Inter-laboratory evaluation of SNP-based forensic identification by massively parallel sequencing using the Ion PGM™

M. Eduardoff[1], C. Santos[2], M. de la Puente[2], T.E. Gross[3], M. Fondevila[2], C. Strobl[1], B. Sobrino[4], P.M. Schneider[3], Á. Carracedo[2,4,5], M.V. Lareu[2], W. Parson[1,6], C. Phillips[2]*

[1] Institute of Legal Medicine, Innsbruck Medical University, Müllerstrasse 44, A-6020 Innsbruck, Austria

[2] Forensic Genetics Unit, Institute of Legal Medicine, University of Santiago de Compostela, ES-15705 Santiago de Compostela, Galicia, Spain

[3] Institute of Legal Medicine, Faculty of Medicine, University of Cologne, Cologne, Germany

[4] Grupo de Medicina Xenómica (GMX), Faculty of Medicine, University of Santiago de Compostela

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

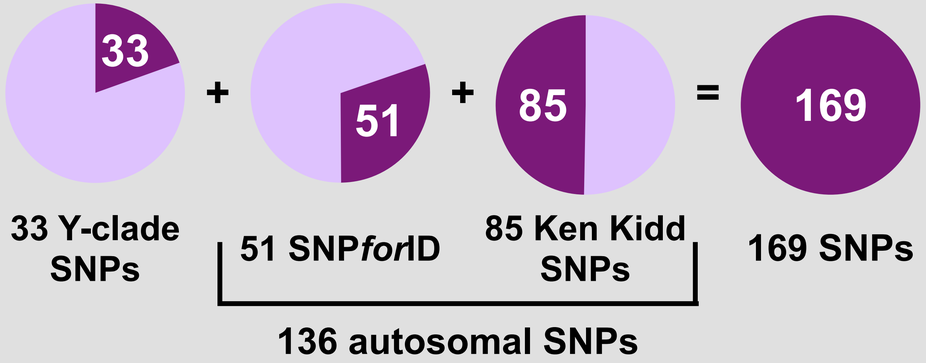
[6] Penn State Eberly College of Science, University Park, PA, USA

* Corresponding author.

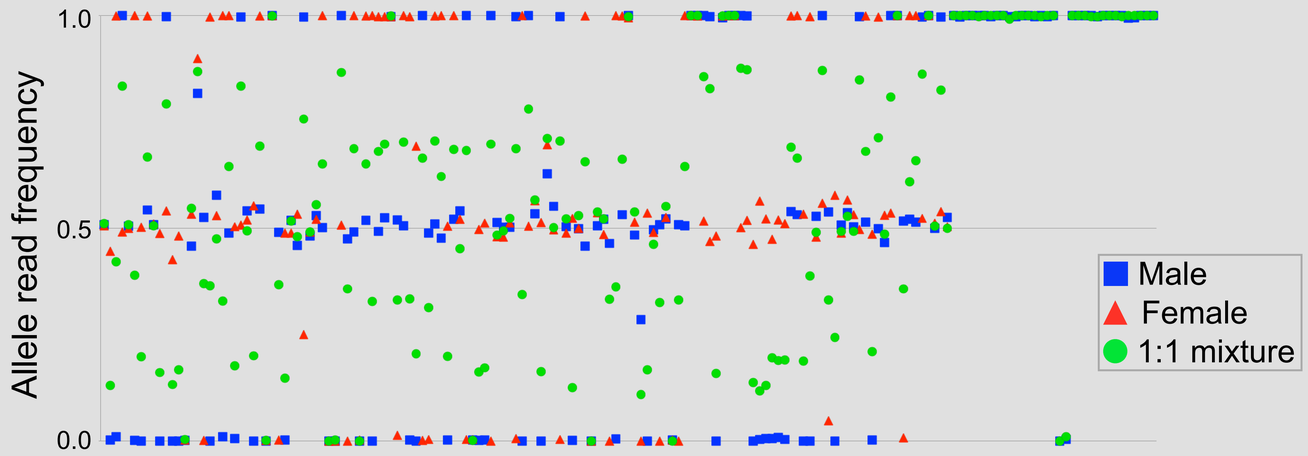
E-mail address: c.phillips@mac.com (C. Phillips).

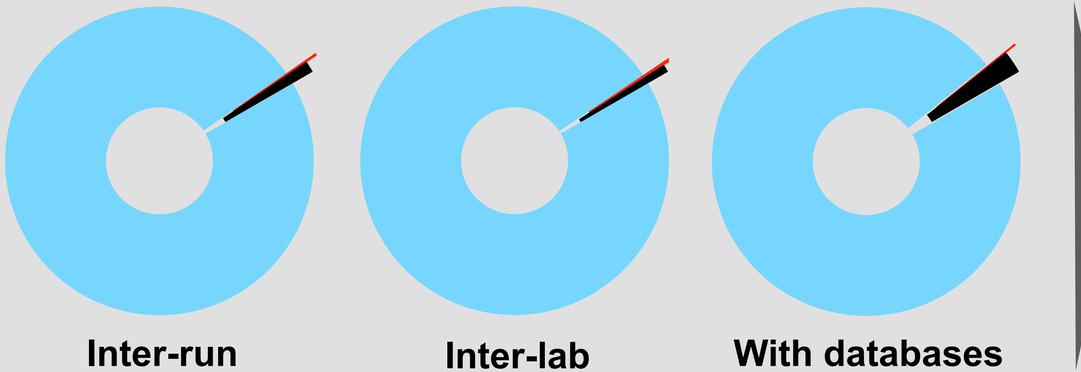
Graphical Abstract (for review)

HID-Ion AmpliSeq™ Identity Panel v2.2



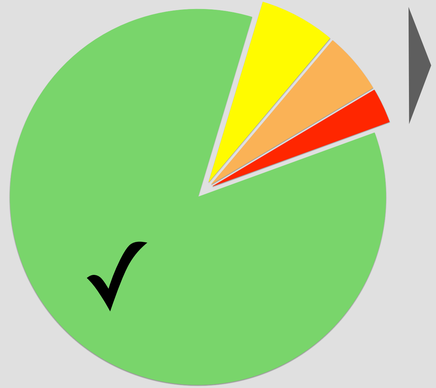
Mixture detection





Overall concordance
99.8%

- Concordant genotypes
- No-call
- Discordant genotypes



Outlier SNPs

Atypical
No-calls
Discordant
rs1979255
rs1004357
rs938283
rs2032597
rs2399332

Highlights:

- Evaluation of the HID-Ion AmpliSeq™ Identity Panel v 2.2 was performed between three laboratories.
- Levels of sequence coverage, sensitivity, ability to detect mixed DNA and genotyping precision were assessed.
- High coverage levels were obtained for the majority of the 169 SNPs studied for input DNA levels as low as 25-100 pg and the overall genotyping concordance rate was 99.8%.
- Mixed source DNAs can be detected but further optimisation of the analysis parameter settings is needed.
- Certain component SNPs underperform so they should be excluded from the panel or their data discounted during the analysis.
- The HID-Ion AmpliSeq™ Identity Panel and Ion PGM™ system provide a sensitive and accurate genotyping assay highly applicable to forensic analysis.

Inter-laboratory evaluation of SNP-based forensic identification by massively parallel sequencing using the Ion PGM™

Abstract

Next generation sequencing (NGS) offers the opportunity to analyse forensic DNA samples and obtain massively parallel coverage of targeted short sequences with the variants they carry. We evaluated the levels of sequence coverage, genotyping precision, sensitivity and mixed DNA patterns of a prototype version of the first commercial forensic NGS kit: the HID-Ion AmpliSeq™ Identity Panel with 169-markers designed for the Ion PGM™ system. Evaluations were made between three laboratories following closely matched Ion PGM™ protocols and a simple validation framework of shared DNA controls. The sequence coverage obtained was extensive for the bulk of SNPs targeted by the HID-Ion AmpliSeq™ Identity Panel. Sensitivity studies showed 90-95% of SNP genotypes could be obtained from 25-100 picograms of input DNA. Genotyping concordance tests included Coriell cell-line control DNA analyses checked against whole-genome sequencing data from 1000 Genomes and Complete Genomics, indicating a very high concordance rate of 99.8%. Discordant genotypes detected in rs1979255, rs1004357, rs938283, rs2032597 and rs2399332 indicate these loci should be excluded from the panel. Therefore, the HID-Ion AmpliSeq™ Identity Panel and Ion PGM™ system provide a sensitive and accurate forensic SNP genotyping assay. However, low-level DNA produced much more varied sequence coverage and in forensic use the Ion PGM™ system will require careful calibration of the total samples loaded per chip to preserve the genotyping reliability seen in routine forensic DNA. Furthermore, assessments of mixed DNA indicate the user's control of sequence analysis parameter settings is necessary to ensure mixtures are detected robustly. Given the sensitivity of Ion PGM™, this aspect of forensic genotyping requires further optimisation before massively parallel sequencing is applied to routine casework.

Keywords: Next generation sequencing; Massively parallel sequencing; Ion PGM™; Ion Torrent; Identification SNPs;

1. Introduction

Next generation sequencing (NGS) systems are becoming available to genotype established forensic markers for identification, inference of genetic ancestry and prediction of externally visible characteristics (EVCs). The two current NGS systems most applicable to forensic analysis are Life Technologies' (LT) Ion Personal Genome Machine® (PGM™) system [1] and Illumina's MiSeq [2]. Both offer compact detectors and massively parallel sequencing chemistries, with comparable accuracy and ease-of-use [3]. As well as expanding the scope of forensic mitochondrial sequencing [4], NGS offers the ability to genotype both STRs and single nucleotide polymorphisms (SNPs) by sequencing hundreds to several thousand copies of short DNA fragments carrying the variation [5]. Initial target amplification of DNA can potentially multiplex several hundred to thousand markers per PCR, so all loci required for a particular forensic purpose: identification or ancestry/EVC inference, are amplifiable in one tube. This large-scale approach extends further since LT and Illumina can use sample-tagging DNA barcodes, allowing multiple samples to be individualised with specific sequence tags then combined in a joint sequencing run.

This report describes inter-laboratory evaluations of the LT Ion PGM™ system (herein Ion PGM™) and the forensic SNP set named HID-Ion AmpliSeq™ Identity Panel (herein HID SNP). Ion PGM™ exploits a sensitive semiconductor-based detection of H^+ ion release during base incorporation onto short template sequences bound to micro-spheres. The HID SNP set version 2.2 evaluated here, comprises 51 SNP*for*ID [6] and 85 Kiddlab autosomal SNPs [7] plus 33 Y-markers [8]. Three aspects of Ion PGM™ and the HID SNP set are important when assessing this system's applicability to forensic analysis: i. performance of the Ion PGM™ sequencing chemistry as a whole, including base misincorporation, sensitivity gauged by capacity to reliably sequence low-level DNA and genotyping accuracy; ii. characteristics of HID SNP markers, including sequence coverage per locus, Y-SNP male specificity and heterozygote balance; iii. characteristics of Ion PGM™ relating to its ability to detect mixtures from the reduced variation of bi-allelic SNPs. Our experiments followed the simple scheme for evaluating any new forensic technique that uses qualified runs. The validation framework genotyped shared staff donor and Coriell cell-line control DNAs amongst three laboratories running closely matched Ion PGM™ protocols. Sensitivity was assessed using simple dilution series and one highly degraded 800 year old DNA from archaeological remains. Mixtures were made to gauge how well Ion PGM™ detected multiple components in male-female mixed DNA.

An important preamble to evaluating heterozygote balance was the measurement of genotype concordance – comparing genotypes assigned by Ion PGM™ to those from alternative SNP typing techniques. While sequencing ambiguities can be accurately detected in mitochondrial sequences by reference to a well-established phylogeny, SNP genotype error is less straightforward to measure. Although the massively parallel coverage of NGS should reduce the probability of error substantially, it is still necessary to confirm the level of

genotyping concordance with this new type of sequencing technology. Concordance studies used Coriell cell-line control DNAs already characterised by 1000 Genomes [9] and Complete Genomics [10] large-scale genome sequencing projects. As well as allele balance, the context sequence around each SNP was checked for closely sited features (e.g. polymeric tracts or Indels): having the potential to interfere with reliable alignment of detected sequences. Although care was taken to avoid such features in the original SNP*for*ID marker choice and primer positioning [6], Indels or low complexity sequence can still occur in amplified fragments and influence their alignment.

2. Materials and methods

All Ion PGM™ protocols followed published laboratory guidelines [11-15]. 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).

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 [9] and CG a proprietary DNA nanoarray method [10]).

Dilutions of 9947A and 007 DNAs assessed the forensic sensitivity of Ion PGM™, using 10 ng, 1 ng, 100 pg, 50 pg and 25 pg of DNA amplified with varying PCR cycle numbers, as outlined in Table 1. 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-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 [16]. 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.

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 [11-13]. Prior quantification of DNAs used Qubit® ds DNA 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® ds DNA HS Assay Kit, Agilent® High Sensitivity DNA Kit or Ion Library Quantitation Kit to equalise a final library of 100 pM in ≥20 µL [12].

Template preparation used Ion OneTouch™ 200 Template Kit v2, following manufacturer's protocols [14]. 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 [14].

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 [17]. 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 [18]. 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) [19,20].

3. Results and discussion

3.1. 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 (Supplementary Fig. S1). It is noticeable that 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 Supplementary File S1 (Fig. S3). 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 ($p=0.029$, $alpha=0.05$) 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 Fig. 1A. The distribution of quartiles reveals variation both within and between runs, but Fig. 1C indicates that runs combining low-level DNA alongside optimum input DNA samples has higher variation between samples. Fig. 1B 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 ($p=0.00045$, $alpha=0.05$) in low-level DNA samples.

We detected an increased number of sequenced multiplex primers from target amplification. 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 (Supplementary File S1). 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 30x and 500x, respectively. The threshold for somatic SNP calling is close to the values cited in whole genome and enrichment variant detection studies [2, 21-25]. 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 ~20x 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 13x and 41x respectively, discounting outlier SNPs. This largely matches results of a recent study by Daniel et al. finding a similar minimum coverage estimate of 20x for reliable SNP genotyping [26]. 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 269x coverage or higher. Y-SNPs gave concordant genotypes with an average of 63x coverage in the 1:9 male-female mixture whereas this value increased to 274x 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 42x to 286x for 95% of bases (Supplementary Table S1). 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 62x for 95% of bases is necessary to accomplish a minimum coverage of 13x 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 Fig. 2 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 (Supplementary File S1, Fig. S5). 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.

3.2. 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 (ARF) 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. The following results are outlined in detail in Supplementary Table S2.

3.2.1. 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 $3^{rd}/4^{th}$ 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%. Supplementary Fig. S6 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 (column N, Supplementary Table S2). These SNP's data should be discounted from assessments of imbalanced homozygote patterns in mixtures, particularly rs2399332 and rs1523537 with disproportionately high allelic misincorporation.

Although allelic and non-specific misincorporation are similar enough to largely discount drop-in, Y-chromosome sequences were observed in female DNA. Supplementary Fig. S7 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 >2 million female-specific sequences indicates extremely low levels of drop-in genotypes from extraneous DNA for the Ion PGM™ system.

3.2.2. 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. Fig. 3A 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 reliable detected (see sections 3.3 and 3.4). A 40% threshold (60:40 heterozygote ratio), indicated by the middle grey box over A-SNPs in Fig. 3A, 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 Fig. 3A and were identified from divergent average heterozygote ARF values (column P, Supplementary Table S2, but rs1029047: cell P19, identified from out-of-range values both sides of midline). 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. [27]. Additionally, rs2107612 was poorly balanced in our study, but not singled out by Børsting. Interestingly, SNPs rs10776839, rs4530059 and rs1031825 found to be problematic by Børsting et al., gave reasonably balanced ARFs here, although Fig. 3A indicates rs4530059 and rs1031825 have small proportions of genotypes lying outside the threshold range.

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 (column P, Supplementary Table S2).

3.2.3. 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: columns Q-S, Supplementary Table **S**2). 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 Supplementary Fig. S8. 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 3.4.2.

3.3. 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 CG 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, whereas the individual rates for each sample used for concordance comparisons are detailed in Supplementary Tables S2 and S3.

3.3.1. 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 3.4.1 (column T, Supplementary Table S2). 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.

3.3.2. 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 2. 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 vs. 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 version 2.2.0.26, except for the genotypes of sample NA06994 where version 2.2.0.19 was used). In addition to 30 no-call genotypes from Ion PGM™ results, 8 no-call genotypes resulted from ambiguous CG genotype calls (Table 2 and row 40, Supplementary Table S2); 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 2. 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, our comparisons of Coriell cell-line control DNA genotype data generated from different SNP genotyping systems indicate a very high concordance rate of 99.8%.

3.4. 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. Fig. 4 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 (Supplementary Table S2). SNPs with atypical

sequencing characteristics were then analysed in detail by examining their VCF files and using IGV sequence visualisation software [28, 29].

3.4.1. Discordant SNPs

Five SNPs were identified showing consistent patterns of discordant genotyping. Section 3.3 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 (Supplementary File S2-SNP 1), 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 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 (Supplementary Fig. S6) as well as the single discordant genotype observed (Table 2). Furthermore, rs2399332 shows a clear deviation from expected ARF ratios in heterozygotes (Fig. 3A and cell P12, Supplementary Table S2). 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 (row 16, Supplementary Table S2).

3.4.2. 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). Fig. 5 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 (Supplementary Fig. S8). In these SNPs sequencing is initiated on both strands but one fails to reach the SNP position, illustrated by the IGV overview of rs13182883 (Supplementary File S2-SNP 2) with 0.994 strand bias. This phenomenon produces the very strong strand bias deviations shown at the ends of the distribution plot of Supplementary Fig. S8 and remains unexplained from all analyses made in IGV.

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 3.1 and 3.2, 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 (Supplementary File S2-SNP 3). 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. [27], while Budowle et al. also reported discordant genotypes [30, 31].

3.4.3. SNPs with mean analysis parameter values deviating from defined thresholds

Despite an absence of genotyping problems affecting the eleven SNPs of this third category (Fig. 4), examination of their mean values showed consistent atypical behaviour with respect to the analysis parameter thresholds we defined, particularly sequence coverage and strand bias (columns O, Q, R, S in Supplementary Table S2). 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 Supplementary File S2-

SNP 4). 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.

3.5. 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 Fig. 5, 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. Supplementary Fig. S11 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 3.4) had missing genotypes in aDNA and lab1-A runs. Five extra library amplification cycles did not increase sensitivity.

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 PGM™ when target sequences are highly degraded or inhibited. Therefore, although good sensitivity to low-level DNA has been recognised in this and other studies [27,30], 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.

Supplementary Table S4 details sequence data from two analyses of aDNA sample S7. Although these 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.

3.6. 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 Supplementary File S3.

Scrutiny of the ARF plots in Supplementary File S3, 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 (Supplementary File S3, Table S5). 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 (here, 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.

3.7. Context sequence examinations with IGV

To further assess HID SNPs for forensic analysis, the context sequence of each marker was scrutinised using IGV [28,29]. 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 3 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 Supplementary File S2-SNP 4). Variants at sequence extremes and next to polymeric tracts tended to produce unreliable reads (see rs1109037 in Supplementary File S2-SNP 5).

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 Supplementary File S2-SNP 4, 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.

4. 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's study as a problematic SNP [27]. 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. Lastly, mention should be made of rs1029047, which gives genotyping inconsistencies in all NGS studies of this SNP made so far [27,30,31]. There are clearly characterised context sequence factors affecting the alignment and therefore the reliability of allele calls for rs1029047 (Supplementary File S2-SNP 3), which have not affected SNaPshot genotyping of this SNP [6]. 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 ~15-20x minimum coverage thresholds can safeguard the reliability of allele calls made with NGS [2,21,26].

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 under-performing 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. [27], 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.

Acknowledgements This work was funded by the European Union Seventh Framework Program (FP7/2007–2013) under grant agreement no. 285487 (EUROFORGEN-NoE) and the Austrian Science Fund (FWF) [P22880-B12]. CS is supported by funding awarded by the Portuguese Foundation for Science and Technology (FCT) and co-financed by the European Social Fund (Human Potential Thematic Operational Program SFRH/BD/75627/2010). 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). The authors wish to thank Jorge Amigo, Grupo de Medicina Xenómica (GMX), University of Santiago de Compostela; David Ballard, Department of Forensic and Analytical Science, Kings College, London; and Matt Phipps of Life Technologies, for their helpful guidance with NGS data analysis.

References

[1] B. Merriman, Ion Torrent R&D Team, J.M. Rothberg, Progress in Ion Torrent semiconductor chip based sequencing, Electrophoresis 33 (2012) 3397–3417.

[2] M.A. Quail, M. Smith, P. Coupland, T.D. Otto, S.R. Harris, T.R. Connor, A. Bertoni, P. Harold, H.P. Swerdlow, Y. Gu, A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers, BMC Genomics 24 (2012) 341-353.

[3] N.J. Loman, R.V. Misra, T.J. Dallman, C. Constantinidou, S.E. Gharbia, J. Wain, M.J. Pallen, Performance comparison of benchtop high-throughput sequencing platforms, Nat. Biotechnol. 30 (2012) 434-439.

[4] W. Parson, C. Strobl, G. Huber, B. Zimmermann, S.M. Gomes, L. Souto, L. Fendt, R. Delport, R. Langit, S. Wootton, et al., Evaluation of next generation mtGenome sequencing using the ion Torrent Personal Genome Machine (PGM™), Forensic Sci. Int. Genet. 7 (2013) 543–549.

[5] B. Budowle, D.H. Warshauer, S.B. Seo, J.L. King, C. Davis, B. LaRue, Deep Sequencing Provides Comprehensive Multiplex Capabilities, Forensic Sci. Int. Genet. Supp. Series 4 (2013) e334-e335.

[6] J.J. Sanchez, C. Phillips, C. Børsting, K. Balogh, M. Bogus, M. Fondevila, C.D. Harrison, E. Musgrave-Brown, A. Salas, D. Syndercombe-Court, et al., A multiplex assay with 52 single nucleotide polymorphisms for human identification, Electrophoresis 27 (2006) 1713–1724.

[7] A.J. Pakstis, W.C. Speed, R. Fang, F.C. Hyland, M.R. Furtado, J.R. Kidd, K.K. Kidd, SNPs for a universal individual identification panel, Hum. Genet. 127 (2010) 315– 324.

[8] T.M. Karafet, F.L. Mendez, M.B. Meilerman, P.A. Underhill, S.L. Zegura, M.F. Hammer, New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree, Genome Res. 18 (2008) 830–838.

[9] The 1000 Genomes Project Consortium, An integrated map of genetic variation from 1,092 human genomes, Nature 491 (2012) 56-65.

[10] R. Drmanac, 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., Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays, Science 327 (2009) 78-81.

[11] Thermo Fisher Scientific, Life Technologies: Sequencing technology solutions. (2014) Accessed June 2014. Available from: https://www.lifetechnologies.com/au/en/home/life-science/sequencing/sequencing-technology-solutions.html

[12] Thermo Fisher Scientific, Life Technologies: Ion PGM™ system for next-generation sequencing. (2014) Accessed June 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

[13] Thermo Fisher Scientific, Life Technologies: Ion AmpliSeq™ library preparation user guide. July 2013.

[14] Thermo Fisher Scientific, Life Technologies: Ion OneTouch™ 200 Template Kit v2 user guide. 2012.

690 [15] Thermo Fisher Scientific, Life Technologies: Ion PGM™ 200 Sequencing Kit user guide.
691 2012.
692 [16] C.M. Bauer, H. Niederstätter, G. McGlynn, H. Stadler, W. Parson, Comparison of
693 morphological and molecular genetic sex-typing on mediaeval human skeletal remains,
694 Forensic Sci. Int. Genet. 7 (2013) 581–586.
695 [17] Thermo Fisher Scientific, Life Technologies: Torrent Suite™ software 4.0.2. user guide.
696 November 2013.
697 [18] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella,
698 D. Altshuler, S. Gabriel, M. Daly, M.A. DePristo, The Genome Analysis Toolkit: a MapReduce
699 framework for analyzing next-generation DNA sequencing data, Genome Res. 20 (2010)
700 1297-1303.
701 [19] R: A language and environment for statistical computing. R Foundation for Statistical
702 Computing, Vienna, Austria. ISBN 3-900051-07-0, Available from: http://www.R-project.org.
703 [20] R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L.
704 Gautier, Y. Ge, J. Gentry, et al., Bioconductor: Open software development for
705 computational biology and bioinformatics, Genome Biol. 5 (2004) R80.
706 [21] D.R. Bentley, S. Balasubramanian, H.P. Swerdlow, G.P. Smith, J. Milton, C.G. Brown, K.P.
707 Hall, D.J. Evers, C.L. Barnes, H.R. Bignell, et al., Accurate whole human genome sequencing
708 using reversible terminator chemistry, Nature 456 (2008) 53–59.
709 [22] D.C. Kobold, L. Ding, E.R. Mardis, R.K. Wilson, Challenges of sequencing human
710 genomes, Brief Bioinform. 11 (2010) 484–498.
711 [23] R. Nielsen, J.S. Paul, A. Albrechtsen, Y.S. Song, Genotype and SNP calling from next-
712 generation sequencing data, Nat. Rev. Genet. 12 (2011) 443-451.
713 [24] A. Elsharawy, M. Forster, N. Schracke, A. Keller, I. Thomsen, B.S. Petersen, B. Stade, P.
714 Stähler, S. Schreiber, P. Rosenstiel, A. Franke, Improving mapping and SNP-calling
715 performance in multiplexed targeted next-generation sequencing, BMC Genomics 13 (2012)
716 417.
717 [25] D. Sims, I. Sudbery, N.E. Ilott, A. Heger, C.P. Ponting, Sequencing depth and coverage:
718 key considerations in genomic analyses, Nat. Rev. Genet. 15 (2014) 121-132.
719 [26] R. Daniel, C. Santos, C. Phillips, M. Fondevila, R.A.H. van Oorschot, Á. Carracedo, M.V.
720 Lareu, D. McNevin, A SNaPshot of next generation sequencing for forensic SNP analysis,
721 Forensic Sci. Int. Genet. 14 (2014) 50-60.
722 [27] C. Børsting, S.L. Fordyce, J. Olofsson, H. Smidt Mogensen, N. Morling, Evaluation of the
723 Ion Torrent™ HID SNP 169-plex: A SNP typing assay developed for human identification by
724 second generation sequencing, Forensic Sci. Int. Genet. 12 (2014) 144–154.
725 [28] J.T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E.S. Lander, G. Getz, J.P.
726 Mesirov, Integrative Genomics Viewer, Nat. Biotechnol. 29 (2011) 24–26.
727 [29] H. Thorvaldsdóttir, J.T. Robinson, J.P. Mesirov, Integrative Genomics Viewer (IGV): high-
728 performance genomics data visualization and exploration, Brief. Bioinform. 14 (2012) 178-
729 192.
730 [30] S.B. Seo, J.L. King, D.H. Warshauer, C.P. Davis, J. Ge, B. Budowle, Single base
731 polymorphism typing with massively parallel sequencing for human identification, Int. J.

732 Legal Med. 127 (2013) 1079–1086.
733 [31] B. Budowle, 'Next Generation Sequencing Provides Comprehensive Multiplex
734 Capabilities', presentation at 25th ISFG Congress, Melbourne 2013, available at:
735 http://isfg2013.org/wp-content/uploads/2013/09/wed-p3-1100-Bruce-Budowle-Y.pdf.
736 Accessed October 2014.

Figure legends

Fig. 1.

(A) Box plots of recorded SNP Target Reads from 101 samples in 12 runs (quartile range boxes, 95% whiskers, means shown as mid-plot bars). Blue numbers are used to identify samples listed in Fig. 2.

(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.

Fig. 2.

Analysis vs. SNP heatmap arranged as: increasing mean sample 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 Fig. 1A.

Fig. 3.

(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. [27] are marked for comparison.

(B) ARF balance observed in the 1:1 mixture (S5-S6 male-female donors), SNPs listed in the same order as (A). Circle and triangle points show replicate values from two independent library runs.

Fig. 4.

Schematic representation of the proportion of HID SNPs with good performance, poor 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's study of the same SNP panel [27].

Another three SNPs identified by Børsting: rs10776839; rs4530059 and rs1031825 did not show problematic characteristics in our study.

Fig. 5.

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).

Supplementary Files

Supplementary File S1

Assessments of sequence coverage obtained with HID SNP markers and the Ion PGM™

Supplementary File S2

IGV overviews of five SNPs (A-E) showing context sequence features

Supplementary File S3

Mixture analysis with the Ion PGM™

Supplementary Figures

Supplementary Fig. S1.

Proportions of four types of sequence reads from 12 Ion PGM™ runs using the full range of available sequencing chips. (Total; Filtered with barcode; Mapped with barcode; SNP Target Reads), indicating that Total Reads and, more importantly, SNP Target Reads varied considerably between runs.

Supplementary Fig. S2.

Concentration of DNA libraries obtained from seven initial input DNA quantities (or UK: unknown) in 101 analyses. We followed the Ion PGM™ guidelines of 10 ng DNA input for most runs, but the more varied input amounts of lab1 shows no relationship to library concentrations.

Supplementary Fig. S3.

Read length histograms of an optimal input DNA sample, low-level DNA sample and a negative control before and after read filtering (right, left). Pronounced peaks at ~ 50bp in low level DNA and negative control samples correspond to adapter dimers.

Supplementary Fig. S4.

Comparison of primer regions reads for rs1005533 in a negative control, low-level and optimum input DNA sample, from IGV graphical summaries.
(**A**) Negative control shows short reads in the primer region of targeted rs1005533.
(**B**) Similar reads can be seen in a low-level DNA sample.
(**C**) The optimum input DNA sample does not show any short reads in the target neighbouring region. For better visualization reads are down-sampled to 100. Pink sequences are forward direction, violet reverse.

Supplementary Fig. S5. HID SNP panel coverage distribution parameters.
(**A**) Ranked mean/median coverage ratios showing discernible skew in rightmost 13 analyses where lower SNP Target Reads were obtained than mean values would predict.
(**B**) Unity-based normalization of mean SNP coverage vs. median SNP coverage per analysis.
(**C**) Normalization of absolute mean SNP coverage vs. median SNP coverage. Both plots show that not all data points lie on the diagonal line, implying a non-normal distribution of mean values. This suggests amplification bias amongst HID SNP components with increasing total coverage (accentuated by raised 169-SNP competition in male PCR).
(**D-E**) Interquartile range of SNP coverage per sample and maximum coverage rise with total coverage.
(**F-G**) Minimum coverage per sample vs total coverage sample. Minimum coverage is not linearly influenced by total coverage levels - when removing outlier SNPs there is a slight improvement in relatedness.

Supplementary Fig. S6.
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.

Supplementary Fig. S7.
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).

Supplementary Fig. S8.
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 3.4.2 and the IGV overview of rs13182883 is given in Supplementary File S2-SNP 2.

Supplementary Fig. S9.

Allele read frequency distributions observed in mixed DNA analyses (red lines: heterozygote balance thresholds).

Supplementary Fig. S10.
Observed and expected ratios of average Y-SNP coverage vs. average A-SNP coverage for the male component S5 and mixtures.

Supplementary Fig. S11.
Reduction in cumulative RMP with increasing no-call rate.

Supplementary Tables

Supplementary Table S1.
Expected sequence throughput of Ion PGM™ based on chosen sample numbers and 3-series chip type used.

Supplementary Table S2.
Details of SNP performance analysis of concordant, discordant and no-call genotypes and SNPs deviating from defined thresholds for coverage, ARF, and strand bias, GT: genotypes, CG: Complete Genomics, 1000G: 1000 Genomes.

Supplementary Table S3.
Detailed concordance, no-call and discordance rates of the genotype concordance study, GT: genotypes, inter-laboratory concordance was based on six voluntary staff donor samples (marked with an asterisk), while four and five Coriell cell-line control DNAs were compared to 1000 Genomes and Complete Genomics genotypes respectively.

Supplementary Table S4.
Genotypes for two different analyses of the aDNA sample S7.

Supplementary Table S5.
A) Proportions of homozygous, heterozygous and no-calls for mixed DNA components S5 and S6 and for the expected genotype mixtures. Counts and percentages only considered 136 A-SNPs. B) Amongst the expected mixtures heterozygous SNPs were divided into: i) balanced – same numbers of each allele; ii) imbalanced – a higher number of one allele over the other (depending on donor genotypes and mixture ratio); and iii) undetermined – when missing genotypes in donor samples means the numbers of each allele cannot be determined.

Table 1. 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			x		
9947A 100 pg	25	●				
9947A 50 pg	25	●		x		
9947A 25 pg	25			x		
9947A 100 pg	21+5				x	
9947A 100 pg	25+5	●				
9947A 50 pg	25+5	●			x	
9947A 25 pg	25+5				x	
007 10 ng	18			●		
007 1 ng	21	●		●		
007 100 pg	21			x		
007 100 pg	25	●	Δ			Δ
007 50 pg	25	●	Δ	x		Δ
007 25 pg	25			x		
007 100 pg	21+5				x	
007 100 pg	25+5	●	Δ			Δ
007 50 pg	25+5	●	Δ		x	Δ
007 25 pg	25+5				x	

x Same sample re-amplified

Δ Library replicates

905 **Table 2.** Concordance details for comparisons made between Ion PGM™ genotype calls and online data for
906 Coriell cell-line control DNAs. Italic-bold genotypes denote suggested discordances on the basis of consensus.
907

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, 3.4.1
Y-rs2032597	NA07029	T	C	(no Y data)	(no Y data)	See sections 3.3.2, 3.4.1
rs2399332	NA18498	***TT***	GT	GT	GT	See sections 3.3.2, 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

908

Table 3. 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-?	

Figure 1

Click here to download high resolution image

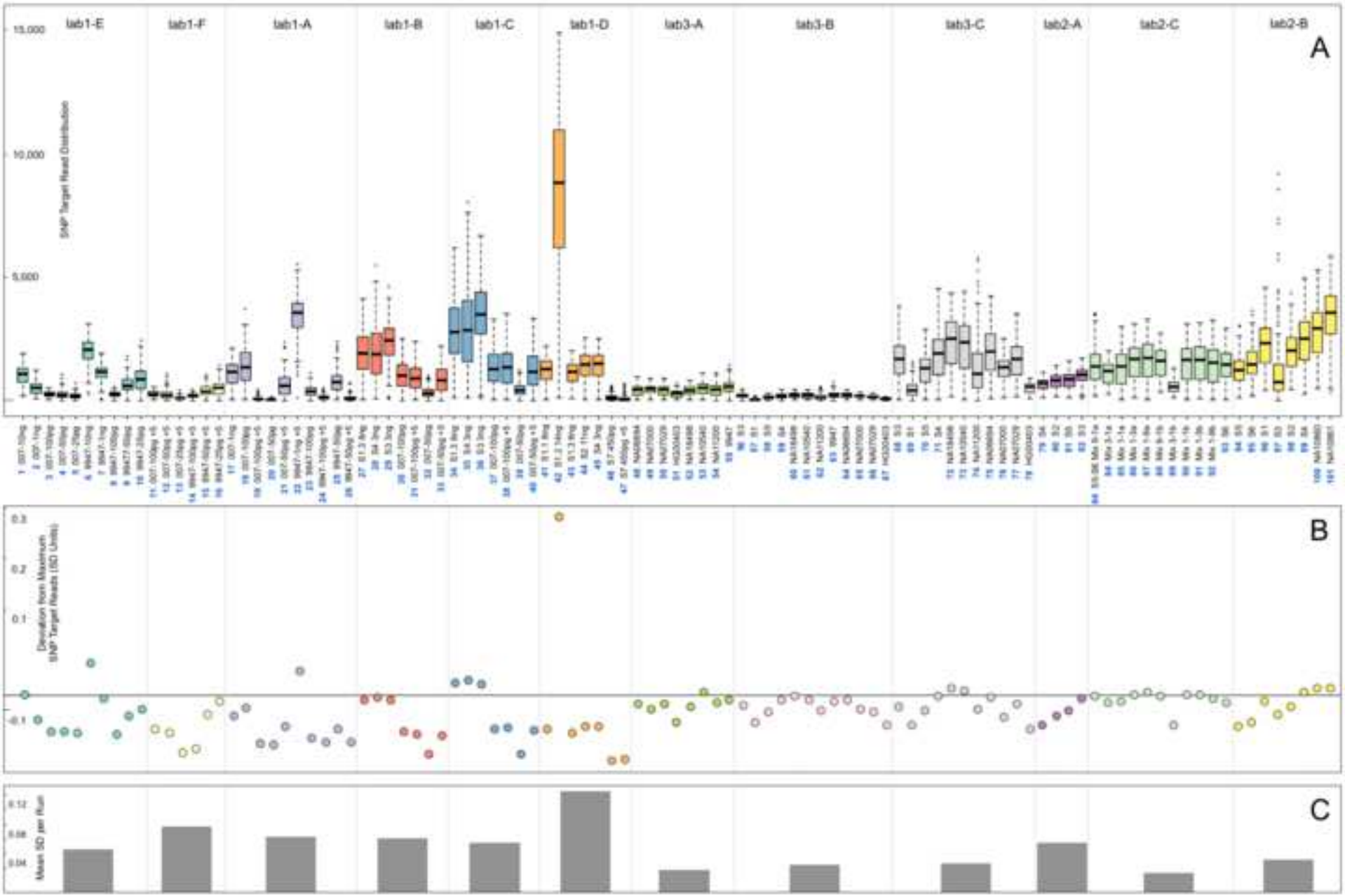


Figure 2

Click here to download high resolution image

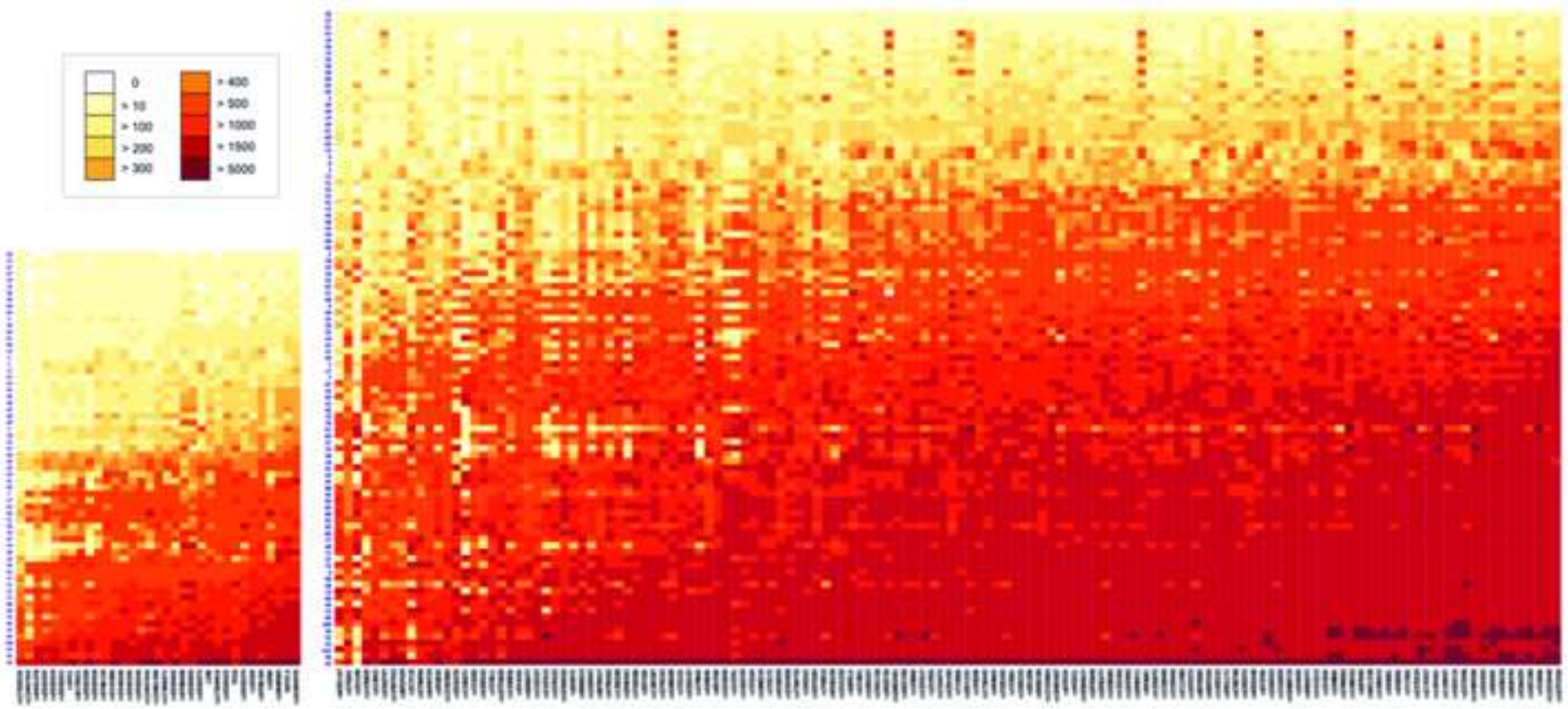


Figure 3

Click here to download high resolution image

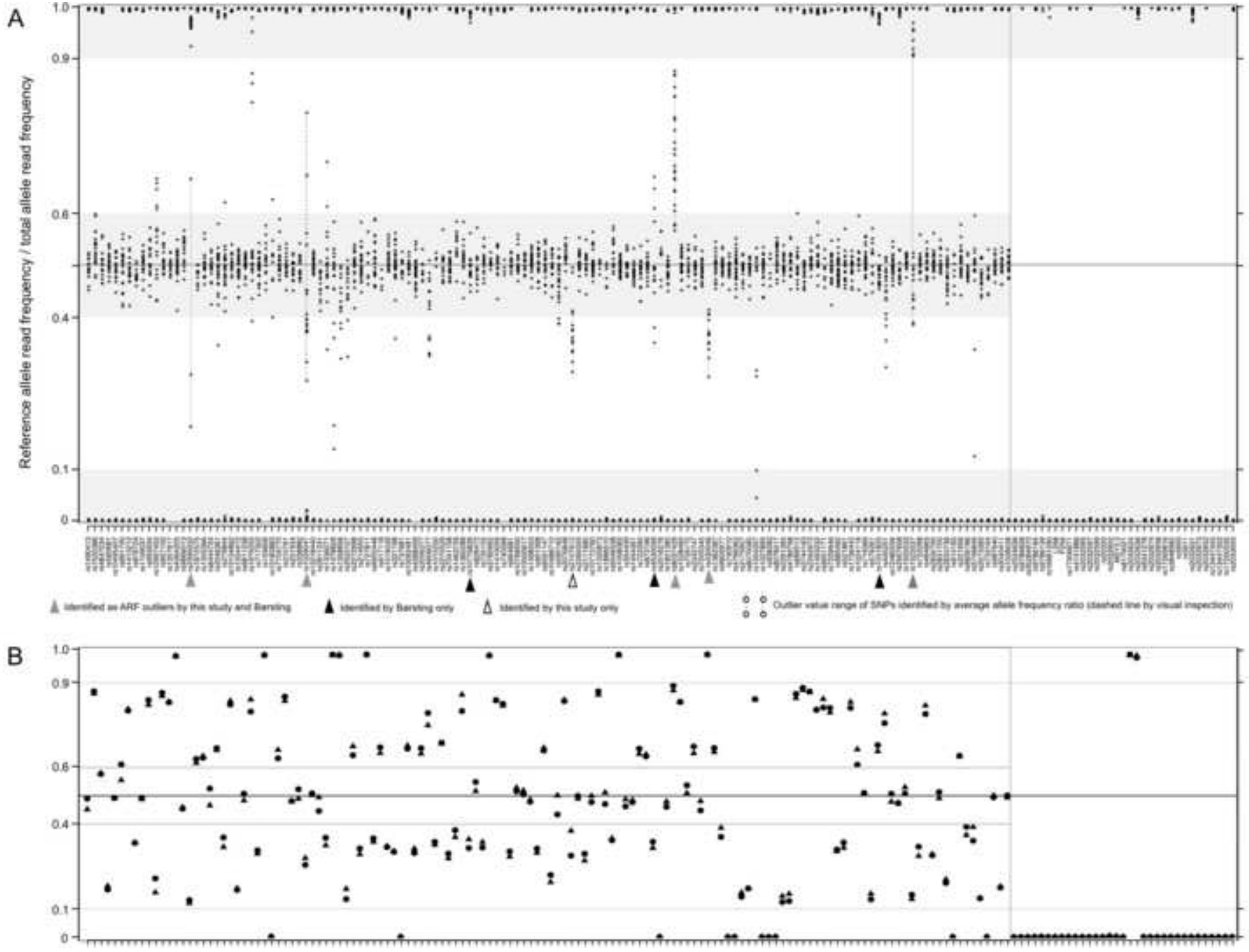


Figure 4

Click here to download high resolution image

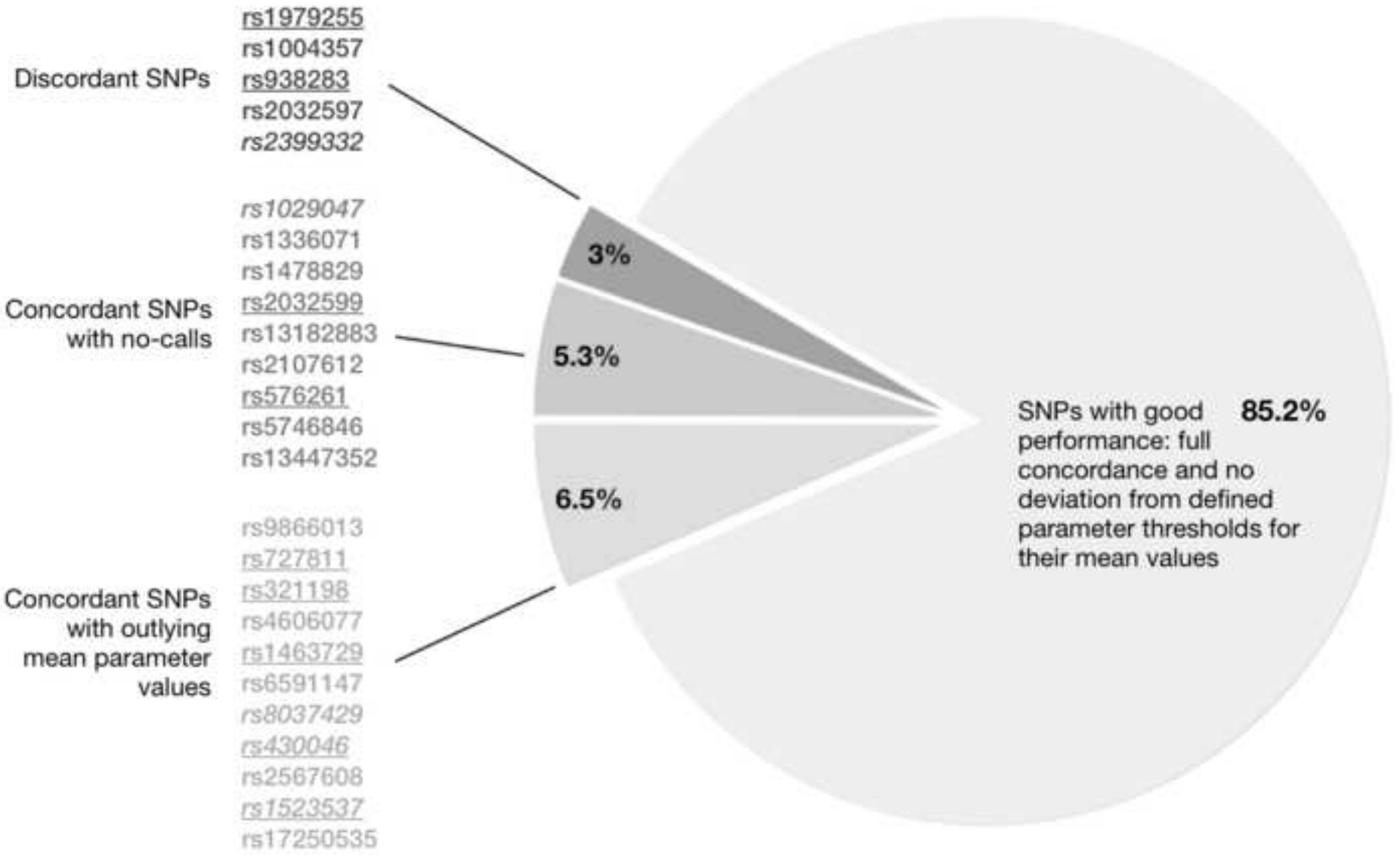
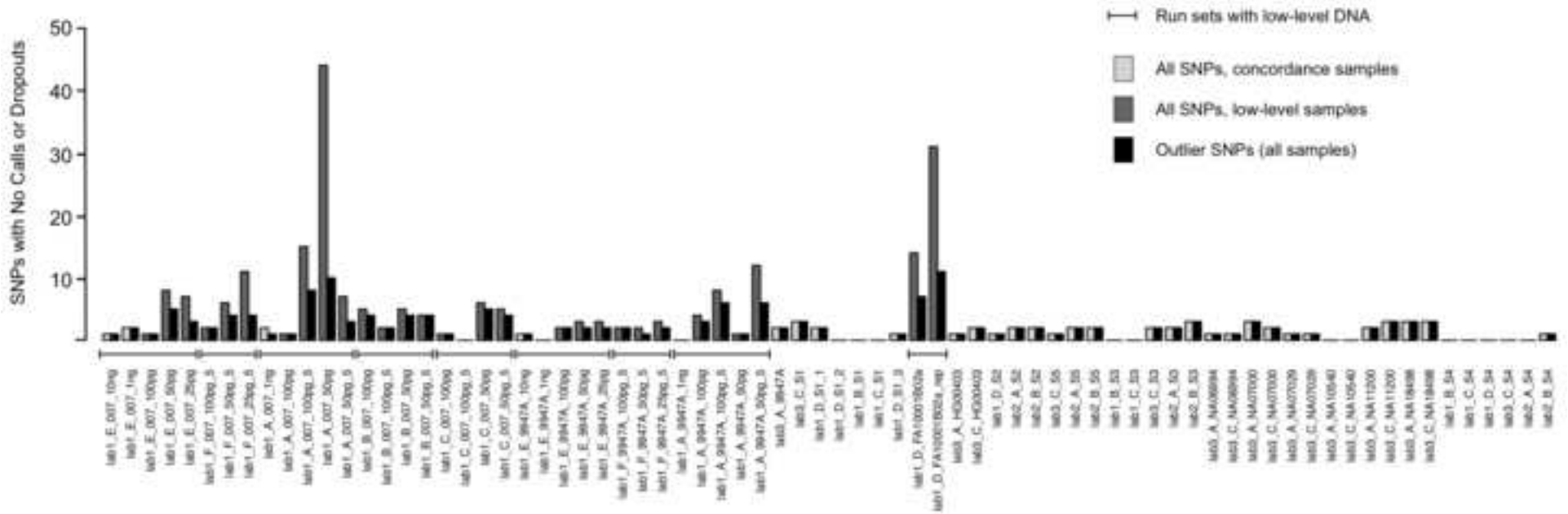


Figure 5

Click here to download high resolution image



Supplementary Tables

Click here to download e-component: Ion PGM Supplementary Tables.xlsx

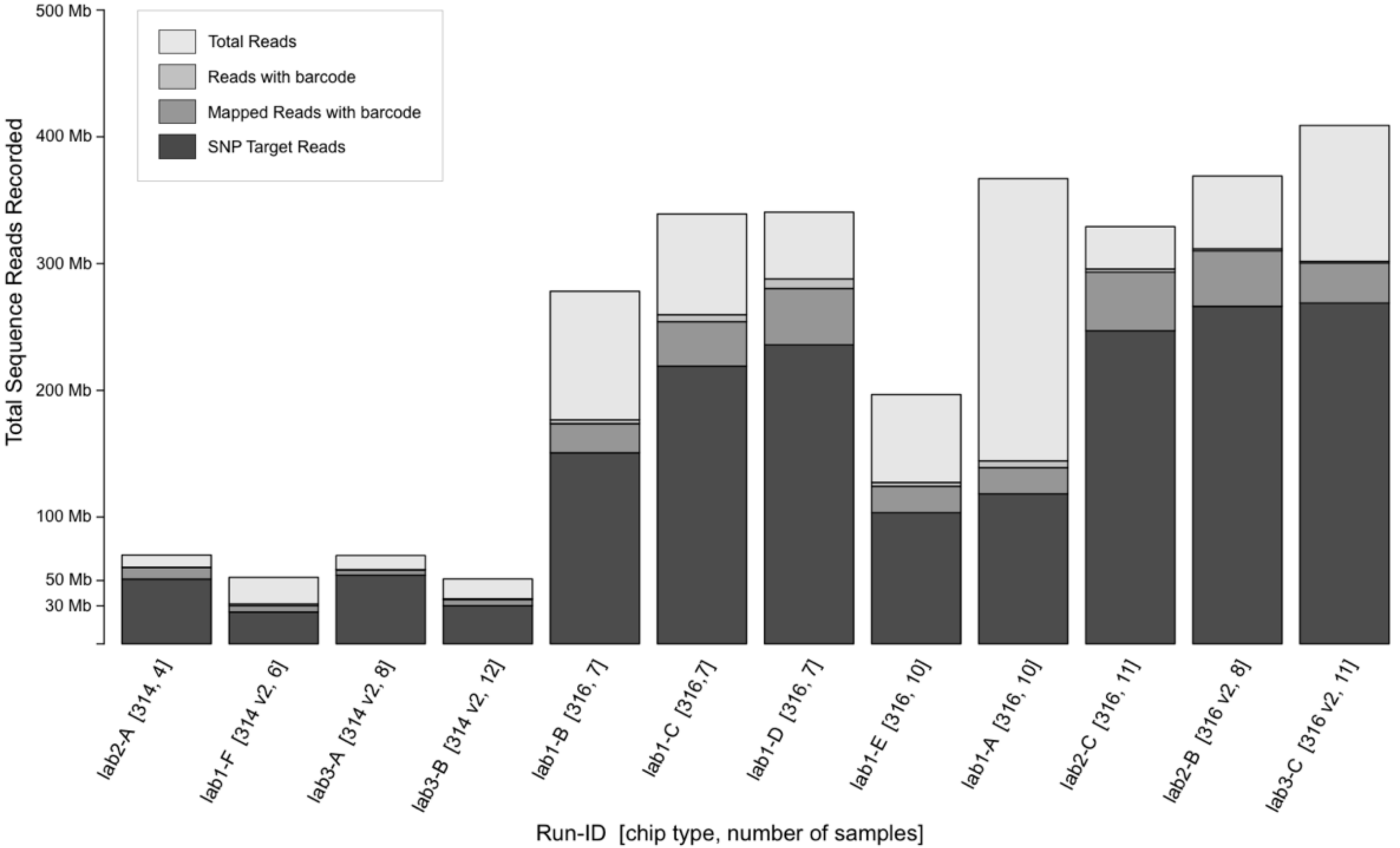
Supplementary Table S1: Expected sequence throughput of Ion PGM™ based on chosen sample numbe

RunID	Chip Type	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

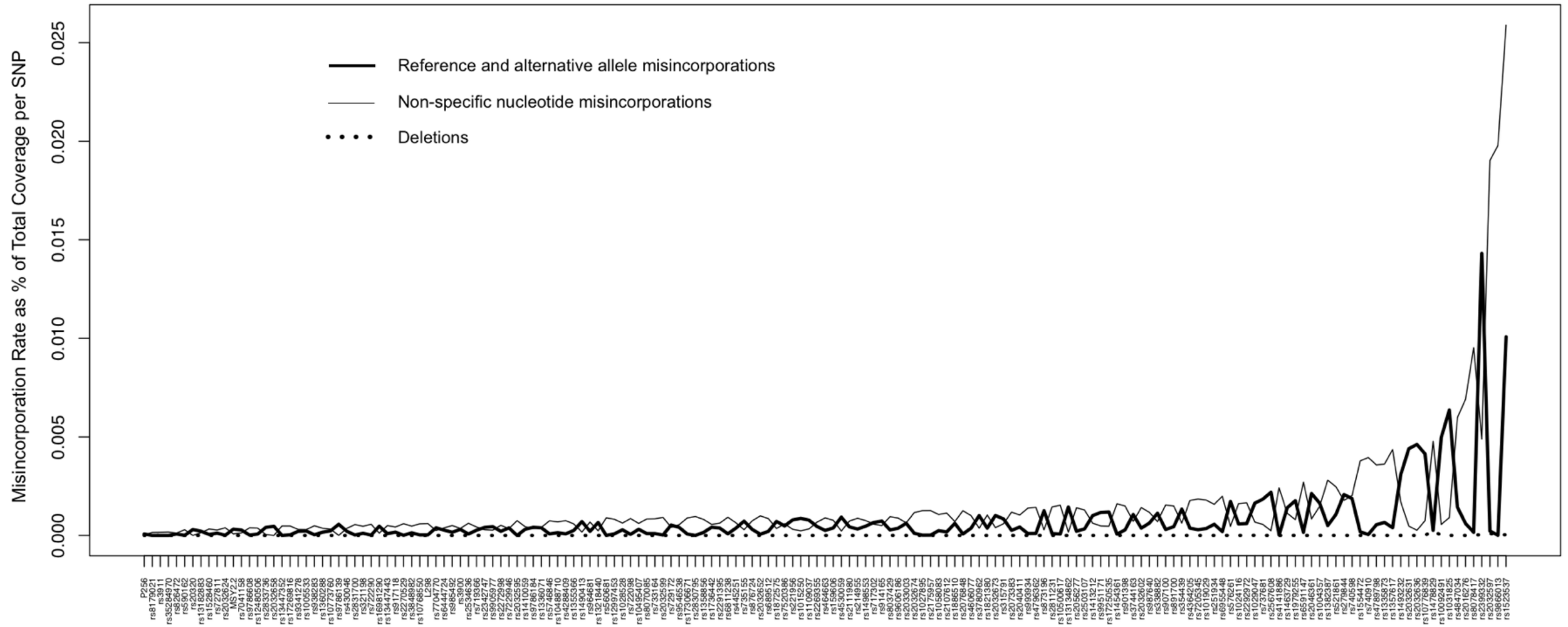
Supplementary Figures

Click here to download e-component: Ion PGM Supplementary Figures.pdf

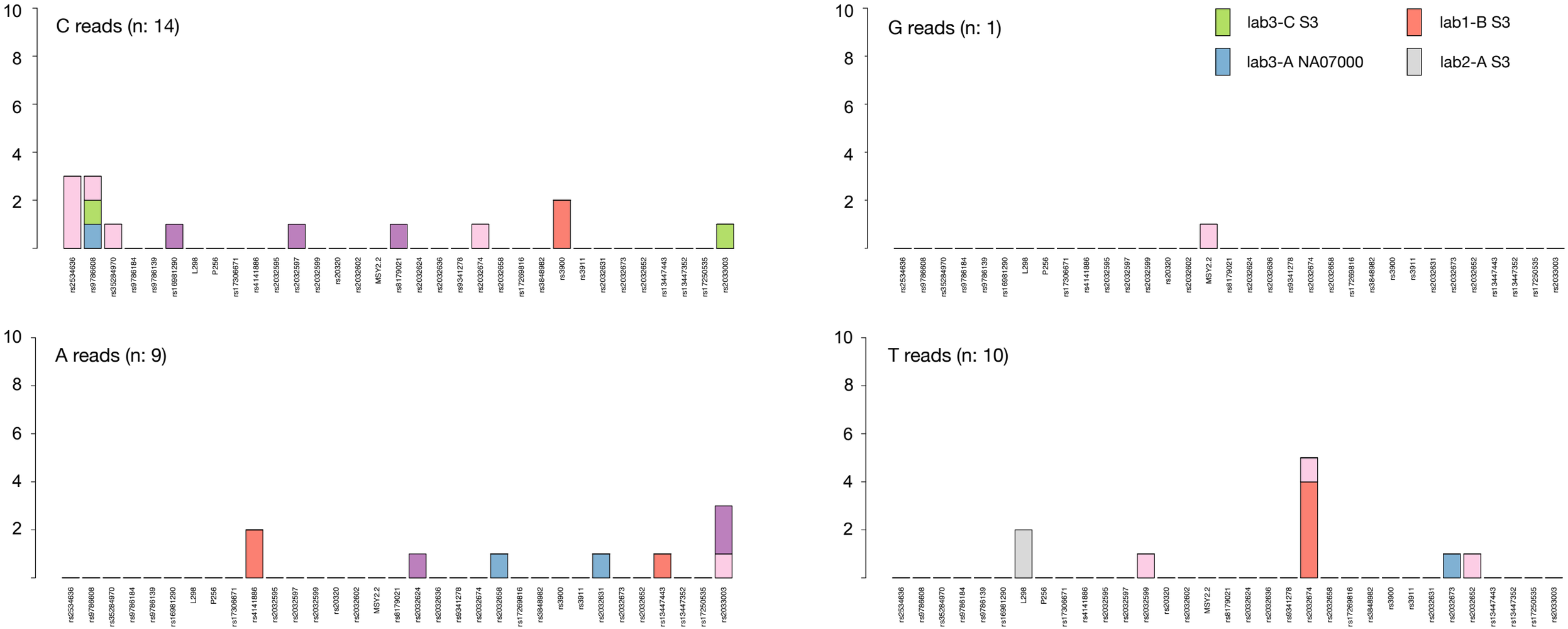
Supplementary Fig. S1 Proportions of four types of sequence reads from 12 Ion PGM™ runs using the full range of available sequencing chips.



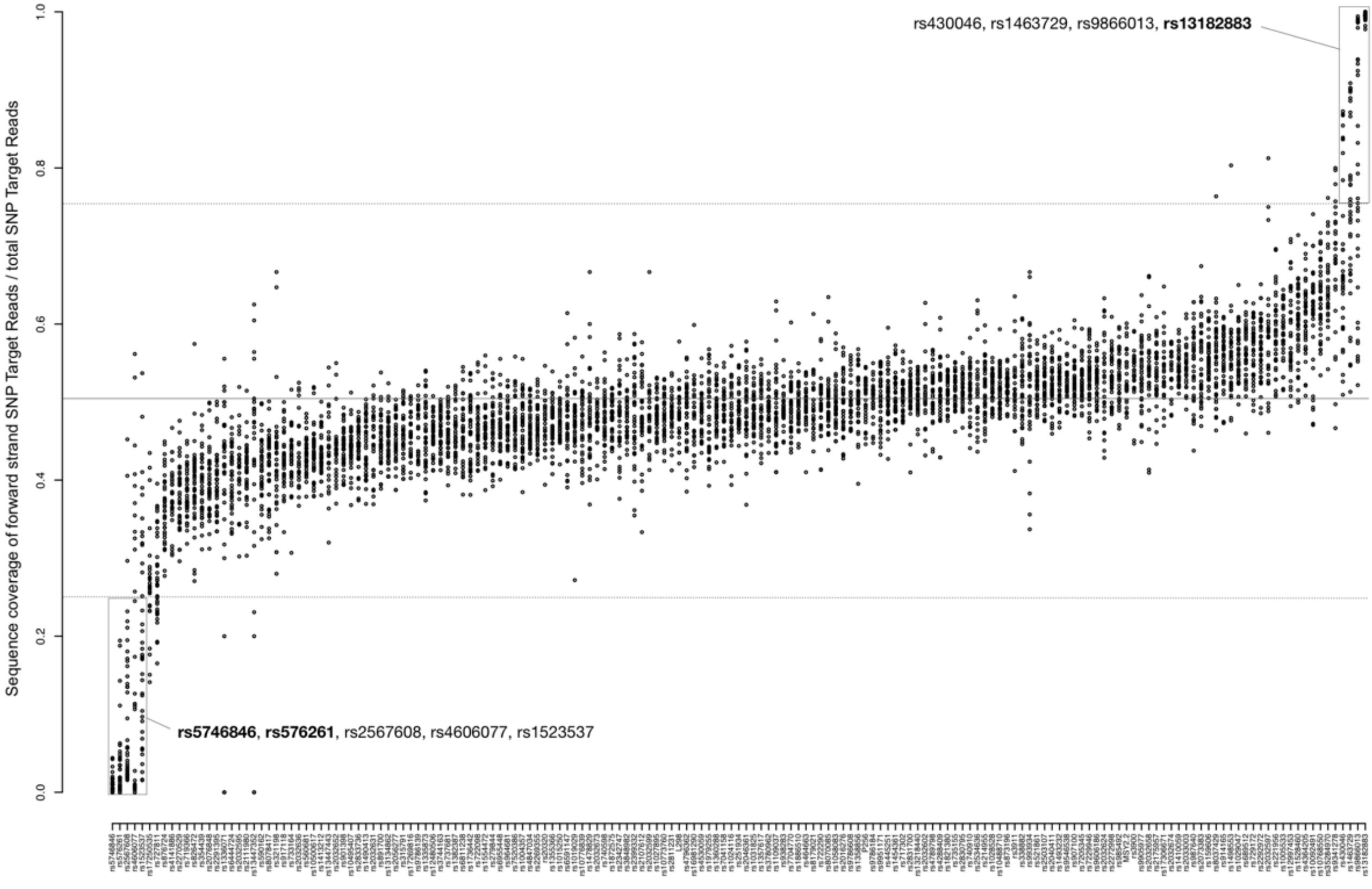
Supplementary Fig. S6 Base misincorporation rates recorded as the presence of non-allelic reference or alternative bases (e.g. low levels of A in G homozygotes plus G in A homozygotes); non-specific base incorporation (e.g. C or T in an A/G SNP) and deletions.



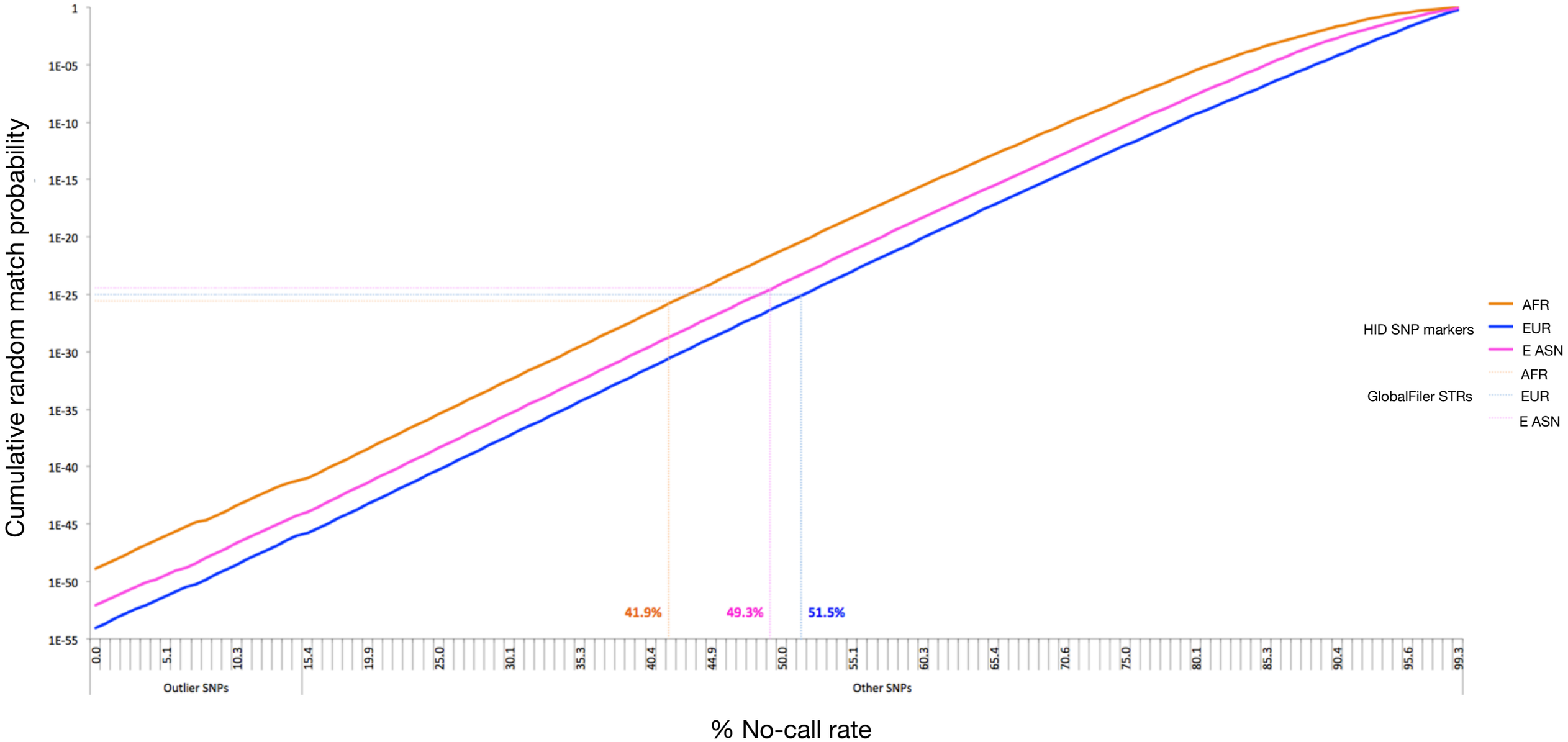
Supplementary Fig. S7 Y-SNP nucleotide reads recorded in analyses of female DNA samples. Numbers of reads indicate very low levels of extraneous male sequences amongst much higher quantities of autosomal SNP target sequence obtained (34 sequences in 6 samples).



Supplementary Fig. S8 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 3.4.2 and the IGV overview of rs13182883 is given in Supplementary File S3-SNP 2.



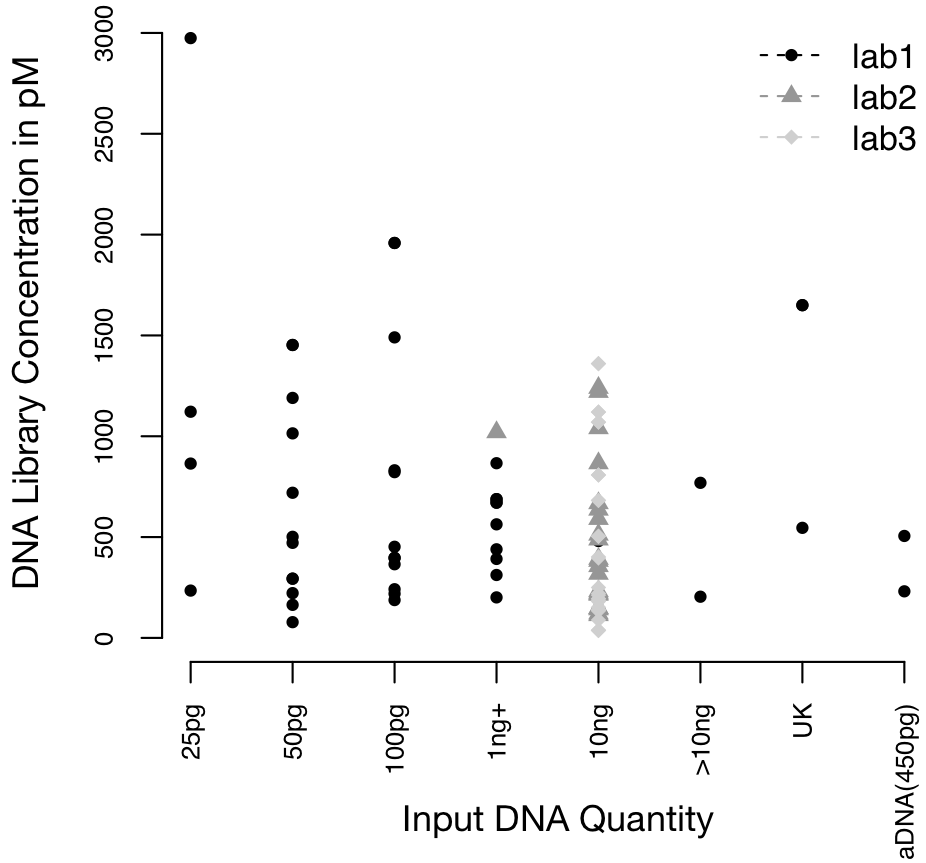
Supplementary Fig. S11 Reduction in cumulative RMP with increasing no-call rate



Supplementary File S1. Assessments of sequence coverage obtained with HID SNP markers and the Ion PGM™

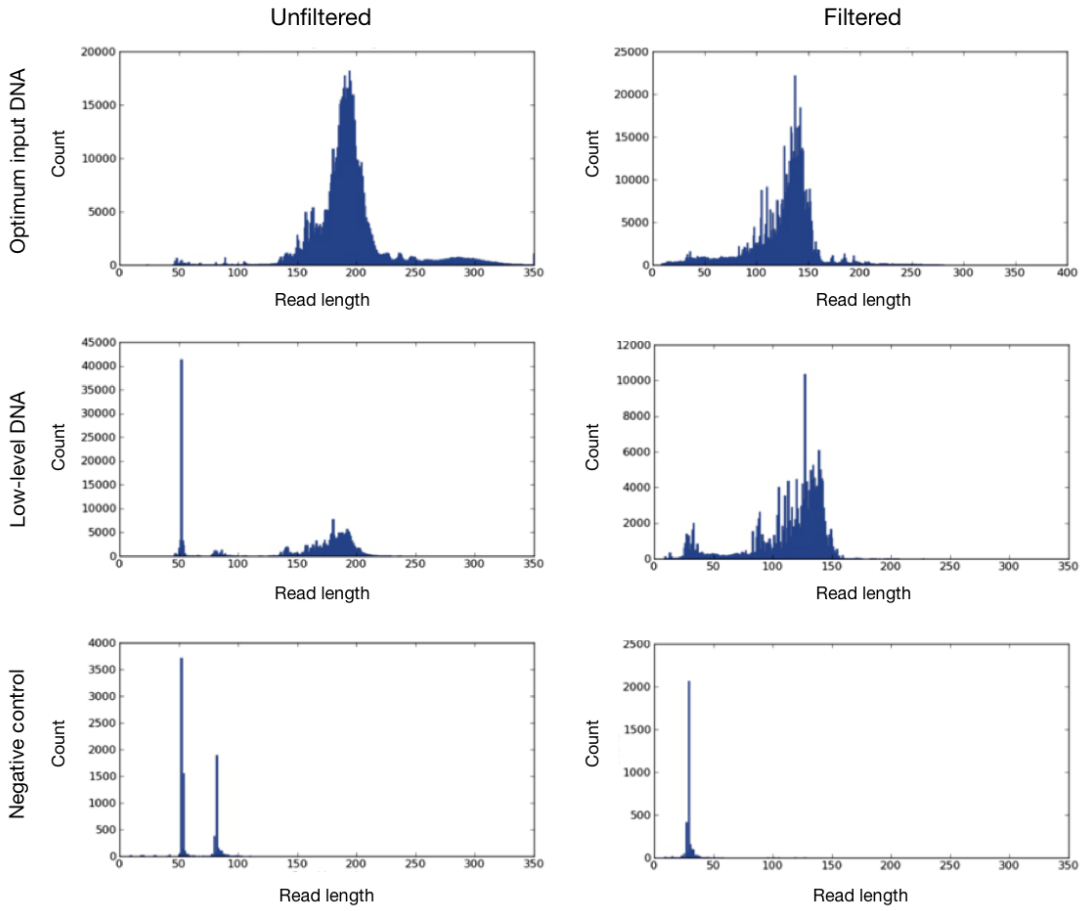
1.1. Comprehensive sequencing output analysis

Dimers can bias quantitation results upwards, especially in low-level DNA samples. In fact, this study found quantitated library concentrations varied in all samples and no relationship was found with DNA input amount, total amplification cycles, laboratory or quantitation method. Supplementary Fig. S2 plots input DNA quantity against the library concentrations obtained from the 101 samples, with no evident link between them. Runs were reanalysed disabling all filters (parameter setting: 'Basecaller Args' =-disable-all-filters off) in order to obtain all reads, unfiltered and untrimmed, per sample. The difference in reads between analyses of each sample, shows that low-level DNA sample reads are significantly more prone to filtering than optimal input DNA quantity samples ($p=2x10^{-6}$, alpha=0.05). Further analysis of reads requires manipulation of bam files outside the Torrent Suite environment and would not be feasible within a forensic setting. Reads around 50 bp in unfiltered bam files correspond to adapter dimers that are significantly higher in low-level DNA samples.



Supplementary Fig. S2. Concentration of DNA libraries obtained from seven initial input DNA quantities (or UK: unknown) in 101 analyses. We followed the Ion PGM™ guidelines of 10 ng DNA input for most runs, but the more varied input amounts of lab1 shows no relationship to library concentrations.

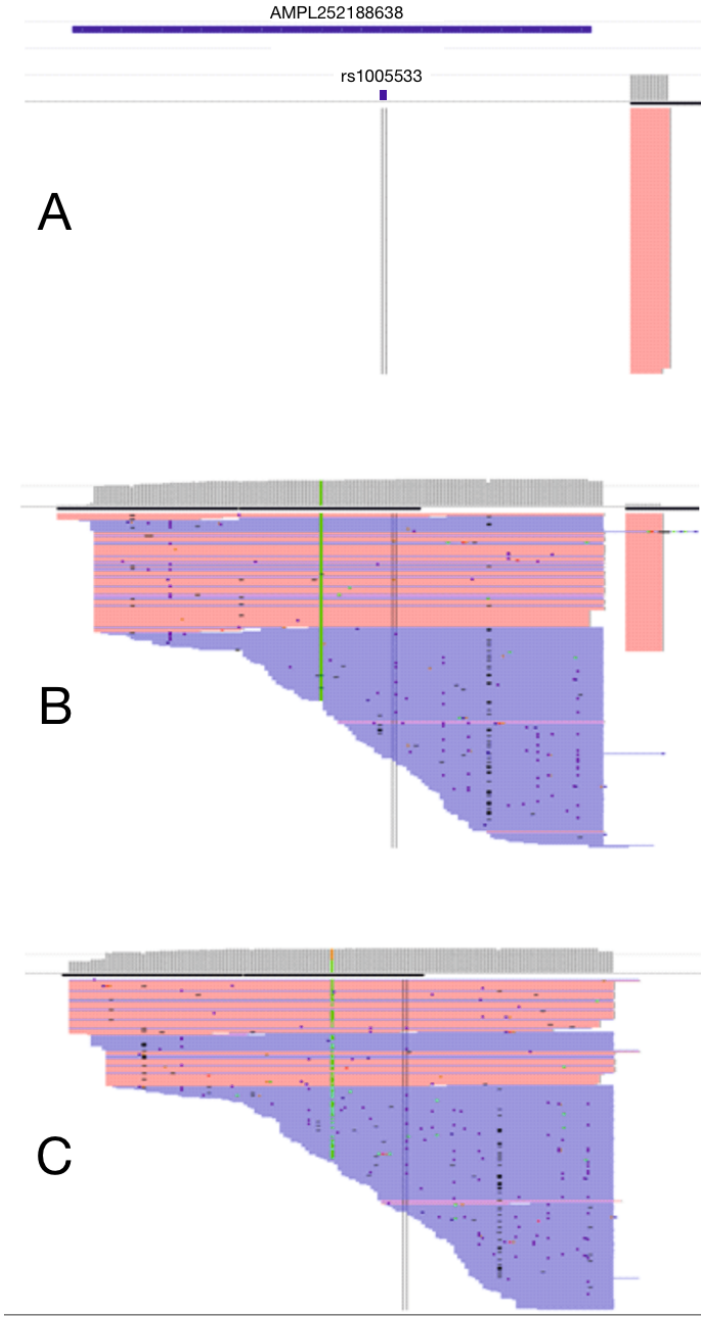
Negative controls were sequenced in two different runs. In the first run two optimum input DNA samples (positive controls) diluted to 100 pM and two undiluted negative controls were pooled and diluted 2:23 for template preparation and run on a 314v2 Chip. The percentage of polyclonal reads rose to 51% compared to averaged 30% (SD 0.8) in the 12 runs used in this study. For the second run, six negative controls and one optimum input DNA sample diluted to 100pM were pooled. To keep the final library pool concentrations between 1-2 pM the undiluted pool was subjected to template preparation. This run yielded 79% of polyclonal reads and only the optimum input DNA sample gave any results.



Supplementary Fig. S3. Read length histograms of an optimal input DNA sample, low-level DNA sample and a negative control before and after read filtering (right, left). Pronounced peaks at ~ 50bp in low level DNA and negative control samples correspond to adapter dimers.

In the sequenced negative controls 64% of total reads (6065/9783) were filtered due to low quality or adapter dimer issues. As with low-level DNA samples, the negative control shows high adapter dimer peaks around ~50bp in the unfiltered read analysis as shown in Supplementary Fig. S3. From the remaining 3650 reads, 76% (2801) mapped to hg19. Out of those reads, five mapped to rs1058083: numbers comparable those observed in low-level Y chromosomal SNP detection in female samples (see section 3.2.1). For the analysis of the remaining reads we compared the negative control to one low-level and one optimum input DNA male sample visually

by using IGV, shown in Supplementary Fig. S4. Another 6% (228/3650) of non-filtered total reads appear to be random matches throughout the genome. 28% of non-filtered total reads (1036/3650) match to 61 genomic regions, which also appear randomly in low-level or optimum input DNA samples.

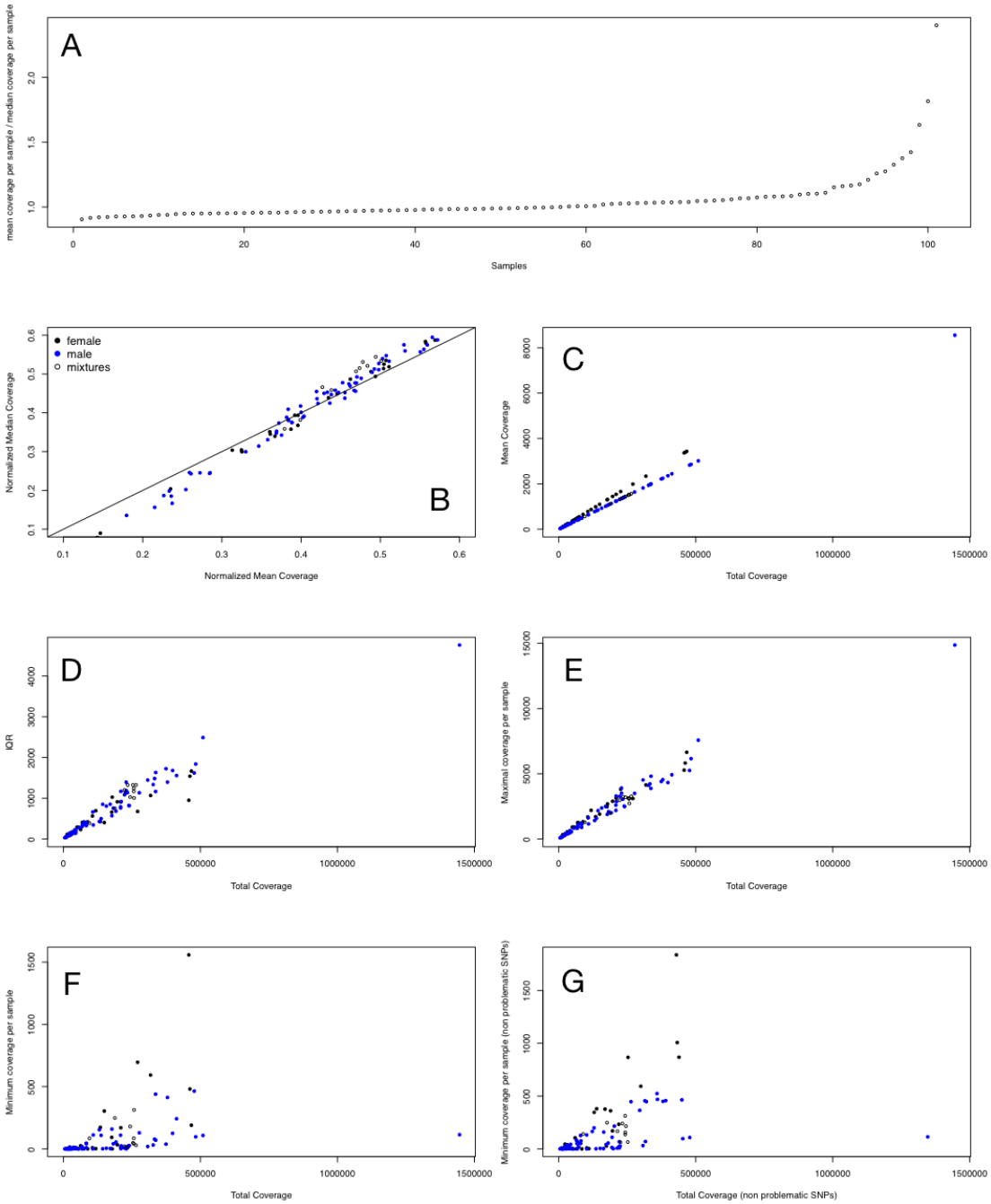


Supplementary Fig. S4. Comparison of primer regions reads for rs1005533 in a negative control, low-level and optimum input DNA sample, from IGV graphical summaries.
(**A**) Negative control shows short reads in the primer region of targeted rs1005533.
(**B**) Similar reads can be seen in a low-level DNA sample.
(**C**) The optimum input DNA sample does not show any short reads in the target neighbouring region. For better visualization reads are down-sampled to 100. Pink sequences are forward direction, violet reverse.

However, the majority of those reads are of low quality (<30). Another 39% (1421/3650) of the non-filtered total reads in the negative control are directly adjacent to SNP target regions suggesting these are sequenced complete or truncated multiplex primers from target amplification. For 25 of these regions, short primer sequence reads were also found in low-level DNA samples but not in optimum input DNA samples. The most prominent example for this is the target region of rs1005533 on chromosome 20 (Supplementary Fig. S4).

1.2. SNP sequence coverage assessments

When mean coverage values are compared to median values, a skew in the distribution of coverage with increasing mean coverage is seen across 101 analyses (Supplementary Fig. S5 A-C), suggesting a certain amplification bias within the multiplex PCR. Maximum coverage, the interquartile range and mean coverage levels all rise as total coverage increases (Supplementary Fig. S5 D-E). However, minimum coverage is not directly related to total or mean coverage in a simple linear fashion. When removing outlier SNPs, increased coverage tends to show a slightly improved positive correlation to increased minimum coverage values throughout the data (Supplementary Fig. S5 F-G).



Supplementary Fig. S5. HID SNP panel coverage distribution parameters.

(**A**) Ranked mean/median coverage ratios showing discernible skew in rightmost 13 analyses where lower SNP Target Reads were obtained than mean values would predict.

(**B**) Unity-based normalization of mean SNP coverage vs. median SNP coverage per analysis.

(**C**) Normalization of absolute mean SNP coverage vs. median SNP coverage. Both plots show that not all data points lie on the diagonal line, implying a non-normal distribution of mean values. This suggests amplification bias amongst HID SNP components with increasing total coverage (accentuated by raised 169-SNP competition in male PCR).

(**D-E**) Interquartile range of SNP coverage per sample and maximum coverage rise with total coverage.

(**F-G**) Minimum coverage per sample vs total coverage sample. Minimum coverage is not linearly influenced by total coverage levels - when removing outlier SNPs there is a slight improvement in relatedness.

Supplementary File S2

Click here to download Component: Idn_PCM_Supplementary_File_S2.pdf

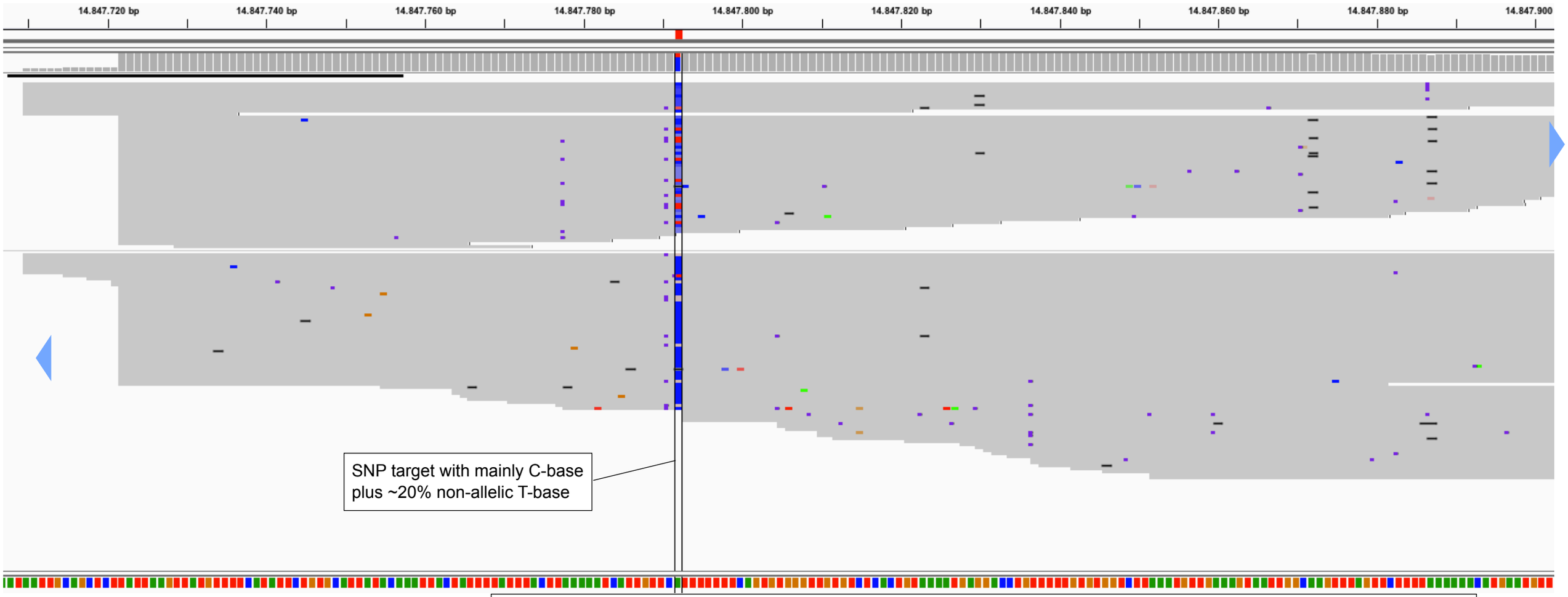
Supplementary File S2

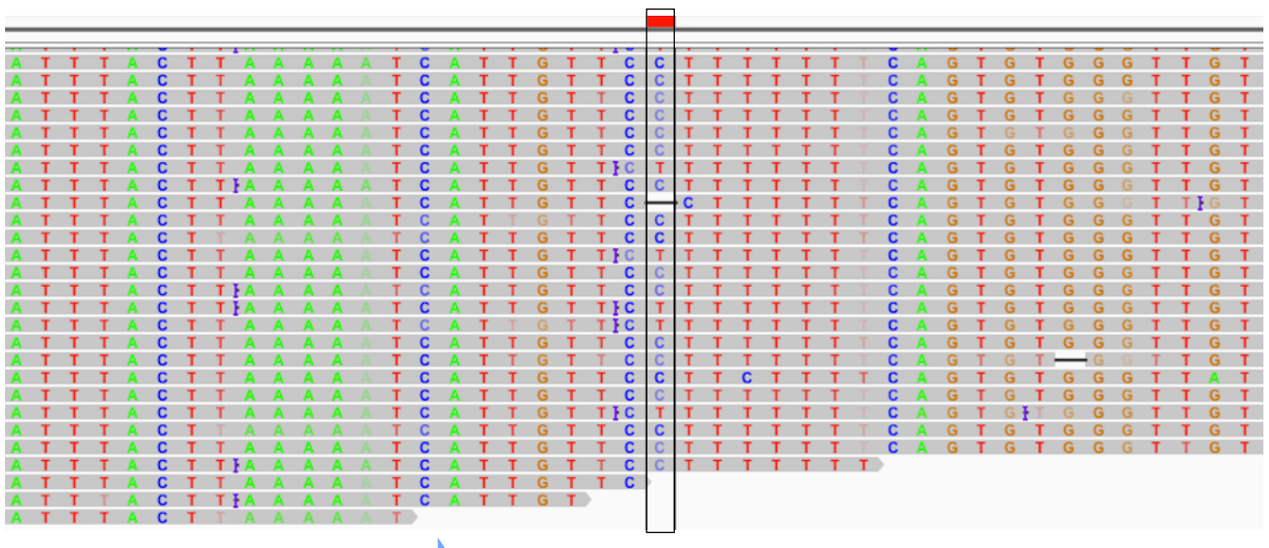
SNP 1: rs2032597

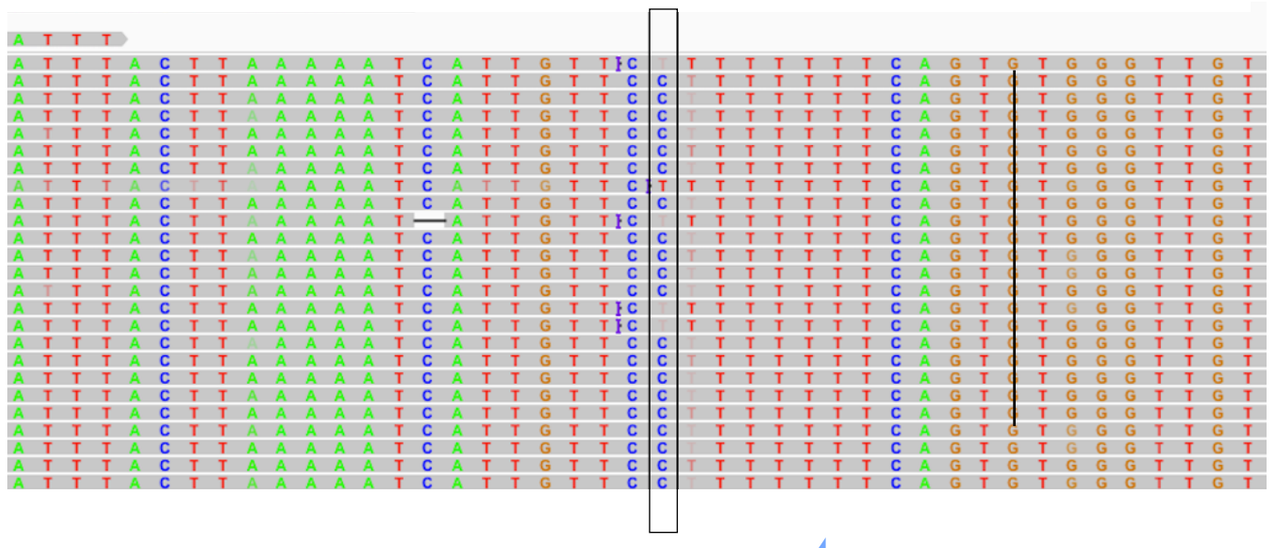
IGV overview of discordant Y-rs2032597 showing misalignment of the SNP site and immediate sequence due to C anchor base (matching one SNP allele) within a poly-T tract

C T A G Direction

Insertion (INS) Deletion (DEL)

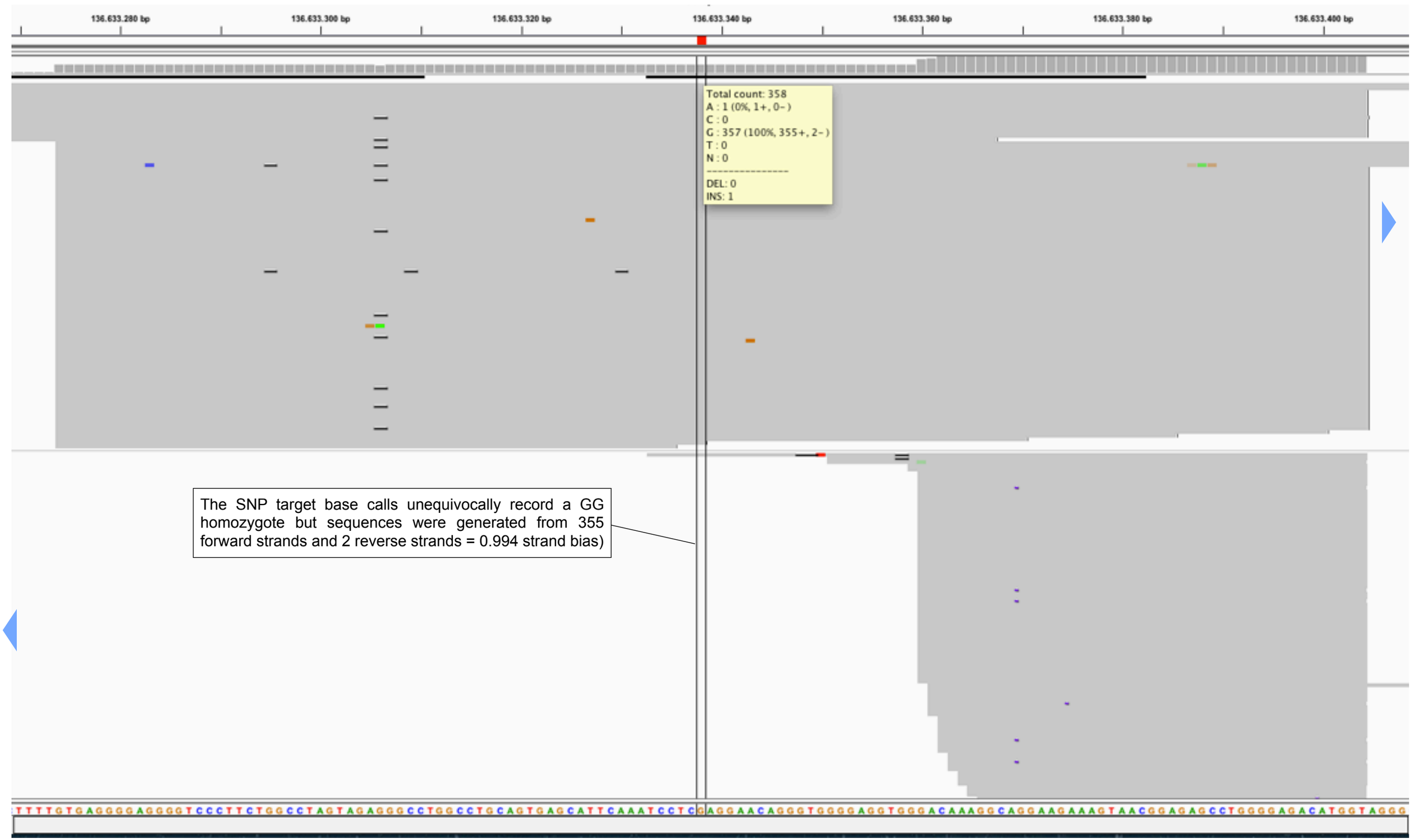






Supplementary File S2
SNP 2: rs13182883

IGV overview of rs13182883 showing very strong strand bias. In this SNP the reverse strand sequencing is initiated but stops after ~40 bp.



Supplementary File S2
SNP 3: rs1029047

IGV overview of A/T SNP rs1029047 sited within poly-A and poly-T tract. AA homozygotes show systematic sequence alignment problems from the upstream 3-T tract creating an overlapping T-base.

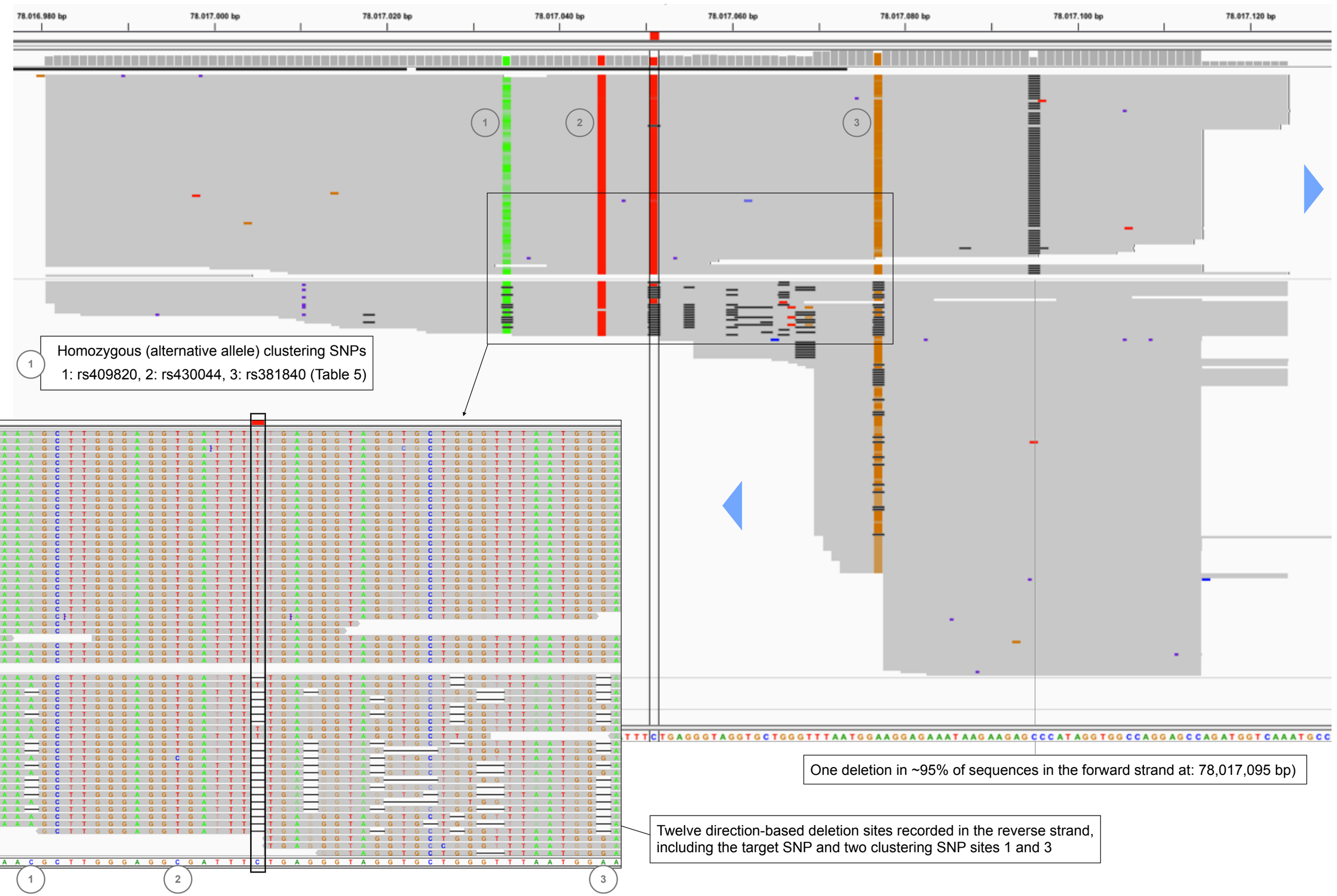
Total count: 408
A : 328 (80%, 196+, 132–)
C : 1 (0%, 0+, 1–)
G : 0
T : 79 (19%, 41+, 38–)
N : 0

DEL: 1
INS: 203

The SNP target base sequence counts indicate 19% spurious T-base calls were made due to misalignment of 3-T tract in the forward strand or misalignment of both 3-T and 8-A tracts in the reverse strand

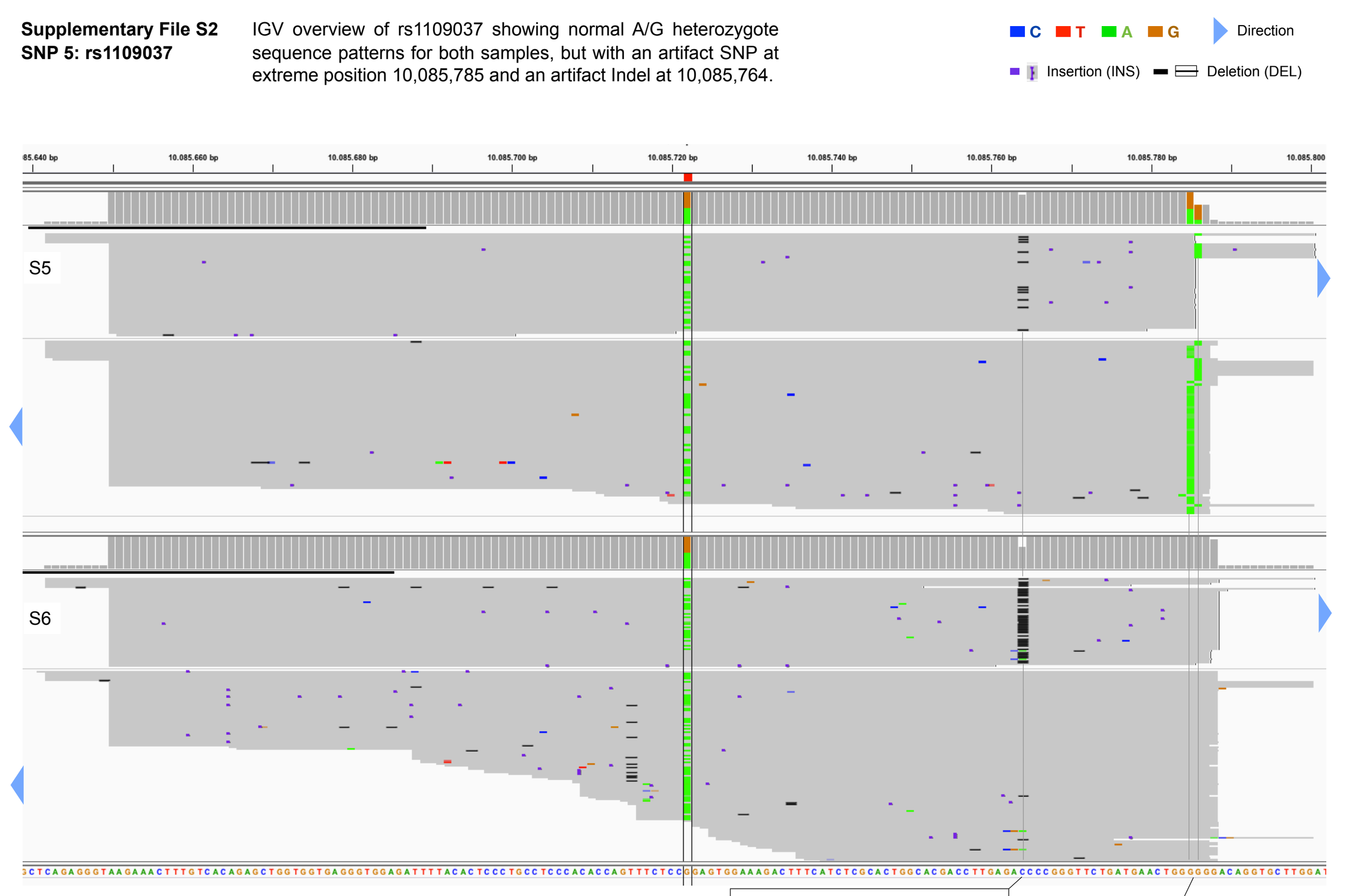
Supplementary File S2
SNP 4: rs430046

IGV overview of rs430046 showing highly directional Indel calls, notably in the forward strand. This SNP also shows three common clustering SNPs within 60 bp of the target site.



Supplementary File S2
SNP 5: rs1109037

IGV overview of rs1109037 showing normal A/G heterozygote sequence patterns for both samples, but with an artifact SNP at extreme position 10,085,785 and an artifact Indel at 10,085,764.



Supplementary File S3. Mixture analysis with the Ion PGM™

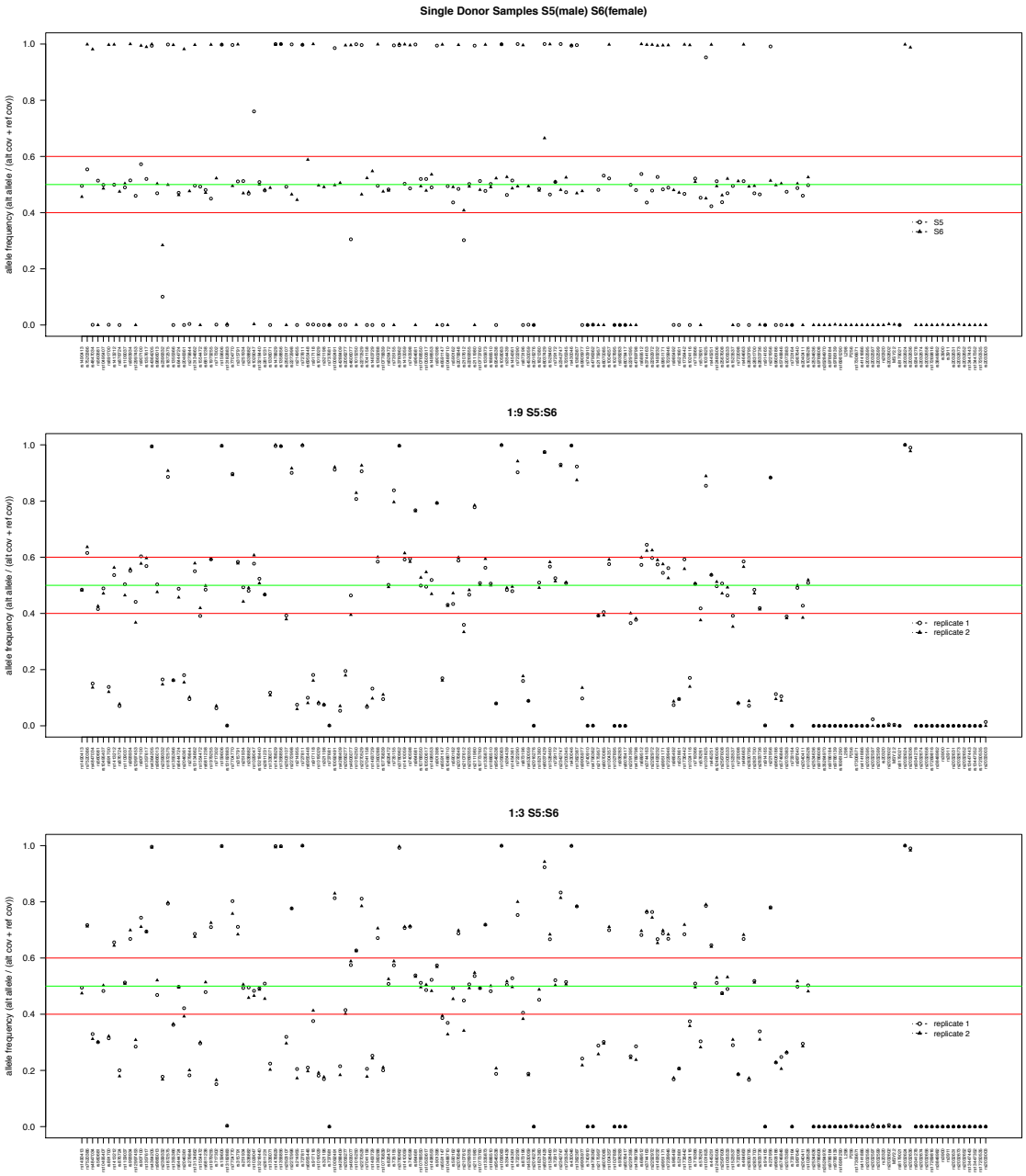
The detection of mixed source DNA samples and the de-convolution of component genetic profiles is difficult when analysing binary SNPs with the commonly used SNaPshot® chemistry. But the use of NGS, in this particular study the Ion PGM™ pipeline and the AmpliSeq™ technology, provides balanced heterozygous genotypes, a characteristic highly valuable for the analysis of mixed source samples. It is important to report a mixture as such and not as a single profile, which would probably originate misleading conclusions during a forensic investigation. Furthermore, enhanced statistical analysis will allow likelihood ratios calculation when one of the component profiles is available (for example, the victim of a sexual assault).

3.1. ARF variation in mixed DNA samples

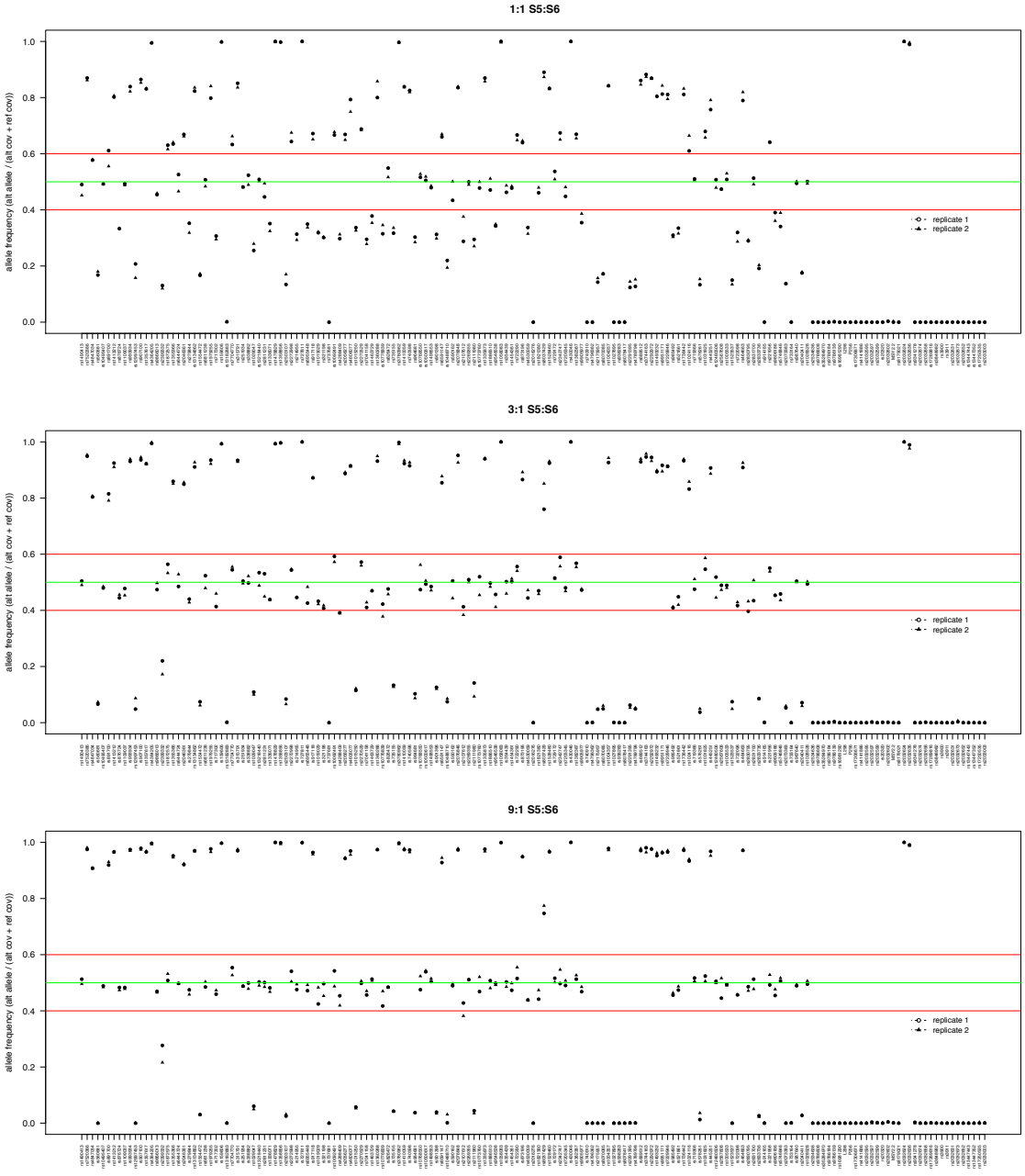
The five mixed samples were first assessed for imbalanced ARF distributions. The distributions observed in the 1:1 mixture are shown in Fig. 6B in the main article, while those of all mixture ratio replicates and donor samples in Supplementary Fig. S9. As described in section 3.3.2, nearly all SNP ARFs in unmixed DNA analyses range from 0-10% and 90-100% for homozygotes and 40-60% for heterozygotes, so the S5 and S6 donor distributions match these expected patterns in all but 2 and 3 SNPs respectively. In contrast, it is not possible to define such limits for homozygous or heterozygous SNPs in the 1:1 mixture as there is very evident scattering and a large proportion of ARFs fall within the 10-40% and 60-90% ranges. Therefore a discernible lack of ARF balance creates a comparable situation to the dye signal peak height ratios in standard STR CE analysis when these deviate from those seen in normal DNA profiles.

Although S5 and S6 have similar ARF distributions, their genotypes are different at the majority of A-SNPs. These differences were observed to affect the genotypes reported in the mixed samples and consequent ARFs. Depending on whether a donor was a minor or major component, minor alleles often went undetected. For example, when S5 is the minor component at 1:3 and 1:9 and heterozygous for a SNP that is homozygous in S6, allele ratios are 1:7 and 1:19 respectively. When S5 is the major component at 3:1 and 9:1 with S6 having an opposite homozygote or heterozygote, allele ratios range from 1:3 to 1:19. The extreme allele ratios can result in a failure to detect the minor allele component, as the minimum 10% value used to call the allele is not reached. Therefore, the 9:1 mixtures in particular, look very similar to the single donor samples, although the opposite ratios of 1:9 mixtures are noticeably more imbalanced. The contrast of 1:9 and 9:1 illustrates that a minor allele will not always escape detection, especially if more stringent ARF analysis parameters are applied. Therefore, mixtures at ratios of ~10% or less may appear more imbalanced than unmixed samples or can be near identical, depending in part, on the particular

combination of homozygotes and heterozygotes and the degree to which they contrast across contributors.



Supplementary Fig. S9. Allele read frequency distributions observed in mixed DNA analyses (red lines: heterozygote balance thresholds).



Supplementary Fig. S9. (Continued)

3.2. Changes to observed levels of heterozygosity

The second approach to assessing mixtures counted the number of heterozygous A-SNPs. Normal unmixed samples can be expected to show ~50% heterozygosity, while from the assessment of the known genotypes in S5 and S6, the expected heterozygosity of the mixture is 86.8%, as shown in Supplementary Table S5. Depending on the donor genotypes and the mixture ratio, heterozygous genotypes can be divided into balanced (equal proportions of opposite homozygote alleles or both components heterozygous for that SNP), or imbalanced categories (all other

combinations that upset a balanced heterozygous allele ratio). Supplementary Table S5 indicates that heterozygosity rises markedly in mixed samples irrespective of the mixture ratio, but the proportion of imbalanced heterozygotes rises from just over 60% observed in 1:1 ratio mixtures to 73% amongst the others.

Supplementary Table S5 (**A)** Proportions of homozygous, heterozygous and no-calls for mixed DNA components S5 and S6 and for the expected genotype mixtures. Counts and percentages only consider 136 A-SNPs.
(**B)** Amongst the expected mixtures the heterozygous SNPs were divided into: i) balanced – same numbers of each allele; ii) imbalanced – a higher number of one allele over the other (depending on donor genotypes and mixture ratio); and iii) undetermined – when missing genotypes in donor samples means the numbers of each allele cannot be determined.

(A)

	Single-donor samples				Expected mixture	
	S5		S6			
	No.	%	No.	%	No.	%
Homozygous	70	51.47	62	45.59	17	12.50
Heterozygous	64	47.06	71	52.21	118	86.77
No Calls	2	1.47	3	2.21	1	0.74

(B)

	Mixture ratios			
	1:1		Other	
	No.	%	No.	%
Balanced	45	38.14	31	26.27
Imbalanced	70	59.32	84	71.19
Undetermined	3	2.54	3	2.54

3.3. Effects of the analysis parameters on ability to detect mixed DNA

The third aspect of Ion PGM™ mixture analysis assessed the effect of different parameter settings and data downsampling limits[1]. Analysis of A-SNP data from mixtures followed the same rationale as concordance analysis. The replicated mixed samples (in this case all in run lab2-C) were analysed with the Germline low stringency

[1] Note that the Genotyper version used in this study allowed for the proportion of sequence data analysed to be adjusted by setting a downsampling value in the analysis parameter set. By default, the number of reads used to call a genotype was randomly reduced to 400 by Genotyper. A comparison of the concordance study genotypes called using the default downsampling of 400 reads vs. genotypes calls with no downsampling (increasing the maximum reads to 10,000 or 20,000 depending on the run) revealed that changes to this parameter have little effect on reported genotypes. The reduction in the number of reads for each SNP is random in effect, so allele proportions are kept almost unchanged. However, mixed samples behave differently when the downsampling parameter is modified, as small changes in the number of minor allele sequence reads may bring them down to levels that fail to reach the minimum ARF necessary for variant detection. However, it is worth mention that the most recent Genotyper version (v 4.2) has a default downsampling value of 1,000,000 so this is no longer a parameter to be considered.

analysis parameters, including the default downsampling setting (downsample_to_coverage=400). This was followed by a much higher downsampling limit of 10,000 so the full number of sequence reads was considered by Genotyper when reporting the observed genotypes. Of the 1,360 possible genotypes for all ratios and replicates, 4.4% of calls were different between downsampling options, 40% of these were due to differences in the no-call rate. In fact, when downsampling is set at 10,000, there are less missing genotypes, but some of the genotypes recovered will still be mistyped as homozygotes when the minor allele remains undetected. For this reason it is important to change this parameter to higher values when analysing mixed source samples.

Comparing the reported genotypes using Germline low stringency analysis parameters (including downsample_to_coverage=10000) with the expected mixture genotypes, there are 87/136 SNPs where discordance is detected for at least one of the replicates of each mixture ratio. Although this corresponds to 64% of the A-SNPs, only 17.35% of the 1,360 mixture genotypes were different to those expected from the known mixture components and 1.76% returned missing data – corresponding to 80.8% genotype accuracy. The discordances fall into four categories: i) 47 SNPs with minor allele dropout or no-calls in the 3:1 and 9:1 ratios – 31/47 show a dropout in both replicates for both ratios and 16/47 show a variety of no-calls and/or dropouts; ii) 29 SNPs with minor allele dropout or no-calls in the 1:9 ratio – 19/29 showed minor allele dropout in both replicates, 5 had dropout in single replicates plus 5/29 had no-calls for one replicate and dropout in the other; iii) 7 SNPs with minor allele dropout in both replicates of the 9:1 mixture; and iv) 4 SNPs with other problems, comprising 2 with only no-calls, 1 with minor allele dropout in both 1:3 and 1:9 replicates and one consistently under-performing SNP rs13182883. This SNP underperformed in 9/10 mixture samples as well as in S5 and S6 donors. As described in section 3.4, rs13182883 is amongst the SNPs recognised to produce lower quality sequence output in unmixed DNA.

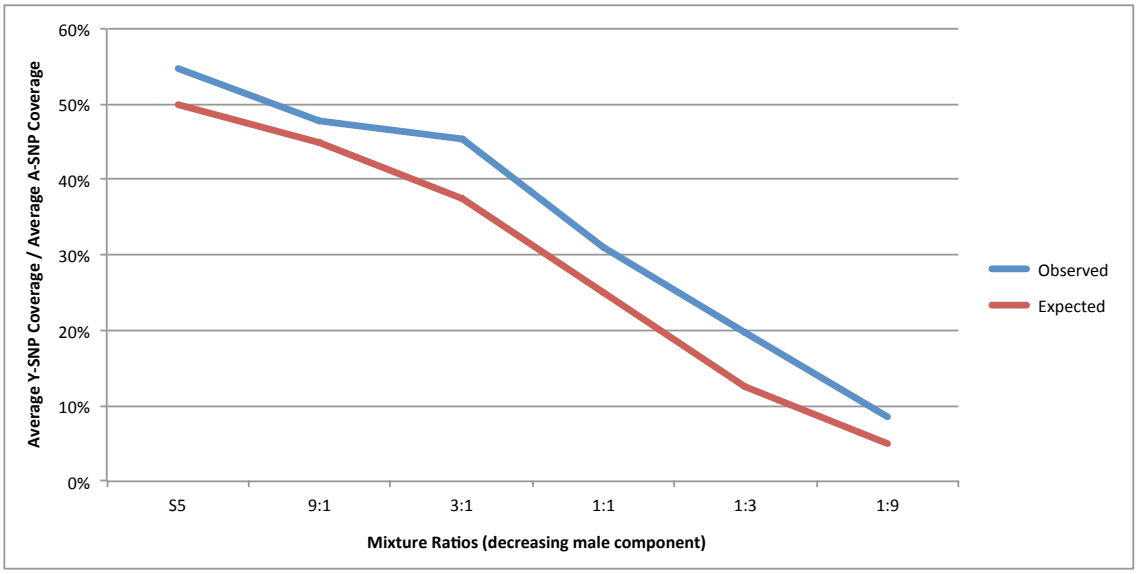
Ion PGM™ applied to medical sequencing has a strong focus on detection of somatic mutations (e.g. cancer genetics) where a novel base is present at a very low frequency compared to the normal reference-genome base. In order to control the false positive rate to manageable levels, Ion PGM™ Somatic analysis parameters are more stringent in setting conditions where a non-reference base is called. In contrast, germline mutations show identical sequence patterns to SNP variants in unmixed DNA by having equal proportions of each base at the mutated site and consequently Germline analysis parameters are the standard approach for forensic SNP analysis with the Ion PGM™. Because SNPs in mixed samples mimic the type of ARF imbalance seen between somatic mutant and reference bases, Somatic analysis parameters optimised to detect low frequency variants are more appropriate for mixtures. We applied

Somatic reduced stringency analysis parameters permitting lower minimum ARF values, as well as reduced limits on quality and coverage-per-strand limits. The default Somatic analysis downsampling is five-fold higher (downsample_to_coverage=2000) and this brings more reliable detection of the low number of sequence reads expected from minor mixture contributors. We first examined if an increase in downsampling would affect the sensitivity of somatic sequence analysis to variant alleles present at extreme ratios. The default setting of 2000 was compared to an increased minimum downsampling of 10000, but unlike Germline analysis, none of the autosomal genotype calls changed. Furthermore, when comparing them with the expected mixture genotypes only 1.32% (of a total 1,360) were different and 1.03% were no-calls. This represents an increase in genotyping accuracy to 97.65% between replicates and from comparisons to expected genotypes. Of the 136 A-SNPs, 14 had minor allele dropout in one or both replicates of the 9:1 mixture ratio and five had at least one no-call (mainly in the 9:1 ratio). Once again, rs13182883 gave missing data for the majority of replicates.

3.4. Y-SNP patterns in mixed DNA

The fourth aspect of mixture analysis examined patterns amongst the Y-SNPs, assessed separately from the A-SNPs. As mixed samples were single male-female mixtures, no second Y-SNP alleles are expected in the mixtures and patterns of mixed Y-SNP genotypes from male-male mixtures was not explored. It is noteworthy that the choice of Y-SNPs in HID-SNP affects the likelihood of finding second Y-SNP alleles in multiple male mixtures that should be explored further in future studies. As unmixed male DNA shows half the Y-SNP coverage of A-SNPs, when the minor component is male, Y-SNP coverage is substantially lower than average autosomal coverage and to a large extent the Y-SNP coverage ratio can be expected to be roughly proportional to average coverage (Supplementary Fig. S10). Observed average Y-SNP coverage goes from 55% of A-SNPs average coverage in the S5 male donor to 9% in the 1:9 mixtures, matching the expected pattern shown in Supplementary Fig. S10. Low levels of Y-SNP coverage can therefore indicate presence of a minor male component in a mixture when analysing forensic samples of unknown origins. Regarding Y-SNP genotyping accuracy, the same parameters used in the analysis of A-SNPs were applied, but no differences were observed between default analysis parameter settings and higher downsampling limits. However, in contrast to the analyses of A-SNPs in mixtures, the Y-SNP no-call rate is higher when Somatic analysis parameters are used, particularly for the 1:9 mixture. The reduction of the minimum allele frequency threshold associated with the lower coverage is responsible for the observed reduction of the Phred quality probabilities associated with the Y-SNPs when using Somatic analysis parameter settings. This particularly applies when the minor component is male. We highlight the fact that when a genotype is reported with both Germline and Somatic analysis

parameter settings, it is always concordant with the expected male genotype. The Y-SNP rs13447352 shows underperformance as it gives no-calls with both analysis parameter settings and was already identified as an outlier SNP in unmixed samples (section 3.4 in main text).



Supplementary Fig. S10. Observed and expected ratios of average Y-SNP coverage vs. average A-SNP coverage for the male component S5 and mixtures.

3.5. Summary considerations for mixture detection with the Ion PGM™

In conclusion, scrutiny of the ARF plots of Supplementary Fig. S9 show mixtures generally have clearly discernible patterns quite distinct from unmixed samples, with high numbers of heterozygous SNPs outside the 40-60% ARF region. Additionally, higher proportions of heterozygotes and a reduction of Y-SNP coverage can give clear indications of the presence of a mixed DNA sample. Our initial analyses of a limited number of mixtures indicate that 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 analysis parameter settings to obtain more accurate genotypes for the A-SNPs. Even then, care is needed with more extreme mixture ratios (here, 1:9 and 9:1) or when the major contributor is below average heterozygosity, as there is increased probability that the minor allele escapes detection. Y-SNPs should be analysed independently with Germline analysis parameter settings as this guarantees higher genotyping rates while maintaining the quality of allele calls made.