

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 precipitation³⁴.

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'ACACTCTTCCCTACACGACGCTCTTCCGATC_xT (x = phosphorothioate bond) and 5'-phosphate-GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG. The adapters for the preparation of paired read libraries comprised an alternative set of two oligonucleotides: 5'ACACTCTTCCCTACACGACGCTCTTCCGATC_xT (x = phosphorothioate bond) and 5'-phosphate-GATCGGAAGAGCGGTTTCAGCAGGAATGCCGAG. Ligation products in the

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'-CAAGCAGAAGACGGCATAACGAGCTCTTCCGATC_xT and 5'-

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC_xT; and for paired-read libraries: 5'-

CAAGCAGAAGACGGCATAACGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATC_xT and 5'-

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC_xT (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 (Dyna). 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 (K₂S₂O₈). Flow cell channels were

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-TTTTTTTTTTCAAGCAGAAGACGGCATAACGA-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-TTTTTTTTTTCAAGCAGAAGACGGCATAACGAGoxoAT-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.

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'-ACACTCTTCCCTACACGACGCTCTCCGATCT -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'-CGGTCTCGGCATTCTGCTGAACCGCTCTCCGATCT -3')(see fig 1c).

Sequencing on the Genome Analyser.

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

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 caller¹⁰ ($Q=-10\log_{10}(p_{\text{error}})$, where p_{error} is the probability of an incorrect base), in that the highest of the four scores - that of the called base - is

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.

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 $\log_{10} p(\text{observed bases} \mid \text{no "A"s are present})$ 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

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 were 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 $\geq Q20$) anomalous pairs, singletons and their non-aligning partner in the region of interest and attempted an assembly using Velvet³⁵. 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

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

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 Velvet³⁵ 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

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.

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).

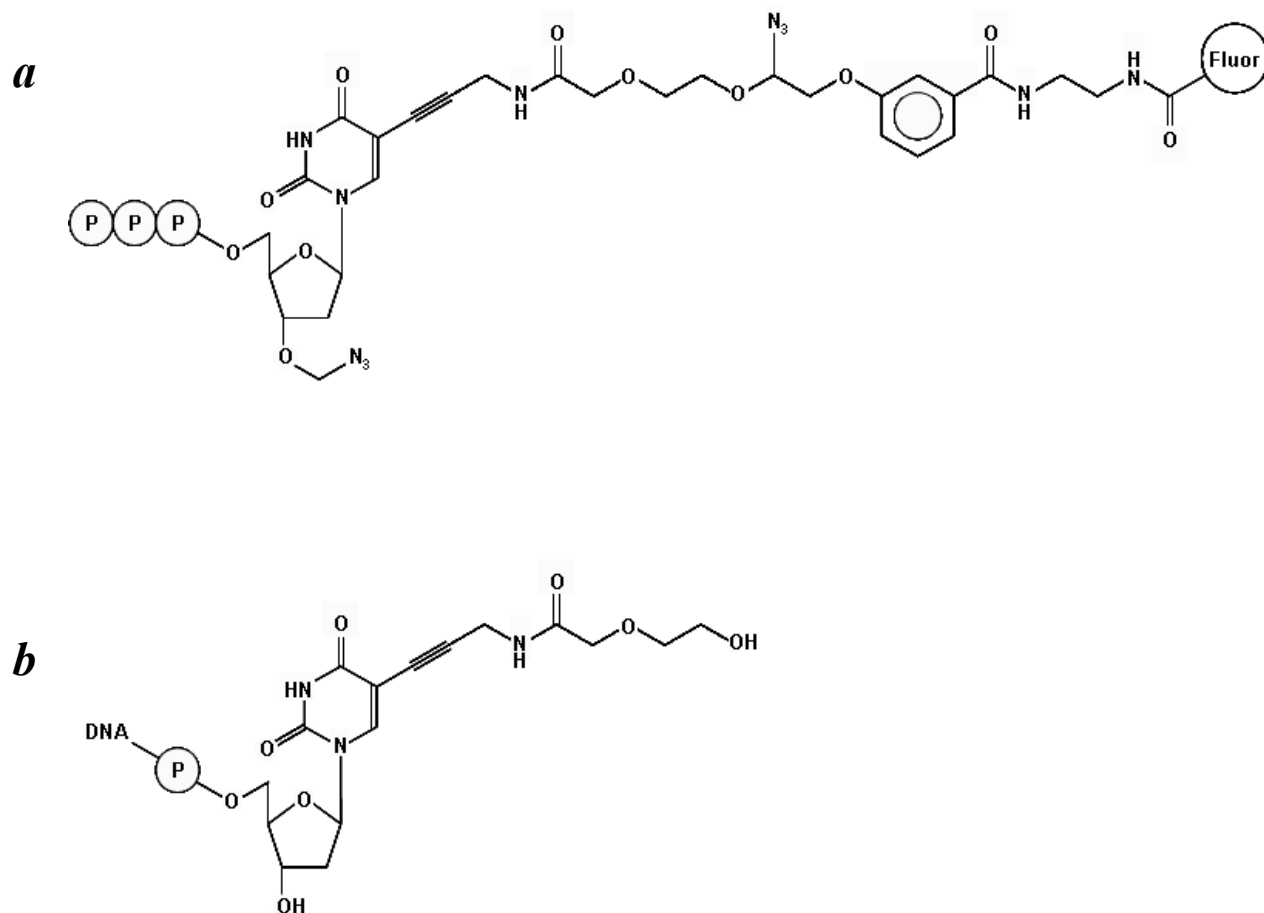


Figure S1. *a*. Structure of the reversible terminator 3'-*O*-azidomethyl 2'-deoxythymine triphosphate (T) labelled with a removable fluorophore. *b*. Structure of the incorporated nucleotide after removal of the fluorophore and terminator group. Each of the four nucleotides have an equivalent structure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

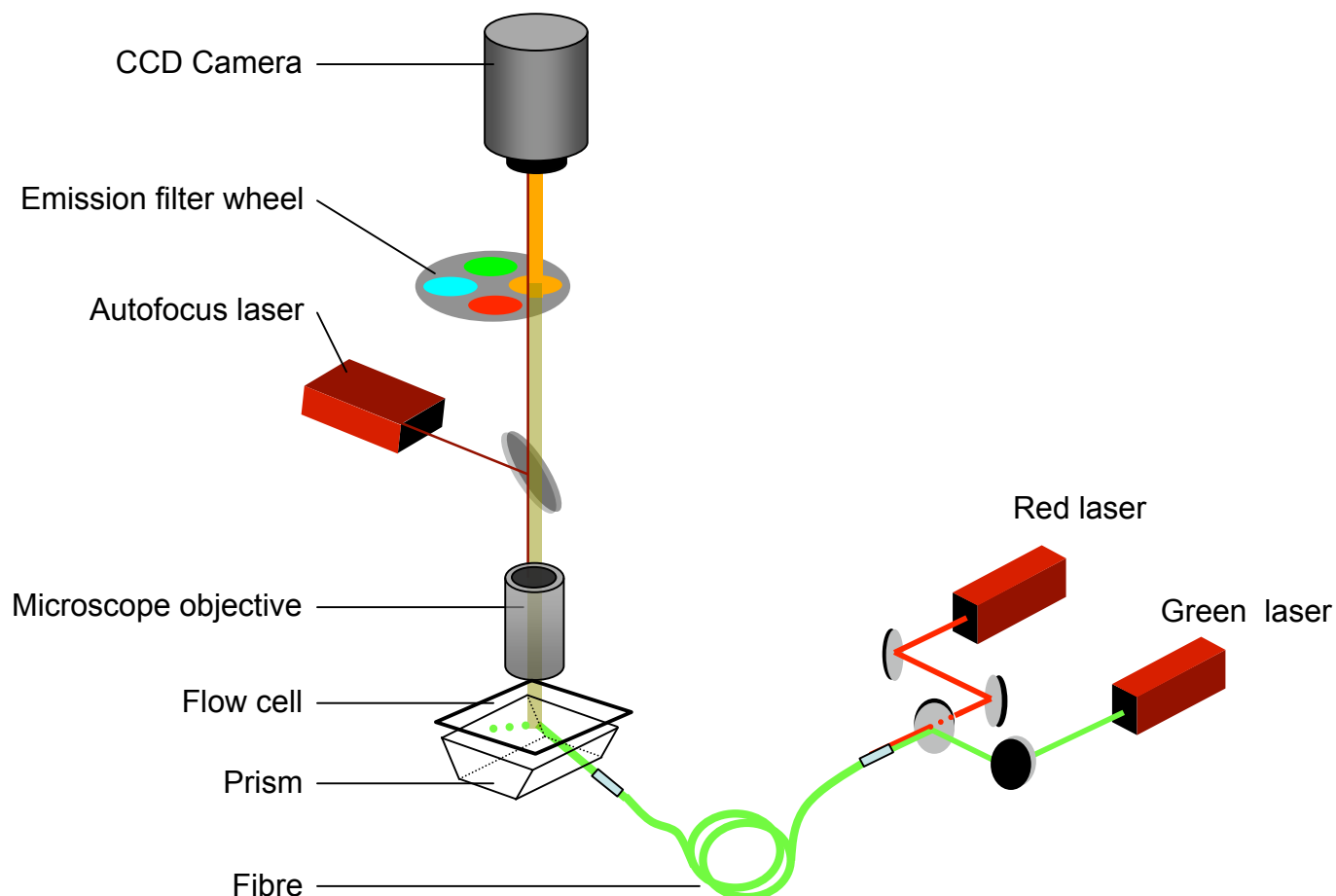


Figure S2a. Optical path of the Genome Analyser (GA1). Red (660 nm) and green (532 nm) lasers provide excitation beams that are directed along an optical fibre and through a prism which is in contact with the flow cell. Excitation of fluorescent nucleotides incorporated into DNA clusters on the inner 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).

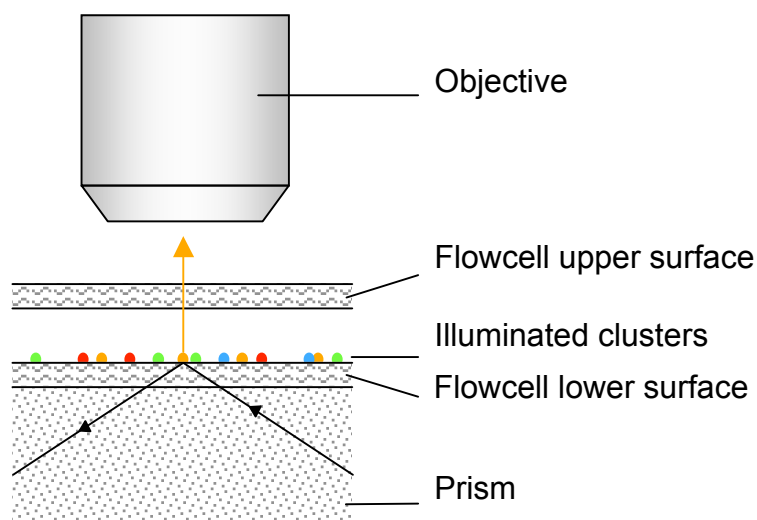


Figure S2b. Total internal reflection of the incident excitation beam at the glass-buffer interface generates an evanescent wave that excites the clusters on the surface. The fluorescence emission is captured by a custom made microscope objective, passed through a filter and is then projected onto a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

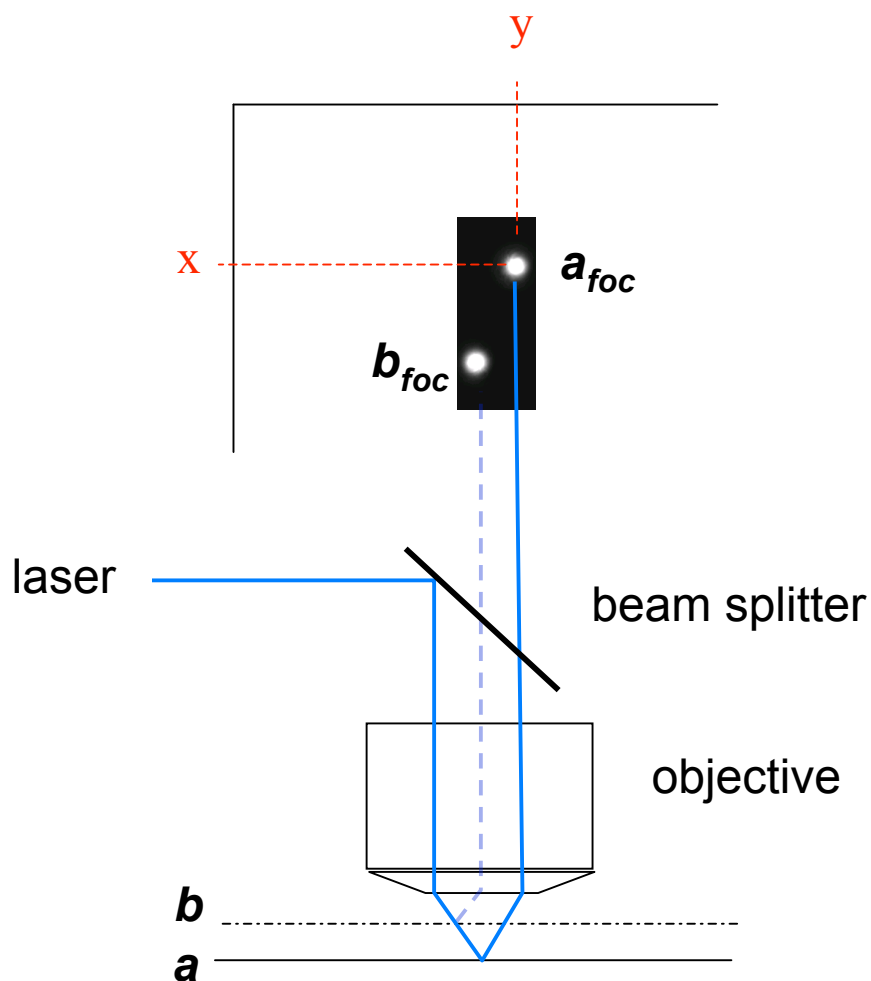
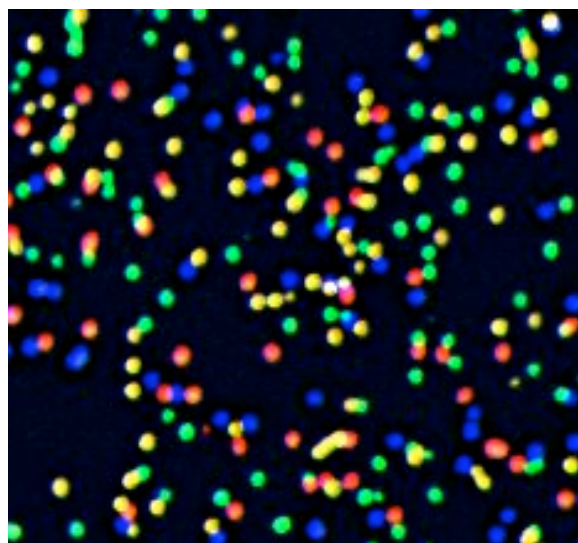


Figure S2c. Autofocus principle of operation. A reference laser beam (solid blue line) is deflected by a beam splitter along a path parallel to the optic axis of the microscope objective and is reflected from the surface of interest. Reflection from a focus plane a generates an autofocus spot a_{foc} . 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 b_{foc} . Focus is automatically adjusted (z movement) as required to return the autofocus spot to the pre-registered correct position before image capture.



20 microns

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.

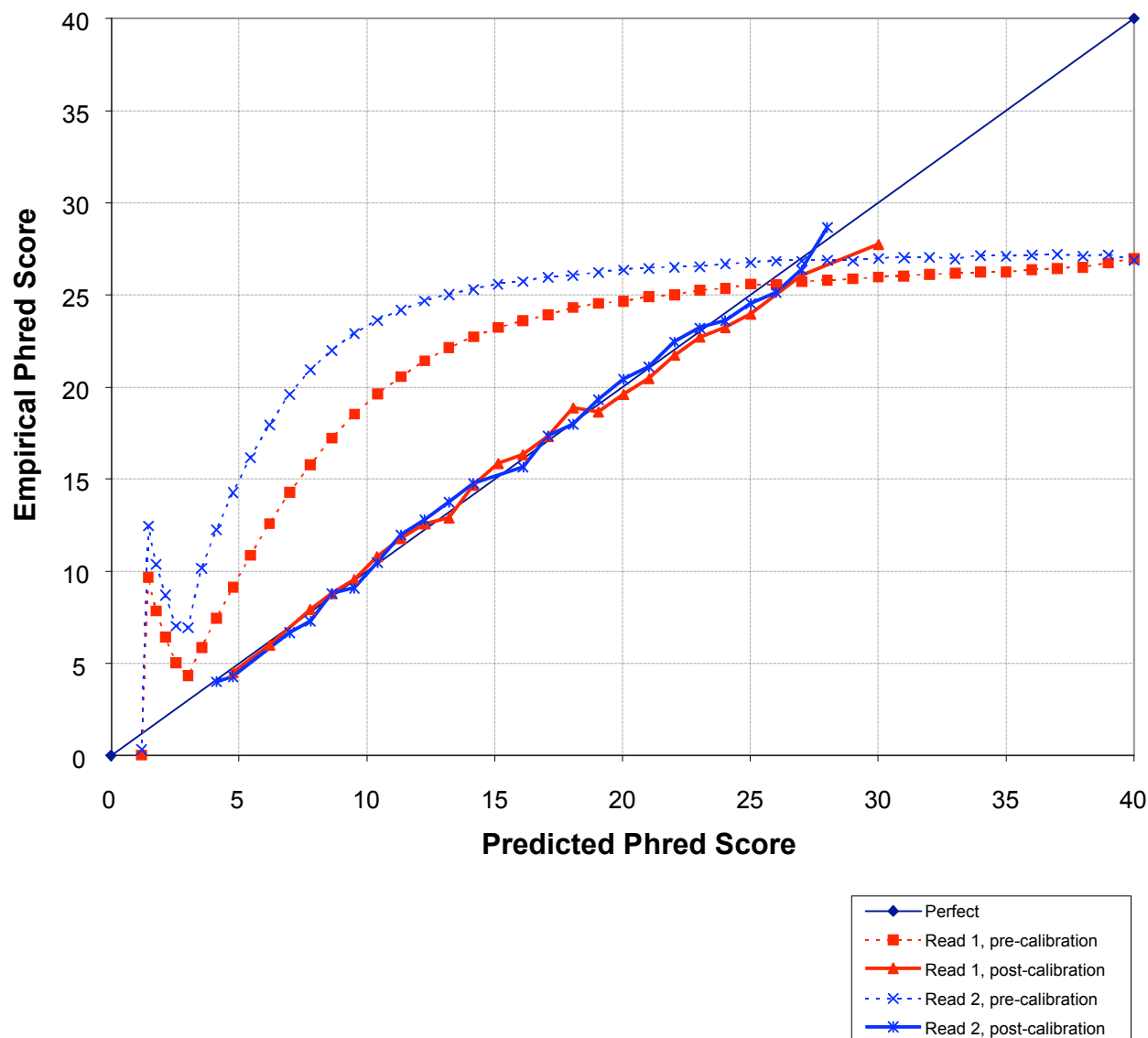


Figure S3. Effect of auto-calibration procedure on base quality scores (see supplementary methods for details). The data set is one lane of paired-read data from the NA18507 individual comprising 8,960,643 purity-filtered clusters, two 35-base reads being obtained from each cluster. Dotted lines show 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 $Q = -10 \log_{10}(\text{perror})$, 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.

a 080531_EAS192_0038_FC20GG5 – lane 5

Read	Lane Info		Tile Mean +/- SD for Lane		1st Cycle Int (PF)	% intensity after 20 cycles (PF)	% PF clusters	& Align (PF)	Alignment Score (PF)	% Error Rate (PF)
	Lane	Lane Yield (kbases)	Clusters (raw)	Clusters (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

Cumulative errors per read by cycle for read 1

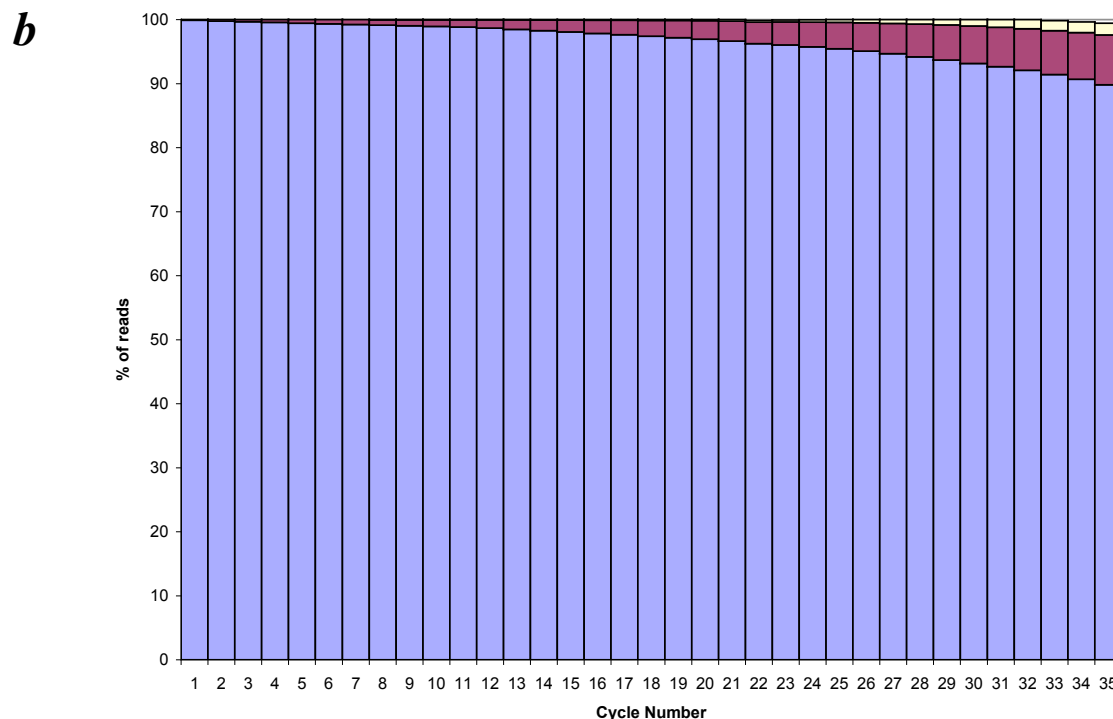


Figure S4. Sequencing summary data for the human BAC from one lane of a flow cell. **a**, summary table for lane 5 (reads 1 and 2). Lane yield = all PF clusters per lane x read length (0.035 kbases). First cycle intensity (Int) is averaged over all PF reads and given in arbitrary units. % intensity after 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.

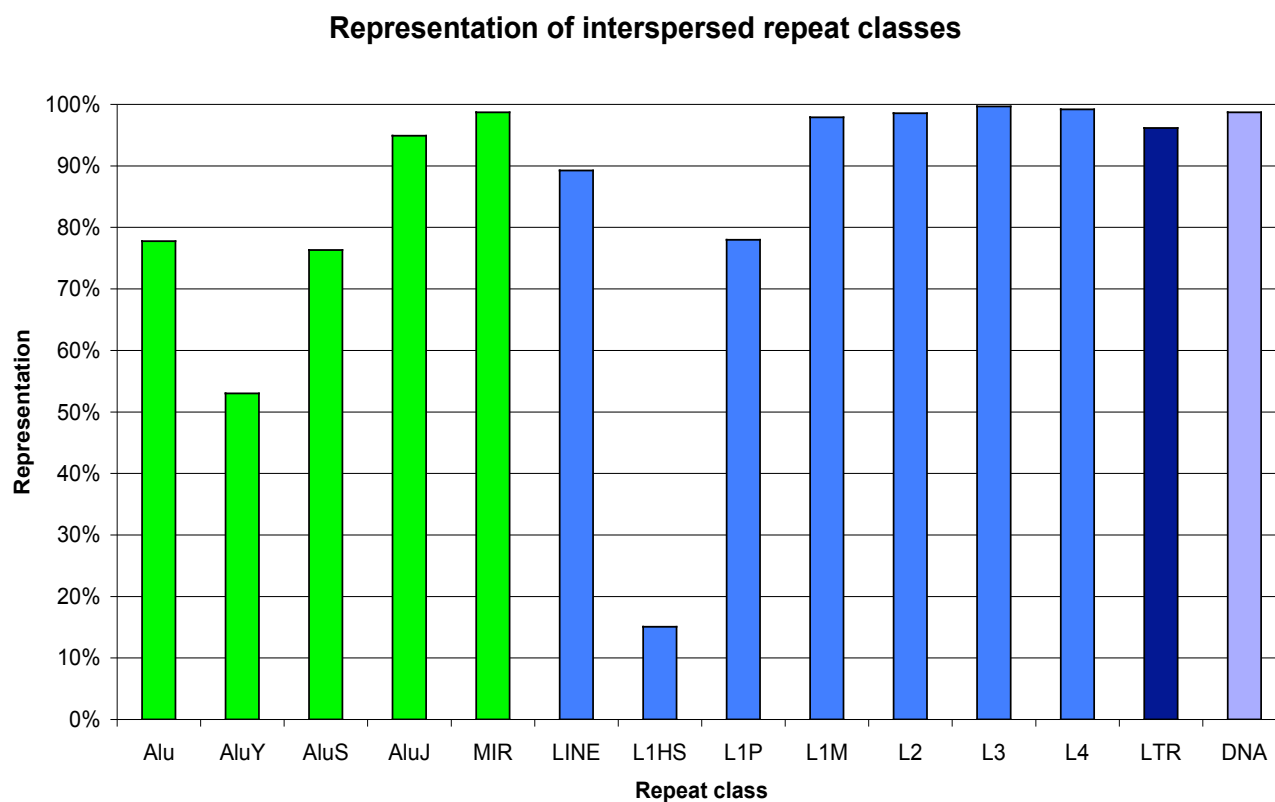


Figure S5. Representation of interspersed repeat classes in the ELAND alignment of the X chromosome dataset. Two-thirds of the low or zero sequence depth in the ELAND alignment is at interspersed repeats. The near-identical repeats arising from recently inserted retroposon elements L1Hs and members of 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.

a

Call	ELAND		MAQ	
	SNPs n	In dbSNP %	SNPs n	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%

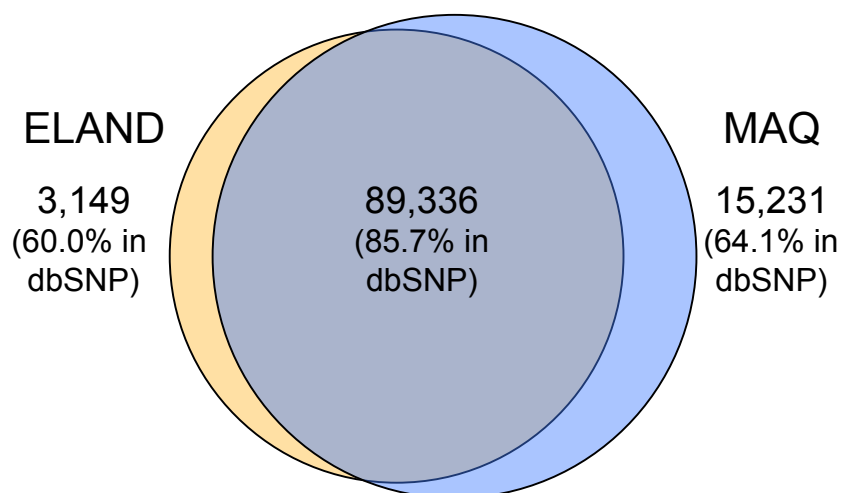
b

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.

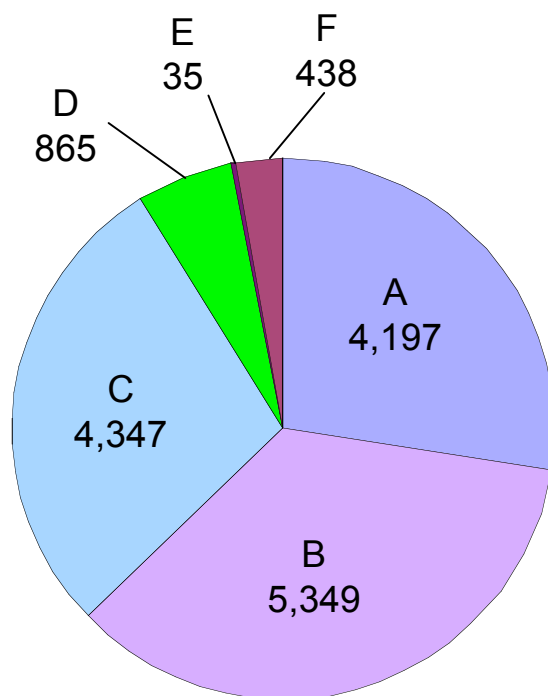
c

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.

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

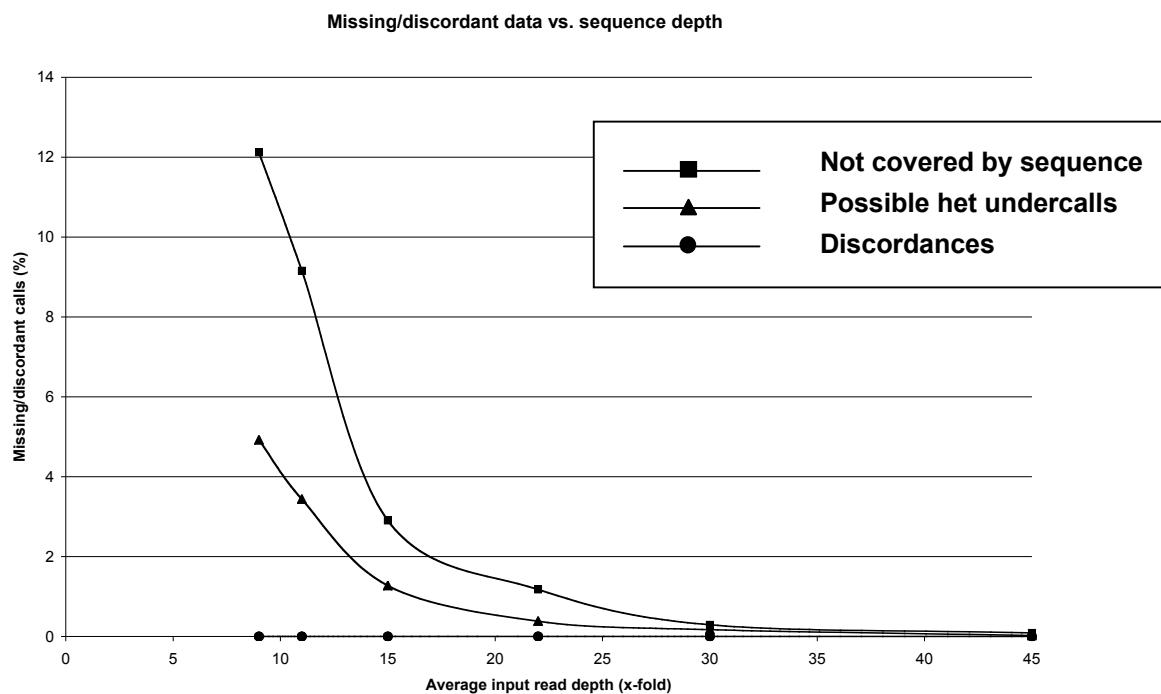
b

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).

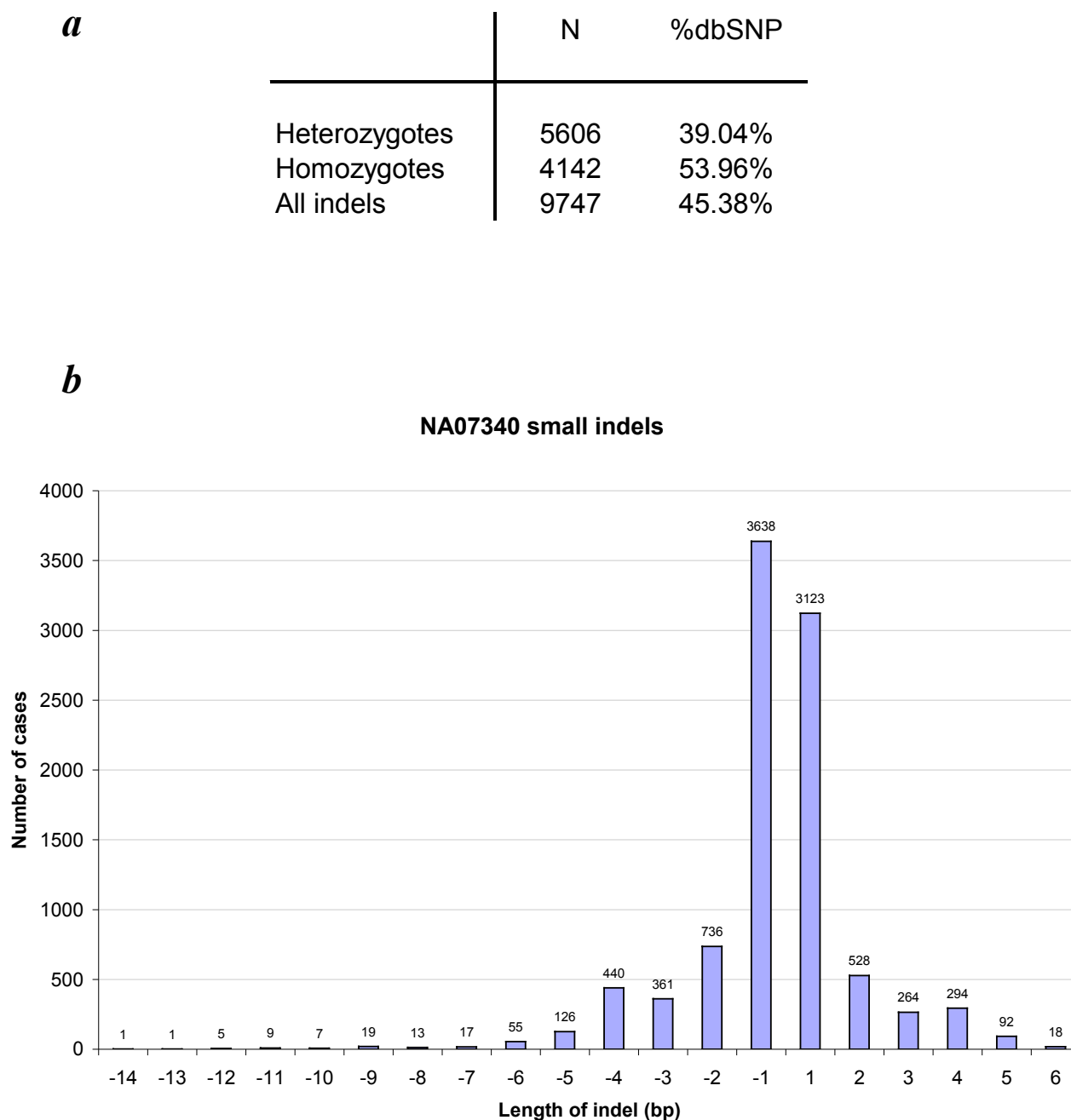


Figure S8. Analysis of short indel calls in X chromosome data for NA07340. **a.** total number of calls and fraction that match previous entries in dbSNP. **b.** Distribution of size in the 9747 indels. + and – values on the x axis correspond to presence or absence of bases in NA07340 relative to the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

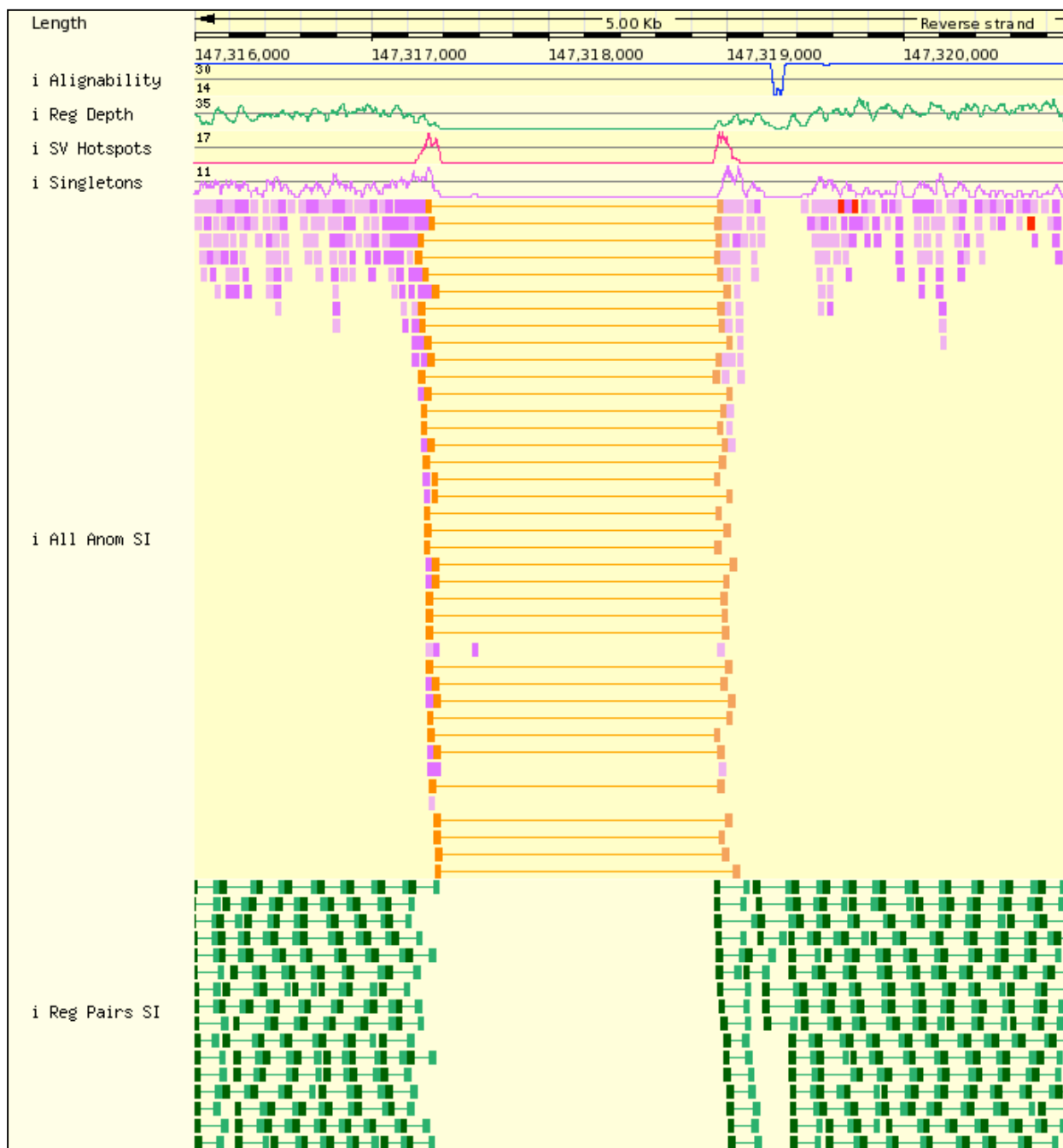


Figure S9a. homozygous 1.6 kb deletion in NA07340 relative to the reference, detected by anomalously spaced read pairs (orange) compared to regularly spaced read pairs (green). Purple reads are singletons, i.e. reads where the other member of the pair does not align to the reference. Note also the alignability 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.

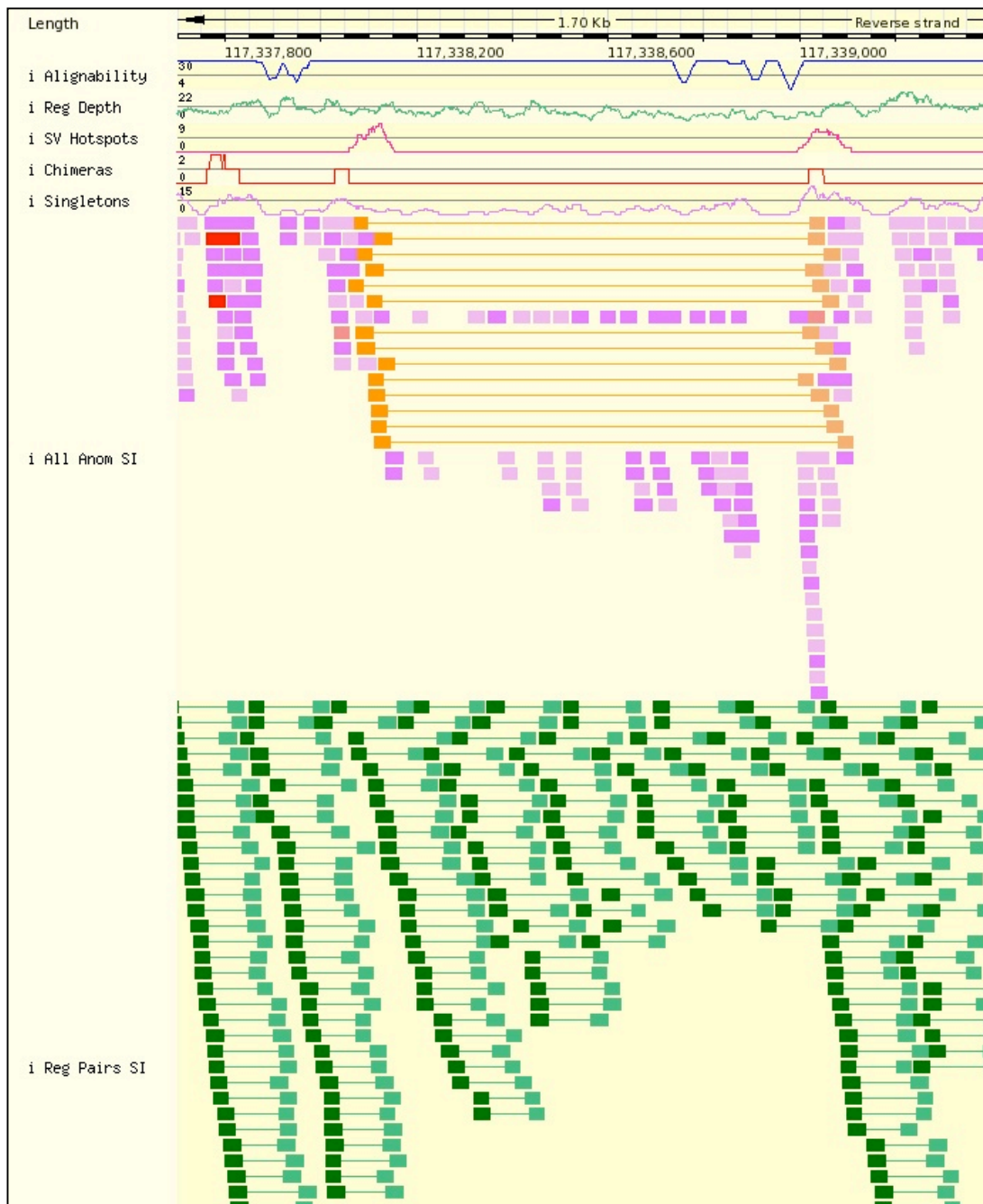


Figure S9b. Heterozygous 844 bp deletion in NA07340 relative to the reference, detected by anomalously spaced read pairs (orange) compared to regularly spaced read pairs (green). The presence of regularly spaced pairs across the entire region is indicative of heterozygosity at this locus. Note that this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

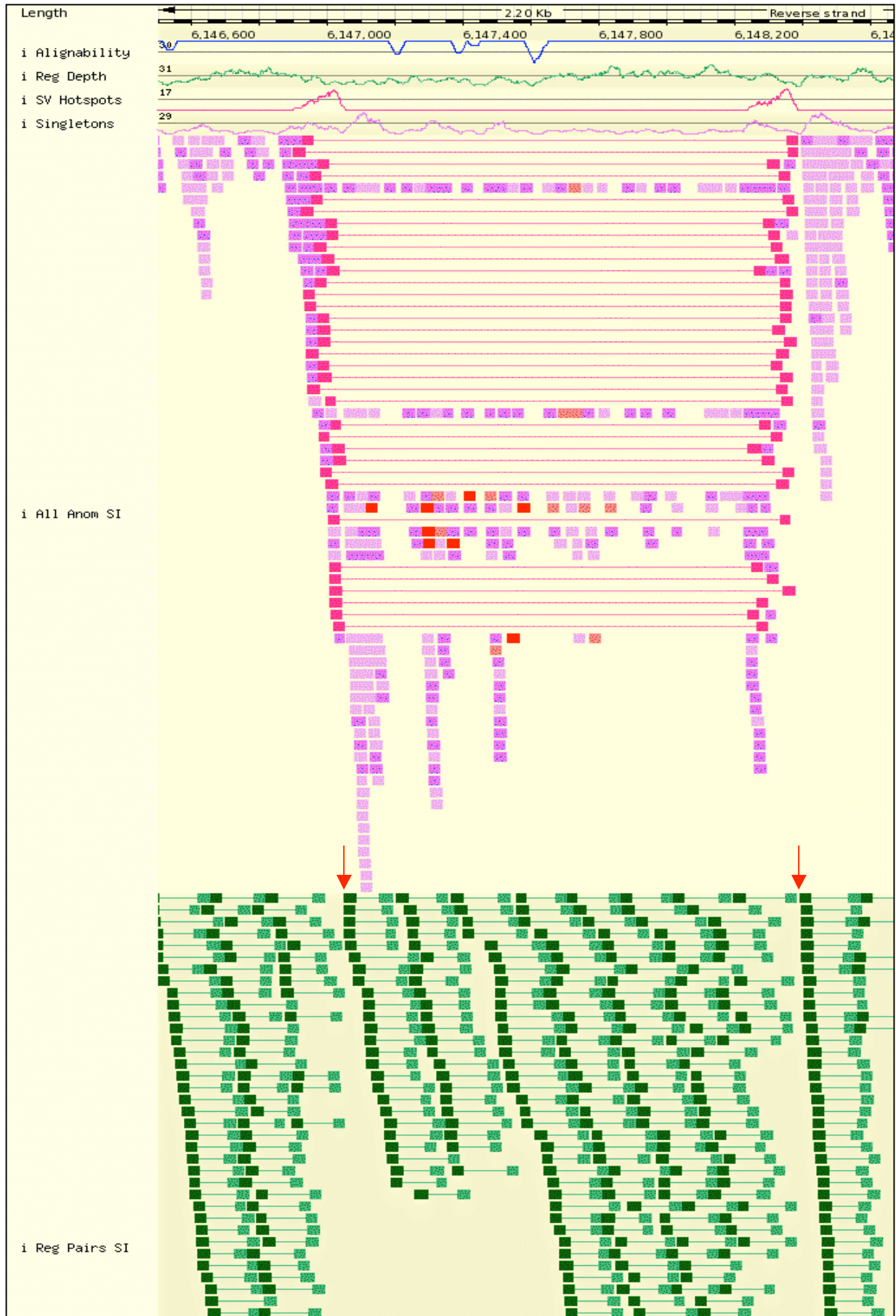


Figure S10. (See next page for legend).

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.

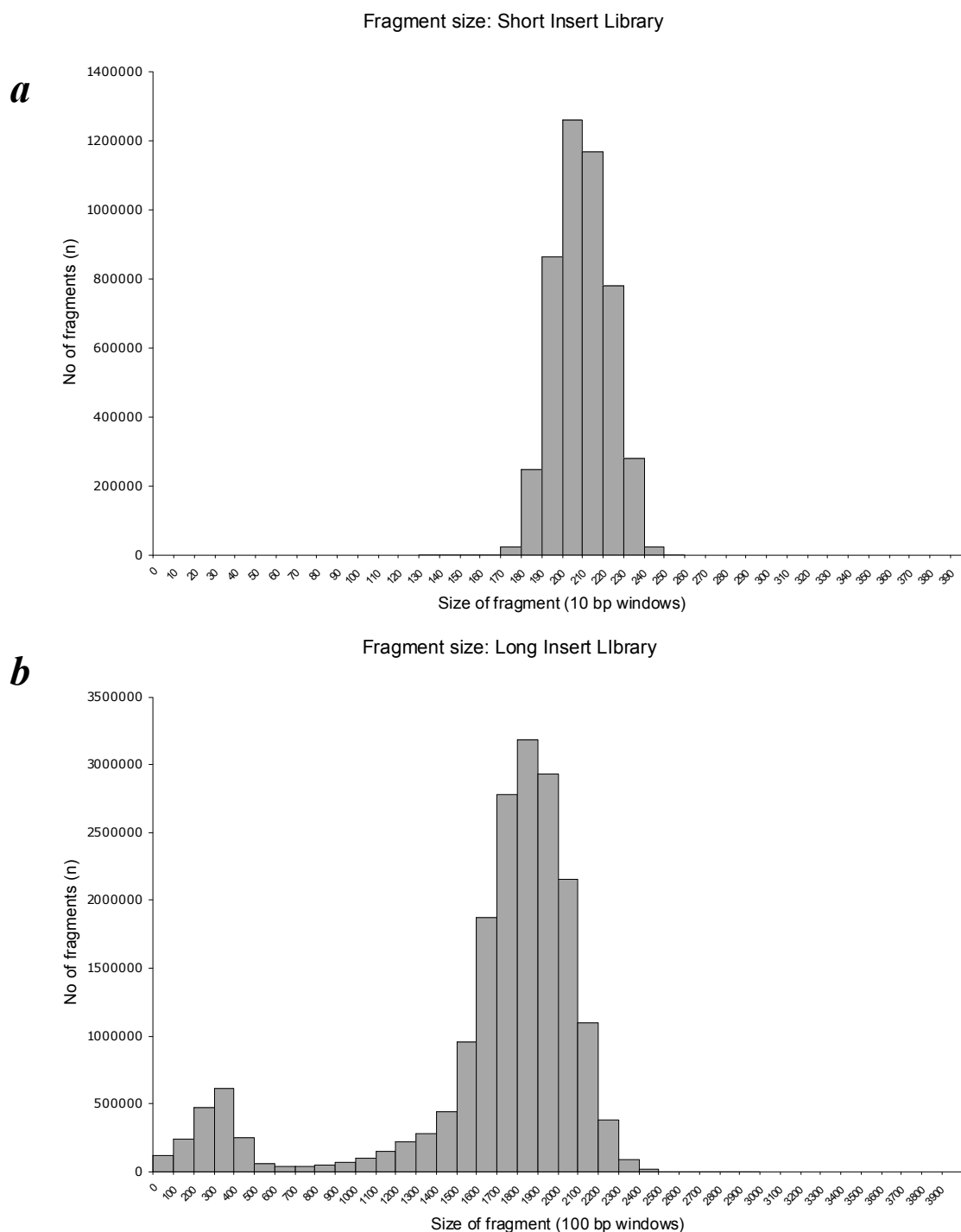


Figure S11 Size distribution for DNA fragment libraries of NA18507. Fragment lengths were determined from the separation between read pairs aligned to the reference. **a.** Short insert library, data from one lane of a flowcell were analysed and fragment lengths binned in 10 bp windows. **b.** Long insert 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).

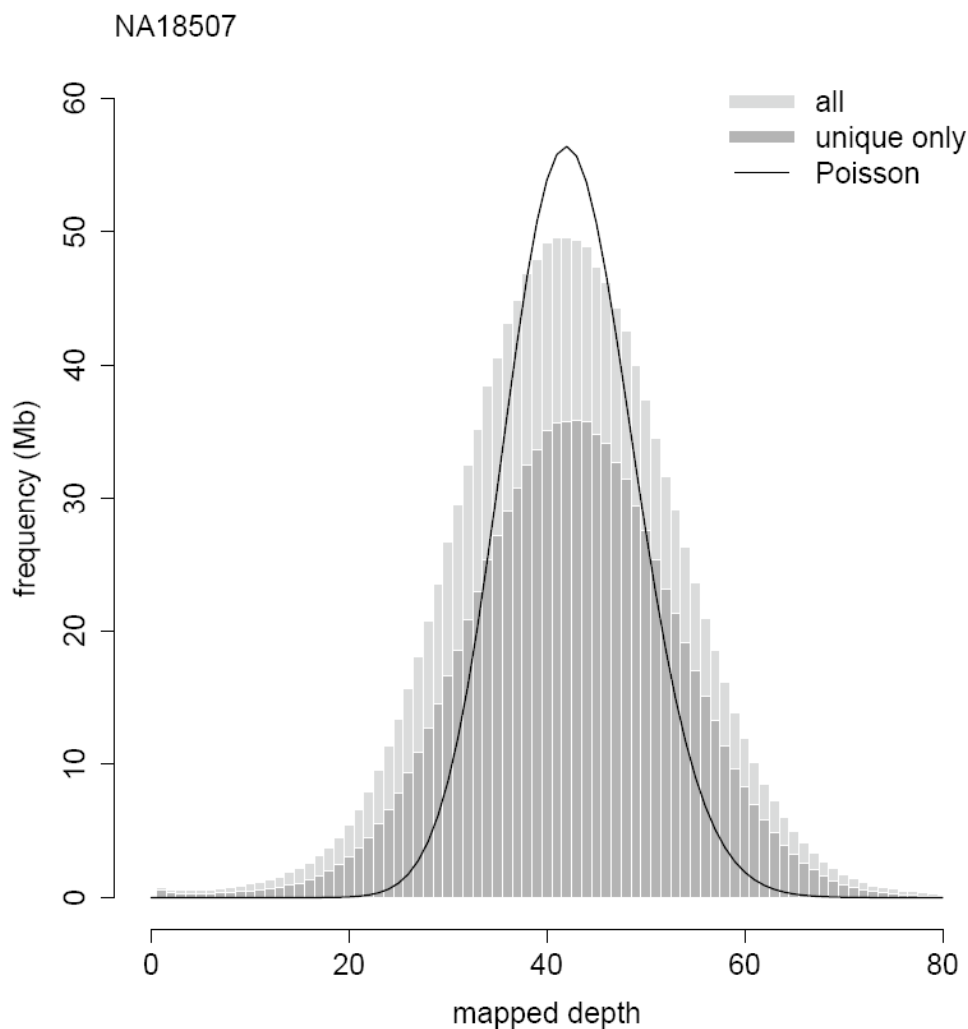


Figure S12a. Distribution of mapped read depth in the human genome dataset, sampled at every 500th position along the genome and displayed as a histogram ('all'). An equivalent analysis of mapped read depth for the unique subset of these positions is also shown ('unique only'). The solid line represents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

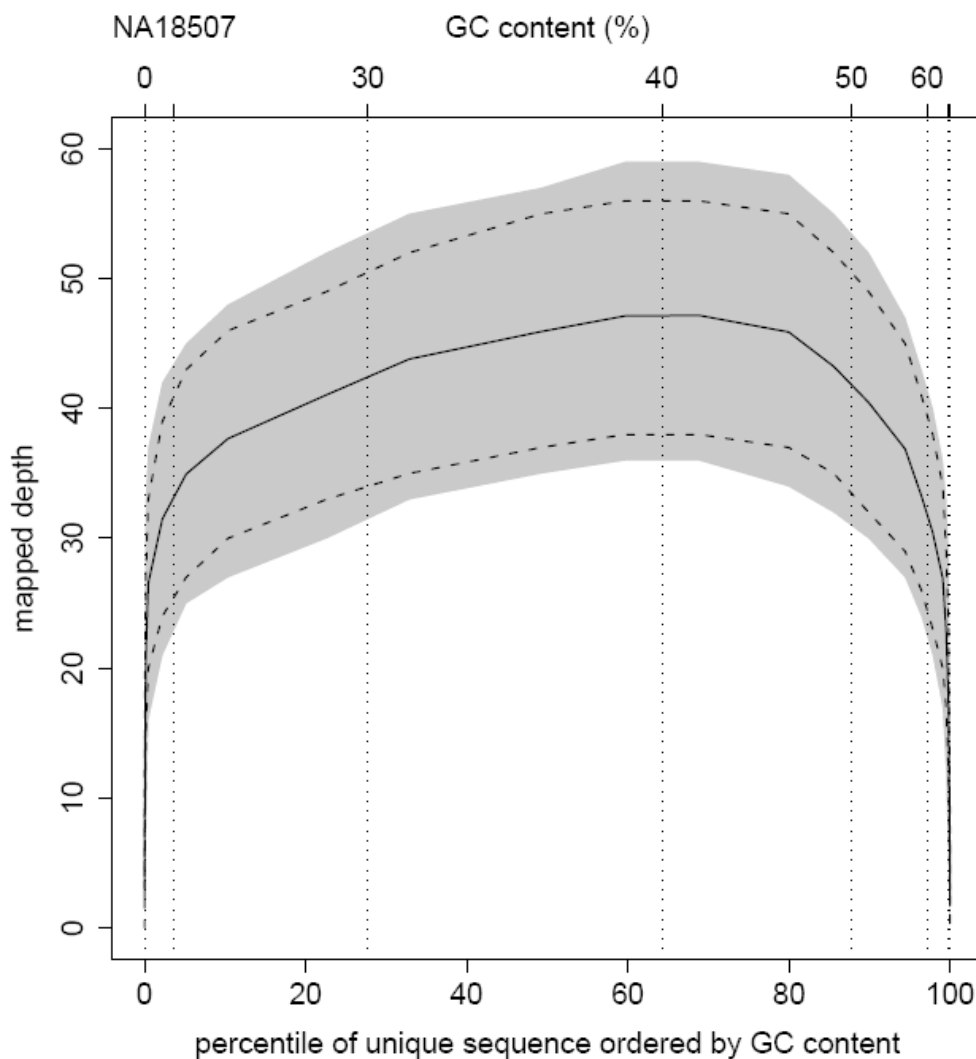


Figure S12b. Distribution of human genome NA18507 uniquely mapped reads as a function of GC content. Note that the x axis is % GC content and is scaled by percentile of unique sequence. The solid line is average mapped depth of unique sequence; the grey region is the central 80% of the data (10th to 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).

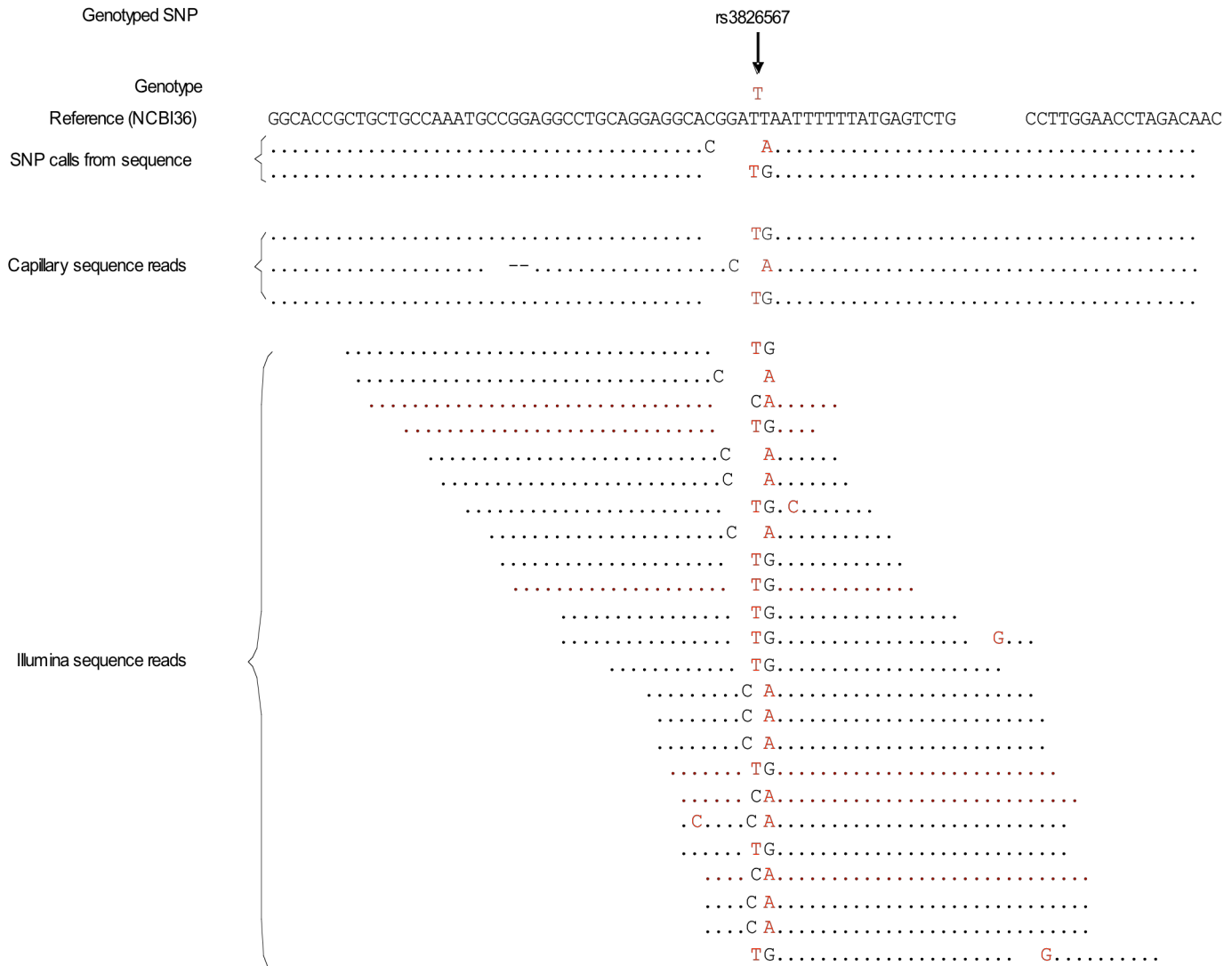


Figure S13. Sequence data reveals ‘hidden’ SNPs that cause genotyping errors. *a.* Example of a ‘hidden’ SNP detected in the sequence data. The HM550 genotyping records a ‘T’ only for NA18507 at SNP rs3826567 (arrowed), indicating that this individual is homozygous ‘TT’ at this position. Illumina sequence 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.

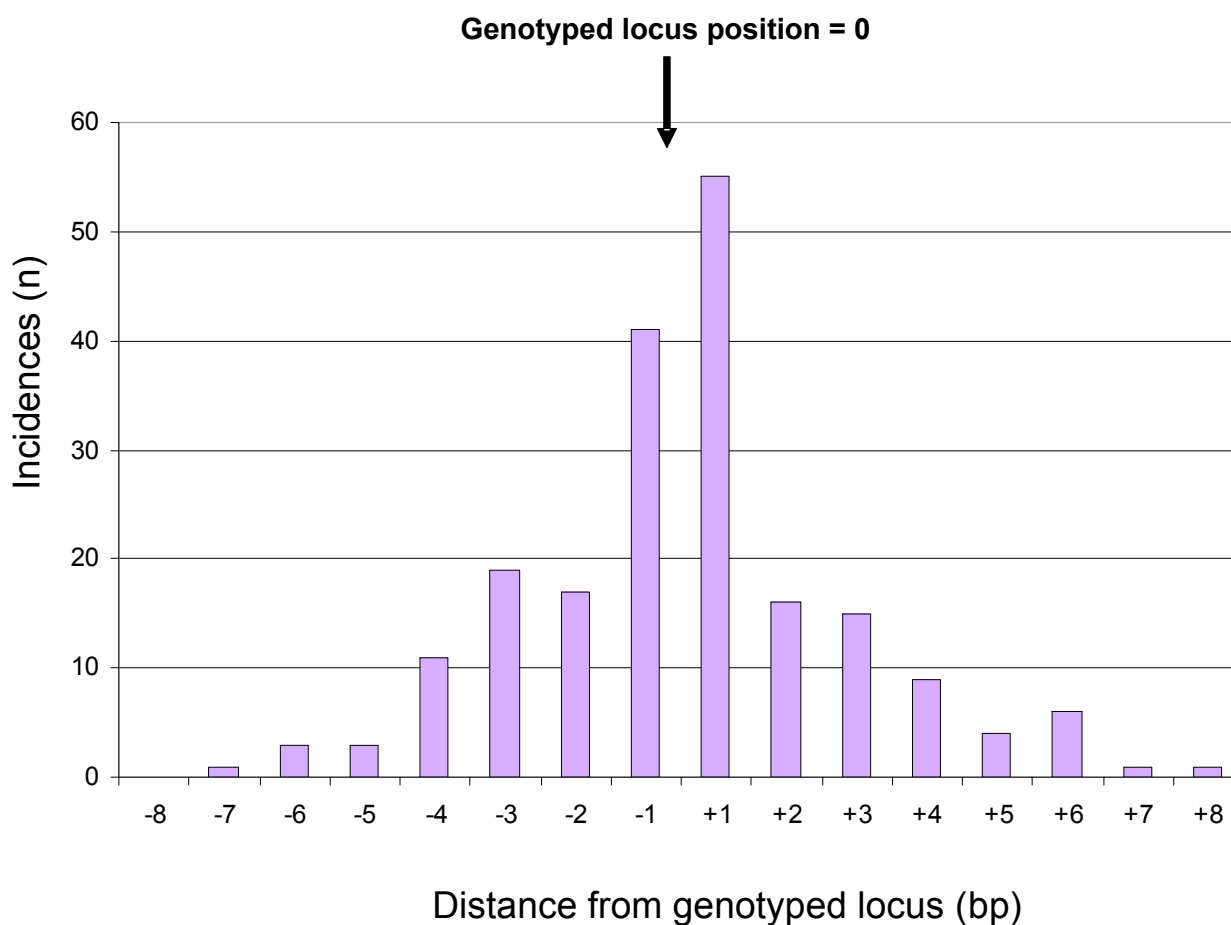
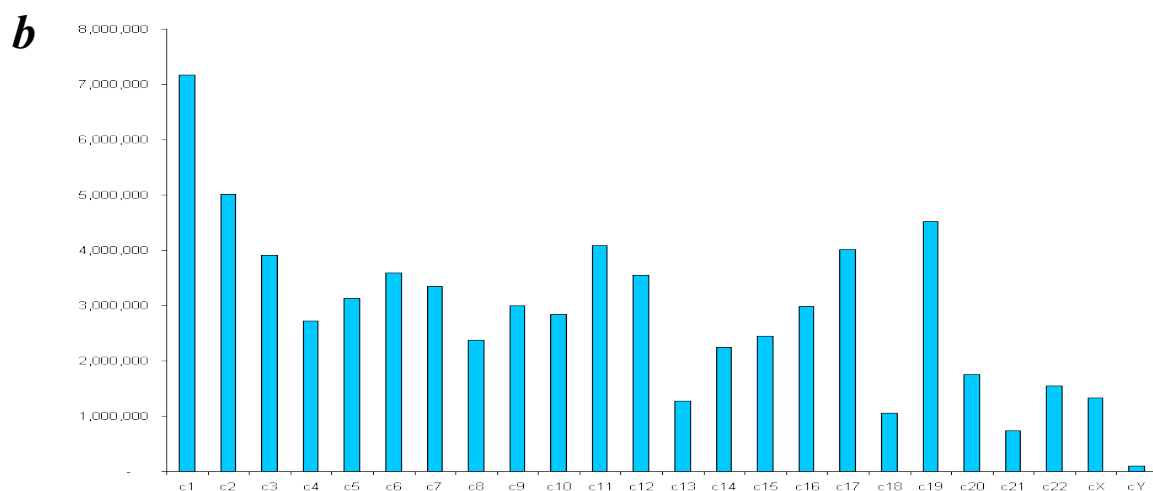
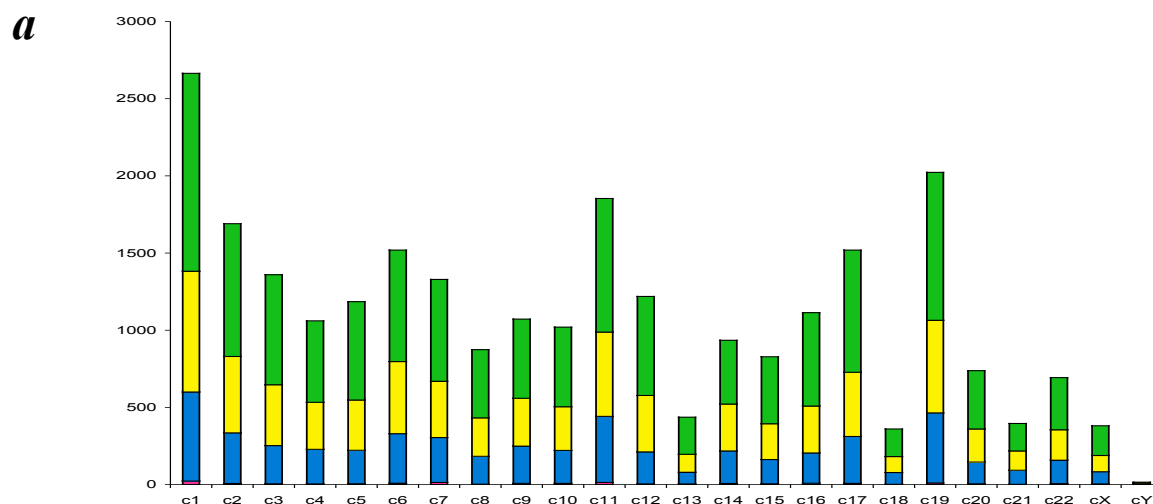


Figure S14. Distribution of ‘hidden’ SNPs in 202 discordant calls (see main text for description of ‘hidden’ SNPs). Note that ‘hidden’ SNPs will affect any genotyping platform. We therefore expect a low, but non-zero, level of genotype undercalling (Seq>GT) that is highly correlated between the platforms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



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’).

a

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

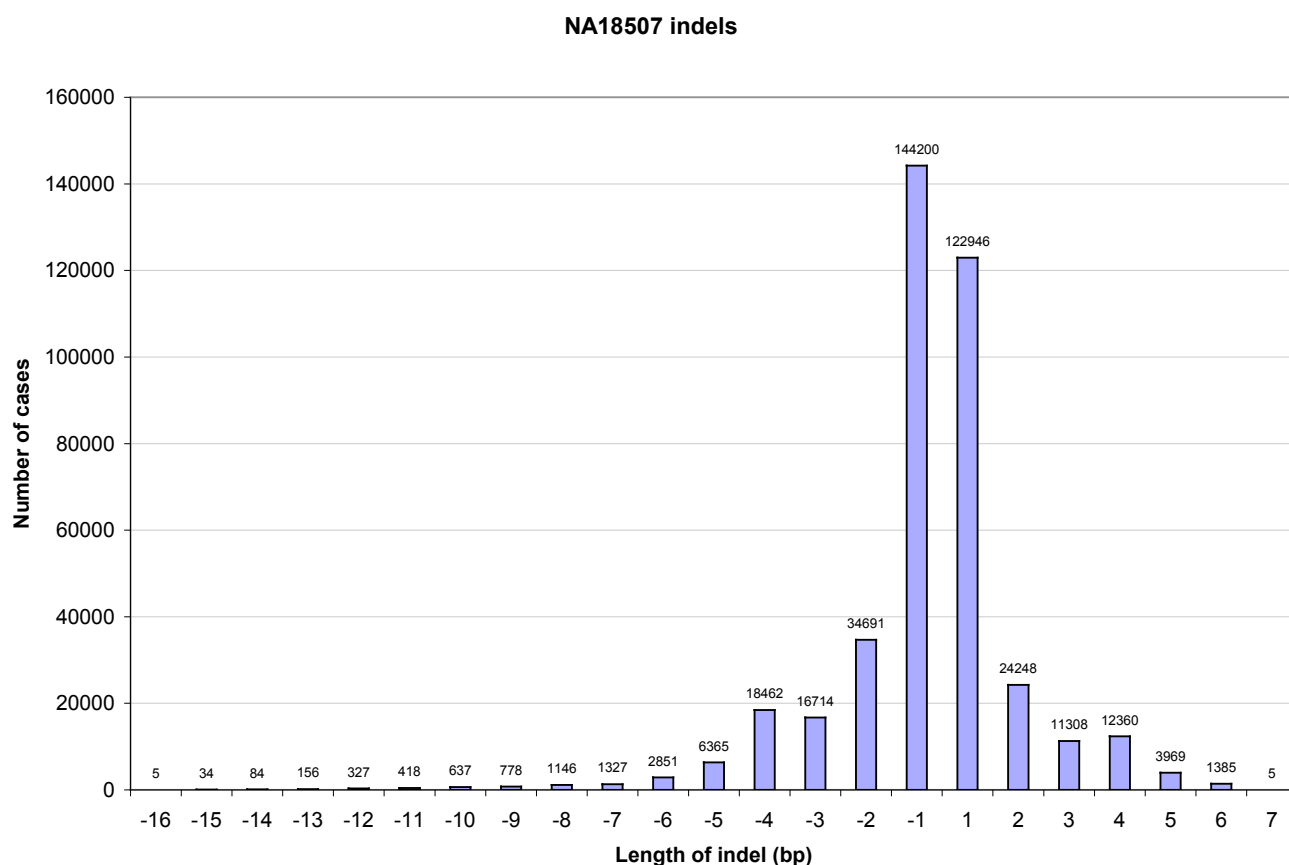
b

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.

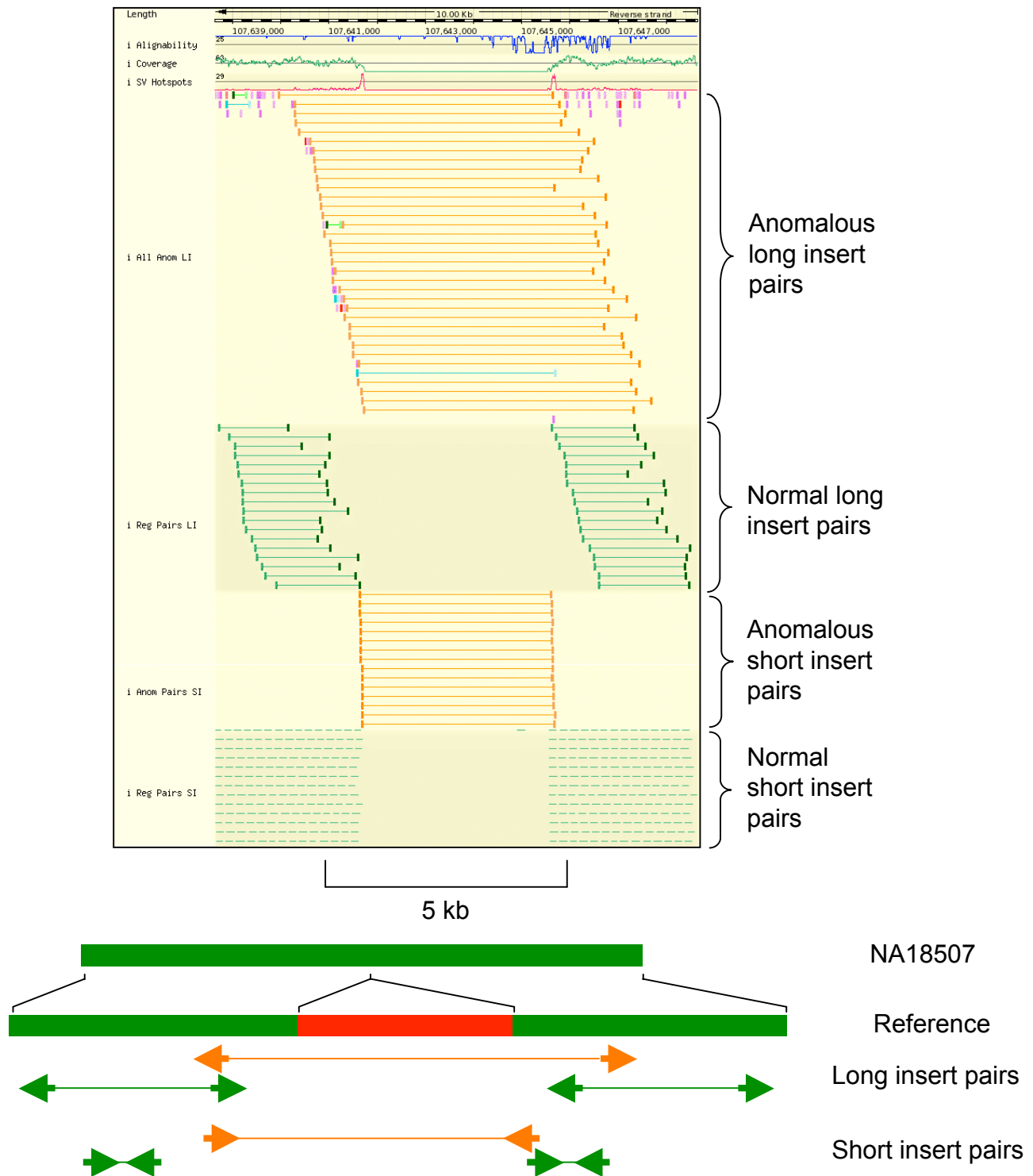


Figure S17a. Homozygous 3.6 kb deletion in NA18507 relative to the reference, supported by anomalously spaced long and short insert read pairs (orange) compared to regularly spaced read pairs (green). See also the schematic below. Note that the correct read orientation for long insert read pairs is different 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.

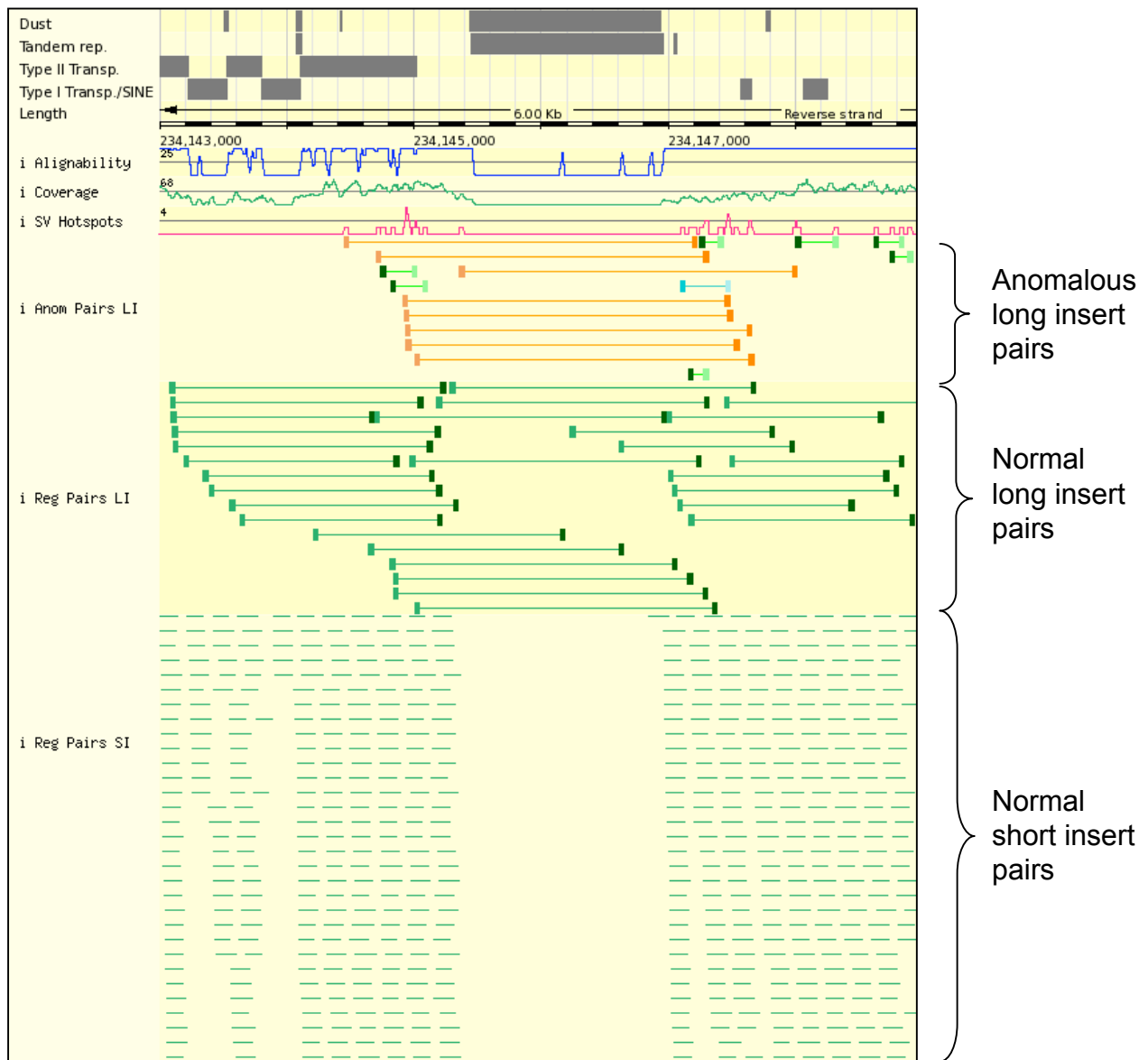


Figure S17b. Deletion in NA18507 relative to the reference, seen with anomalous long insert pairs but not short insert data. Tandem repeats (second row of display) cover ~1500 bp on the right hand side of the deletion, preventing unique alignment of reads. Short insert pairs are too short to span the 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.

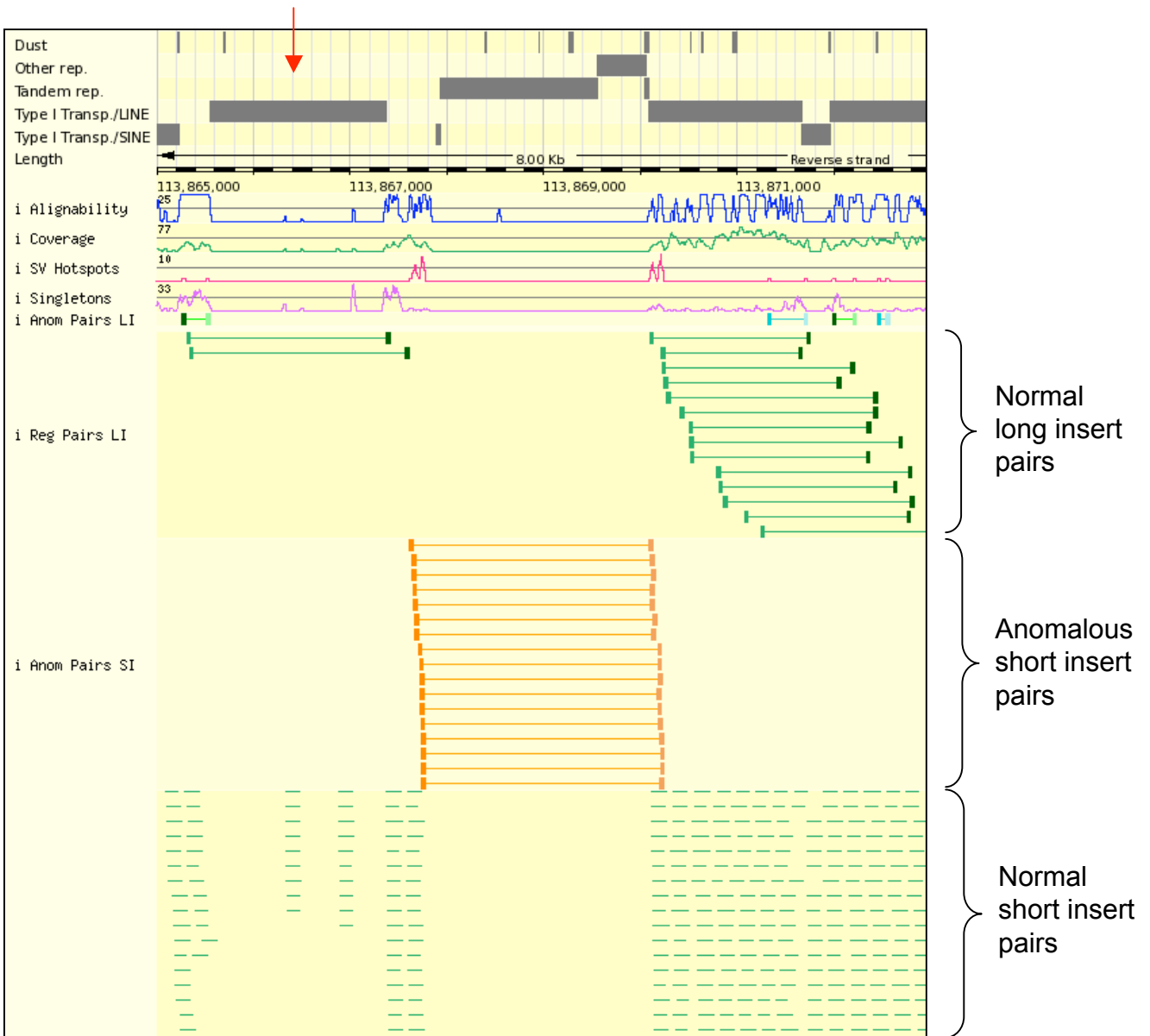


Figure S17c. Deletion in NA18507 relative to the reference, seen with anomalous short insert pairs but not long insert data. The LINE repeat on the left (see red arrow) covers 1.8 kb on the left-hand side of the deletion, preventing unique alignment of long insert reads. The higher density of short insert 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.

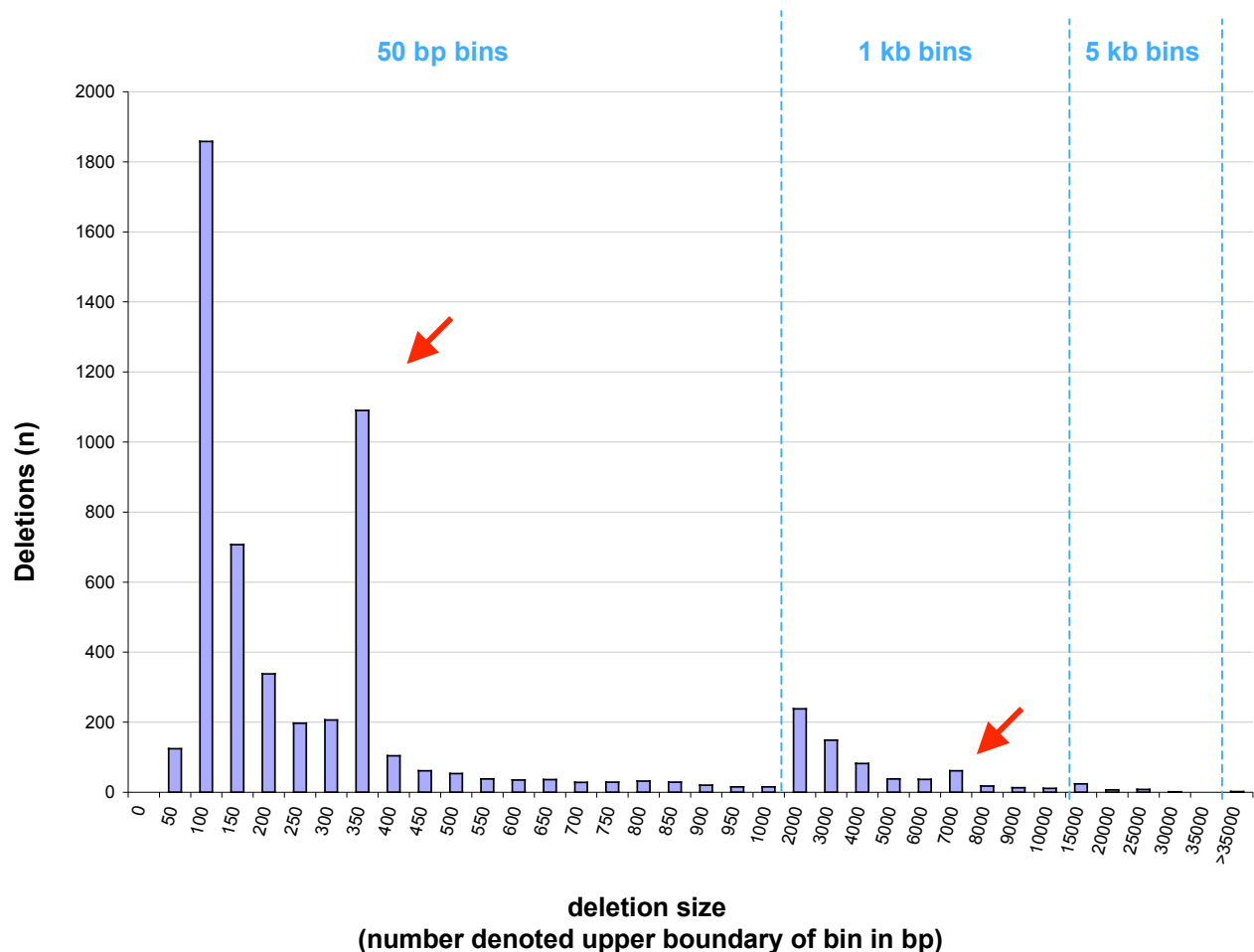


Figure S18. Size distribution of deletions in NA18507 relative to the reference. Two peaks in the distribution (arrowed) contain a substantial fraction of sequences corresponding to recently active retroposon insertions. The AluY family of SINEs are found in 90% (987/1092) of the 300-350bp fraction, 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.

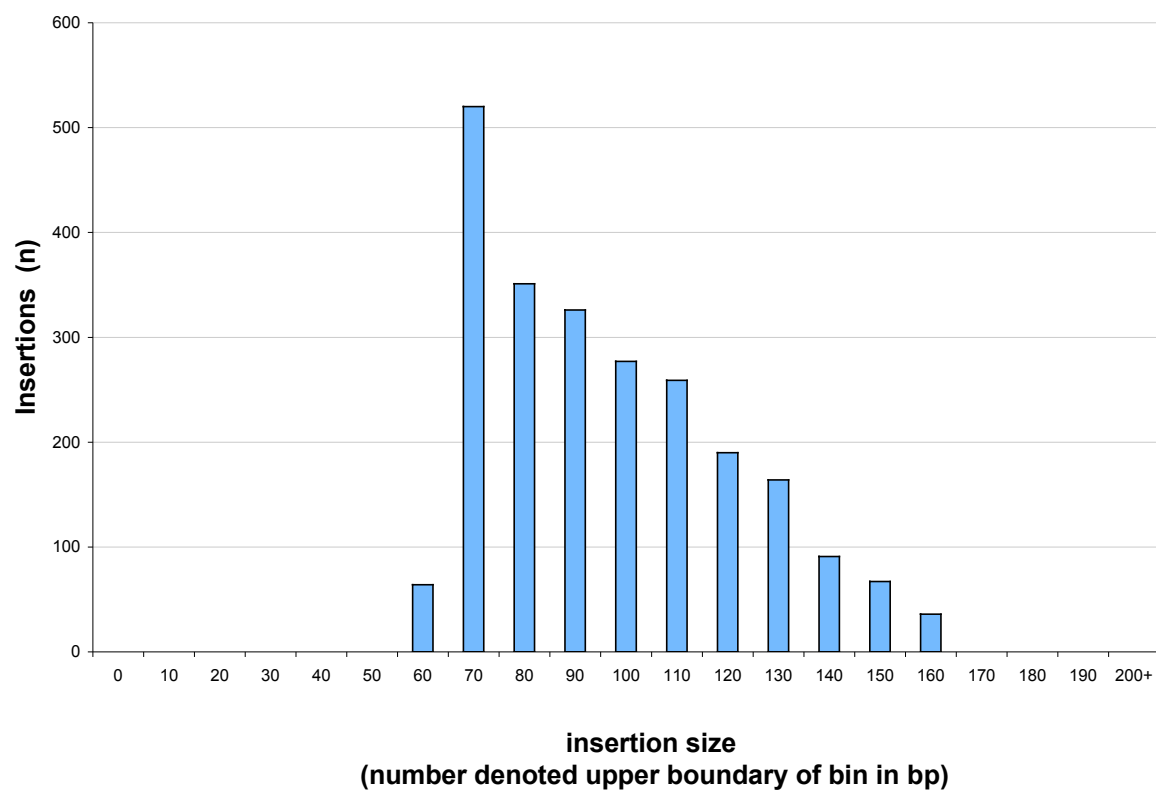


Figure S19. Size distribution of insertions in NA18507 relative to the reference, detected in the short insert dataset. Size distribution is limited by the resolution of analysis of the short insert fragment size range.

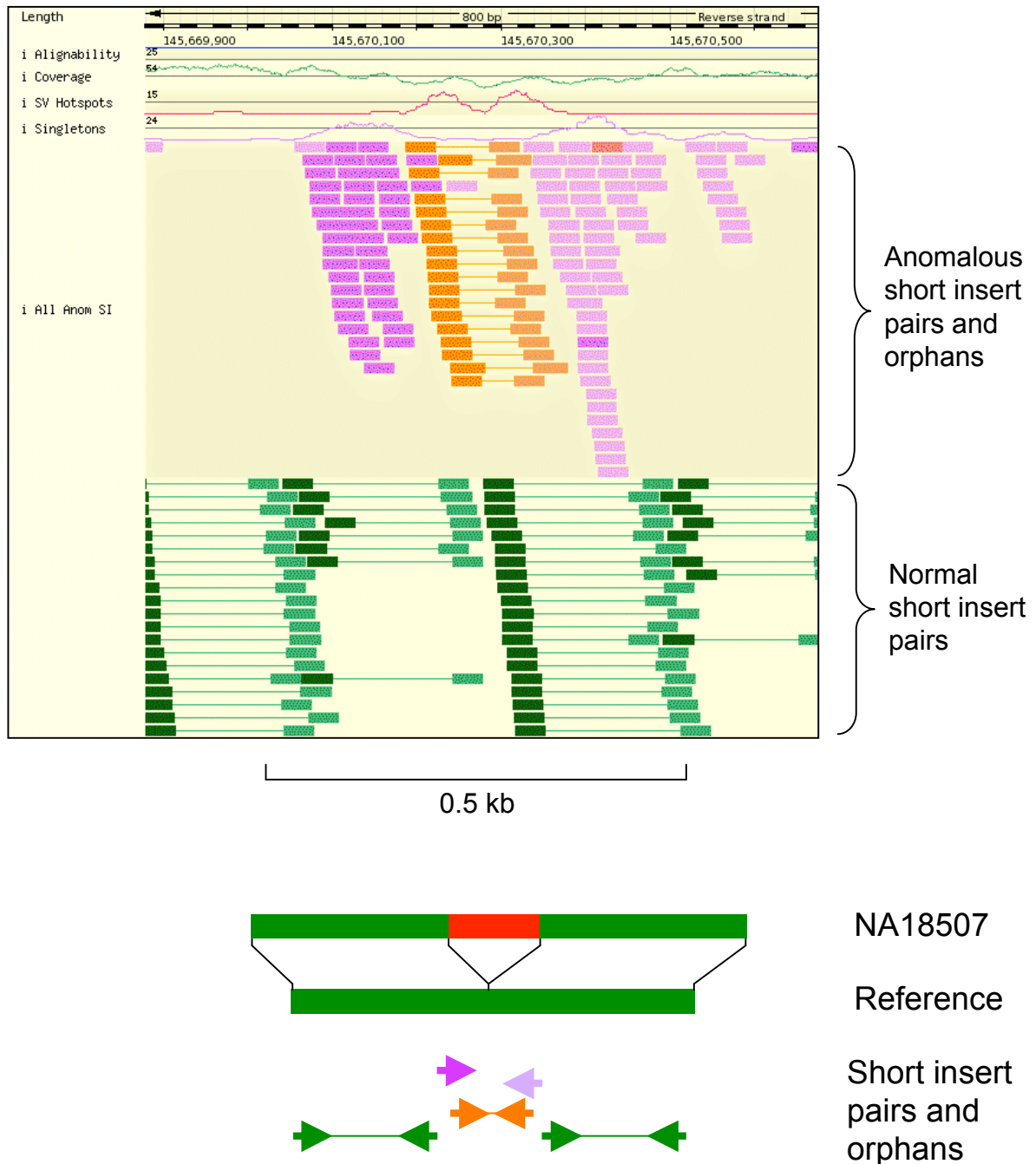


Figure S20. Homozygous 81 bp insertion detected by short insert data only. Orange bars show the anomalously short read pair spacing. Note the accumulation of singleton reads on either side, where each read has an unaligned pair. Dark and light purple denote rightwards and leftwards orientation (see schematic). 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.

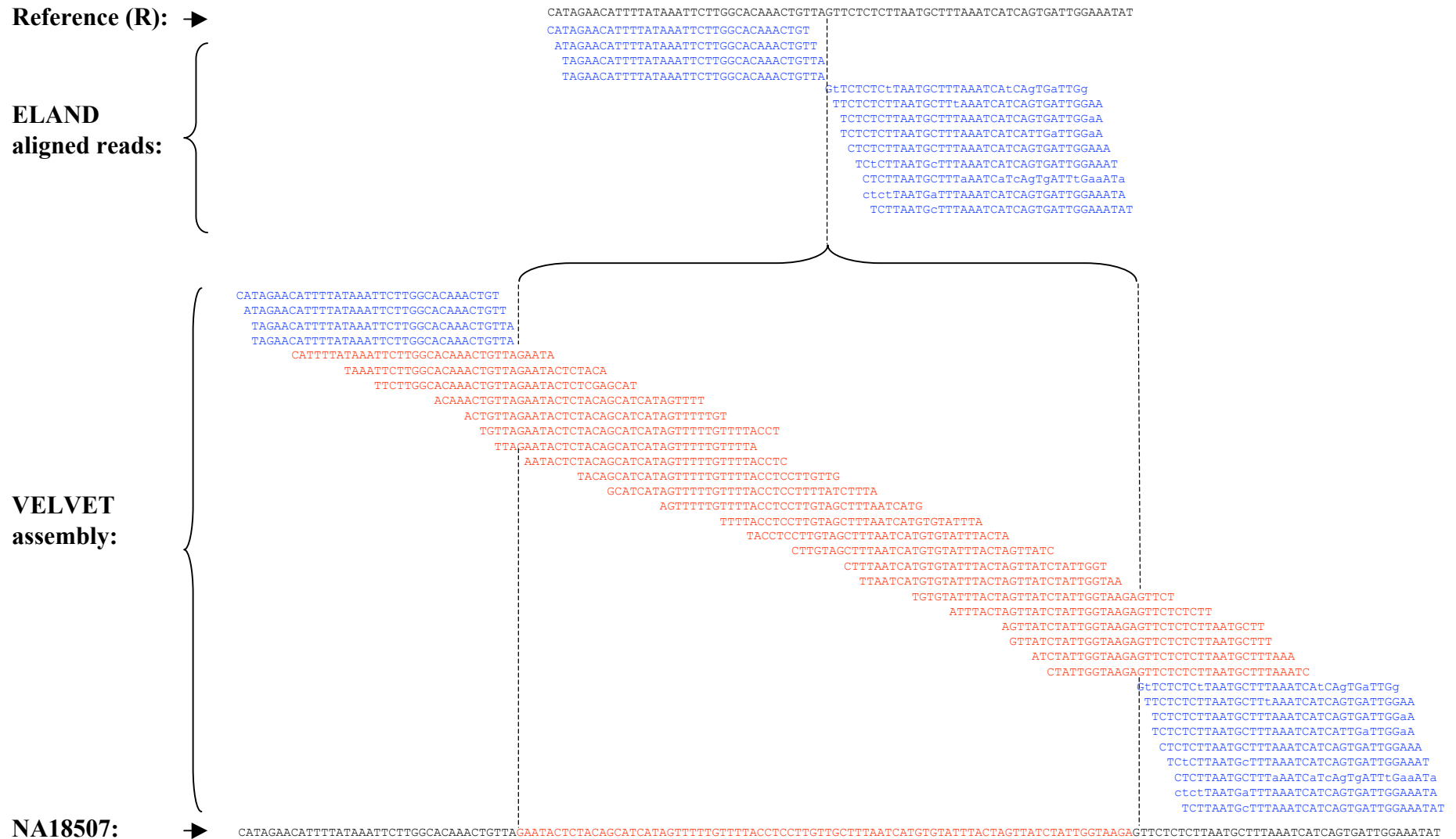


Figure S21. *De novo* assembly defines an inserted sequence in NA18507 absent from the reference. Bases present in the reference are in black, bases in NA18507 reads that align to the reference using ELAND are in blue, and bases in additional NA18507 reads that are assembled by VELVET are in red. The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

		Genotyped n	ELAND		MAQ	
			Loci n	%	Loci 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.

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.

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.

start ¹	length ¹	Reference copy number in region at read-length (32 bp) resolution	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

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.

a

HM550 COMPARISON		Genotyped n	ELAND		MAQ	
			Loci n	%	Loci 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 n	ELAND		MAQ	
			Loci n	%	Loci 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.

		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).

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	MOBKLC2	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)

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	W145STOP	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)

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	Q6ZNV3	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.