



King's Research Portal

DOI:

[10.1016/j.fsigen.2015.04.007](https://doi.org/10.1016/j.fsigen.2015.04.007)

Document Version

Peer reviewed version

[Link to publication record in King's Research Portal](#)

Citation for published version (APA):

Eduardoff, M., Santos, C., De La Puente, M., Gross, T. E., Fondevila, M., Strobl, C., Sobrino, B., Ballard, D., Schneider, P. M., Carracedo, Lareu, M. V., Parson, W., & Phillips, C. (2015). Inter-laboratory evaluation of SNP-based forensic identification by massively parallel sequencing using the Ion PGM™. *Forensic Science International-Genetics*, 17, 110-121. [1342]. <https://doi.org/10.1016/j.fsigen.2015.04.007>

Citing this paper

Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights

Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the Research Portal

Take down policy

If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

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

M. Eduardoff¹, C. Santos², M. de la Puente², T.E. Gross³, M. Fondevila², C. Strobl¹, B. Sobrino⁴, D. Ballard⁵, P.M. Schneider³, Á. Carracedo^{2,4}, M.V. Lareu², 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, 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 Department of Forensic and Analytical Science, Faculty of Life Science, King's College, London UK

⁶ Forensic Science Program, The Pennsylvania State University, PA, USA

* Corresponding author.

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

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.

1 **Inter-laboratory evaluation of SNP-based forensic identification by massively** 2 **parallel sequencing using the Ion PGM™**

3

4 **Abstract**

5

6 Next generation sequencing (NGS) offers the opportunity to analyse forensic DNA samples
7 and obtain massively parallel coverage of targeted short sequences with the variants they
8 carry. We evaluated the levels of sequence coverage, genotyping precision, sensitivity and
9 mixed DNA patterns of a prototype version of the first commercial forensic NGS kit: the HID-
10 Ion AmpliSeq™ Identity Panel with 169-markers designed for the Ion PGM™ system.
11 Evaluations were made between three laboratories following closely matched Ion PGM™
12 protocols and a simple validation framework of shared DNA controls. The sequence
13 coverage obtained was extensive for the bulk of SNPs targeted by the HID-Ion AmpliSeq™
14 Identity Panel. Sensitivity studies showed 90-95% of SNP genotypes could be obtained from
15 25-100 picograms of input DNA. Genotyping concordance tests included Coriell cell-line
16 control DNA analyses checked against whole-genome sequencing data from 1000 Genomes
17 and Complete Genomics, indicating a very high concordance rate of 99.8%. Discordant
18 genotypes detected in rs1979255, rs1004357, rs938283, rs2032597 and rs2399332 indicate
19 these loci should be excluded from the panel. Therefore, the HID-Ion AmpliSeq™ Identity
20 Panel and Ion PGM™ system provide a sensitive and accurate forensic SNP genotyping assay.
21 However, low-level DNA produced much more varied sequence coverage and in forensic use
22 the Ion PGM™ system will require careful calibration of the total samples loaded per chip to
23 preserve the genotyping reliability seen in routine forensic DNA. Furthermore, assessments
24 of mixed DNA indicate the user's control of sequence analysis parameter settings is
25 necessary to ensure mixtures are detected robustly. Given the sensitivity of Ion PGM™, this
26 aspect of forensic genotyping requires further optimisation before massively parallel
27 sequencing is applied to routine casework.

28

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

31 1. Introduction

32

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

47

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

66

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

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

81 2. Materials and methods

82

83 All Ion PGM™ protocols followed published laboratory guidelines [11-15]. The term *sample* is
84 used here for DNA extracts that were amplified then prepared for Ion PGM™ in different
85 ways (i.e. several samples may be used from one donor). The term *run* refers to sequencing
86 tests made using one Ion PGM™ chip, combining multiple samples. The term *analysis* is used
87 to describe sequencing of a specific DNA sample forming part of a run. Somatic and Germline
88 *analysis parameter settings* are distinguished from the biological terms using capitals. The
89 term *allele frequency* is used in Ion PGM™ analysis software, describing how many *sequence*
90 *reads* carry each allele per SNP. To avoid confusion with the population genetics term we use
91 *allele read frequency* (ARF).

92

93 2.1. DNA samples, extraction of DNA and preparation of artificial mixtures

94

95 Common DNAs were used to measure genotyping concordance or assess consistency of
96 sequence quality across three laboratories. These DNAs comprised: i. six voluntary staff
97 donors (S1-S6) that could be repeatedly analysed and exchanged between laboratories; ii.
98 standard 9947A and 007 forensic controls; iii. Coriell cell-line control DNAs that allowed
99 checks against online genotype data published by 1000 Genomes and Complete Genomics
100 (CG) projects (comprising: NA06994; NA07000; NA07029; NA18498; HG00403; NA10540;
101 NA11200). These DNAs provide comparisons of three independent SNP genotyping systems
102 using NGS sequencing (1000 Genomes mainly used Illumina HiSeq [9] and CG a proprietary
103 DNA nanoarray method [10]).

104

105 Dilutions of 9947A and 007 DNAs assessed the forensic sensitivity of Ion PGM™, using 10 ng,
106 1 ng, 100 pg, 50 pg and 25 pg of DNA amplified with varying PCR cycle numbers, as outlined
107 in Table 1. Two runs used eight picomolar (pM) library pools (i.e. following standard Ion
108 AmpliSeq™ library preparation guidelines). Another three runs used libraries pooled at ~26
109 pM dilution to determine if increasing library concentrations enhanced genotyping of low-
110 level DNA. Input DNA <1 ng was either amplified in 25 cycles alone, or with 5 extra
111 amplification cycles after library preparation. Two approaches assessed re-amplification: i.
112 re-amplify half the prepared library per sample and compare to no re-amplification; ii.
113 prepare separate libraries with and without re-amplification for each sample. Samples were
114 quantified for pooling with LT Ion Library Quantitation Kit.

115

116 The ability of Ion PGM™ to detect mixed DNA was evaluated with mixtures of male-female
117 DNAs S5-S6 at ratios 1:9, 1:3, 1:1, 3:1, 9:1. Each mixture ratio was prepared once, then two
118 libraries constructed for each. The two differently-barcoded libraries of each ratio were
119 combined in one template preparation step and sequenced on a single Ion 316™ chip.

120

121 One ancient male DNA sample extracted from 12th Century archaeological remains (S7 or
122 aDNA) was analysed. The skeletal preservation conditions from the site in Volders, Tyrol,

123 Austria are detailed in [16]. Sample S7 was analysed in two separate PCRs with maximum
124 input DNA (450 pg quantified with LT Quantifiler Duo), using 25 PCR cycles and 25 PCR + 5
125 library re-amplification cycles. Although this sample lacked reference genotypes, consistency
126 of SNP genotyping was checked between analyses.

127

128 *2.2. Ion PGM™ library and template preparation, enrichment and sequencing*

129

130 HID-Ion Ampliseq™ Identity Panel v2.2 libraries were constructed with Ion AmpliSeq™
131 Library Kit 2.0 following manufacturer's protocols [11-13]. Prior quantification of DNAs used
132 Qubit® ds DNA HS Assay Kit, diluting samples (not all) to guidance inputs of 10 ng in ≤6 µL.
133 Targets were amplified as recommended for 196 primer pairs with 18-21 cycles of PCR. After
134 partial digestion of primer sequences, Ion Xpress™ Barcode Adapters were ligated for
135 tagging and resulting ligation products purified with Agencourt AMPure XP magnetic beads.
136 Library quality was checked with either Qubit® ds DNA HS Assay Kit, Agilent® High Sensitivity
137 DNA Kit or Ion Library Quantitation Kit to equalise a final library of 100 pM in ≥20 µL [12].

138

139 Template preparation used Ion OneTouch™ 200 Template Kit v2, following manufacturer's
140 protocols [14]. After recovering template-positive Ion Sphere particles (ISPs), Ion Sphere™
141 Quality Control Kit was used to ensure 10-30% templated ISPs before enrichment with Ion
142 PGM™ Enrichment Beads, following manufacturer's protocols. Sequencing was performed
143 using Ion PGM™ Sequencing 200 Kit v2 and Ion 314™ or 316™ chips (both types either v1 or
144 v2) following manufacturer's protocols [14].

145

146 *2.3. Data Analysis*

147

148 Data analysis used Torrent Suite™ 4.0.2 (herein TS) and HID_SNP_Genotyper 4.0.1 plugin
149 (herein Genotyper) with low stringency parameter settings [17]. We applied
150 HID_SNP_v2.2.2_hotspots.bed plus HID_SNP_v2.2.2_targets.bed files, identifying SNPs with
151 genome build hg19. Genotyper makes variant calls using posterior probabilities calculated
152 for each possible genotype in similar fashion to GATK [18]. Posterior probabilities are
153 computed from genotype likelihoods (using Phred quality scores and prior probabilities),
154 accounting for read depth and minimum allele frequency thresholds to report quality scores
155 (QUAL values of 0 to several thousand). SNP genotypes are called when they pass a quality
156 score plus user-defined sequence filter thresholds, or are given as “NN” / “N” no-calls.

157

158 Genotyper output comprises a web-based graphical overview and two report files: a custom-
159 format text file plus a variant call format (vcf) file with SNP details. The text file lists
160 genotype calls with corresponding quality P-values, total sequence coverage from forward
161 and reverse sequence reads, number of calls for all four bases and number of no-calls at
162 each SNP position. For this study all SNP data processing of both Genotyper files was made
163 using R (v3.0.3, 2014-03-06) [19,20].

164 **3. Results and discussion**

165

166 *3.1. Sequence coverage from Ion PGM™*

167

168 Sequencing depth (depth of coverage or simply 'coverage') has a direct bearing on the
169 sensitivity and genotyping accuracy of NGS systems applied to forensic SNP typing. Its value
170 specifies the number of times each base has been read in the sequencing run. For whole
171 genome applications it is usually stated as an average value per base. However, for SNP
172 detection applications such as HID SNP, actual depth of coverage at the targeted SNP site is
173 more relevant and is given in number of reads targeting the site (herein SNP Target Reads).
174 This final number will depend on sequencing technology, raw read filtering methods and
175 how variant calls are processed. In Ion PGM™ sequencing runs, the number of wells per chip
176 that can be filled with ISPs defines the number of possible reads. Sample pooling, template
177 preparation (influencing the number of non-templated and polyclonal ISPs) and loading
178 efficiency (influencing the number of empty wells) determine the final number of
179 successfully read ISPs (monoclonal reads). During the base calling steps of TS data processing
180 monoclonal reads are further filtered for low quality and adapter dimer reads. When
181 sequencing multiple barcoded samples, equimolar pooling ahead of template preparation
182 aims for a homogenous distribution of reads between samples of the same run.

183

184 In this study, all 12 runs reached overall sequencing throughput, measured in Mb per run, in
185 compliance with TS guidelines for each chip version used (Supplementary Fig. S1). It is
186 noticeable that for runs pooling low-level and optimum input DNA samples (lab1), more
187 reads are filtered during the base calling process. A more comprehensive description of
188 primer sequence and primer dimer issues in low-level DNA samples as well as sequencing
189 results of negative controls is summarised in Supplementary File S1 (Fig. S3). While the
190 amount of filtered low quality reads per run is similar for all runs, the percentage of filtered
191 primer dimer reads is slightly higher ($p=0.029$, $\alpha=0.05$) in lab1 runs with low-level DNA
192 and optimum input DNA samples combined on the same chip. This is indicated by the SNP
193 Target Read distributions for all 101 analyses in Fig. 1A. The distribution of quartiles reveals
194 variation both within and between runs, but Fig. 1C indicates that runs combining low-level
195 DNA alongside optimum input DNA samples has higher variation between samples. Fig. 1B
196 shows the deviation from maximum achievable SNP Target Reads (see figure legend for this
197 metric's definition). In comparison to low-level DNA samples the analysis of optimum input
198 DNA samples (68 high quantity/quality DNAs of 1-10 ng) gave less deviation from expected
199 SNP Target Reads. Furthermore, Ion PGM™ coverage analysis shows significantly higher off-
200 target reads ($p=0.00045$, $\alpha=0.05$) in low-level DNA samples. We detected an increased
201 number of sequenced multiplex primers from target amplification in low-level DNA samples.
202 These primer sequences are aligned to the reference genome and account for the total
203 number of monoclonal reads in TS, but are not considered part of the amplicon, thus
204 increasing the amount of off-target reads (Supplementary File S1).

205

206 There are two main considerations for multiplex SNP typing in massively parallel sequencing
207 analyses: minimum coverage thresholds for reliable genotyping and number of samples that
208 can be sequenced in parallel to meet those thresholds. LT guidelines suggest minimum
209 coverage thresholds for germline and somatic SNP detection of 30x and 500x, respectively.
210 The threshold for somatic SNP calling is close to the values cited in whole genome and
211 enrichment variant detection studies [2, 21-25]. However, minimum coverage thresholds
212 generally depend on the sequencing application, the SNP variant-calling algorithms used and
213 analysis parameter settings. For forensic applications, a threshold of ~20x could be sufficient
214 coverage to reliably detect variants in high quality single source DNA samples, whereas
215 mixture detection and low-level DNA samples will require much higher coverage. In this
216 study, the lowest coverage values with concordant genotypes in autosomal and Y-
217 chromosome SNPs (herein A- and Y-SNPs) were 13x and 41x respectively, discounting outlier
218 SNPs. This largely matches results of a recent study by Daniel et al. finding a similar
219 minimum coverage estimate of 20x for reliable SNP genotyping [26]. In mixtures, however,
220 minimum coverage should be set higher to reliably identify minor alleles in heterozygous
221 markers. For A-SNPs, concordance between the expected genotypes in the mixture and
222 those of the components was obtained with an average 269x coverage or higher. Y-SNPs
223 gave concordant genotypes with an average of 63x coverage in the 1:9 male-female mixture
224 whereas this value increased to 274x in the 9:1 male-female mixture.

225
226 To gauge samples loaded per run, LT provides guidelines for pooling samples to reach the
227 estimated minimum coverage for 95% of bases. In this study samples were pooled in a run to
228 aim for a minimum coverage between 42x to 286x for 95% of bases (Supplementary Table
229 S1). Information on minimum coverage per sample for 95% of bases is not included in TS
230 output. In the HID SNP panel the targeted 95% base minimum coverage thresholds were
231 only reached for all SNPs in 8 samples (all optimum input DNA). When accounting for outlier
232 SNPs, 31 optimum input DNA samples reach the desired minimum coverage threshold. From
233 the general coverage assessments made we infer that a targeted minimum coverage of at
234 least 62x for 95% of bases is necessary to accomplish a minimum coverage of 13x for all SNPs
235 in the panel, which is in agreement with minimum coverage threshold values for
236 concordance samples. For this reason, Run Lab3-B was omitted from further concordance
237 studies since none of the optimum input DNA samples reached this threshold. The heatmaps
238 in Fig. 2 outline differences between analyses by ranking cells with increasing coverage per
239 analysis (top to bottom, topmost analyses comprising mainly low-level DNA), and per SNP
240 (left to right). Although a similar SNP coverage pattern across samples is discernible, the
241 leftmost columns show more heterogeneity than average. In fact, further analysis shows
242 that per sample coverage distribution of all SNPs in the panel is not uniform across samples
243 (Supplementary File S1, Fig. S5). In conclusion, LT guidelines are useful for initial estimation
244 of sample numbers per chip and minimum coverage. However, the guidelines do not
245 function well when estimating minimum coverage for all HID SNPs in the panel, as well as
246 when considering low-level DNA samples. For this reason, it is important to adjust numbers

247 of samples loaded on each chip to a particular SNP set and to carefully gauge the quality and
248 quantity of DNA samples to be sequenced.

249

250 *3.2. Sequencing characteristics Ion PGM™ that impact forensic SNP genotyping*

251

252 Considering the sequence data in Genotyper output or obtained from this study's
253 comparisons amongst runs and laboratories, we focused on sequence coverage, base
254 misincorporation, allele read frequency (ARF) balance and strand bias, as factors impacting
255 the reliable differentiation of SNP heterozygotes from homozygotes. While artificial mixtures
256 can help assess how mixed DNA changes standard Ion PGM™ sequence data and creates
257 atypical patterns, it is important to assess the range of values observed in HID SNP
258 sequences with unmixed DNA. From the value ranges recorded, outlier SNPs were identified
259 which either should be removed from the HID SNP set or excluded from the data analysis
260 applied to more complex forensic analyses, including genotyping low-level and extremely
261 degraded DNA or detecting mixtures. The following results are outlined in detail in
262 Supplementary Table S2.

263

264 *3.2.1. Base misincorporation rates*

265

266 To gauge the overall rate of base misincorporation of Ion PGM™ (incorrect bases detected at
267 the SNP site in small proportions of sequence reads), the incidence of non-specific 3rd/4th
268 base incorporation (e.g. G and T in an A/C SNP) was compared to incidences of incorrect
269 alleles in homozygotes (e.g. very low occurrence of A bases in C homozygotes). If such rates
270 are comparable then a simple baseline rate of misincorporation can be established. If
271 different, then levels of extraneous target DNA detected by Ion PGM™, akin to allele drop-in,
272 can be estimated by how much more allelic misincorporation is seen. In either case, any
273 outlying SNPs with above-average misincorporation can be identified and appropriate
274 safeguards applied when detecting mixed DNA with minor components below 10%.
275 Supplementary Fig. S6 records frequencies of misincorporated bases in ranked order and
276 shows allelic and non-specific misincorporation rates were similar in nearly all SNPs and
277 hardly rose above 0.2% in all but 12/169 SNPs. Amongst the twelve SNPs with higher
278 misincorporation (on the right-hand side), only rs8078417, rs2399332, Y-rs2032597,
279 rs9866013 and rs1523537 reach 1% or more (column N, Supplementary Table S2). These
280 SNP's data should be discounted from assessments of imbalanced homozygote patterns in
281 mixtures, particularly rs2399332 and rs1523537 with disproportionately high allelic
282 misincorporation.

283

284 Although allelic and non-specific misincorporation are similar enough to largely discount
285 drop-in, Y-chromosome sequences were observed in female DNA. Supplementary Fig. S7
286 shows 34 Y-SNP nucleotide reads made in six analyses of two female samples. This data
287 represents male SNP target sequence in processed Genotyper output, but only 34 sequences

288 amongst >2 million female-specific sequences indicates extremely low levels of drop-in
289 genotypes from extraneous DNA for the Ion PGM™ system.

290

291 *3.2.2. Allele read frequency balance*

292

293 All forensic genotyping approaches must reliably differentiate imbalanced heterozygote
294 signals, created by stochastic effects in PCR, from the combined allele signals of mixed DNA.
295 This is particularly important for the 136 binary A-SNPs of the HID SNP set, as mixtures can
296 only be detected by measuring the signal of one allele against its alternative. Furthermore,
297 the Y-SNPs, chosen to help infer population divergent male phylogenies, are much more
298 restricted in detecting multiple genotypes (i.e. males from the same population are
299 minimally differentiated). We defined ARF settings that could equate to signal ratios
300 commonly observed in forensic markers and then assessed their effect on genotype calls.
301 Allele reads were reviewed from 38 analyses, comprising 169 SNPs in 28 male DNAs, 136 A-
302 SNPs in 10 female DNAs. Fig. 3A shows the distributions obtained from the ratio of reference
303 and total ARF values. A-SNP heterozygotes mostly showed good levels of clustering around
304 the 0.5 'perfect balance' midline. Homozygote data at the top and bottom is even more
305 regular in distribution, indicating ratios do not cross 0.1/0.9 thresholds.

306

307 Applying an 'aggressive' 45% allele balance thresholds, (i.e. a maximum 55:45 heterozygote
308 ratio) was assessed, but marked too many SNPs as imbalanced when in all other respects
309 their genotypes were concordant and reliable detected (see sections 3.3 and 3.4). A 40%
310 threshold (60:40 heterozygote ratio), indicated by the middle grey box over A-SNPs in Fig.
311 3A, gave better equilibrium between gaining the highest proportion of reliable genotypes
312 and balanced signals in optimum input DNA samples. Several SNPs with atypical ARF
313 distributions are evident in Fig. 3A and were identified from divergent average heterozygote
314 ARF values (column P, Supplementary Table S2, but rs1029047: cell P19, identified from out-
315 of-range values both sides of midline). SNPs rs2399332, rs1029047, rs8037428, rs430046
316 and rs1523537 were identified as poorly balanced ARF markers, in common with the analysis
317 of HID SNP performance by Børsting et al. [27]. Additionally, rs2107612 was poorly balanced
318 in our study, but not singled out by Børsting. Interestingly, SNPs rs10776839, rs4530059 and
319 rs1031825 found to be problematic by Børsting et al., gave reasonably balanced ARFs here,
320 although Fig. 3A indicates rs4530059 and rs1031825 have small proportions of genotypes
321 lying outside the threshold range.

322

323 Allele read frequency ratios also apply to homozygotes but in a different way. The presence
324 of other bases at a low proportion in the Ion PGM™ data arise from non-specific
325 incorporation, but the proportion of a second allele must exceed 10% for Genotyper to call
326 the genotype. For this reason, when ARFs reach $\geq 90\%$ samples cannot be mistyped as
327 heterozygotes (column P, Supplementary Table S2).

328

329 *3.2.3. Strand bias*

330

331 Ion PGM™ measures strand bias from forward strand SNP Target Reads divided by total SNP
332 Target Reads, indicating the ratio of sequencing in each direction. Arguably, sequence output
333 heavily biased towards one strand direction is less reliable, but we observed a large range of
334 strand bias from 0.5 (no discernible bias, equal sequencing of both strands) to values
335 occasionally close to one or zero (output exclusively from forward or reverse strand
336 respectively: columns Q-S, Supplementary Table S2). We set strand bias to 25%-75%:
337 equating to three-fold differences in output from each direction. The range of strand bias
338 values observed is summarised in Supplementary Fig. S8. Nine SNPs are marked at the plot
339 ends with average strand bias values outside the threshold set, three of these SNPs gave a
340 small proportion of genotype no-calls and this is discussed in more detail in section 3.4.2.

341

342 *3.3. Genotype concordance*

343

344 Genotype concordance was assessed in three ways: i. between replicate runs of the same
345 sample in each laboratory (inter-run concordance, 13 samples, 38 analyses); ii. between
346 laboratories running identical samples (inter-lab concordance, 6 samples, 24 analyses), and
347 iii. by comparing Ion PGM™ genotypes of Coriell cell-line control DNAs to those listed for HID
348 SNPs in 1000 Genomes and CG public databases. The individual concordance rate for each
349 sample is based on the number of called genotypes, to account for varying numbers of
350 replicates for different samples and varying numbers of no-call results (one or more runs
351 with NN calls for a SNP or ambiguous genotypes in project data). In the following section the
352 total values for no-call, concordance and discordance rates are given, whereas the individual
353 rates for each sample used for concordance comparisons are detailed in Supplementary
354 Tables S2 and S3.

355

356 *3.3.1. Inter-run and inter-lab concordance*

357

358 The no-call rate for inter-run samples was as low as 1.2% (70/6092) from eleven SNPs, while
359 99.8% of called genotypes were concordant in between runs of the same sample, with only
360 0.2% discordant genotypes (13/6022). Discordances were observed in rs2399332, rs1004357,
361 rs938283, rs1979255 and rs2032597 in six different samples. Possible explanations for the
362 discordances and no-calls are discussed in section 3.4.1 (column T, Supplementary Table S2).
363 In addition to the 38 analyses for inter-run concordance we observed a complete absence of
364 discordances and no-calls between library replicate analyses lab1_B and lab1_C. These
365 replicates correspond to Ion PGM™ libraries, prepared from the same original sample, but
366 processed separately in two distinct template preparations and sequencing runs.

367

368 Inter-lab concordance of called genotypes was 99.7% (3751/3763), with a no-call rate of
369 0.8% (29/3792). The same five SNPs as those from inter-run comparisons accounted for the
370 inter-lab discordances of 0.3% (12/3763) in five samples. No discordances were seen in
371 9947A analyses.

372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412

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

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

415

416 *3.4.1. Discordant SNPs*

417

418 Five SNPs were identified showing consistent patterns of discordant genotyping. Section 3.3
419 listed SNPs rs2032597 and rs2399332 as showing genotype differences between replicates in
420 more than one sample and it is notable that they share the characteristic of having closely
421 sited polymeric tracts around the target SNP site.

422

423 The A/C Y-SNP rs2032597 gave a non-allelic T base in ~20% of male analyses. As shown in the
424 IGV graphics (Supplementary File S2-SNP 1), the base immediately upstream of the SNP
425 position is C (the anchor base). An rs2032597-C genotype leads to a large proportion of
426 misaligned reads in both sequencing directions, with a false C insertion being generated and
427 the SNP's C allele becoming the anchor base. This displaces the downstream poly-T tract one
428 base into the SNP position and as it is hemizygous, when the number of T reads exceeds the
429 minimum ARF, the genotype is called T instead of C.

430

431 The G/T A-SNP rs2399332 is sited within a poly-T tract. Examination in IGV showed many G
432 reads had an extra T in the poly-T tract downstream of the SNP position. This caused
433 misalignments and the G allele was considered an insertion, incorrectly placing a T in the
434 SNP position. As this usually happens at a frequency <10%, Genotyper correctly reports GG
435 for most samples, but discordant genotypes can occur when the T frequency exceeds the
436 10% threshold. This phenomenon explains the above-average allelic misincorporation rate of
437 rs2399332 (Supplementary Fig. S6) as well as the single discordant genotype observed (Table
438 2). Furthermore, rs2399332 shows a clear deviation from expected ARF ratios in
439 heterozygotes (Fig. 3A and cell P12, Supplementary Table S2). Those samples can reach ARF
440 imbalances of 0.2:0.8 (20% of sequences G), however IGV shows these are not caused by
441 misalignment from the poly-T tract. For this reason, context sequence was scrutinised for
442 possible primer binding site polymorphisms that could hinder production of sequences
443 carrying the G allele. Several SNPs were found in the region encompassing the amplicon plus
444 30 bp upstream/downstream of the amplicon ends. In particular SNP rs2399333 is very likely
445 to be in the forward primer-binding site as it is located ~10 bp within the inferred 5'-
446 amplicon end. Furthermore, if the reverse primer is long enough, rs9866331 could also
447 interfere with balanced PCR of each allele as it is ~25 bp within the inferred 3'-amplicon end.
448 Depending on the PCR efficiency and the degree to which neighbour SNPs affect primer
449 binding, the rs9866331-G ARF may drop to ≤10%, causing heterozygotes to be reported as
450 homozygous T genotypes, as seen in discordant S5 replicates.

451

452 The remaining three SNPs had discordant genotypes in 1-2 analyses of single samples. In
453 rs1979255 and rs1004357, heterozygotes had balanced ARFs in all but the single discordant
454 sample. The third SNP rs938283 showed balanced heterozygote allele distributions including

455 the discordant sample. IGV context sequence analysis failed to indicate distinct features that
456 could create misalignments and produce mistyping in the samples analysed (row 16,
457 Supplementary Table S2).

458

459 3.4.2. SNPs with no-calls

460

461 Genotyper reports no-calls when SNPs fail to fulfil Germline analysis parameter settings, but
462 additionally dropouts were observed, defined here as SNPs with nil sequence output
463 (QUAL=0). Fig. 5 summarises total SNPs with no-calls or dropouts in 74 analyses (mixtures
464 and lab3-B run excluded). In the 38 concordance analyses, no-calls were recorded in nine
465 SNPs. First, rs5746846, rs576261, and rs13182883 had insufficient coverage in one strand
466 (Supplementary Fig. S8). In these SNPs sequencing is initiated on both strands but one fails
467 to reach the SNP position, illustrated by the IGV overview of rs13182883 (Supplementary File
468 S2-SNP 2) with 0.994 strand bias. This phenomenon produces the very strong strand bias
469 deviations shown at the ends of the distribution plot of Supplementary Fig. S8 and remains
470 unexplained from all analyses made in IGV.

471

472 Second, rs13447352 and rs1336071 consistently showed low numbers of sequence reads;
473 failing to reach minimum values for both strands and total coverage. The same observation
474 was made for SNPs rs2032599, rs2107612 and rs1478829, but only in single analyses.
475 Notably, rs1478829 had zero reads in one analysis.

476

477 Lastly, as well as the coverage-related analysis parameters and sequence quality thresholds
478 detailed in sections 3.1 and 3.2, analysis parameter settings: *VCF minimum quality*
479 (*min_variant_score=10*) plus *maximum common signal shift* (*filter_unusual_predictions=0.3*)
480 affected genotype reporting and occasionally caused no-calls, the latter most strongly in
481 rs1029047. Comprehensive review of rs1029047 data in IGV revealed uncertainty about
482 Genotyper heterozygote calls, even when all replicates were concordant (Supplementary File
483 S2-SNP 3). This A/T SNP lies between poly-T tract and long poly-A tracts plus several indels,
484 highly likely to produce systematic misalignments. This same SNP was identified as poorly
485 performing by Børsting et al. [27], while Budowle et al. also reported discordant genotypes
486 [30, 31].

487

488 3.4.3. SNPs with mean analysis parameter values deviating from defined thresholds

489

490 Despite an absence of genotyping problems affecting the eleven SNPs of this third category
491 (Fig. 4), examination of their mean values showed consistent atypical behaviour with respect
492 to the analysis parameter thresholds we defined, particularly sequence coverage and strand
493 bias (columns O, Q, R, S in Supplementary Table S2). IGV files from all analyses of the eleven
494 SNPs were scrutinised, but failed to indicate sequence problems. An example is rs430046
495 that, despite strong strand bias and a high frequency of base deletion calls at the target site,
496 gave consistent genotypes across all replicates (typical IGV data in Supplementary File S2-

497 SNP 4). There is no strong reason to doubt SNP genotype calls predominantly based on
498 sequences in one direction, despite an increased rate of no-calls observed in such markers.
499

500 *3.5. Assessments of Ion PGM™ sensitivity*

501
502 Assessing sequence data from input DNA well below recommended quantities, the Ion
503 PGM™ system is evidently a very sensitive SNP detection system. Levels of SNP data
504 completeness in low-level DNA analyses are indicated by dark grey columns in Fig. 5,
505 counting SNPs with no-calls and dropouts. At 100-50-25 pg inputs, SNPs generally show
506 more no-calls/dropouts than optimum input DNA, although runs lab1-E and –F maintain
507 good genotyping performance at these lowest inputs. Only rs2016276 appeared
508 disproportionately amongst failing markers in 100-50-25 pg dilutions, giving 6/23 male and
509 6/13 female no-calls. Although concordance study DNAs mainly had missing genotypes in
510 only 1-3 SNPs, low-level DNA rarely exceeded 8-12 SNPs with missing genotypes.
511 Furthermore, this has little impact on random match probability (RMP) values.
512 Supplementary Fig. S11 indicates approximately 40-50% of missing data (including outlier
513 SNPs) is needed to decrease the cumulative RMP to a value similar to GlobalFiler. Half or less
514 of outlier SNPs (using each category defined in section 3.4) had missing genotypes in aDNA
515 and lab1-A runs. Five extra library amplification cycles did not increase sensitivity.
516

517 The highly degraded aDNA sample gave more SNP failures than most dilution series analyses.
518 Although this is limited initial NGS data, these results indicate very high sensitivity for Ion
519 PGM™ when target sequences are highly degraded or inhibited. Therefore, although good
520 sensitivity to low-level DNA has been recognised in this and other studies [27,30], specific
521 effects of aggressive degradation need to be comprehensively assessed to properly test the
522 effectiveness of NGS analysing skeletal remains typical of missing person identification.
523

524 Supplementary Table S4 details sequence data from two analyses of aDNA sample S7.
525 Although these gave relatively low levels of SNP Target Reads and the lowest mean read
526 lengths of any samples (data not shown), genotypes had very good levels of agreement. In
527 all, 128/169 genotype pairs were called identically (75.7%) and a further 23 genotypes called
528 from one analysis (totalling 89.3% genotypes). More no-calls and dropouts (QUAL=0) were
529 recorded applying library re-amplification. The 25-cycle PCR gave 10 no-calls, 4 dropouts,
530 whereas 25 + 5 cycles gave 18 no-calls, 13 dropouts (6 no-calls, 4 dropouts in common).
531 Unmodified PCR also achieved higher average sequence coverage and quality scores: 128
532 sequences and QUAL=422.7, compared to 72 sequences and QUAL=286.5 in 25 + 5 analysis,
533 plus just 1/23 singleton genotypes.
534

535 The slight rise in numbers of common genotypes to ‘common results’ (same SNPs giving
536 genotypes *or* no-calls/dropouts in both analyses) from 75.7% to 81.7%, suggests some locus
537 dropout in Ion PGM™ SNP genotyping may be systematic rather than random, but many
538 more highly degraded DNA samples must be assessed to test this assumption. Despite

539 lacking reference genotypes, the aDNA heterozygosity of 51% compares to an expected 46%
540 heterozygosity for these SNPs (1000 Genomes CEU data), suggesting very little allele dropout.

541

542 *3.6. Mixture analysis*

543

544 Detection of mixed source DNA and possible identification of components in simple mixtures
545 is challenging when genotyping binary SNPs with the commonly used SNaPshot® system. In
546 contrast, NGS data from this study of Ion PGM™ and AmpliSeq™ technology gave balanced
547 heterozygous genotypes, providing a more secure basis for analysing mixtures. It is
548 important to reliably recognise SNP data as originating from a mixture and not a single
549 profile. Furthermore, development of enhanced statistical analyses, prompted by our results
550 from Ion PGM™ runs, will allow likelihood ratio calculations when one of the component
551 DNAs is known. For these reasons, our assessment of NGS data from artificial mixtures was
552 more comprehensive than for the other DNAs. Detailed descriptions of these mixed
553 sequence data analyses are given in Supplementary File S3.

554

555 Scrutiny of the single 1:1 mixture ARF plot of Fig. 3B and all other ARF plots in
556 Supplementary File S3, Fig. S9, shows mixtures generally have patterns quite distinct from
557 unmixed samples, with more heterozygous SNPs outside the 40-60% ARF range. Additionally,
558 increased heterozygosity and reduced Y-SNP coverage provide clear indications of the
559 presence of mixed DNA in HID SNP data (Supplementary File S3, Table S5). Our initial
560 analyses of limited numbers of mixtures indicate Germline analysis parameter settings
561 should be used for forensic samples of unknown origin. If any of the described mixture
562 indicators is found, data should then be re-analysed with Somatic settings to improve
563 accuracy of A-SNP genotyping. Even with this two-tier approach, care is needed with more
564 extreme mixture ratios (here, 1:9 and 9:1), as there is increased probability minor alleles
565 escape detection. Y-SNPs should be analysed independently with Germline analysis
566 parameter settings as this guarantees higher genotyping rates while maintaining allele call
567 quality.

568

569 *3.7. Context sequence examinations with IGV*

570

571 To further assess HID SNPs for forensic analysis, the context sequence of each marker was
572 scrutinised using IGV [28,29]. This provided checks on characteristics that could influence
573 alignment, including Indels or polymeric tracts, but also screened for extra polymorphisms
574 close to target sites. In staff donors, we detected clustering polymorphisms associated with
575 target SNPs. Table 3 summarises data for these additional polymorphisms. In SNP rs430046
576 there are three well-characterised and closely sited SNPs adding discrimination power (all
577 variant allele homozygotes in Supplementary File S2-SNP 4). Variants at sequence extremes
578 and next to polymeric tracts tended to produce unreliable reads (see rs1109037 in
579 Supplementary File S2-SNP 5).

580

581 In contrast to SNPs close to target sites, Indel discovery and genotyping with Ion PGM™
582 sequence data remains more restricted. Small sequencing errors, usually linked to short
583 polymeric tracts of four or more bases, tend to produce artefact Indels at high frequency.
584 Mostly deletions were observed in such cases, but insertions occasionally occur in
585 misaligned polymeric tracts. Two other observations made from IGV sequences are worth
586 noting. First, Indel artefacts are affected by sequence directionality and tend to occur
587 exclusively on one strand, aiding the differentiation of true from artefact Indels (rs430046 in
588 Supplementary File S2-SNP 4, shows 12 direction-dependent Indel calls). Second, false Indels
589 can be generated from misaligned sequences containing repetitive motifs, although handling
590 of short tandem repeat alignments is being refined and such artefacts will be better
591 controlled as sequence analysis software improves.

592 **4. Concluding remarks**

593

594 The evaluation of Ion PGM™ sensitivity and genotyping accuracy made here, give strong
595 support for the application of NGS technology to forensic DNA analysis. Sequence data
596 obtained in all three laboratories had sufficiently high coverage and gave reliable SNP
597 genotyping for most loci in HID SNP. We discovered five SNPs with discordances that should
598 be excluded from the panel. We note rs1004357 and rs2032597 are already removed from
599 the revised version of HID SNP, while rs2399332 was identified in Børsting's study as a
600 problematic SNP [27]. We also found discordant genotypes in rs1979255 and rs938283, and
601 their continued inclusion in HID SNP needs critical review. Furthermore, rs2107612 showed
602 imbalanced heterozygote reads and should also be removed from the panel in addition to
603 the eight problematic markers identified by Børsting. Lastly, mention should be made of
604 rs1029047, which gives genotyping inconsistencies in all NGS studies of this SNP made so far
605 [27,30,31]. There are clearly characterised context sequence factors affecting the alignment
606 and therefore the reliability of allele calls for rs1029047 (Supplementary File S2-SNP 3),
607 which have not affected SNaPshot genotyping of this SNP [6]. Therefore, careful scrutiny of
608 sequence characteristics is required of any SNP chosen for forensic use. This is particularly
609 important for coding SNPs in forensic phenotype predictive tests, since these must work well
610 for the SNP analyses to be sufficiently informative.

611

612 The estimation of optimum sample numbers for each of the six Ion PGM™ chip versions,
613 presented this study with the biggest challenge, both in harmonising NGS runs across three
614 laboratories and ensuring the coverage obtained was appropriate for assessing forensic
615 sensitivity. Since low-level DNA appears to accentuate coverage variability in HID SNP
616 markers, this will be a major problem when initially optimising NGS for routine forensic use.
617 As Ion PGM™ chip capacities have now reached very reasonable levels of sequence output,
618 users can be cautious by loading fewer samples than coverage estimation guidelines suggest.
619 Furthermore, there is some consensus that ~15-20x minimum coverage thresholds can
620 safeguard the reliability of allele calls made with NGS [2,21,26].

621

622 Although the Torrent Suite™ software provides several sequence quality parameters in the
623 data output, we found there was little or no scope for changing the default analysis
624 parameters settings to more aggressive thresholds. Setting such thresholds would provide a
625 way to exclude miscalled genotypes from under-performing SNPs or mixed DNA. This finding
626 has consequences for the average forensic scientist's capacity to properly scrutinise the
627 extensive data that Ion PGM™ produces. Since mixture detection with binary markers is
628 severely restricted compared to multi-allele STRs, it is all the more important to properly
629 assess deviations from balanced heterozygote patterns. We largely agree with the
630 conclusions of Børsting et al. [27], that the Ion PGM™ analysis software needs further
631 optimisation to be fully suitable for forensic application, although it is being constantly
632 revised to this end. In particular, there is an evident need to apply Somatic analysis
633 parameter settings to properly analyse mixtures, even though Germline analysis parameter

634 settings are set in place for forensic SNP analysis with Ion PGM™. This reduces the capacity
635 of the system to alert the analyst to mixtures and represents a critical shortfall when the
636 very high sensitivity of Ion PGM™ is borne in mind.

637

638 **Acknowledgements** This work was funded by the European Union Seventh Framework
639 Program (FP7/2007–2013) under grant agreement no. 285487 (EUROFORGEN-NoE) and the
640 Austrian Science Fund (FWF) [P22880-B12]. CS is supported by funding awarded by the
641 Portuguese Foundation for Science and Technology (FCT) and co-financed by the European
642 Social Fund (Human Potential Thematic Operational Program SFRH/BD/75627/2010). MdIP is
643 supported by funding awarded by the Consellería de Cultura, Educación e Ordenación
644 Universitaria of the Xunta de Galicia as part of the Plan Galego de Investigación, Innovación e
645 Crecemento 2011-2015 (Plan I2C). The authors wish to thank Jorge Amigo, Grupo de
646 Medicina Xenómica (GMX), University of Santiago de Compostela and Matt Phipps of Life
647 Technologies, for their helpful guidance with NGS data analysis.

648 **References**

- 649 [1] B. Merriman, Ion Torrent R&D Team, J.M. Rothberg, Progress in Ion Torrent
650 semiconductor chip based sequencing, *Electrophoresis* 33 (2012) 3397–3417.
- 651 [2] M.A. Quail, M. Smith, P. Coupland, T.D. Otto, S.R. Harris, T.R. Connor, A. Bertoni, P.
652 Harold, H.P. Swerdlow, Y. Gu, A tale of three next generation sequencing platforms:
653 comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers, *BMC*
654 *Genomics* 24 (2012) 341-353.
- 655 [3] N.J. Loman, R.V. Misra, T.J. Dallman, C. Constantinidou, S.E. Gharbia, J. Wain, M.J. Pallen,
656 Performance comparison of benchtop high-throughput sequencing platforms, *Nat.*
657 *Biotechnol.* 30 (2012) 434-439.
- 658 [4] W. Parson, C. Strobl, G. Huber, B. Zimmermann, S.M. Gomes, L. Souto, L. Fendt, R.
659 Delpont, R. Langit, S. Wootton, et al., Evaluation of next generation mtGenome sequencing
660 using the ion Torrent Personal Genome Machine (PGM™), *Forensic Sci. Int. Genet.* 7 (2013)
661 543–549.
- 662 [5] B. Budowle, D.H. Warshauer, S.B. Seo, J.L. King, C. Davis, B. LaRue, Deep Sequencing
663 Provides Comprehensive Multiplex Capabilities, *Forensic Sci. Int. Genet. Supp. Series 4*
664 (2013) e334-e335.
- 665 [6] J.J. Sanchez, C. Phillips, C. Børsting, K. Balogh, M. Bogus, M. Fondevila, C.D. Harrison, E.
666 Musgrave-Brown, A. Salas, D. Syndercombe-Court, et al., A multiplex assay with 52 single
667 nucleotide polymorphisms for human identification, *Electrophoresis* 27 (2006) 1713–1724.
- 668 [7] A.J. Pakstis, W.C. Speed, R. Fang, F.C. Hyland, M.R. Furtado, J.R. Kidd, K.K. Kidd, SNPs for a
669 universal individual identification panel, *Hum. Genet.* 127 (2010) 315– 324.
- 670 [8] T.M. Karafet, F.L. Mendez, M.B. Meilerman, P.A. Underhill, S.L. Zegura, M.F. Hammer,
671 New binary polymorphisms reshape and increase resolution of the human Y chromosomal
672 haplogroup tree, *Genome Res.* 18 (2008) 830–838.
- 673 [9] The 1000 Genomes Project Consortium, An integrated map of genetic variation from 1,092
674 human genomes, *Nature* 491 (2012) 56-65.
- 675 [10] R. Drmanac, A.B. Sparks, M.J. Callow, A.L. Halpern, N.L. Burns, B.G. Kermani, P.
676 Carnevali, I. Nazarenko, G.B. Nilsen, G. Yeung, et al., Human genome sequencing using
677 unchained base reads on self-assembling DNA nanoarrays, *Science* 327 (2009) 78-81.
- 678 [11] Thermo Fisher Scientific, Life Technologies: Sequencing technology solutions. (2014)
679 Accessed June 2014. Available from: [https://www.lifetechnologies.com/au/en/home/life-](https://www.lifetechnologies.com/au/en/home/life-science/sequencing/sequencing-technology-solutions.html)
680 [science/sequencing/sequencing-technology-solutions.html](https://www.lifetechnologies.com/au/en/home/life-science/sequencing/sequencing-technology-solutions.html)
- 681 [12] Thermo Fisher Scientific, Life Technologies: Ion PGM™ system for next-generation
682 sequencing. (2014) Accessed June 2014. Available from:
683 [https://www.lifetechnologies.com/au/en/home/life-science/sequencing/next-generation-](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)
684 [sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-](https://www.lifetechnologies.com/au/en/home/life-science/sequencing/next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-run-sequence/ion-PGM™-system-for-next-generation-sequencing.html)
685 [sequencing-run-sequence/ion-PGM™-system-for-next-generation-sequencing.html](https://www.lifetechnologies.com/au/en/home/life-science/sequencing/next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-run-sequence/ion-PGM™-system-for-next-generation-sequencing.html)
- 686 [13] Thermo Fisher Scientific, Life Technologies: Ion AmpliSeq™ library preparation user
687 guide. July 2013.
- 688 [14] Thermo Fisher Scientific, Life Technologies: Ion OneTouch™ 200 Template Kit v2 user
689 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.

737 **Figure legends**

738

739 **Fig. 1.**

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

743 **(B)** Deviation from expected maximum SNP Target Reads adjusted for a wide range of
744 chip types, sample numbers per chip and DNA quality amongst the runs made. The deviation
745 metric is calculated as: $[(\text{SNP Target Reads} - \text{Expected SNP Target Reads}) / \text{Sum of Clonal}$
746 $\text{Reads per Chip}]$, where: Clonal Reads=number of reads passing the polyclonal filter;
747 Expected SNP Target Reads=Clonal Reads/number of samples.

748 **(C)** Summary bar plots of mean standard deviation of SNP Target Reads to expected SNP
749 Target Reads per run. Generally, runs combining optimum input and low-level DNA samples
750 show higher variation from expected SNP Target Reads than runs with optimum input DNA
751 only.

752

753 **Fig. 2.**

754 Analysis vs. SNP heatmap arranged as: increasing mean sample coverage levels top to
755 bottom, increasing mean SNP coverage levels left to right. Left map shows Y-SNPs and for
756 brevity, blue run identifiers are as detailed in Fig. 1A.

757

758 **Fig. 3.**

759 **(A)** ARF balance in 169 SNPs (listed in Genotyper order, Y-SNPs rightmost) with this study's
760 analysis parameter thresholds marked with grey boxes denoting reference/total ARF ratios
761 of: 0-0.1 and 0.9-1 for A-SNP homozygotes/Y-SNP hemizygotes and 0.4-0.6 for A-SNP
762 heterozygotes. The marked outlier SNPs were identified by recording average ARF ratios
763 (solid lines) or for rs10129047 by visual inspection, as values positioned each side of midline
764 affect the average. Outlier SNPs identified from the study of the same HID SNPs by Børsting
765 et al. [27] are marked for comparison.

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

769

770 **Fig. 4.**

771 Schematic representation of the proportion of HID SNPs with good performance, poor
772 performance or outlier characteristics. Markers listed left were identified as: five SNPs with
773 genotype discordances; nine concordant SNPs with no-calls; eleven concordant SNPs
774 showing deviation from analysis parameter thresholds defined in this study. Underlined
775 SNPs are still retained in the HID SNP set, to the best of the authors' knowledge. Italic SNPs
776 show 5/8 markers recommended for removal by Børsting's study of the same SNP panel [27].
777 Another three SNPs identified by Børsting: rs10776839; rs4530059 and rs1031825 did not
778 show problematic characteristics in our study.

779

780 **Fig. 5.**

781 Numbers of SNPs showing no-calls (sequence quality outlying analysis parameter thresholds)
782 or dropouts (QUAL=0) in concordance study or low-level DNA analyses (marked by horizontal
783 bars for each dilution series or for aDNA S7).

784

785 -----

786

787 **Supplementary Files**

788

789 **Supplementary File S1**

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

791

792 **Supplementary File S2**

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

794

795 **Supplementary File S3**

796 Mixture analysis with the Ion PGM™

797

798 **Supplementary Figures**

799

800 **Supplementary Fig. S1.**

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

805

806 **Supplementary Fig. S2.**

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

811

812 **Supplementary Fig. S3.**

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

816

817 **Supplementary Fig. S4.**

818 Comparison of primer regions reads for rs1005533 in a negative control, low-level and
819 optimum input DNA sample, from IGV graphical summaries.

820 **(A)** Negative control shows short reads in the primer region of targeted rs1005533.

821 (B) Similar reads can be seen in a low-level DNA sample.
822 (C) The optimum input DNA sample does not show any short reads in the target
823 neighbouring region. For better visualization reads are down-sampled to 100. Pink
824 sequences are forward direction, violet reverse.

825

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

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

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

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

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

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

839

840 **Supplementary Fig. S6.**

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

844

845 **Supplementary Fig. S7.**

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

849

850 **Supplementary Fig. S8.**

851 Distribution of strand bias (forward strand SNP Target Reads / total SNP Target Reads) for
852 136 autosomal HID SNPs. The midline represents no discernible strand bias and dotted lines
853 the 25%-75% value range used to identify nine SNP outliers with mean strand bias values
854 outside this range (extreme values marked by boxes). SNPs in bold gave several no-calls and
855 are discussed in section 3.4.2 and the IGV overview of rs13182883 is given in Supplementary
856 File S2-SNP 2.

857

858 **Supplementary Fig. S9.**

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

861

862 **Supplementary Fig. S10.**

863 Observed and expected ratios of average Y-SNP coverage vs. average A-SNP coverage for the
864 male component S5 and mixtures.

865

866 **Supplementary Fig. S11.**

867 Reduction in cumulative RMP with increasing no-call rate.

868

869 **Supplementary Tables**

870

871 **Supplementary Table S1.**

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

874

875 **Supplementary Table S2.**

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

879

880 **Supplementary Table S3.**

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

885

886 **Supplementary Table S4.**

887 Genotypes for two different analyses of the aDNA sample S7.

888

889 **Supplementary Table S5.**

890 A) Proportions of homozygous, heterozygous and no-calls for mixed DNA components S5
891 and S6 and for the expected genotype mixtures. Counts and percentages only considered
892 136 A-SNPs. B) Amongst the expected mixtures heterozygous SNPs were divided into: i)
893 balanced – same numbers of each allele; ii) imbalanced – a higher number of one allele over
894 the other (depending on donor genotypes and mixture ratio); and iii) undetermined – when
895 missing genotypes in donor samples means the numbers of each allele cannot be
896 determined.

897
898
899
900

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	

901
902
903

x Same sample re-amplified

Δ Library replicates

904
905
906

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

SNP ID	Coriell cell-line control DNA No.	Ion PGM™ genotype	CG genotype	1000 Genomes-Phase I genotype	1000 Genomes-Phase III genotype	Comments on discordance
Y-rs2032597	NA06994	T	C	(no Y data)	(no Y data)	See sections 3.3.2, 3.4.1
Y-rs2032597	NA07029	T	C	(no Y data)	(no Y data)	See sections 3.3.2, 3.4.1
rs2399332	NA18498	<i>TT</i>	GT	GT	GT	See sections 3.3.2, 3.4.1
rs2342747	NA07000	AG	<i>NN</i>	AG	AG	no call on either allele in CG
rs4288409	NA18498	AC	<i>NN</i>	AC	AC	no call on either allele in CG
rs4847034	NA07000	GG	<i>NG</i>	GG	GG	no call for 1st allele in CG
rs4847034	NA07029	GG	<i>GN</i>	GG	GG	no call for 2nd allele in CG
rs8078417	HG00403	TT	TT	<i>CT</i>	TT	Likely 1000 Genomes-Phase I error
rs10768550	NA18498	CT	CT	<i>CC</i>	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

907

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](#)

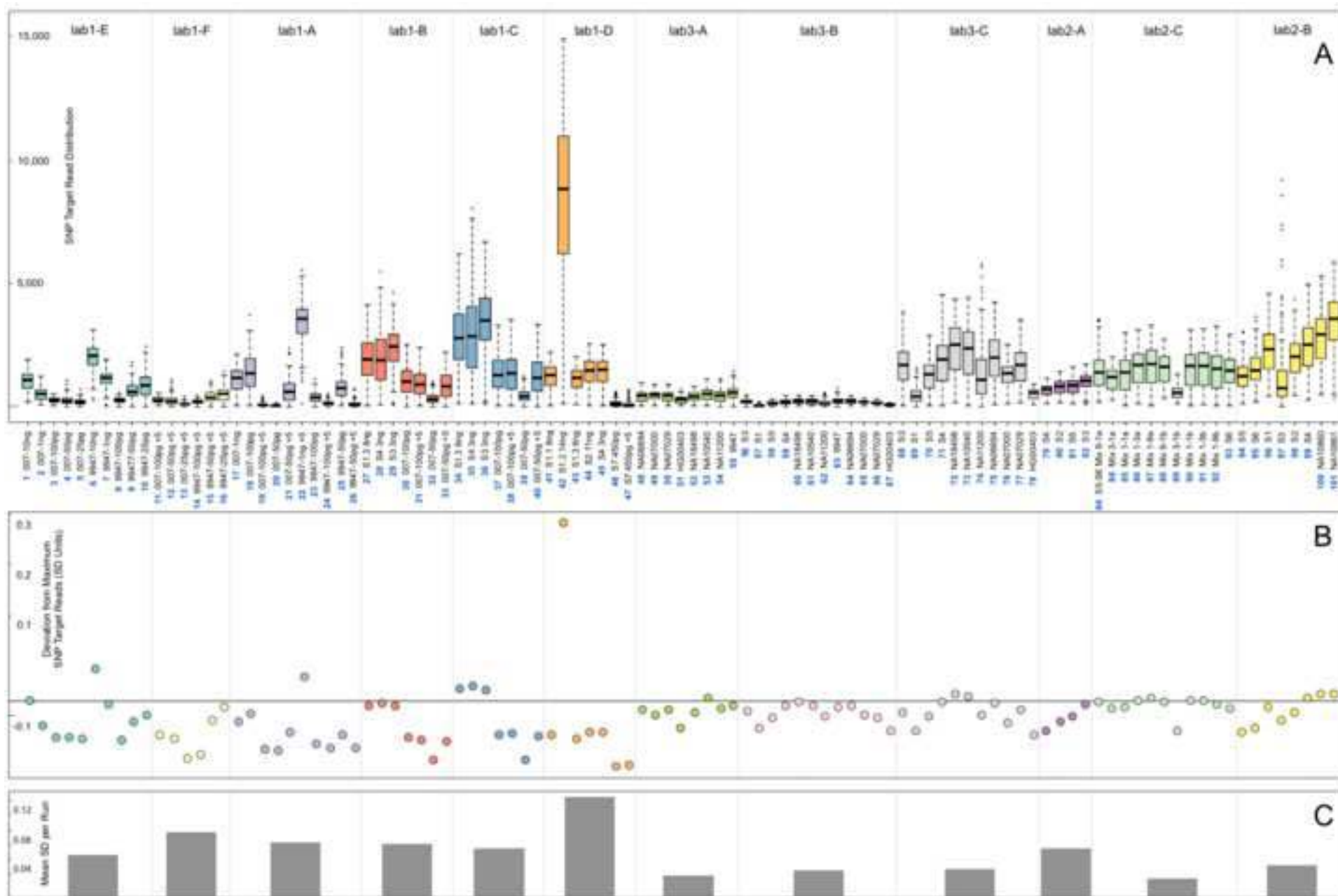


Figure 2
[Click here to download high resolution image](#)

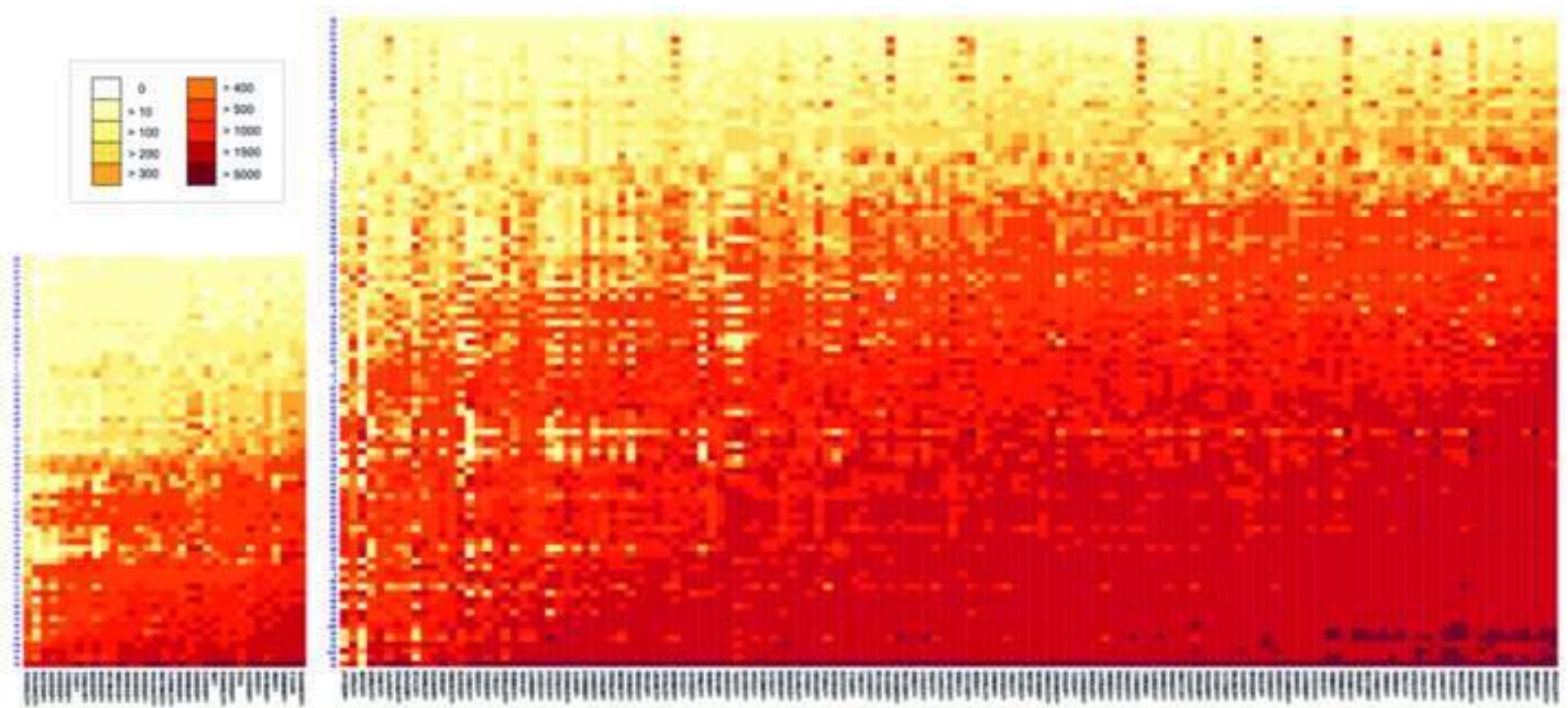


Figure 3
[Click here to download high resolution image](#)

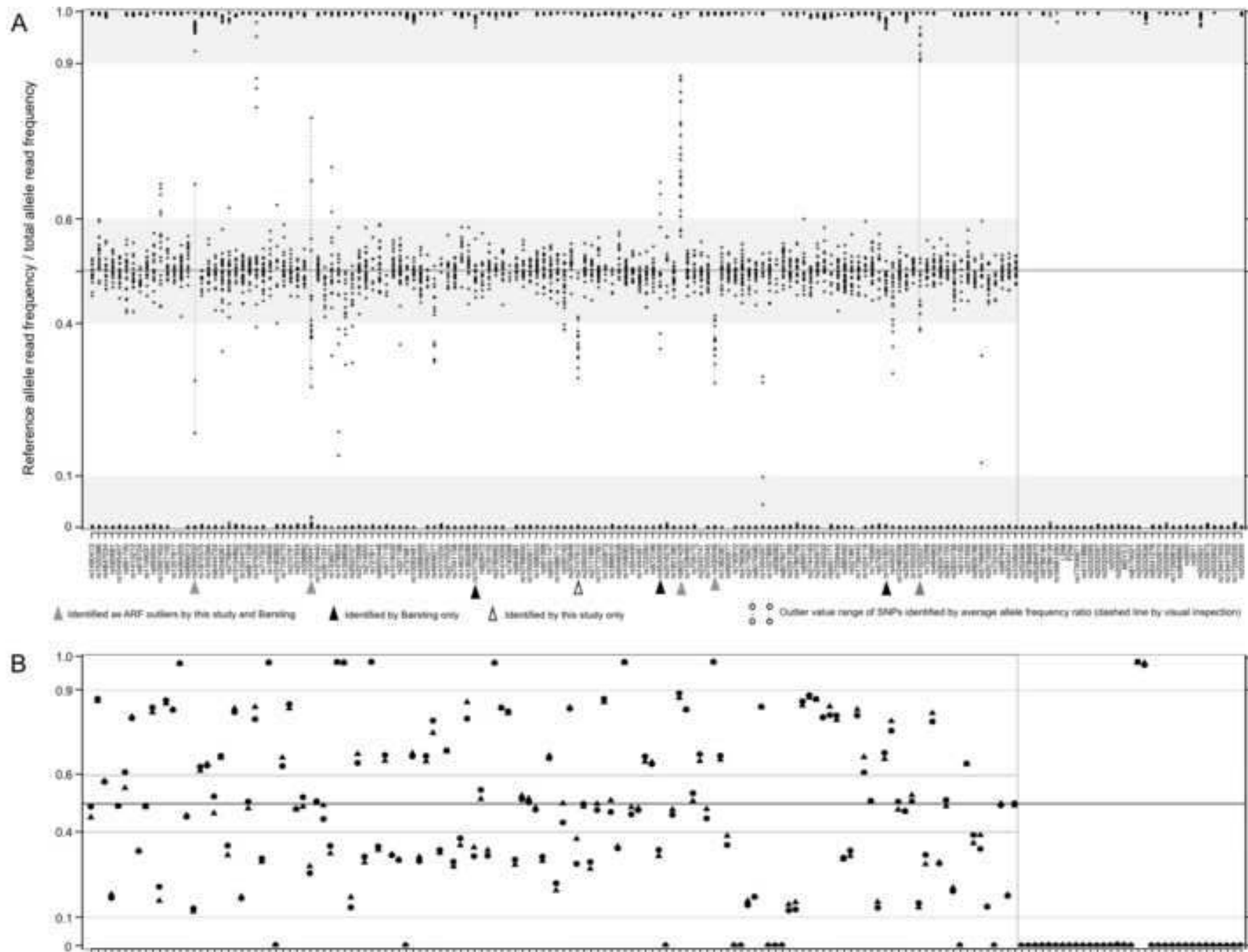


Figure 4
[Click here to download high resolution image](#)

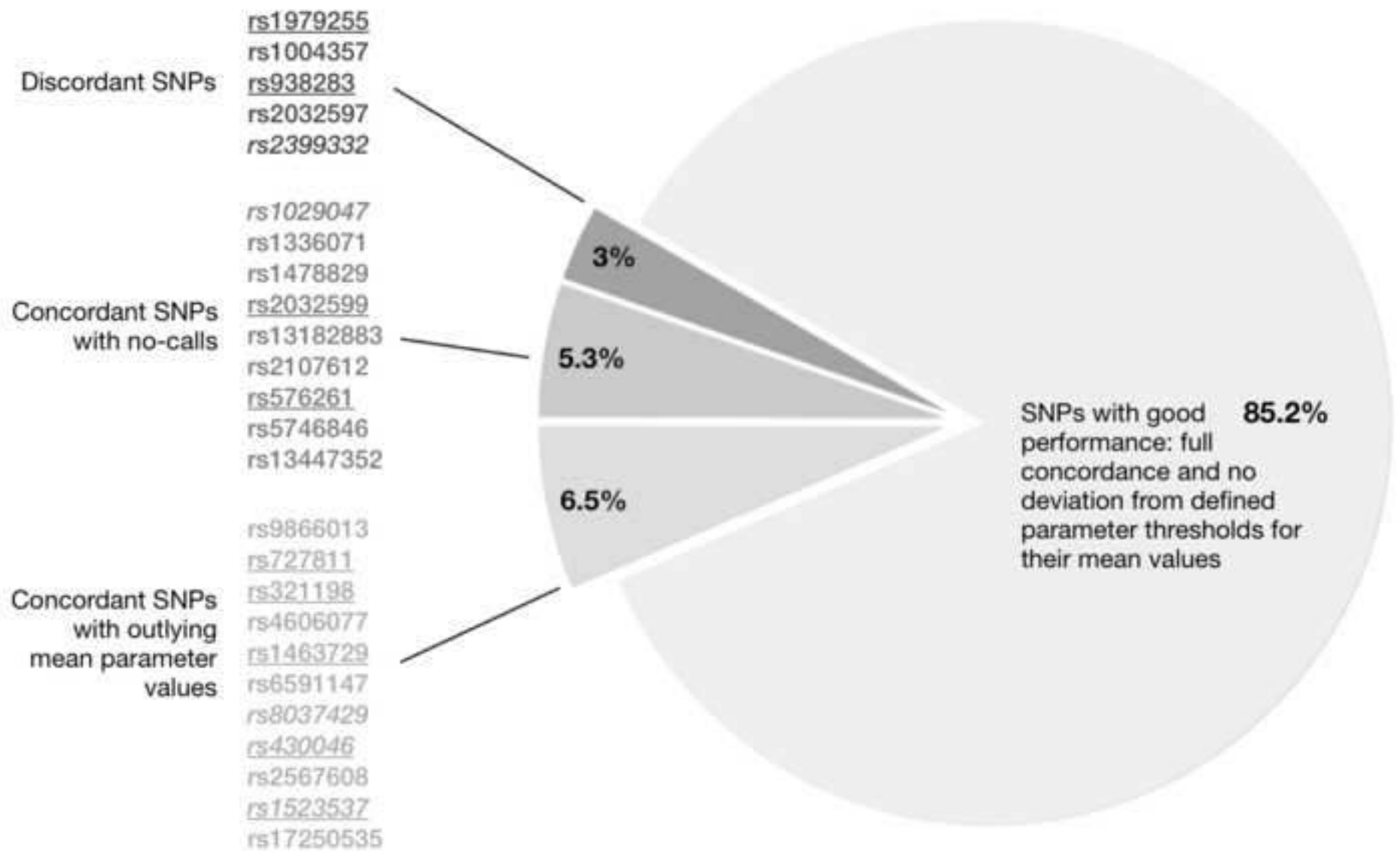
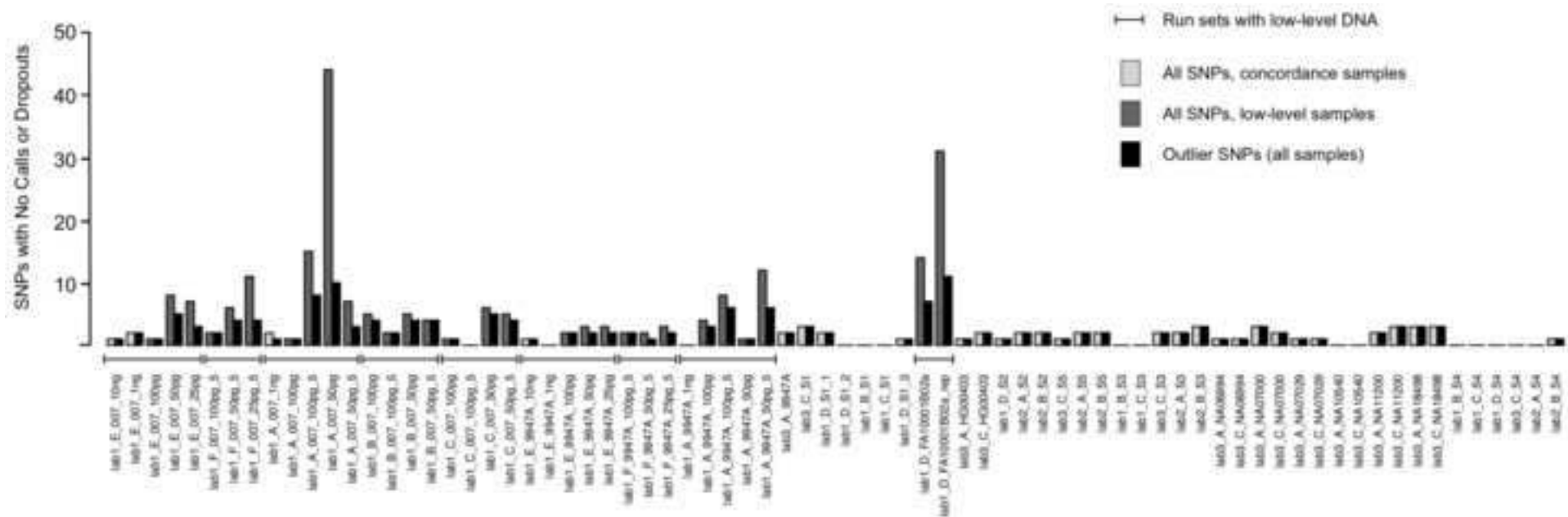


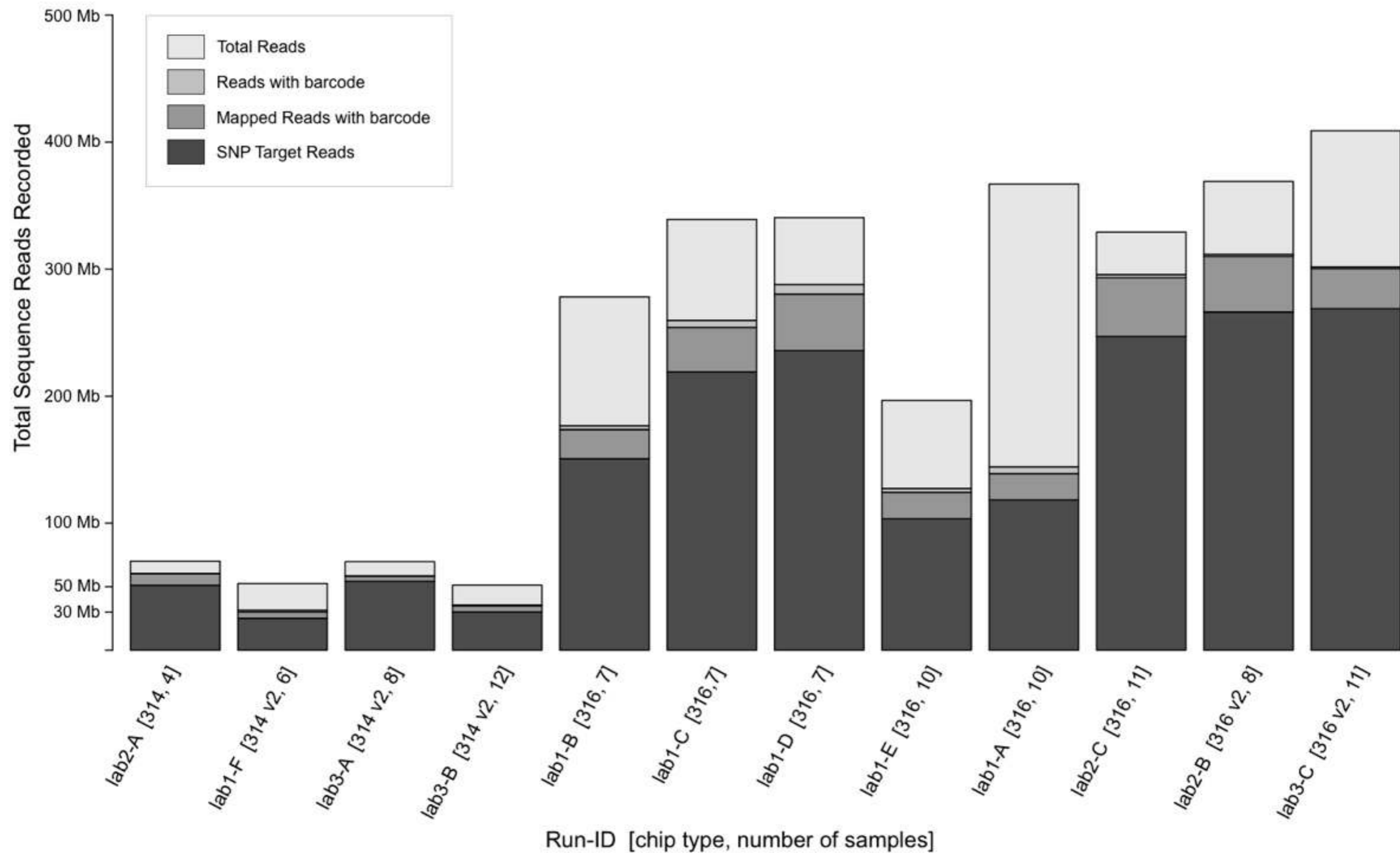
Figure 5
[Click here to download high resolution image](#)



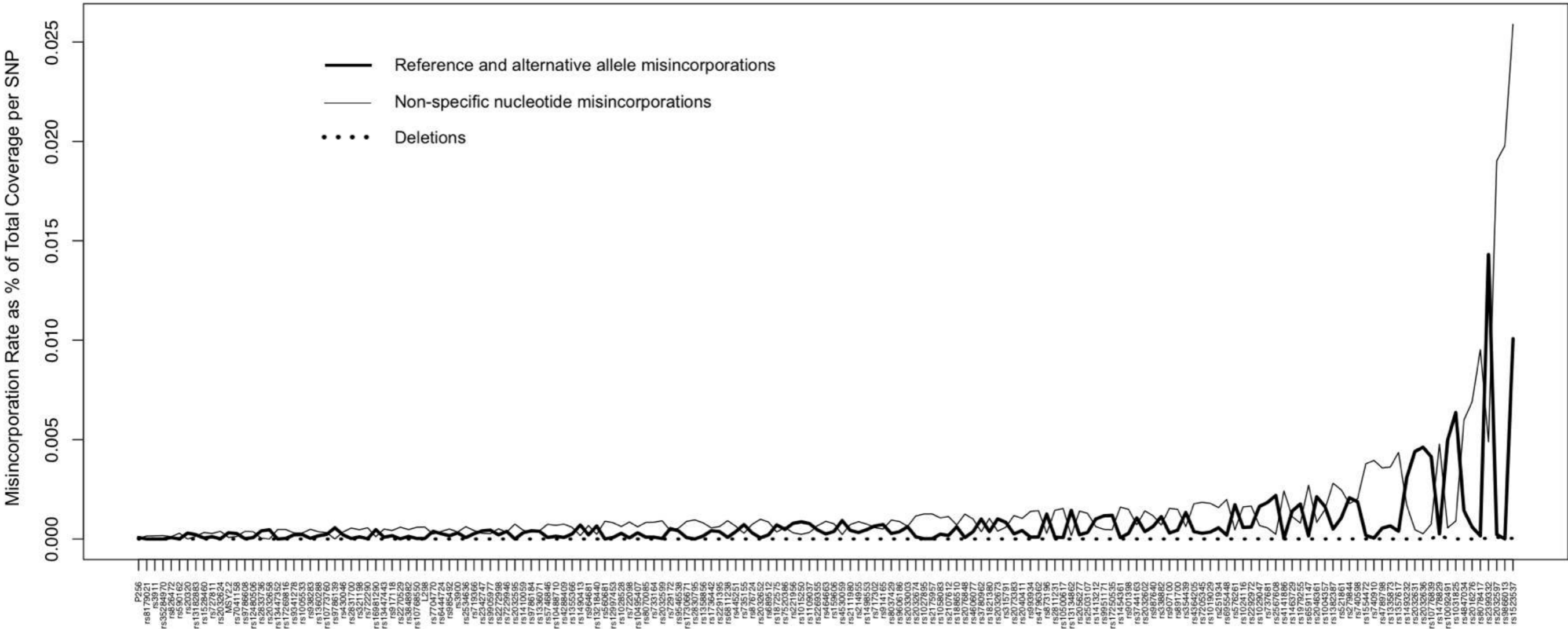
Supplementary Table S1: Expected sequence throughput of Ion PGM™ based on chosen sample numbe

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

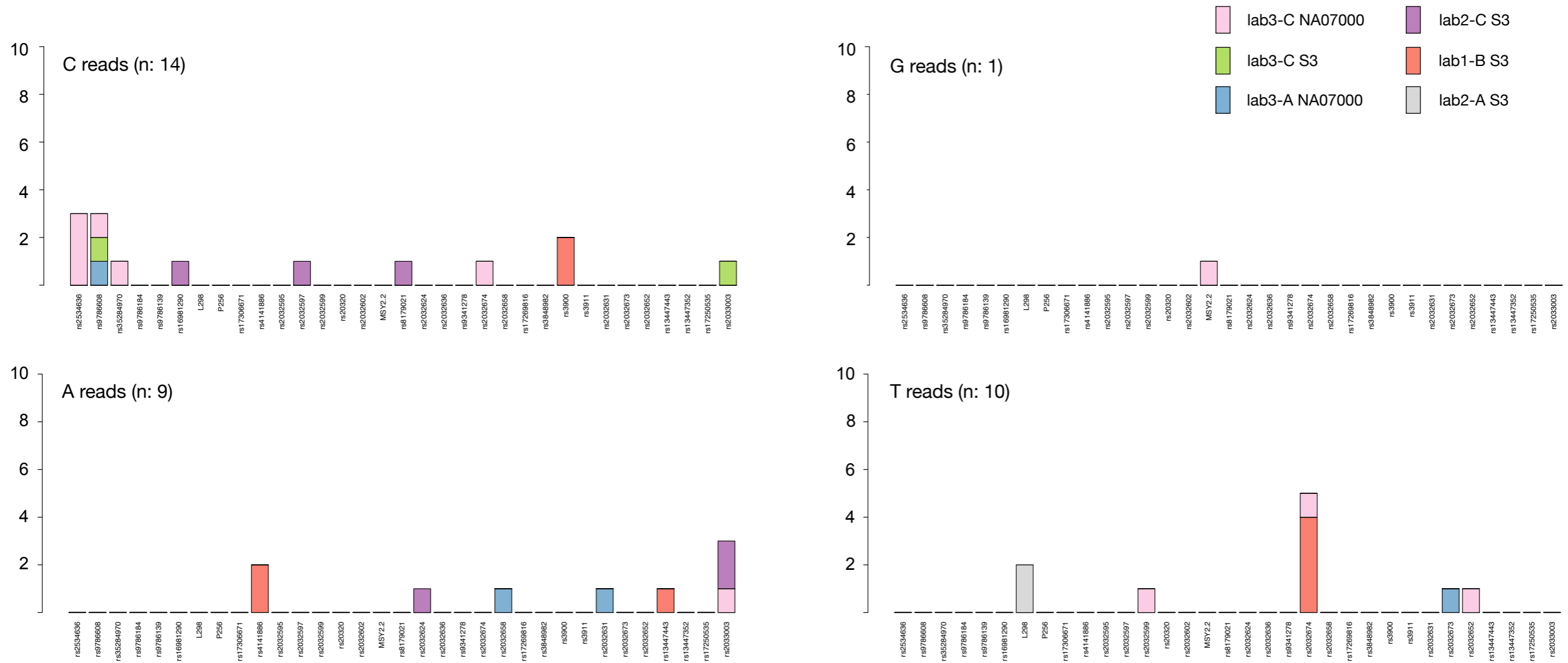
Supplementary Fig. S1 Proportions of four types of sequence reads from 12 Ion PGM™ runs using the full range of available sequencing chips.



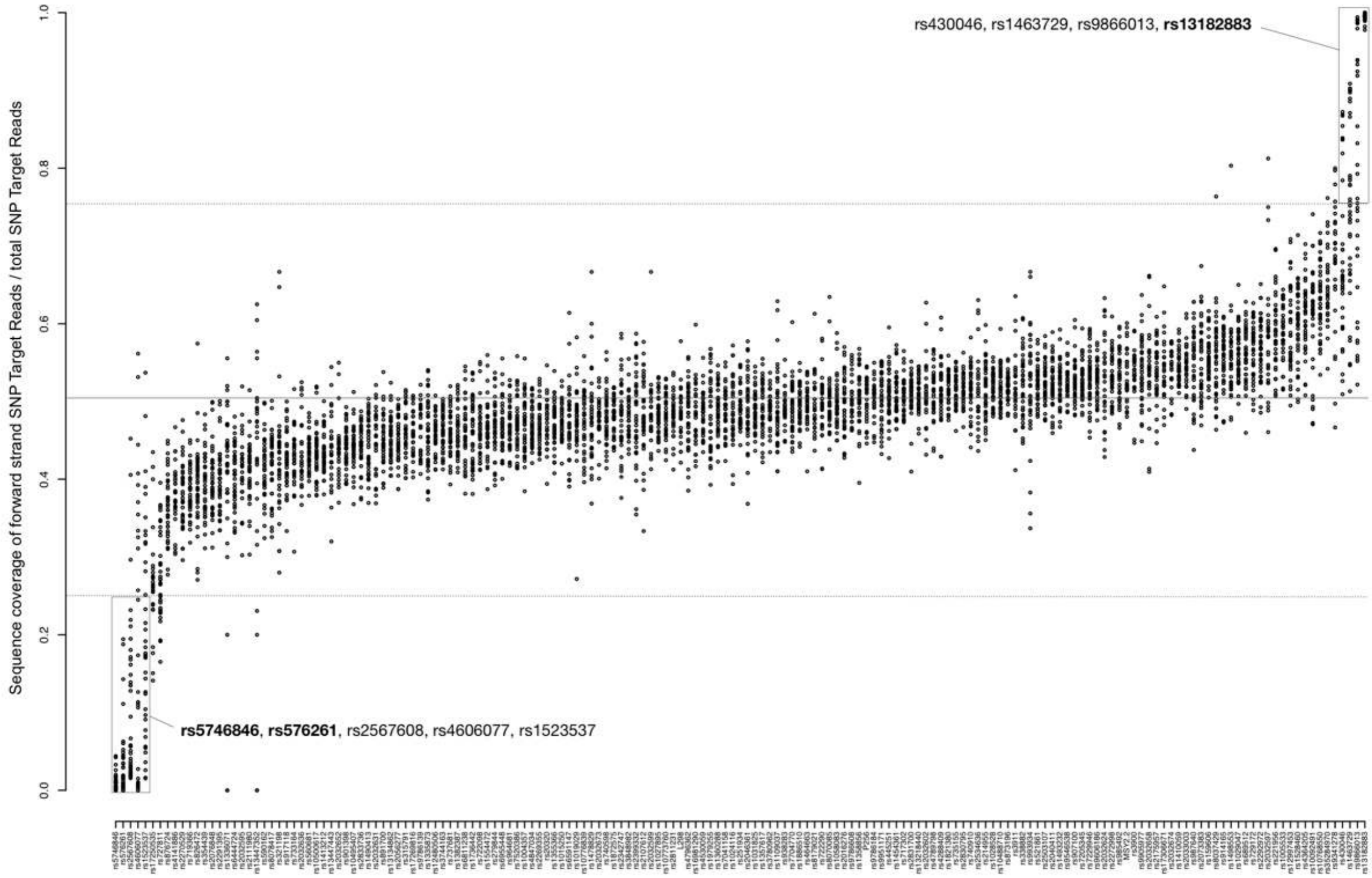
Supplementary Fig. S6 Base misincorporation rates recorded as the presence of non-allelic reference or alternative bases (e.g. low levels of A in G homozygotes plus G in A homozygotes); non-specific base incorporation (e.g. C or T in an A/G SNP) and deletions.



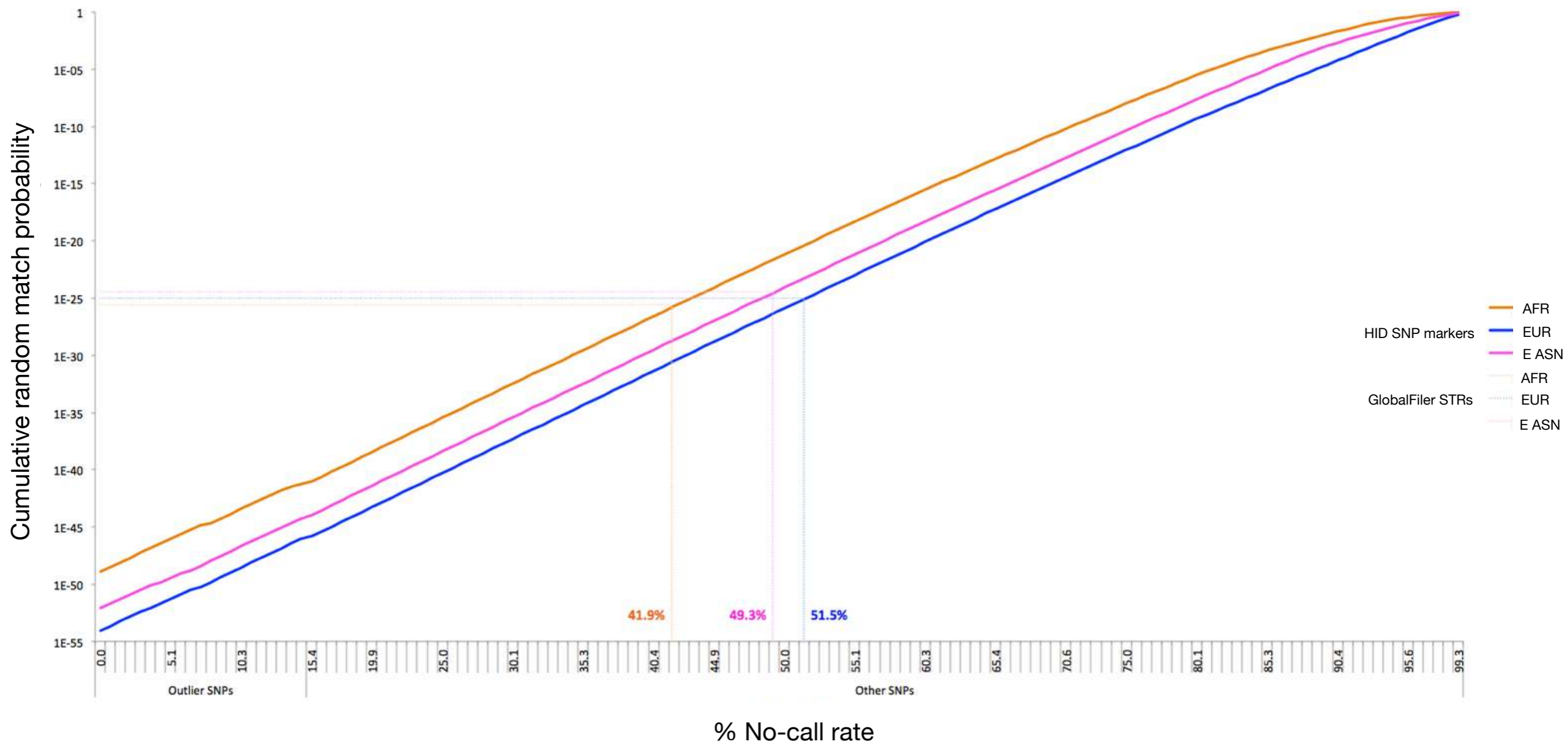
Supplementary Fig. S7 Y-SNP nucleotide reads recorded in analyses of female DNA samples. Numbers of reads indicate very low levels of extraneous male sequences amongst much higher quantities of autosomal SNP target sequence obtained (34 sequences in 6 samples).



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.



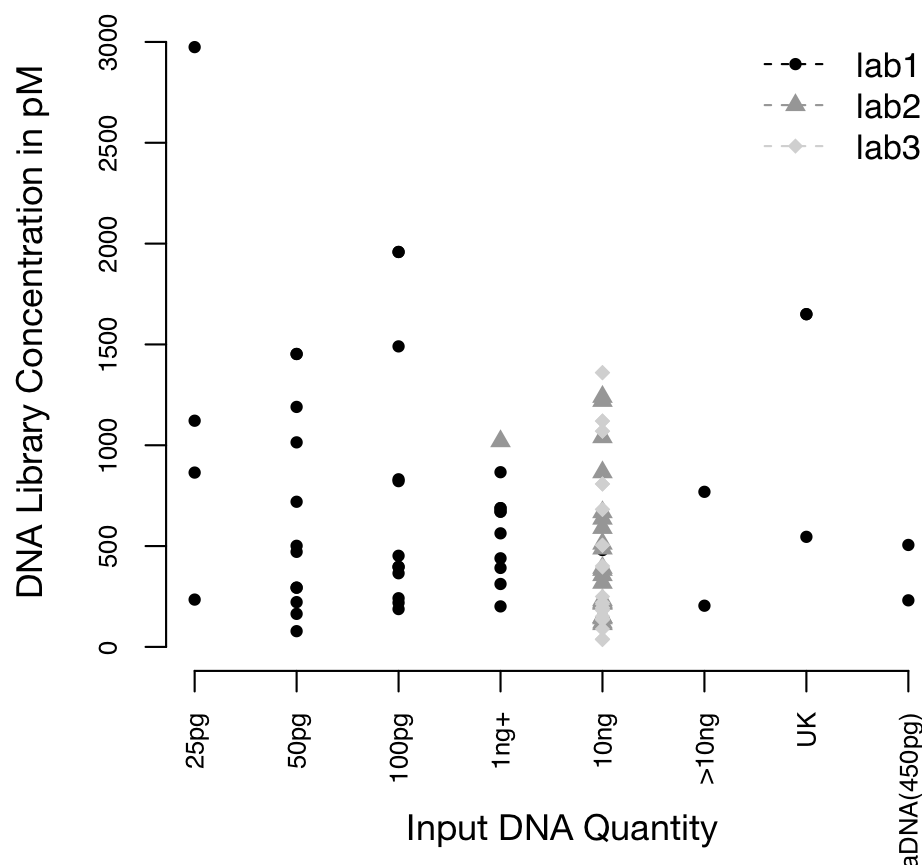
Supplementary Fig. S11 Reduction in cumulative RMP with increasing no-call rate



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

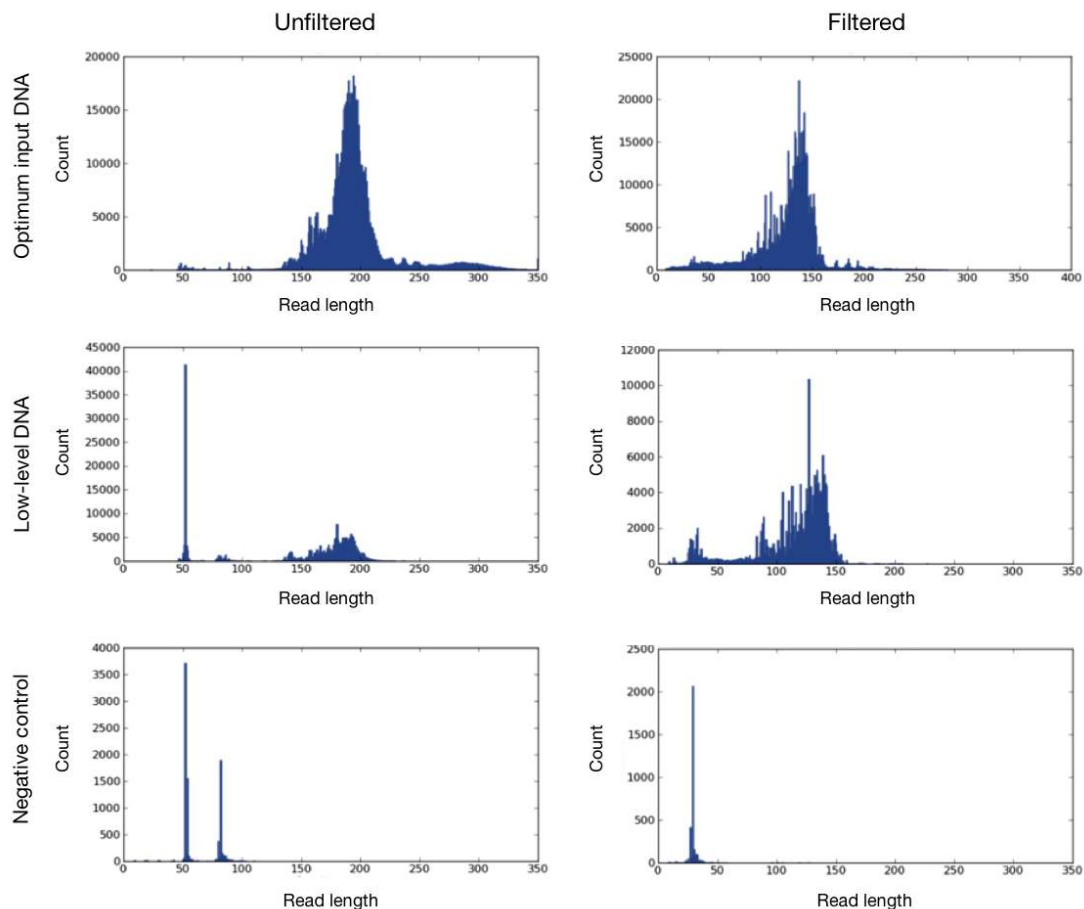
1.1. Comprehensive sequencing output analysis

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 (Command line arguments: '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 ($p=2 \times 10^{-6}$, $\alpha=0.05$). 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.

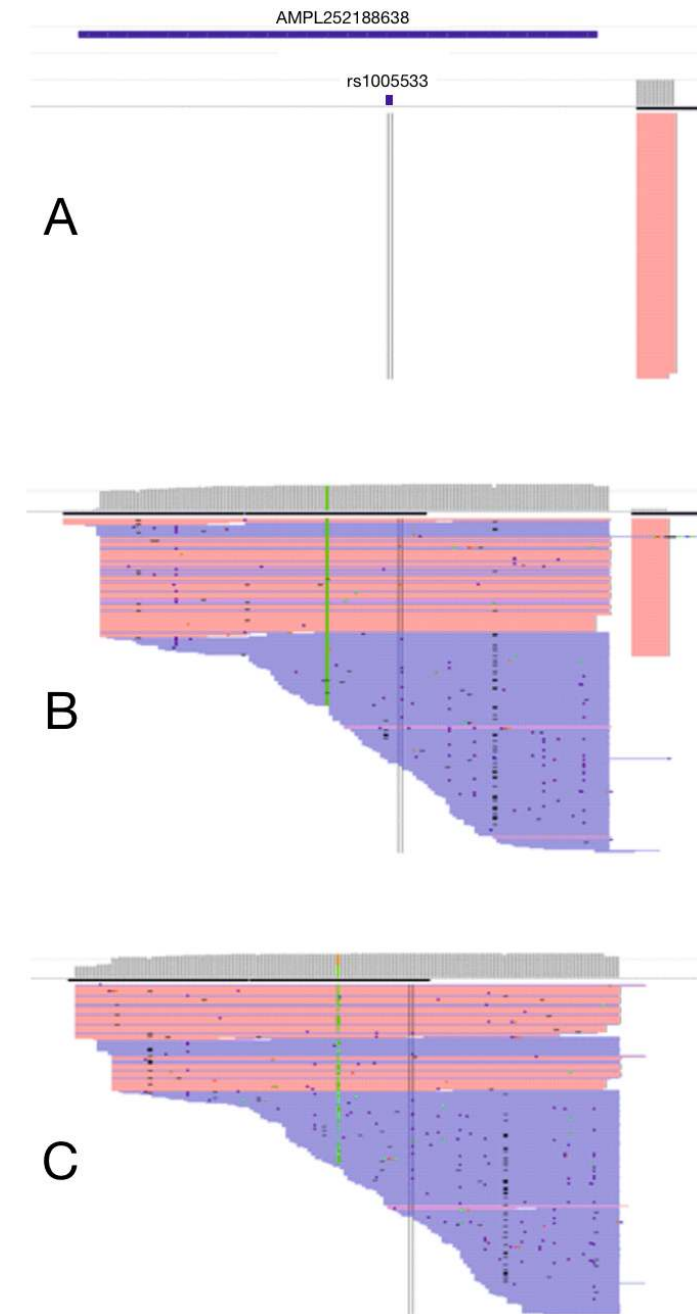
Negative controls were sequenced in two different runs. In the first run two optimum input DNA samples (positive controls) diluted to 100 pM and two undiluted negative controls were pooled and diluted 2:23 for template preparation and run on a 314v2 Chip. The percentage of polyclonal reads rose to 51% 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.

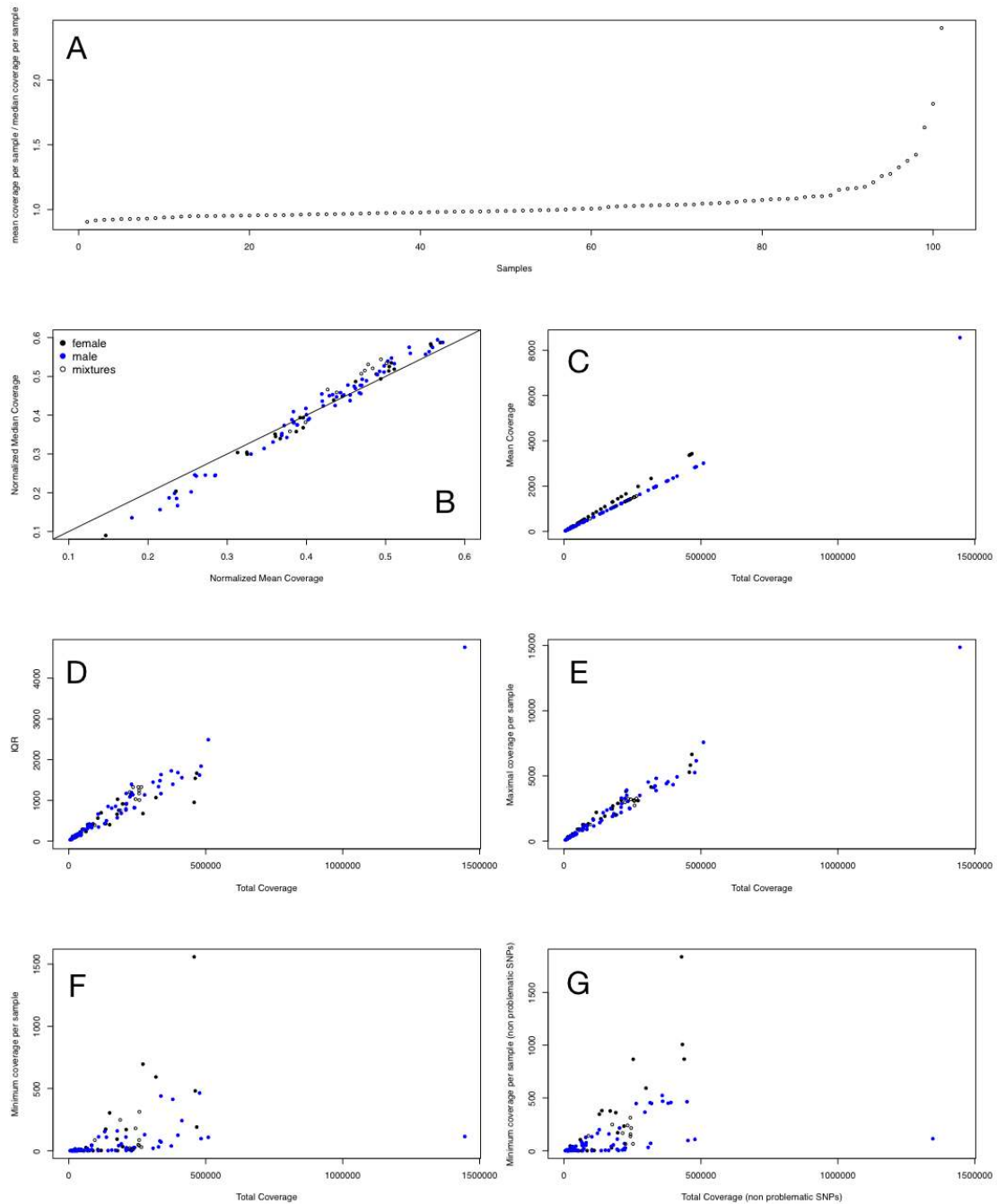
(B) Similar reads can be seen in a low-level DNA sample.

(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).

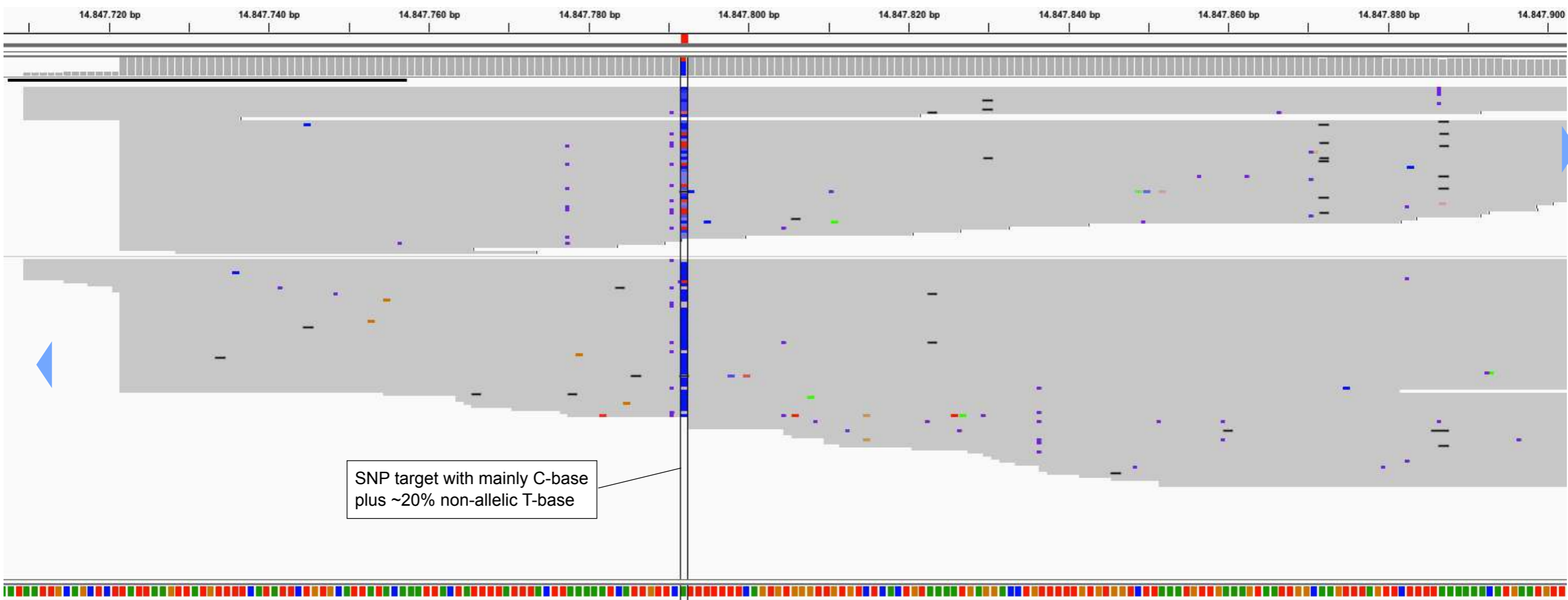
(D-E) Interquartile range of SNP coverage per sample and maximum coverage rise with total coverage.

(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.

SNP 1: rs2032597

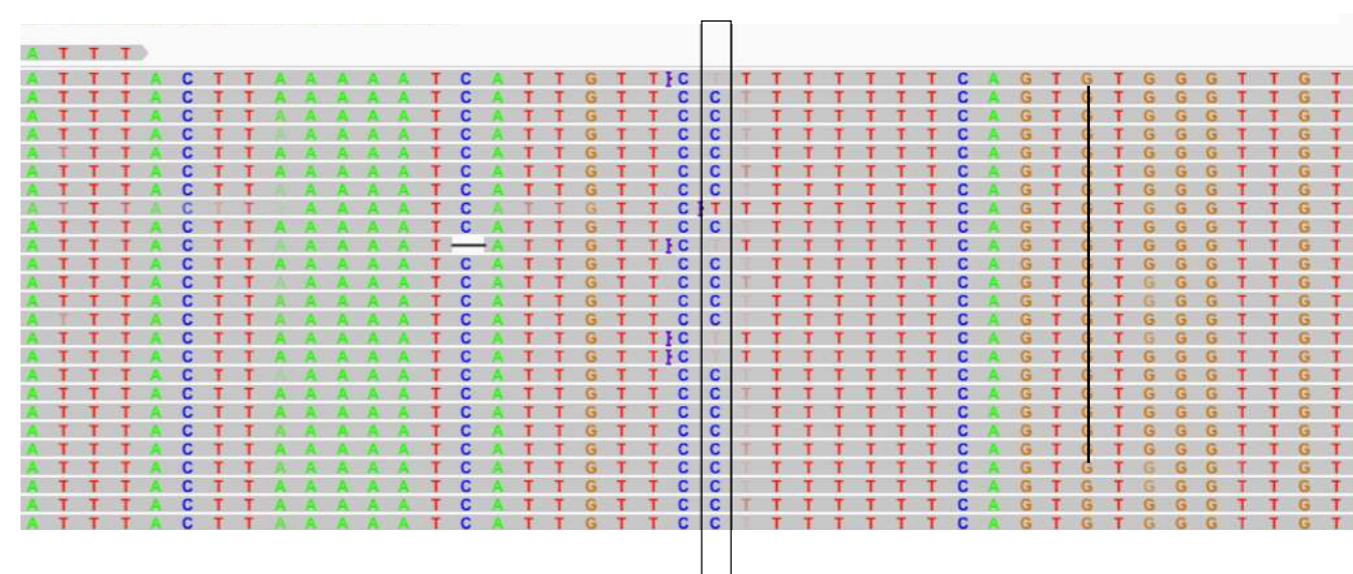
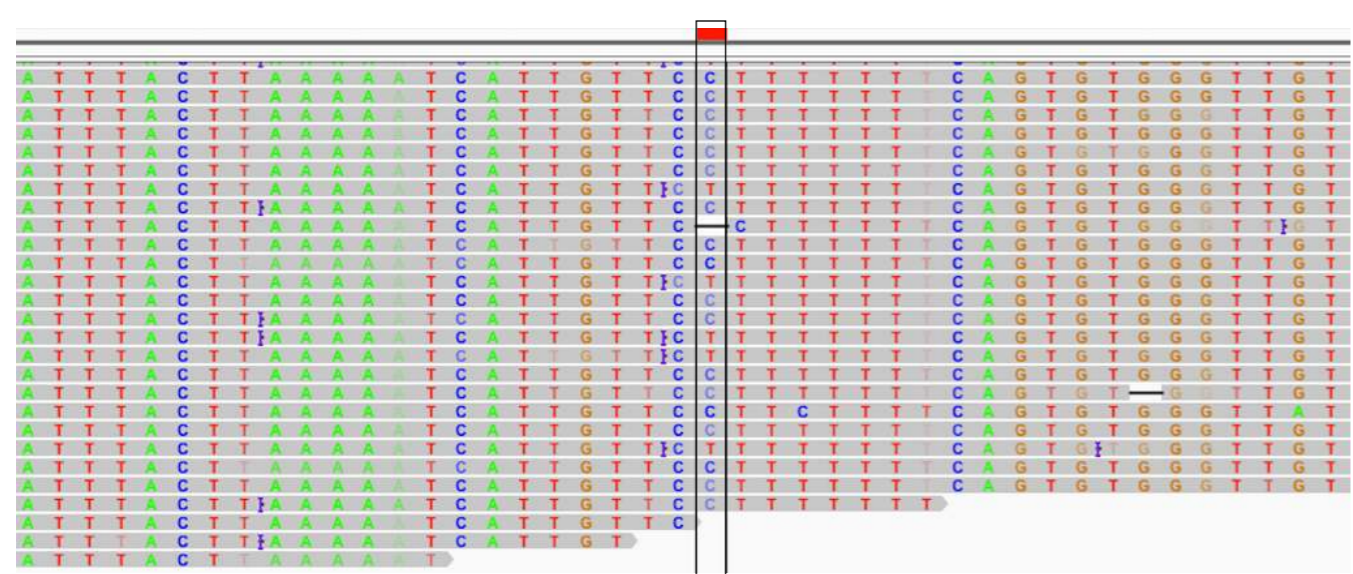
IGV overview of discordant Y-rs2032597 showing misalignment of the SNP site and immediate sequence due to C anchor base (matching one SNP allele) within a poly-T tract

■ C ■ T ■ A ■ G ▶ Direction
■ } Insertion (INS) ▬ Deletion (DEL)



SNP target with mainly C-base plus ~20% non-allelic T-base

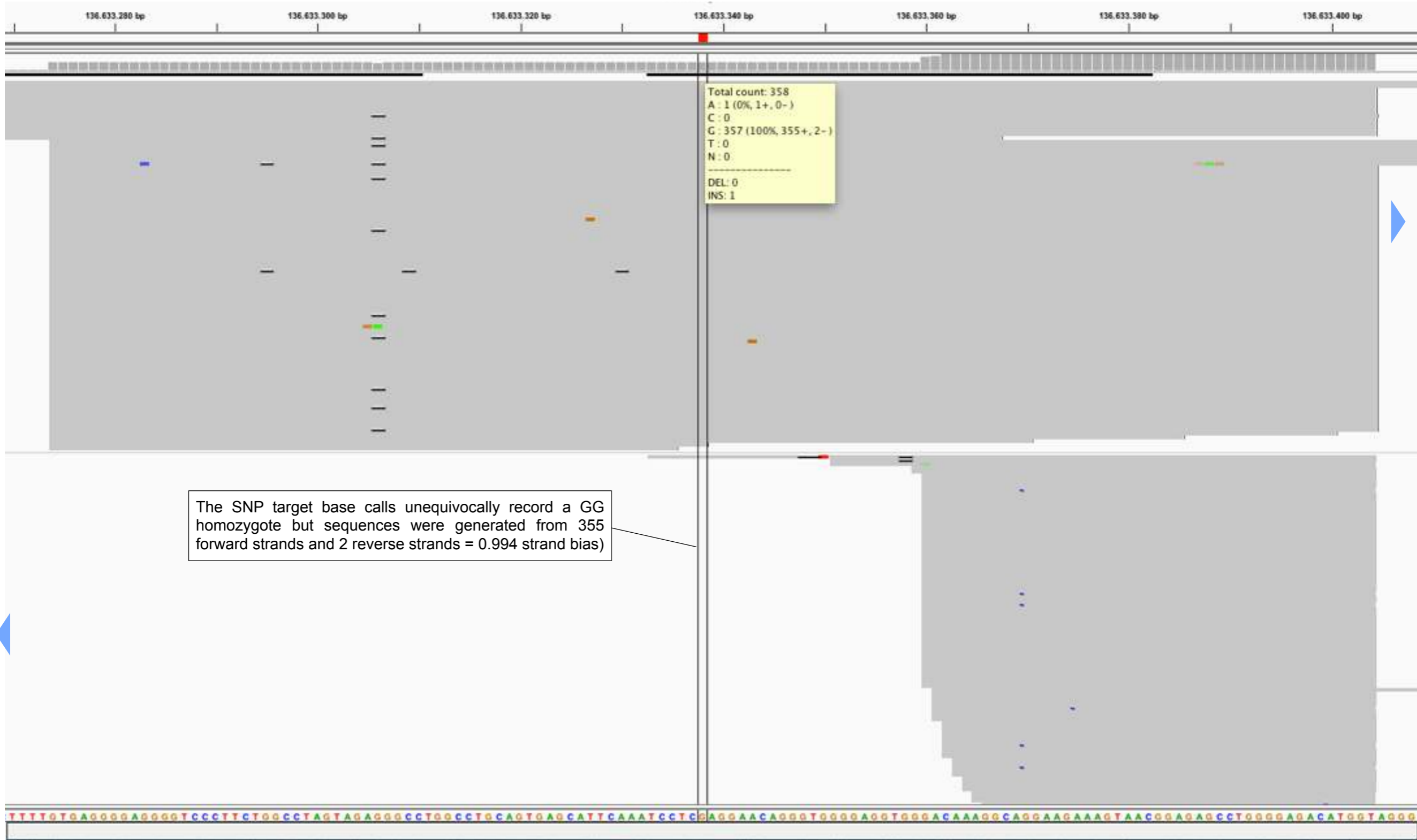
A-base in the genome reference sequence used to make the alignment (gray boxes above denote identical bases in the analysis)



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.

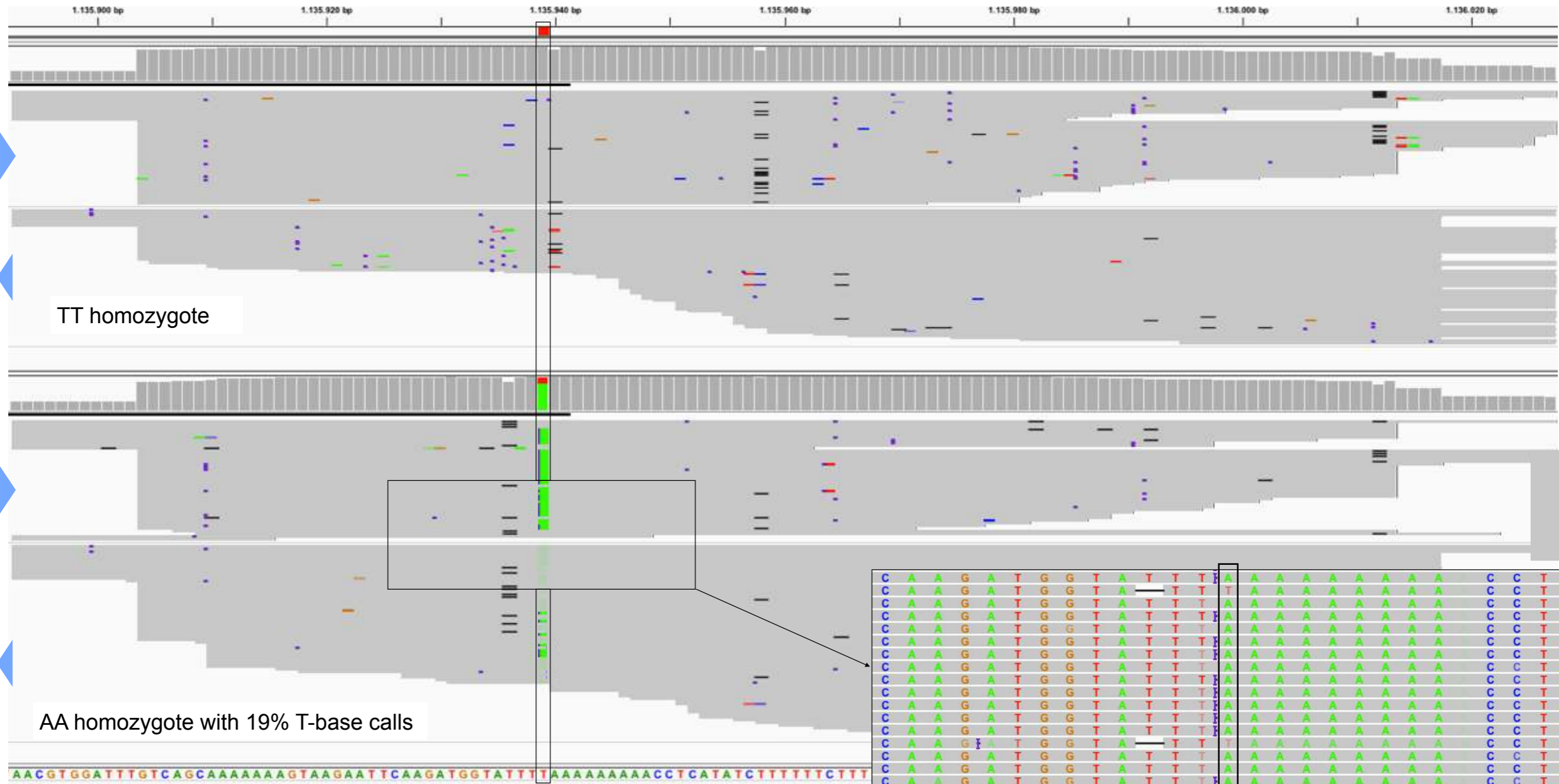
■ C ■ T ■ A ■ G ▶ Direction
■ Insertion (INS) ■ Deletion (DEL)



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.

■ C ■ T ■ A ■ G ▶ Direction
 ■ Insertion (INS) ■ Deletion (DEL)

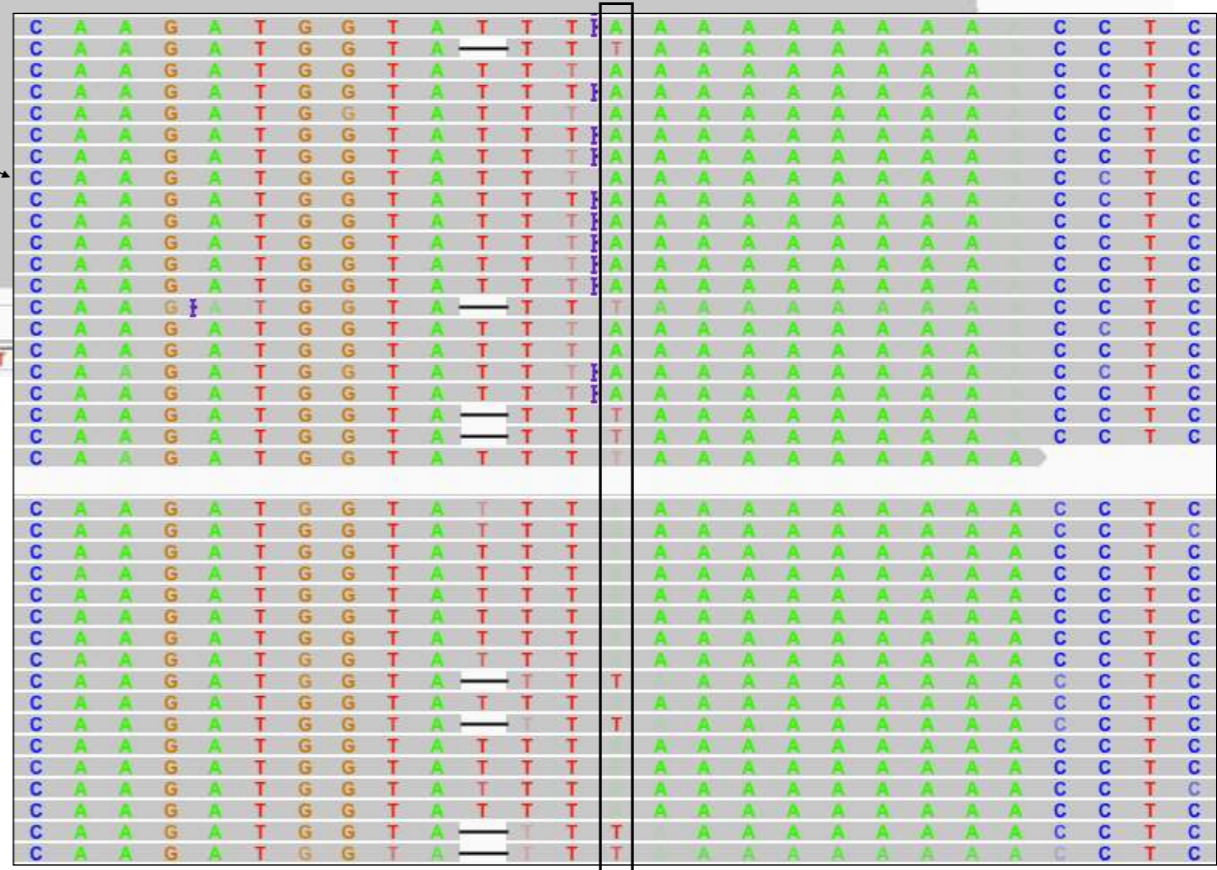


AA homozygote with 19% T-base calls

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

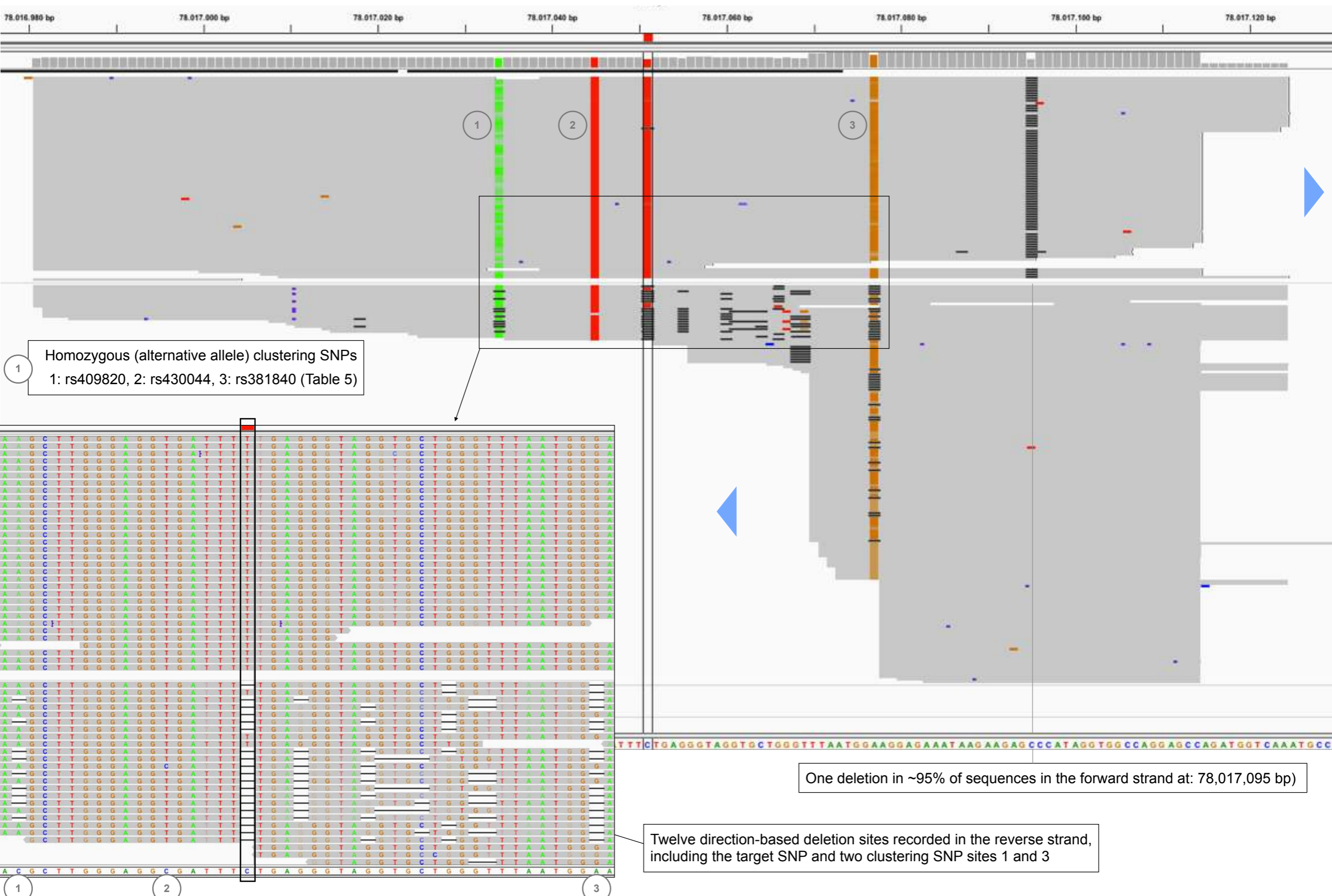
The SNP target base sequence counts indicate 19% spurious T-base calls were made due to misalignment of 3-T tract in the forward strand or misalignment of both 3-T and 8-A tracts in the reverse strand



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.

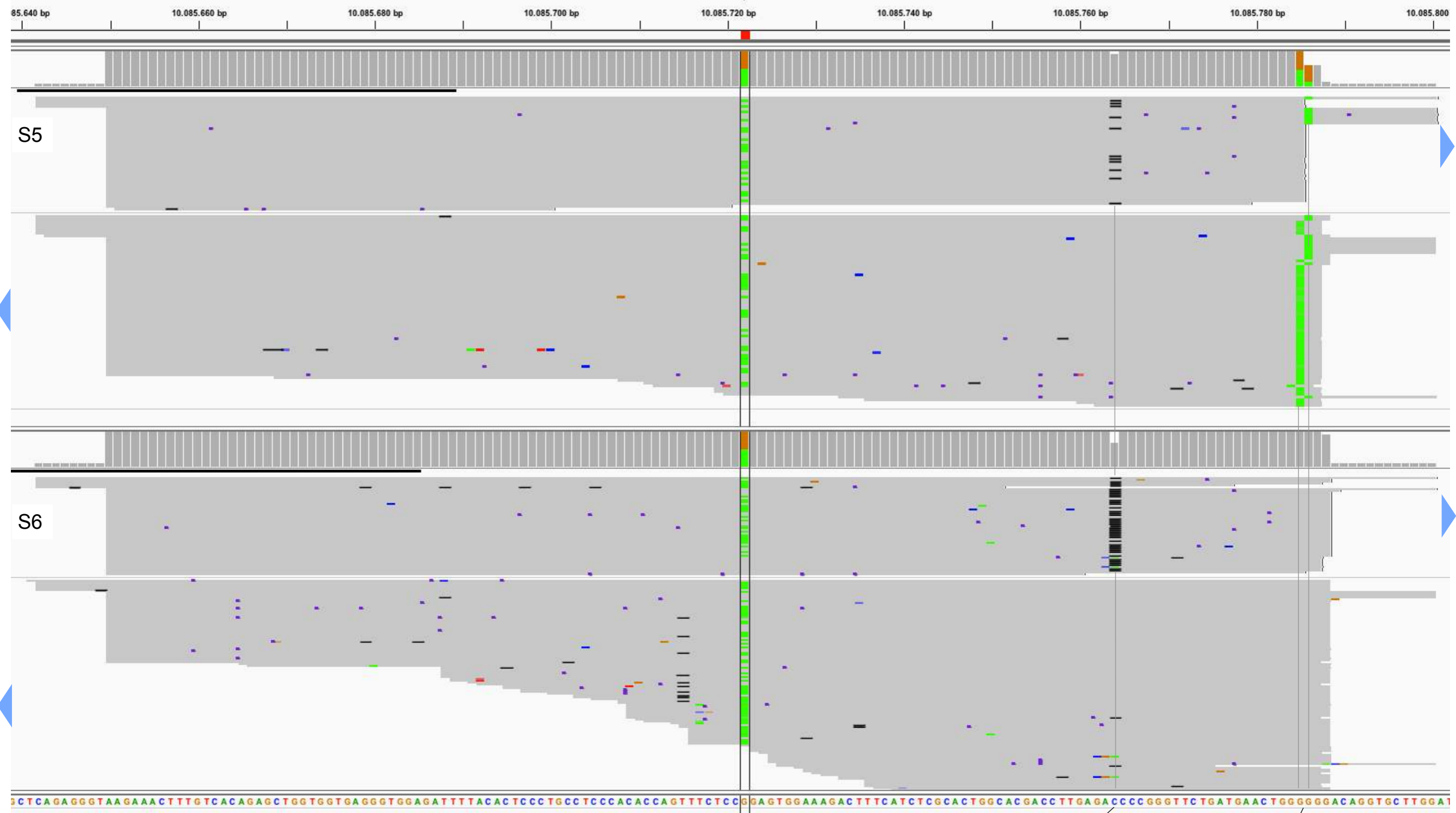
C T A G Direction
Insertion (INS) Deletion (DEL)



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.

■ C ■ T ■ A ■ G ▶ Direction
■ Insertion (INS) ■ Deletion (DEL)



An artifact Indel is created in the misaligned 4-C tract in the forward strand

An artifact variant is created in the misaligned 6-G tract in the reverse strand at 10,085,785 (and A misreads made in the adjacent 10,085,786 site in both strands)

Supplementary File S3. Mixture analysis with the Ion PGM™

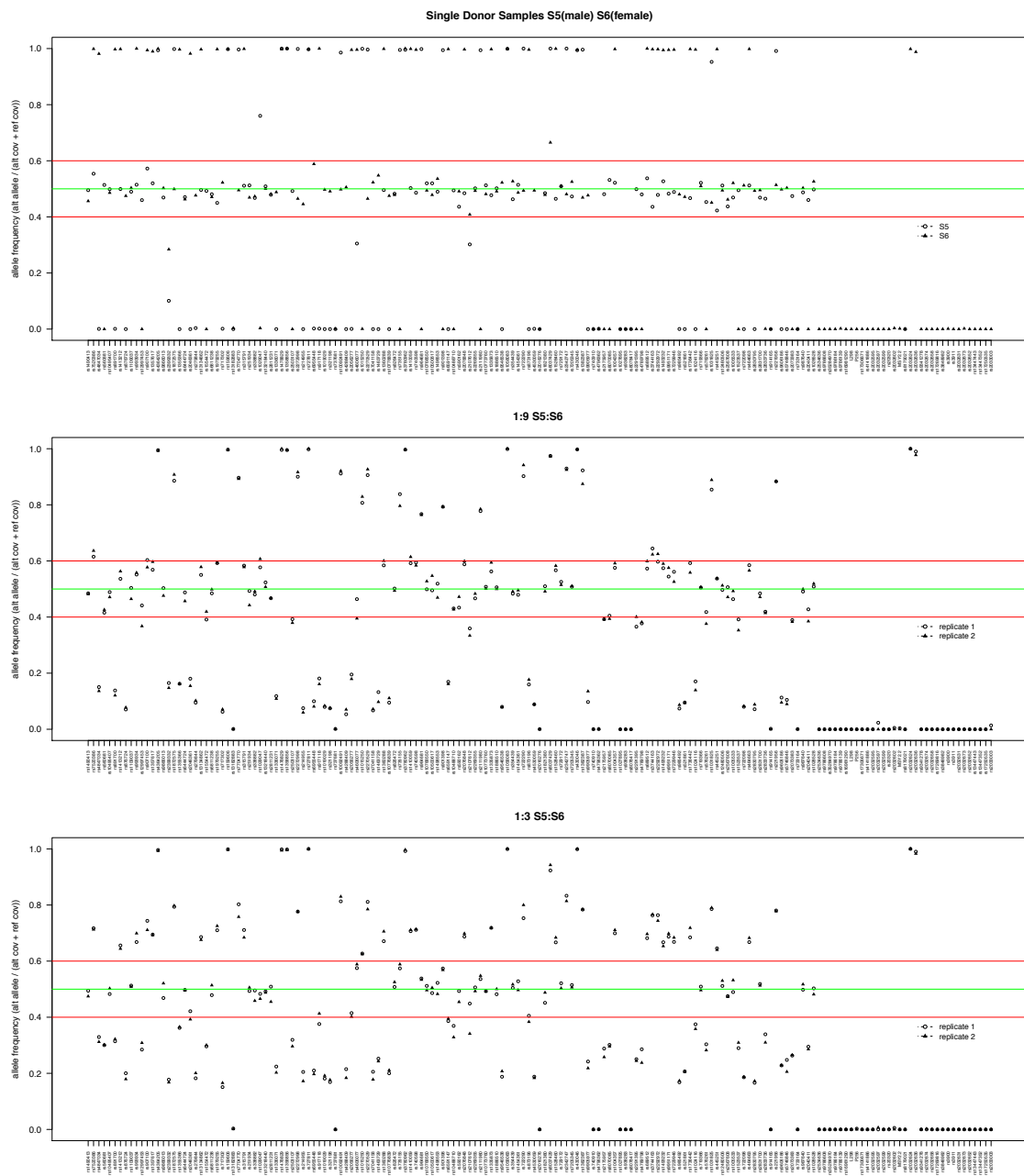
The detection of mixed source DNA samples and the de-convolution of component genetic profiles is difficult when analysing binary SNPs with the commonly used SNaPshot® chemistry. But the use of NGS, in this particular study the Ion PGM™ pipeline and the AmpliSeq™ technology, provides balanced heterozygous 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

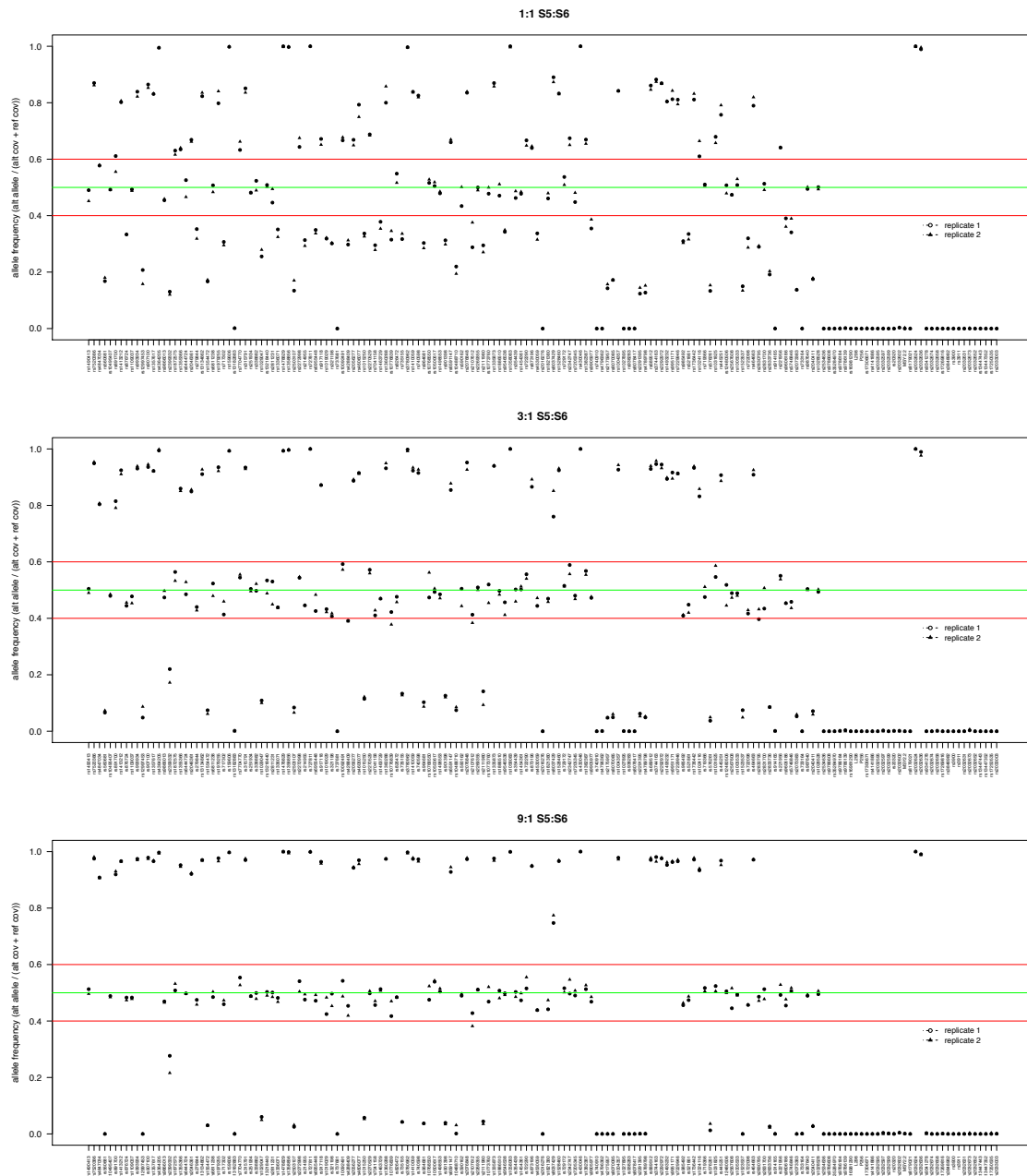
The five mixed samples were first assessed for imbalanced ARF distributions. The distributions observed in the 1:1 mixture are shown in Fig. 3B in the main article, while those of all mixture ratio replicates and donor samples in Supplementary Fig. S9. As described in section 3.2.2, nearly all SNP ARFs 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 limits¹. 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.

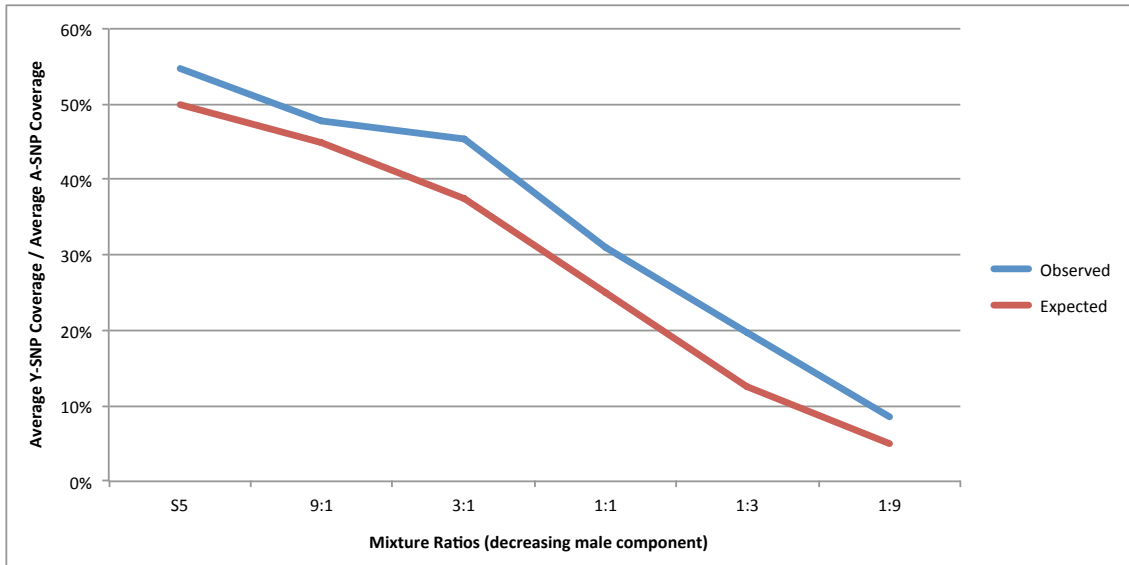
Ion PGM™ applied to medical sequencing has a strong focus on detection of somatic mutations (e.g. cancer genetics) where a novel base is present at a very low frequency compared to the normal reference-genome base. In order to control the false positive rate to manageable levels, Ion PGM™ Somatic analysis 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.