Elsevier Editorial System(tm) for Forensic Science International: Genetics
Manuscript Draft

Manuscript Number: FSIGEN-D-15-00279R1

Title: The Global AIMs Nano set: A 31-plex SNaPshot assay of ancestry-informative SNPs

Article Type: Research Paper

Keywords: SNPs; AIMs; Biogeographical ancestry; SNaPshot; Population-specific Divergence

Corresponding Author: Dr. Chris Phillips, PhD

Corresponding Author's Institution: Forensic Genetics Unit

First Author: Maria de la Puente

Order of Authors: Maria de la Puente; Carla Santos; Manuel Fondevila; Laura Manzo; Ángel Carracedo; María Victoria Lareu; Chris Phillips, PhD

Abstract: A 31-plex SNaPshot assay, named 'Global AIMs Nano', has been developed by reassembling the most differentiated markers of the EUROFORGEN Global AIM-SNP set. The SNPs include three tri-allelic loci and were selected with the goal of maintaining a balanced differentiation of: Africans, Europeans, East Asians, Oceanians and Native Americans. The Global AIMs Nano SNP set provides higher divergence between each of the five continental population groups than previous small-scale AIM sets developed for forensic ancestry analysis with SNaPshot. Both of these characteristics minimise potential bias when estimating co-ancestry proportions in individuals with admixed ancestry; more likely to be observed when using markers disproportionately informative for only certain population group comparisons. The optimised multiplex is designed to be easily implemented using standard capillary electrophoresis regimes and has been used to successfully genotype challenging forensic samples from highly degraded material with low level DNA. The ancestry predictive performance of the Global AIMs Nano set has been evaluated by the analysis of samples previously characterised with larger AIM sets.

Suggested Reviewers:

The Global AIMs Nano set: a 31-plex SNaPshot assay of ancestry-informative SNPs

M. de la Puente[1], C. Santos[1], M. Fondevila[1], L. Manzo[1]; The EUROFORGEN-NoE Consortium; Á. Carracedo[1,2], M.V. Lareu[1], C. Phillips[1]*.

[1] Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain

[2] Center of Excellence in Genomic Medicine Research, King Abdulaziz University, Jeddah, Saudi Arabia

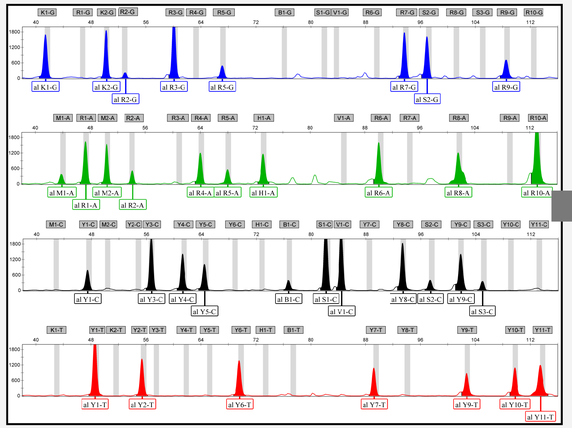
* Corresponding author.

E-mail address: c.phillips@mac.com (C. Phillips).

Graphical Abstract (for review)

G-AIMs Nano

31 SNPs

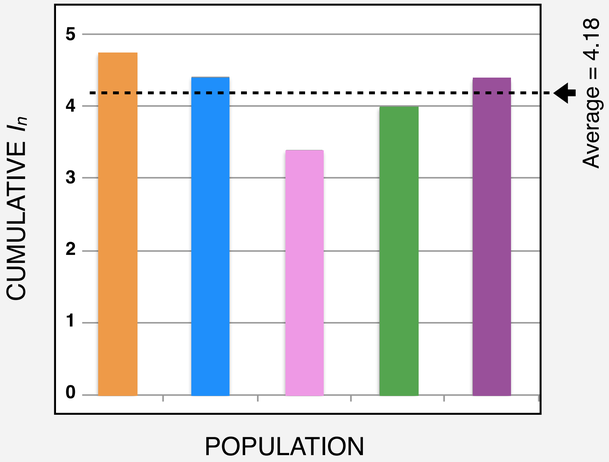


Single CE assay

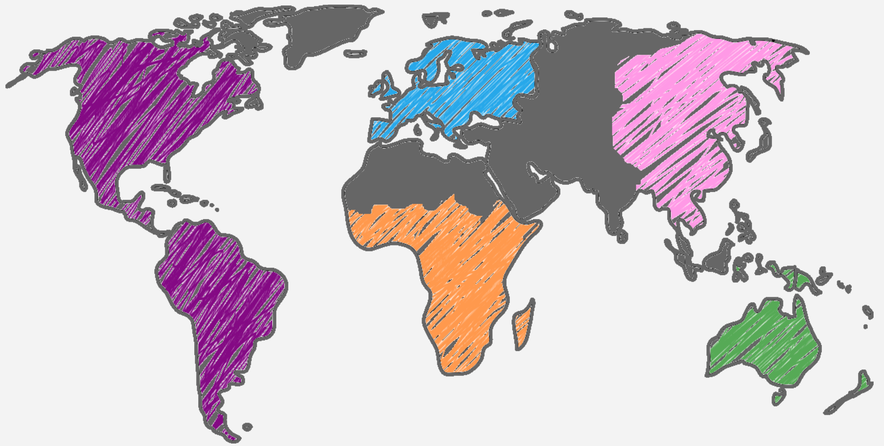
✓ Sensitivity **0.064 ng** of DNA

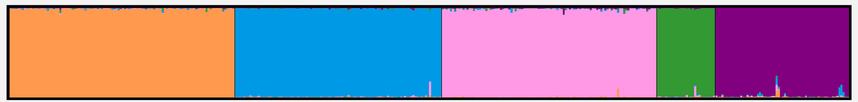
✓ Degraded samples

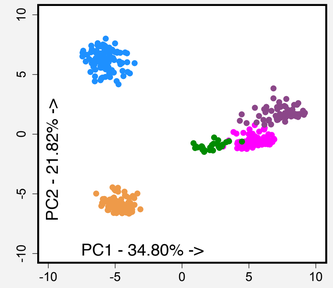
Balanced population-specific Divergence

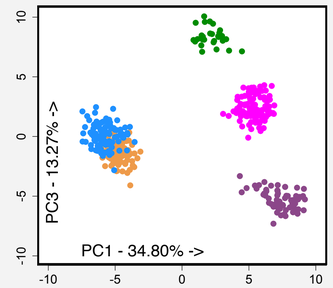


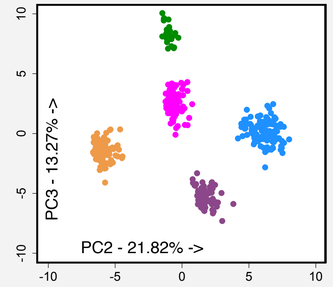
5 GROUPS











The Global AIMs Nano set: a 31-plex SNaPshot assay of ancestry-informative SNPs

● A 31-plex AIM-SNP set was assembled from the most differentiated markers in the Global forensic ancestry panel.

● The SNPs extend population differentiations to Native Americans and Oceanians.

● SNPs preserve balanced population-specific Divergence, reducing estimation bias of individual co-ancestry.

● Assessments of the assay's ability to type degraded or low level DNA indicate good sensitivity for routine forensic use.

● The SNaPshot assay gives a compact but powerful multiplex to infer ancestry for labs yet to adopt massively parallel sequencing regimes.

The Global AIMs Nano set: A 31-plex SNaPshot assay of ancestry-informative SNPs

Abstract

A 31-plex SNaPshot assay, named 'Global AIMs Nano', has been developed by reassembling the most differentiated markers of the EUROFORGEN Global AIM-SNP set. The SNPs include three tri-allelic loci and were selected with the goal of maintaining a balanced differentiation of: Africans, Europeans, East Asians, Oceanians and Native Americans. The Global AIMs Nano SNP set provides higher divergence between each of the five continental population groups than previous small-scale AIM sets developed for forensic ancestry analysis with SNaPshot. Both of these characteristics minimise potential bias when estimating co-ancestry proportions in individuals with admixed ancestry; more likely to be observed when using markers disproportionately informative for only certain population group comparisons. The optimised multiplex is designed to be easily implemented using standard capillary electrophoresis regimes and has been used to successfully genotype challenging forensic samples from highly degraded material with low level DNA. The ancestry predictive performance of the Global AIMs Nano set has been evaluated by the analysis of samples previously characterised with larger AIM sets.

Keywords

SNPs; AIMs; Biogeographical ancestry; SNaPshot; Population-specific Divergence

1. Introduction

Although STR profiling has been successfully applied to the majority of forensic DNA analyses for many years, there are still situations when STR typing is unable to inform criminal investigations, for example, with no matching profile found in DNA database searches or when no suspect is apprehended. For this reason, there is interest in developing DNA tests that can provide investigative leads, focused on panels of single nucleotide polymorphisms (SNPs) to predict external visible characteristics (ECVs), including common variation in pigmentation [1], or to infer an individual's biogeographical ancestry [2].

With the recent availability of bench-top systems for massively parallel sequencing (MPS) that are applicable to forensic DNA analysis, it is now possible to assemble multiplexes of 400-500 markers [3]. Such enlarged forensic multiplexes can include a portion of carefully chosen ancestry informative markers (AIMs), e.g. the Illumina ForenSeq panel [4], or can be exclusively composed of AIMs [5-7]. Both approaches raise the level of geographic resolution that can be obtained from tests that keep the necessary forensic sensitivity. However, the forensic community will take time to adopt, optimise and validate MPS technology as a routine analysis system. Therefore, it is important to continue to develop small-scale AIM sets suited to short-amplicon marker genotyping with validated, universally applicable capillary electrophoresis (CE) analysis regimes [8-10]. However, one drawback with use of small-scale AIM sets is the potential for over-estimation of co-ancestry proportions in individuals with admixed ancestry, stemming from the analysis of genotypes strongly differentiated for some populations but not others. The phenomenon of biased estimation of co-ancestry components was detected in a study of Bolivian populations [11] using 46 ancestry-informative Indels [8] compared with a much larger panel of 446 AIM-SNPs [12]. The Indel set consistently over-estimated European co-ancestry and under-estimated Native American co-ancestry using STRUCTURE-based analyses, indicating that the higher European differentiation of the Indel genotypes inflated the estimates of European co-ancestry proportions. Bearing in mind this effect, construction of a

dedicated AIM-SNP set for MPS by the EUROFORGEN Consortium [5] sought to carefully balance the cumulative population-specific Divergence values for the five continental population groups of Africa, Europe, East Asia, Native America and Oceania.

The emphasis on keeping balanced population group Divergence values provided the main focus for the new ancestry informative SNP panel reported here. We took the most differentiated AIM-SNPs from the EUROFORGEN Global AIMs panel [5] and assembled a compact 31-plex assay genotyped with SNaPshot® single base extension technology. The SNP set, named 'Global AIMs Nano' (herein Nano) was designed to be applicable to forensic analyses where several different admixture combinations may be commonly encountered, e.g. in Australia; where comparisons of European, Oceanian and East Asian co-ancestry components will be routinely necessary. As well as preserving a comparable level of differentiation amongst the five population groups, the Nano assay aimed to provide a single CE-based test that is sufficiently informative for all five groups.

2. Materials and methods

2.1. Reference population SNP genotype data and DNA samples

SNP variation data from representative populations without high levels of admixture was obtained from 1000 Genomes Phase III [13] and from the Stanford University HGDP-CEPH SNP analysis [14] using the SPSmart frequency browser [15]. SNP genotype data was compiled from 108 YRI Africans (AFR: Yoruba in Ibadan, Nigeria); 99 CEU Europeans (EUR: Utah Residents with North and Western European ancestry); 103 CHB East Asians (EAS: Han Chinese in Beijing, China); 28 HGDP-CEPH Oceanians (OCE: 17 Papuan from New Guinea and 11 Melanesian from Bougainville); and 64 HGDP-CEPH Native Americans (AMR: 14 Karitiana, 8 Surui from Brazil; 21 Maya, 14 Pima from Mexico; and 7 Piapoco from Colombia). Phase III 1000 Genomes populations were also analysed, comprising: as a test set, 99 AFR LWK (Luhya in Webuye, Kenya); 113 AFR GWD (Gambian in Western

Divisions in the Gambia); 85 AFR MSL (Mende in Sierra Leone); 99 AFR ESN (Esan in Nigeria); 107 EUR TSI (Toscani in Italia); 99 EUR FIN (Finnish in Finland); 91 EUR GBR (British in England and Scotland); 107 EUR IBS (Iberian Population in Spain); 104 EAS JPT (Japanese in Tokyo, Japan); 105 EAS CHS (Southern Han Chinese); 99 EAS KHV (Kinh in Ho Chi Minh City, Vietnam); 93 EAS CDX (Chinese Dai in Xishuangbanna, China); plus admixed populations 61 ASW (Americans of African Ancestry in SW USA); 96 ACB (African Caribbeans in Barbados); 104 PUR (Puerto Ricans from Puerto Rico), 94 CLM (Colombians from Medellin, Colombia); 64 MXL (individuals with Mexican Ancestry from Los Angeles USA); 85 PEL (Peruvians from Lima, Peru).

To evaluate the forensic sensitivity of the Nano assay, challenging casework samples plus control DNAs were analysed, comprising: (i) five DNA samples each from separate population groups, previously used in an ancestry analysis collaborative exercise [10]; (ii) highly degraded skeletal DNA extracts; (iii) a doubling dilution series of 1 ng/μL; 0.5 ng/μL; 0.25 ng/μL; 0.125 ng/μL; 0.064 ng/μL; 0.032 ng/μL; and 0.016 ng/μL of the 9947A forensic kit DNA standard.

2.2. AIM-SNP selection and SNaPshot assay design

Ancestry-informative SNPs were selected directly from the EUROFORGEN Global AIM-SNP set according to the following criteria: (i) differentiation of five population groups to comparable levels to produce population-specific Divergence (PSD) values as balanced as possible (use of capitalised Divergence distinguishes the metric from the phenomenon of population divergence); (ii) inclusion of certain informative tri-allelic SNPs to allow a level of mixed DNA detection; (iii) genomic separation of component SNPs by a minimum inter-marker distance of 1 Mb to minimise the effects of linkage on likelihood calculations that assume independence for the loci analysed.

From the selected SNPs, a 31-plex SNaPshot® single base extension assay was designed and optimised following established guidelines [16]. Locus

details and summary allele frequencies for component SNPs are summarised in Table 1. PCR and single base extension (SBE) primers are detailed in Supplementary Table S1.

PCR reactions comprised: 1 µL Buffer II (100 mM Tris-HCl, pH 8.3, 500 mM KCl); 1.8 µL 25 mM $MgCl_2$; 0.1 µL AmpliTaq Gold® DNA Polymerase (at 5 U/µL); 0.4 µL of GeneAmp® 10 mM dNTP Mix with dTTP (Applied Biosystems, AB); 1 µL of 3.2 mg/ml bovine serum albumin; 1.5 µL PCR primer mix (Supplementary Table S1); 1 ng of target DNA adjusted to total reaction volume of 10 µL. PCR cycling with GeneAmp® PCR System 9700 or 2700 (AB) thermocyclers used conditions: 10 mins at 95ºC, 32 cycles of 30 secs at 95°C, 40 secs at 62°C and 1 min at 72°C with a final extension of 20 mins at 72°C. PCR primer clean up combined 2.5 µL PCR product with 1 µL of 1 in 3 diluted Illustra™ ExoStar™ 1-Step (GE Healthcare) then incubation at 37°C for 45 mins and enzyme inactivation at 85°C for 15 mins. SBE reactions comprised: 1.25 µL of 1 in 2 diluted SNaPshot® Multiplex Ready Reaction Mix; 0.75 µL SBE primers (Supplementary Table S1) and 1 µL purified PCR product in a total volume of 3 µL. SBE cycling used conditions: 33 cycles of 10 secs at 96°C, 5 secs at 59°C and 30 secs at 60°C. SBE primer clean up combined the full SBE volume with 1 µL of 1 in 2 diluted Illustra™ Shrimp Alkaline Phosphatase (GE Healthcare) then incubating at 37°C for 80 mins and enzyme inactivation at 85°C for 15 mins. Purified SBE products were then prepared for CE detection by adding 1 µL of product to 9.5 µL of Hi-Di™ Formamide (AB) and 0.25 µL of GeneScan™-120 LIZ® Size Standard (AB). Electrophoresis was performed in an ABI Prism 3130*xl* Genetic Analyser, with 36 cm capillary arrays and POP-4™ polymer using standard conditions. Electropherograms were visualised using AB GeneMapper® ID Software v. 3.2.1.

2.3. Analysis of population variation in the selected SNPs

Population-specific Divergence and simple pairwise Divergence values were calculated using the Snipper cross-validation option (http://mathgene.usc.es/snipper/analysispopfile2_new.html) by marking SNP genotype profiles as AFR and non-AFR, etc., or by comparing each pair of

population groups in turn. Output from Snipper lists Shannon's Divergence values for each SNP from the comparisons made by cross-validation [17]. These values were converted to the more widely-used Rosenberg's informativeness-for-assignment metric: I_n [18], by multiplication with 0.693 (i.e. converting the natural log to ln(2)). The Snipper portal was also used to cross-validate the reference population data or calculate classification likelihood ratios (LRs) by uploading an Excel file of reference data (provided ready to use as Supplementary File S1) or by choosing analysis options available in the portal.

Population analyses with STRUCTURE v. 2.3.4 [19] were performed following recommendations outlined previously [20]. Parameters comprised: five iterations (for K=1 to K=9) of 100,000 burnin steps and 100,000 MCMC steps, correlated allele frequencies under the Admixture model (no POPFLAG for just reference populations and POPFLAG for analyses of reference populations plus test or admixed populations). The optimum K value was estimated by computing results with STRUCTURE HARVESTER [21] and following previous guidelines [22]. Ancestry membership plots were constructed with CLUMPAK v. 1.1 [23] or a combination of CLUMPP v. 1.1.2 [24] and distruct v. 1.1 [25]. PCA analysis was performed using R software v. 3.1.2 [26] and executing a homemade script. F_{ST} calculations and graphics were computed using Arlequin v. 3.5 [27].

To assess the ancestry inference performance of the Nano SNP set, comparisons were made with two previously developed biallelic AIM sets comprising 46 Indels [8] and 34 SNPs [9], by applying STRUCTURE and PCA analyses of the same 1000 Genomes African, European and East Asian genotypes plus HGDP-CEPH Native American and Oceanian genotypes compiled from each set (data used from 44 of 46 Indels currently listed by 1000 Genomes Phase III).

3. Results

3.1. Characteristics and PSD balance of the Nano SNP set

198 The 31 SNPs selected showed highly contrasting allele frequency
199 distributions. In each of the 28 biallelic SNPs one allele was close to fixation
200 (allele frequencies between 0.9 and 1) in at least one population group, as
201 shown by the raster plot of Fig. 1. Population group summary allele frequency
202 pie charts are also shown in Supplementary Fig. S1. All SNPs were distributed
203 in the genome with sufficient distance between syntenic marker sets to be free
204 from the effects of close physical linkage (Supplementary Fig. S2).
205
206 With the reduction in scale from an MPS multiplex of 128 SNPs to the
207 SNaPshot 31-plex, it is important to ensure the cumulative PSD values remain
208 at comparable levels for each group. Reference population comparisons
209 produced the cumulative PSD values listed in Table 1 and summarised in Fig.
210 2. These indicate the reduced East Asian PSD of 3.39 was noticeably lower
211 than the average of 4.18 and the African PSD of 4.74 was the highest but
212 comparable to three population groups. The reduced differentiation of East
213 Asians from Native American and Oceanian populations is illustrated by their
214 similar allele frequency distributions for many SNPs, exemplified by rs4657449
215 and rs9809818. The close relationship of East Asian and Native American
216 population variability is highlighted by Fig. 1, with little divergence between the
217 two groups evident in rs3827760, rs6437783 and rs12594144. This suggests
218 one or two extra East Asian-informative SNPs could address this slight PSD
219 imbalance in future adjustments of the Nano SNP set.
220
221 The F_{ST} and pairwise genotype difference data from Arlequin analyses
222 are summarised in Supplementary Fig. S3. Results indicate a high average
223 number of pairwise genotype differences between the population groups and
224 consequently F_{ST} values are low within-groups and high between-groups.
225 Admixed populations from 1000 Genomes give high within-population average
226 number of pairwise genotype differences and F_{ST} values, as would be expected
227 from the complex patterns of variation that are characteristic of population
228 admixture.
229
230 *3.2. Ancestry inference capabilities of the Nano SNP set*
231

Cross-validation of reference populations gave 100% ancestry assignment success (Supplementary Table S2). Additionally, assignment success remained at 100% for all populations when excluding the 14 most informative SNPs (those with highest overall Divergence values), indicating this marker set maintains a high level of informativeness even when many components fail to be reliably genotyped, e.g. analysing low-level DNA.

STRUCTURE analysis of reference population data (no POPFLAG) produced a pattern of five distinct clusters matching the known origin of the individuals. The 34 SNP and 46 Indel forensic ancestry assays with which Nano was compared also distinguish the five genetic clusters, but in contrast, both give estimated optimum numbers of clusters below five (Supplementary Fig. S4). The PCA plots shown in Fig. 3A reveal improved separation and clusteredness of populations from the Nano set compared to the other two assays. This is especially evident in the PC2 vs. PC3 plots for Nano genotypes, where no set of points overlap and the five population groups have almost equidistant cluster positions.

Analysis of other genotype data from 1000 Genomes which were marked as ‘study populations’ (POPFLAG=0) gave patterns consistent with analyses made during the development of the 128 Global ancestry set [5] or subsequent analysis of admixed 1000 Genomes ACB, ASW, PEL, MXL, PUR, CLM samples using the same SNPs (Fig. 7 in [2]). The STRUCTURE cluster plots (Fig. 3B) in particular match well with results of both previous analyses using many more SNPs, indicating PEL have the highest Native American co-ancestry proportions and PUR show predominantly European co-ancestry. The PCA plots arranged individually along with STRUCTURE cluster plots for the six admixed 1000 Genomes populations in Fig. 3C also give very similar cluster distributions that are positioned between, or close to, the expected population admixture contributor clusters.

As an additional simple gauge of the ancestry informativeness of the 31 Nano SNPs, this assay was used to analyse the control DNAs used in a collaborative EDNAP exercise that assessed the 34 SNPs and 46 Indels

described above. The five control DNAs each have confirmed ancestry from one of the five continental regions and they were analysed using Bayes analyses and PCA in Snipper. The five control DNAs are positioned in the middle of their respective population group clusters describing the correct ancestry in each case. The Bayes analysis likelihoods obtained from Snipper are listed in Table 2. Moreover, as shown in Supplementary Fig. S5, high likelihood ratio values can be obtained from partial profiles even when the fourteen most informative markers are missing.

3.3. Forensic performance of the Nano SNPs set

A typical SNaPshot profile from analysis of 1 ng of the 9947A control DNA with the Nano assay is shown in Fig. 4. The dilutions series of 9947A gave full profiles with 0.5, 0.25, 0.125 and 0.064 ng of DNA. Locus and allele drop out occurred with 0.032 and 0.016 ng of input DNA, however >80% profile completeness was obtained for these analyses.

The sensitivity of the Nano assay was assessed with paternity test samples, comprising biopsy, bones (cranium, femur and tibia) and teeth identified as degraded or PCR-inhibited (~35% to 95% STR profile completeness). Nano profile completeness ranged from ~20% to 70% and produced ancestry assignment likelihoods above 60,000.

4. Discussion

When originally rebuilding a SNP-based forensic ancestry multiplex for MPS analysis, we started to recognise subsets of the most informative markers that would be well suited to development of smaller SNaPshot tests. The present study describes the compilation of 31 SNPs, mainly representing the most informative markers from the full set of 128 Global AIMs. These SNPs maintain the capacity to differentiate five continentally-defined population groups. Therefore, the Nano assay extends the three group comparisons possible with existing SNaPshot forensic ancestry tests [9,28-30] to the two additional population groups of Native Americans and Oceanians. Each of

these population groups contribute to admixture patterns commonly seen in large parts of the regions they occupy. For this reason, we prioritised the preservation of balanced PSD, although this was difficult to maintain when accomplishing the 75% reduction in multiplex size. One outcome of this process was a disproportionate lowering of the East Asian cumulative PSD compared to the other groups that we aim to address by careful choice of 1-2 additional AIMs.

The selection of SNPs informative for Native American and Oceanian populations requires use of much smaller population sample sizes from the HGDP-CEPH panel compared to those of 1000 Genomes. Therefore, SNP ascertainment bias could reduce the power of the 31 SNPs to differentiate novel populations not yet characterised from each of these regions. However, such bias is unlikely to lead to the discovery of new SNPs as divergent as the 22 Native American informative and 28 Oceanian informative SNPs assembled in the original 128-plex set. The inclusion in the 31-plex Nano set of the five most informative SNPs for both Native Americans and Oceanians, comprising SNPs with alleles near to fixation, ensures the Nano set is almost equally informative for all five groups and the assignment likelihoods obtained for populations from the two additional population groups exceed the values possible with previous AIM-SNP sets developed for SNaPshot genotyping.

When the same control samples with known ancestries are tested with established multiplexes of 34 SNPs, 46 Indels and the Nano 31-plex, higher assignment likelihoods are obtained for the 31 SNPs than 80 markers combined in three of five population groups, and all likelihoods exceed those obtained from 34-plex SNP data by considerable margins (between 3 and 16 orders of magnitude). Therefore, for the bulk of forensic samples that require an ancestry analysis in laboratories without MPS systems in place, the Nano SNP set represents the best option as a stand-alone CE test. The use of the Nano SNPs alongside Indels, with their enhanced capacity to detect mixed DNA profiles [8], will provide a particularly powerful approach to forensic ancestry analysis from the use of conventional capillary electrophoresis techniques, already optimised for routine DNA analyses in every forensic laboratory.

In conclusion, the Nano ancestry assay has brought together a highly informative set of markers with well-balanced population-specific Divergence for the five population groups it is designed to analyse. This characteristic minimises the co-ancestry proportion estimation bias when analysing admixed samples. Moreover, the single SNaPshot reaction shows high sensitivity (complete profiles obtained down to 64 pg of input DNA), with the potential to analyse degraded samples, making it an ideal forensic ancestry assay for the full range of casework applications where DNA is analysed with capillary electrophoresis.

Acknowledgements

This work was funded by the EUROFORGEN Node of Excellence (Grant Agreement No. 285487). MdlP is supported by funding awarded by the Consellería de Cultura, Educación e Ordenación Universitaria of the Xunta de Galicia as part of the Plan Galego de Investigación, Innovación e Crecemento 2011-2015 (Plan I2C). CS is supported by 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). MVL was supported by funding from Xunta de Galicia, Incite 09208163 PR.

References

[1] M. Kayser, Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes, Forensic Science International: Genetics 18 (2015) 33-48.

[2] C. Phillips, Forensic genetic analysis of bio-geographical ancestry, Forensic Science International: Genetics 18 (2015) 49-65.

[3] B. Budowle, D.H. Warshauer, S.B. Seo, et al., Deep sequencing provides comprehensive multiplex capabilities, Forensic Science International: Genetics Supplement Series 4 (2013) e334-e335.

[4] Illumina, Forenseq DNA signature prep kit (2014) http://www.illumina.com/products/forenseq-dna-signature-kit.ilmn.

[5] C. Phillips, W. Parson, B. Lundsberg, et al., Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set, Forensic Sci Int Genet 11 (2014) 13-25.

[6] K.K. Kidd, W.C. Speed, A.J. Pakstis, et al., Progress toward an efficient panel of SNPs for ancestry inference, Forensic Science International: Genetics 10 (2014) 23-32.

[7] J. Kidd, F. Friedlaender, W. Speed, et al., Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples, Investigative Genetics 2 (2011) 1-13.

[8] R. Pereira, C. Phillips, N. Pinto, et al., Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing, PLoS One 7 (2012) e29684.

[9] M. Fondevila, C. Phillips, C. Santos, et al., Revision of the SNP*for*ID 34-plex forensic ancestry test: Assay enhancements, standard reference sample genotypes and extended population studies, Forensic Sci Int Genet 7 (2013) 63-74.

[10] C. Santos, M. Fondevila, D. Ballard, et al., Forensic ancestry analysis with two capillary electrophoresis ancestry informative marker (AIM) panels: Results of a collaborative EDNAP exercise, Forensic Science International: Genetics 19 (2015) 56-67.

[11] P. Taboada-Echalar, V. Álvarez-Iglesias, T. Heinz, et al., The genetic legacy of the pre-colonial period in contemporary Bolivians, PLoS ONE 8 (2013) e58980.

[12] J.M. Galanter, J.C. Fernandez-Lopez, C.R. Gignoux, et al., Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas, PLoS Genet 8 (2012) e1002554.

[13] An integrated map of genetic variation from 1,092 human genomes, Nature 491 (2012) 56-65.

[14] J.Z. Li, D.M. Absher, H. Tang, et al., Worldwide human relationships inferred from genome-wide patterns of variation, Science 319 (2008) 1100-1104.

[15] J. Amigo, A. Salas, C. Phillips, et al., SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access, BMC Bioinformatics 9 (2008) 428.

[16] J.J. Sanchez, P. Endicott, Developing multiplexed SNP assays with special reference to degraded DNA templates, Nat Protocols 1 (2006) 1370-1378.

[17] C. Phillips, A. Salas, J.J. Sánchez, et al., Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs, Forensic Science International: Genetics 1 (2007) 273-280.

[18] N.A. Rosenberg, L.M. Li, R. Ward, et al., Informativeness of Genetic Markers for Inference of Ancestry, American Journal of Human Genetics 73 (2003) 1402-1422.
[19] J.K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data, Genetics 155 (2000) 945-959.
[20] L. Porras-Hurtado, Y. Ruiz, C. Santos, et al., An overview of STRUCTURE: applications, parameter settings, and supporting software, Front Genet 4 (2013) 98.
[21] D. Earl, B. vonHoldt, STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method, Conservation Genetics Resources 4 (2012) 359-361.
[22] G. Evanno, S. Regnaut, J. Goudet, Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study, Mol Ecol 14 (2005) 2611-2620.
[23] N.M. Kopelman, J. Mayzel, M. Jakobsson, et al., Clumpak: a program for identifying clustering modes and packaging population structure inferences across K, Mol Ecol Resour (2015) doi: 10.1111/1755-0998.12387. [Epub ahead of print].
[24] M. Jakobsson, N.A. Rosenberg, CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure, Bioinformatics 23 (2007) 1801-1806.
[25] N.A. Rosenberg, distruct: a program for the graphical display of population structure, Molecular Ecology Notes 4 (2004) 137-138.
[26] R.C. Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2014.
[27] L. Excoffier, H.E. Lischer, Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows, Mol Ecol Resour 10 (2010) 564-567.
[28] Y.-L. Wei, L. Wei, L. Zhao, et al., A single-tube 27-plex SNP assay for estimating individual ancestry and admixture from three continents, International Journal of Legal Medicine (2015) 1-11.
[29] U. Rogalla, E. Rychlicka, M.V. Derenko, et al., Simple and cost-effective 14-loci SNP assay designed for differentiation of European, East Asian and African samples, Forensic Science International: Genetics 14 (2015) 42-49.
[30] O. Lao, P.M. Vallone, M.D. Coble, et al., Evaluating Self-declared Ancestry of U.S. Americans with Autosomal, Y-chromosomal and Mitochondrial DNA, Human Mutation 31 (2010) E1875-E1893.

Figure legends

Fig. 1. Raster plot summarising the allele frequency distributions of the Nano SNPs in five population groups. AFR: African; EUR: European; EAS: East Asian; OCE: Oceanian; AMR: Native American.

Fig. 2. Bar charts indicating cumulative, population-specific and pairwise I_n values for Nano SNPs.

Fig. 3. (A) Three AIM panels compared with PCA analyses (triallelic SNPs not included in analyses, so Nano=28 SNPs, 34-plex=32 SNPs and AIM-indels=44 markers) for reference samples. PC: principal component; AFR: African; EUR: European; EAS: East Asian; OCE: Oceanian; AMR: Native American. Control samples with known ancestry are shown as black points. (B) STRUCTURE analysis for admixed and non-admixed populations included in the study. (C) Mixed populations are represented below in black in the corresponding PCA analyses (PC1 vs PC2). AFR: African; EUR: European; EAS: East Asian; OCE: Oceanian; AMR: Native American; ESN: Esan in Nigeria; MSL: Mende in Sierra Leone; GWD: Gambian in Western Divisions in the Gambia; LWK: Luhya in Webuye, Kenya; ACB: African Caribbeans in Barbados; ASW: Americans of African Ancestry in SW USA; GBR: British in England and Scotland; TSI: Toscani in Italia; FIN: Finnish in Finland; IBS: Iberian Population in Spain; JPT: Japanese in Tokyo, Japan; CHS: Southern Han Chinese; CDX: Chinese Dai in Xishuangbanna, China; KHV: Kinh in Ho Chi Minh City, Vietnam; PEL: Peruvians from Lima, Peru); MXL: Mexican Ancestry from Los Angeles, USA; CLM: Colombians from Medellin, Colombia and PUR: Puerto Ricans from Puerto Rico.

Fig. 4. Electropherogram of 31 Nano SNP genotypes obtained from the control DNA 9947A.

Table 1. Description, reference allele frequencies and population-specific/pairwise Divergence values (I_n) of the 31 Nano SNPs. Chr: chromosome; RA: reference allele. All positions from genome build 37.1 (GRCh37). SNPs are ranked according to their Pop vs. Other Pop I_n (highlighted in grey) inside each population informative (Informat.) group.

SNP details					Reference allele frequency					Population-specific Divergence					Pairwise Divergence									
Informat.	SNP ID	Chr	Position	RA	AFR	EUR	EAS	OCE	AMR	AFR	EUR	EAS	OCE	AMR	AFR - EUR	AFR - EAS	AFR - OCE	AFR - AMR	EUR - EAS	EUR - OCE	EUR - AMR	EAS - OCE	EAS - AMR	OCE - AMR
AFR	**rs2814778**	1	159174683	A	0.005	1.000	1.000	1.000	0.992	0.672	0.131	0.134	0.083	0.108	0.663	0.663	0.634	0.656	0.000	0.672	0.000	0.000	0.000	0.001
	rs1871534	8	145639681	C	0.981	0.000	0.000	0.000	0.000	0.641	0.128	0.131	0.081	0.107	0.631	0.632	0.603	0.624	0.000	0.641	0.000	0.000	0.000	0.000
	rs2789823	9	136769888	G	0.935	0.000	0.000	0.000	0.000	0.565	0.121	0.123	0.076	0.100	0.556	0.557	0.527	0.548	0.000	0.565	0.000	0.000	0.000	0.000
EUR	**rs1426654**	15	48426484	A	0.014	1.000	0.029	0.000	0.039	0.117	0.622	0.093	0.081	0.069	0.641	0.001	0.000	0.003	0.611	0.117	0.595	0.001	0.000	0.002
	rs16891982	5	33951693	C	1.000	0.020	0.985	1.000	0.984	0.126	0.620	0.104	0.073	0.087	0.629	0.001	0.000	0.002	0.606	0.126	0.603	0.000	0.000	0.000
	rs12142199	1	1249187	G	0.977	0.177	0.971	1.000	1.000	0.076	0.402	0.068	0.061	0.083	0.393	0.000	0.000	0.002	0.383	0.076	0.423	0.001	0.003	0.000
	rs8072587	17	19211073	C	0.986	0.182	1.000	1.000	0.817	0.099	0.358	0.113	0.069	0.003	0.405	0.001	0.000	0.047	0.425	0.099	0.218	0.000	0.058	0.043
	rs9522149	13	111827167	T	0.972	0.237	0.995	1.000	0.977	0.063	0.353	0.092	0.055	0.056	0.334	0.004	0.001	0.000	0.377	0.063	0.341	0.002	0.003	0.000
	rs4749305	10	28391596	A	0.389	0.909	0.078	0.036	0.008	0.001	0.309	0.095	0.100	0.155	0.162	0.072	0.105	0.141	0.404	0.001	0.514	0.004	0.017	0.005
EAS	**rs17822931**	16	48258198	C	1.000	0.869	0.029	0.875	0.650	0.190	0.053	0.432	0.039	0.000	0.039	0.613	0.037	0.129	0.428	0.190	0.034	0.434	0.251	0.036
	rs1229984	4	100239319	A	0.000	0.015	0.709	0.071	0.000	0.089	0.070	0.320	0.018	0.071	0.002	0.336	0.018	0.000	0.313	0.089	0.001	0.238	0.328	0.015
	rs3827760	2	109513601	T	1.000	1.000	0.063	0.946	0.109	0.219	0.209	0.319	0.098	0.203	0.000	0.559	0.012	0.500	0.558	0.219	0.499	0.471	0.003	0.415
	rs6437783	3	108172817	C	0.259	0.146	0.995	0.589	0.891	0.078	0.153	0.268	0.001	0.104	0.010	0.359	0.057	0.223	0.459	0.078	0.311	0.157	0.031	0.062
	rs12594144	15	64161351	C	1.000	0.889	0.121	0.607	0.177	0.237	0.097	0.222	0.000	0.130	0.032	0.487	0.149	0.430	0.334	0.237	0.283	0.136	0.003	0.101
	rs4657449	1	165465281	G	0.912	0.909	0.102	0.000	0.117	0.176	0.164	0.177	0.210	0.130	0.000	0.380	0.497	0.364	0.376	0.176	0.360	0.017	0.000	0.022
OCE	**rs9908046**	17	53563782	C	0.958	0.929	0.883	0.018	0.992	0.020	0.008	0.000	0.528	0.042	0.002	0.010	0.562	0.006	0.003	0.020	0.015	0.463	0.030	0.626
	rs3751050	11	9091244	A	0.972	0.924	0.966	0.089	0.961	0.020	0.002	0.016	0.451	0.011	0.006	0.000	0.477	0.000	0.004	0.020	0.003	0.467	0.000	0.459
	rs2139931	1	84590527	A	0.898	0.753	0.879	0.018	0.898	0.017	0.002	0.011	0.433	0.014	0.019	0.000	0.480	0.000	0.013	0.017	0.019	0.458	0.000	0.480
	rs715605	22	30640308	T	0.866	0.914	0.985	0.089	1.000	0.000	0.003	0.039	0.422	0.041	0.003	0.030	0.345	0.036	0.015	0.000	0.020	0.502	0.001	0.516
	rs6054465	20	6673018	T	0.972	0.859	0.743	0.036	0.859	0.062	0.005	0.005	0.408	0.004	0.022	0.061	0.553	0.022	0.011	0.062	0.000	0.306	0.011	0.408
	rs9809818	3	71480566	C	0.019	0.116	0.869	0.982	0.820	0.245	0.120	0.172	0.227	0.102	0.021	0.446	0.603	0.399	0.319	0.245	0.276	0.026	0.002	0.042
AMR	**rs12498138**	3	121459589	G	1.000	0.949	0.922	0.911	0.094	0.085	0.034	0.020	0.011	0.443	0.011	0.020	0.024	0.519	0.002	0.085	0.437	0.000	0.401	0.387
	rs10483251	14	21671277	G	0.921	0.798	0.898	0.712	0.024	0.055	0.006	0.040	0.000	0.429	0.016	0.001	0.038	0.497	0.010	0.055	0.370	0.028	0.469	0.301
	rs2080161	7	13331150	T	0.981	0.758	0.689	0.820	0.000	0.144	0.007	0.001	0.036	0.424	0.064	0.091	0.044	0.624	0.003	0.144	0.366	0.033	0.314	0.462
	rs8137373	22	41729216	G	0.833	0.707	0.927	0.982	0.023	0.020	0.000	0.068	0.093	0.406	0.011	0.011	0.037	0.402	0.043	0.020	0.298	0.009	0.506	0.593
	rs1557553	22	44760984	C	0.949	0.904	0.714	0.786	0.094	0.076	0.042	0.000	0.002	0.325	0.004	0.053	0.031	0.436	0.030	0.076	0.380	0.003	0.219	0.270
	rs12402499	1	101528954	G	1.000	0.919	1.000	1.000	0.258	0.061	0.007	0.059	0.032	0.324	0.021	0.000	0.000	0.361	0.021	0.061	0.252	0.000	0.361	0.335
	rs4792928	17	42105174	T	1.000	0.960	0.345	0.804	0.195	0.171	0.111	0.108	0.011	0.178	0.008	0.297	0.064	0.413	0.239	0.171	0.350	0.113	0.014	0.199
Triallelic	**rs2069945**	20	33761837	CG	0.153 / 0.796	0.480 / 0.420	0.680 / 0.277	0.036 / 0.536	0.766 / 0.234	0.114	0.002	0.045	0.201	0.087	0.078	0.156	0.118	0.210	0.022	0.114	0.065	0.289	0.017	0.384
	rs4540055	4	38803255	AC	0.069 / 0.514	0.793 /0.010	0.301 / 0.068	0.536 / 0.250	0.605 / 0.000	0.217	0.148	0.055	0.024	0.084	0.355	0.147	0.142	0.300	0.130	0.217	0.027	0.098	0.063	0.101
	rs5030240	11	32424389	CA	0.278 / 0.389	0.712 /0.055	0.228 / 0.039	0.093 / 0.278	0.258 / 0.023	0.083	0.117	0.062	0.062	0.054	0.125	0.123	0.052	0.135	0.135	0.083	0.124	0.067	0.001	0.085
								CUMULATIVE VALUES		**4.739**	**4.404**	**3.392**	**3.986**	**4.374**	**5.263**	**6.111**	**6.210**	**8.029**	**6.274**	**4.739**	**7.184**	**4.323**	**3.106**	**6.350**

Table 2. Ancestry assignment likelihood ratios from a five-population group comparison in Snipper for control samples A-E of known ancestry. All assignments were correct except for Sample A misclassified as AMR with 34-plex data. Data for 34-plex, AIM-indel markers and the 80 markers combined was obtained directly using the Snipper portal options and for Nano AIMs, by applying the data in Supplementary File S1 as a custom training set.

		Assignment likelihood ratios (from two highest likelihoods)			
Sample	Known Ancestry	34-plex	AIM indels	80 Markers	G-AIMs Nano
A	East Asian (EAS)	6.8E+00 more likely to be AMR*	5.5E+06 more likely to be EAS	9.7E+06 EAS	2.9E+16 EAS
B	European (EUR)	4.4E+16 EUR	1.7E+11 EUR	1.0E+28 EUR	1.3E+29 EUR
C	Oceanian (OCE)	1.0E+07 OCE	4.0E+07 OCE	1.5E+14 OCE	1.5E+16 OCE
D	Native American (AMR)	1.0E+05 AMR	1.2E+09 AMR	1.1E+14 AMR	4.2E+08 AMR
E	African (AFR)	6.1E+19 AFR	2.8E+21 AFR	3.6E+40 AFR	2.0E+29 AFR

*Likelihood ratio of AMR / EAS likelihoods

Supplementary material legends

Supplementary Fig. S1. Pie charts indicating the frequency of each Nano SNP allele in five population groups (AFR: Africans; EUR: Europeans; EAS: East Asians; OCE: Oceanians; AMR: Native Americans).

Supplementary Fig. S2. Autosomal chromosome ideograms representing the genomic positions of the Nano SNPs.

Supplementary Fig. S3. Pairwise F_{ST} and number of pairwise genotype differences between and within populations. AFR: African; EUR: European; EAS: East Asian; OCE: Oceanian; AMR: American; ESN: Esan in Nigeria; MSL: Mende in Sierra Leone; GWD: Gambian in Western Divisions in the Gambia; LWK: Luhya in Webuye, Kenya; ACB: African Caribbeans in Barbados; ASW: Americans of African Ancestry in SW USA; GBR: British in England and Scotland; TSI: Toscani in Italia; FIN: Finnish in Finland; IBS: Iberian Population in Spain; JPT: Japanese in Tokyo, Japan; CHS: Southern Han Chinese; CDX: Chinese Dai in Xishuangbanna, China; KHV: Kinh in Ho Chi Minh City, Vietnam; PEL: Peruvians from Lima, Peru); MXL: Mexican Ancestry from Los Angeles, USA; CLM: Colombians from Medellin, Colombia and PUR: Puerto Ricans from Puerto Rico. Reference populations are highlighted in red, and admixed populations in blue.

Supplementary Fig. S4. Three panel comparison of STRUCTURE analysis (Nano - 31 SNPs, 34-plex - 34 SNPs and AIM-indels – 44 of 46 markers) for the reference population samples. Graphics on the right indicate the mean of estimated Ln probability-of-data and Delta K according to Evanno's method across different numbers of clusters (K). Red arrows indicate the most probable number of clusters taking into account both graphs and black arrows indicate other probable numbers of clusters represented on the left. K: number of clusters; AFR: African; EUR: European; EAS: East Asian; OCE: Oceanian; AMR: Native American.

Supplementary Fig. S5. Likelihoods ratios of ancestry estimation for control samples of known ancestry (A-E) and 9947A, removing sequentially the most informative markers. The grey box represents a threshold likelihood value of 1,000 and the red line indicates the most informative fourteen SNPs.

Supplementary Table S1. PCR and SBE primer sequences and their mixture concentrations for the 31-SNP Nano multiplex reactions. Internal locus identifiers use the IUPAC code for the extended base and a number indicating the position in the multiplex. The SBE primer sequence and non-specific mobility modifying tails are distinguished by upper and lower case nucleotides respectively.

Supplementary Table S2. Cross-validation analyses from Snipper for the 31 Nano SNPs. Grey-highlighted values correspond to the percentage of correctly classified individuals from the reference training set of each population.

Supplementary File S1. Custom training set file that can be uploaded unmodified for five-group analyses in Snipper. SNP rs-numbers and internal codes are provided as identifiers of the markers, genotypes are provided in the strand direction that allows a direct classification when using the primers listed in Supplementary Table S1. To classify a profile: choose the '*Classification with a custom Excel file of populations'* option; upload this Excel file in '*Data input (population)'*; choose *'Naive Bayes (Hardy-Weinberg principle applies)'* as the classifier and input the genotype profile of the individual to classify in the same SNP order as this file.

Figure 1


rs2789823
rs1871534
rs2814778
rs1426654
rs16891982
rs12142199
rs8072587
rs9522149
rs4749305
rs17822931
rs1229984
rs3827760
rs6437783
rs12594144
rs4657449
rs9908046
rs3751050
rs2139931
rs715605
rs6054465
rs9809818
rs12498138
rs10483251
rs2080161
rs8137373
rs1557553
rs12402499
rs4792928
rs2069945
rs4540055
rs5030240
AFR
EUR
EAS
OCE
AMR
Biallelic SNPs
Homozygous population specific informative allele
Homozygous other allele
Heterozygous
No data
Triallelic SNPs
Homozygous allele 1
Homozygous allele 2
Homozygous allele 3
Heterozygous 1/2
Heterozygous 1/3
Heterozygous 2/3

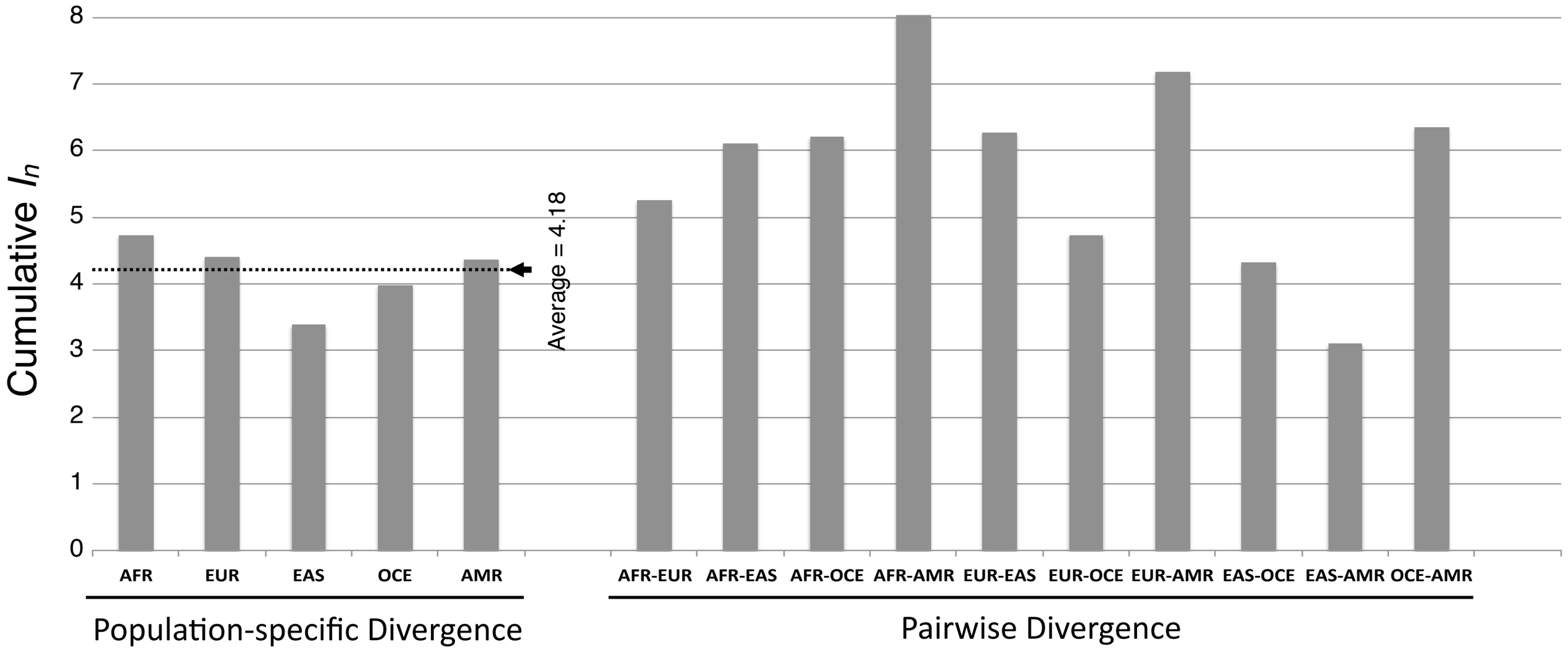

Cumulative I_n
8
7
6
5
4
3
2
1
0
Average = 4.18
AFR
EUR
EAS
OCE
AMR
AFR-EUR
AFR-EAS
AFR-OCE
AFR-AMR
EUR-EAS
EUR-OCE
EUR-AMR
EAS-OCE
EAS-AMR
OCE-AMR
Population-specific Divergence
Pairwise Divergence

Figure 3

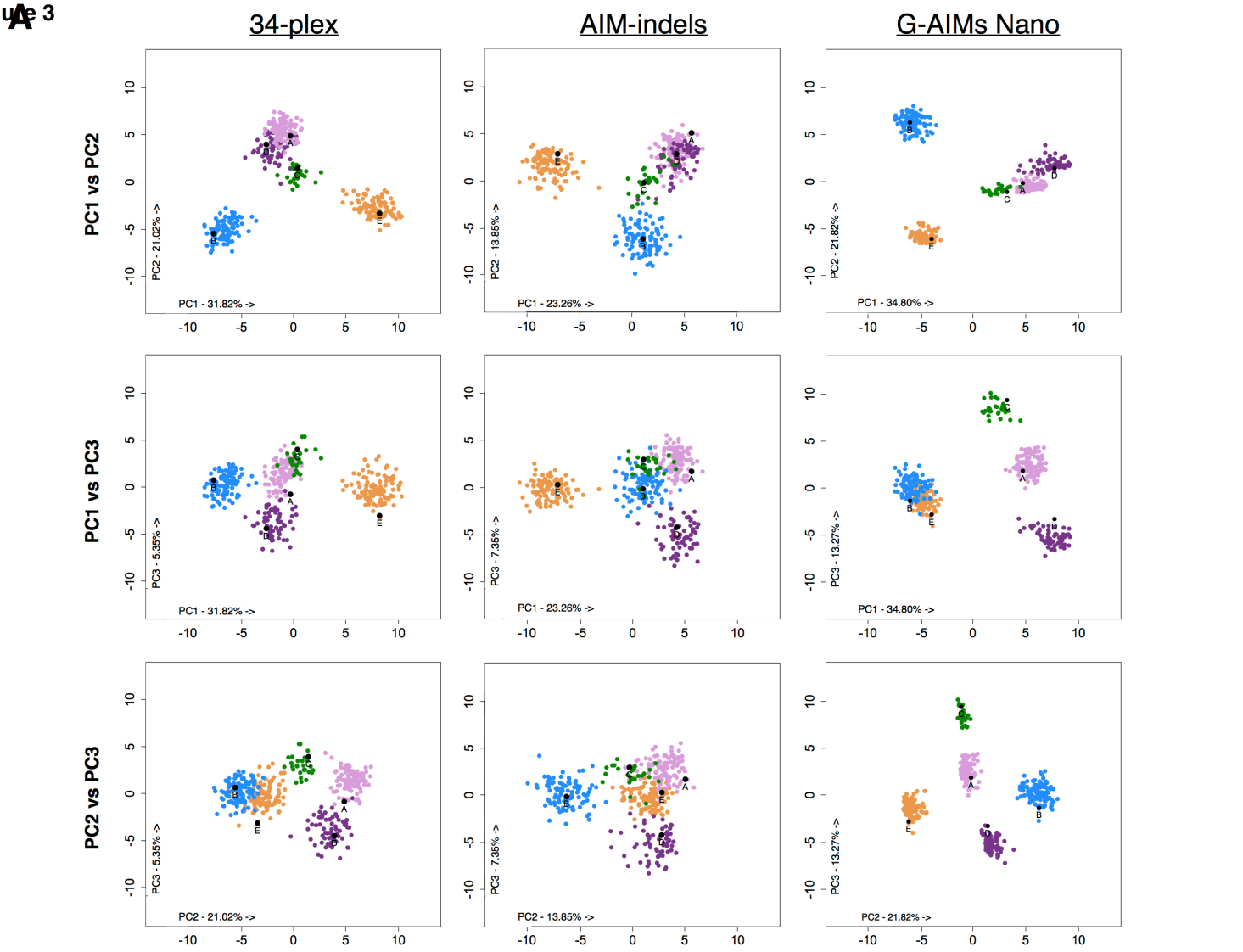
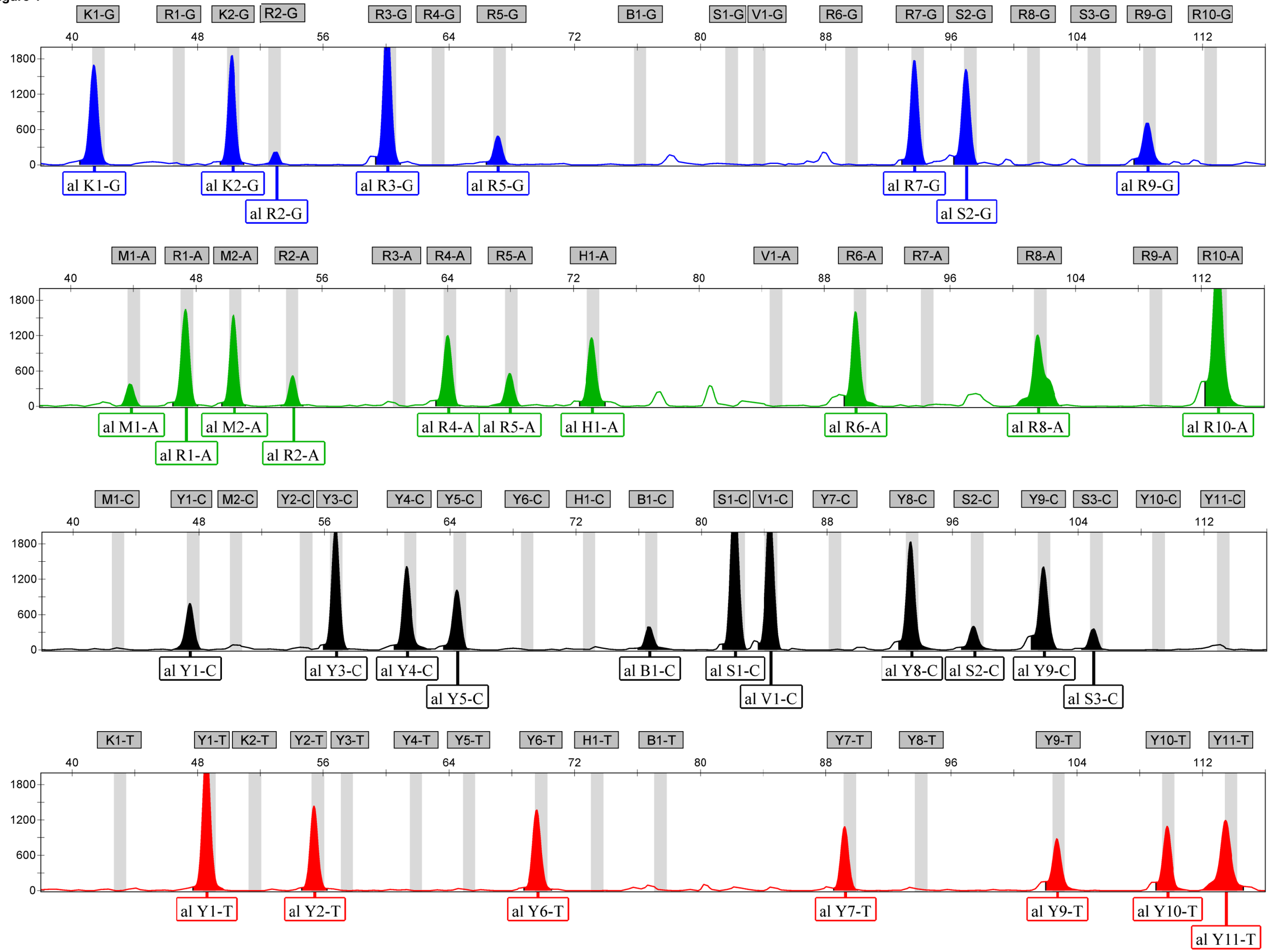

A
34-plex
AIM-indels
G-AIMs Nano
PC1 vs PC2
PC1 vs PC3
PC2 vs PC3
PC1 - 31.82% ->
PC2 - 21.02% ->
PC3 - 5.35% ->
PC1 - 23.26% ->
PC2 - 13.85% ->
PC3 - 7.35% ->
PC1 - 34.80% ->
PC2 - 21.82% ->
PC3 - 13.27% ->
B
K: 5
AFR
EUR
EAS
OCE
AMR
ESN
MSL
GWD
LWK
ACB
ASW
GBR
TSI
FIN
IBS
JPT
CHS
CDX
KHV
PEL
MXL
CLM
PUR
REFERENCE SET
C
ACB
PEL
CLM
ASW
MXL
PUR
AFR
EUR
EAS
OCE
AMR

Figure 4


K1-G
R1-G
K2-G
R2-G
R3-G
R4-G
R5-G
B1-G
S1-G
V1-G
R6-G
R7-G
S2-G
R8-G
S3-G
R9-G
R10-G
40
48
56
64
72
80
88
96
104
112
1800
1200
600
0
al K1-G
al K2-G
al R2-G
al R3-G
al R5-G
al R7-G
al S2-G
al R9-G
M1-A
R1-A
M2-A
R2-A
R3-A
R4-A
R5-A
H1-A
V1-A
R6-A
R7-A
R8-A
R9-A
R10-A
al M1-A
al R1-A
al M2-A
al R2-A
al R4-A
al R5-A
al H1-A
al R6-A
al R8-A
al R10-A
M1-C
Y1-C
M2-C
Y2-C
Y3-C
Y4-C
Y5-C
Y6-C
H1-C
B1-C
S1-C
V1-C
Y7-C
Y8-C
S2-C
Y9-C
S3-C
Y10-C
Y11-C
al Y1-C
al Y3-C
al Y4-C
al Y5-C
al B1-C
al S1-C
al V1-C
al Y8-C
al S2-C
al Y9-C
al S3-C
K1-T
Y1-T
K2-T
Y2-T
Y3-T
Y4-T
Y5-T
Y6-T
H1-T
B1-T
Y7-T
Y8-T
Y9-T
Y10-T
Y11-T
al Y1-T
al Y2-T
al Y6-T
al Y7-T
al Y9-T
al Y10-T
al Y11-T

Supplementary Table S1. PCR and SBE primer sequences and their mixture concentrations for the 31-SNP Nano multiplex reactions. Internal locus identifiers use the IUPAC code for the extended base and a number indicating the position in the multiplex. The SBE primer sequence and non-specific mobility modifying tails are distinguished by upper and lower case nucleotides respectively.

SNP ID	Internal code	PRIMER F	PRIMER R	Concentration (µM)	SBE probe	SBE length	Concentration (µM)
rs10483251	K2	GCACGTTCTTAACCTTGGCTAT	TTCTGAATATCCCACCCACAA	0.80	gtgccacgtcgtgaaagtctgacaaGGAAAAAGTTATGTGACCAGATT	48	0.45
rs12142199	R4	AGGCCTTGATGTGCTTGAAC	CGAGAAGGCCAACCACTACT	0.60	aacaactgactaaactaggtgccacgtcgtgaaagtctgacaaTCAAACATGTTCCTCTGCAC	63	0.35
rs1229984	Y4	ATTCTGTAGATGGTGGCTGTAGGA	CTGCCTCATGGCCTAAAATCA	1.50	caactgactaaactaggtgccacgtcgtgaaagtctgacaaACCACGTGGTCATCTGTG	59	1.20
rs12402499	R2	TGAAGGGTATTACTAGTGGC	TTGACAGACTTCTGCTTTTG	2.70	aggtgccacgtcgtgaaagtctgacaaCTGCTTTTGATTTCAAGTATCAGT	51	1.70
rs12498138	Y5	TCTTCTTCAGGGAATCCTGT	GAGTTACATAGGATTTGCGAG	3.50	caactgactaaactaggtgccacgtcgtgaaagtctgacaaGGGAATCCTGTTATTCACATTA	63	4.00
rs12594144	K1	CCTACAAGACCACCCACCAG	GGACCCATGGTCATTCCATA	0.80	cacgtcgtgaaagtctgacaaGCTCCCACCCTGAAAAAGA	40	1.00
rs1426654	Y6	AATTCAGGAGCTGAACTGCC	TGTTCAGCCCTTGGATTGTC	3.00	ctctctctctctctctctctctctctctctctctctctctctctctctttCGCTGCCATGAAAGTTG	67	3.00
rs1557553	Y8	TAATACAAGAGCCGCCTGGA	CTTGCAAGGAACTGCAGCTAT	1.30	taaactaggtgccacgtcgtgaaagtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaaCCCAAAGCCCCTGGAAAAA	93	0.80
rs16891982	S1	GAATAAAGTGAGGAAAACACGGAGT	GTTTCTCATCTACGAAAGAGGAGTC	1.00	cacgtcgtgaaagtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaaGGTTGGATGTTGGGGCTT	80	1.00
rs17822931	R3	CCTAGAGTCCCCCAAACCTC	CACTTCTGGGCATCTGCTTC	1.00	aactgactaaactaggtgccacgtcgtgaaagtctgacaaCTGCATTGCCAGTGTACTC	59	0.80
rs1871534	S3	ACATCCTGCAGACCTTCCTG	CAGACCTTGGGCGTCAGAT	1.30	ctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaaCCTGGCAGTGGGTGCA	105	1.80
rs2069945	B1	GCAAACCTTGGCTCTGCTAC	CCTTTCCCCAGTGGCTTAAT	1.20	aaagtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaaTATTATTAACCCCAGTCTACATG	76	1.10
rs2080161	M2	GAGTATGATATAATTTTGTTCCTGCTG	TGGACTTTATGGGTTGTTGTTTT	1.00	cacgtcgtgaaagtctgacaaTTTTTTGTTTTTTTTTTGCACTCATCA	48	0.80
rs2139931	Y10	AGTCTTGGCTAGGGCGTTAGTA	CTCCTAGTCATGGTTGATGTGG	3.00	acaacaactgactaaactaggtgccacgtcgtgaaagtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaaTTCGTGTTGATGAGAAAATTTCA	109	1.10
rs2789823	Y2	AGAGGGCTTCTGTTCACACC	ATGCACCACTACTGTCCAAG	0.50	aaactaggtgccacgtcgtgaaagtctgacaaGGAGGTGAGCTTCACGGGG	51	1.10
rs2814778	Y7	AACCTGATGGCCCTCATTAGT	ATGGCACCGTTTGGTTCAG	0.80	agtctgacaactaggtgccacgtcgtgaaagtctgacaactaggtgccacgtcgtgaaagtctgacatCTCATTAGTCCTTGGCTCTTA	89	0.80
rs3751050	R1	GAAGGCTCCCAACTCGTTAG	GTCATTAAAGTCAACCTAGGC	2.70	ctctctctctctctctctctctgtTTAGGAGAGTTGAGACATC	44	1.40
rs3827760	R6	TGCTCAGCTCCACGTACAAC	CTCTTCAGGCCGAAGCTCT	0.60	actaggtgccacgtcgtgaaagtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaaTGGCGCCACGTTTTCACA	89	1.00
rs4540055	H1	TGTGCCTCTGATCACTTTTGAATAC	CCTAGCCAACTCCAGAGTTCAT	0.80	ctctctctctctctctctctctctctctctctctctctctctctctctctctctcGAAGCAGTGATCAGCAC	72	0.50
rs4657449	R7	CCCCTCGGGAGAAAACATAG	TTCTAGAGTTGAATGAGGGTCAGA	1.70	aaactaggtgccacgtcgtgaaagtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaaGAGCTAAGGAAAGATACGTG	93	0.50
rs4749305	R5	CAGCCCAACCTACTCCTCTG	TCCCTACAAAGTGGCAAACC	2.50	caacaactgactaaactaggtgccacgtcgtgaaagtctgacaaCAGTAAATAGTAACTCCATCTTC	67	1.30
rs4792928	R8	TCTCTCAGGATATCCCTTTGG	AAAATCTTGATTCTGTATCGCAGTC	6.00	caactgactaaactaggtgccacgtcgtgaaagtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaaCGCAGTCTACTAGTTGTCC	101	1.80
rs5030240	V1	CCAAAGTGCCAGGATCACAG	TCCCTAGAAATCCTTCAGCC	1.70	ctctctctctctctctctctctctctctctctctctctctctctctctctctctctctctctctcCACAGGAGTGAGCCACTGC	84	1.30
rs6054465	Y11	TATGGCCTCAGGTTCTCCAC	CACATGATCTCACCGTTTCCT	4.00	tgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaaCACATGCAAAATCAGGATAATAATG	113	1.60
rs6437783	Y1	GCAATGAGATTAGTTGCACTGG	ATTATATGCCCACCCTGCTC	0.60	tgccacgtcgtgaaagtctgacaaCTGGTTGAGGCACACTATTA	44	0.90
rs715605	R10	CCCAGCTAGGGCTAGACACC	TCAAAGACTGAGCCATGCAC	0.80	gaaagtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaaCCACCCTAAGGGGACAGA	113	0.90
rs8072587	S2	TGGCAACCTCACATGGTAGA	CCAGGGGAGGTAGAAAGAGG	4.00	aactgactaaactaggtgccacgtcgtgaaagtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaaCAGTCTCCTGCCCGGC	97	2.00
rs8137373	Y9	CCAGAGCTTTGCAGCACTTT	CAAGGACGCAGCTCTCTCA	4.00	acaactgactaaactaggtgccacgtcgtgaaagtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaaGAGTGTTTTGTGGGCCTC	101	2.30
rs9522149	Y3	AGAAAGGAGAGGAAACACCG	TCAGCAACTTCTAGTCCTCG	0.60	ctctctctctctctctctctctctctgacaatctGAGGTCCTTGCAGCTCC	55	1.00
rs9809818	M1	TGTGTGGTTTTCTCAGCGAC	AGCATGGTATGAGCACTGAG	10.00	tctctctctctctctctctccTCCTAATAAGAGCTGGC	40	3.00
rs9908046	R9	CCTTGGCATGTTCCTCTCTC	TCAGAGGAATTAGAAAGGCCTAAA	0.60	agtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaacaactgactaaactaggtgccacgtcgtgaaagtctgacaaGGAGGTAGGAGCACCCA	109	0.60
						$(NH_4)_2SO_4$:	4.00 mM

Supplementary Table 2. Cross-validation analyses from Snipper for the 31 Nano SNPs. Grey-highlighted values correspond to the percentage of correctly classified individuals from the reference training set of each population.

	AFR	EUR	EAS	OCE	AMR
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 %	0.00 %	100.00 %	0.00 %	0.00 %
Population of OCE origin	0.00 %	0.00 %	0.00 %	100.00 %	0.00 %
Population of AMR origin	0.00 %	0.00 %	0.00 %	0.00 %	100.00 %

Supplementary File S1
Click here to download e-component: Supplementary File S1.xlsx

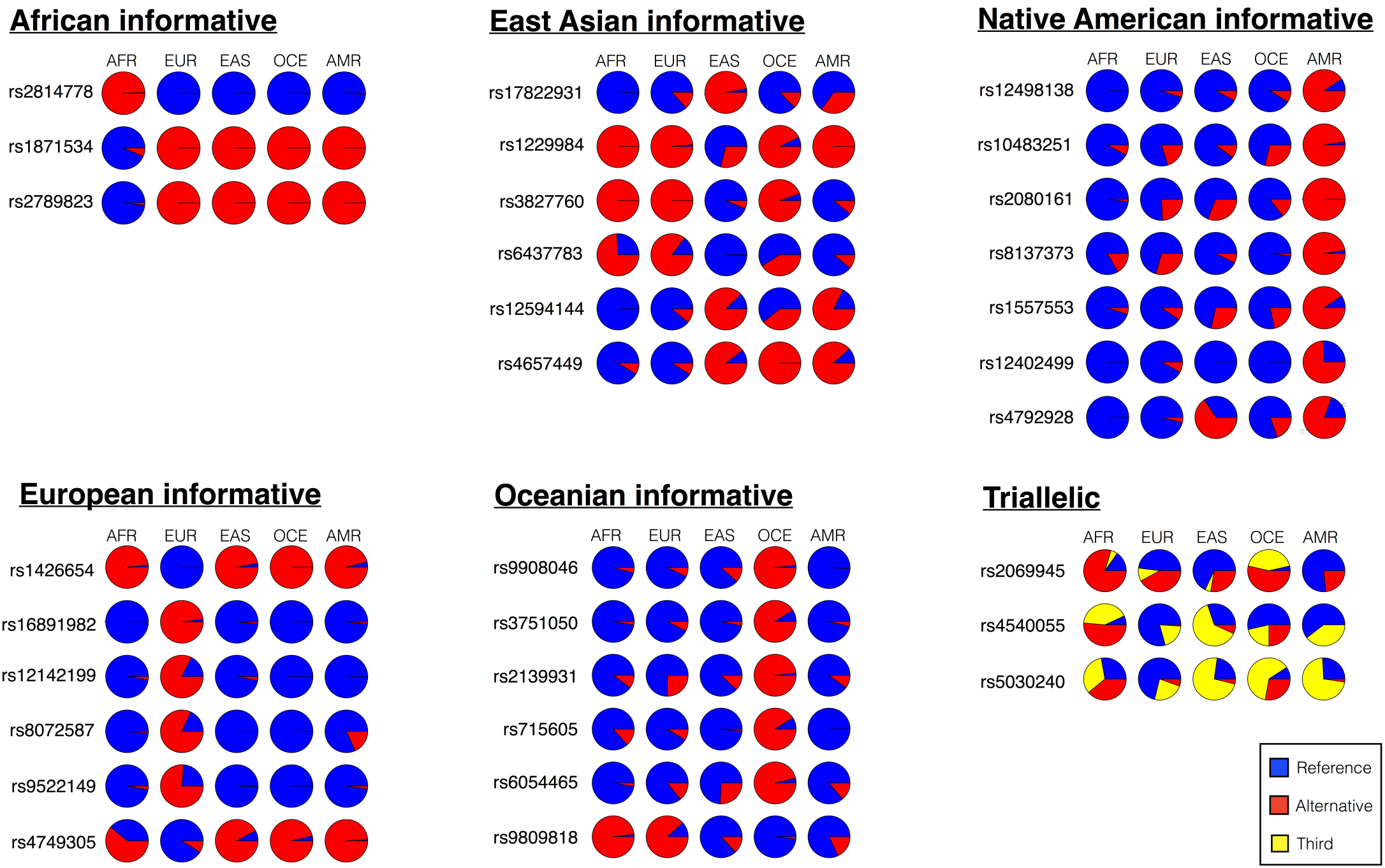
402	31	5	rs3751050_R1	rs12402499_R2
(sample no.)	(SNPs)	(populations)	1	2

Supplementary File S1. Custom training set file that can be uploaded unmodified for

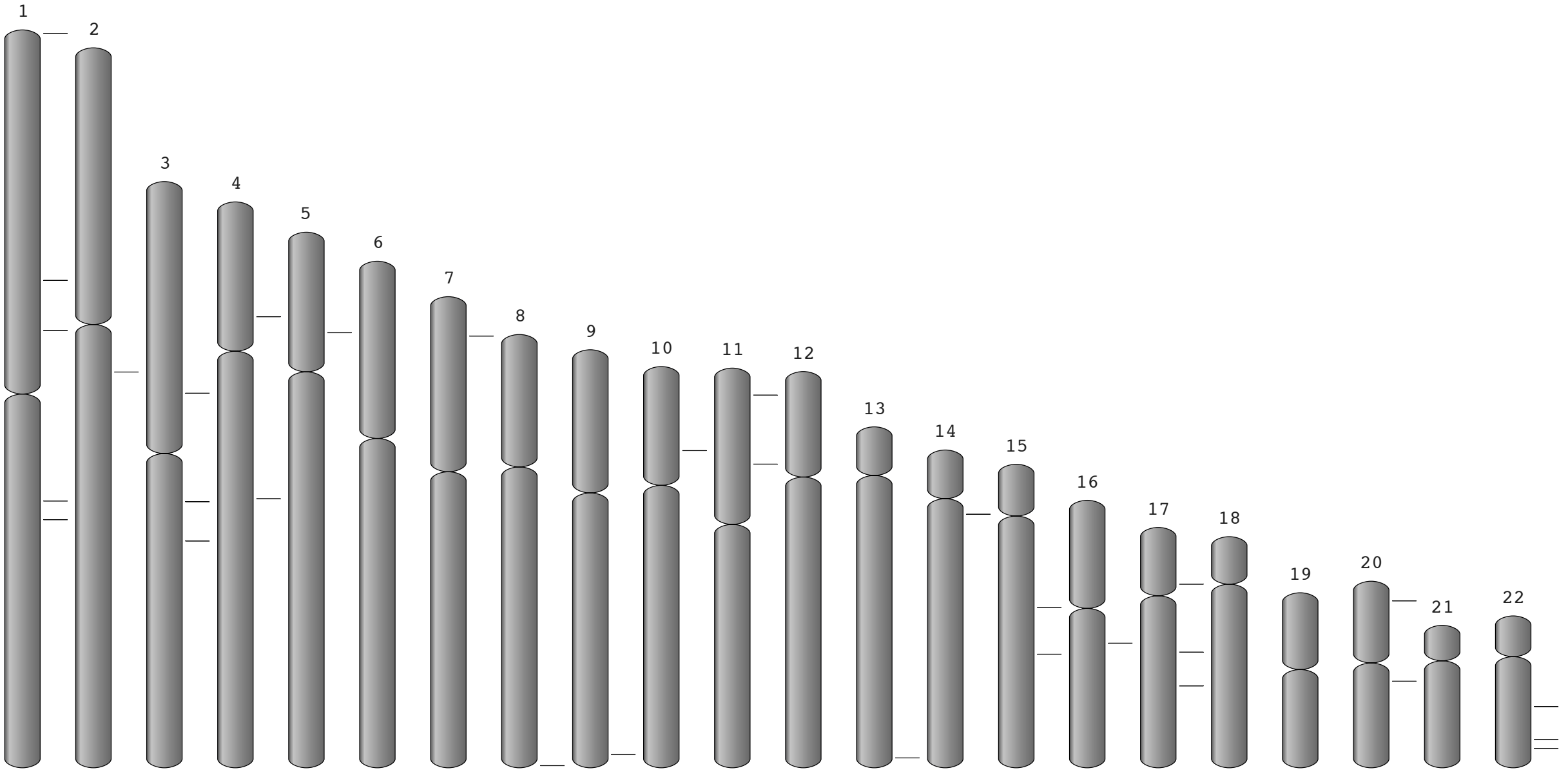
1	AFR	NA18486	AA	GG
2	AFR	NA18488	AA	GG
3	AFR	NA18489	AA	GG
4	AFR	NA18498	AA	GG
5	AFR	NA18499	AA	GG
6	AFR	NA18501	AA	GG
7	AFR	NA18502	AA	GG
8	AFR	NA18504	AA	GG
9	AFR	NA18505	AA	GG
10	AFR	NA18507	AA	GG
11	AFR	NA18508	AA	GG
12	AFR	NA18510	AG	GG
13	AFR	NA18511	AA	GG
14	AFR	NA18516	AA	GG
15	AFR	NA18517	AA	GG
16	AFR	NA18519	AA	GG
17	AFR	NA18520	AA	GG
18	AFR	NA18522	AG	GG
19	AFR	NA18523	AA	GG
20	AFR	NA18853	AA	GG
21	AFR	NA18856	AA	GG
22	AFR	NA18858	AA	GG
23	AFR	NA18861	AA	GG
24	AFR	NA18864	AA	GG
25	AFR	NA18865	AA	GG
26	AFR	NA18867	AA	GG
27	AFR	NA18868	AA	GG
28	AFR	NA18870	AA	GG
29	AFR	NA18871	AA	GG
30	AFR	NA18873	AG	GG
31	AFR	NA18874	AA	GG
32	AFR	NA18876	AA	GG
33	AFR	NA18877	AA	GG
34	AFR	NA18878	AA	GG
35	AFR	NA18879	AA	GG
36	AFR	NA18881	AA	GG

Supplementary Fig. S1
Click here to download e-component: Supplementary Fig S1.pdf

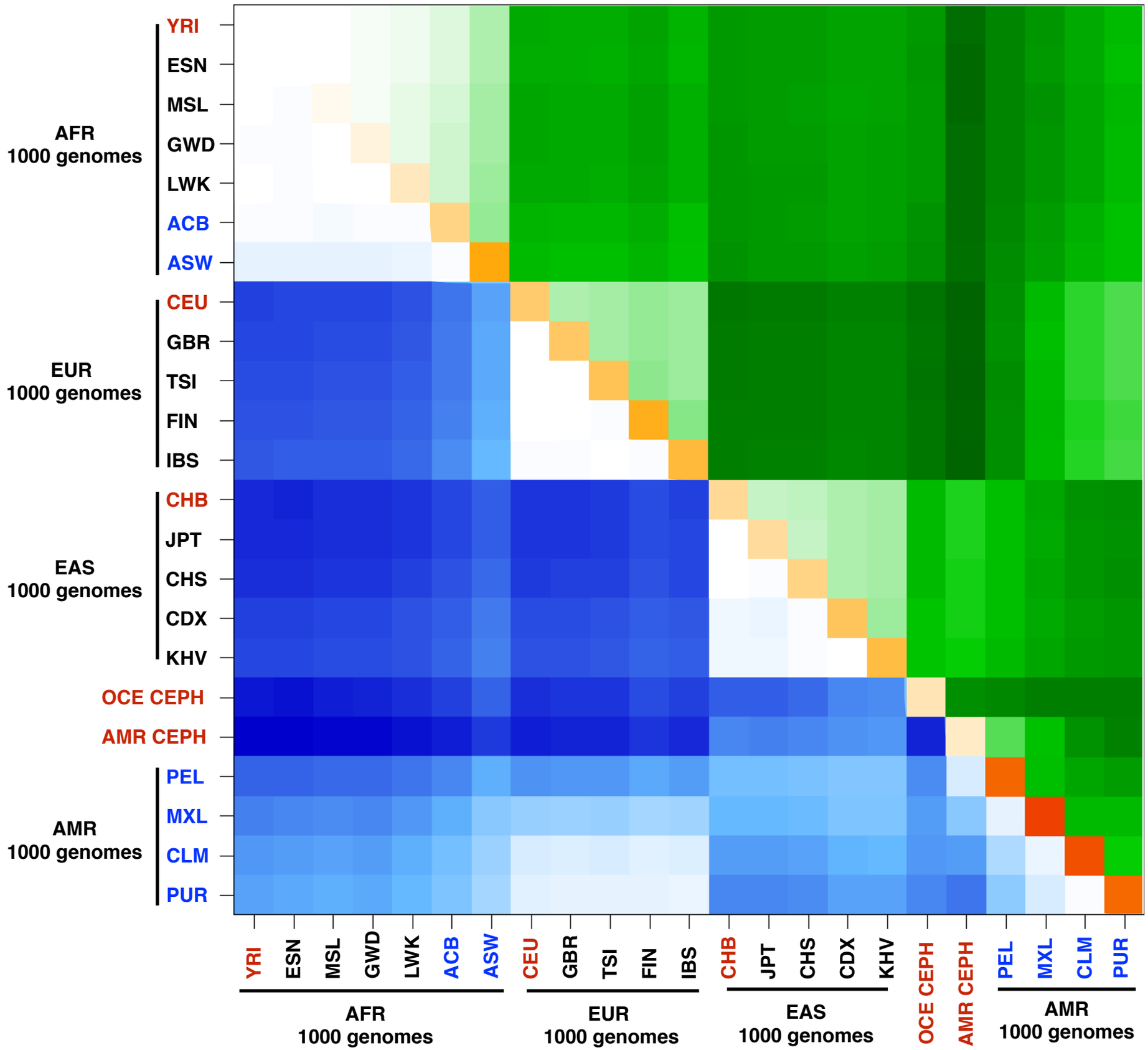
Supplementary Fig. S1. Pie charts indicating the frequency of each Nano SNP allele in five population groups (AFR: Africans; EUR: Europeans; EAS: East Asians; OCE: Oceanians; AMR: Native Americans).

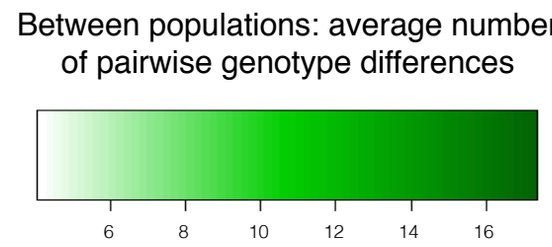


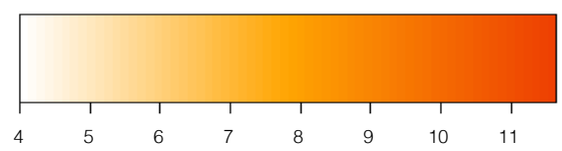
Supplementary Fig. S2. Autosomal chromosome ideograms representing the positions of the Nano SNPs.

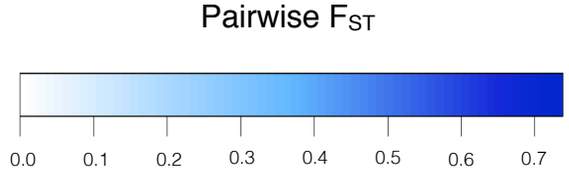


Supplementary Fig. S3. Pairwise F_{ST} and number of pairwise genotype differences between and within populations. AFR: African; EUR: European; EAS: East Asian; OCE: Oceanian; AMR: American; ESN: Esan in Nigeria; MSL: Mende in Sierra Leone; GWD: Gambian in Western Divisions in the Gambia; LWK: Luhya in Webuye, Kenya; ACB: African Caribbeans in Barbados; ASW: Americans of African Ancestry in SW USA; GBR: British in England and Scotland; TSI: Toscani in Italia; FIN: Finnish in Finland; IBS: Iberian Population in Spain; JPT: Japanese in Tokyo, Japan; CHS: Southern Han Chinese; CDX: Chinese Dai in Xishuangbanna, China; KHV: Kinh in Ho Chi Minh City, Vietnam; PEL: Peruvians from Lima, Peru); MXL: Mexican Ancestry from Los Angeles, USA; CLM: Colombians from Medellin, Colombia and PUR: Puerto Ricans from Puerto Rico. Reference populations are highlighted in red, and admixed populations in blue.



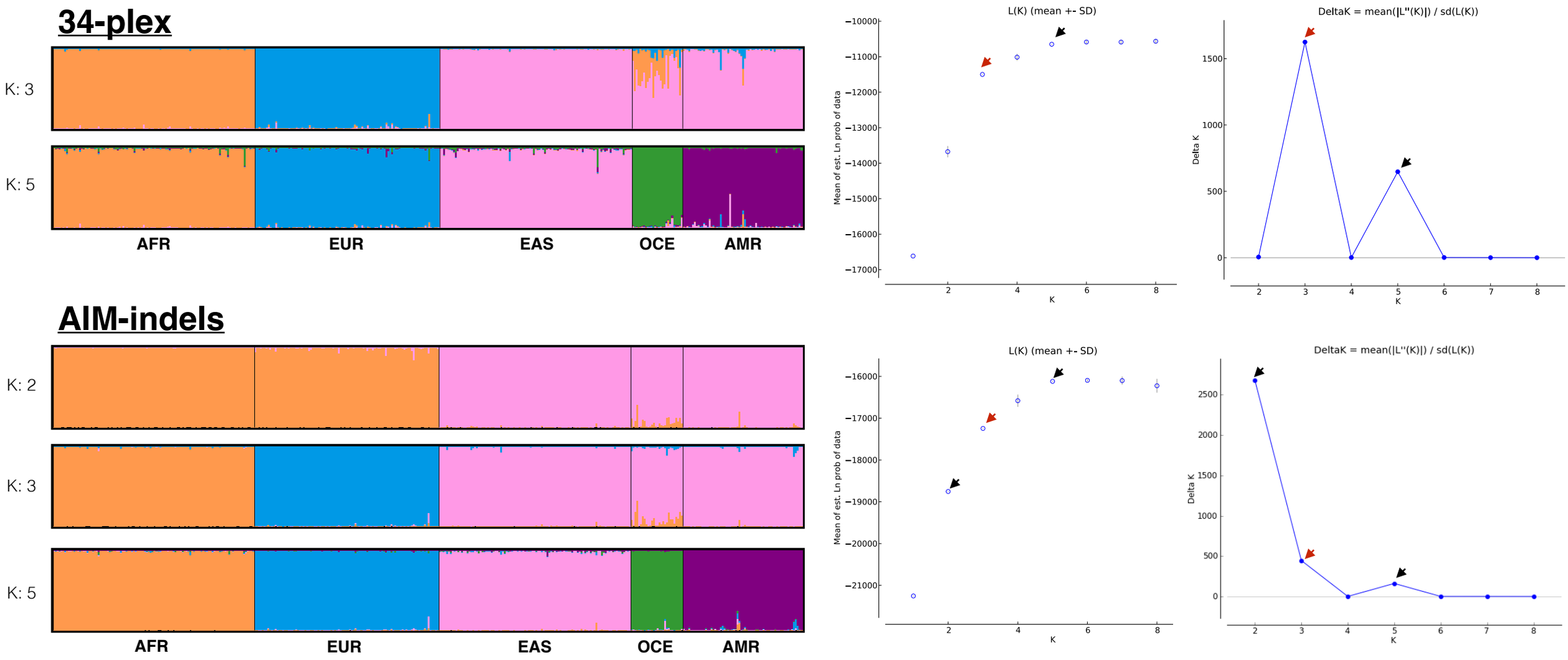






here to download e-component: Supplementary Fig S4.pdf

Supplementary Fig. S4. Three panel comparison of STRUCTURE analysis (Nano - 31 SNPs, 34-plex - 34 SNPs and AIM-indels - 44 markers) for the reference population samples. Graphics on the right indicate mean of estimated Ln probability of data and Delta K according to Evanno's method across different numbers of clusters (K). Red arrows indicate the most probable number of clusters taking into account both graphs and black arrows indicate other probable numbers of clusters represented on the left. K: number of clusters; AFR: African; EUR: European; EAS: East Asian; OCE: Oceanian; AMR: Native American.



Supplementary Fig. S5. Likelihoods ratios of ancestry estimation for control samples of known ancestry (A-E) and 9947A, removing sequentially the most informative markers. Grey box represents a threshold likelihood value of 1,000 and the red line indicates the most informative fourteen SNPs.