Vol 456 | 6 November 2008 | doi:10.1038/nature07517

ARTICLES

Accurate whole human genome sequencing using reversible terminator chemistry

A list of authors and their affiliations appears at the end of the paper

DNA sequence information underpins genetic research, enabling discoveries of important biological or medical benefit. Sequencing projects have traditionally used long (400–800 base pair) reads, but the existence of reference sequences for the human and many other genomes makes it possible to develop new, fast approaches to re-sequencing, whereby shorter reads are compared to a reference to identify intraspecies genetic variation. Here we report an approach that generates several billion bases of accurate nucleotide sequence per experiment at low cost. Single molecules of DNA are attached to a flat surface, amplified *in situ* and used as templates for synthetic sequencing with fluorescent reversible terminator deoxyribonucleotides. Images of the surface are analysed to generate high-quality sequence. We demonstrate application of this approach to human genome sequencing on flow-sorted X chromosomes and then scale the approach to determine the genome sequence of a male Yoruba from Ibadan, Nigeria. We build an accurate consensus sequence from >30× average depth of paired 35-base reads. We characterize four million single-nucleotide polymorphisms and four hundred thousand structural variants, many of which were previously unknown. Our approach is effective for accurate, rapid and economical whole-genome re-sequencing and many other biomedical applications.

DNA sequencing yields an unrivalled resource of genetic information. We can characterize individual genomes, transcriptional states and genetic variation in populations and disease. Until recently, the scope of sequencing projects was limited by the cost and throughput of Sanger sequencing. The raw data for the three billion base (3 gigabase (Gb)) human genome sequence, completed in 2004 (ref. 1), was generated over several years for ~$300 million using several hundred capillary sequencers. More recently an individual human genome sequence has been determined for ~$10 million by capillary sequencing[2]. Several new approaches at varying stages of development aim to increase sequencing throughput and reduce cost[3–6]. They increase parallelization markedly by imaging many DNA molecules simultaneously. One instrument run produces typically thousands or millions of sequences that are shorter than capillary reads. Another human genome sequence was recently determined using one of these approaches[7]. However, much bigger improvements are necessary to enable routine whole human genome sequencing in genetic research.

We describe a massively parallel synthetic sequencing approach that transforms our ability to use DNA and RNA sequence information in biological systems. We demonstrate utility by re-sequencing an individual human genome to high accuracy. Our approach delivers data at very high throughput and low cost, and enables extraction of genetic information of high biological value, including single-nucleotide polymorphisms (SNPs) and structural variants.

DNA sequencing using reversible terminators

We generated high-density single-molecule arrays of genomic DNA fragments attached to the surface of the reaction chamber (the flow cell) and used isothermal 'bridging' amplification to form DNA 'clusters' from each fragment. We made the DNA in each cluster single-stranded and added a universal primer for sequencing. For paired read sequencing, we then converted the templates to double-stranded DNA and removed the original strands, leaving the complementary strand as template for the second sequencing reaction (Fig. 1a–c). To obtain paired reads separated by larger distances, we circularized DNA fragments of the required length (for example, 2 ± 0.2 kb) and obtained short junction fragments for paired end sequencing (Fig. 1d).

We sequenced DNA templates by repeated cycles of polymerase-directed single base extension. To ensure base-by-base nucleotide incorporation in a stepwise manner, we used a set of four reversible terminators, 3′-*O*-azidomethyl 2′-deoxynucleoside triphosphates (A, C, G and T), each labelled with a different removable fluorophore (Supplementary Fig. 1a)[8]. The use of 3′-modified nucleotides allowed the incorporation to be driven essentially to completion without risk of over-incorporation. It also enabled addition of all four nucleotides simultaneously rather than sequentially, minimizing risk of misincorporation. We engineered the active site of 9°N DNA polymerase to improve the efficiency of incorporation of these unnatural nucleotides[9]. After each cycle of incorporation, we determined the identity of the inserted base by laser-induced excitation of the fluorophores and imaging. We added tris(2-carboxyethyl)phosphine (TCEP) to remove the fluorescent dye and side arm from a linker attached to the base and simultaneously regenerate a 3′ hydroxyl group ready for the next cycle of nucleotide addition (Supplementary Fig. 1b). The Genome Analyzer (GA1) was designed to perform multiple cycles of sequencing chemistry and imaging to collect the sequence data automatically from each cluster on the surface of each lane of an eight-lane flow cell (Supplementary Fig. 2).

To determine the sequence from each cluster, we quantified the fluorescent signal from each cycle and applied a base-calling algorithm. We defined a quality (Q) value for each base call (scaled as by the phred algorithm[10]) that represents the likelihood of each call being correct (Supplementary Fig. 3). We used the Q-values in subsequent analyses to weight the contribution of each base to sequence alignment and detection of sequence variants (for example, SNP

©2008 Macmillan Publishers Limited. All rights reserved

Illumina Exhibit 1006, Page 1 of 67

calling). We discarded all reads from mixed clusters and used the remaining 'purity filtered' reads for analysis. Typically we generated 1–2 Gb of high-quality purity filtered sequence per flow cell from ~30–60-million single 35-base reads, or 2–4 Gb in a paired read experiment (Supplementary Table 1).

To demonstrate accurate sequencing of human DNA, we sequenced a human bacterial artificial chromosome (BAC) clone (bCX98J21) that contained 162,752 bp of the major histocompatibility complex on human chromosome 6 (accession AL662825.4, previously determined using capillary sequencing by the Wellcome Trust Sanger Institute). We developed a fast global alignment algorithm ELAND that aligns a read to the reference only if the read can be assigned a unique position with 0, 1 or 2 differences. We collected 0.17 Gb of aligned data for the BAC from one lane of a flow cell. Approximately 90% of the 35-base reads matched perfectly to the reference, demonstrating high raw read accuracy (Supplementary Fig. 4). To examine consensus coverage and accuracy, we used 5 Mb of 35-base purity filtered reads (30-fold average input depth of the BAC) and obtained 99.96% coverage of the reference. There was one consensus miscall, at a position of very low coverage (just above our cutoff threshold), yielding an overall consensus accuracy of >99.999%.

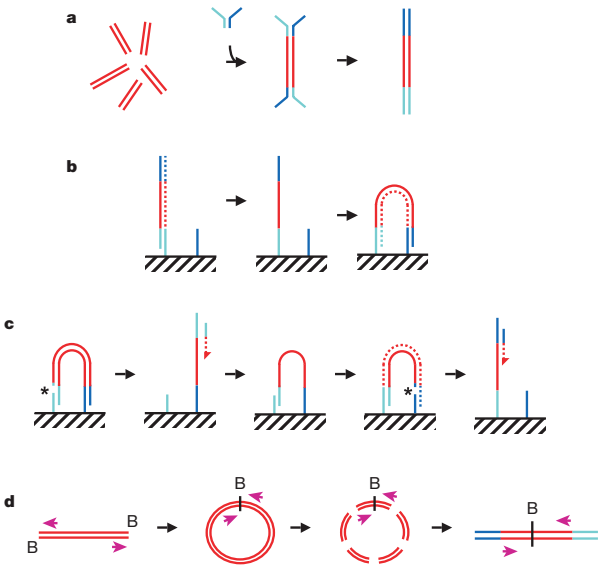


Figure 1 | Preparation of samples. **a**, DNA fragments are generated, for example, by random shearing and joined to a pair of oligonucleotides in a forked adaptor configuration. The ligated products are amplified using two oligonucleotide primers, resulting in double-stranded blunt-ended material with a different adaptor sequence on either end. **b**, Formation of clonal single-molecule array. DNA fragments prepared as in **a** are denatured and single strands are annealed to complementary oligonucleotides on the flow-cell surface (hatched). A new strand (dotted) is copied from the original strand in an extension reaction that is primed from the 3′ end of the surface-bound oligonucleotide; the original strand is then removed by denaturation. The adaptor sequence at the 3′ end of each copied strand is annealed to a new surface-bound complementary oligonucleotide, forming a bridge and generating a new site for synthesis of a second strand (dotted). Multiple cycles of annealing, extension and denaturation in isothermal conditions result in growth of clusters, each ~1 μm in physical diameter. This follows the basic method outlined in ref. 33. **c**, The DNA in each cluster is linearized by cleavage within one adaptor sequence (gap marked by an asterisk) and denatured, generating single-stranded template for sequencing by synthesis to obtain a sequence read (read 1; the sequencing product is dotted). To perform paired-read sequencing, the products of read 1 are removed by denaturation, the template is used to generate a bridge, the second strand is re-synthesized (shown dotted), and the opposite strand is then cleaved (gap marked by an asterisk) to provide the template for the second read (read 2). **d**, Long-range paired-end sample preparation. To sequence the ends of a long (for example, >1 kb) DNA fragment, the ends of each fragment are tagged by incorporation of biotinylated (B) nucleotide and then circularized, forming a junction between the two ends. Circularized DNA is randomly fragmented and the biotinylated junction fragments are recovered and used as starting material in the standard sample preparation procedure illustrated in **a**. The orientation of the sequence reads relative to the DNA fragment is shown (magenta arrows). When aligned to the reference sequence, these reads are oriented with their 5′ ends towards each other (in contrast to the short insert paired reads produced as shown in **a**–**c**). See Supplementary Fig. 17a for examples of both. Turquoise and blue lines represent oligonucleotides and red lines represent genomic DNA. All surface-bound oligonucleotides are attached to the flow cell by their 5′ ends. Dotted lines indicate newly synthesized strands during cluster formation or sequencing. (See Supplementary Methods for details.)

Detecting genetic variation of the human X chromosome

For an initial study of genetic variation, we sequenced flow-sorted X chromosomes of a Caucasian female (sample NA07340 originating from the Centre d'Etude du Polymorphisme Humain (CEPH)). We generated 278-million paired 30–35-bp purity filtered reads and aligned them to the human genome reference sequence. We carried out separate analyses of the data using two alignment algorithms: ELAND (see above) or MAQ (Mapping and Assembly with Qualities)[11]. Both algorithms place each read pair where it best matches the reference and assign a confidence score to the alignment. In cases where a read has two or more equally likely positions (that is, in an exact repeat), MAQ randomly assigns the read pair to one position and assigns a zero alignment quality score (these reads are excluded from SNP analysis). ELAND rejects all non-unique alignments, which are mostly in recently inserted retrotransposons (see Supplementary Fig. 5). MAQ therefore provides an opportunity to assess the properties of a data set aligned to the entire reference, whereas ELAND effectively excludes ambiguities from the short read alignment before further analysis.

We obtained comprehensive coverage of the X chromosome from both analyses. With MAQ, 204 million reads aligned to 99.94% of the X chromosome at an average depth of 43×. With ELAND, 192 million reads covered 91% of the reference sequence, showing what can be covered by unique best alignments. These results were obtained after excluding reads aligning to non-X sequence (impurities of flow sorting) and apparently duplicated read pairs (Supplementary Table 2). We reasoned that these duplicates (~10% of the total) arose during initial sample amplification.

The sampling of sequence fragments from the X chromosome is close to random. This is evident from the distribution of mapped read depth in the MAQ alignment in regions where the reference is unique (Fig. 2a): the variance of this distribution is only 2.26 times that of a Poisson distribution (the theoretical minimum). Half of this excess variance can be accounted for by a dependence on G+C content. However, the average mapped read depth only falls below 10× in regions with G+C content less than 4% or greater than 76%, comprising in total just 1% of unique chromosome sequence and 3% of coding sequence (Fig. 2b).

We identified 92,485 candidate SNPs in the X chromosome using ELAND (Supplementary Fig. 6). Most calls (85%) match previous entries in the public database dbSNP. Heterozygosity (π) in this data set is 4.3×10^{-4} (that is, one substitution per 2.3 kb), close to a previously published X chromosome estimate (4.7×10^{-4})[12]. Using MAQ we obtained 104,567 SNPs, most of which were common to the results of the ELAND analysis. The differences between the two sets of SNP calls are largely the consequence of different properties of the alignments as described earlier. For example, most of the SNPs found only by the MAQ-based analysis were at positions of low or zero sequence depth in the ELAND alignment (Supplementary Fig. 6c).

We assessed accuracy and completeness of SNP calling by comparison to genotypes obtained for this individual using the Illumina HumanHap550 BeadChip (HM550). The sequence data covered >99.8% of the 13,604 genotyped positions and we found excellent

©2008 Macmillan Publishers Limited. All rights reserved

Illumina Exhibit 1006, Page 2 of 67

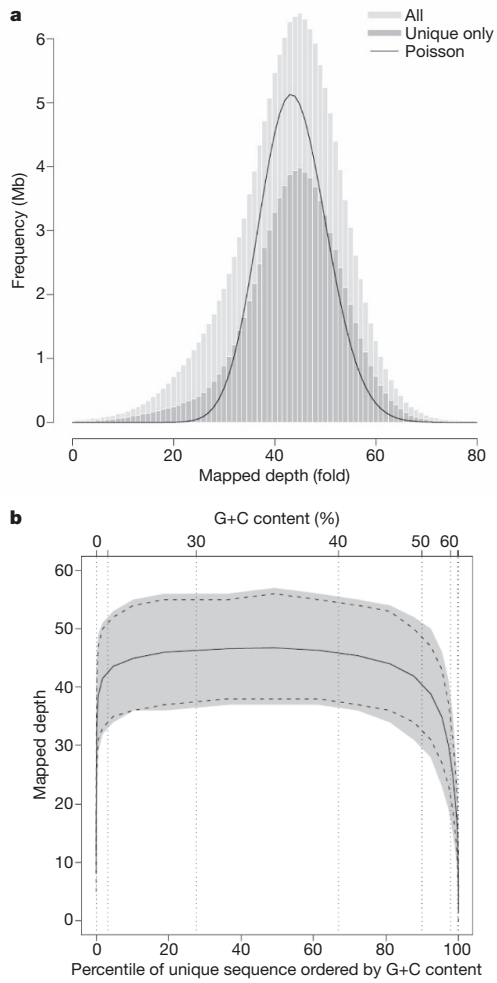


Figure 2 | X chromosome data. a, Distribution of mapped read depth in the X chromosome data set (NA07340), sampled at every 50th position along the chromosome and displayed as a histogram ('All'). An equivalent analysis of mapped read depth for the unique subset of these positions is also shown ('Unique only'). The solid line represents a Poisson distribution with the same mean. **b**, Distribution of X chromosome uniquely mapped reads as a function of G+C content. Note that the *x* axis is per cent G+C content and is scaled by percentile of unique sequence. The solid line is average mapped depth of unique sequence; the grey region is the central 80% of the data (10th to 90th centiles); the dashed lines are 10th and 90th centiles of a Poisson distribution with the same mean as the data.

agreement between sequence-based SNP calls and genotyping data (99.52% or 99.99% using ELAND or MAQ, respectively; Supplementary Table 3). There was complete concordance of all homozygous calls and a low level of 'under-calling' from the sequence data (denoted as 'GT>Seq' in Table 1) at a small number of the heterozygous sites, caused by inadequate sampling of one of the two alleles. The depth of input sequence influences the coverage and accuracy of SNP calling. We found that reducing the read depth to 15× still gives 97% coverage of genotype positions and only 1.27% of the heterozygous sites are under-called. We observed no other types of disagreement at any input depth (Supplementary Fig. 7).

We detected structural variants (defined as any variant other than a single base substitution) as follows. We found 9,747 short insertions/deletions ('short indels'; defined here as less than the length of the read) by performing a gapped alignment of individual reads (Supplementary Fig. 8). We identified larger indels based on read depth and/or anomalous read pair spacing, similar to previous approaches[13–15]. We detected 115 indels in total, 77 of which were visible from anomalous read-pair spacing (see Supplementary Tables 4 and 5). We developed Resembl, an extension to the Ensembl browser[16], to view all variants (Supplementary Fig 9). Inversions can be detected when the orientation of one read in a pair is reversed (for example, see Supplementary Fig. 10). In general, inversions occur as the result of non-allelic homologous recombination, and are therefore flanked by repetitive sequence that can compromise alignments. We found partial evidence for other inversion events, but characterization of inversions from short read data is complex because of the repeats and requires further development.

Sequencing and analysis of a whole human genome

Our X chromosome study enabled us to develop an integrated set of methods for rapid sequencing and analysis of whole human genomes. We sequenced the genome of a male Yoruba from Ibadan, Nigeria (YRI, sample NA18507). This sample was originally collected for the HapMap project[17,18] through a process of community engagement and informed consent[19] and has also been studied in other projects[20,21]. We were therefore able to compare our results with publicly available data from the same sample. We constructed two libraries: one of short inserts (~200 bp) with similar properties to the previous X chromosome library and one from long fragments (~2 kb) to provide longer-range read-pair information (see Supplementary Fig. 11 for size distributions). We generated 135 Gb of sequence (~4 billion paired 35-base reads; see Supplementary Table 6) over a period of 8 weeks (December 2007 to January 2008) on six GA1 instruments averaging 3.3 Gb per production run (see Supplementary Table 1 for example). The approximate consumables cost (based on full list price of reagents) was $250,000. We aligned 97% of the reads using MAQ and found that 99.9% of the human reference (NCBI build 36.1) was covered with one or more reads at an average of 40.6-fold depth. Using ELAND, we aligned 91% of the reads over 93% of the reference sequence at sufficient depth to call a strong consensus (>three Q30 bases). The distribution of mapped read depth was close to random, with slight over-dispersion as seen for the X chromosome data. We observed comprehensive representation across a wide range of G+C content, dropping only at the very extreme ends, but with a different pattern of distribution compared to the X chromosome (see Supplementary Fig. 12).

We identified ~4 million SNPs, with 74% matching previous entries in dbSNP (Fig. 3). We found excellent agreement of our SNP calls with genotyping results: sequence-based SNP calls covered almost all of the 552,710 loci of HM550, with >99.5% concordance of sequencing versus genotyping calls (Table 1 and Supplementary Table 7a). The few disagreements were mostly under-calls of heterozygous positions (GT>Seq) in areas of low sequence depth, providing us with a false-negative rate of <0.35% from the ELAND analysis (see Table 1). The other disagreements (0.09% of all genotypes) included errors in genotyping plus apparent tri-allelic SNPs (Supplementary Table 7a). The main cause of genotype error (0.05% of all genotypes) is the existence of a second 'hidden' SNP close to the assayed locus that disrupts the genotyping assay, leading to loss of one allele and an erroneous homozygous genotype (Supplementary Figs 13 and 14).

To examine the accuracy of SNP calling in more detail, we compared our sequence-based SNP calls with 3.7 million genotypes (HM-All) generated for this sample during the HapMap project (Table 1 and Supplementary Table 7b)[18] and found excellent concordance between the data sets. Disagreements included sequence-based under-calls of heterozygous positions in regions of low read depth. The slightly higher level of other disagreements (0.76%) seen in this analysis compared to that of the HM550 data (0.09%) is in line with the higher level of underlying genotype error rate of 0.7% for the HapMap data[18]. To refine this analysis further, we generated a set of 530,750 very high confidence reference genotypes comprising

©2008 Macmillan Publishers Limited. All rights reserved

Illumina Exhibit 1006, Page 3 of 67

Table 1 | Comparison of SNP calls made from sequence versus genotype data for the human genome (NA18507) and X chromosome (NA07340)

	ELAND			MAQ				
	X	Human	Human	X	Human	Human	Human	
	HM550 (13,604 SNPs) (%)	HM550 (552,710 SNPs) (%)	HM-All (3,699,592 SNPs) (%)	HM550 (13,604 SNPs) (%)	HM550 (552,710 SNPs) (%)	HM-All (3,699,592 SNPs) (%)	Combined (530,750 SNPs) (%)	(*n*)
Covered by sequence	99.77	99.60	99.24	99.91	99.74	99.29	99.78	529,589
Concordant calls	99.52	99.57	98.80	99.99	99.90	99.12	99.94	529,285
All disagreements	0.48	0.43	1.20	0.01	0.1	0.88	0.06	304
GT>Seq	0.48	0.35	0.46	0.01	0.03	0.15	0.02	130
Seq>GT	0	0.05	0.52	0	0.05	0.54	0.02	130
Other discordances	0	0.03	0.22	0	0.02	0.2	0.01	44

SNP panels referred to are HM550 (Illumina Infinium HumanHap550 BeadChip) and HM-All (complete data from phase 1 and phase 2 of the International HapMap Project). 'Combined' is a set of concordant genotypes from both sets (HM550 and HM-All; see text). GT>Seq denotes a heterozygous genotyping SNP call where there is a homozygous sequencing SNP call (one of the two alleles); Seq>GT denotes the converse (that is, a heterozygous sequencing SNP call where there is a homozygous genotyping call). Other discordances are differences in the two SNP calls that cannot be accounted for by one allele being missing from one call.

concordant calls in both the HM550 and HM-All genotype data sets. Comparing the results of the MAQ analysis to this high confidence set (see Table 1), we found 130 heterozygote under-calls GT>Seq (that is, a false-negative rate of 0.025%). There were also 130 heterozygote over-calls Seq>GT, but most of these are probably genotype errors as 82 have a nearby 'hidden' SNP and 3 have a nearby indel. A further 41 are tri-allelic loci, leaving at most 4 potential wrong calls by sequencing (that is, false-positive rate of 4 per 529,589 positions). Finally we selected a subset of novel SNP calls from the sequence data and tested them by genotyping. We found 96.1% agreement between sequence and genotype calls (Supplementary Table 8). However, the 47 disagreements included 10 correct sequencing calls (genotyping under-calls owing to hidden SNPs) and 7 sequencing under-calls. On this basis, therefore, the false-positive discovery rate for the one million novel SNPs is 2.5% (30 out of 1,206). For the entire data set of four million SNPs detected in this analysis, the false-positive and -negative rates both average <1%.

This genome from a Yoruba individual contains significantly more polymorphism than a genome of European descent. The autosomal heterozygosity (π) of NA18507 is 9.94×10^{-4} (1 SNP per 1,006 bp), higher than previous values for Caucasians (7.6×10^{-4}, ref. 12). Heterozygosity in the pseudoautosomal region 1 (PAR1) is substantially higher (1.92×10^{-3}) than the autosomal value. PAR1 (2.7 Mb) at the tip of the short arm of chromosomes X and Y undergoes obligatory recombination in male meiosis, which is equivalent to 20× the autosome average. This illustrates a clear correlation between recombination and nucleotide diversity. By contrast, the 0.33-Mb PAR2 region has a much lower recombination rate than PAR1; we observed that heterozygosity in PAR2 is identical to that of the autosomes in NA18507. Heterozygosity in coding regions is lower (0.54×10^{-3}) than the total autosome average, consistent with the model that some coding changes are deleterious and are lost as the result of natural selection[22]. Nevertheless, the 26,140 coding SNPs (Supplementary Fig. 15) include 5,361 non-conservative amino acid substitutions plus 153 premature termination codons (Supplementary Table 9), many of which are expected to affect protein function.

a

Call	ELAND SNPs (*n*)	ELAND In dbSNP (%)	MAQ SNPs (*n*)	MAQ In dbSNP (%)
Homozygote	1,417,320	90.1	1,503,420	90.8
Heterozygote	2,411,022	63.9	2,635,776	63.8
All	3,828,342	73.6	4,139,196	73.6

b

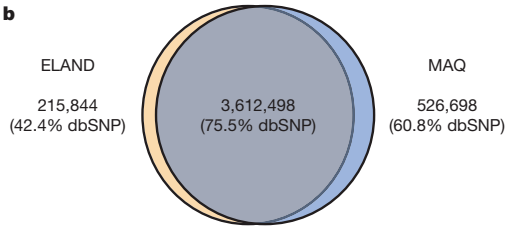


Figure 3 | SNPs identified in the human genome sequence of NA18507. **a**, Number of SNPs detected by class and percentage in dbSNP (release 128). Results from ELAND and MAQ alignments are reported separately. **b**, Analysis of SNPs detected in each analysis reveals extensive overlap. The percentage of NA18507 SNP calls that match previous entries in dbSNP is lower than that of our X chromosome study (see Supplementary Fig. 6). We expect this because individual NA07340 (from the X chromosome study) was also previously used for discovery and submission of SNPs to dbSNP during the HapMap project, in contrast to NA18507.

We performed a genome-wide survey of structural variation in this individual and found excellent correlation with variants that had been reported in previous studies, as well as detecting many new variants. We found 0.4 million short indels (1–16 bp; Supplementary Fig. 16), most of which are length polymorphisms in homopolymeric tracts of A or T. Half of these events are corroborated by entries in dbSNP, and 95 of 100 examined were present in amplicons sequenced from this individual in ENCODE regions, confirming the high specificity of this method of short indel detection. For larger structural variants (detected by anomalously spaced paired ends) we found that some were detected by both long and short insert data sets (Supplementary Fig. 17a), but most were unique to one or other data set. We observed two reasons for this: first, small events (<400 bp) are within the normal size variance of the long insert data; second, nearby repetitive structures can prevent unique alignment of read pairs (see Supplementary Fig. 17b, c). In some cases, the high resolution of the short insert data permits detection of additional complexity in a structural rearrangement that is not revealed by the long insert data. For example, where the long insert data indicate a 1.3-kb deletion in NA18507 relative to the reference, the short insert data reveal an inversion accompanied by deletions at both breakpoints (Fig. 4). We carried out *de novo* assembly of reads in this region and constructed a single contig that defines the exact structure of the rearrangement (data not shown).

We discovered 5,704 structural variants ranging from 50 bp to >35 kb where there is sequence absent from the genome of NA18507 compared to the reference genome. We observed a steadily decreasing number of events of this type with increasing size, except for two peaks (Supplementary Fig. 18). Most of the events represented by the large peak at 300–350 bp contain a sequence of the AluY family. This is consistent with insertion of short interspersed nuclear elements (SINEs) that are present in the reference genome but missing from the genome of NA18507. Similarly, the second, smaller peak at 6–7 kb is the consequence of insertion of the long interspersed nuclear element (LINE) L1 *Homo sapiens* (L1Hs) in

©2008 Macmillan Publishers Limited. All rights reserved

Illumina Exhibit 1006, Page 4 of 67

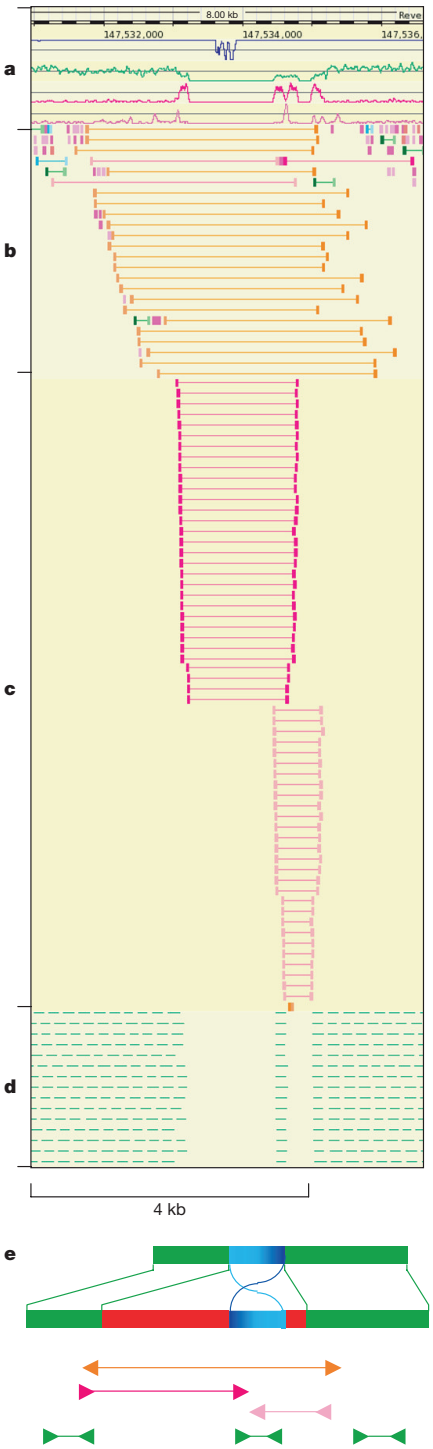


Figure 4 | Homozygous complex rearrangement detected by anomalous paired reads. The rearrangement involves an inversion of 369 bp (blue–turquoise bar in the schematic diagram) flanked by deletions (red bars) of 1,206 and 164 bp, respectively, at the left- and right-hand breakpoints. **a**, Summary tracks in the Resembl browser, denoting scale, simulated alignability of reads to reference (blue plot), actual aligned depth of coverage by NA18507 reads (green plot), density of anomalous reads indicating structural variants (red plot; peaks denote 'hotspots') and density of singleton reads (pink plot). **b**, Anomalous long-insert read pairs (orange lines denote DNA fragment; blocks at either end denote each read); the data indicate loss of ~1.3 kb in NA18507 relative to the reference. **c**, Anomalous short-insert pairs of two types (red and pink) indicate an inverted sequence flanked by two deletions. **d**, Normal short-insert read-pair alignments (each green line denotes the extent of the reference that is covered by the short fragment, including the two reads). **e**, The schematic diagram depicts the arrangement of normal and anomalous read pairs relative to the rearrangement. Top line, structure of NA18507; second line, structure of reference sequence. Green bars denote sequence that is collinear in the reference and NA18507 genomes. The turquoise–blue bar illustrates the inverted segment. Red bars indicate the sequences present in the reference but absent in NA18507. Arrows denote orientation of reads when aligned to the reference. The display in **a**–**d** is a composite of screen shots of the same window, overlapped for display purposes.

many cases. We found good correspondence between our results and the data of ref. 23, which reported 148 deletions of <100 kb in this individual on the basis of abnormal fosmid paired-end spacing. We found supporting evidence for 111 of these events. We detected a further 2,345 indels in the range 60–160 bp which are sequences present in the genome of NA18507 and absent from the reference genome (Supplementary Fig. 19). One example is shown in Supplementary Fig. 20. The 'singleton' reads on either side of the event, which have partners that do not align to the reference, form part of a *de novo* assembly that precisely defines the novel sequence and breakpoint (Supplementary Fig. 21).

Effect of sequence depth on coverage and accuracy

We investigated the impact of varying input read depth (and hence cost) on SNP calling using chromosome 2 as a model. SNP discovery increases with increasing depth: essentially all homozygous positions are detected at 15×, whereas heterozygous positions accumulate more gradually to 33× (Fig. 5a). This effect is influenced by the stringency of the SNP caller. To call each allele in this analysis we required the equivalent of two high-quality Q30 bases (as opposed to three used in full depth analyses). Homozygotes could be detected at read depth of 2× or higher, whereas heterozygote detection required at least double this depth for sampling of both alleles. Missing calls (not covered by sequence) and discordances between sequence-based SNP calls and genotype loci (mostly under-calls of heterozygotes due to low depth) progressively reduced with increasing depth (Fig. 5b). We observed very few other types of discordance at any depth; many of these are genotyping errors as described above.

Concluding remarks

Reversible terminator chemistry is a defining feature of this sequencing approach, enabling each cycle to be driven to completion while minimizing misincorporation. The result is a system that generates accurate data at very high throughput and low cost. We determined an accurate whole human genome sequence in 8 weeks to an average depth of ~40×. We built a consensus sequence, optimized methods for analysis, assessed accuracy and characterized the genetic variation of this individual in detail.

We assessed accuracy relative to genotype data over the entire fraction of the human sequence where SNP calling was possible (>90%). We established very low false-positive and -negative rates for the ~four million SNPs detected (<1% over-calls and under-calls). This compares favourably with previous individual genome analyses which reported a 24% under-calling of heterozygous positions[2,7].

Paired reads were very powerful in all areas of the analysis. They provided very accurate read alignment and thus improved the accuracy and coverage of consensus sequence and SNP calling. They were essential for developing our short indel caller, and for detecting larger structural variants. Our short-insert paired-read data set introduced a new level of resolution in structural variation detection, revealing thousands of variants in a size range not characterized previously. In

©2008 Macmillan Publishers Limited. All rights reserved

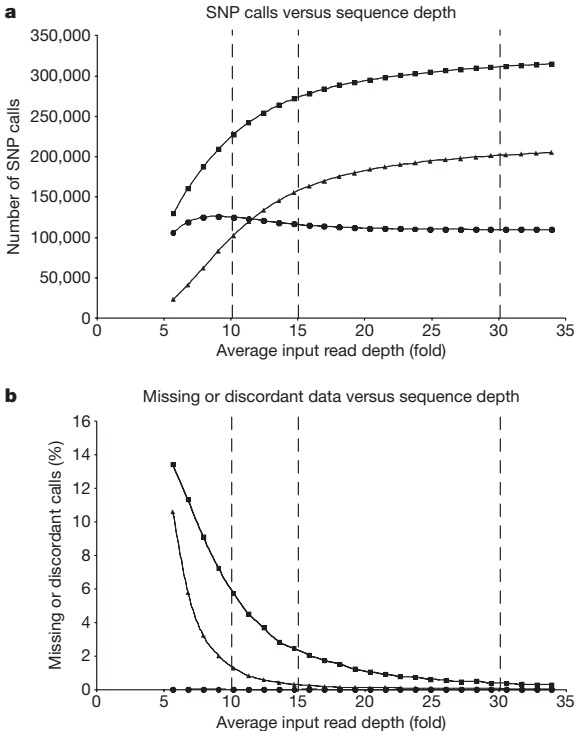


Figure 5 | Effect of sequence depth on coverage and accuracy of human genome sequencing. ELAND alignments were used for this analysis. **a**, Accumulation of sequence-based SNP calls, including all SNPs (squares), heterozygous SNPs (triangles) and homozygous SNPs (circles) with increasing input read depth. **b**, Decrease in genotype positions not covered by sequence (squares), heterozygote under-calls in sequence data relative to genotype data (triangles) and discordant SNP calls compared to genotypes (circles) with increasing input read depth. Vertical dotted lines indicate various input read depths (10×, 15×, 30× haploid genome).

some cases we determined the exact sequence of structural variants by *de novo* assembly from the same paired-read data set. Interpreting events that are embedded in repetitive sequence tracts will require further work.

Massively parallel sequencing technology makes it feasible to consider whole human genome sequencing as a clinical tool in the near future. Characterizing multiple individual genomes will enable us to unravel the complexities of human variation in cancer and other diseases and will pave the way for the use of personal genome sequences in medicine and healthcare. Accuracy of personal genetic information from sequence will be critical for life-changing decisions.

In addition to the large-scale genomic projects exemplified by the present study and others[15,24–26], the system described here is being used to explore biological phenomena in unprecedented detail, including transcriptional activity, mechanisms of gene regulation and epigenetic modification of DNA and chromatin[27–32]. In the future, DNA sequencing will be the central tool for unravelling how genetic information is used in living processes.

METHODS SUMMARY

DNA and sequencing. DNA samples (NA07340 and NA18507) and cell line (GM07340) were obtained from Coriell Repositories. DNA samples were genotyped on the HM550 array and the results compared to publicly available data to confirm their identity before use. Methods for DNA manipulation, including sample preparation, formation of single-molecule arrays, cluster growth and sequencing were all developed during this study and formed the basis for the standard protocols now available from Illumina, Inc. All sequencing was performed on Illumina GA1s equipped with a one-megapixel camera. All purity filtered read data are available for download from the Short Read Archive at NCBI or from the European Short Read Archive (ERA) at the EBI.

Analysis software. Image analysis software and the ELAND aligner are provided as part of the Genome Analyzer analysis software. SNP and structural variant detectors will be available as future upgrades of the analysis pipeline. The Resembl extension to Ensembl is available on request. The MAQ (Mapping and Assembly with Qualities) aligner is freely available for download from http://maq.sourceforge.net.

Data access. Sequence data for NA18507 are freely available from the NCBI short read archive, accession SRA000271 (ftp://ftp.ncbi.nih.gov/pub/TraceDB/ShortRead/SRA000271). X chromosome data are freely available from ERA, accession ERA000035. Links to Resembl displays for chromosome X and human data, plus information on other available data, are provided at http://www.illumina.com/HumanGenome.

See Supplementary Methods for a detailed Methods section.

Received 24 June; accepted 2 October 2008.

1. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
2. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
3. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
4. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).
5. Harris, T. D. *et al.* Single-molecule DNA sequencing of a viral genome. *Science* **320**, 106–109 (2008).
6. Lundquist, P. M. *et al.* Parallel confocal detection of single molecules in real time. *Opt. Lett.* **33**, 1026–1028 (2008).
7. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
8. Milton, J. *et al.* Modified nucleotides. World Intellectual Property Organization WO/2004/018497 (2004).
9. Smith, G. P. *et al.* Modified polymerases for improved incorporation of nucleotide analogues. World Intellectual Property Organization WO/2005/024010 (2005).
10. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
11. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*. doi:10.1101/gr.078212.108 (25 September 2008).
12. The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
13. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nature Genet.* **37**, 727–732 (2005).
14. Korbel, J. O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
15. Campbell, P. J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genet.* **40**, 722–729 (2008).
16. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res.* **30**, 38–41 (2002).
17. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
18. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
19. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
20. The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
21. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
22. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**, 231–238 (1999).
23. Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
24. Hillier, L. W. *et al.* Whole-genome sequencing and variant discovery in *C. elegans*. *Nature Methods* **5**, 183–188 (2008).
25. Hodges, E. *et al.* Genome-wide *in situ* exon capture for selective resequencing. *Nature Genet.* **39**, 1522–1527 (2007).
26. Porreca, G. J. *et al.* Multiplex amplification of large sets of human exons. *Nature Methods* **4**, 931–936 (2007).
27. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
28. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
29. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).

©2008 Macmillan Publishers Limited. All rights reserved

30. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132,** 311–322 (2008).
31. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in *Arabidopsis. Cell* **133,** 523–536 (2008).
32. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5,** 585–587 (2008).
33. Fedurco, M., Romieu, A., Williams, S., Lawrence, I. & Turcatti, G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res.* **34,** e22 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors acknowledge the advice of A. Williamson, T. Rink, S. Benkovic, J. Berriman, J. Todd, R. Waterston, S. Eletr, W. Jack, M. Cooper, T. Brown, C. Reece and R. Cook during this work; E. Margulies for assistance with data analysis; M. Shumway for assistance with data submission; and the contributions of the administrative and support staff at all the institutions. This research was supported in part by The Wellcome Trust (to H.L., A.Sc., K.W., N.P.C, B.N.L., J.R., M.E.H. and R.D.), the Biotechnology and Biological Sciences Research Council (BBSRC) (to S.B. and D.K.), the BBSRC Applied Genomics LINK Programme (to A.Sp. and C.L.B.) and the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health (to N.F.H. and J.C.M.). S. Balasubramanian and D. Klenerman are inventors and founders of Solexa Ltd.

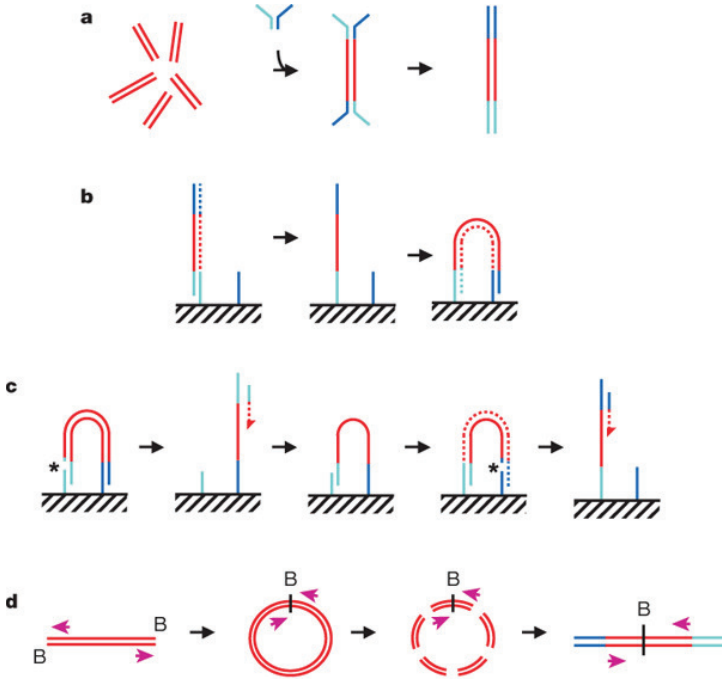
Author Information Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and is freely available to all readers at www.nature.com/nature. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to D.R.B. (dbentley@illumina.com).

David R. Bentley[1], Shankar Balasubramanian[2], Harold P. Swerdlow[1]†, Geoffrey P. Smith[1], John Milton[1]†, Clive G. Brown[1]†, Kevin P. Hall[1], Dirk J. Evers[1], Colin L. Barnes[1,2], Helen R. Bignell[1], Jonathan M. Boutell[1], Jason Bryant[1], Richard J. Carter[1], R. Keira Cheetham[1], Anthony J. Cox[1], Darren J. Ellis[1], Michael R. Flatbush[3], Niall A. Gormley[1], Sean J. Humphray[1], Leslie J. Irving[1], Mirian S. Karbelashvili[3], Scott M. Kirk[3], Heng Li[4], Xiaohai Liu[1,2], Klaus S. Maisinger[1], Lisa J. Murray[1], Bojan Obradovic[1], Tobias Ost[1], Michael L. Parkinson[1], Mark R. Pratt[3], Isabelle M. J. Rasolonjatovo[1], Mark T. Reed[3], Roberto Rigatti[1], Chiara Rodighiero[1], Mark T. Ross[1], Andrea Sabot[1], Subramanian V. Sankar[3], Aylwyn Scally[4], Gary P. Schroth[3], Mark E. Smith[1], Vincent P. Smith[1], Anastassia Spiridou[1], Peta E. Torrance[1], Svilen S. Tzonev[3], Eric H. Vermaas[3], Klaudia Walter[4], Xiaolin Wu[1], Lu Zhang[3], Mohammed D. Alam[3], Carole Anastasi[1], Ify C. Aniebo[1], David M. D. Bailey[1], Iain R. Bancarz[1], Saibal Banerjee[3], Selena G. Barbour[1], Primo A. Baybayan[3], Vincent A. Benoit[1], Kevin F. Benson[1], Claire Bevis[1], Phillip J. Black[1], Asha Boodhun[1], Joe S. Brennan[1], John A. Bridgham[3], Rob C. Brown[1], Andrew A. Brown[1], Dale H. Buermann[3], Abass A. Bundu[1], James C. Burrows[3], Nigel P. Carter[4], Nestor Castillo[3], Maria Chiara E. Catenazzi[1], Simon Chang[3], R. Neil Cooley[1], Natasha R. Crake[1], Olubunmi O. Dada[1], Konstantinos D. Diakoumakos[1], Belen Dominguez-Fernandez[1], David J. Earnshaw[1,2], Ugonna C. Egbujor[1], David W. Elmore[3], Sergey S. Etchin[3], Mark R. Ewan[3], Milan Fedurco[5], Louise J. Fraser[1], Karin V. Fuentes Fajardo[1], W. Scott Furey[2], David George[3], Kimberley J. Gietzen[6], Colin P. Goddard[1], George S. Golda[3], Philip A. Granieri[3], David E. Green[1], David L. Gustafson[3], Nancy F. Hansen[7], Kevin Harnish[1], Christian D. Haudenschild[3], Narinder I. Heyer[1], Matthew M. Hims[1], Johnny T. Ho[3], Adrian M. Horgan[1], Katya Hoschler[1], Steve Hurwitz[3], Denis V. Ivanov[3], Maria Q. Johnson[3], Terena James[1], T. A. Huw Jones[1], Gyoung-Dong Kang[1], Tzvetana H. Kerelska[3], Alan D. Kersey[1], Irina Khrebtukova[3], Alex P. Kindwall[3], Zoya Kingsbury[1], Paula I. Kokko-Gonzales[1], Anil Kumar[1], Marc A. Laurent[6], Cynthia T. Lawley[6], Sarah E. Lee[1], Xavier Lee[3], Arnold K. Liao[3], Jennifer A. Loch[1], Mitch Lok[3], Shujun Luo[3], Radhika M. Mammen[1], John W. Martin[3], Patrick G. McCauley[1], Paul McNitt[3], Parul Mehta[1], Keith W. Moon[3], Joe W. Mullens[3], Taksina Newington[1], Zemin Ning[4], Bee Ling Ng[4], Sonia M. Novo[1], Michael J. O'Neill[3], Mark A. Osborne[1,2], Andrew Osnowski[1], Omead Ostadan[3,6], Lambros L. Paraschos[3], Lea Pickering[1], Andrew C. Pike[1], Alger C. Pike[3], D. Chris Pinkard[3], Daniel P. Pliskin[3], Joe Podhasky[3], Victor J. Quijano[3], Come Raczy[1], Vicki H. Rae[1], Stephen R. Rawlings[1], Ana Chiva Rodriguez[1], Phyllida M. Roe[1], John Rogers[1], Maria C. Rogert Bacigalupo[1], Nikolai Romanov[1], Anthony Romieu[5], Rithy K. Roth[3], Natalie J. Rourke[1], Silke T. Ruediger[1], Eli Rusman[3], Raquel M. Sanches-Kuiper[1], Martin R. Schenker[1], Josefina M. Seoane[3], Richard J. Shaw[1], Mitch K. Shiver[3], Steven W. Short[3], Ning L. Sizto[3], Johannes P. Sluis[3], Melanie A. Smith[1], Jean Ernest Sohna Sohna[1], Eric J. Spence[3], Kim Stevens[1], Neil Sutton[1], Lukasz Szajkowski[1], Carolyn L. Tregidgo[1], Gerardo Turcatti[5], Stephanie vandeVondele[1], Yuli Verhovsky[3], Selene M. Virk[3], Suzanne Wakelin[3], Gregory C. Walcott[3], Jingwen Wang[1], Graham J. Worsley[1], Juying Yan[3], Ling Yau[3], Mike Zuerlein[3], Jane Rogers[4]†, James C. Mullikin[7], Matthew E. Hurles[4], Nick J. McCooke[1]†, John S. West[3], Frank L. Oaks[3], Peter L. Lundberg[3], David Klenerman[2], Richard Durbin[4] & Anthony J. Smith[1]

[1]Illumina Cambridge Ltd. (Formerly Solexa Ltd), Chesterford Research Park, Little Chesterford, Nr Saffron Walden, Essex CB10 1XL, UK. [2]Department of Chemistry, University of Cambridge, The University Chemical Laboratory, Lensfield Road, Cambridge CB2 1EW, UK. [3]Illumina Hayward (Formerly Solexa Inc.), 23851 Industrial Boulevard, Hayward, California 94343, USA. [4]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [5]Manteia Predictive Medicine S.A. Zone Industrielle, Coinsins, CH-1267, Switzerland. [6]Illumina Inc., Corporate Headquarters, 9883 Towne Centre Drive, San Diego, California 92121, USA. [7]National Human Genome Research Institute, National Institutes of Health, 41 Center Drive, MSC 2132, 9000 Rockville Pike, Bethesda, Maryland 20892-2132, USA. †Present addresses: The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK (H.P.S.); Oxford Nanopore Technologies, Begbroke Science Park, Sandy Lane, Kidlington OX5 1PF, UK (J.M., C.G.B.); BBSRC Genome Analysis Centre, John Innes Centre, Norwich Research Park, Colney, Norwich NR4 7UH, UK (J.R.); Pronota, NV, VIB Bio-Incubator, Technologiepark 4, B-9052 Zwijnaarde/Ghent, Belgium (N.J.M.).

©2008 Macmillan Publishers Limited. All rights reserved

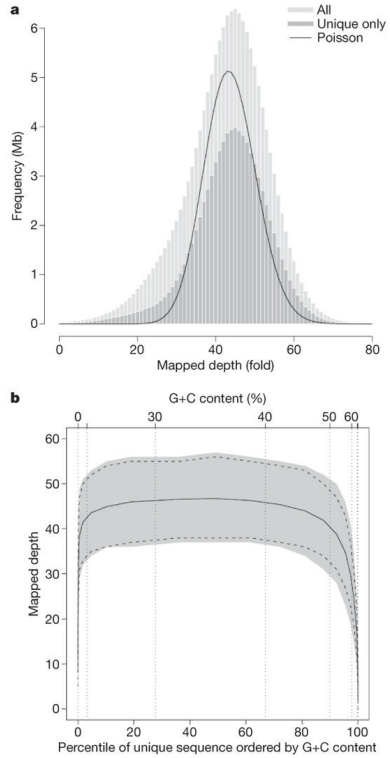
Preparation of samples.



DR Bentley of the paper *et al. Nature* **456**, 53-59 (2008) doi:10.1038/nature07484

nature

X chromosome data.



DR Bentley of the paper *et al. Nature* **456**, 53-59 (2008) doi:10.1038/nature07484

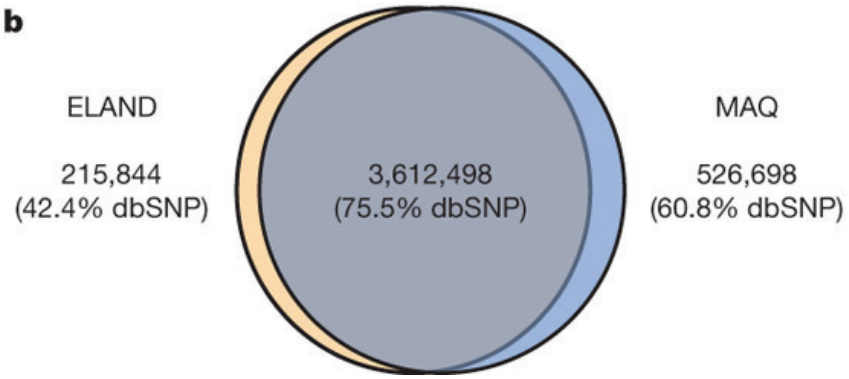
nature

SNPs identified in the human genome sequence of NA18507.

a

Call	ELAND SNPs (*n*)	ELAND In dbSNP (%)	MAQ SNPs (*n*)	MAQ In dbSNP (%)
Homozygote	1,417,320	90.1	1,503,420	90.8
Heterozygote	2,411,022	63.9	2,635,776	63.8
All	3,828,342	73.6	4,139,196	73.6

b



DR Bentley of the paper *et al. Nature* **456**, 53-59 (2008) doi:10.1038/nature07484

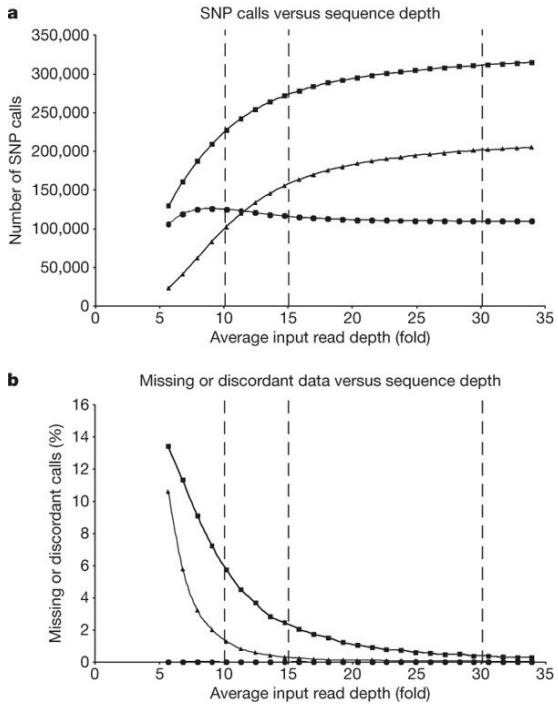
nature

Homozygous complex rearrangement detected by anomalous paired reads.

DR Bentley of the paper *et al. Nature* **456**, 53-59 (2008) doi:10.1038/nature07484

nature

Effect of sequence depth on coverage and accuracy of human genome sequencing.



DR Bentley of the paper *et al. Nature* **456**, 53-59 (2008) doi:10.1038/nature07484

nature

Experimental methods

Preparation of DNA

DNA for a bacterial artificial chromosome (bCX98J21) clone was prepared from a single colony by standard alkaline extraction method and purified by equilibrium density gradient fractionation using standard procedures. X chromosome DNA was prepared by flow-sorting X chromosomes from a culture of human lymphoblastoid cell line GM07340. Cells were arrested in metaphase using demecolcine, harvested, treated with hypotonic solution, transferred to polyamine isolation buffer for lysis. Chromosomes were harvested, stained overnight with Hoechst and Chromomycin A3 (both Sigma) and analysed on a flow cytometer (MoFlo®, DAKO) equipped with two Innova 300 series lasers (Coherent). Approximately 1.3 million chromosomes were collected and treated with proteinase K and sodium lauroyl sarcosine followed by phenylmethylsulphonyl fluoride, before recovering DNA by precipitation[34].

Samples of human genomic DNA NA07340 and NA18507 were obtained from Coriell Repositories, Camden NJ, USA.

Construction of short insert single- and paired-end libraries

Purified DNA (50 ng – 5 ug) was fragmented (6 minutes at 32 psi) through a disposable nebulizer (Invitrogen) and purified on a single Qiaquick column (Qiagen). Recovered DNA was repaired using a cocktail of T4 DNA polymerase, DNA polymerase (large fragment) and T4 polynucleotide kinase (all enzymes from New England Biolabs) and then treated with Klenow fragment (3' to 5' exo$^-$) in the presence of 0.2 mM dATP to add a dAMP to the 3' ends of the fragments. A single adaptor was ligated to the ends of the DNA. Adapters for the preparation of single read libraries comprised the oligonucleotides: 5'ACACTCTTTCCCTACACGACGCTCTTCCGATC$_x$T (x = phosphorothioate bond) and 5'-phosphate-GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG. The adapters for the preparation of paired read libraries comprised an alternative set of two oligonucleotides: 5'ACACTCTTTCCCTACACGACGCTCTTCCGATC$_x$T (x = phosphorothioate bond) and 5'-phosphate-GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG. Ligation products in the

Illumina Exhibit 1006, Page 13 of 67

desired size range were gel-purified and amplified with 6-18 cycles of PCR using Phusion High-Fidelity system (New England Biolabs) and the appropriate primers: for single read libraries: 5'-CAAGCAGAAGACGGCATACGAGCTCTTCCGATC$_x$T and 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC$_x$T; and for paired-read libraries: 5'-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC$_x$T and 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC$_x$T (x = phosphorothioate bond resistant to excision by 3'-5' exonucleases). These primers include the sequences that anneal to the complementary oligonucleotides bound to the flowcell surface, plus the sequencing primer sites. Samples were diluted to a concentration of 10 nM in 10 mM Tris pH 8.5 and 0.1% Tween 20 prior to cluster formation (see below).

Construction of long range paired end libraries

Purified DNA (10 ug) was fragmented (30 seconds at 7.5 psi) through a disposable nebulizer (Invitrogen). The recovered DNA was end repaired and the 3' ends labelled using a mixture of biotin-dNTPs (PerkinElmer), dNTPs and end repair enzymes (as described above). 2 kb DNA fragments were gel purified and circularised using T3 DNA ligase (Enzymatics). The linear DNA was removed using Plasmid Safe (Epicentre) and the circular DNA fragmented by nebulization (6 minutes at 32 psi). The recovered DNA was end repaired, dAMP tailed and the ends of the DNA ligated to the paired-read adapters (see above). Biotinylated fragments were purified on M280 streptavidin beads (Dynal). The DNA was amplified from the beads using Phusion High-Fidelity system (New England Biolabs) and the appropriate primers for paired-read libraries (as described above) (see fig 1d). Library fragments in the desired size range (400-600 bp) were gel purified and diluted for cluster formation.

Preparation of flowcells

Glass 8-channel flow cells (Silex Microsystems, Sweden) were thoroughly washed and then coated for 90 min at 20°C with 2% acrylamide containing ~3.9 mg/ml *N*-(5-bromoacetamidylpentyl) acrylamide, 0.85 mg/ml tetramethylethylenediamine (TEMED) and 0.48 mg/ml potassium persulfate ($K_2S_2O_8$). Flow cell channels were

Illumina Exhibit 1006, Page 14 of 67

rinsed thoroughly before further use. The coated surface was then functionalised by reaction for 1 hour at 50°C with a mixture containing 0.5 uM each of two priming oligonucleotides in 10 mM potassium phosphate buffer pH 7; either A and B for single read experiments or C and D for paired read experiments (see below for sequences). Grafted flow cells were stored in 5xSSC until required.

Cluster creation for single read experiments

Cluster creation was carried out using an Illumina Cluster Station. To obtain single stranded templates, adapted DNA was first denatured in NaOH (to a final concentration of 0.1M) and subsequently diluted in cold (4 °C) hybridisation buffer (5x SSC + 0.05 % Tween 20) to working concentrations of 2 – 4 pM, depending on the desired cluster density / tile. 85 ul of each sample were primed through each lane of a flowcell at 96°C (60 ul / min). The temperature was then slowly decreased to 40 °C at a rate of 0.05 °C/sec to enable annealing to complementary adapter oligonucleotides immobilised on the flowcell surface (oligo 'A': 5´-PS-TTTTTTTTTT-(diol)3-AATGATACGGCGACCACCGA-3´; oligo 'B': 5´-PS-TTTTTTTTTTCAAGCAGAAGACGGCATACGA-3´). Hybridised template strands were extended using Taq polymerase to generate their surface-bound complement. The samples were then denatured using formamide to remove the initial seeded template. The remaining single stranded copy was the starting point for cluster creation. Clusters were amplified under isothermal conditions at 60 °C for 35 cycles using Bst polymerase for extension and formamide for denaturation during each cycle (see fig 1b). Clusters were washed with storage buffer (5x SSC) and either stored at 4 °C or used directly.

Cluster creation for paired read experiments

Paired read flowcells contained the two oligonucleotides
oligo 'C': 5'-PS-TTTTTTTTTTAATGATACGGCGACCACCGAGAUCTACAC-3'
(U = 2-deoxyuridine) and
oligo 'D': 5'-PS-TTTTTTTTTTCAAGCAGAAGACGGCATACGAGoxoAT-3',
(Goxo = 8-oxoguanine) immobilised on the surface in a ratio C:D = 1:1. Other than the use of a paired-end specific library, cluster creation was the same as described above.

Illumina Exhibit 1006, Page 15 of 67

Processing of clusters for single read experiments

Linearisation of surface-immobilised complementary oligonucleotide 'A' was achieved by incubation with linearization mix (100 mM sodium periodate, 10 mM 3-aminopropan-1-ol, 20 mM Tris pH 8.0, 50 % v/v formamide) for 20 minutes at 20 °C followed by a water wash. All exposed 3'-OH termini of DNA, either from the extended template or unextended surface oligonucleotides were blocked by dideoxy chain termination using a terminal transferase and ddNTPs. Linearised and blocked clusters were denatured with 0.1M NaOH prior to hybridisation of the sequencing primer (see fig 1c). Processed flowcells were transferred to the Illumina Genome Analyser for sequencing.

Processing of clusters for paired read experiments

For read 1, linearisation of surface immobilised oligonucleotide 'C' to retain strand 1 of each cluster was achieved by incubation with USER enzyme (as shown above). After blocking, clusters were denatured with 0.1M NaOH prior to hybridisation of the read 1 specific sequencing primer (5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCT -3'). Processed flowcells were transferred to the Illumina Genome Analyser for sequencing.

Following the successful completion of sequencing of read 1 on the Genome Analyser, flowcells remained mounted and were automatically prepared for read 2 *in situ* using the Illumina Paired End module (according to operating manual). Clusters were denatured with 0.1 M NaOH to remove the products of read 1. Clusters were 3'-dephosphorylated using T4 polynucleotide kinase, and the strand that had been linearised as part of the read 1 preparation was re-synthesized isothermally as previously described for cluster creation. Linearisation to remove strand 1 of the re-synthesised clusters was achieved by the excision of 8-oxoguanine from oligo 'D' using Fpg (formamidopyrimidine DNA glycosylase, New England Biolabs). Linearised and blocked clusters were denatured with 0.1M NaOH prior to hybridisation of the read 2 specific sequencing primer (5'-CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT -3')(see fig 1c).

Sequencing on the Genome Analyser.

Illumina Exhibit 1006, Page 16 of 67

All sequencing runs were performed as described in the Illumina Genome Analyser operating manual. Flowcells were sequenced using standard recipes (see User Guide) in order to generate 25 and 35 base single and paired reads. Typically a single read run producing 1-2 Gb of PF data required 72 hours; paired read runs required approximately 150 hours including the time taken for automated preparation of the template for the second read.

Data Analysis

Image analysis

The image analysis program (Firecrest) first identifies the position of the DNA clusters on the images taken from the first sequencing cycle. Each initial image was band-pass filtered to remove background fluorescence and large-scale structure on the image, as well as enhance the signal-to-noise. Cluster positions were identified from a search for local maxima on the filtered image. Because of the finite accuracy of the movements of the motion stage, images taken at different sequencing cycles have random translational offsets with respect to each other. Furthermore, images taken in different frequency channels have different optical paths and wavelengths and experience further, albeit smaller, translations and scale transformations. In order to correct for the image shifts and scalings, the cluster positions that were extracted from the four images taken in the first cycle were super-imposed to construct a 'reference image' containing all detected clusters. Transformations of the image coordinates to later cycles were then obtained from a cross-correlation of the images in later cycles to the reference image. In this way, we obtained a set of four intensity measurements for each cluster and sequencing cycle. These series of intensities for each cluster are analogous to the intensity traces from Sanger sequencing. In addition to estimates of the intensities, the image analysis also extracts an estimate of the local noise or image background dispersion around the cluster for each image.

Base calling

The signals detected for the four different dye-labelled dNTPs are not independent, as the emission spectra of their dyes and the transmission and detection frequency windows may overlap between nucleotides. The relative intensities and cross-talk are described by a frequency cross-talk matrix, which characterises the intensity response

Illumina Exhibit 1006, Page 17 of 67

of the system to each nucleotide. We developed a method to auto-calibrate this matrix from the intensity traces and apply a correction to the extracted intensities. Because the rates of phasing and prephasing were small and consistent, the resulting intensity dissipation into different frequency cycles was also small and the accumulated loss roughly linear. Therefore we estimated phasing and prephasing rates by measuring the build-up of correlated signal between different cycles over early sequencing cycles. From these rates we derived the expected correlation of signals for each cycle and de-correlated them. The end result of these computations was a set of matrix-corrected, phasing-corrected intensity values for each cluster, from which we took whichever of the four channels gave the highest value at a given cycle to be the base call for that cycle.

Purity filtering

In order to discriminate between good reads without errors and reads derived from mixed clusters that overlap their nearest neighbours, we defined a measure of signal purity at a given cycle by taking the brightest of the four corrected intensities at that cycle as a fraction of the sum of the brightest and next brightest intensities. We discarded all reads whose corrected brightest intensity in any of the first 12 sequencing cycles was less than 60% of the sum of brightest intensity and the next brightest. We find that this criterion provides reasonable discrimination between good and bad data. Depending on the loading density, we typically kept between 50% and 70% of the raw reads. All figures quoted for accuracy and yield per flowcell refer only to this purity-filtered subset of the raw data.

Quality scoring of bases

The base caller (Bustard) provides a first estimate of the uncertainty of the base call. This is computed by propagating the noise estimates from the image analysis and integrating the resulting likelihood functions to obtain probability estimates for each of the four possible base calls. The probability estimates are transformed to scores by first converting to a log-odds scale via the formula $Q=10\log_{10}(p(X)/(1-p(X))$, where p(X) for A,C,G,T is the estimated probability of the base call being X, and then rounding to integers. This scoring scheme can be thought of as a generalization of the scoring scheme made popular by the Phred base caller[10] ($Q=-10\log10(p_{error})$, where p_{error} is the probability of an incorrect base), in that the highest of the four scores - that of the called base - is

Illumina Exhibit 1006, Page 18 of 67

asymptotic to the Phred score, but the scheme also enables meaningful integer scores to be assigned to the three non-called bases.

This initial base quality estimate is refined by an implementation of the Phred algorithm, taking as predictors the initial confidence score for the called base together with the sequencing cycle, the purity of the called base and the minimum purity over the first 12 bases of the read. The sequence data for each flowcell lane in an experiment can be used as a training set for the same lane, with the alignment to the reference being used to determine whether or not a base is correct. Since a lane may produce several hundred million base pairs, each lane contains enough data to serve as a reasonable training set. After this procedure, the observed error rate for each base closely matches the error rate implied by the adjusted quality score (fig S3).

This "auto-calibration" procedure was used to estimate the base quality for the X chromosome and Yoruba genome analyses. However, we note that this method will underestimate the true quality of the bases since genuine differences between the sample DNA and the reference are treated as errors by the algorithm. One way of ameliorating this effect would be to exclude from the auto-calibration all reads that overlap positions known to be highly variable in the genome being sequenced – in human DNA, dbSNP positions would be obvious candidates. An alternative method is to allow one lane of the flowcell to serve as a control by using it to sequence DNA from a sample whose reference is exactly known (we have used DNA from the bacteriophage PhiX 174RF1 for this purpose). The Phred calibration table obtained from this data is then used to recalibrate the quality scores of the bases coming from the unknown samples in the other lanes. By breaking the dependency on alignments, this method has the advantage of removing the need for any knowledge at all of the reference sequence of the sample(s) in the non-control lanes, albeit at a cost of a lane's worth of data.

PhageAlign alignment

PhageAlign is a program for exhaustive alignment of reads of length k against a known reference. Both strands of the reference sequence are split into overlapping k-mers which are then sorted lexicographically. The reads are also sorted and then compared to each genomic k-mer in turn allowing any number of substitution errors.

Illumina Exhibit 1006, Page 19 of 67

Commonality of prefixes between lexicographically adjacent prefixes is exploited to minimise the number of base to base comparisons required, however its exhaustive nature renders it too slow for high-throughput alignment of datasets where either the number of reads or the size of the reference genome is large.

The principal use of PhageAlign is to enable an accurate measure of raw read error by allowing even noisy reads to participate in the error rate calculation, provided the aligner is able to find a unique best match for them in the reference.

ELAND alignment

In order to remap the sequence reads to a large reference genome, we developed a fast short-read alignment program called ELAND. The first few bases (32 by default, or the entire read length for reads shorter than this) are aligned to the genome allowing up to two substitution differences to obtain a set of candidate match positions for each read in the reference. Sequence from each of these positions is then used to extend the candidate alignments along the full length of each read. Finally base quality values are used to choose, where possible, the most probable of the candidate alignments.

For paired reads, a set of candidate alignments is obtained for each of the two reads as described above. Read pairs having a unique alignment of each read are first used to determine the nominal strand orientation and insert size distribution of the sample then, on a second pass, this information is used to resolve repeats and determine the anomalously paired reads that are possible indicators of structural variation.

ELAND SNP calling

For allele calling, we used only read pairs having alignments that were correctly oriented and that indicated a template insert size of within 3 standard deviations of the sample median. We further required a paired alignment score ≥6 (indicating the quality of the paired mapping). The basecalls and their associated quality values were sent to a Bayesian allele caller, which produced one or two allele calls and scores for each position in the genome. At each position, the allele caller computes log10 p(observed bases | no “A”s are present) and similarly for C,G and T. The highest two scores are then normalized by subtracting the third highest, thus obtaining log-odds scores for the two most probable alleles for which an increment in score of 3

Illumina Exhibit 1006, Page 20 of 67

approximately corresponds to an increase in coverage of a single base of Phred quality 30. SNPs were called where a non-reference base allele was observed, the allele call score was ≥10, and the depth at this position was no greater than three times the chromosomal mean. For heterozygous calls, we additionally required both alleles to have an allele-call score ≥10 and the ratio of their scores to be ≤3. For the reduced depth analysis (fig 5) we required an allele-call score ≥6. We excluded SNP calls that were within 15 bp of an apparent small indel (detected as described below).

ELAND structural variant detection

Hierarchical clustering of anomalous readpairs was used to identify groupings of five or more readpairs that had a similar size and position. Read pairs were defined as anomalous if they had high-confidence alignments of each individual read that nevertheless were either incorrectly oriented or implied an insert size of at least 3 standard deviations outside the sample median. These groupings were combined with other information such as depth changes, alignability and gaps in expected coverage, and a ranking system was applied. Higher (positive) ranks were assigned where the event supporting evidence was seen; negative ranks where used for regions where it would be difficult to call variants, such as the centromere, or where contradictory evidence was seen.

Some structural variants were characterised using local *de novo* assembly. We took high quality (23 of the first 25 bases had ≥Q20) anomalous pairs, singletons and their non-aligning partner in the region of interest and attempted an assembly using Velvet[35]. Contigs were aligned back to the chromosome using BLAST, in order to look for discontinuities in the alignment, indicating breakpoints.

MAQ alignment

MAQ first searches for the ungapped match with lowest mismatch score, defined as the sum of qualities at mismatching bases. To speed up the alignment, MAQ only considers positions that have 2 or fewer mismatches in the first 28bp.
Sequences that fail to reach a mismatch score threshold but whose read pair is mapped are searched with a gapped alignment algorithm in the regions defined by the read pair. To evaluate the reliability of alignments, MAQ assigns each alignment a Phred-scaled quality score which measures the probability that the true alignment is not the

Illumina Exhibit 1006, Page 21 of 67

one found by MAQ. MAQ always reports a single alignment, and if a read can be aligned equally well to multiple positions, MAQ will randomly pick one position and give it a mapping quality zero.

MAQ fully utilizes the read-pair information of paired reads. It is able to use this information to correct wrong alignments, to add confidence to correct alignments, and to accurately map a read to repetitive sequences if its mate is confidently aligned. With paired-end reads, MAQ also finds short insertions/deletions (indels) from the gapped alignment described above.

Calculation of Mapped read depth and distribution

For NA07340-X: After aligning the data to the reference sequence for chromosome X, the depth of mapped reads was sampled at every 50th position. The distribution of this depth is shown by the 'all' histogram in fig 2a. Then all the positions from this sample were discarded where the reference is not unique on the scale of the read length (as determined by mapping the reference to itself). The distribution of these uniquely mapped reads is shown by the 'unique' histogram in fig 2a. Comparison with the Poisson distribution having the same mean shows that there is some extra variance or overdispersion relative to the theoretical minimum.

At each of the unique positions the GC content of the reference in a surrounding window of length twice the read length was calculated. This gives an estimate of the GC content of all the reads that could have overlapped that position. Then we binned the positions by GC content, and within each bin calculated the mean depth and the 10th and 90th centiles of both the depth and a Poisson distribution with the same mean. The resulting depth-GC variation is shown in fig 2b, where the lower x-axis represents the proportion of unique reference sequence corresponding to the GC content values on the upper axis. In this plot, overdispersion at a given GC content is indicated by the Poisson 10th and 90th centile lines lying inside the shaded area at that GC value.

MAQ SNP calling

MAQ produces a consensus genotype sequence from the alignment. The consensus sequence is inferred from a Bayesian statistical model and each consensus genotype is

Illumina Exhibit 1006, Page 22 of 67

associated with a Phred quality which measures the probability that the consensus genotype is incorrect. Potential SNPs are detected by comparing the consensus sequence to the reference and are further filtered by a set of predefined rules. These rules are:

i) discard SNPs within the 3bp flanking region around a potential indel;
ii) discard SNPs covered by three or fewer reads;
iii) discard SNPs covered by no read with a mapping quality higher than 60;
iv) in any 10bp window, if there are 3 or more SNPs, discard them all;
v) discard SNPs with consensus quality smaller than 20; and
vi) discard a SNP if a base with consensus quality lower than 20 occurs within 3bp on either side of the target SNP.

MAQ small indel detection

MAQ regards an indel is reliable, if at least three reads contain the exact indel (identical position and indel size). MAQ only keeps one most evident indel in any 10bp window because close indels may indicate alignment artefacts.

Genome-wide de novo assembly of unaligned reads

De novo assembly was performed using Velvet[35] and unaligned read pairs. The read pairs were duplicate filtered and then further quality filtered to obtain the best 11,067,318 reads, to allow for computer memory limitations. Version 0.5.05 of velvet was used with a hash length of 23, a coverage cutoff of 5 and a maximum insert length of 350. The contigs were then filtered to ensure a minimum length of 100 bases.

MAQ structural variant detection

Anomalous read pairs with a mapping quality of at least 20 were ordered first by start position, then two anomalous pairs were allocated to the same cluster if they were overlapping and if their end positions were not further apart than a given threshold (mean of the insert lengths plus three times their standard deviation). This procedure was followed by merging neighbouring overlapping clusters. To obtain a final set of putative deletions, these candidate clusters were filtered for the number of read pairs per cluster (at least 5), for distance between leftmost and rightmost forward reads (distance is less than a given threshold, i.e. mean of the insert lengths plus three times their standard deviation), similarly for distance between leftmost and rightmost

Illumina Exhibit 1006, Page 23 of 67

reverse reads and for deletion size (greater than a given threshold, i.e. mean of the insert lengths plus three times their standard deviation); also read depth and repeat structures were considered. Putative deletions greater than 100kb were removed.

Mapped read depth was used to infer copy-number variants between the sample (NA07340) and reference sequences. At any given sequence position the depth of reads aligned by Maq is expected to be Poisson distributed, with mean determined by the copy numbers of both the sample and the reference at that position, the mean overall depth of coverage, and any GC or other biases. We constructed a Hidden Markov model with hidden states representing copy number differences between sample and reference and an emission or observed variable representing the mapped depth accounting for GC content. (We use a negative binomial distribution rather than a Poisson for the latter, to account for the overdispersion evident in fig 2b.) Standard HMM methods (e.g. ref 36) were then applied to infer the most probable sequence of copy-number states in the sample given the depth data.

Resembl

Resembl is an extended version of Ensembl developed at Illumina as part of a collaborative project with EBI, Sanger and Imperial College under a UK government funded LINK framework grant. It allows storage, query and viewing of Illumina re-sequencing data in a genomic context. The human X chromosome (NA07340) and the whole human genome (NA18507) re-sequencing datasets from ELAND-based alignment data were loaded into Resembl databases and Resembl websites were used for interactive data mining and QC within Illumina.

The Resembl back-end database was designed to allow storage and retrieval of the large amounts of re-sequencing data in an efficient way. It supports paired-end alignments at high coverage, as well as per-base and summary-type data on coverage and alignability. Parsing scripts were written to pre-process sort.txt files from the build process and adapted to run on a Linux clustered environment. Import scripts were written for very large-scale data loading. A system of clustered indexes was developed to allow fast retrieval of data from multi billion record tables.

Illumina Exhibit 1006, Page 24 of 67

The website extensions to the Ensembl browser allow visualization of re-sequencing data in a genomic context and support easy navigation back to the raw data. Extensions were made to Ensembl's Karyotype, Map and Contig View, and a new view (called ReadView) was added for closer examination of reads. Resembl websites were set-up for browsing the X and whole human genome paired-end alignments, SNPs and structural variations, as well as coverage and alignability graphical plots. . Paired alignments are categorised into Regular and Anomalous (i.e. anomalous-gapped, misoriented, chimeras, singletons) and displayed in tracks, accordingly. Suitable colour coding is used to distinguish the different types of variation. The displayed paired alignments are also filtered according to the rules used during the pipeline analysis. Candidate structural variants are highlighted by peaks in a graphical plot within ContigView. SNPs produced from the pipeline are loaded in Resembl as a user track. All the extensions were seamlessly integrated within Ensembl, and were flexibly implemented as a plugin.

34 Ng, B.L. and Carter, N.P., Factors affecting flow karyotype resolution *Cytometry A* **Vol** (9), 1028 (2006).

35 Zerbino D.R. and Birney, E. , Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs." *Genome Res* **18**(5): 821-9. (2008).

36 Rabiner L.R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**(2): 257-286. (1989).

Illumina Exhibit 1006, Page 25 of 67

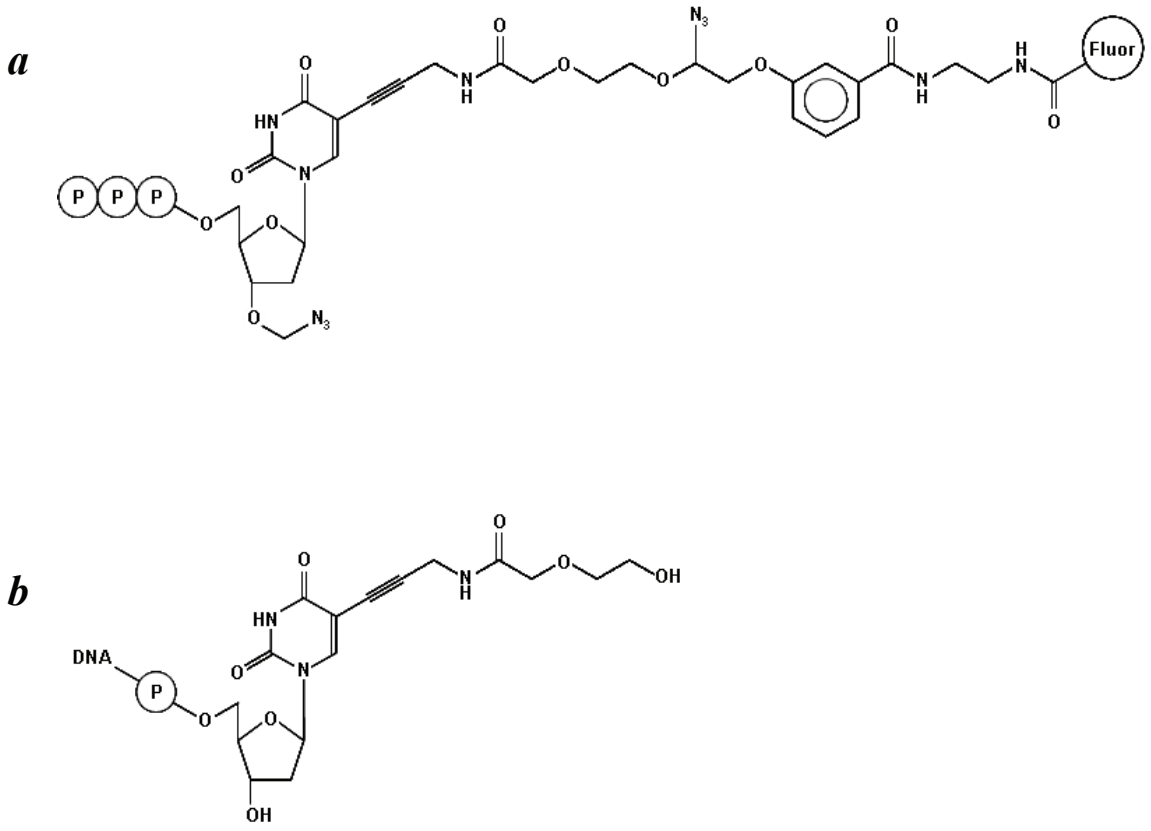


Figure S1. *a*. Structure of the reversible terminator 3'-*O*-azidomethyl 2'-deoxythymine triphosphate (T) labelled with a removable fluorophore. *b*. Structure of the incorporated nucleotide after removal of the fluorophore and terminator group. Each of the four nucleotides have an equivalent structure to the one shown here, except for the different base and a corresponding base-specific fluor.

Illumina Exhibit 1006, Page 26 of 67

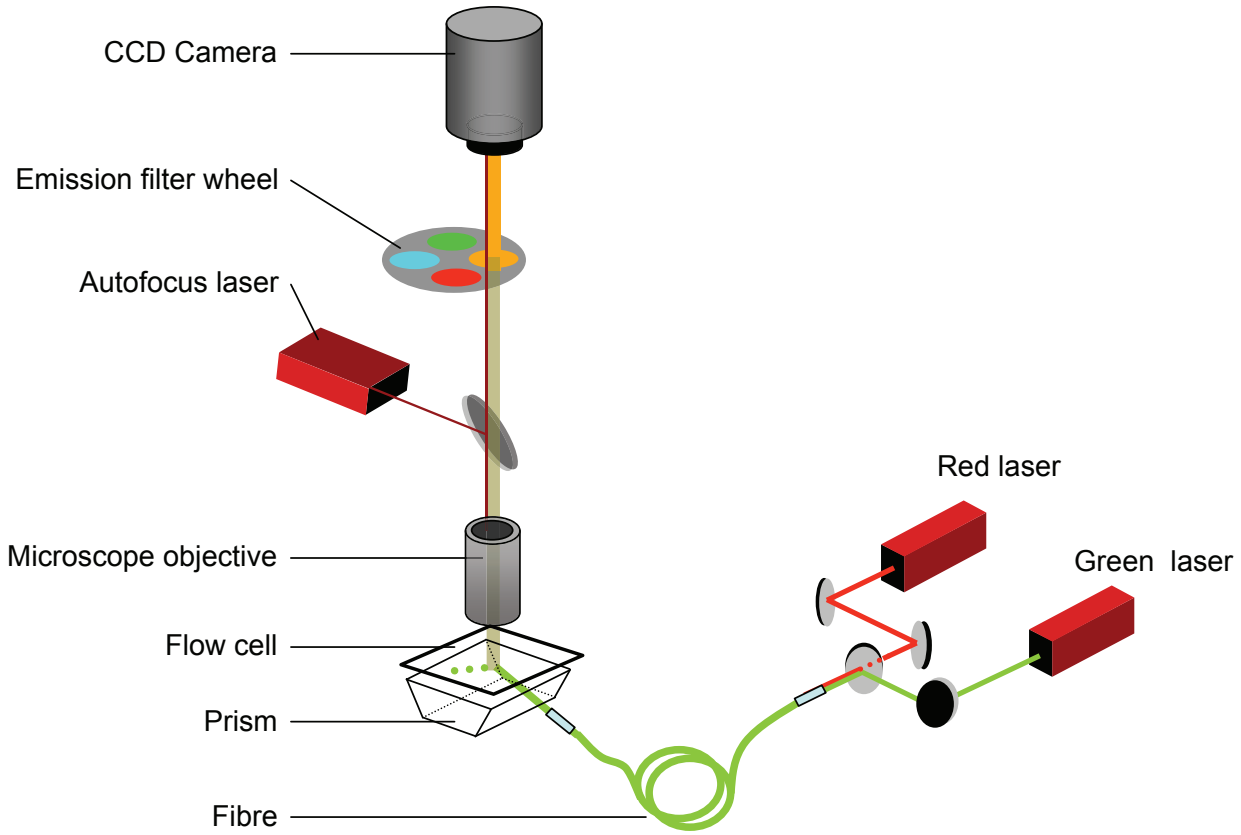


Figure S2*a*. Optical path of the Genome Analyser (GA1). Red (660 nm) and green (532 nm) lasers provide excitation beams that are directed along an optical fibre and through a prism which is in contact with the flow cell. Excitation of fluorescent nucleotides incorporated into DNA clusters on the inner surface of the flow cell (see fig S2b) leads to a base-specific emission that passes through an objective and a filter wheel and the signal is collected by a CCD camera. Autofocus utilises a third laser (635 nm) that is projected through the objective onto the flowcell (see fig S2c for details). Images are collected sequentially for each square 0.33 mm x 0.33 mm 'tile' on the flowcell surface (see fig S2d).

Illumina Exhibit 1006, Page 27 of 67

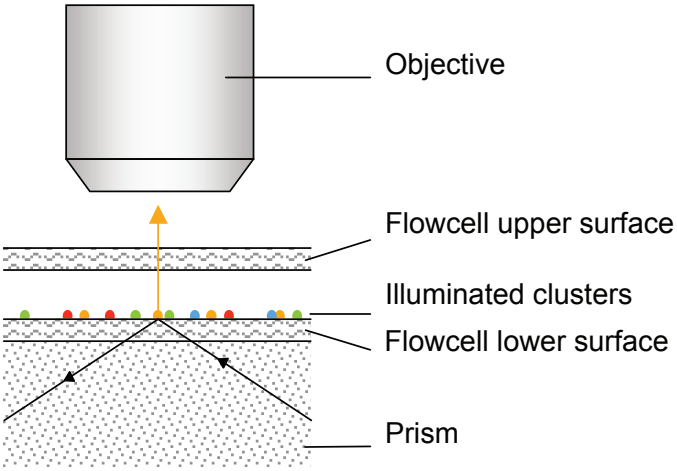


Figure S2b. Total internal reflection of the incident excitation beam at the glass-buffer interface generates an evanescent wave that excites the clusters on the surface. The fluorescence emission is captured by a custom made microscope objective, passed through a filter and is then projected onto a CCD. The evanescent wave excitation technique maximises the sensitivity of signal detection while minimising background noise.

Illumina Exhibit 1006, Page 28 of 67

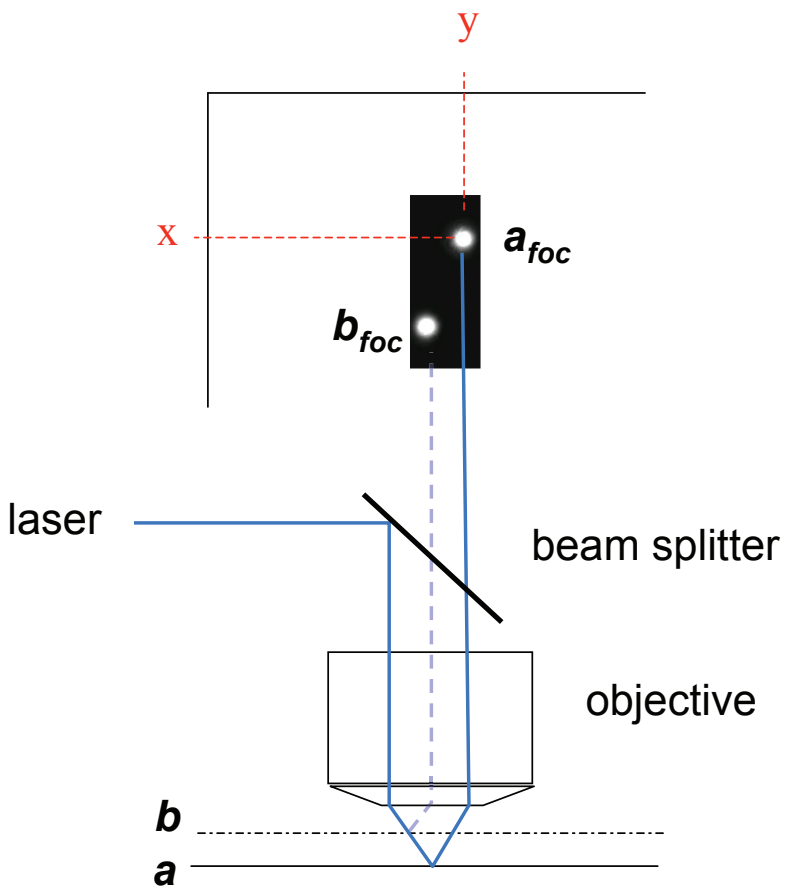


Figure S2c. Autofocus principle of operation. A reference laser beam (solid blue line) is deflected by a beam splitter along a path parallel to the optic axis of the microscope objective and is reflected from the surface of interest. Reflection from a focus plane ***a*** generates an autofocus spot $\boldsymbol{a}_{foc}$. Focus is checked manually at the beginning of the run, after which the autofocus set-up registers the correct (x,y) co-ordinates for the autofocus spot relative to the boundaries of the tile. If the surface of interest is out of focus in a subsequent cycle (as in ***b***), the laser beam follows a different path (blue dashed line) and the autofocus spot is displaced at $\boldsymbol{b}_{foc}$ Focus is automatically adjusted (z movement) as required to return the autofocus spot to the pre-registered correct position before image capture.

Illumina Exhibit 1006, Page 29 of 67

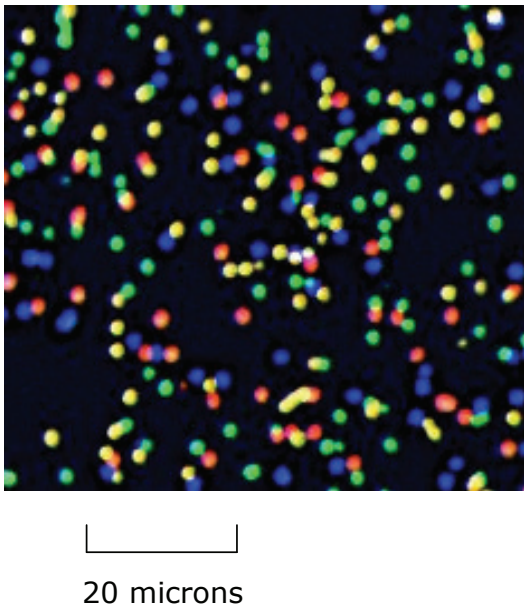


Figure S2d. Imaging clusters during the sequencing process. Part of the image of a tile with a low density of clusters is shown. Four images, one for each emission wavelength, have been artificially coloured and superimposed to show the four-colour detection system that provides the raw signal intensity for calling the individual bases.

Illumina Exhibit 1006, Page 30 of 67

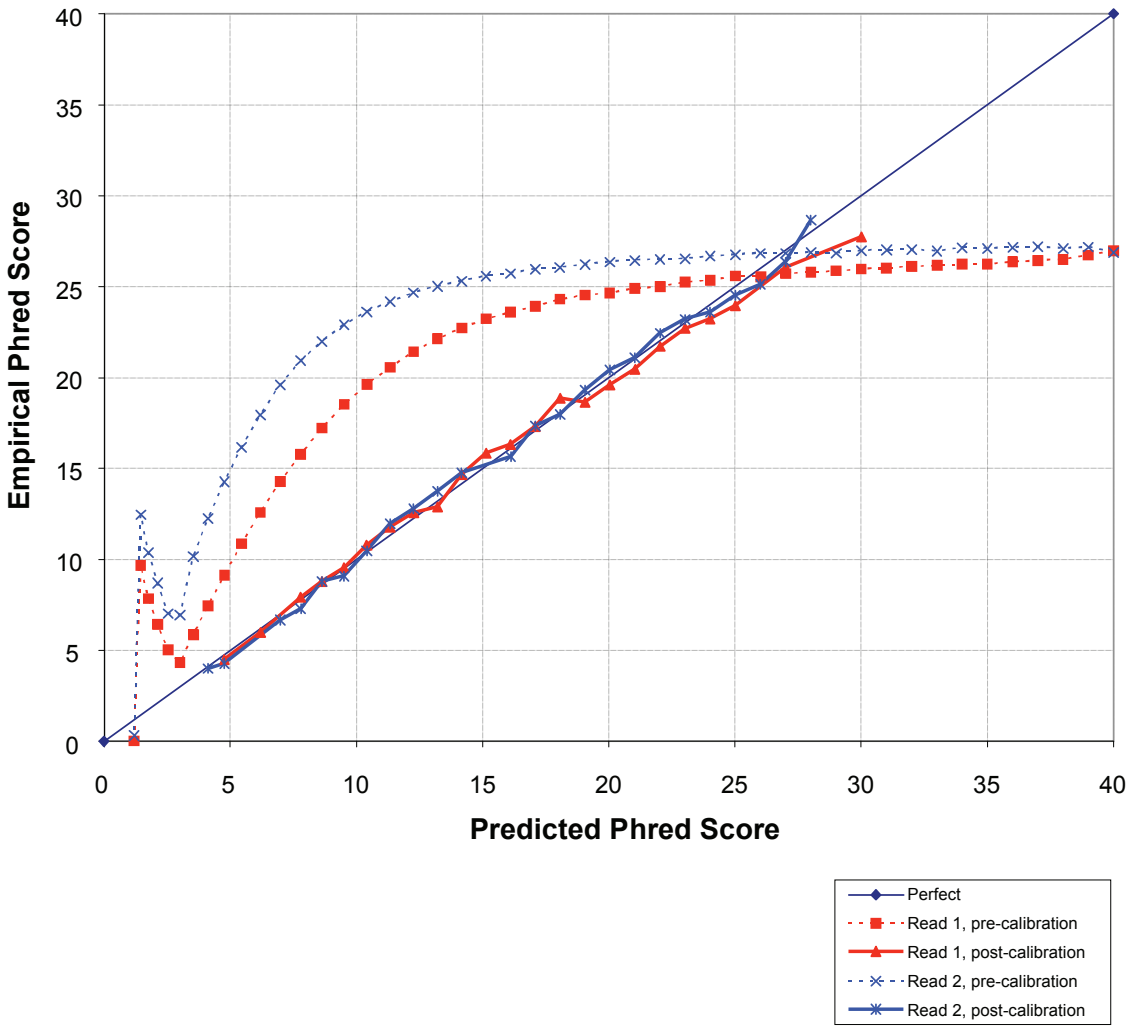


Figure S3. Effect of auto-calibration procedure on base quality scores (see supplementary methods for details). The data set is one lane of paired-read data from the NA18507 individual comprising 8,960,643 purity-filtered clusters, two 35-base reads being obtained from each cluster. Dotted lines show the original quality estimates obtained from the base caller for each read, the solid lines show the adjusted quality scores generated by the auto-calibration procedure. The probability estimates have been re-mapped from the log-odds scoring scheme in which they were originally expressed to the Phred scoring scheme of Q=-10log10(perror), where perror is the probability of an incorrect base. No attempt was made to exclude genuine differences between the sample DNA and the human reference sequence from the calibration procedure, so the calibrated quality scores likely represent a lower bound on the true quality of the data set.

Illumina Exhibit 1006, Page 31 of 67

a 080531_EAS192_0038_FC20GG5 – lane 5

	Lane Info		Tile Mean +/- SD for Lane							
Read	Lane	Lane Yield (kbases)	Clusters (raw)	Clusters (PF)	1st Cycle Int (PF)	% intensity after 20 cycles (PF)	% PF clusters	& Align (PF)	Alignment Score (PF)	% Error Rate (PF)
1	5	297306	38689 +/- 3121	25740 +/- 1646	4211 +/- 867	87.22 +/- 11.11	66.67 +/- 2.40	91.29 +/- 0.20	53.86 +/- 1.44	0.57 +/- 0.07
2	5	297306	38689 +/- 3121	25740 +/- 1646	2274 +/- 504	79.47 +/- 9.69	66.67 +/- 2.40	89.80 +/- 0.40	49.65 +/- 1.78	0.69 +/- 0.09

b

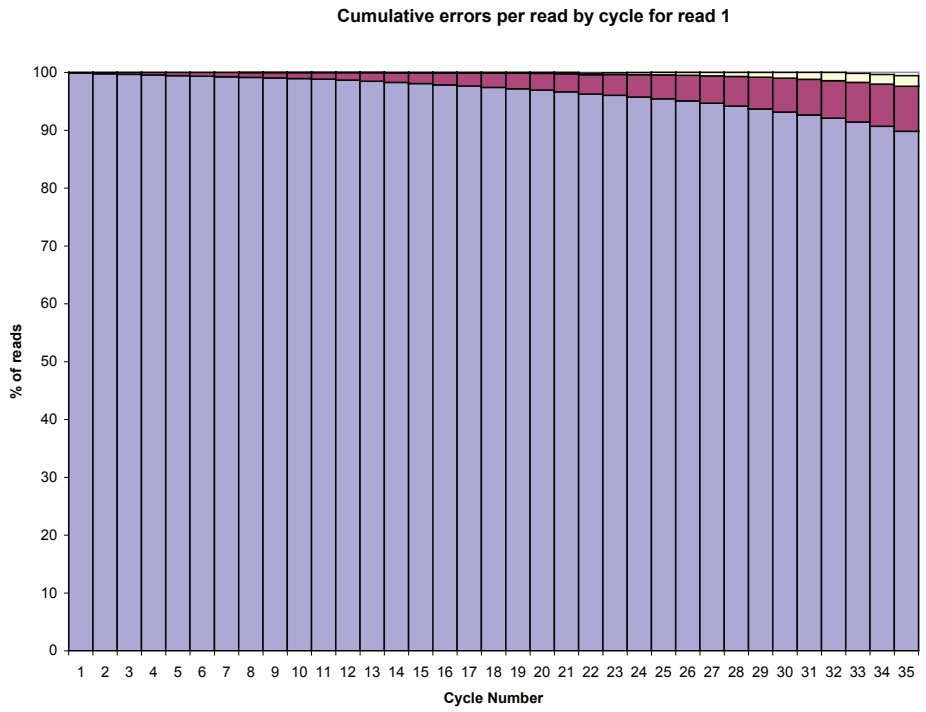


Figure S4. Sequencing summary data for the human BAC from one lane of a flow cell. ***a***, summary table for lane 5 (reads 1 and 2). Lane yield = all PF clusters per lane x read length (0.035 kbases). First cycle intensity (Int) is averaged over all PF reads and given in arbitrary units. % intensity after 20 cycles provides a measure of signal loss during the run. The % PF clusters indicates the fraction of raw clusters that provide PF sequence data after purity filtering (described in supplementary methods). In this example 91.29% of reads (read 1) aligned to the BAC reference. Note that the BAC sample also contains *E. coli* DNA; we found that a further 7% of reads aligned to the *E coli* reference. ***b***. Plot of cumulative errors per read by cycle for read 1 of the experiment summarised in a. The plot shows the % reads with 0 (blue), 1 (burgundy) or 2 (yellow) differences from the reference.

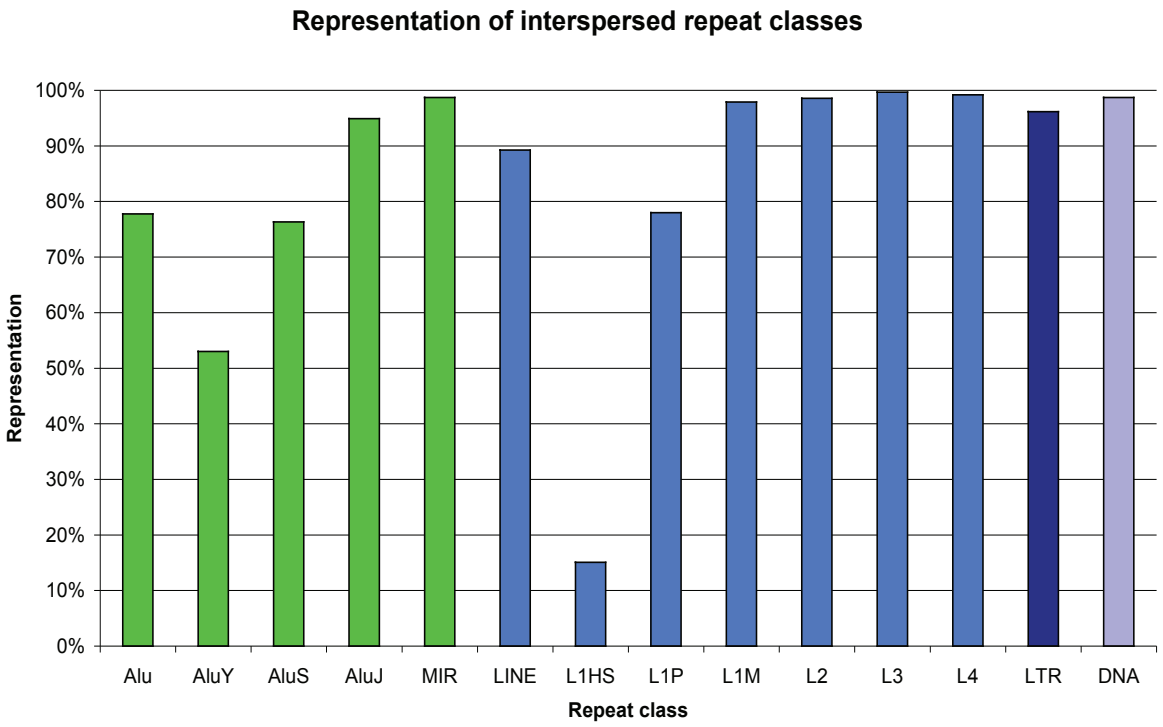


Figure S5. Representation of interspersed repeat classes in the ELAND alignment of the X chromosome dataset. Two-thirds of the low or zero sequence depth in the ELAND alignment is at interspersed repeats. The near-identical repeats arising from recently inserted retroposon elements L1Hs and members of the AluY subfamily of long and short interspersed nuclear element (SINE and LINE) classes are under-represented, while the rest are well covered with reads. Key: Alu, all Alu repeat sequences combined. Different subfamilies are as labelled: AluY - AluJ. LINE, all elements; subsets are defined by bars L1HS – L4. LTR, long terminal repeat retroposons. DNA, DNA transposons.

Illumina Exhibit 1006, Page 33 of 67

a

Call	ELAND SNPs n	ELAND In dbSNP %	MAQ SNPs n	MAQ In dbSNP %
Homozygote	36,402	91.7%	40,953	90.3%
Heterozygote	56,003	80.3%	63,614	77.6%
All	92,485	84.9%	104,567	82.6%

b

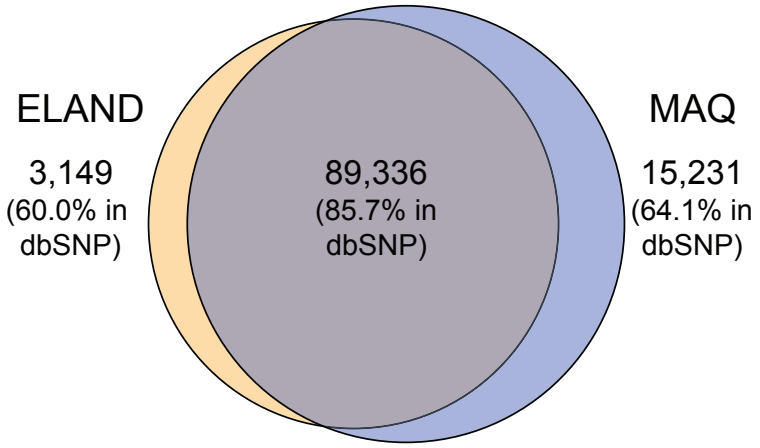


Figure S6. SNPs identified in the X chromosome sequence of NA07340. ***a***. Number of SNPs detected by class and % in dbSNP (release 128). Results from ELAND and MAQ alignments are reported separately. ***b.*** Overlap of SNPs detected in each analysis. See figure S6c (next page) for a more detailed breakdown of SNPs called by MAQ but not ELAND.

Illumina Exhibit 1006, Page 34 of 67

c

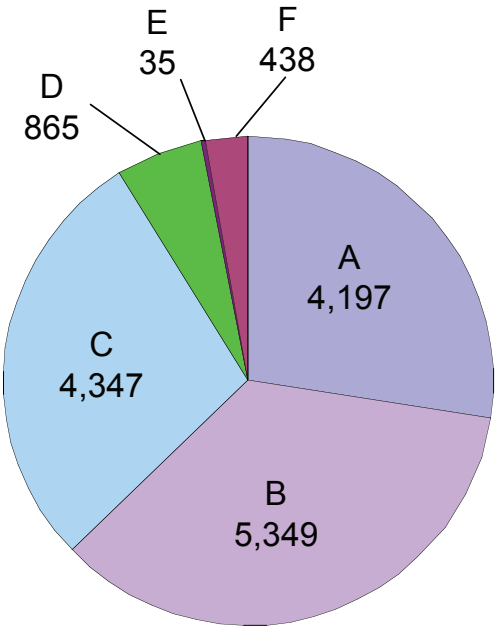


Figure S6c. Breakdown of NA07340 X-chromosome SNP calls present in MAQ but not the ELAND analysis. A: No call possible in ELAND analysis; B: MAQ calls a heterozygous position - both alleles are seen in the ELAND alignment but read depth is too low for one allele to reach the required score threshold; C: MAQ calls a heterozygous position - only the reference allele is seen in the ELAND alignment; D: Different SNP calls are made by ELAND and MAQ; E: MAQ calls a homozygous difference from the reference where ELAND calls a heterozygous position; F. MAQ calls a homozygous difference from the reference where ELAND calls the reference allele.

Illumina Exhibit 1006, Page 35 of 67

a

Input read depth x-fold	GT posns not covered % (of 13604)	hets undercalled % (of 4717 hets)	Discordant calls %
9	12.12	4.92	0
11	9.16	3.43	0
15	2.90	1.27	0
22	1.18	0.38	0
30	0.29	0.17	0
45	0.09	0.02	0

b

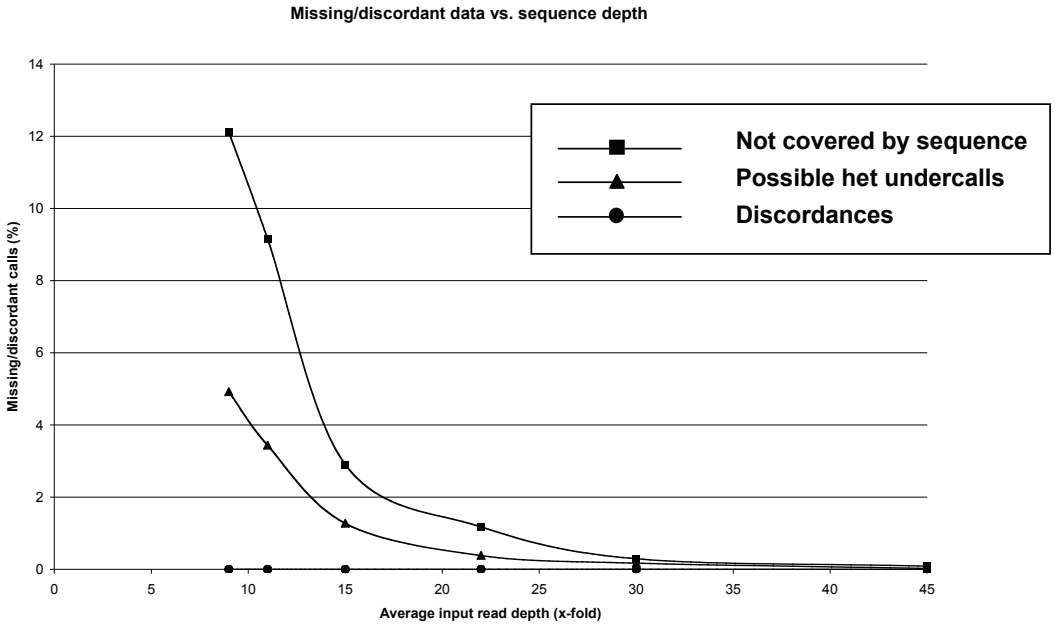


Figure S7. Analysis of SNP calls in X chromosome data for NA07340 using MAQ at different input read depths, compared to genotype data. ***a.*** Tabulated values and ***b.*** plot of percent coverage of 13604 genotype positions, percent of heterozygotes undercalled compared to genotype results, and discordant calls (all 0) as a function of Input read depth (from 9x to 45x).

Illumina Exhibit 1006, Page 36 of 67

a

	N	%dbSNP
Heterozygotes	5606	39.04%
Homozygotes	4142	53.96%
All indels	9747	45.38%

b

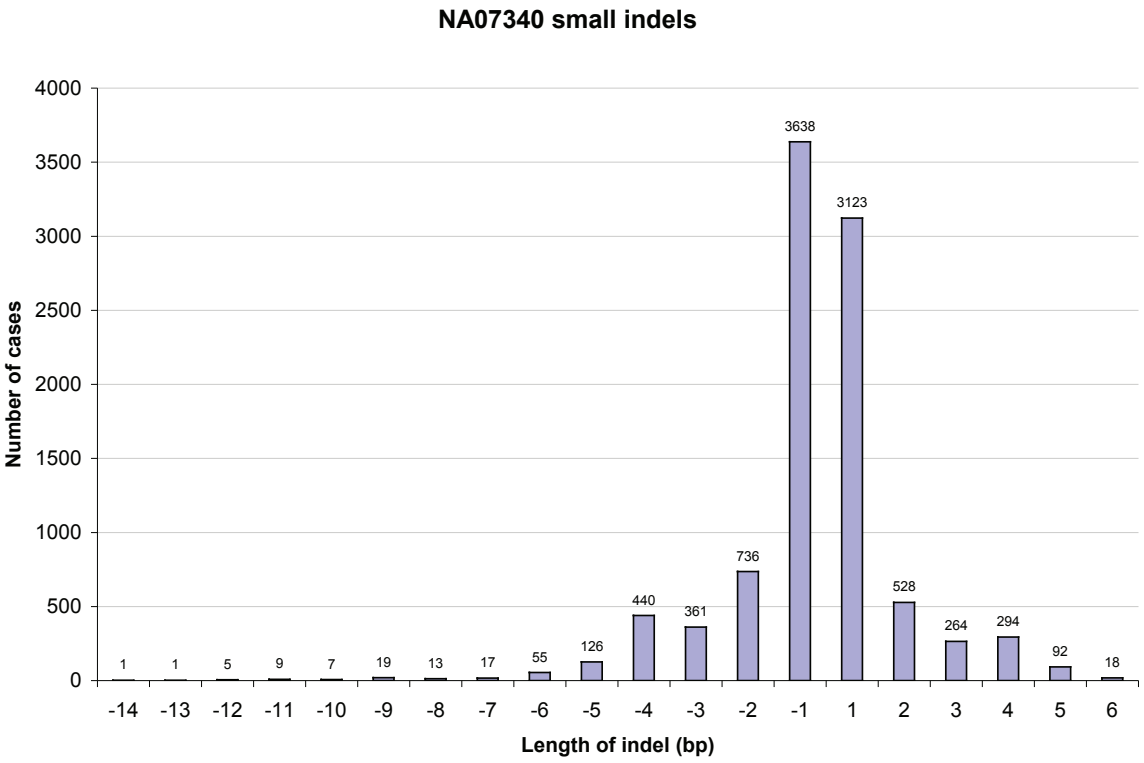


Figure S8. Analysis of short indel calls in X chromosome data for NA07340. ***a.*** total number of calls and fraction that match previous entries in dbSNP. ***b.*** Distribution of size in the 9747 indels. + and – values on the x axis correspond to presence or absence of bases in NA07340 relative to the reference sequence.

Illumina Exhibit 1006, Page 37 of 67

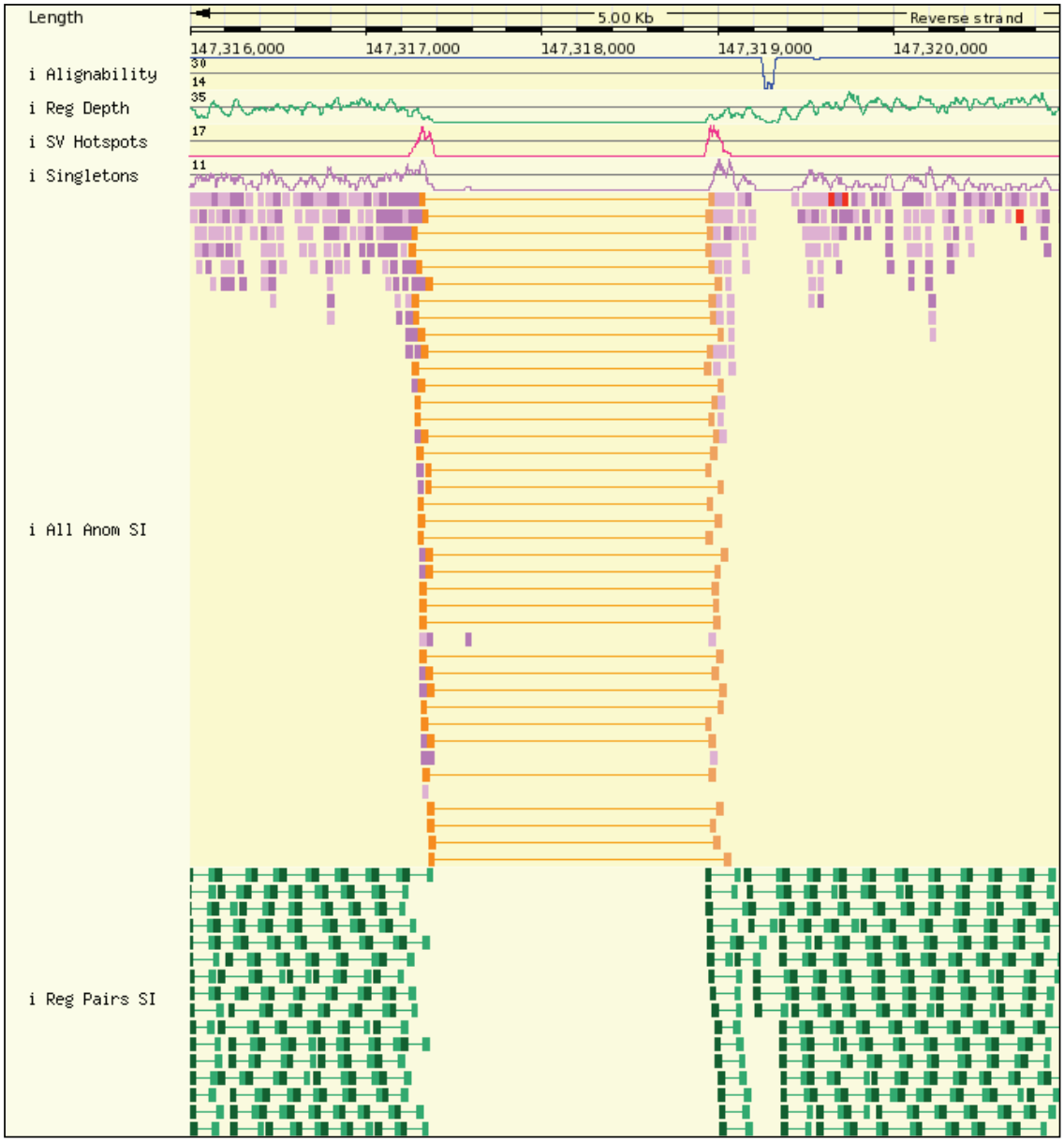


Figure S9a. homozygous 1.6 kb deletion in NA07340 relative to the reference, detected by anomalously spaced read pairs (orange) compared to regularly spaced read pairs (green). Purple reads are singletons, i.e. reads where the other member of the pair does not align to the reference. Note also the alignability score (blue line) is high across the region. Read depth (green line) falls to zero within the deletion. The red line depicts SV hotspots, and provides a quick look-up for regions where there is a concentration of anomalous events detected on the basis of the paired read alignments. Note that this display is a composite of screen shots of the same window, overlapped for display purposes in this figure.

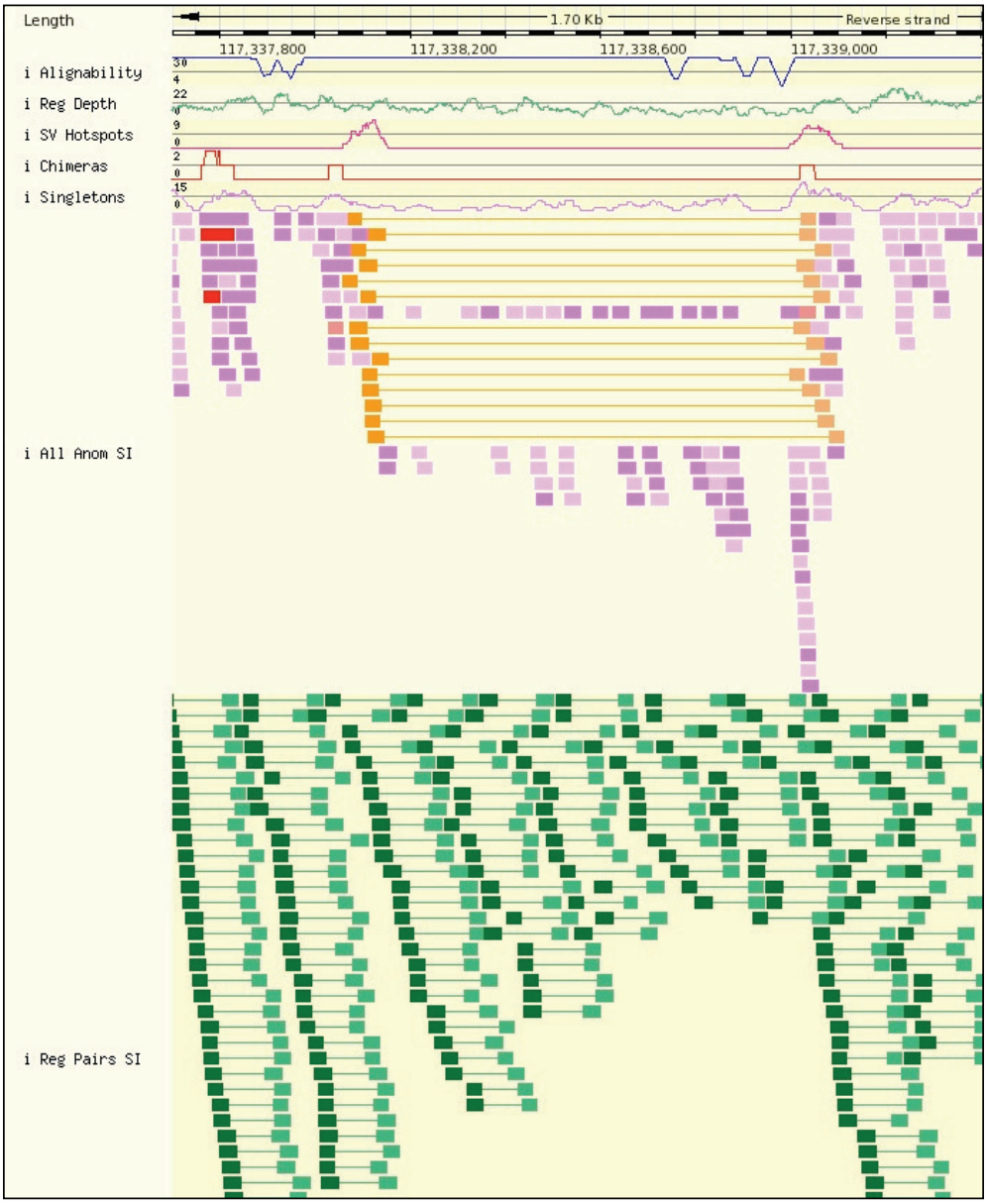


Figure S9b. Heterozygous 844 bp deletion in NA07340 relative to the reference, detected by anomalously spaced read pairs (orange) compared to regularly spaced read pairs (green). The presence of regularly spaced pairs across the entire region is indicative of heterozygosity at this locus. Note that this display is a composite of screen shots of the same window, overlapped for display purposes in this figure.

Illumina Exhibit 1006, Page 39 of 67

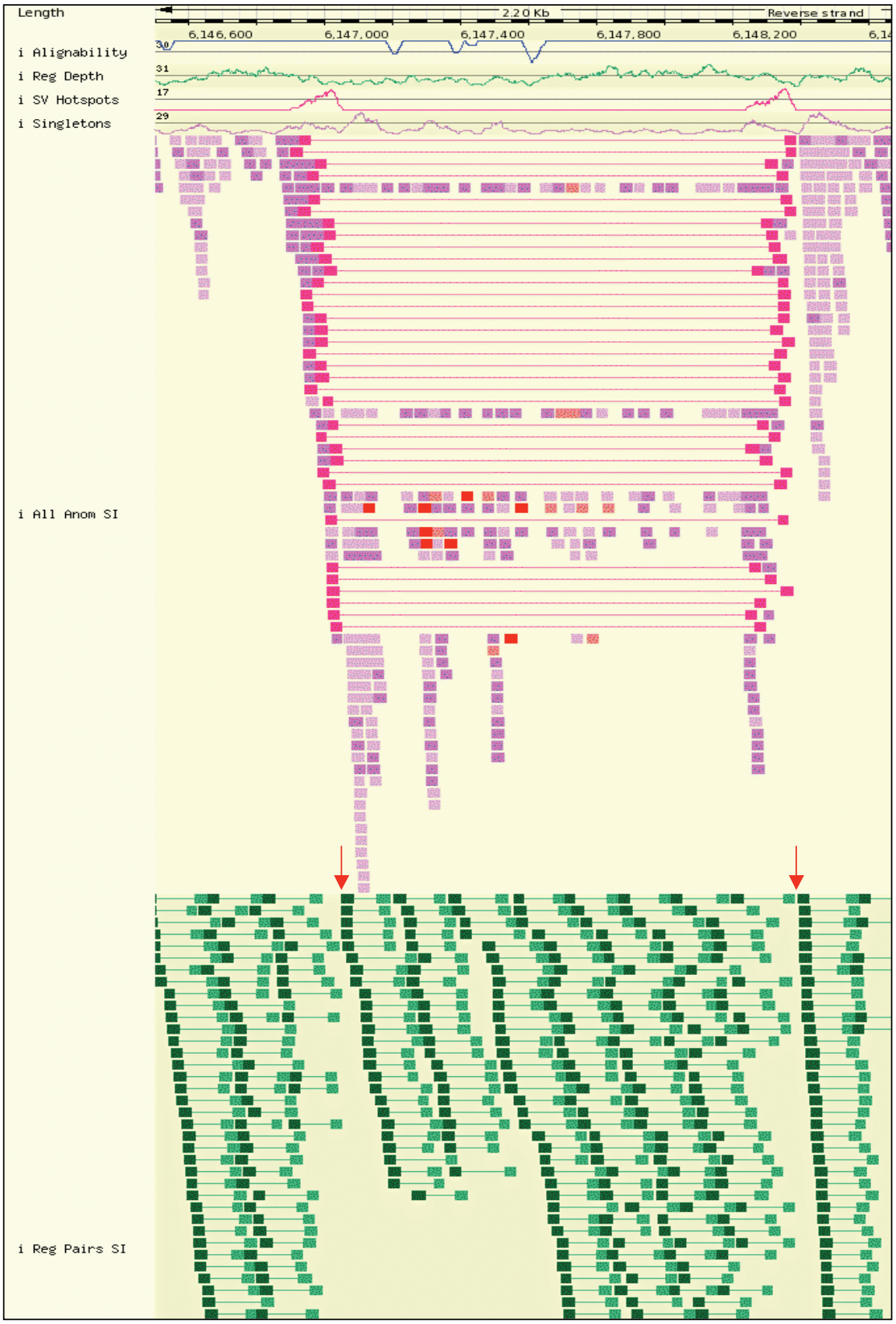


Figure S10. (See next page for legend).

Figure S10 (see previous page for figure). Homozygous 1.5 kb inversion in NA07340 relative to the reference, detected by anomalously spaced and inverted read pairs (red) compared to regularly spaced read pairs (green). Note the gaps in continuity of coverage in the normal short inserts, indicated by the red arrows, indicating that this variant is homozygous. No read pair spans either breakpoint. The reader is asked to refer to the website (link below) and to zoom in to visualise the gap running through the complete dataset. Note also the alignability score (blue line) is high across most of the region. Read depth (green line) remains high across the inversion. Purple reads are singletons, which accumulate as the other member of the read pair does not match to the reference. Singletons at each inversion breakpoint result from the disruption in alignment of the other member of the pair at or near the breaks. Red arrows indicate homozygous breaks in continuity with normal paired end mapping. Note that this display is a composite of screen shots of the same window, overlapped for display purposes in this figure.

Illumina Exhibit 1006, Page 41 of 67

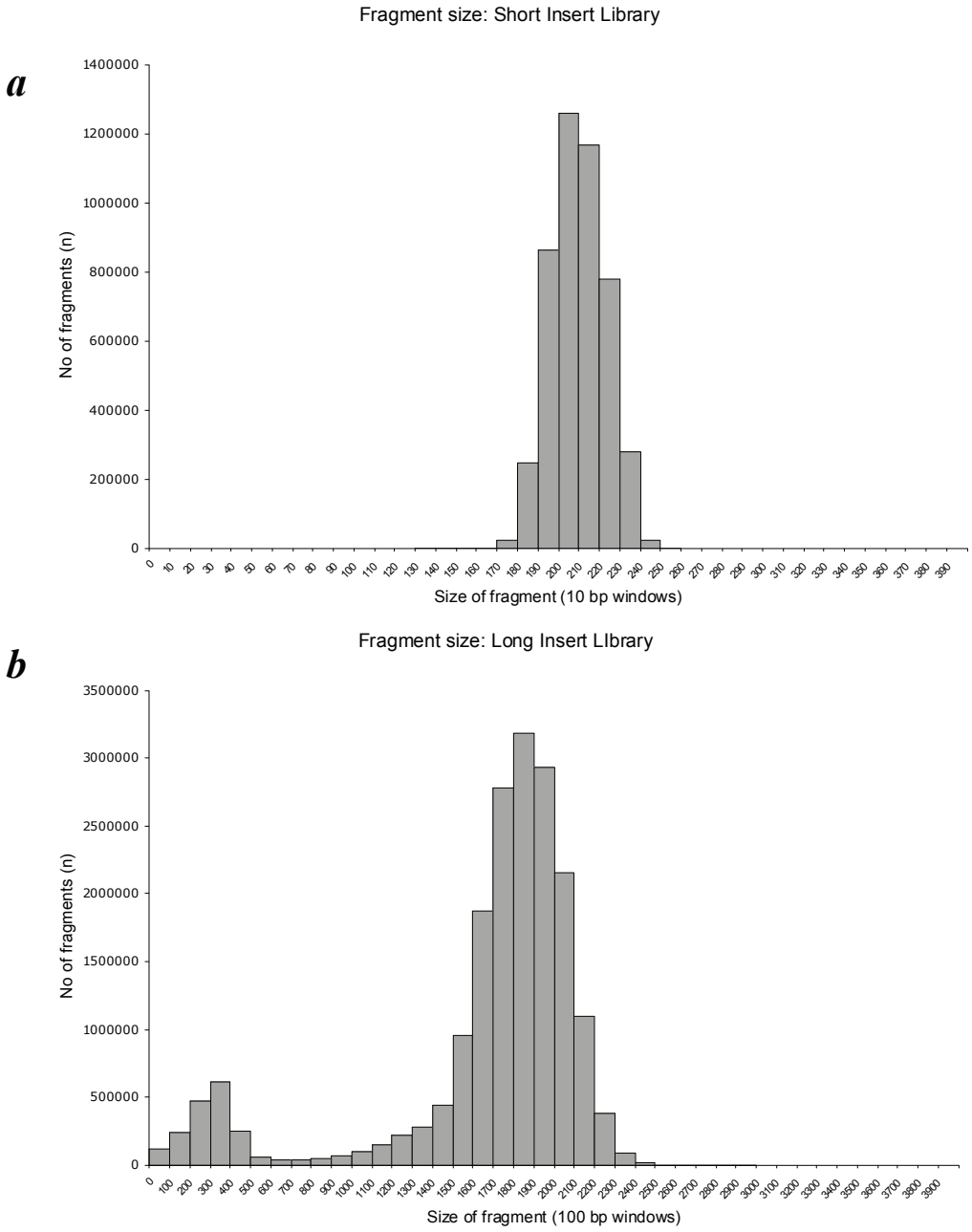


Figure S11 Size distribution for DNA fragment libraries of NA18507. Fragment lengths were determined from the separation between read pairs aligned to the reference. **a**. Short insert library, data from one lane of a flowcell were analysed and fragment lengths binned in 10 bp windows. **b.** Long insert library, data from six lanes of a flowcell with lengths binned in 100 bp windows. Note that the smaller peak on the left reflects the presence of a small fraction of short, non-junction fragments of genomic DNA that are not biotinylated. These are readily eliminated during analysis as the orientation of the paired reads is opposite to that of the real junction fragments (see fig 1 for more details).

Illumina Exhibit 1006, Page 42 of 67

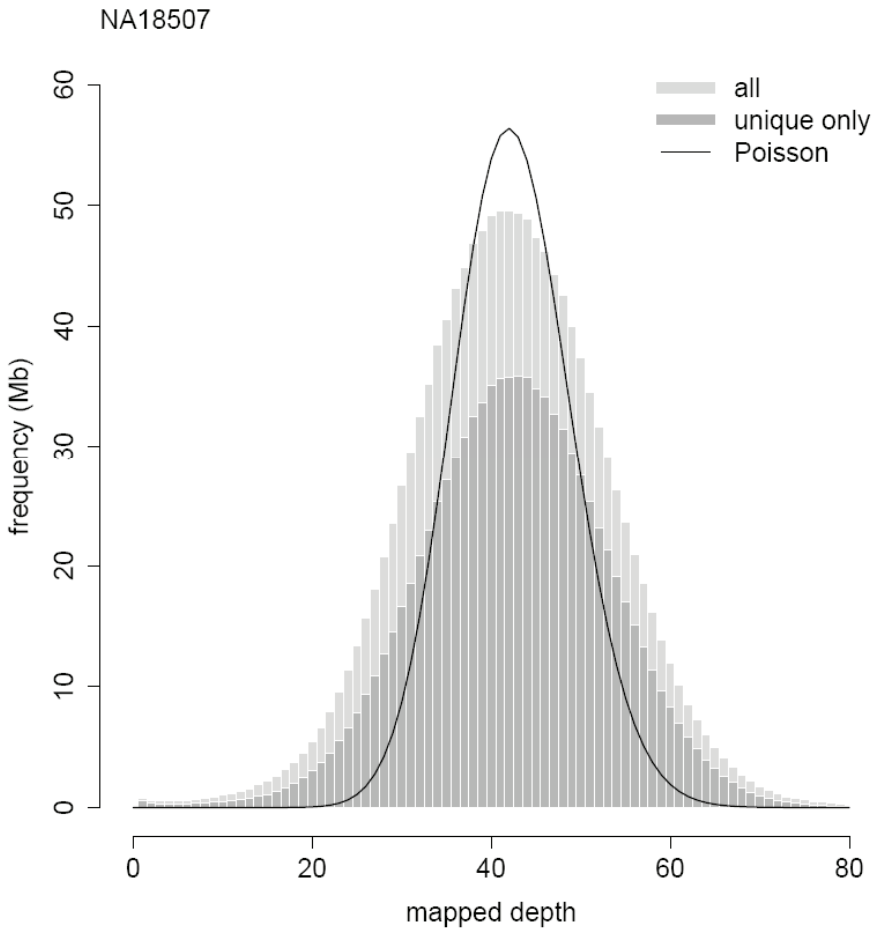


Figure S12a. Distribution of mapped read depth in the human genome dataset, sampled at every 500th position along the genome and displayed as a histogram ('all'). An equivalent analysis of mapped read depth for the unique subset of these positions is also shown ('unique only'). The solid line represents a Poisson distribution for unique human genome sequence

Illumina Exhibit 1006, Page 43 of 67

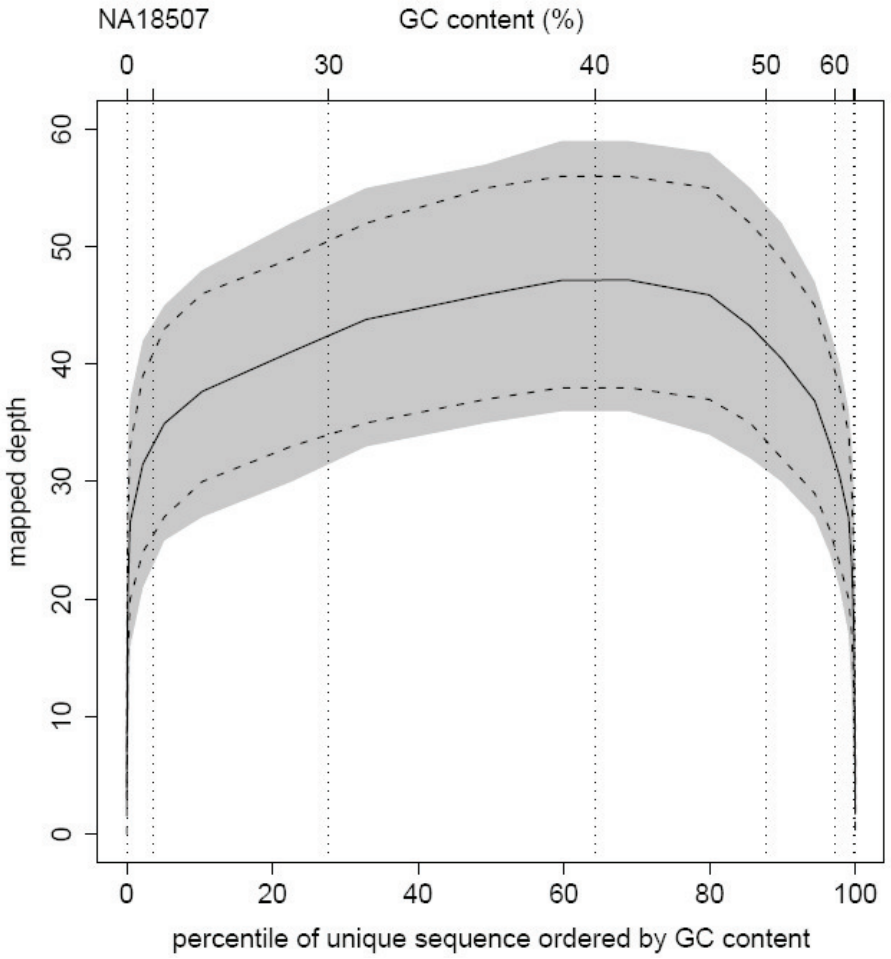


Figure S12b. Distribution of human genome NA18507 uniquely mapped reads as a function of GC content. Note that the x axis is % GC content and is scaled by percentile of unique sequence. The solid line is average mapped depth of unique sequence; the grey region is the central 80% of the data (10th to 90th centiles); the dashed lines are 10th and 90th centiles of a Poisson distribution with the same mean as the data. GC biases may be introduced during sample pre-amplification, cluster growth and/or sequencing. Possible contributors include reduced efficiency of GC-rich fragment amplification or loss of very short AT-rich fragments due to denaturation during sample preparation, both of which might vary from library to library and result in the difference in distribution seen here compared to the X-chromosome library (see fig. 2b).

Illumina Exhibit 1006, Page 44 of 67

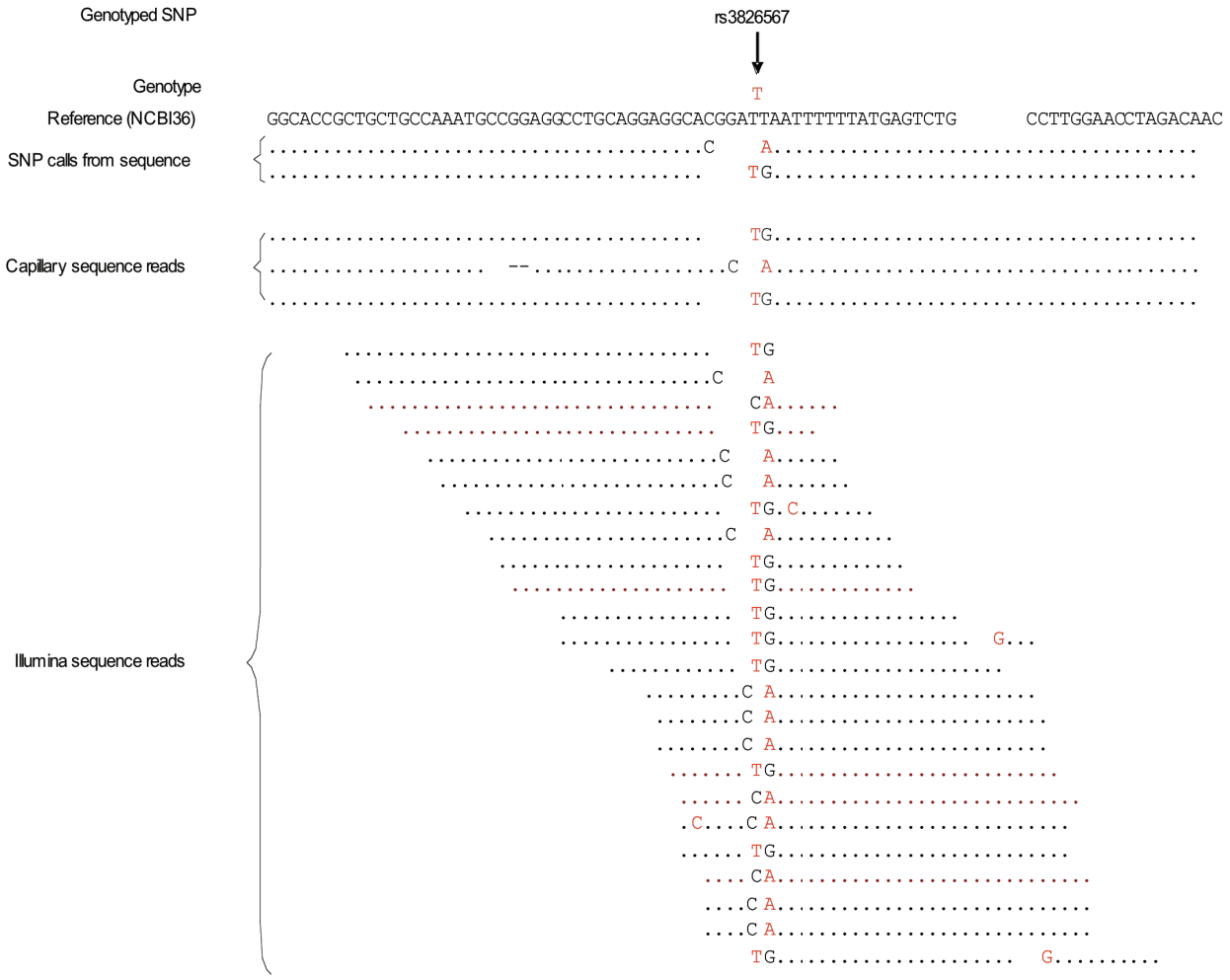


Figure S13. Sequence data reveals 'hidden' SNPs that cause genotyping errors. *a*. Example of a 'hidden' SNP detected in the sequence data. The HM550 genotyping records a 'T' only for NA18507 at SNP rs3826567 (arrowed), indicating that this individual is homozygous 'TT' at this position. Illumina sequence reads detect two alleles at this position, T and C, indicating a heterozygous position 'TC'. Capillary sequence reads obtained previously for this individual were available in the trace archive and confirm the existence of both T and C alleles. The sequence data identify the existence of a second SNP (the 'hidden SNP') directly adjacent to rs3826567 and indicate that NA18507 is heterozygous 'AG' at this position. As the genotyping assay utilises the reference sequence (in this case 'G') for design of the oligonucleotides used in the assay, the genotyping assay does not accommodate the existence of the non-reference 'A' allele and only interacts with DNA containing the reference 'G' allele, resulting in a 'T' call at rs3826567.

Illumina Exhibit 1006, Page 45 of 67

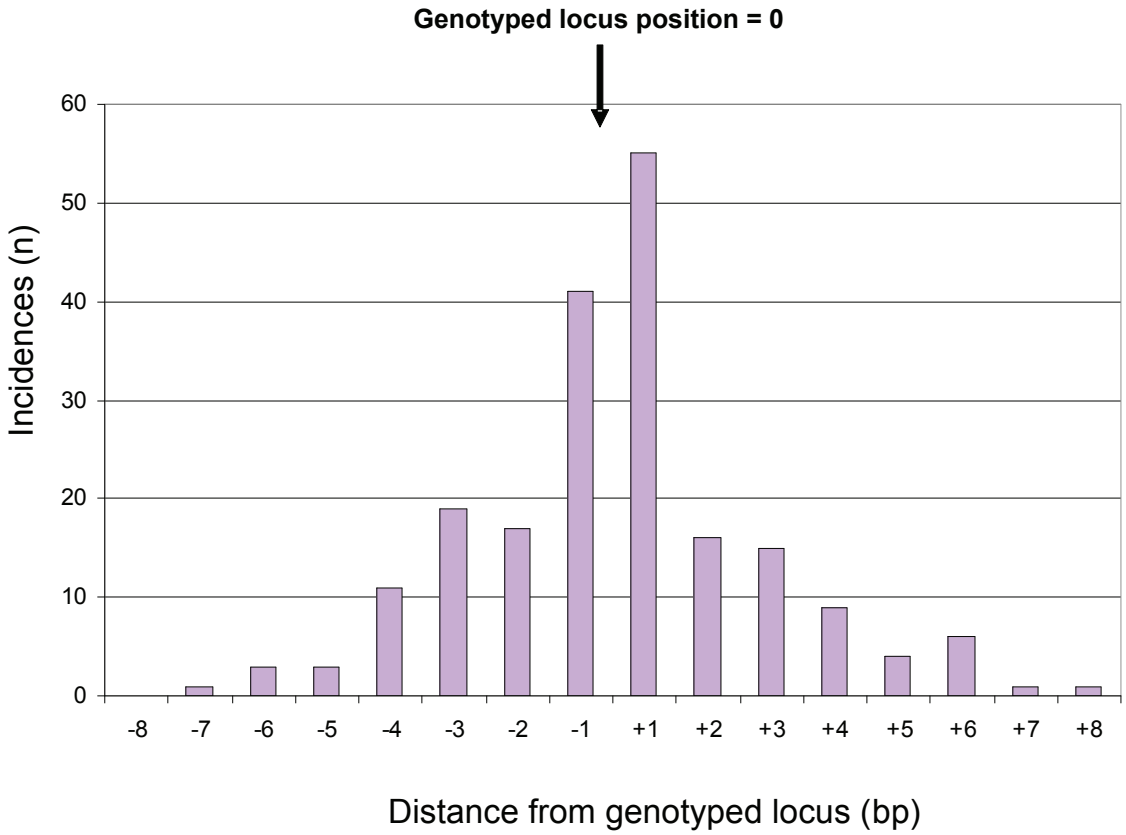
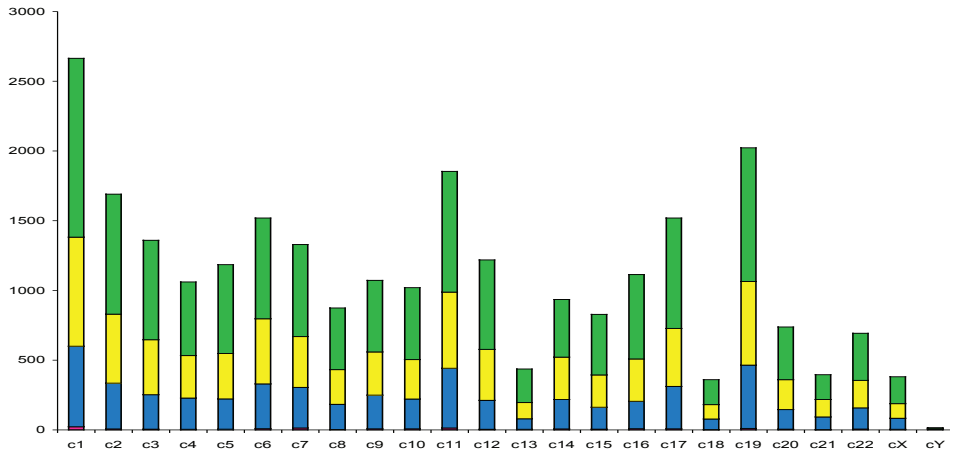


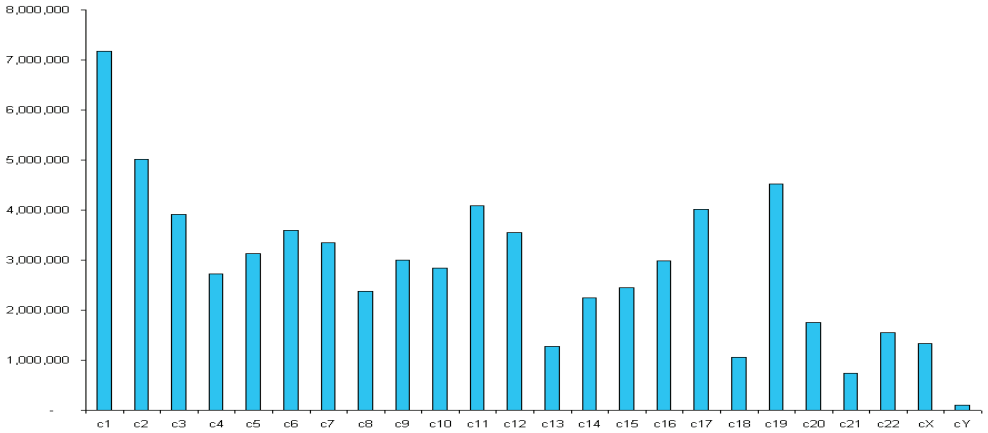
Figure S14. Distribution of 'hidden' SNPs in 202 discordant calls (see main text for discription of 'hidden' SNPs). Note that 'hidden' SNPs will affect any genotyping platform. We therefore expect a low, but non-zero, level of genotype undercalling (Seq>GT) that is highly correlated between the platforms that use these flanking sequences in assay design.

Illumina Exhibit 1006, Page 46 of 67

a



b



c

Chromosome	1	2	3	4	5	6	7	8	9	10	11	12
Heterozygosity (no. of hets/covered coding bases x 10-3)	0.54	0.46	0.46	0.54	0.48	0.57	0.57	0.55	0.51	0.51	0.66	0.45
Chromosome	13	14	15	16	17	18	19	20	21	22	All coding	All genome
Heterozygosity (no. of hets/covered coding bases x 10-3)	0.49	0.58	0.45	0.57	0.53	0.42	0.63	0.62	0.65	0.61	0.54	0.99

Figure S15. Coding SNPs in NA18507, based on the gene annotation in Ensembl (release 49). ***a***. distribution of coding SNPs and their consequences by chromosome. Green: synonymous, yellow: non-synonymous, conservative, blue: non-conservative, red: premature stop codons. ***b***. distribution of coding bases by chromosome. Note close correlation of coding SNP and coding sequence distribution by chromosome (as opposed to overall chromosome length). ***c.*** Correlation of heterozygosity with coding length between chromosomes. Average autosomal coding region heterozygosity ('All coding') is approximately half that of genome average for autosomes ('All genome').

Illumina Exhibit 1006, Page 47 of 67

a

	n	%dbSNP
All indels	404,416	49.82%
Heterozygotes	260,620	41.44%
Homozygotes	143,796	64.99%

b

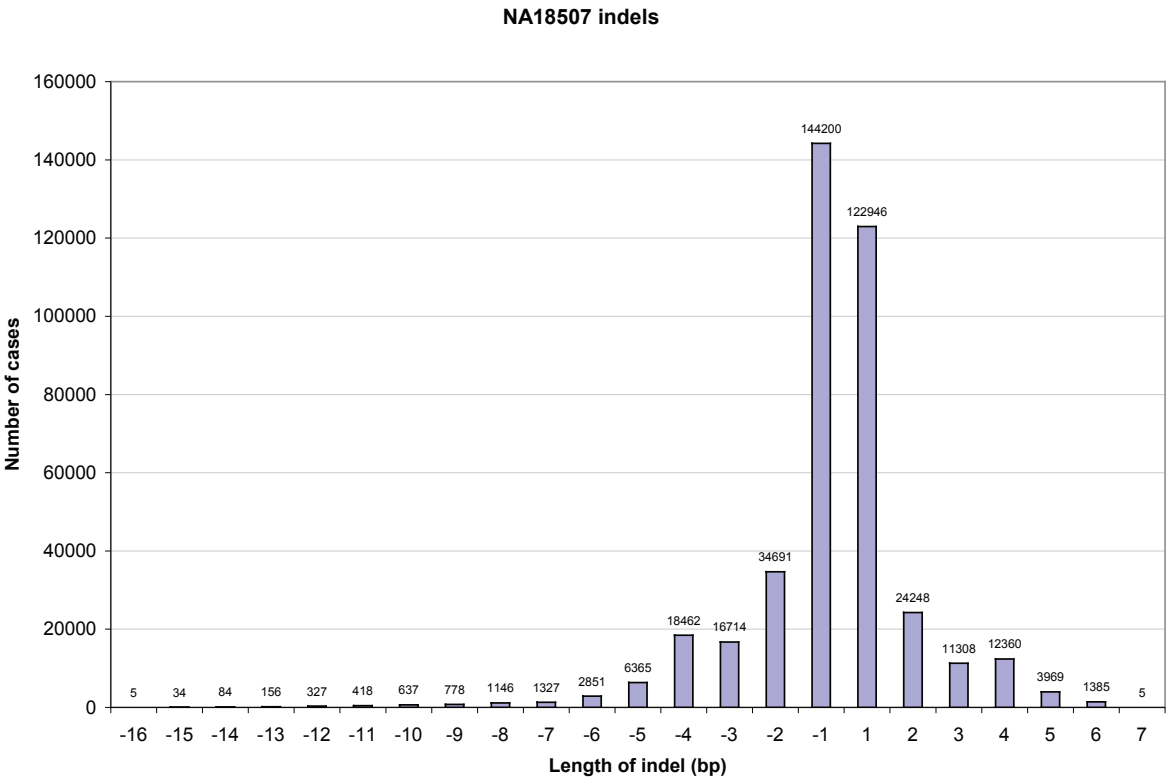


Figure S16. Analysis of short indel calls in human genome data for N18507. ***a.*** total number of calls and fraction that match previous entries in dbSNP. ***b.*** Distribution of size in the 404,416 indels. + and – values on the x axis correspond to presence or absence of bases in NA18507 relative to the reference sequence.

Illumina Exhibit 1006, Page 48 of 67

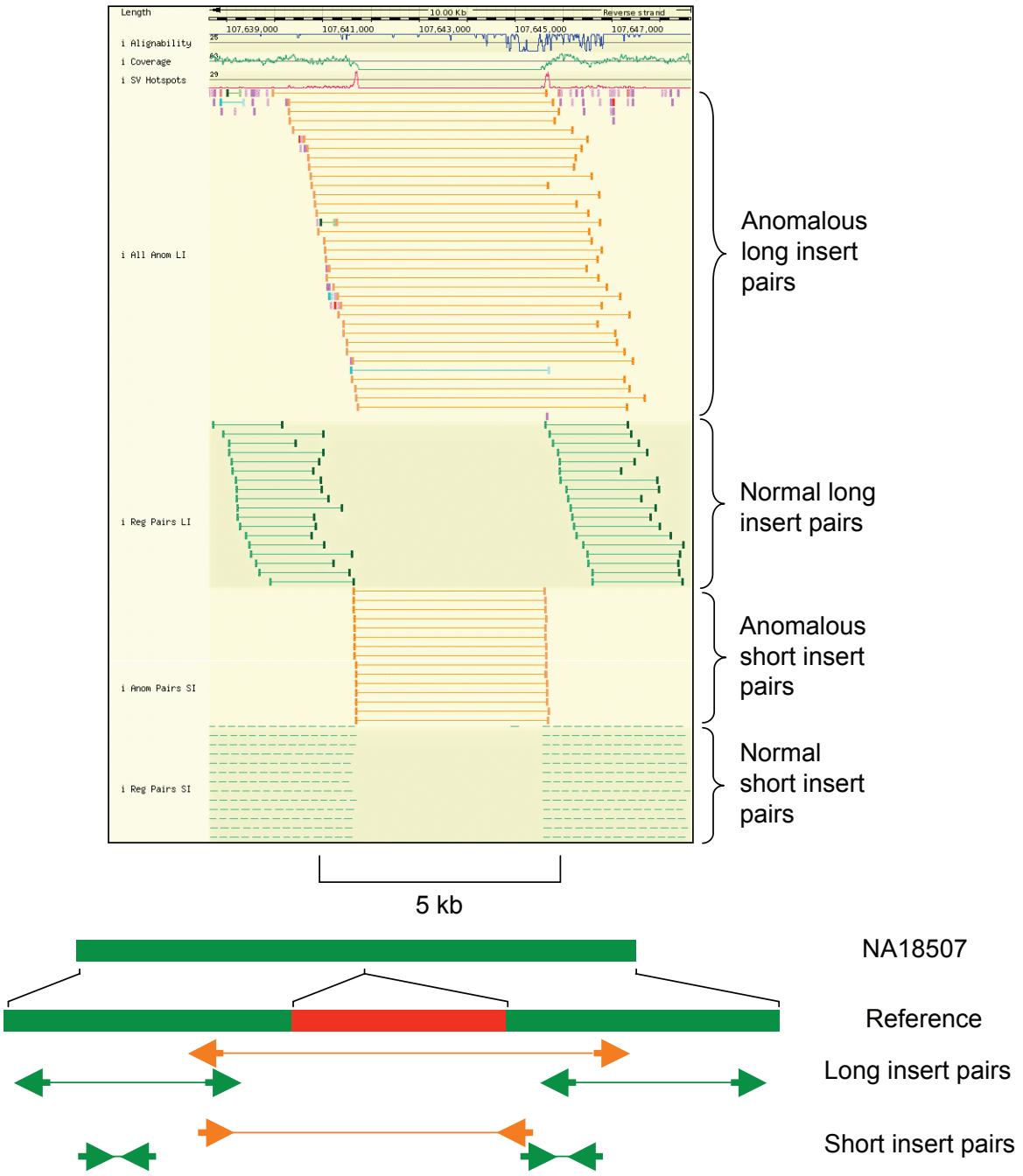


Figure S17*a*. Homozygous 3.6 kb deletion in NA18507 relative to the reference, supported by anomalously spaced long and short insert read pairs (orange) compared to regularly spaced read pairs (green). See also the schematic below. Note that the correct read orientation for long insert read pairs is different from that of the short insert read pairs and is denoted by the arrows which point in a 5' to 3' direction. See fig 1 for explanation. Note that this display is a composite of screen shots of the same window, overlapped for display purposes in this figure.

Illumina Exhibit 1006, Page 49 of 67

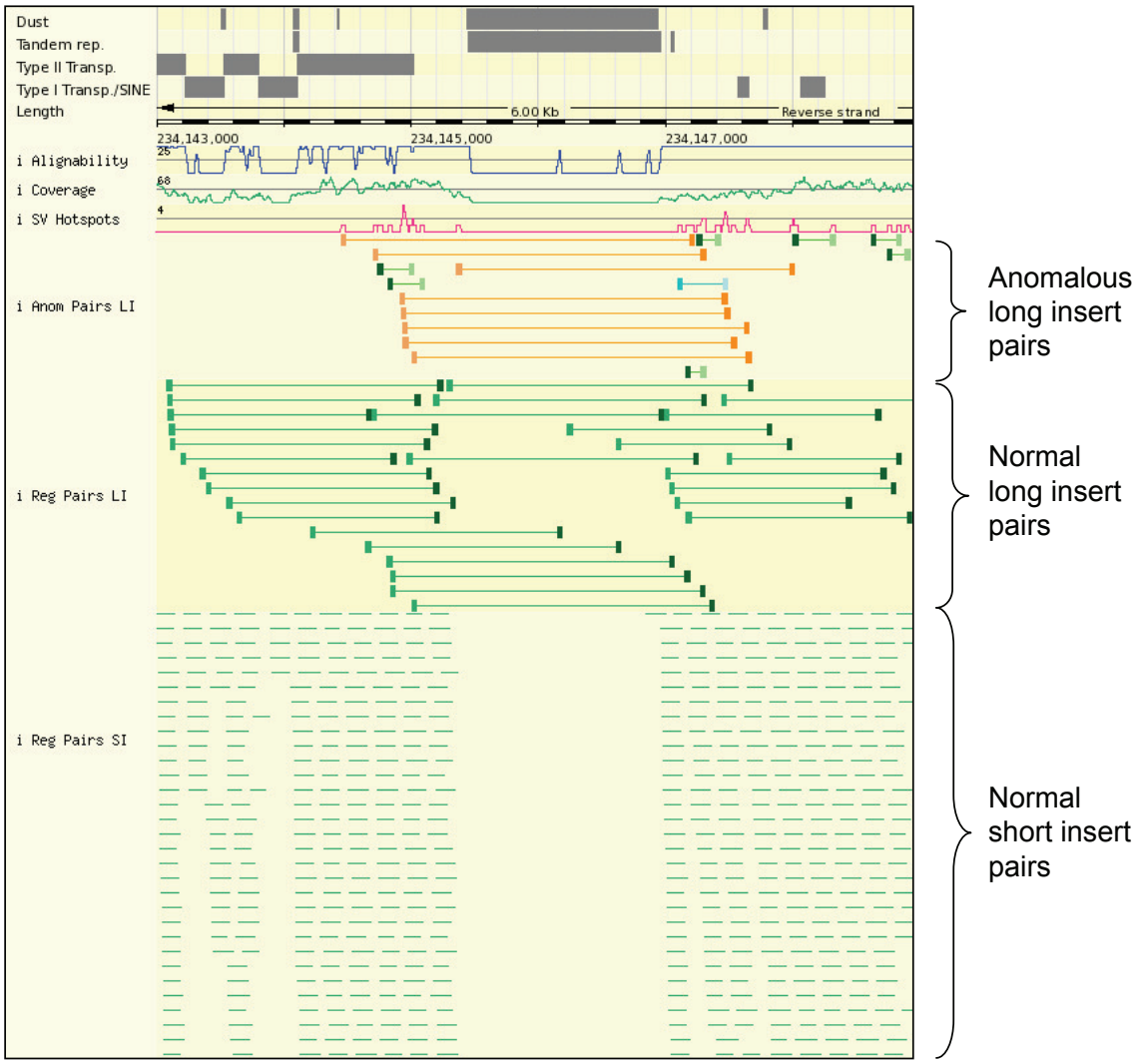


Figure S17b. Deletion in NA18507 relative to the reference, seen with anomalous long insert pairs but not short insert data. Tandem repeats (second row of display) cover ~1500 bp on the right hand side of the deletion, preventing unique alignment of reads. Short insert pairs are too short to span the tandem repeat region. Long insert pairs are longer range and right-hand reads map in unique sequence further away from the deletion (see orange blocks representing reads on right hand side of each of the long insert anomalous pairs). Note that this display is a composite of screen shots of the same window, overlapped for display purposes in this figure.

Illumina Exhibit 1006, Page 50 of 67

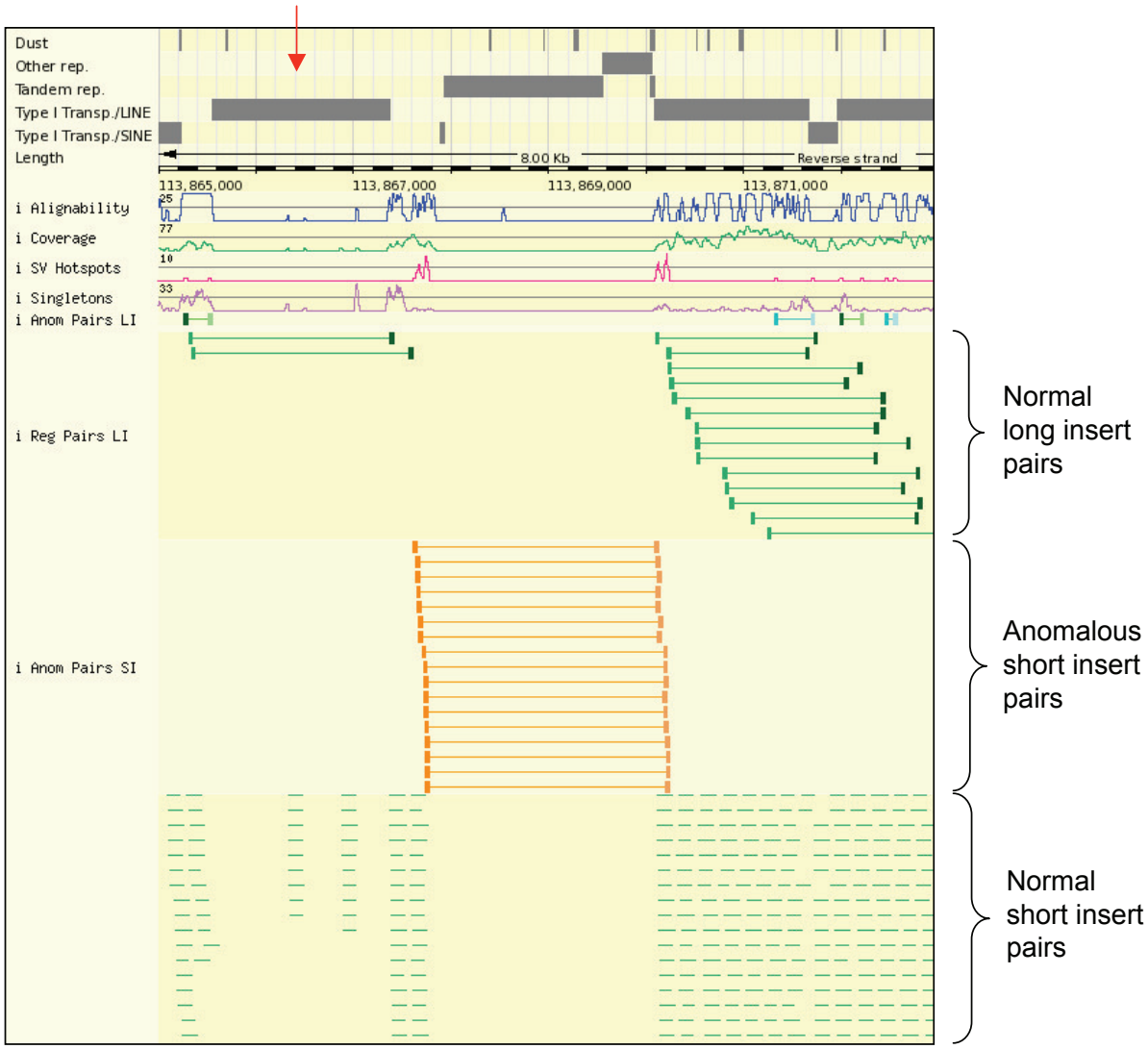


Figure S17c. Deletion in NA18507 relative to the reference, seen with anomalous short insert pairs but not long insert data. The LINE repeat on the left (see red arrow) covers 1.8 kb on the left-hand side of the deletion, preventing unique alignment of long insert reads. The higher density of short insert pairs includes a number of examples that map directly to either side of the deletion, with left ends mapping between the deletion breakpoint and the LINE repeat. Note that the LINE repeat on the right of the deletion does not appear to significantly compromise alignability (c.f. the LINE repeat on the left). Note that this display is a composite of screen shots of the same window, overlapped for display purposes in this figure.

Illumina Exhibit 1006, Page 51 of 67

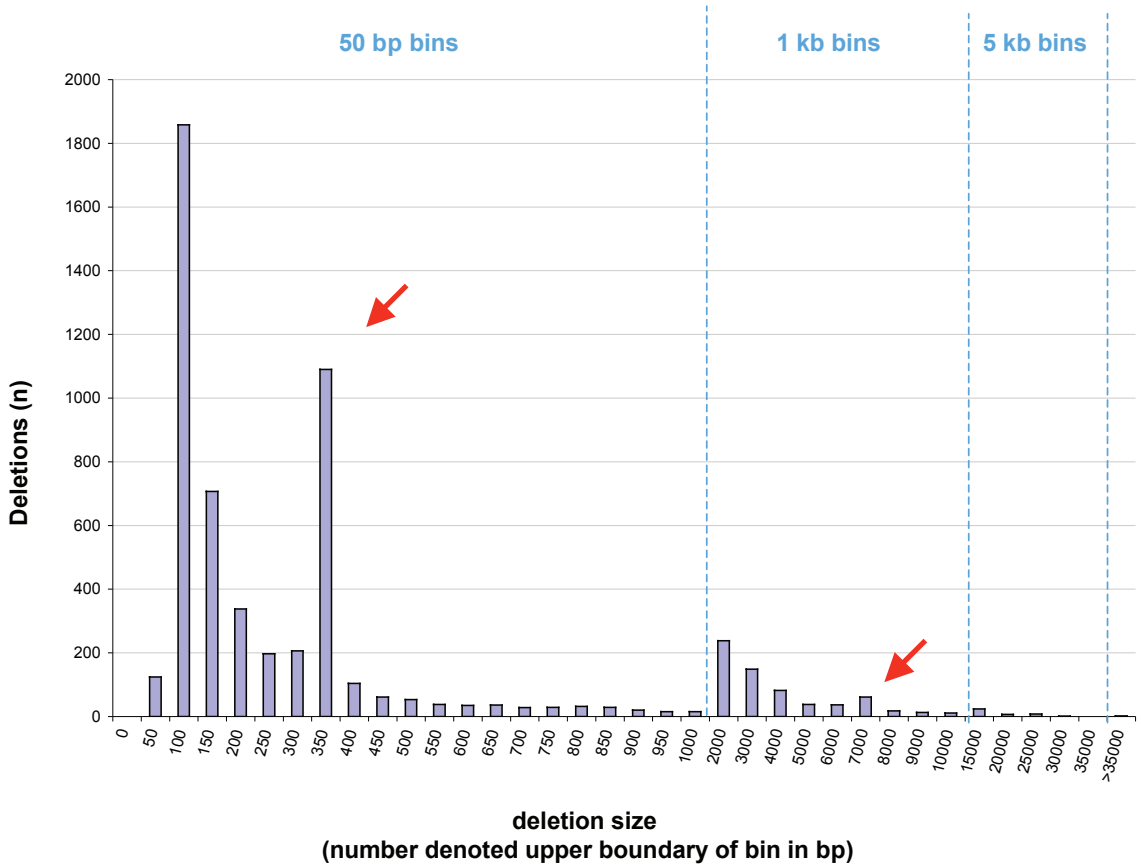


Figure S18. Size distribution of deletions in NA18507 relative to the reference. Two peaks in the distribution (arrowed) contain a substantial fraction of sequences corresponding to recently active retroposon insertions. The AluY family of SINEs are found in 90% (987/1092 of the 300-350bp fraction, and also contribute slightly to the 250-300 bp fraction. The L1HS LINE element accounts for 67% (42/61) of the 6-7 kb fraction. Note that other peaks are artefacts in the plot caused by the changes in size range of the bins.

Illumina Exhibit 1006, Page 52 of 67

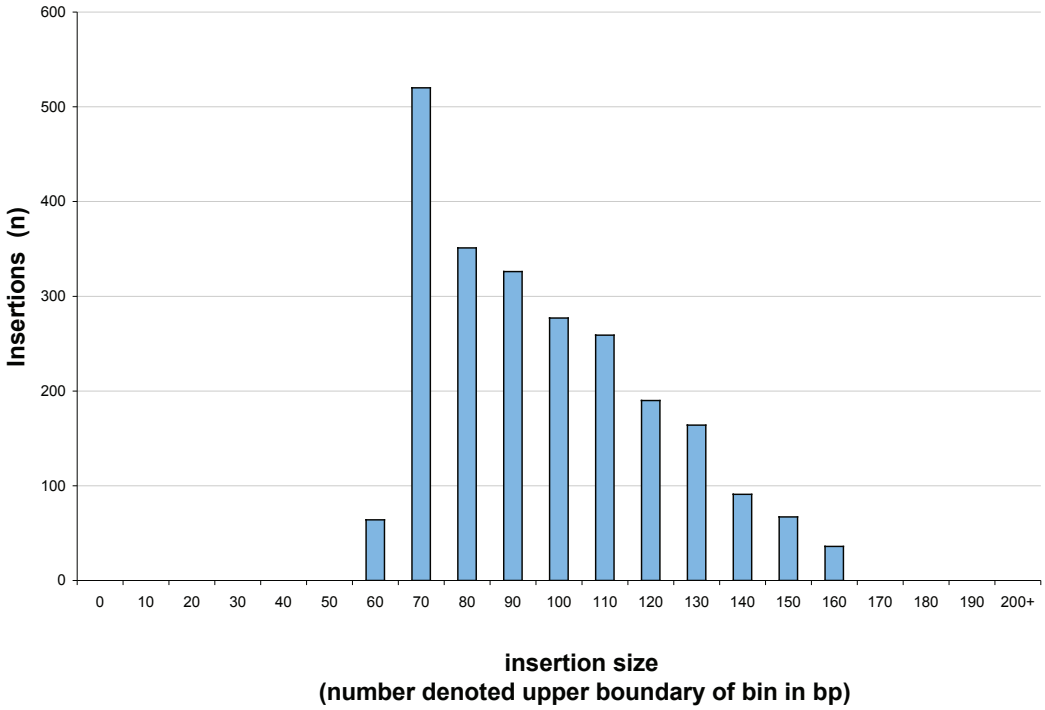


Figure S19. Size distribution of insertions in NA18507 relative to the reference, detected in the short insert dataset. Size distribution is limited by the resolution of analysis of the short insert fragment size range.

Illumina Exhibit 1006, Page 53 of 67

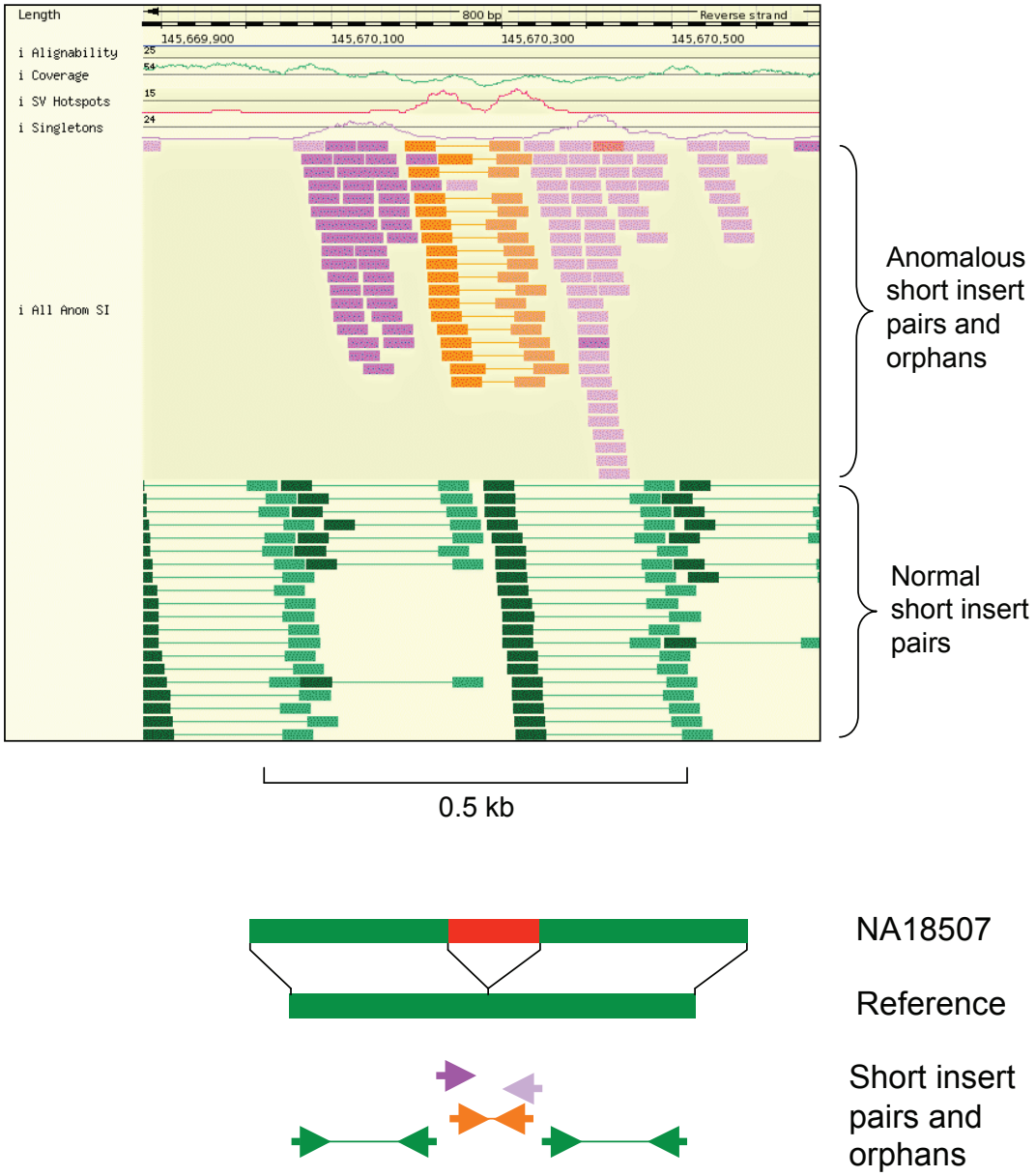


Figure S20. Homozygous 81 bp insertion detected by short insert data only. Orange bars show the anomalously short read pair spacing. Note the accumulation of singleton reads on either side, where each read has an unaligned pair. Dark and light purple denote rightwards and leftwards orientation (see schematic). Thus all singleton reads (dark purple) have unaligned pairs to the right of them; and vice versa for pale purple singletons. These unaligned pairs form part of the *de novo* assembly that results in a single contig across the sequence inserted in NA18507 relative to the reference (see fig S21). Note that this display is a composite of screen shots of the same window, overlapped for display purposes in this figure.

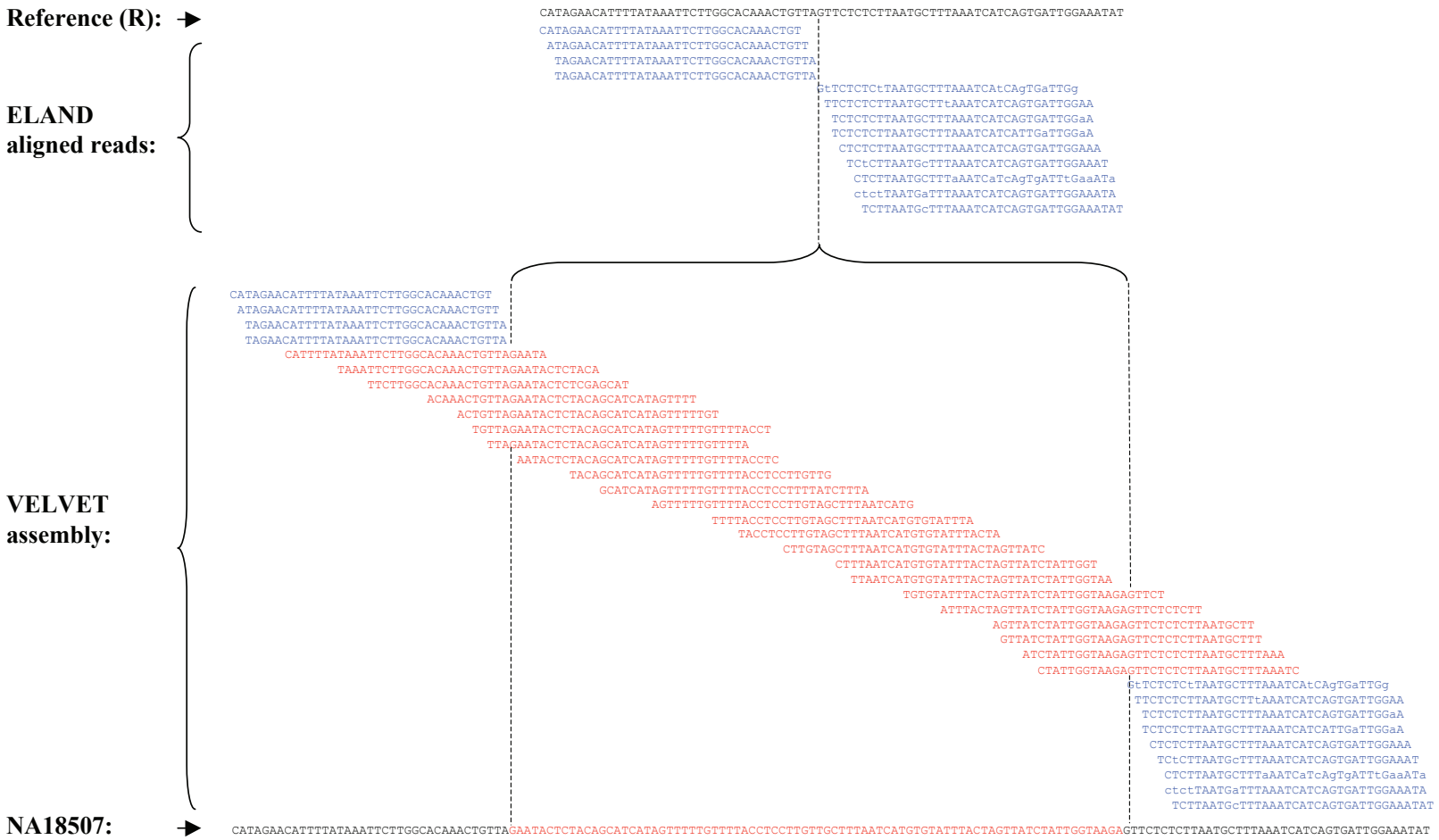


Figure S21. *De novo* assembly defines an inserted sequence in NA18507 absent from the reference. Bases present in the reference are in black, bases in NA18507 reads that align to the reference using ELAND are in blue, and bases in additional NA18507 reads that are assembled by VELVET are in red. The extra sequence in the NA18507 consensus (bottom line) is also shown in red.

a 071126_EAS56_0057_FC200433 – lanes 1-8 read 1

Lane Info		Tile Mean +/- SD for Lane							
Lane	Lane Yield (kbases)	Clusters (raw)	Clusters (PF)	1st Cycle Int (PF)	% intensity after 20 cycles (PF)	% PF Clusters	% Align (PF)	Alignment Score (PF)	% Error Rate (PF)
1	266021	31482 +/- 1612	23032 +/- 1162	3646 +/- 318	99.39 +/- 3.25	73.17 +/- 1.17	86.05 +/- 0.35	49.50 +/- 1.73	0.50 +/- 0.04
2	258725	30395 +/- 1596	22400 +/- 1396	3803 +/- 385	92.23 +/- 2.98	73.68 +/- 1.84	86.03 +/- 0.39	48.68 +/- 2.21	0.53 +/- 0.07
3	262340	30436 +/- 1382	22713 +/- 1128	3610 +/- 370	98.04 +/- 3.36	74.62 +/- 1.07	86.05 +/- 0.30	49.82 +/- 1.61	0.49 +/- 0.04
4	264384	30633 +/- 1697	22890 +/- 1321	4111 +/- 372	94.06 +/- 3.44	74.77 +/- 2.69	86.23 +/- 0.29	50.01 +/- 1.57	0.48 +/- 0.06
5	266557	30793 +/- 1492	23078 +/- 1039	4278 +/- 388	94.15 +/- 3.64	74.97 +/- 0.94	86.09 +/- 0.27	50.36 +/- 1.01	0.48 +/- 0.02
6	261586	30493 +/- 1656	22648 +/- 1194	4294 +/- 403	94.10 +/- 3.25	74.30 +/- 1.68	86.14 +/- 0.44	48.53 +/- 3.00	0.50 +/- 0.15
7	261867	30502 +/- 1835	22672 +/- 1180	4056 +/- 469	96.64 +/- 3.85	74.39 +/- 1.76	86.12 +/- 0.29	50.04 +/- 1.52	0.49 +/- 0.03
8	67955	30168 +/- 1697	21815 +/- 1276	3629 +/- 950	96.75 +/- 4.11	72.36 +/- 2.69	86.11 +/- 0.30	49.81 +/- 1.89	0.50 +/- 0.04

b 071126_EAS56_0057_FC200433 – lanes 1-8 read 2

Lane Info		Tile Mean +/- SD for Lane							
Lane	Lane Yield (kbases)	Clusters (raw)	Clusters (PF)	1st Cycle Int (PF)	% intensity after 20 cycles (PF)	% PF Clusters	% Align (PF)	Alignment Score (PF)	% Error Rate (PF)
1	266021	31482 +/- 1612	23032 +/- 1162	1954 +/- 174	92.80 +/- 6.94	73.17 +/- 1.17	84.17 +/- 0.88	42.70 +/- 2.70	0.75 +/- 0.11
2	258725	30395 +/- 1596	22400 +/- 1396	2109 +/- 206	89.61 +/- 6.56	73.68 +/- 1.84	84.45 +/- 1.06	42.62 +/- 3.25	0.74 +/- 0.14
3	262340	30436 +/- 1382	22713 +/- 1128	2045 +/- 190	95.16 +/- 5.93	74.62 +/- 1.07	84.65 +/- 0.48	44.50 +/- 2.08	0.67 +/- 0.06
4	264384	30633 +/- 1697	22890 +/- 1321	2338 +/- 220	90.64 +/- 5.90	74.77 +/- 2.69	84.97 +/- 0.39	45.20 +/- 2.11	0.64 +/- 0.05
5	266557	30793 +/- 1492	23078 +/- 1039	2260 +/- 220	92.23 +/- 5.66	74.97 +/- 0.94	84.59 +/- 0.55	44.61 +/- 2.19	0.68 +/- 0.10
6	261586	30493 +/- 1656	22648 +/- 1194	2345 +/- 216	90.92 +/- 5.23	74.30 +/- 1.68	84.81 +/- 0.74	44.47 +/- 2.63	0.67 +/- 0.08
7	261867	30502 +/- 1835	22672 +/- 1180	2209 +/- 244	93.00 +/- 5.65	74.39 +/- 1.76	84.56 +/- 0.50	44.30 +/- 2.51	0.70 +/- 0.07
8	67955	30168 +/- 1697	21815 +/- 1276	1945 +/- 524	95.72 +/- 6.79	72.36 +/- 2.69	84.41 +/- 0.78	43.70 +/- 2.72	0.72 +/- 0.08

Table S1. Summary tables for ***a*** read 1 and ***b*** read 2 of a production sequencing run of human DNA sample NA18507. For each of lanes 1-7, 330 tiles of data were collected and each tile contained approximately 22,000-23,000 clusters providing PF sequence. Lane yield = all PF clusters per lane (n) x read length (0.035 kbases). First cycle intensity (Int) is averaged over all PF reads and given in arbitrary units. % intensity after 20 cycles provides a measure of signal loss during the run. The % PF clusters indicates the fraction of raw clusters that provide PF sequence data after purity filtering (described in supplementary methods). The % align indicates what fraction of the PF reads align to the reference (in this case human genome ncbi36) using ELAND. This value will vary for different reference sequences. The % Error Rate is the average % differences for all bases in the PF dataset when compared to the reference, and includes all differences due true polymorphisms as well as alignment errors and sequence errors. The yield for this run was 3.8 Gb of PF sequence data.

	ELAND			MAQ		
	N	%		N	%	
All PF reads	278.6	100.0%		278.6	100.0%	
PF reads aligning to genome	239.3	85.9%		267.4	96.0%	
Non duplicate PF reads aligning to genome	216.8	77.8%	(90.6%)	236.7	85.0%	(88.5%)
Non duplicate PF reads aligning to X chromosome	191.9	68.9%	(88.5%)	203.9	73.2%	(86.2%)

Table S2. Distribution of reads in flow-sorted X chromosome analysis Number of reads N is given in millions. Percent values not in brackets are fraction of the total number of reads (top line value for N). Percent values in (brackets) express the proportion of reads as a fraction of the category on the previous line. Thus using ELAND, non duplicate PF reads aligning to genome are 90.6% of all PF reads aligning to genome. They represent 77.8% of all PF reads.

Illumina Exhibit 1006, Page 57 of 67

		Genotyped	ELAND		MAQ	
			Loci		Loci	
		n	n	%	n	%
Covered by sequence	Homozygote	8,887	8,865	99.75%	8,885	99.98%
	Heterozygote	4,717	4,708	99.81%	4,707	99.79%
	All	13,604	13,573	99.77%	13,592	99.91%
Concordant calls	Homozygote		8,865	100.00%	8,885	100.00%
	Heterozygote		4,643	98.62%	4,706	99.98%
	All		13,508	99.52%	13,591	99.99%
All disagreements			65	0.48%	1	0.01%
GT>Seq			65	0.48%	1	0.01%
Seq>GT			0	0.00%	0	0.00%
Other discordances			0	0.00%	0	0.00%

Table S3. Comparison of X chromosome SNP calls made from sequence vs. genotype data. Genotyping of the NA07340 sample was carried out using the Illumina HumanHap550 BeadChip. GT>Seq denotes a heterozygous genotyping SNP call where there is a homozygous sequencing SNP call (one of the two alleles); Seq>GT denotes the converse, i.e. a heterozygous sequencing SNP call where there is a homozygous genotyping call. Other discordances are differences in the two SNP calls that cannot be accounted for by one allele being missing from one call.

Illumina Exhibit 1006, Page 58 of 67

Left end	Right end	Length	MAQ APs	ELAND APs	Type	Repeat	Depth	Ext. evi
402924	403049	130	9	5	hom		1	3,4
435713	436369	630	11	3	hom		0	4
823102	823947	888	7	8	het		1	
827894	828076	167	9	10	hom		0	
830122	830892	735	5	3	het		1	
1130746	1131006	348	17	9	het		0	
1574560	1574825	226	10	7	hom		0	
1871088	1871680	628	4	9	het		1	4
2204889	2204968	173	5	7	het		1	4
2618720	2618890	201	12	6	hom		1	4
3044385	3044590	194	9	3	hom		0	4
4536875	4537115	248	13	11	hom		1	3,4
4914723	4915023	328	29	24	hom	AluYa5	0	2,3
5065353	5067496	2164	27	24	het		1	4
5580355	5580671	356	58	50	hom	AluYa5	1	1,2,3,4
6730947	6731159	182	5	2	hom		0	3,4
8746525	8748362	2349	6	12	het		1	4
9720010	9720194	158	8	7	hom		1	3
11635293	11641318	6069	36	35	hom	L1HS	1	1,3,4
11863127	11869352	6258	57	42	hom	L1HS	1	1,4
16337623	16338127	550	47	37	hom	L1HS	1	1,3
17430556	17430853	336	61	47	hom	AluY	1	1,3
22582505	22582796	334	34	25	hom	AluY	1	
32380501	32380550	172	6	3	hom		0	
33118510	33118820	332	31	29	hom	AluYb8	1	2,3
38281972	38282205	222	3	6	het		1	
38469315	38473453	4170	3	6	het		1	3,4
39520762	39521076	295	5	5	hom		1	1,2,3
45472524	45472804	320	58	54	hom	AluY	1	2,3
52904522	52906807	2288	11	8	het		0	
58406935	58407152	200	1	6	hom		0	3
66020510	66020630	154	18	15	het		0	3
66250543	66250693	197	51	45	hom		1	1,3,4
71039962	71040256	342	72	51	hom	AluYa5	1	2,3,4
75946512	75946687	200	20	14	het		0	
77583524	77584097	614	63	46	hom	L1HS	1	3
78833108	78835525	2408	16	25	hom		1	4
80929136	80929745	610	7	8	het		1	1,3,4
80983316	80989344	6067	13	11	hom	L1HS	0	4
84229771	84229872	134	11	6	hom		1	
86165708	86166809	1116	23	24	hom		1	1
86173414	86173872	478	7	7	het		1	
88239424	88239563	166	13	27	het		1	
89794851	89795060	159	5	5	het		0	4

Table S4 (part 1 of 2). See part 2 for legend.

Illumina Exhibit 1006, Page 59 of 67

Table S4 (part 2 of 2). Analysis of structural variants in X chromosome data for NA07340. Maq or Eland APs refer to the number of anomalous pairs (APs) that support the variant. A variant was called if the number of APs was 5 or more in one analysis. Thereafter, supporting data in the other analysis is recorded if there are 1 or more APs in the other analysis. Depth indicates if there was evidence for the variant on the basis of read depth (1) or not (0) in addition to the AP data. Overlap with independent datasets is indicated in the final column by 1: Evidence for a variant obtained from split fosmid reads from the same DNA detected by alignment to the reference using Ssaha; 2: entry in dbRIP; 3, entry in DGVindel (see refs 15, 23) or 4: Database of Genomic Variants. We found 39 variants in the present dataset that were supported by evidence from independent analyses of category 1, 2 or 3. Note in the case of 4 that the overlap can be across a very large region and provides corroboration but does not necessarily constitute strong supporting evidence for detection of the same event.

Illumina Exhibit 1006, Page 60 of 67

start[1]	length[1]	Reference copy number in region at read-length (32 bp) resolutior	Copy number difference in NA07340 (diploid)[2]	Corroborated by external data[3]
4000	12000	Mostly 2	loss (-2)	1
16375	18350	Variable between 20 and 120	gain (+++)	
80000	15000	Mostly 2	loss (-2)	
97000	9000	Variable between 20 and 120	gain (+++)	
408025	1200	Approximately 40	gain (+++)	2
930225	750	Variable between 10 and 120	gain (+++)	
950000	180000	At least 2	loss (-2)	
1715775	4500	Approximately 25	loss (---)	
3750000	120000	At least 2	loss (-2)	2
9332000	11000	Mixture of 5 and 10	gain (+++)	1,2
19375625	4000	Mostly unique	loss (-1)	1
48771725	3900	Mostly 2	loss (-2)	2
49050000	210000	Tandem repeat of 16 units	gain (+++)	
49708725	300	Approximately 15	gain (+++)	1
52100000	500000	Mixture of 2 and 5	loss (--)	
52844000	6000	Variable between 2 and 10	loss (--)	
55220000	50000	Variable between 2 and 10	gain (+++)	
56810000	14000	Variable between 3 and 5	gain (++)	1,2
62250000	180000	Mixture of 2 and 10	gain(+2)	1,2
69378925	8450	Mostly unique	gain (+1)	1
73025525	200	At least 10	gain (+++)	
77402000	10000	Mostly unique	loss (-1)	1
81868975	5950	At least 5	loss (---)	1
88344000	6000	Variable up to 45	gain (+++)	1
96493275	2050	Mostly unique	loss (-2)	1
101253375	150	At least 10	gain (+++)	
114860000	60000	Tandem repeat of 20 units	gain (+++)	1,2
118758025	3500	Mostly unique	loss (-2)	1
119893875	54650	Tandem repeat of 12 units	loss (---)	2
125433375	1550	Variable between 2 and 6	gain (+++)	2
130832975	3300	Mostly unique	loss (-2)	1
134670000	130000	Variable between 2 and 12	gain (++)	2
139633075	400	Variable between 2 and 10	gain (+++)	
146168000	9000	Mostly 2	gain (++)	2
150831000	12000	Variable up to 10	loss (-1)	
154430975	8650	Mostly unique	loss (-1)	1,2
154866825	4050	At least 2	gain (++)	
154913225	350	At least 90	gain (+++)	

Table S5. Analysis of structural variants in X chromosome data for NA07340 based on read depth only.

[1] There is some imprecision in the definition of the breakpoints as genomic windows are assessed.

[2] ++/-- = 3 to 8 additional/fewer copies; +++/--- = more than 8 additional/fewer copies.

[3] 1 = CGH, 2 = Database of Genomic Variants

Illumina Exhibit 1006, Page 61 of 67

a

	ELAND			MAQ		
	n (billions)	%		n (billions)	%	
All PF reads	3.77	100.0%		3.77	100.0%	
PF reads aligning to genome	3.41	90.5%		3.66	97.1%	
Non duplicate PF reads aligning to genome	3.14	83.3%	(92.0%)	3.33	88.3%	(91.0%)

b

	ELAND		
	n (millions)	%	
All PF reads	296.1	100.0%	
PF reads aligning to genome	270.7	91.4%	
Non duplicate PF reads aligning to genome	228.6	77.2%	(84.4%)

Table S6. Summary of reads in human genome sequencing: ***a*** short insert library (208+/- 13 bp); ***b.*** long insert library (1,840+/- 200bp). The numbers in parentheses shows % retained from the category on the preceding line.

Illumina Exhibit 1006, Page 62 of 67

a

HM550 COMPARISON		Genotyped	ELAND Loci		MAQ Loci	
		n	n	%	n	%
Covered by sequence	Homozygote	390,585	389,139	99.63%	389,909	99.83%
	Heterozygote	162,125	161,357	99.53%	161,359	99.53%
	All	552,710	550,496	99.60%	551,268	99.74%
Concordant calls	Homozygote		388,798	99.91%	389,609	99.92%
	Heterozygote		159,317	98.74%	161,115	99.85%
	All		548,115	99.57%	550,724	99.90%
All disagreements			2,381	0.43%	544	0.10%
GT>Seq			1,940	0.35%	170	0.03%
Seq>GT			258	0.05%	290	0.05%
Other discordances			183	0.03%	84	0.02%

b

HM-All COMPARISON		Genotyped	ELAND Loci		MAQ Loci	
		n	n	%	n	%
Covered by sequence	Homozygote	2,901,392	2,876,278	99.13%	2,882,021	99.33%
	Heterozygote	798,200	795,041	99.60%	791,434	99.15%
	All	3,699,592	3,671,319	99.24%	3,673,455	99.29%
Concordant calls	Homozygote		2,851,483	99.14%	2,855,192	99.07%
	Heterozygote		775,708	97.57%	785,757	99.28%
	All		3,627,191	98.80%	3,640,949	99.12%
All disagreements			44,136	1.20%	32,506	0.88%
GT>Seq			16,821	0.46%	5,602	0.15%
Seq>GT			19,268	0.52%	19,679	0.54%
Other discordances			8,047	0.22%	7,225	0.20%

Table S7. Detailed breakdown of comparison of human genome SNP calls made from sequence vs. genotype data. Genotype data were generated using the Infinium HumanHap550 BeadChip (HM550) (***a***); or obtained from the International HapMap Project (HM-All) (***b***). GT>Seq denotes a heterozygous genotyping SNP call where there is a homozygous sequencing SNP call (one of the two alleles); Seq>GT denotes the converse, i.e. a heterozygous sequencing SNP call where there is a homozygous genotyping call. Other discordances are differences in the two SNP calls that cannot be accounted for by one allele being missing from one call.

Illumina Exhibit 1006, Page 63 of 67

		SNP calls n	%
Genotyped loci	Homozygote	309	
	Heterozygote	897	
	All	1206	
Concordant calls	Homozygote	296	95.8%
	Heterozygote	863	96.2%
	All	1159	96.1%
All disagreements		47	3.9%
GT>Seq		7	0.6%
Seq>GT		22	1.8%
Other discordances		18	1.5%

Table S8. Validation of novel SNP calls by genotyping. Results for 1,206 randomly chosen novel SNP calls. There is 96.1% (1159/1206) concordance overall between sequencing and genotype calls. In all there were 47 discrepancies. GT>Seq denotes a heterozygous call by genotyping, where one of the alleles is detected by sequencing. These seven undercalls are considered to be false negatives. Seq>GT denotes the converse, i.e. a heterozygous sequencing SNP call where there is a homozygous genotyping call. On manual inspection we found that there were two classes of Seq>GT: Ten have nearby hidden SNPs and are therefore considered correct sequence calls. The other twelve lie in small indels that were missed in by the indel caller. These are therefore counted as false positive SNP calls in the present analysis. Other discordances are differences in the two SNP calls that cannot be accounted for by one allele being missing from one call and are also counted as false positives. The total false positive discovery rate among the novel SNPs in this analysis is therefore 2.4% (30/1206).

Illumina Exhibit 1006, Page 64 of 67

chromosome	position	call	ref	exon info (ensembl id)	gene	change	hom/het	known/novel
c1	16895539	A	C	exon:ENST00000270691	ESPNP	E638STOP	HOM	Known
c1	20374169	GA	G	exon:ENST00000247992	Q5R387	R27STOP	HET	Known
c1	32034879	GA	G	exon:ENST00000360482	SPOCD1	Q724STOP	HET	Novel
c1	46853266	AG	G	exon:ENST00000271139	MOBKL2C	R24STOP	HET	Known
c1	54954888	TA	T	exon:ENST00000395691	C1orf175	C1290STOP	HET	Known
c1	55023902	GA	G	exon:ENST00000371274	TTC22	R342STOP	HET	Novel
c1	86873963	G	C	exon:ENST00000284054	CLCA3	Y84STOP	HOM	Known
c1	89221456	AT	A	exon:ENST00000321792	NP62556.2	Y214STOP	HET	Novel
c1	100454229	CA	C	exon:ENST00000370132	DBT	E224STOP	HET	Novel
c1	110268078	AG	G	exon:ENST00000369803	CSF1	W491STOP	HET	Known
c1	111769644	TC	C	exon:ENST00000369728	OVGP1	W131STOP	HET	Known
c1	116028154	CA	C	exon:ENST00000355485	VANGL1	S338STOP	HET	Novel
c1	120138308	CT	C	exon:ENST00000369402	REG4	W114STOP	HET	Known
c1	143563747	CT	C	exon:ENST00000369356	PDE4DIP	W2351STOP	HET	Known
c1	143626918	GA	G	exon:ENST00000369356	PDE4DIP	R622STOP	HET	Known
c1	156816116	T	C	exon:ENST00000368150	OR10X1	W66STOP	HOM	Known
c1	159359477	GA	G	exon:ENST00000392188	DEDD	Q198STOP	HET	Novel
c1	169379114	T	C	exon:ENST00000236166	FMO6P	Q105STOP	HOM	Known
c1	230165470	CT	C	exon:ENST00000359647	Q6ZRE3	R116STOP	HET	Known
c1	234772923	AT	T	exon:ENST00000352231	LGALS8	L212STOP	HET	Known
c1	246179649	AT	T	exon:ENST00000357191	OR2L8	Y289STOP	HET	Known
c1	246789346	A	G	exon:ENST00000328570	OR2T29	Q24STOP	HOM	Novel
c2	31252918	CG	G	exon:ENST00000398824	N/A	Y708STOP	HET	Known
c2	89400921	GA	G	exon:ENST00000390263	IGKV1-39	R18STOP	HET	Known
c2	95883315	AG	G	exon:ENST00000357042	Q53S06	Q32STOP	HET	Known
c2	159224979	AG	G	exon:ENST00000342892	Q6ZTQ2	Q103STOP	HET	Novel
c2	207528443	AG	G	exon:ENST00000272852	CPO	W59STOP	HET	Novel
c3	15468185	GT	G	exon:ENST00000383782	COLQ	S285STOP	HET	Novel
c3	16243990	AC	C	exon:ENST00000388817	GALNTL2	S162STOP	HET	Known
c3	139505118	AG	G	exon:ENST00000383180	TXNDC6	Q171STOP	HET	Novel
c3	150980779	AG	G	exon:ENST00000383050	N/A	R130STOP	HET	Novel
c4	12948066	GA	G	exon:ENST00000382444	HSP90AB2P	W252STOP	HET	Novel
c4	70933511	AT	T	exon:ENST00000381057	HTN3	Y47STOP	HET	Known
c4	77879853	CA	C	exon:ENST00000296043	SHROOM3	Y501STOP	HET	Novel
c4	102301286	T	C	exon:ENST00000399145	N/A	Q115STOP	HOM	Known
c5	1089962	AC	C	exon:ENST00000382730	NKD2	S23STOP	HET	Known
c5	1293757	CG	C	exon:ENST00000324642	SLC6A18	Y319STOP	HET	Known
c5	32185000	AG	G	exon:ENST00000332294	Q9UI72	W17STOP	HET	Known
c5	75000878	A	G	exon:ENST00000344149	N/A	W419STOP	HOM	Known
c5	134810349	A	T	exon:ENST00000333591	C5orf20	R117STOP	HOM	Known
c6	33156444	T	G	exon:ENST00000395267	HLA-DPB1	G17STOP	HOM	Known
c6	41210331	T	C	exon:ENST00000338759	Q6ZRD8	R20STOP	HOM	Known
c6	43032225	A	G	exon:ENST00000326586	Q5T8W0	R74STOP	HOM	Known
c6	57506229	CT	C	exon:ENST00000389488	PRIM2	Q325STOP	HET	Known
c6	74076059	AG	G	exon:ENST00000370384	C6orf148	Q34STOP	HET	Known
c6	139618237	A	G	exon:ENST00000367652	TXLNB	Q123STOP	HOM	Known
c6	150253727	GA	G	exon:ENST00000367363	RAET1E	R2STOP	HET	Novel
c6	154402262	TC	C	exon:ENST00000330432	OPRM1	Q26STOP	HET	Known
c6	160823771	C	G	exon:ENST00000297289	LPAL2	Y69STOP	HOM	Novel
c7	21549488	TG	G	exon:ENST00000328843	DNAH11	E34STOP	HET	Known
c7	39615800	AG	G	exon:ENST00000357751	Q8N8G3	W43STOP	HET	Known
c7	55508201	GA	G	exon:ENST00000285279	NM30796.3	R83STOP	HET	Known
c7	63806451	A	T	exon:ENST00000344930	ZNF107	C778STOP	HOM	Novel
c7	64076102	AG	G	exon:ENST00000398695	ZNF117	R428STOP	HET	Known
c7	80138385	G	T	exon:ENST00000309881	CD36	Y325STOP	HOM	Known
c7	99072835	A	T	exon:ENST00000379712	Q6ZVR2	C18STOP	HOM	Known
c7	138727857	A	T	exon:ENST00000397584	Q8WYX0	L80STOP	HOM	Known
c7	141589238	A	G	exon:ENST00000389113	N/A	Q390STOP	HOM	Known
c7	141918182	T	C	exon:ENST00000390382	TRBV7-9	R55STOP	HOM	Novel
c7	142459539	TC	C	exon:ENST00000391508	OR6V1	Q80STOP	HET	Known
c7	143995851	TG	G	exon:ENST00000378099	TPK1	S24STOP	HET	Novel
c7	149123706	CT	C	exon:ENST00000262089	Q76B61	R1193STOP	HET	Known
c7	149153299	G	T	exon:ENST00000378016	NM198455.2	L331STOP	HOM	Known

Table S9 (part 1 of 3)

Illumina Exhibit 1006, Page 65 of 67

chromosome	position	call	ref	exon info (ensembl id)	gene	change	hom/het	known/novel
c8	222801	CT	C	exon:ENST00000382855	N/A	Q133STOP	HET	Novel
c8	12480423	TA	T	exon:ENST00000359701	Q8NAJ9	L30STOP	HET	Known
c8	101791075	CA	C	exon:ENST00000318607	PABPC1	E345STOP	HET	Novel
c9	19347166	CT	C	exon:ENST00000380424	DENND4C	Q246STOP	HET	Novel
c9	44114866	GC	G	exon:ENST00000377553	Q8NGA9	Y224STOP	HET	Novel
c9	45060265	CA	C	exon:ENST00000377537	Q6ZVI3	C148STOP	HET	Novel
c9	66297827	TG	T	exon:ENST00000359897	Q6ZS19	Y41STOP	HET	Known
c9	125012928	CA	C	exon:ENST00000398660	N/A	S87STOP	HET	Novel
c9	134946938	GT	G	exon:ENST00000314220	Q8N7A6	S57STOP	HET	Known
c9	138754316	AG	G	exon:ENST00000341040	LCN10	Q148STOP	HET	Known
c9	139057620	A	G	exon:ENST00000371600	NPDC1	R25STOP	HOM	Known
c10	38225046	CA	C	exon:ENST00000374537	N/A	S190STOP	HET	Novel
c10	38934622	CT	C	exon:ENST00000399687	N/A	R26STOP	HET	Novel
c10	46825773	AT	T	exon:ENST00000342900	N/A	L340STOP	HET	Novel
c10	47239426	T	C	exon:ENST00000335083	N/A	R81STOP	HOM	Novel
c10	73634249	CG	G	exon:ENST00000394915	ASCC1	S78STOP	HET	Known
c10	96437552	AT	T	exon:ENST00000285979	CYP2C18	Y68STOP	HET	Known
c10	126668082	GA	G	exon:ENST00000337195	CTBP2	Q445STOP	HET	Known
c11	1662913	C	G	exon:ENST00000382165	FAM99A	S25STOP	HOM	Novel
c11	2278405	TC	C	exon:ENST00000381153	C11orf21	W23STOP	HET	Novel
c11	5733060	AT	A	exon:ENST00000317254	OR52N4	R172STOP	HET	Known
c11	6107204	GA	G	exon:ENST00000316517	Q8NH77	W263STOP	HET	Novel
c11	45029583	GA	G	exon:ENST00000378793	Q6ZS37	R107STOP	HET	Novel
c11	48242807	AT	T	exon:ENST00000320048	OR4X1	Y273STOP	HET	Known
c11	48303590	GA	G	exon:ENST00000395239	OR4C3	W14STOP	HET	Known
c11	48344426	GT	G	exon:ENST00000319813	OR4C5	Y56STOP	HET	Known
c11	60021578	CT	C	exon:ENST00000016913	MS4A12	Q71STOP	HET	Known
c11	92781120	TC	C	exon:ENST00000298050	CCDC67	R460STOP	HET	Novel
c11	104268327	A	G	exon:ENST00000375726	CASP12	R125STOP	HOM	Novel
c11	123561942	TG	T	exon:ENST00000318666	Q8NH80	Y252STOP	HET	Known
c12	3591206	T	A	exon:ENST00000382608	Q6UY24	L71STOP	HOM	Known
c12	50490305	TA	T	exon:ENST00000313893	Q8NBA4	L27STOP	HET	Novel
c12	54231266	TC	C	exon:ENST00000394256	OR6C4	R5STOP	HET	Known
c12	120699406	A	G	exon:ENST00000333310	Q9H9T1	W5STOP	HOM	Novel
c13	17941446	CT	C	exon:ENST00000342944	N/A	W241STOP	HET	Known
c13	18897913	A	G	exon:ENST00000341326	TPTE2	R324STOP	HOM	Known
c13	52020061	TC	C	exon:ENST00000258589	Q5JV89	R138STOP	HET	Novel
c13	111076351	AG	G	exon:ENST00000375713	Q5T400	W3STOP	HET	Novel
c14	22668816	GA	G	exon:ENST00000334354	SLC7A8	R179STOP	HET	Known
c14	24173254	GA	G	exon:ENST00000382542	GZMB	Q19STOP	HET	Known
c14	49872477	C	A	exon:ENST00000356146	CDKL1	L989STOP	HOM	Known
c14	74228760	GA	G	exon:ENST00000338772	Q86TS6	R95STOP	HET	Known
c14	105161972	T	G	exon:ENST00000390543	IGHG4	S290STOP	HOM	Known
c15	18726530	T	C	exon:ENST00000400226	LOC440368	W124STOP	HOM	Known
c15	19000266	CA	C	exon:ENST00000400222	NM1001413.2	E487STOP	HET	Novel
c15	53510174	AC	C	exon:ENST00000321149	DYX1C1	E417STOP	HET	Novel
c15	67114791	GA	G	exon:ENST00000388866	NOX5	W300STOP	HET	Known
c15	95102199	TC	C	exon:ENST00000398303	N/A	R167STOP	HET	Known
c16	764474	A	G	exon:ENST00000293892	NP1020361.1	Q590STOP	HOM	Novel
c16	977150	CT	C	exon:ENST00000305973	Q96S05	Q39STOP	HET	Novel
c16	1373566	AG	G	exon:ENST00000397462	UNKL	Q744STOP	HET	Novel
c16	33537201	AG	G	exon:ENST00000354689	N/A	W7STOP	HET	Known
c16	33537474	AG	G	exon:ENST00000354689	N/A	W66STOP	HET	Known
c16	33868594	GA	G	exon:ENST00000380143	Q6PQ33	R450STOP	HET	Known
c16	88172869	CG	C	exon:ENST00000268720	CPNE7	S154STOP	HET	Novel
c16	88655877	AG	G	exon:ENST00000325921	Q9NQW5-2	R73STOP	HET	Novel

Table S9 (part 2 of 3)

Illumina Exhibit 1006, Page 66 of 67

chromosome	position	call	ref	exon info (ensembl id)	gene	change	hom/het	known/novel
c17	3142353	TA	T	exon:ENST00000323404	OR3A1	K92STOP	HET	Known
c17	4744490	A	G	exon:ENST00000381365	Q6ZR85	W152STOP	HOM	Known
c17	20710491	GT	G	exon:ENST00000327925	NP1004306.1	S178STOP	HET	Known
c17	21144803	CT	C	exon:ENST00000342679	MAP2K3	Q102STOP	HET	Known
c17	36559089	GA	G	exon:ENST00000343246	KRTAP4-5	R153STOP	HET	Known
c17	37237375	C	G	exon:ENST00000393910	NT5C3L	S137STOP	HOM	Known
c17	64723626	TG	A	exon:ENST00000269081	ABCA10	Y261STOP	HET	other
c17	70100401	AC	C	exon:ENST00000328023	C17orf77	C207STOP	HET	Known
c18	50112154	AC	A	exon:ENST00000307844	STARD6	L114STOP	HET	Novel
c19	776155	TC	C	exon:ENST00000334630	AZU1	Q20STOP	HET	Novel
c19	9098263	GA	G	exon:ENST00000305444	OR7G3	R122STOP	HET	Known
c19	40410860	T	C	exon:ENST00000324675	TMEM162	W188STOP	HOM	Known
c19	48463827	GA	G	exon:ENST00000270077	PSG9	R127STOP	HET	Novel
c19	55246050	GA	G	exon:ENST00000377006	Q6ZNZ3	W93STOP	HET	Known
c19	56822743	GA	G	exon:ENST00000222107	SIGLEC5	R356STOP	HET	Known
c19	62334594	A	C	exon:ENST00000391707	USP29	Y913STOP	HOM	Known
c19	63065900	AT	A	exon:ENST00000396146	Q9NWJ2	R58STOP	HET	Novel
c19	63785276	TC	C	exon:ENST00000312426	NP76428.2	Q26STOP	HET	Known
c20	25703913	CA	C	exon:ENST00000376403	Q5T319	E15STOP	HET	Novel
c20	26032175	GA	G	exon:ENST00000356439	C20orf191	R81STOP	HET	Known
c20	55497442	AT	A	exon:ENST00000243919	HMG1L1	Y16STOP	HET	Known
c21	27137937	T	C	exon:ENST00000382884	ADAMTS1	W4STOP	HOM	Known
c21	45588725	GA	G	exon:ENST00000215202	C21orf111	Q188STOP	HET	Known
c22	15849049	AC	C	exon:ENST00000400588	GAB4	G163STOP	HET	Known
c22	21037728	CT	C	exon:ENST00000390293	IGLV5-48	Q106STOP	HET	Novel
c22	25398153	TC	C	exon:ENST00000382641	Q6ZVA3	W143STOP	HET	Known
c22	34886714	AG	G	exon:ENST00000349314	APOL3	Q58STOP	HET	Known
c22	37900350	CA	C	exon:ENST00000330414	YV004	G9STOP	HET	Known
c22	41000237	GT	G	exon:ENST00000332965	Q8IXR4	G223STOP	HET	Known
cX	30999550	A	C	exon:ENST00000359202	FTHL17	E148STOP	HOM	Known
cX	118489029	GT	G	exon:ENST00000317881	SLC25A5	E293STOP	HET	Known
cX	138492281	C	G	exon:ENST00000370573	MCF2	S857STOP	HOM	Novel
cX	142795134	G	T	exon:ENST00000370494	UBE2NL	L89STOP	HOM	Known
cMT	9378	A	G	exon:ENST00000391562	Q14Y83	W57STOP	HOM	Known

Table S9 (part 3 of 3). Coding SNPs predicted to cause premature termination of translation in NA18507. Note that this list excludes positions where the reference sequence also contains a termination codon within annotated coding sequence.

Illumina Exhibit 1006, Page 67 of 67