-Miranda, A. V. Contreras, L. U. Figueroa, P. Raska, et al. (2012). "*Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas*." PLoS Genet **8**(3): e1002554.

Garcia-Diaz, M. and T. A. Kunkel (2006). "*Mechanism of a genetic glissando: structural biology of indel mutations*." Trends Biochem Sci **31**(4): 206-214.

Gelardi, C., E. Rockenbauer, S. Dalsgaard, C. Borsting and N. Morling (2014). "*Second generation sequencing of three STRs D3S1358, D12S391 and D21S11 in Danes and a new nomenclature for sequenced STR alleles*." Forensic Sci Int Genet **12**: 38-41.

Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, et al. (2004). "*Bioconductor: open software development for computational biology and bioinformatics*." Genome Biol **5**(10): R80.

Gill, P. (2001a). "*Application of low copy number DNA profiling*." Croat Med J **42**(3): 229-232.

Gill, P. (2001b). "*An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes*." Int J Legal Med **114**(4-5): 204-210.

Gill, P., A. J. Jeffreys and D. J. Werrett (1985). "*Forensic application of DNA 'fingerprints'*." Nature **318**(6046): 577-579.

Gill, P., C. Phillips, C. McGovern, J. A. Bright and J. Buckleton (2012). "*An evaluation of potential allelic association between the STRs vWA and D12S391: implications in criminal casework and applications to short pedigrees*." Forensic Sci Int Genet **6**(4): 477-486.

Gill, P., D. J. Werrett, B. Budowle and R. Guerrieri (2004). "*An assessment of whether SNPs will replace STRs in national DNA databases*." Science & Justice **44**(1): 51-53.

Gill, P., J. Whitaker, C. Flaxman, N. Brown and J. Buckleton (2000). "*An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA*." Forensic Sci Int **112**(1): 17-40.

Goebel, T. (1999). "*Pleistocene human colonization of Siberia and peopling of the Americas: An ecological approach*." Evolutionary Anthropology: Issues, News, and Reviews **8**(6): 208-227.

Goebel, T., M. R. Waters and D. H. O'Rourke (2008). "*The late Pleistocene dispersal of modern humans in the Americas*." Science **319**(5869): 1497-1502.

Goldstein, D. B., A. Ruiz Linares, L. L. Cavalli-Sforza and M. W. Feldman (1995a). "*An evaluation of genetic distances for use with microsatellite loci*." Genetics **139**(1): 463-471.

Goldstein, D. B., A. Ruiz Linares, L. L. Cavalli-Sforza and M. W. Feldman (1995b). "*Genetic absolute dating based on microsatellites and the origin of modern humans*." Proc Natl Acad Sci U S A **92**(15): 6723-6727.

Gonzalez, J. R., L. Armengol, X. Sole, E. Guino, J. M. Mercader, X. Estivill and V. Moreno (2007). "*SNPassoc: an R package to perform whole genome association studies*." Bioinformatics **23**(5): 644-645.

Goodman, M., C. A. Porter, J. Czelusniak, S. L. Page, H. Schneider, J. Shoshani, G. Gunnell and C. P. Groves (1998). "*Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence*." Mol Phylogenet Evol **9**(3): 585-598.

Gray, R. D., Q. D. Atkinson and S. J. Greenhill (2011). "*Language evolution and human history: what a difference a date makes*." Philos Trans R Soc Lond B Biol Sci **366**(1567): 1090-1100.

Gray, R. D., A. J. Drummond and S. J. Greenhill (2009). "*Language phylogenies reveal expansion pulses and pauses in Pacific settlement*." Science **323**(5913): 479-483.

Gray, R. D. and F. M. Jordan (2000). "*Language trees support the express-train sequence of Austronesian expansion*." Nature **405**(6790): 1052-1055.

Green, R. C. (1991). *The Lapita Cultural Complex: Current evidence and proposed models*. Indo-Pacific Prehistory 1990 : Proceedings of the 14th Congress of the Indo-Pacific prehistory Association. P. Bellwood. Canberra and Jakarta, Bulletin of the Indo-Pacific Prehistory Association**:** 295-305.

Green, R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H. Fritz, et al. (2010). "*A draft sequence of the Neandertal genome*." Science **328**(5979): 710-722.

Green, T. J., B. Cochran, T. W. Fenton, J. C. Woods, G. L. Titmus, L. Tieszen, M. A. Davis and S. J. Miller (1998). "*The Buhl burial: a paleoindian woman from southern Idaho*." American Antiquity **63**(3): 437-456.

Griffiths, A. J. F., S. R. Wessler, R. C. Lewontin, W. M. Gelbart, D. T. Suzuki and J. H. Miller (2005). *Introduction to Genetic Analysis*, W. H. Freeman and Company.

Grimes, E. A., P. J. Noake, L. Dixon and A. Urquhart (2001). "*Sequence polymorphism in the human melanocortin 1 receptor gene as an indicator of the red hair phenotype*." Forensic Sci Int **122**(2): 124-129.

Grün, R., C. Stringer, F. McDermott, R. Nathan, N. Porat, S. Robertson, L. Taylor, G. Mortimer, S. Eggins and M. McCulloch (2005). "*U-series and ESR analyses of bones and teeth relating to the human burials from Skhul*." J Hum Evol **49**(3): 316-334.

Guillen, M., M. V. Lareu, C. Pestoni, A. Salas and A. Carracedo (2000). "*Ethical-legal problems of DNA databases in criminal investigation*." J Med Ethics **26**(4): 266-271.

Guillot, G. and M. Foll (2009). "*Correcting for ascertainment bias in the inference of population structure*." Bioinformatics **25**(4): 552-554.

Haas, C., B. Klesser, C. Maake, W. Bar and A. Kratzer (2009). "*mRNA profiling for body fluid identification by reverse transcription endpoint PCR and realtime PCR*." Forensic Sci Int Genet **3**(2): 80-88.

Halder, I., M. Shriver, M. Thomas, J. R. Fernandez and T. Frudakis (2008). "*A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: Utility and applications*." Hum Mutat **29**(5): 648-658.

Han, J., P. Kraft, H. Nan, Q. Guo, C. Chen, A. Qureshi, S. E. Hankinson, F. B. Hu, D. L. Duffy, Z. Z. Zhao, et al. (2008). "*A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation*." PLoS Genet **4**(5): e1000074.

Hardt, T. and W. Henke (2006). "*Facial patterns in Cercopithecoidea and Hominoidea: a geometric approach*." Anthropol Anz **64**(2): 129-146.

Hardy, G. H. (1908). "*Mendelian Proportions in a Mixed Population*." Science **28**(706): 49-50.

He, J. D., M. S. Peng, H. H. Quang, K. P. Dang, A. V. Trieu, S. F. Wu, J. Q. Jin, R. W. Murphy, Y. G. Yao and Y. P. Zhang (2012). "*Patrilineal perspective on the Austronesian diffusion in Mainland Southeast Asia*." PLoS One **7**(5): e36437.

Helmuth, R., N. Fildes, E. Blake, M. C. Luce, J. Chimera, R. Madej, C. Gorodezky, M. Stoneking, N. Schmill, W. Klitz, et al. (1990). "*HLA-DQ alpha allele and genotype frequencies in various human populations, determined by using enzymatic amplification and oligonucleotide probes*." Am J Hum Genet **47**(3): 515-523.

Hering, S., C. Augustin, J. Edelmann, M. Heidel, J. Dressler, H. Rodig, E. Kuhlisch and R. Szibor (2006). "*DXS10079, DXS10074 and DXS10075 are STRs located within a 280-kb region of Xq12 and provide stable haplotypes useful for complex kinship cases*." Int J Legal Med **120**(6): 337-345.

Hernandez, R. D., J. L. Kelley, E. Elyashiv, S. C. Melton, A. Auton, G. McVean, P. Genomes, G. Sella and M. Przeworski (2011). "*Classic selective sweeps were rare in recent human evolution*." Science **331**(6019): 920-924.

Hershey, A. D. and M. Chase (1952). "*Independent functions of viral protein and nucleic acid in growth of bacteriophage*." J Gen Physiol **36**(1): 39-56.

Hill, C., P. Soares, M. Mormina, V. Macaulay, D. Clarke, P. B. Blumbach, M. Vizuete-Forster, P. Forster, D. Bulbeck, S. Oppenheimer, et al. (2007). "*A mitochondrial stratigraphy for island southeast Asia*." Am J Hum Genet **80**(1): 29-43.

Hill, C., P. Soares, M. Mormina, V. Macaulay, W. Meehan, J. Blackburn, D. Clarke, J. M. Raja, P. Ismail, D. Bulbeck, et al. (2006). "*Phylogeography and ethnogenesis of aboriginal Southeast Asians*." Mol Biol Evol **23**(12): 2480-2491.

Hill, C. R., J. M. Butler and P. M. Vallone (2009). "*A 26plex autosomal STR assay to aid human identity testing*." J Forensic Sci **54**(5): 1008-1015.

Hofer, T., N. Ray, D. Wegmann and L. Excoffier (2009). "*Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection*." Ann Hum Genet **73**(1): 95-108.

Homer, N., M. Lyons and M. Shah (2014). *Torrent Mapping Alignment Program*, GitHub, Inc., San Francisco, CA, Available from: https://github.com/iontorrent/TMAP (accessed 27.06.14).

HPAC, M. A. Abdulla, I. Ahmed, A. Assawamakin, J. Bhak, S. K. Brahmachari, G. C. Calacal, A. Chaurasia, C. H. Chen, J. Chen, et al. (2009). "*Mapping human genetic diversity in Asia*." Science **326**(5959): 1541-1545.

Hubisz, M. J., D. Falush, M. Stephens and J. K. Pritchard (2009). "*Inferring weak population structure with the assistance of sample group information*." Mol Ecol Resour **9**(5): 1322-1332.

Hundertmark, T., S. Hering, J. Edelmann, C. Augustin, I. Plate and R. Szibor (2008). "*The STR cluster DXS10148-DXS8378-DXS10135 provides a powerful tool for X-chromosomal haplotyping at Xp22*." Int J Legal Med **122**(6): 489-492.

IHGSC (2001). "*Initial sequencing and analysis of the human genome*." Nature **409**(6822): 860-921.

IHMC (2003). "*The International HapMap Project*." Nature **426**(6968): 789-796.

IHMC (2005). "*A haplotype map of the human genome*." Nature **437**(7063): 1299-1320.

IHMC, D. M. Altshuler, R. A. Gibbs, L. Peltonen, D. M. Altshuler, R. A. Gibbs, L. Peltonen, E. Dermitzakis, S. F. Schaffner, F. Yu, et al. (2010). "*Integrating common and rare genetic variation in diverse human populations*." Nature **467**(7311): 52-58.

IHMC, K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, et al. (2007). "*A second generation human haplotype map of over 3.1 million SNPs*." Nature **449**(7164): 851-861.

Inagaki, S., Y. Yamamoto, Y. Doi, T. Tomoyo, I. Takaki, I. Kiyomi, Y. Kei, M. Satoru and I. Hideo (2004). "*A new 39-plex analysis method for SNPs including 15 blood group loci*." Forensic Sci Int **144**(1): 45-57.

Ingman, M., H. Kaessmann, S. Paabo and U. Gyllensten (2000). "*Mitochondrial genome variation and the origin of modern humans*." Nature **408**(6813): 708-713.

Innan, H. and M. Nordborg (2002). "*Recombination or Mutational Hot Spots in Human mtDNA?*" Mol Biol Evol **19**(7): 1122-1127.

Jakobsson, M. and N. A. Rosenberg (2007). "*CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure*." Bioinformatics **23**(14): 1801-1806.

Jeffreys, A. J., V. Wilson and S. L. Thein (1985a). "*Hypervariable 'minisatellite' regions in human DNA*." Nature **314**(6006): 67-73.

Jeffreys, A. J., V. Wilson and S. L. Thein (1985b). "*Individual-specific 'fingerprints' of human DNA*." Nature **316**(6023): 76-79.

Jinam, T. A., L. C. Hong, M. E. Phipps, M. Stoneking, M. Ameen, J. Edo, H. P.-A. S. Consortium and N. Saitou (2012). "*Evolutionary history of continental southeast Asians: "early train" hypothesis based on genetic analysis of mitochondrial and autosomal DNA data*." Mol Biol Evol **29**(11): 3513-3527.

Jobling, M., E. Hollox, M. Hurles, T. Kivisild and C. Tyler-Smith (2014). *Human Evolutionary Genetics: Origins, Peoples & Disease* NY, Garland Science - Taylor & Francis Group.

Jobling, M. A. and P. Gill (2004). "*Encoded evidence: DNA in forensic analysis*." Nat Rev Genet **5**(10): 739-751.

Jobling, M. A., I. C. Lo, D. J. Turner, G. R. Bowden, A. C. Lee, Y. Xue, D. Carvalho-Silva, M. E. Hurles, S. M. Adams, Y. M. Chang, et al. (2007). "*Structural variation on the short arm of the human Y chromosome: recurrent multigene deletions encompassing Amelogenin Y*." Hum Mol Genet **16**(3): 307-316.

Jobling, M. A. and C. Tyler-Smith (2003). "*The human Y chromosome: an evolutionary marker comes of age*." Nat Rev Genet **4**(8): 598-612.

Johansen, P., J. D. Andersen, C. Borsting and N. Morling (2013). "*Evaluation of the iPLEX(R) Sample ID Plus Panel designed for the Sequenom MassARRAY(R) system. A SNP typing assay developed for human identification and sample tracking based on the SNPforID panel*." Forensic Sci Int Genet **7**(5): 482-487.

Jolliffe, I. (2002). *Principal Component Analysis*. New York, Springer Verlag.

Jorde, L. B., W. S. Watkins, M. J. Bamshad, M. E. Dixon, C. E. Ricker, M. T. Seielstad and M. A. Batzer (2000). "*The distribution of human genetic diversity: A comparison of mitochondrial, autosomal, and Y-chromosome data*." Am J Hum Genet **66**(3): 979-988.

Juusola, J. and J. Ballantyne (2005). "*Multiplex mRNA profiling for the identification of body fluids*." Forensic Sci Int **152**(1): 1-12.

Juusola, J. and J. Ballantyne (2007). "*mRNA profiling for body fluid identification by multiplex quantitative RT-PCR*." J Forensic Sci **52**(6): 1252-1262.

Kalinowski, S. T. (2011). "*The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure*." Heredity **106**(4): 625-632.

Karafet, T. M., B. Hallmark, M. P. Cox, H. Sudoyo, S. Downey, J. S. Lansing and M. F. Hammer (2010). "*Major east-west division underlies Y chromosome stratification across Indonesia*." Mol Biol Evol **27**(8): 1833-1844.

Kasai, K., Y. Nakamura and R. White (1990). "*Amplification of a variable number of tandem repeats (VNTR) locus (pMCT118) by the polymerase chain reaction (PCR) and its application to forensic science*." J Forensic Sci **35**(5): 1196-1200.

Kastelic, V., B. Budowle and K. Drobnic (2009). "*Validation of SRY marker for forensic casework analysis*." J Forensic Sci **54**(3): 551-555.

Kaye, D. H. and M. E. Smith (2003). "*DNA identification databases: legality, legitimacy, and the case for population-wide coverage*." Wisconsin Law Review **3**: 413-459.

Kayser, M. (2010). "*The human genetic history of Oceania: near and remote views of dispersal*." Curr Biol **20**(4): R194-201.

Kayser, M. (2015). "*Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes*." Forensic Sci Int Genet **18**: 33-48.

Kayser, M., S. Brauer, R. Cordaux, A. Casto, O. Lao, L. A. Zhivotovsky, C. Moyse-Faurie, R. B. Rutledge, W. Schiefenhoevel, D. Gil, et al. (2006). "*Melanesian and Asian Origins of Polynesians: mtDNA and Y Chromosome Gradients Across the Pacific*." Mol Biol Evol **23**(11): 2234-2244.

Kayser, M., S. Brauer, G. Weiss, P. Underhill, L. Roewer, W. Schiefenhövel and M. Stoneking (2000a). "*Melanesian origin of Polynesian Y chromosomes*." Curr Biol **10**(20): 1237-1246.

Kayser, M., A. Caglia, D. Corach, N. Fretwell, C. Gehrig, G. Graziosi, F. Heidorn, S. Herrmann, B. Herzog, M. Hidding, et al. (1997). "*Evaluation of Y-chromosomal STRs: a multicenter study*." Int J Legal Med **110**(3): 125-133, 141-129.

Kayser, M., Y. Choi, M. van Oven, S. Mona, S. Brauer, R. J. Trent, D. Suarkia, W. Schiefenhovel and M. Stoneking (2008a). "*The impact of the Austronesian expansion: evidence from mtDNA and Y chromosome diversity in the Admiralty Islands of Melanesia*." Mol Biol Evol **25**(7): 1362-1374.

Kayser, M. and P. de Knijff (2011). "*Improving human forensics through advances in genetics, genomics and molecular biology*." Nat Rev Genet **12**(3): 179-192.

Kayser, M., O. Lao, K. Saar, S. Brauer, X. Wang, P. Nurnberg, R. J. Trent and M. Stoneking (2008b). "*Genome-wide analysis indicates more Asian than Melanesian ancestry of Polynesians*." Am J Hum Genet **82**(1): 194-198.

Kayser, M., F. Liu, A. C. Janssens, F. Rivadeneira, O. Lao, K. van Duijn, M. Vermeulen, P. Arp, M. M. Jhamai, W. F. van Ijcken, et al. (2008c). "*Three genome-wide association studies and a linkage analysis identify HERC2 as a human iris color gene*." Am J Hum Genet **82**(2): 411-423.

Kayser, M., L. Roewer, M. Hedman, L. Henke, J. Henke, S. Brauer, C. Kruger, M. Krawczak, M. Nagy, T. Dobosz, et al. (2000b). "*Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs*." Am J Hum Genet **66**(5): 1580-1588.

Kayser, M. and P. M. Schneider (2009). "*DNA-based prediction of human externally visible characteristics in forensics: Motivations, scientific challenges, and ethical considerations*." Forensic Sci Int Genet **3**(3): 154-161.

Kersbergen, P., K. van Duijn, A. Kloosterman, J. den Dunnen, M. Kayser and P. de Knijff (2009). "*Developing a set of ancestry-sensitive DNA markers reflecting continental origins of humans*." BMC Genet **10**(1): 69.

Kidd, J. R., F. R. Friedlaender, W. C. Speed, A. J. Pakstis, F. M. De La Vega and K. K. Kidd (2011). "*Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples*." Investig Genet **2**(1): 1.

Kidd, K. K., W. C. Speed, A. J. Pakstis, M. R. Furtado, R. Fang, A. Madbouly, M. Maiers, M. Middha, F. R. Friedlaender and J. R. Kidd (2014). "*Progress toward an efficient panel of SNPs for ancestry inference*." Forensic Sci Int Genet **10**(0): 23-32.

Kim, R. N., A. Kim, D. W. Kim, S. H. Choi, D. S. Kim, S. H. Nam, A. Kang, M. Y. Kim, K. H. Park, B. H. Yoon, et al. (2012). "*Analysis of indel variations in the human disease-associated genes CDKN2AIP, WDR66, USP20 and OR7C2 in a Korean population*." J Genet **91**(1): e1-e11.

Kimbel, W. H. and L. K. Delezene (2009). "*"Lucy" redux: a review of research on Australopithecus afarensis*." Am J Phys Anthropol **140 Suppl 49**: 2-48.

Kimpton, C. P., P. Gill, A. Walton, A. Urquhart, E. S. Millican and M. Adams (1993). "*Automated DNA profiling employing multiplex amplification of short tandem repeat loci*." PCR Methods Appl **3**(1): 13-22.

Kimura, M. and T. Ohta (1973). "*Mutation and evolution at the molecular level*." Genetics **73**: Suppl 73:19-35.

King, T. E., E. J. Parkin, G. Swinfield, F. Cruciani, R. Scozzari, A. Rosa, S.-K. Lim, Y. Xue, C. Tyler-Smith and M. A. Jobling (2007). "*Africans in Yorkshire? - the deepest-rooting clade of the Y phylogeny within an English genealogy*." Eur J Hum Genet **15**(3): 288-293.

King, W. (1864). "*The reputed fossil man of the Neanderthal*." Quarterly Review of Science **1**: 88-97.

Knapp, M., K. A. Horsburgh, S. Prost, J. A. Stanton, H. R. Buckley, R. K. Walter and E. A. Matisoo-Smith (2012). "*Complete mitochondrial DNA genome sequences from the first New Zealanders*." Proc Natl Acad Sci U S A **109**(45): 18350-18354.

Koboldt, D. C., L. Ding, E. R. Mardis and R. K. Wilson (2010). "*Challenges of sequencing human genomes*." Brief Bioinform **11**(5): 484-498.

Kong, A., M. L. Frigge, G. Masson, S. Besenbacher, P. Sulem, G. Magnusson, S. A. Gudjonsson, A. Sigurdsson, A. Jonasdottir, A. Jonasdottir, et al. (2012). "*Rate of de novo mutations and the importance of father's age to disease risk*." Nature **488**(7412): 471-475.

Kosoy, R., R. Nassir, C. Tian, P. A. White, L. M. Butler, G. Silva, R. Kittles, M. E. Alarcon-Riquelme, P. K. Gregersen, J. W. Belmont, et al. (2009). "*Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America*." Hum Mutat **30**(1): 69-78.

Krause, J., Q. Fu, J. M. Good, B. Viola, M. V. Shunkov, A. P. Derevianko and S. Paabo (2010). "*The complete mitochondrial DNA genome of an unknown hominin from southern Siberia*." Nature **464**(7290): 894-897.

Krawczak, M. (1999). "*Informativity assessment for biallelic single nucleotide polymorphisms*." Electrophoresis **20**(8): 1676-1681.

Kruglyak, S., R. T. Durrett, M. D. Schug and C. F. Aquadro (1998). "*Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations*." Proc Natl Acad Sci U S A **95**(18): 10774-10778.

La Spada, A. R., E. M. Wilson, D. B. Lubahn, A. E. Harding and K. H. Fischbeck (1991). "*Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy*." Nature **352**(6330): 77-79.

Lai, Y. and F. Sun (2003). "*The relationship between microsatellite slippage mutation rate and the number of repeat units*." Mol Biol Evol **20**(12): 2123-2131.

Landsteiner, K. (1900). "*Zur Kenntnis der antifermentativen, lytisichen und agglutinierenden Wirkungen des Blutserums und der Lymphe*." Zentralbl Bakteriol **27**: 57-362.

Landsteiner, K. and P. Levine (1927). "*A New Agglutinable Factor Differentiating Individual Human Bloods*." Exp Biol Med **24**(6): 600-602.

Landsteiner, K. and A. S. Wiener (1940). "*An Agglutinable Factor in Human Blood Recognized by Immune Sera for Rhesus Blood*." Exp Biol Med **43**(1): 223.

Langergraber, K. E., K. Prufer, C. Rowney, C. Boesch, C. Crockford, K. Fawcett, E. Inoue, M. Inoue-Muruyama, J. C. Mitani, M. N. Muller, et al. (2012). "*Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution*." Proc Natl Acad Sci U S A **109**(39): 15716-15721.

Lango Allen, H., K. Estrada, G. Lettre, S. I. Berndt, M. N. Weedon, F. Rivadeneira, C. J. Willer, A. U. Jackson, S. Vedantam, S. Raychaudhuri, et al. (2010). "*Hundreds of variants clustered in genomic loci and biological pathways affect human height*." Nature **467**(7317): 832-838.

Lao, O., J. M. de Gruijter, K. van Duijn, A. Navarro and M. Kayser (2007). "*Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms*." Ann Hum Genet **71**(Pt 3): 354-369.

Lao, O., P. M. Vallone, M. D. Coble, T. M. Diegoli, M. van Oven, K. J. van der Gaag, J. Pijpe, P. de Knijff and M. Kayser (2010). "*Evaluating self-declared ancestry of U.S. Americans with autosomal, Y-chromosomal and mitochondrial DNA*." Hum Mutat **31**(12): E1875-1893.

Lao, O., K. van Duijn, P. Kersbergen, P. de Knijff and M. Kayser (2006). "*Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry*." Am J Hum Genet **78**(4): 680-690.

Lattanzi, W., M. C. Di Giacomo, G. M. Lenato, G. Chimienti, G. Voglino, N. Resta, G. Pepe and G. Guanti (2005). "*A large interstitial deletion encompassing the amelogenin gene on the short arm of the Y chromosome*." Hum Genet **116**(5): 395-401.

Lepre, C. J., H. Roche, D. V. Kent, S. Harmand, R. L. Quinn, J. P. Brugal, P. J. Texier, A. Lenoble and C. S. Feibel (2011). "*An earlier origin for the Acheulian*." Nature **477**(7362): 82-85.

Levine, P. and R. E. Stetson (1939). "*An unusual case of intra-group agglutination*." JAMA **113**(2): 126-127.

Lewis, R. (2001). *Human Genetics: concepts and applications*, McGraw-Hill International Edition.

Lewontin, R. C. (1972). "*The apportionment of human diversity*." Evol Biol **6**: 381-398.

Li, C., S. Zhao, S. Zhang, L. Li, Y. Liu, J. Chen and J. Xue (2011). "*Genetic polymorphism of 29 highly informative InDel markers for forensic use in the Chinese Han population*." Forensic Sci Int Genet **5**(1): e27-30.

Li, H., K. Cho, J. R. Kidd and K. K. Kidd (2009). "*Genetic landscape of Eurasia and "admixture" in Uyghurs*." Am J Hum Genet **85**(6): 934-937; author reply 937-939.

Li, J. Z., D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S. Ramachandran, H. M. Cann, G. S. Barsh, M. Feldman, L. L. Cavalli-Sforza, et al. (2008). "*Worldwide human relationships inferred from genome-wide patterns of variation*." Science **319**(5866): 1100-1104.

Lieberman, D. E., B. M. McBratney and G. Krovitz (2002). "*The evolution and development of cranial form in Homo sapiens*." Proc Natl Acad Sci U S A **99**(3): 1134-1139.

Lipson, M., P. R. Loh, N. Patterson, P. Moorjani, Y. C. Ko, M. Stoneking, B. Berger and D. Reich (2014). "*Reconstructing Austronesian population history in Island Southeast Asia*." Nat Commun **5**: 4689.

Liu, B., J. R. Lague, D. P. Nunes, P. Toselli, F. G. Oppenheim, R. V. Soares, R. F. Troxler and G. D. Offner (2002). "*Expression of membrane-associated mucins MUC1 and MUC4 in major human salivary glands*." J Histochem Cytochem **50**(6): 811-820.

Liu, F., F. van der Lijn, C. Schurmann, G. Zhu, M. M. Chakravarty, P. G. Hysi, A. Wollstein, O. Lao, M. de Bruijne, M. A. Ikram, et al. (2012a). "*A Genome-Wide Association Study Identifies Five Loci Influencing Facial Morphology in Europeans*." PLoS Genet **8**(9): e1002932.

Liu, F., K. van Duijn, J. R. Vingerling, A. Hofman, A. G. Uitterlinden, A. C. Janssens and M. Kayser (2009). "*Eye color and the prediction of complex phenotypes from genotypes*." Curr Biol **19**(5): R192-193.

Liu, F., A. Wollstein, P. G. Hysi, G. A. Ankra-Badu, T. D. Spector, D. Park, G. Zhu, M. Larsson, D. L. Duffy, G. W. Montgomery, et al. (2010a). "*Digital quantification of human eye color highlights genetic association of three new loci*." PLoS Genet **6**(5): e1000934.

Liu, L., Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu and M. Law (2012b). "*Comparison of next-generation sequencing systems*." J Biomed Biotechnol **2012**: 251364.

Liu, W., C. Z. Jin, Y. Q. Zhang, Y. J. Cai, S. Xing, X. J. Wu, H. Cheng, R. L. Edwards, W. S. Pan, D. G. Qin, et al. (2010b). "*Human remains from Zhirendong, South China, and modern human emergence in East Asia*." Proc Natl Acad Sci U S A **107**(45): 19201-19206.

Liu, X. and S. Harada (2013). "*DNA Isolation from Mammalian Samples*." Curr Protoc Mol Bio **102**: 2.14.11–12.14.13.

Loman, N. J., R. V. Misra, T. J. Dallman, C. Constantinidou, S. E. Gharbia, J. Wain and M. J. Pallen (2012). "*Performance comparison of benchtop high-throughput sequencing platforms*." Nat Biotech **30**(5): 434-439.

Londin, E. R., M. A. Keller, C. Maista, G. Smith, L. A. Mamounas, R. Zhang, S. J. Madore, K. Gwinn and R. A. Corriveau (2010). "*CoAIMs: a cost-effective panel of ancestry informative markers for determining continental origins*." PLoS One **5**(10): e13443.

Lowe, A. L., A. Urquhart, L. A. Foreman and I. W. Evett (2001). "*Inferring ethnic origin by means of an STR profile*." Forensic Sci Int **119**(1): 17-22.

Lu, G. D. and J. Needham (1988). "*A history of forensic medicine in China*." Med Hist **32**(4): 357-400.

Lu, T., Y. Pan, S. Y. Kao, C. Li, I. Kohane, J. Chan and B. A. Yankner (2004). "*Gene regulation and DNA damage in the ageing human brain*." Nature **429**(6994): 883-891.

Makova, K. D., S. Yang and F. Chiaromonte (2004). "*Insertions and deletions are male biased too: a whole-genome analysis in rodents*." Genome Res **14**(4): 567-573.

Malausa, T., A. Gilles, E. Meglecz, H. Blanquart, S. Duthoy, C. Costedoat, V. Dubut, N. Pech, P. Castagnone-Sereno, C. Delye, et al. (2011). "*High-throughput microsatellite isolation through 454 GS-FLX Titanium pyrosequencing of enriched DNA libraries*." Mol Ecol Resour **11**(4): 638-644.

Manica, A., W. Amos, F. Balloux and T. Hanihara (2007). "*The effect of ancient population bottlenecks on human phenotypic variation*." Nature **448**(7151): 346-348.

Manica, A., F. Prugnolle and F. Balloux (2005). "*Geography is a better determinant of human genetic differentiation than ethnicity*." Hum Genet **118**(3-4): 366-371.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, et al. (2005). "*Genome sequencing in microfabricated high-density picolitre reactors*." Nature **437**(7057): 376-380.

Maroñas, O., C. Phillips, J. Sochtig, A. Gomez-Tato, R. Cruz, J. Alvarez-Dios, M. C. de Cal, Y. Ruiz, M. Fondevila, A. Carracedo, et al. (2014). "*Development of a forensic skin colour predictive test*." Forensic Sci Int Genet **13**: 34-44.

Marques-Bonet, T., O. A. Ryder and E. E. Eichler (2009). "*Sequencing primate genomes: what have we learned?*" Annu Rev Genomics Hum Genet **10**: 355-386.

Marth, G. T., E. Czabarka, J. Murvai and S. T. Sherry (2004). "*The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations*." Genetics **166**(1): 351-372.

Mattick, J. S. and I. V. Makunin (2006). "*Non-coding RNA*." Hum Mol Genet **15 Spec No 1**: R17-29.

McDougall, I., F. H. Brown and J. G. Fleagle (2005). "*Stratigraphic placement and age of modern humans from Kibish, Ethiopia*." Nature **433**(7027): 733-736.

McKeigue, P. (2007). "*ADMIXMAP: a program to model admixture using marker genotype data*." 3.7. Retrieved 25/04/2013, 2013, from http://homepages.ed.ac.uk/pmckeigu/admixmap/admixmap_manual.pdf.

McKeigue, P. M., J. R. Carpenter, E. J. Parra and M. D. Shriver (2000). "*Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations*." Ann Hum Genet **64**(Pt 2): 171-186.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, et al. (2010). "*The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data*." Genome Res **20**(9): 1297-1303.

McPherron, S. P., Z. Alemseged, C. W. Marean, J. G. Wynn, D. Reed, D. Geraads, R. Bobe and H. A. Bearat (2010). "*Evidence for stone-tool-assisted consumption of animal tissues before 3.39 million years ago at Dikika, Ethiopia*." Nature **466**(7308): 857-860.

McVean, G., C. C. Spencer and R. Chaix (2005). "*Perspectives on human genetic variation from the HapMap Project*." PLoS Genet **1**(4): e54.

Medland, S. E., D. R. Nyholt, J. N. Painter, B. P. McEvoy, A. F. McRae, G. Zhu, S. D. Gordon, M. A. Ferreira, M. J. Wright, A. K. Henders, et al. (2009). "*Common variants in the trichohyalin gene are associated with straight hair in Europeans*." Am J Hum Genet **85**(5): 750-755.

Mellars, P. (2006). "*A new radiocarbon revolution and the dispersal of modern humans in Eurasia*." Nature **439**(7079): 931-935.

Mellars, P. (2011). "*Palaeoanthropology: the earliest modern humans in Europe*." Nature **479**(7374): 483-485.

Melton, T., R. Peterson, A. J. Redd, N. Saha, A. S. Sofro, J. Martinson and M. Stoneking (1995). "*Polynesian genetic affinities with Southeast Asian populations as identified by mtDNA analysis*." Am J Hum Genet **57**(2): 403-414.

Mills, R. E., C. T. Luttig, C. E. Larkins, A. Beauchamp, C. Tsui, W. S. Pittard and S. E. Devine (2006). "*An initial map of insertion and deletion (INDEL) variation in the human genome*." Genome Res **16**(9): 1182-1190.

Mirabal, S., A. M. Cadenas, R. Garcia-Bertrand and R. J. Herrera (2013). "*Ascertaining the role of Taiwan as a source for the Austronesian expansion*." Am J Phys Anthropol **150**(4): 551-564.

Mirabal, S., K. J. Herrera, T. Gayden, M. Regueiro, P. A. Underhill, R. L. Garcia-Bertrand and R. J. Herrera (2012). "*Increased Y-chromosome resolution of haplogroup O suggests genetic ties between the Ami aborigines of Taiwan and the Polynesian Islands of Samoa and Tonga*." Gene **492**(2): 339-348.

Mitchell, R. J., M. Kreskas, E. Baxter, L. Buffalino and R. A. Van Oorschot (2006). "*An investigation of sequence deletions of amelogenin (AMELY), a Y-chromosome locus commonly used for gender determination*." Ann Hum Biol **33**(2): 227-240.

Mona, S., K. E. Grunz, S. Brauer, B. Pakendorf, L. Castri, H. Sudoyo, S. Marzuki, R. H. Barnes, J. Schmidtke, M. Stoneking, et al. (2009). "*Genetic admixture history of Eastern Indonesia as revealed by Y-chromosome and mitochondrial DNA analysis*." Mol Biol Evol **26**(8): 1865-1877.

Monzon, F. A., J. M. Hagenkord, M. A. Lyons-Weiler, J. P. Balani, A. V. Parwani, C. M. Sciulli, J. Li, U. R. Chandran, S. I. Bastacky and R. Dhir (2008). "*Whole genome SNP arrays as a potential diagnostic tool for the detection of characteristic chromosomal aberrations in renal epithelial tumors*." Mod Pathol **21**(5): 599-608.

Mou, C., H. A. Thomason, P. M. Willan, C. Clowes, W. E. Harris, C. F. Drew, J. Dixon, M. J. Dixon and D. J. Headon (2008). "*Enhanced ectodysplasin-A receptor (EDAR) signaling alters multiple fiber characteristics to produce the East Asian hair form*." Hum Mutat **29**(12): 1405-1411.

Mourant, A. E. (1946). "*A new human blood group antigen of frequent occurrence*." Nature **158**: 237.

Mulero, J. J., C. W. Chang, L. M. Calandro, R. L. Green, Y. Li, C. L. Johnson and L. K. Hennessy (2006). "*Development and validation of the AmpFlSTR Yfiler PCR amplification kit: a male specific, single amplification 17 Y-STR multiplex system*." J Forensic Sci **51**(1): 64-75.

Mullis, K., F. Faloona, S. Scharf, R. Saiki, G. Horn and H. Erlich (1986). "*Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction*." Cold Spring Harb Symp Quant Biol **51 Pt 1**: 263-273.

Musgrave-Brown, E., D. Ballard, K. Balogh, K. Bender, B. Berger, M. Bogus, C. Borsting, M. Brion, M. Fondevila, C. Harrison, et al. (2007). "*Forensic validation of the SNPforID 52-plex assay*." Forensic Sci Int Genet **1**(2): 186-190.

Nachman, M. W. and S. L. Crowell (2000). "*Estimate of the Mutation Rate per Nucleotide in Humans*." Genetics **156**(1): 297-304.

Nakamura, Y., M. Carlson, K. Krapcho and R. White (1988). "*Isolation and mapping of a polymorphic DNA sequence (pMCT118) on chromosome 1p [D1S80]*." Nucleic Acids Res **16**(19): 9364.

Nalls, M. A., J. G. Wilson, N. J. Patterson, A. Tandon, J. M. Zmuda, S. Huntsman, M. Garcia, D. Hu, R. Li, B. A. Beamer, et al. (2008). "*Admixture mapping of white cell count: genetic locus responsible for lower white blood cell count in the Health ABC and Jackson Heart studies*." Am J Hum Genet **82**(1): 81-87.

Nassir, R., R. Kosoy, C. Tian, P. White, L. Butler, G. Silva, R. Kittles, M. Alarcon-Riquelme, P. Gregersen, J. Belmont, et al. (2009). "*An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels*." BMC Genet **10**(1): 39.

Ngo, V. N., R. M. Young, R. Schmitz, S. Jhavar, W. Xiao, K.-H. Lim, H. Kohlhammer, W. Xu, Y. Yang, H. Zhao, et al. (2011). "*Oncogenically active MYD88 mutations in human lymphoma*." Nature **470**(7332): 115-119.

Nielsen, R., J. S. Paul, A. Albrechtsen and Y. S. Song (2011). "*Genotype and SNP calling from next-generation sequencing data*." Nat Rev Genet **12**(6): 443-451.

Nievergelt, C., A. Maihofer, T. Shekhtman, O. Libiger, X. Wang, K. Kidd and J. Kidd (2013). "*Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel*." Investig Genet **4**(1): 13.

Novembre, J. and S. Ramachandran (2011). "*Perspectives on human population structure at the cusp of the sequencing era*." Annu Rev Genomics Hum Genet **12**: 245-274.

Nussbaumer, C., E. Gharehbaghi-Schnell and I. Korschineck (2006). "*Messenger RNA profiling: a novel method for body fluid identification by real-time PCR*." Forensic Sci Int **157**(2-3): 181-186.

Ohta, T. and M. Kimura (1973). "*A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population*." Genet Res **22**(2): 201-204.

Oppenheimer, S. (2012). "*A single southern exit of modern humans from Africa: Before or after Toba?*" Quat Int **258**(0): 88-99.

Orban, G. A. and F. Caruana (2014). "*The neural basis of human tool use*." Front Psychol **5**: 310.

Ossorio, P. N. (2006). "*About face: forensic genetic testing for race and visible traits*." J Law Med Ethics **34**(2): 277-292.

Ou, X., H. Zhao, H. Sun, Z. Yang, B. Xie, Y. Shi and X. Wu (2011). "*Detection and quantification of the age-related sjTREC decline in human peripheral blood*." Int J Legal Med **125**(4): 603-608.

Ou, X. L., J. Gao, H. Wang, H. S. Wang, H. L. Lu and H. Y. Sun (2012). "*Predicting Human Age with Bloodstains by sjTREC Quantification*." PLoS ONE **7**(8): e42412.

Oz, C., N. Zaken, M. Amiel and A. Zamir (2008). "*A Y-chromosome STR marker should be added to commercial multiplex STR kits*." J Forensic Sci **53**(4): 858-861.

Padmanabhan, R. and R. Wu (1972). "*Nucleotide sequence analysis of DNA. IX. Use of oligonucleotides of defined sequence as primers in DNA sequence analysis*." Biochem Biophys Res Commun **48**(5): 1295-1302.

Pakendorf, B. and M. Stoneking (2005). "*Mitochondrial DNA and human evolution*." Annu Rev Genomics Hum Genet **6**(1): 165-183.

Parson, W., C. Strobl, G. Huber, B. Zimmermann, S. M. Gomes, L. Souto, L. Fendt, R. Delport, R. Langit, S. Wootton, et al. (2013). "*Evaluation of next generation mtGenome sequencing using the Ion Torrent Personal Genome Machine (PGM)*." Forensic Sci Int Genet **7**(5): 543-549.

Paschou, P., J. Lewis, A. Javed and P. Drineas (2010). "*Ancestry informative markers for fine-scale individual assignment to worldwide populations*." J Med Genet **47**(12): 835-847.

Paternoster, L., Alexei I. Zhurov, Arshed M. Toma, John P. Kemp, B. St. Pourcain, Nicholas J. Timpson, G. McMahon, W. McArdle, Susan M. Ring, George D. Smith, et al. (2012). "*Genome-wide Association Study of Three-Dimensional Facial Morphology Identifies a Variant in PAX3 Associated with Nasion Position*." Am J Hum Genet **90**(3): 478-485.

Patterson, N., A. L. Price and D. Reich (2006). "*Population structure and eigenanalysis*." PLoS Genet **2**(12): e190.

Pearce, J. M. (2006). "*Minding the gap: frequency of indels in mtDNA control region sequence data and influence on population genetic analyses*." Mol Ecol **15**(2): 333-341.

Pemberton, T. J., M. DeGiorgio and N. A. Rosenberg (2013). "*Population structure in a comprehensive genomic data set on human microsatellite variation*." G3 (Bethesda) **3**(5): 891-907.

Perego, U. A., A. Achilli, N. Angerhofer, M. Accetturo, M. Pala, A. Olivieri, B. Hooshiar Kashani, K. H. Ritchie, R. Scozzari, Q. P. Kong, et al. (2009). "*Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups*." Curr Biol **19**(1): 1-8.

Pereira, R., V. Pereira, I. Gomes, C. Tomas, N. Morling, A. Amorim, M. J. Prata, A. Carracedo and L. Gusmao (2012a). "*A method for the analysis of 32 X chromosome insertion deletion polymorphisms in a single PCR*." Int J Legal Med **126**(1): 97-105.

Pereira, R., C. Phillips, C. Alves, A. Amorim, A. Carracedo and L. Gusmao (2009). "*A new multiplex for human identification using insertion/deletion polymorphisms*." Electrophoresis **30**(21): 3682-3690.

Pereira, R., C. Phillips, N. Pinto, C. Santos, S. E. dos Santos, A. Amorim, Á. Carracedo and L. Gusmão (2012b). "*Straightforward inference of ancestry and admixture*

proportions through ancestry-informative insertion deletion multiplexing." PLoS One **7**(1): e29684.

Perkins, T. T., C. Y. Tay, F. Thirriot and B. Marshall (2013). "*Choosing a benchtop sequencing machine to characterise Helicobacter pylori genomes*." PLoS ONE **8**(6): e67539.

Petkovski, E., C. Keyser-Tracqui, D. Niemeyer, R. Hienne and B. Ludes (2004). "*SNPs: tools for individual identification*." Int Congr Ser **1261**: 21-23.

Pettersson, E., J. Lundeberg and A. Ahmadian (2009). "*Generations of sequencing technologies*." Genomics **93**(2): 105-111.

Pfaff, C. L., J. Barnholtz-Sloan, J. K. Wagner and J. C. Long (2004). "*Information on ancestry from genetic markers*." Genet Epidemiol **26**(4): 305-315.

Phillips, C. (2009). "*SNP databases*." Methods Mol Biol **578**: 43-71.

Phillips, C. (2013). *Ancestry Informative Markers*. Encyclopedia of Forensic Sciences. J. Siegel and P. Saukko, Waltham: Academic Press. **1:** 323-331.

Phillips, C. (2015). "*Forensic genetic analysis of bio-geographical ancestry*." Forensic Sci Int Genet **18**(49-65).

Phillips, C., D. Ballard, P. Gill, D. S. Court, A. Carracedo and M. V. Lareu (2012). "*The recombination landscape around forensic STRs: Accurate measurement of genetic distances between syntenic STR pairs using HapMap high density SNP data*." Forensic Sci Int Genet **6**(3): 354-365.

Phillips, C., R. Fang, D. Ballard, M. Fondevila, C. Harrison, F. Hyland, E. Musgrave-Brown, C. Proff, E. Ramos-Luis, B. Sobrino, et al. (2007a). "*Evaluation of the Genplex SNP typing system and a 49plex forensic marker panel*." Forensic Sci Int Genet **1**(2): 180-185.

Phillips, C., L. Fernandez-Formoso, M. Garcia-Magariños, L. Porras, T. Tvedebrink, J. Amigo, M. Fondevila, A. Gomez-Tato, J. Álvarez-Dios, A. Freire-Aradas, et al. (2011). "*Analysis of global variability in 15 established and 5 new European Standard Set (ESS) STRs using the CEPH human genome diversity panel*." Forensic Sci Int Genet **5**(3): 155-169.

Phillips, C., L. Fernandez-Formoso, M. Gelabert-Besada, M. Garcia-Magariños, C. Santos, M. Fondevila, Á. Carracedo and M. V. Lareu (2013a). "*Development of a novel forensic STR multiplex for ancestry analysis and extended identity testing*." Electrophoresis **34**(8): 1151-1162.

Phillips, C., M. Fondevila, M. Garcia-Magariños, A. Rodriguez, A. Salas, Á. Carracedo and M. V. Lareu (2008a). "*Resolving relationship tests that show ambiguous STR results using autosomal SNPs as supplementary markers*." Forensic Sci Int Genet **2**(3): 198-204.

Phillips, C., A. Freire-Aradas, A. K. Kriegel, M. Fondevila, O. Bulbul, C. Santos, F. Serrulla Rech, M. D. Perez Carceles, Á. Carracedo, P. M. Schneider, et al. (2013b). "*Eurasiaplex: a forensic SNP assay for differentiating European and South Asian ancestries*." Forensic Sci Int Genet **7**(3): 359-366.

Phillips, C., M. Lareu, J. Sanchez, M. Brion, B. Sobrino, N. Morling, P. Schneider, D. S. Court and A. Carracedo (2004a). "*Selecting single nucleotide polymorphisms for forensic applications*." Int Congr Ser **1261**: 18-20.

Phillips, C., M. V. Lareu, A. Salas and Á. Carracedo (2004b). "*Nonbinary single-nucleotide polymorphism markers*." Int Congr Ser **1261**: 27-29.

Phillips, C., W. Parson, B. Lundsberg, C. Santos, A. Freire-Aradas, M. Torres, M. Eduardoff, C. Borsting, P. Johansen, M. Fondevila, et al. (2014). "*Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set*." Forensic Sci Int Genet **11**: 13-25.

Phillips, C., L. Prieto, M. Fondevila, A. Salas, A. Gomez-Tato, J. Álvarez-Dios, A. Alonso, A. Blanco-Verea, M. Brion, M. Montesino, et al. (2009). "*Ancestry analysis in the 11-M Madrid bomb attack investigation*." PLoS One **4**(8): e6583.

Phillips, C., A. Rodriguez, A. Mosquera-Miguel, M. Fondevila, L. Porras-Hurtado, F. Rondon, A. Salas, A. Carracedo and M. V. Lareu (2008b). "*D9S1120, a simple STR with a common Native American-specific allele: forensic optimization, locus characterization and allele frequency studies*." Forensic Sci Int Genet **3**(1): 7-13.

Phillips, C., A. Salas, J. J. Sánchez, M. Fondevila, A. Gomez-Tato, J. Álvarez-Dios, M. Calaza, M. C. de Cal, D. Ballard, M. V. Lareu, et al. (2007b). "*Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs*." Forensic Sci Int Genet **1**(3-4): 273-280.

Pickrell, J. K. and D. Reich (2014). "*Toward a new history and geography of human genes informed by ancient DNA*." Trends Genet **30**(9): 377-389.

Pimenta, J. R. and S. D. Pena (2010). "*Efficient human paternity testing with a panel of 40 short insertion-deletion polymorphisms*." Genet Mol Res **9**(1): 601-607.

Porras-Hurtado, L., Y. Ruiz, C. Santos, C. Phillips, Á. Carracedo and M. V. Lareu (2013). "*An overview of STRUCTURE: applications, parameter settings and supporting software*." Front Genet **4**.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick and D. Reich (2006). "*Principal components analysis corrects for stratification in genome-wide association studies*." Nat Genet **38**(8): 904-909.

Prinz, M., A. Carracedo, W. R. Mayr, N. Morling, T. J. Parsons, A. Sajantila, R. Scheithauer, H. Schmitter, P. M. Schneider and G. International Society for Forensic (2007). "*DNA Commission of the International Society for Forensic Genetics (ISFG): recommendations regarding the role of forensic genetics for disaster victim identification (DVI)*." Forensic Sci Int Genet **1**(1): 3-12.

Pritchard, J. K., M. Stephens and P. Donnelly (2000). "*Inference of population structure using multilocus genotype data*." Genetics **155**(2): 945-959.

Pruetz, J. D. and P. Bertolani (2007). "*Savanna chimpanzees, Pan troglodytes verus, hunt with tools*." Curr Biol **17**(5): 412-417.

Prugnolle, F., A. Manica and F. Balloux (2005). "*Geography predicts neutral genetic diversity of human populations*." Curr Biol **15**(5): R159-160.

Pugach, I., F. Delfin, E. Gunnarsdottir, M. Kayser and M. Stoneking (2013). "*Genome-wide data substantiate Holocene gene flow from India to Australia*." Proc Natl Acad Sci U S A **110**(5): 1803-1808.

Quail, M. A., M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow and Y. Gu (2012). "*A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers*." BMC Genomics **13**: 341.

Ramachandran, S., O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman and L. L. Cavalli-Sforza (2005). "*Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa*." Proc Natl Acad Sci U S A **102**(44): 15942-15947.

Rasmussen, M., X. Guo, Y. Wang, K. E. Lohmueller, S. Rasmussen, A. Albrechtsen, L. Skotte, S. Lindgreen, M. Metspalu, T. Jombart, et al. (2011). "*An Aboriginal Australian genome reveals separate human dispersals into Asia*." Science **334**(6052): 94-98.

Ratan, A., W. Miller, J. Guillory, J. Stinson, S. Seshagiri and S. C. Schuster (2013). "*Comparison of sequencing platforms for single nucleotide variant calls in a human sample*." PLoS One **8**(2): e55089.

Ratan, A., T. L. Olson, T. P. Loughran Jr and W. Miller (2015). "*Identification of indels in next-generation sequencing data*." BMC Bioinformatics **16**: 42-49.

Ray, S. and R. H. Turi. (1999). "*Determination of number of clusters in k-means clustering and application in colour image segmentation*." Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques: 137–143.

Redd, A. J., J. Roberts-Thomson, T. Karafet, M. Bamshad, L. B. Jorde, J. M. Naidu, B. Walsh and M. F. Hammer (2002). "*Gene flow from the Indian subcontinent to Australia: Evidence from the Y chromosome*." Curr Biol **12**(8): 673-677.

Redd, A. J. and M. Stoneking (1999). "*Peopling of Sahul: mtDNA variation in aboriginal Australian and Papua New Guinean populations*." Am J Hum Genet **65**(3): 808-828.

REGWG (2005). "*The Use of Racial, Ethnic, and Ancestral Categories in Human Genetics Research*." Am J Hum Genet **77**(4): 519-532.

Reich, D., R. E. Green, M. Kircher, J. Krause, N. Patterson, E. Y. Durand, B. Viola, A. W. Briggs, U. Stenzel, P. L. Johnson, et al. (2010). "*Genetic history of an archaic hominin group from Denisova Cave in Siberia*." Nature **468**(7327): 1053-1060.

Reich, D., N. Patterson, D. Campbell, A. Tandon, S. Mazieres, N. Ray, M. V. Parra, W. Rojas, C. Duque, N. Mesa, et al. (2012). "*Reconstructing Native American population history*." Nature **488**(7411): 370-374.

Reich, D., N. Patterson, M. Kircher, F. Delfin, M. R. Nandineni, I. Pugach, A. M. Ko, Y. C. Ko, T. A. Jinam, M. E. Phipps, et al. (2011). "*Denisova admixture and the first modern*

human dispersals into Southeast Asia and Oceania." Am J Hum Genet **89**(4): 516-528.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward, et al. (2001). "*Linkage disequilibrium in the human genome*." Nature **411**(6834): 199-204.

Relethford, J. H. (2004). "*Global patterns of isolation by distance based on genetic and morphological data*." Hum Biol **76**(4): 499-513.

Ribeiro-Rodrigues, E. M., N. P. dos Santos, A. K. dos Santos, R. Pereira, A. Amorim, L. Gusmao, M. A. Zago and S. E. dos Santos (2009). "*Assessing interethnic admixture using an X-linked insertion-deletion multiplex*." Am J Hum Biol **21**(5): 707-709.

Ricaut, F.-X., T. Thomas, M. Mormina, M. P. Cox, M. Bellatti, R. A. Foley and M. Mirazon-Lahr (2010). "*Ancient Solomon Islands mtDNA: assessing Holocene settlement and the impact of European contact*." J Archaeol Sci **37**(6): 1161-1170.

Roach, J. C., C. Boysen, K. Wang and L. Hood (1995). "*Pairwise end sequencing: a unified approach to genomic mapping and sequencing*." Genomics **26**(2): 345-353.

Robinson, J. T., H. Thorvaldsdottir, W. Winckler, M. Guttman, E. S. Lander, G. Getz and J. P. Mesirov (2011). "*Integrative genomics viewer*." Nat Biotechnol **29**(1): 24-26.

Robledo, R., W. Beggs and P. Bender (2003). "*A simple and cost-effective method for rapid genotyping of insertion/deletion polymorphisms*." Genomics **82**(5): 580-582.

Rockenbauer, E., S. Hansen, M. Mikkelsen, C. Borsting and N. Morling (2014). "*Characterization of mutations and sequence variants in the D21S11 locus by next generation sequencing*." Forensic Sci Int Genet **8**(1): 68-72.

Roewer, L. (2009). "*Y chromosome STR typing in crime casework*." Forensic Sci Med Pathol **5**(2): 77-84.

Roffey, P. E., C. I. Eckhoff and J. L. Kuhl (2000). "*A rare mutation in the amelogenin gene and its potential investigative ramifications*." J Forensic Sci **45**(5): 1016-1019.

Rohlfs, R. V., S. M. Fullerton and B. S. Weir (2012). "*Familial identification: population structure and relationship distinguishability*." PLoS Genet **8**(2): e1002469.

Romero, I. G., A. Manica, J. Goudet, L. L. Handley and F. Balloux (2009). "*How accurate is the current picture of human genetic variation?*" Heredity **102**(2): 120-126.

Rosenberg, N. A. (2004). "*DISTRUCT: a program for the graphical display of population structure*." Mol Ecol Notes **4**(1): 137-138.

Rosenberg, N. A. (2006). "*Standardized subsets of the HGDP-CEPH human genome diversity cell line panel, accounting for atypical and duplicated samples and pairs of close relatives*." Ann Hum Genet **70**: 841-847.

Rosenberg, N. A., L. M. Li, R. Ward and J. K. Pritchard (2003). "*Informativeness of genetic markers for inference of ancestry*." Am J Hum Genet **73**(6): 1402-1422.

Rosenberg, N. A., S. Mahajan, S. Ramachandran, C. Zhao, J. K. Pritchard and M. W. Feldman (2005). "*Clines, clusters, and the effect of study design on the inference of human population structure*." PLoS Genet **1**(6): e70.

Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky and M. W. Feldman (2002). "*Genetic structure of human populations*." Science **298**(5602): 2381-2385.

Ross, K. A. (2014). "*Coherent somatic mutation in autoimmune disease*." PLoS One **9**(7): e101093.

Ross, M. G., C. Russ, M. Costello, A. Hollinger, N. J. Lennon, R. Hegarty, C. Nusbaum and D. B. Jaffe (2013). "*Characterizing and measuring bias in sequence data*." Genome Biol **14**(5): R51.

Rothberg, J. M., W. Hinz, T. M. Rearick, J. Schultz, W. Mileski, M. Davey, J. H. Leamon, K. Johnson, M. J. Milgrew, M. Edwards, et al. (2011). "*An integrated semiconductor device enabling non-optical genome sequencing*." Nature **475**(7356): 348-352.

Royal, C. D., J. Novembre, S. M. Fullerton, D. B. Goldstein, J. C. Long, M. J. Bamshad and A. G. Clark (2010). "*Inferring genetic ancestry: opportunities, challenges, and implications*." Am J Hum Genet **86**(5): 661-673.

Ruff, C. (2009). "*Relative limb strength and locomotion in Homo habilis*." Am J Phys Anthropol **138**(1): 90-100.

Ruiz, Y., C. Phillips, A. Gomez-Tato, J. Álvarez-Dios, M. Casares de Cal, R. Cruz, O. Maronas, J. Söchtig, M. Fondevila, M. J. Rodriguez-Cid, et al. (2013). "*Further development of forensic eye color predictive tests*." Forensic Sci Int Genet **7**(1): 28-40.

Salas, A., C. Phillips and Á. Carracedo (2006). "*Ancestry vs physical traits: the search for ancestry informative markers (AIMs)*." Int J Legal Med **120**(3): 188-189.

Sánchez, J. J., C. Phillips, C. Børsting, K. Balogh, M. Bogus, M. Fondevila, C. D. Harrison, E. Musgrave-Brown, A. Salas, D. Syndercombe Court, et al. (2006). "*A multiplex assay with 52 single nucleotide polymorphisms for human identification*." Electrophoresis **27**(9): 1713-1724.

Sanger, F., S. Nicklen and A. R. Coulson (1977). "*DNA sequencing with chain-terminating inhibitors*." Proc Natl Acad Sci U S A **74**(12): 5463-5467.

Sanger, R., R. R. Race and J. Jack (1955). "*The Duffy Blood Groups of New York Negroes: The Phenotype Fy (a−b−)*." Br J Haematol **1**(4): 370-374.

Sankararaman, S., N. Patterson, H. Li, S. Paabo and D. Reich (2012). "*The date of interbreeding between Neandertals and modern humans*." PLoS Genet **8**(10): e1002947.

Santos, C., M. Fondevila, D. Ballard, R. Banemann, A. M. Bento, C. Borsting, W. Branicki, F. Brisighelli, M. Burrington, T. Capal, et al. (2015a). "*Forensic ancestry analysis with two capillary electrophoresis ancestry informative marker (AIM) panels: Results of a collaborative EDNAP exercise*." Forensic Sci Int Genet **19**: 56-67.

Santos, C., C. Phillips, F. Oldoni, J. Amigo, M. Fondevila, R. Pereira, Á. Carracedo and M. V. Lareu (2015b). "*Completion of a worldwide reference panel of samples for an ancestry informative Indel assay*." Forensic Sci Int Genet **17**(75-80).

Santos, N. P., E. M. Ribeiro-Rodrigues, A. K. Ribeiro-Dos-Santos, R. Pereira, L. Gusmao, A. Amorim, J. F. Guerreiro, M. A. Zago, C. Matte, M. H. Hutz, et al. (2010). "*Assessing*

individual interethnic admixture and population substructure using a 48-insertion-deletion (INSEL) ancestry-informative marker (AIM) panel." Hum Mutat **31**(2): 184-190.

Scheible, M., O. Loreille, R. Just and J. Irwin (2014). "*Short tandem repeat typing on the 454 platform: strategies and considerations for targeted sequencing of common forensic markers*." Forensic Sci Int Genet **12**: 107-119.

Semaw, S., M. J. Rogers, J. Quade, P. R. Renne, R. F. Butler, M. Dominguez-Rodrigo, D. Stout, W. S. Hart, T. Pickering and S. W. Simpson (2003). "*2.6-Million-year-old stone tools and associated bones from OGS-6 and OGS-7, Gona, Afar, Ethiopia*." J Hum Evol **45**(2): 169-177.

Semsei, I., S. Maier, J. Workman-Azbill, L. Urban, K. Moser, M. Zeher, M. Bachmann and A. D. Farris (2007). "*Detection of a rare oligo(A) repeat tract mutation (8As-->7As) in the sequence encoding the La/SS-B autoantigen*." Anal Biochem **370**(1): 47-53.

Seo, S. B., J. L. King, D. H. Warshauer, C. P. Davis, J. Ge and B. Budowle (2013). "*Single nucleotide polymorphism typing with massively parallel sequencing for human identification*." Int J Legal Med **127**(6): 1079-1086.

Serra, A., A. M. Bento, M. Carvalho, L. Andrade, L. Batista, M. C. Oliveira, V. Lopes, F. Balsa, F. Corte-Real and M. J. Anjos (2008). "*X-chromosome STR typing in deficiency paternity cases*." Forensic Sci Int Genet Suppl Ser **1**(1): 162-163.

Serre, D. and S. Paabo (2004). "*Evidence for gradients of human genetic diversity within and among continents*." Genome Res **14**(9): 1679-1685.

Shang, H., H. Tong, S. Zhang, F. Chen and E. Trinkaus (2007). "*An early modern human from Tianyuan Cave, Zhoukoudian, China*." Proc Natl Acad Sci U S A **104**(16): 6573-6578.

Shendure, J. and H. Ji (2008). "*Next-generation DNA sequencing*." Nat Biotechnol **26**(10): 1135-1145.

Sherry, S. T., M. Ward and K. Sirotkin (1999). "*dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation*." Genome Res **9**(8): 677-679.

Sherry, S. T., M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski and K. Sirotkin (2001). "*dbSNP: the NCBI database of genetic variation*." Nucleic Acids Res **29**(1): 308-311.

Shi, W., Q. Ayub, M. Vermeulen, R. G. Shao, S. Zuniga, K. van der Gaag, P. de Knijff, M. Kayser, Y. Xue and C. Tyler-Smith (2010). "*A worldwide survey of human male demographic history based on Y-SNP and Y-STR data from the HGDP-CEPH populations*." Mol Biol Evol **27**(2): 385-393.

Shriver, M. D., G. C. Kennedy, E. J. Parra, H. A. Lawson, V. Sonpar, J. Huang, J. M. Akey and K. W. Jones (2004). "*The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs*." Hum Genomics **1**(4): 274-286.

Shriver, M. D., M. W. Smith, L. Jin, A. Marcini, J. M. Akey, R. Deka and R. E. Ferrell (1997). "*Ethnic-affiliation estimation by use of population-specific DNA markers*." Am J Hum Genet **60**(4): 957-964.

Silva, N. M., L. Pereira, E. S. Poloni and M. Currat (2012). "*Human neutral genetic variation and forensic STR data*." PLoS One **7**(11): e49666.

Sims, D., I. Sudbery, N. E. Ilott, A. Heger and C. P. Ponting (2014). "*Sequencing depth and coverage: key considerations in genomic analyses*." Nat Rev Genet **15**(2): 121-132.

Smith, L. M., J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. H. Kent and L. E. Hood (1986). "*Fluorescence detection in automated DNA sequence analysis*." Nature **321**(6071): 674-679.

Smith, P. J. and J. Ballantyne (2007). "*Simplified low-copy-number DNA analysis by post-PCR purification*." J Forensic Sci **52**(4): 820-829.

Soares, P., T. Rito, J. Trejaut, M. Mormina, C. Hill, E. Tinkler-Hundal, M. Braid, D. J. Clarke, J. H. Loo, N. Thomson, et al. (2011). "*Ancient voyaging and Polynesian origins*." Am J Hum Genet **88**(2): 239-247.

Sobrino, B., M. Brion and A. Carracedo (2005). "*SNPs in forensic genetics: a review on SNP typing methodologies*." Forensic Sci Int **154**(2-3): 181-194.

Staden, R. (1979). "*A strategy of DNA sequencing employing computer programs*." Nucleic Acids Res **6**(7): 2601-2610.

Stenson, P. D., E. V. Ball, M. Mort, A. D. Phillips, J. A. Shiel, N. S. T. Thomas, S. Abeysinghe, M. Krawczak and D. N. Cooper (2003). "*Human Gene Mutation Database (HGMD®): 2003 update.*" Human Mutation **21**(6): 577-581.

Stokowski, R. P., P. V. Pant, T. Dadd, A. Fereday, D. A. Hinds, C. Jarman, W. Filsell, R. S. Ginger, M. R. Green, F. J. van der Ouderaa, et al. (2007). "*A genomewide association study of skin pigmentation in a South Asian population*." Am J Hum Genet **81**(6): 1119-1132.

Stout, D., J. Quade, S. Semaw, M. J. Rogers and N. E. Levin (2005). "*Raw material selectivity of the earliest stone toolmakers at Gona, Afar, Ethiopia*." J Hum Evol **48**(4): 365-380.

Strauss, K. A., E. G. Puffenberger, D. W. Craig, C. B. Panganiban, A. M. Lee, D. Hu-Lince, D. A. Stephan and D. H. Morton (2005). "*Genome-wide SNP arrays as a diagnostic tool: clinical description, genetic mapping, and molecular characterization of Salla disease in an Old Order Mennonite population*." Am J Med Genet A **138A**(3): 262-267.

Stringer, C. (2012). "*Evolution: What makes a modern human*." Nature **485**(7396): 33-35.

Stringer, C. and R. McKie (1996). *African exodus: the origins of modern humanity*. New York, Henry Holt.

Sulem, P., D. F. Gudbjartsson, S. N. Stacey, A. Helgason, T. Rafnar, M. Jakobsdottir, S. Steinberg, S. A. Gudjonsson, A. Palsson, G. Thorleifsson, et al. (2008). "*Two newly identified genetic determinants of pigmentation in Europeans*." Nat Genet **40**(7): 835-837.

Sulem, P., D. F. Gudbjartsson, S. N. Stacey, A. Helgason, T. Rafnar, K. P. Magnusson, A. Manolescu, A. Karason, A. Palsson, G. Thorleifsson, et al. (2007). "*Genetic determinants of hair, eye and skin pigmentation in Europeans*." Nat Genet **39**(12): 1443-1452.

Sun, J. X., A. Helgason, G. Masson, S. S. Ebenesersdottir, H. Li, S. Mallick, S. Gnerre, N. Patterson, A. Kong, D. Reich, et al. (2012). "*A direct characterization of human mutation based on microsatellites*." Nat Genet **44**(10): 1161-1165.

Sundstrom, H., M. T. Webster and H. Ellegren (2003). "*Is the rate of insertion and deletion mutation male biased?: Molecular evolutionary analysis of avian and primate sex chromosome sequences*." Genetics **164**(1): 259-268.

Susman, R. L. (1998). "*Hand function and tool behavior in early hominids*." J Hum Evol **35**(1): 23-46.

Sutherland, G. R. and R. I. Richards (1995). "*Simple tandem DNA repeats and human genetic disease*." Proc Natl Acad Sci U S A **92**(9): 3636-3641.

Sutovsky, P. (2003). "*Ubiquitin-dependent proteolysis in mammalian spermatogenesis, fertilization, and sperm quality control: Killing three birds with one stone*." Microsc Res Tech **61**(1): 88-102.

Sutovsky, P., K. V. Leyen, T. McCauley, B. Day and M. Sutovsky (2004). "*Degradation of paternal mitochondria after fertilization: implications for heteroplasmy, assisted reproductive technologies and mtDNA inheritance*." Reprod Biomed Online **8**(1): 24-33.

Szibor, R. (2007). "*X-chromosomal markers: past, present and future*." Forensic Sci Int Genet **1**(2): 93-99.

Szibor, R., M. Krawczak, S. Hering, J. Edelmann, E. Kuhlisch and D. Krause (2003a). "*Use of X-linked markers for forensic purposes*." Int J Legal Med **117**(2): 67-74.

Szibor, R., I. Plate, J. Edelmann, S. Hering, E. Kuhlisch, M. Michael and D. Krause (2003b). "*Chromosome X haplotyping in deficiency paternity testing principles and case report*." Int Congr Ser **1239**(0): 815-820.

Takahata, N., S. H. Lee and Y. Satta (2001). "*Testing multiregionality of modern human origins*." Mol Biol Evol **18**(2): 172-183.

Tandon, A., N. Patterson and D. Reich (2011). "*Ancestry informative marker panels for African Americans based on subsets of commercially available SNP arrays*." Genet Epidemiol **35**(1): 80-83.

Tang, H., J. Peng, P. Wang and N. J. Risch (2005). "*Estimation of individual admixture: Analytical and study design considerations*." Genet Epidemiol **28**(4): 289-301.

Taylor, M. S., C. P. Ponting and R. R. Copley (2004). "*Occurrence and consequences of coding sequence insertions and deletions in Mammalian genomes*." Genome Res **14**(4): 555-566.

Tazi, J., N. Bakkour and S. Stamm (2009). "*Alternative splicing and disease*." Biochim Biophys Acta **1792**(1): 14-26.

Team, R. C. (2011). *R: A language and environment for statistical computing*. Vienna, Austria, R Foundation for Statistical Computing.

Terrell, J. (1988). " *History as a family tree, history as an entangled bank: constructing images and interpretations of prehistory in the South Pacific.*" Antiquity **62**(237): 642-657.

Terrell, J. E. (2004). "*The 'sleeping giant' hypothesis and New Guinea's place in the prehistory of Greater Near Oceania*." World Archaeol **36**(4): 601-609.

Teschendorff, A. E., U. Menon, A. Gentry-Maharaj, S. J. Ramus, D. J. Weisenberger, H. Shen, M. Campan, H. Noushmehr, C. G. Bell, A. P. Maxwell, et al. (2010). "*Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer*." Genome Res **20**(4): 440-446.

Thanakiatkrai, P. and L. Welch (2011). "*Evaluation of nucleosome forming potentials (NFPs) of forensically important STRs*." Forensic Sci Int Genet **5**(4): 285-290.

Thangaraj, K., A. G. Reddy and L. Singh (2002). "*Is the amelogenin gene reliable for gender identification in forensic casework and prenatal diagnosis?*" Int J Legal Med **116**(2): 121-123.

THDCRG (1993). "*A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Reasearch Group.*" Cell **72**(6): 971-983.

Thieme, H. (1997). "*Lower Palaeolithic hunting spears from Germany*." Nature **385**(6619): 807-810.

Thomson, R., J. K. Pritchard, P. Shen, P. J. Oefner and M. W. Feldman (2000). "*Recent common ancestry of human Y chromosomes: evidence from DNA sequence data*." Proc Natl Acad Sci U S A **97**(13): 7360-7365.

Thorsby, E. (2012). "*The Polynesian gene pool: an early contribution by Amerindians to Easter Island*." Philos Trans R Soc Lond B Biol Sci **367**(1590): 812-819.

Thorvaldsdottir, H., J. T. Robinson and J. P. Mesirov (2013). "*Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration*." Brief Bioinform **14**(2): 178-192.

Tian, D., Q. Wang, P. Zhang, H. Araki, S. Yang, M. Kreitman, T. Nagylaki, R. Hudson, J. Bergelson and J.-Q. Chen (2008). "*Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes*." Nature **455**(7209): 105-108.

Toscanini, U., G. Berardi, A. Gómez and E. Raimondi (2009). "*X-STRs analysis in paternity testing when the alleged father is related to the biological father*." Forensic Sci Int Genet Suppl Ser **2**(1): 234-235.

Trejaut, J. A., T. Kivisild, J. H. Loo, C. L. Lee, C. L. He, C. J. Hsu, Z. Y. Lee and M. Lin (2005). "*Traces of archaic mitochondrial lineages persist in Austronesian-speaking Formosan populations*." PLoS Biol **3**(8): e247.

Trindade-Filho, A., S. Ferreira and S. F. Oliveira (2013). "*Impact of a chromosome X STR Decaplex in deficiency paternity cases*." Genet Mol Biol **36**(4): 507-510.

Tsunoda, T., G. M. Lathrop, A. Sekine, R. Yamada, A. Takahashi, Y. Ohnishi, T. Tanaka and Y. Nakamura (2004). "*Variation of gene-based SNPs and linkage disequilibrium patterns in the human genome*." Hum Mol Genet **13**(15): 1623-1632.

Underhill, P. A. and T. Kivisild (2007). "*Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations*." Annu Rev Genet **41**: 539-564.

Vallone, P. M. and J. M. Butler (2004). "*AutoDimer: a screening tool for primer-dimer and hairpin structures*." Biotechniques **37**(2): 226-231.

Valverde, P., E. Healy, I. Jackson, J. L. Rees and A. J. Thody (1995). "*Variants of the melanocyte-stimulating hormone receptor gene are associated with red hair and fair skin in humans*." Nat Genet **11**(3): 328-330.

van Dijk, E. L., H. Auger, Y. Jaszczyszyn and C. Thermes (2014). "*Ten years of next-generation sequencing technology*." Trends Genet **30**(9): 418-426.

van Neste, C., F. van Nieuwerburgh, D. van Hoofstat and D. Deforce (2012). "*Forensic STR analysis using massive parallel sequencing*." Forensic Sci Int Genet **6**(6): 810-818.

van Oven, M., A. Van Geystelen, M. Kayser, R. Decorte and M. H. D. Larmuseau (2014). "*Seeing the Wood for the Trees: A Minimal Reference Phylogeny for the Human Y Chromosome*." Human Mutation **35**(2): 187-191.

van Schaik, C. P., M. Ancrenaz, G. Borgen, B. Galdikas, C. D. Knott, I. Singleton, A. Suzuki, S. S. Utami and M. Merrill (2003). "*Orangutan cultures and the evolution of material culture*." Science **299**(5603): 102-105.

Varela, I., P. Tarpey, K. Raine, D. Huang, C. K. Ong, P. Stephens, H. Davies, D. Jones, M. L. Lin, J. Teague, et al. (2011). "*Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma*." Nature **469**(7331): 539-542.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, et al. (2001). "*The sequence of the human genome*." Science **291**(5507): 1304-1351.

Visser, M., D. Zubakov, K. N. Ballantyne and M. Kayser (2011). "*mRNA-based skin identification for forensic applications*." Int J Legal Med **125**(2): 253-263.

von Dungern, E. and L. Hirschfeld (1910). "*Ueber Vererbung gruppenspezifischer Strukturen des Blutes*." Z Immunforsch **6**: 284-292.

Wagner, G. A., M. Krbetschek, D. Degering, J. J. Bahain, Q. Shao, C. Falgueres, P. Voinchet, J. M. Dolo, T. Garcia and G. P. Rightmire (2010). "*Radiometric dating of the type-site for Homo heidelbergensis at Mauer, Germany*." Proc Natl Acad Sci U S A **107**(46): 19726-19730.

Walsh, P. S., N. J. Fildes and R. Reynolds (1996). "*Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA*." Nucleic Acids Res **24**(14): 2807-2812.

Walsh, S., A. Lindenbergh, S. B. Zuniga, T. Sijen, P. de Knijff, M. Kayser and K. N. Ballantyne (2011a). "*Developmental validation of the IrisPlex system: Determination of blue and brown iris colour for forensic intelligence*." Forensic Sci Int Genet **5**(5): 464-471.

Walsh, S., F. Liu, K. N. Ballantyne, M. van Oven, O. Lao and M. Kayser (2011b). "*IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information*." Forensic Sci Int Genet **5**(3): 170-180.

Walsh, S., F. Liu, A. Wollstein, L. Kovatsi, A. Ralf, A. Kosiniak-Kamysz, W. Branicki and M. Kayser (2013). "*The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA*." Forensic Sci Int Genet **7**(1): 98-115.

Wang, S., C. M. Lewis, M. Jakobsson, S. Ramachandran, N. Ray, G. Bedoya, W. Rojas, M. V. Parra, J. A. Molina, C. Gallo, et al. (2007). "*Genetic variation and population structure in native Americans*." PLoS Genet **3**(11): e185.

Waples, R. S. and O. Gaggiotti (2006). "*What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity*." Mol Ecol **15**(6): 1419-1439.

Warren, S. T., F. Zhang, G. R. Licameli and J. F. Peters (1987). "*The fragile X site in somatic cell hybrids: an approach for molecular cloning of fragile sites*." Science **237**(4813): 420-423.

Watson, J. D. and F. H. Crick (1953). "*Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid*." Nature **171**(4356): 737-738.

Weber, J. L., D. David, J. Heil, Y. Fan, C. Zhao and G. Marth (2002). "*Human diallelic insertion/deletion polymorphisms*." Am J Hum Genet **71**(4): 854-862.

Weber, J. L. and C. Wong (1993). "*Mutation of human short tandem repeats*." Hum Mol Genet **2**(8): 1123-1128.

Weidner, C. I., Q. Lin, C. M. Koch, L. Eisele, F. Beier, P. Ziegler, D. O. Bauerschlag, K. H. Jockel, R. Erbel, T. W. Muhleisen, et al. (2014). "*Aging of blood can be tracked by DNA methylation changes at just three CpG sites*." Genome Biol **15**(2): R24.

Weinberg, W. (1908). "*Über den Nachweis der Vererbung beim Menschen*." Jahresh Wuertt Ver vaterl Natkd **64**(369-382).

Westen, A. A., A. S. Matai, J. F. J. Laros, H. C. Meiland, M. Jasper, W. J. F. de Leeuw, P. de Knijff and T. Sijen (2009). "*Tri-allelic SNP markers enable analysis of mixed and degraded DNA samples*." Forensic Sci Int Genet **3**(4): 233-241.

Whiten, A., J. Goodall, W. C. McGrew, T. Nishida, V. Reynolds, Y. Sugiyama, C. E. Tutin, R. W. Wrangham and C. Boesch (1999). "*Cultures in chimpanzees*." Nature **399**(6737): 682-685.

Wilkins, J. F. and F. W. Marlowe (2006). "*Sex-biased migration in humans: what should we expect from genetic data?*" Bioessays **28**(3): 290-300.

Wilmshurst, J. M., T. L. Hunt, C. P. Lipo and A. J. Anderson (2011). "*High-precision radiocarbon dating shows recent and rapid initial human colonization of East Polynesia*." Proc Natl Acad Sci U S A **108**(5): 1815-1820.

Winkler, C. A., G. W. Nelson and M. W. Smith (2010). "*Admixture mapping comes of age*." Annu Rev Genomics Hum Genet **11**: 65-89.

Wollstein, A., O. Lao, C. Becker, S. Brauer, R. J. Trent, P. Nurnberg, M. Stoneking and M. Kayser (2010). "*Demographic history of Oceania inferred from genome-wide data*." Curr Biol **20**(22): 1983-1992.

Wolpoff, M. H., J. Hawks and R. Caspari (2000). "*Multiregional, not multiple origins*." Am J Phys Anthropol **112**(1): 129-136.

Wolpoff, M. H., J. N. Spuhler, F. H. Smith, J. Radovcic, G. Pope, D. W. Frayer, R. Eckhardt and G. Clark (1988). "*Modern human origins*." Science **241**(4867): 772-774.

Wolpoff, M. H., W. X. Zhi and A. G. Thorne (1984). *Modern Homo sapiens origins; a general theory of hominid evolution involving the fossil evidence from East Asia*. The origins of modern humans: a world survey of the fossil evidence. F. H. Smith and F. Spencer. New York, Alan Liss**:** 411-483.

Wood, A. R., T. Esko, J. Yang, S. Vedantam, T. H. Pers, S. Gustafsson, A. Y. Chu, K. Estrada, J. Luan, Z. Kutalik, et al. (2014). "*Defining the role of common variation in the genomic and biological architecture of adult human height*." Nat Genet **46**(11): 1173-1186.

Wood, B. and J. Baker (2011). "*Evolution in the Genus Homo*." Annu Rev Ecol Evol Syst **42**: 47-69.

Wood, B. and M. Collard (1999). "*The human genus*." Science **284**(5411): 65-71.

Wood, B. and B. G. Richmond (2000). "*Human evolution: taxonomy and paleobiology*." J Anat **197 (Pt 1)**: 19-60.

Wright, S. (1949). "*The genetical structure of populations*." Ann Hum Genet **15**(1): 323-354.

Wu, R. (1972). "*Nucleotide sequence analysis of DNA*." Nat New Biol **236**(68): 198-200.

Xu, X., M. Peng and Z. Fang (2000). "*The direction of microsatellite mutations is dependent upon allele length*." Nat Genet **24**(4): 396-399.

Yang, N., H. Z. Li, L. A. Criswell, P. K. Gregersen, M. E. Alarcon-Riquelme, R. Kittles, R. Shigeta, G. Silva, P. I. Patel, J. W. Belmont, et al. (2005). "*Examination of ancestry and ethnic affiliation using highly informative diallelic DNA markers: application to diverse and admixed populations and implications for clinical epidemiology and forensic medicine*." Hum Genet **118**(3-4): 382-392.

Yang, S. and R. E. Rothman (2004). "*PCR-based diagnostics for infectious diseases: uses, limitations, and future applications in acute-care settings*." Lancet Infect Dis **4**(6): 337-348.

Yang, Y., B. Xie and J. Yan (2014). "*Application of next-generation sequencing technology in forensic science*." Genomics Proteomics Bioinformatics **12**(5): 190-197.

Yauk, C. L., A. Polyzos, A. Rowan-Carroll, I. Kortubash, A. Williams and O. Kovalchuk (2008). "*Tandem repeat mutation, global DNA methylation, and regulation of DNA methyltransferases in cultured mouse embryonic fibroblast cells chronically exposed to chemicals with different modes of action*." Environ Mol Mutagen **49**(1): 26-35.

Zaumsegel, D., M. A. Rothschild and P. M. Schneider (2013). "*A 21 marker insertion deletion polymorphism panel to study biogeographic ancestry*." Forensic Sci Int Genet **7**(2): 305-312.

Zbiec-Piekarska, R., M. Spolnicka, T. Kupiec, Z. Makowska, A. Spas, A. Parys-Proszek, K. Kucharczyk, R. Ploski and W. Branicki (2015). "*Examination of DNA methylation status of the ELOVL2 marker may be useful for human age prediction in forensic science*." Forensic Sci Int Genet **14**: 161-167.

Zeng, X., J. L. King, M. Stoljarova, D. H. Warshauer, B. L. LaRue, A. Sajantila, J. Patel, D. R. Storts and B. Budowle (2015). "*High sensitivity multiplex short tandem repeat loci analyses with massively parallel sequencing*." Forensic Sci Int Genet **16**: 38-47.

Zeng, Z., D. J. Rowold, R. Garcia-Bertrand, S. Calderon, M. Regueiro, L. Li, M. Zhong and R. J. Herrera (2014). "*Taiwanese aborigines: genetic heterogeneity and paternal contribution to Oceania*." Gene **542**(2): 240-247.

Zhang, Z. and M. Gerstein (2003). "*Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes*." Nucleic Acids Res **31**(18): 5338-5348.

Zhou, H., D. H. Alexander and K. Lange (2011). "*A quasi-Newton acceleration for high-dimensional optimization algorithms*." Stat Comput **21**(2): 261-273.

Zubakov, D., M. Kokshoorn, A. Kloosterman and M. Kayser (2009). "*New markers for old stains: stable mRNA markers for blood and saliva identification from up to 16-year-old stains*." Int J Legal Med **123**(1): 71-74.

Zubakov, D., F. Liu, M. C. van Zelm, J. Vermeulen, B. A. Oostra, C. M. van Duijn, G. J. Driessen, J. J. van Dongen, M. Kayser and A. W. Langerak (2010). "*Estimating human age from T-cell DNA rearrangements*." Curr Biol **20**(22): R970-971.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA