Supplementary Material: Wheeler, Srinivasan, Egholm, Shen, et al.

CONTENTS

UNABRIDGED METHODS

Library Preparation and Large Volume Emulsion PCR

Five μg of genomic DNA extracted from blood using the Flexigene DNA kit (Qiagen) and a DNA library was prepared according to previously published instructions[3] except the agarose gel fractionation step to size select fragments was replaced by a Solid Phase Reversible Immobilization (SPRI) step to remove DNA fragments[21] less than 300 bp. Briefly, the genomic DNA was fragmented using a nebulizer and the DNA was recovered in exactly 50 μl. SPRI beads (35 μl) were added to the DNA and the beads were washed to recover the fragmented DNA that predominantly contains fragments that are greater than 300 bases. When the volume guidance is strictly adhered to, fragments less than 300 bases would be less than 10% of the total fragments and the mean size of fragments is 450-550 base pairs as measured by the Agilent BioAnalyzer DNA 7500 LabChip.

Preparation of DNA Capture Beads

Size selected (25-30 μm) N-hydroxysuccinimide ester (NHS)-activated Sepharose HP beads were purchased from Amersham Biosciences. 1 mM amine-labelled HEG capture primer (5'-Amine-3 sequential 18-atom hexaethyleneglycol spacers CCATCTGTTGCGTGCGTGTC-3') (IDT Technologies, Coralville, IA, USA) in 20 mM phosphate buffer, pH 8.0, were bound to the beads and the beads were collected in bead storage buffer (50 mM Tris, 0.02% Tween, 0.02% sodium azide, pH 8), quantified with a Multisizer 3 Coulter Counter (Beckman Coulter, Fullerton, CA, USA) and stored at 4ºC until needed.

PCR Reaction Mix Preparation and Formulation

To reduce the possibility of contamination, the PCR reaction mix was prepared in a UV-treated laminar flow hood located in a PCR clean room. For each LVE emulsion PCR

reaction, 6 mls of reaction mix (1X Platinum HiFi Buffer (Invitrogen), 1mM dNTPs (Pierce), 2.5 mM MgSO4 (Invitrogen), 0.1% Acetylated, molecular biology grade BSA (Sigma, St. Louis, MO), 0.01% Tween-80 (Acros Organics, Morris Plains, NJ), 0.003 U/μL thermostable pyrophosphatase (NEB), 0.625 μM forward (5' – Biotin-CGTTTCCCCTGTGTGCCTTG-3') and 0.039 μM reverse primers (5'-CCATCTGTTGCG TGCGTGTC-3') (IDT Technologies) and 0.15 U/μL Platinum Hi-Fi Taq Polymerase (Invitrogen)) were prepared in a 1.5 mL tube. Additionally, 7 mls of mock amplification mix (1X Platinum HiFi Buffer (Invitrogen), 2.5 mM MgSO4 (Invitrogen), 0.1% BSA, 0.01% Tween were prepared in a 15 ml tube, and similarly stored at room temperature until needed.

Emulsification

The magnitude of the sequencing effort demanded that we develop a modified emulsion PCR procedure[3] that can process emulsions in large volumes. While the concentration of the various reagents remained unchanged, there were significant procedural changes to the amplification procedure. Capture beads that contain one of the PCR primers covalently bound (B primer)[3] were washed three times in a 1.7 ml microfuge tube using 1 ml of Annealing Buffer (20 mM Tris, pH 7.5 and 5 mM magnesium acetate), centrifuged in a minifuge for 10 seconds and the supernatant carefully removed. The beads were washed with 1 ml annealing buffer, vortexed for 5 seconds to resuspend the beads, and pelleted as above. All but approximately 25 μL of the supernatant above the beads were removed and an additional 1 ml of Annealing Buffer was added. The beads were vortexed again for 5 seconds, allowed to sit for 1 minute, then pelleted as above. All but 25 μL of supernatant were discarded, and library was added to the beads at 0.6 molecule/bead. The tube was vortexed for 5 seconds to mix the contents, and the contents of the tube were dispensed at 200 μl/well of 8-well strip-cap tube and the templates were annealed to the beads in a controlled denaturation/annealing program preformed in an MJ

thermocycler (5 minutes at 80° C, followed by a decrease by 0.1° C /sec to 70° C, 1 minute at 70° C, decrease by 0.1° C /sec to 60° C, hold at 60° C for 1 minute, decrease by 0.1° C /sec to 50° C, hold at 50° C for 1 minute, decrease by 0.1° C /sec to 20° C, hold at 20° C). Upon completion of the annealing process the beads were stored on ice until needed.

The emulsion making process involves two steps – preparation of microemulsions and macroemulsions. The components that make up the emulsion oil were unchanged from what was reported previously[3]. The emulsion oil was stored as 16 ml aliquots in 60 ml specimen cups (Fisher Scientific). To prepare large volume emulsions, the emulsion oil was vigorously mixed in the 60 ml specimen cup to remove any phase separation. Mock amplification mix (7 mls) was prepared as described previously[3] and was added to the emulsion oil in the specimen cup and the cup was inserted into the clamps of a TissueLyser (Qiagen) using adapters that were specially made in-house. The specimen cup was shaken at 28 Hz for 5 minutes to make micro-emulsions.

Preparation of macroemulsions (PCR droplets) was initiated with 6 mls of live amplification mix[3] in a 15 ml Falcon tube and the annealed template beads were transferred to the Falcon tube. Any remaining annealed template beads were retrieved by the addition of 100 µl of live amplification mix into each of the PCR tubes and were added to the 15 ml Falcon tube. The contents of the Falcon tube were then mixed with the microemulsions in the 60 ml specimen cup, which was then fitted into the TissueLyzer and shaken at 15 Hz for 2 minutes. This process will create an emulsion with aqueous phase micelles of the appropriate size to contain single beads with amplification mix. The emulsions were aliquoted into 96-well semi-skirted plates (Eppendorf) at 200 µl/well and emulsion PCR was initiated[3].

Emulsion Breaking

The 200 µl/ well reaction volume necessitates the retrieval of beads without the addition of isopropyl alcohol directly into filters as described previously[3]. After the beads were retrieved in filters, 100 µl isopropyl alcohol (3 times) washes were performed to collect the remaining beads in the filter[3]. The filter assembly was carefully dismantled and gently dropped into a 50 ml Falcon tube containing 35 mls of enhancing fluid (1 M NaCl, 10 mM Tris-HCl, 1 mM EDTA, pH 7.5). The beads were centrifuged and the supernatant carefully removed. The beads were divided equally into two 1.7 ml Eppendorf tubes containing ~750 µl total volume of enhancing buffer-bead slurry/tube

Template Bead Recovery

At this point the recovered beads contain a population of beads that contain amplified DNA as well as beads with no amplified DNA. To recover only beads that contain amplified DNA, a stock solution of SeraMag-30 magnetic streptavidin beads (Seradyn, Indianapolis, IN, USA) was resuspended by gentle swirling, and 320 µl of SeraMag beads were added to each of two 1.7 ml microcentrifuge tubes containing 1 mL of Enhancing Fluid. The SeraMag bead mix was vortexed for 5 seconds, and the tube placed in a Dynal MPC-S magnet, pelletting the paramagnetic beads against the side of the microcentrifuge tube. The supernatant was carefully removed and discarded without disturbing the SeraMag beads, the tube removed from the magnet, and 320 µl of enhancing fluid were added to each of the tubes. The tubes were vortexed for 3 seconds to resuspend the beads, and the tubes stored on ice until needed.

The beads recovered from emulsion PCR were mixed with the washed SeraMag beads and incubated for 5 min at room temperature on a LabQuake tube roller. Since the emPCR uses biotin containing forward primer, all amplified template beads will contain biotin, which can be selectively enriched using the SeraMag beads. After the incubation, the tubes are placed in a Dynal MPC-S magnet and the SeraMag beads were allowed to

completely adhere to the sides of the tube before the supernatant that contains the null beads was removed. The tubes were removed from the magnet and the bead pellet was resuspended with enhancing fluid. This wash procedure is repeated as described above until the supernatant is clear of null beads.

After the last wash, the beads were left in the Dynal MPC-S magnet and 800 µl of freshly prepared Melting Solution (0.125 M NaOH, 0.2 M NaCl) was added to each tube and the pellet resuspended by vortexing for 2 seconds. The tubes were then placed back in the Dynal MPC-S magnet and the SeraMag beads were allowed to completely adhere to the sides of the tube before the supernatant that contains the template beads with ssDNA was removed and placed in another 1.7 ml tube. This step was repeated one more time to collect the template beads. The tubes containing the single stranded template beads were centrifuged and the melt solution was removed. The beads were washed three times in 1 ml Annealing Buffer. The beads from the two 1.7 ml tubes were pooled in a 0.2 ml PCR tube.

Sequencing Primer Annealing

Approximately 1.4-1.8 million library beads were prepared for 454 Sequencing by annealing sequencing primer to the library beads as follows. The enriched beads were centrifuged at 2,000 RPM for 3 minutes and the supernatant decanted, after which 96 µL of 100 µM sequencing primer (5'-CCATCTGTTCCCTCCCTGTC -3', IDT Technologies), were added. The tube was then vortexed for 5 seconds, and transferred to 200 µl PCR tubes and placed in an MJ thermocycler for the following 4 stage annealing program: 5 minutes @ 65^{o}C, decrease by 0.1^{o}C /sec to 50^{o}C, 1 minute @ 50^{o}C, decrease by 0.1^{o}C /sec to 40^{o}C, hold at 40^{o}C for 1 minute, decrease by 0.1^{o}C /sec to 15^{o}C, hold at 15^{o}C.

Upon completion of the annealing program, the beads were removed from thermocycler and pelleted by centrifugation for 10 seconds, rotating the tube 180˚, and spun for an additional 10 seconds. The supernatant was discarded, and 200 µL of annealing buffer were added and beads pooled into a 1.7 ml microfuge tube. The beads were resuspended with a 5 second vortex, and the beads pelleted as before. The supernatant was removed, and the beads resuspended in 500 µL annealing buffer, at which point the beads were quantified with a Multisizer 3 Coulter Counter. Beads were stored at 4^{o}C and were stable for at least one week.

Incubation of DNA Beads

Bead buffer 2 (200 ml) was prepared by the addition of 34 µl of apyrase solution (Biotage, Uppsala Sweden) to pre-chilled 200 ml of "Buffer CB for Bead Buffer 1" (final apyrase activity 8.5 units/litre). The PicoTiterPlate was incubated in bead buffer 2 at room temperature for at least 10 minutes. Approximately 1.4 million to 1.8 million of the previously prepared DNA beads were centrifuged and the supernatant was carefully removed. The beads were then incubated in 1740 µl of bead buffer 2 containing 25 mM tricine, pH 7.8, 0.4 mg/mL polyvinyl pyrrolidone (MW 360,000), 8.8 mM magnesium acetate, 0.1% tween 20, 323 µg of E. coli single strand binding protein (SSB) (United States Biochemicals Cleveland, OH) and 13000 units of Bst DNA polymerase, Large Fragment (New England Biolabs). The beads were incubated at room temperature on a rotator for 30 minutes.

Preparation of Enzyme Beads and Micro-Particle Fillers

UltraGlow Luciferase (Promega, Madison WI) and Bst ATP sulfurylase were prepared in house as biotin carboxyl carrier protein (BCCP) fusions. The 87-aminoacid BCCP region contains a lysine residue to which a biotin is covalently linked during the in vivo expression of the fusion proteins in E. coli. The biotinylated luciferase (1.2 mg) and

sulfurylase (0.4 mg) were premixed and bound at 4ºC to 2.0 mL of Dynal M280 paramagnetic beads (10 mg/mL, Dynal SA) according to the manufacturer's instructions. The enzyme bound beads were washed 3 times in 2000 µL of bead buffer 2 and resuspended in 2000 µL of bead wash buffer.

Seradyn microparticles (Powerbind SA, 0.8 µm, 10 mg/mL, Seradyn Inc, Indianapolis, IN) were prepared as follows: 720 µL of the stock were washed with 1000 µL of bead buffer 2. The microparticles were centrifuged at 9300 g for 10 minutes and the supernatant removed. The wash was repeated 2 more times and the microparticles were then incubated in 720 µl of bead buffer 2 containing 25 mM tricine, pH 7.8, 0.4 mg/mL polyvinyl pyrrolidone (MW 360,000), 8.8 mM magnesium acetate, 0.1% tween 20, 30 µg of *E. coli* single strand binding protein (SSB) (United States Biochemicals Cleveland, OH) and 1200 units of *Bst* DNA polymerase, Large Fragment (New England Biolabs). The microparticles were incubated at room temperature on a rotator for 30 minutes.

Bead Deposition

The three different layers of beads are deposited onto the PicoTiterPlate in a stepwise fashion. The 1st layer is the library beads, followed by the Seradyn microparticles (2nd layer) followed by the enzyme beads (3rd layer). Approximately 1.4-1.8 million library beads were deposited by allowing the beads to settle in the PicoTiterPlate wells using gravity for 10 minutes. Following this step, the packing beads were added and the PicoTiterPlate was centrifuged at 1430*g* for 10 minutes. After the deposition of the packing beads, the enzyme-beads were deposited and the PicoTiterPlate was centrifuged at 1430g for 10 minutes. The prepared PicoTiterPlate is then loaded onto the GS-FLX instrument to initiate a sequencing run.

Wetting the Pico Titer Plate

Remove the PicoTiterPlate (PTP) from the shipping tray where it has been soaking and place it onto the Bead Deposition Device (BDD) base. Carefully place a gasket that creates two 30x60 mm^2 active areas over the surface of a 60x60 mm^2 PTP. Secure the BDD top onto the assembled BDD base/PTP/Gasket. Pre-wet each loading regions of the PTP with 1860 µl of bead buffer 2 and centrifuge the assembled BDD for 5 minutes at 1430 x g RCF. After centrifugation the supernatant was removed with a pipette

Layer 1: When the 30 minutes of DNA bead incubation is complete, add 1920 µl of bead buffer 2 to the 1800 µl of DNA beads. Vortex the DNA beads and carefully draw 1860 µl of first layer solution and pipette onto region 1 of the PTP. Repeat the same for region 2 of the PTP. Leave the PTP on the bench for at least 10 minutes to allow the DNA beads to settle into the wells by gravity.

Layer 2: After the 10 minute deposition of the DNA beads was complete, the supernatant was carefully removed from each PTP region and transferred into two 2.0 ml microfuge tubes. The supernatant was centrifuged for 10 seconds at 9300 x g RCF to pellet any residual DNA beads. Without disturbing the pellet, 2920 µl of supernatant was carefully removed and mixed with 880 µl the Seradyn microparticles. The microparticles were vortexed and 1860 µl of diluted microparticles were loaded onto each region of the PTP. The PTP was then centrifuged for 10 minutes at 1430 x g RCF

Layer 3: After centrifugation of the second layer was completed, the supernatant was discarded. Dynal enzyme beads (1840 µL) were mixed with 1960 µL of bead buffer 2 and 3400 µL of enzyme-bead suspension was loaded on the PTP. The PTP was centrifuged for 10 minutes at 1430 x g RCF and the supernatant decanted. The PTP was removed from the BDD and stored in bead buffer 2 until ready to be loaded on the instrument.

Sequencing on the 454 Instrument

All sequencing was performed using the LR70 sequencing kits. All flow reagents were prepared as follows. GS-FLX common buffer contains 25 mM Tricine, pH 7.8, 0.4 mg/mL polyvinyl pyrrolidone (MW 360,000), 1 mM DTT and 0.1% Tween 20. Substrate (18 mM D-luciferin (Regis, Morton Grove, IL), 150 μM adenosine phophosulfate (Sigma) and Inhibitor (100 μM 2-(4-carboxymethylaminophenyl)-6-methylbenzothiazole (Promega)) was prepared in GS-FLX common buffer. Apyrase wash is prepared by the addition of apyrase to a final activity of 10 units per litre in GS-FLX common buffer. Deoxynucleotides dCTP, dGTP and dTTP (Pierce) were prepared to a final concentration of 130 μM in 25 mM Tricine, pH 7.8 and 0.01% Tween 20, α-thio deoxyadenosine triphosphate (dATPαS, Biolog, Hayward, CA) and sodium pyrophosphate (Sigma) were prepared to a final concentration of 1 mM and 2 μM, respectively, in the 25 mM Tricine and 0.01% Tween 20.

The 454 sequencing instrument consists of three major assemblies: a fluidics subsystem, a PTP cartridge/flow chamber, and an imaging subsystem. Reagents inlet lines, a multi-valve manifold, and a peristaltic pump form part of the fluidics subsystem. The individual reagents are connected to the appropriate reagent inlet lines, which allows for reagent delivery into the flow chamber, at a pre-programmed flow rate and duration. The flow chamber also included means for temperature control of the reagents and PTP, as well as a light-tight housing. The polished (unetched) side of the slide was placed directly in contact with the imaging system.

The Genome Sequencing FLX System has sequencing scripts for performing 100 cycles of each deoxynucleotide (dNTP) flow. The scripts specify the reagent name (dATPαS, dCTP, dGTP, dTTP, and PPi standard), flow rate and duration of each script step. The flow order for the dNTPs is T-A-C-G. Sequencing on the GS-FLX is accomplished by the use of concentrated reagents (substrate, inhibitor, apyrase and dNTPs) in the system where GS-FLX common buffer is flowed constantly at 4.2 mls/min

and the concentrated dNTPs, substrate, inhibitor or apyrase are mixed at the beginning of the common buffer flow stream by computer-controlled valve actuations that each last for 200 milliseconds. Each dNTP flow order is organized into kernels – programmed reagent flow order. Each kernel is composed of ordered flows of apyrase wash, substrate wash, dNTPs followed by another substrate wash. The apyrase wash, substrate wash, dNTP flow and substrate wash are 21, 10, 12 and 6 seconds, respectively. Each concentrated reagent actuation is punctuated by an air plug to remove mixing of the various reagent fronts. This air plug is removed as the reagents flow through a debubbler - a small volume vertical chamber that takes advantage of the difference in density of air and the rest of the fluids and aspirates off air from the top at a rate of 0.7mls/min. Aspiration of the air plug removes the air gap, rejoins the sections of fluids, and prevents any air bubbles to interfere with the sequencing reactions on the PTP that is maintained at a constant temperature of 35°C. When the respective reagents reach the PTP, the individual reagents would have mixed with the common buffer stream and attained their final concentration. A single 38-second image is captured per kernel step.

During each dNTP flow, the *Bst* DNA polymerase bound to the DNA strands on the library beads will incorporate complementary dNTPs thereby releasing inorganic pyrophosphate (PPi) as a by-product. This PPi is captured by the signal-transduction enzyme system (*Bst* ATP Sulfurylase and UltraGlow luciferase) to produce light. Signal generated during each dNTP flow is captured in real-time using a CCD camera. The wash steps are required for removing the residual unincorporated nucleotide, quenching the signal and re-generating the signal-transduction enzymes for the next cycle of dNTP incorporation[3].

Image and Signal Processing

For each dNTP addition, an image was recorded and processed numerically “on-the-fly” while the sequencer was recording the image for the next dNTP addition. This image-processing step generated a set of raw data composed of signal intensities of DNA-containing beads, whose numerical computation completed almost immediately following the completion of the sequencing experiment. The set of raw data was then transferred to a Linux server for signal processing, while the sequencer was performing next sequencing runs or undergoing routine maintenance.

Signal processing contained multiple steps, including (i) de-convolution of signal cross-talk among DNA-containing beads, (ii) correction for phase errors due to a-synchronicity of dNTP additions and the primer extensions, (iii) read-quality estimation and filtering, and (iv) base-calling to sequence read-out for each DNA-containing bead. These processes were either done in a single Linux server, or were performed through a parallelization scheme across several Linux servers linked by a gigabit network distributed internally in 454 Life Sciences. Both image and signal processing were performed using commercial software designed by 454 Life Sciences.

The sequencing data were first filtered using a stringent setting in the quality filter, which allowed less than 4 "ambiguous" calls over the first 320 dNTP flow-additions for each read. These ambiguous calls had their signal intensities fall close to the "valley" positions of the signal intensity histogram of the sequencing run data, and were the flow signals for which the base-caller had less confidence in calling the bases reliably. The reads that passed this stringent filter were categorized as "Tier 1" reads.

For each read that failed the stringent filter, its flow signals were scanned from 5' to 3' end to determine a trim position, such that from the 5' to this trim position, the read spanned the longest-possible flow window within which its "valley-per-flow" ratio still satisfied "4 / 320", or equivalently, the 0.0125 valley-per-flow condition used in the

previous stringent filter. If the trim position was not set at less than 84 flows, the read was categorized as "Tier 2" read, and the sequence of bases was determined (base-called) up to this trim position. These tier 2 reads thus had shorter lengths than the tier 1 reads (see below) but still had the desired low read-errors guarded by the trimming mechanism.

454 Reads and Criteria for Mapping

Reads from each of 234 sequencing runs were analyzed as FASTA formatted sequence plus quality files. From each run, both Tier 1 and Tier 2 reads (see p. 12 "Image and Signal Processing") were analyzed. The length distributions of the reads shows that tier 1 reads were longer in general, with a modal value of 257, whereas the tier 2 reads exhibited a trimodal distribution with peaks at 195, 145 and 80 bases (Supplementary Fig. 9A)

Reads were mapped to the human reference genome, NCBI build 36 (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/bigZips/), using BLAT, with the -ooc option, which prevented seeding alignments with 11-mers that appeared in the reference genome more than 1024 times. Reads within one run that shared a common start position on the reference genome sequence and were within 2bp of the same length were judged to be duplicate reads produced from the rare accidental inclusion of two beads within a single emulsion droplet. Only one of the duplicates was kept for further analysis. Approximately 2.79% of the tier 1 reads and 0.139% of tier 2 reads were rejected by these criteria. Due to the structure and distribution of repeat sequences in the genome, many of the reads aligned to multiple sites (Supplementary Fig. 9B). The location of the best hit in the genome was used as the map location for the read.

Several properties of a given read's best hit match were used to judge whether the read was mapped to the correct genomic location. Approximately 7% of reads that found matches in the genome have one or more alternative hits within 1% of the match score of

the best hit (Fig. 9C). The majority of reads found genomic positions that aligned their full length. Less than 5% of reads aligned less than 90% of their length to the genome (Supplementary Fig. 9D). Analysis of the BLAT matches revealed a wide range of percent identity and frequency of indels within the alignment. Supplementary Fig. 9E and 9F show the distribution of mismatches and indels respectively in aligned reads. Based on these results, a given read was deemed ambiguously mapped if it failed to meet the following criteria: 0 alternate hits; 90% of the read aligned to the genome; less than 5 mismatches; less than 5 indels. Reads failing these criteria and were set aside from subsequent analysis of SNPs and indels.

Ninety-three million reads passed the filtering criteria and were used for subsequent analysis of SNPs and indels (see Supplementary Table 1). Coverage of reads was calculated in 5000 base consecutive windows across the entire reference genome. Figure 1, main text, shows a final coverage distribution with mean 7.4 fold redundancy overall with a shoulder at 3.6 fold representing the X chromosome. The coverage in 5 kb intervals across the chromosomes (e.g., chromosome 1, Supplementary Fig. 2) showed uniform coverage across greater than 95% of the chromosome, excluding telomeres and centromeres.

All reads satisfying the mapping criteria were realigned to the genome using cross_match with *-minscore* 24, -raw, *-masklevel* 80, and *-discrep_lists*. Alignments were achieved by breaking the reference chromosome into 41 kb segments, with 1 kb overlaps between neighbouring pieces. The cross-match alignment was important in optimizing the alignment, although it limits the length of deletion that can be detected within a given read: the longer the deletion the less likely it is to be captured on a given 41 kb segment. Each mapped read was assigned to a reference segment by the start position of its BLAT result on the genome. The variation lists produced by cross_match were used for subsequent analysis of SNPs and indels.

Errors in 454 Reads

We conducted independent evaluation of error rates in FLX reads using sequencing runs from a strain of *Staphylococcus aureus* USA300-HOU-MR (also known as TCH1516) for which a high-quality reference sequence was available (accession: CP000730) . A set of 353,192 reads (91012736 bp) were aligned to the reference using cross_match (v0.990319) with "-masklevel 80," "-raw," and "-discrep_lists" options. All base variation were assumed to be sequencing error in this comparison. Although the error rate was low, approximately 0.025% mismatch and 0.27% indels, the distribution of errors across the reads showed a gradually increasing rate, for both substitutions and insertions or deletions (indels) starting from position 100 to the end of the reads (Supplementary Fig. 13). The vast majority of indel errors were single base events associated with homopolymer runs (data not shown).

Further analysis revealed the mismatch error rate, measured from the 3' end of the read, decreased exponentially to100 base from the end of the reads. This gave rise to a linear increase in the $\log_{10}$ error probability from 0 to 100 bases *from the 3' end of the read* (Supplementary Fig. 14) from 25 to 40, which reached a plateau of approximately 45 across the remaining 150 bases to the beginning of the read. The 454 base calling software provided quality scores (Q) that strive to incorporate the substitution and indel error rates and therefore the Q values are lower than the substitution error rate alone. Q values provided by 454 failed to reflect the drop off expected from these empirical error rates (data not shown).

The Q values, taken across all reads, were bi-modally distributed with the most prominent peak at Q=27, and another minor peak at Q=32 (Supplementary Fig. 15). For Q <20 the true error rates correlated poorly; for Q>25 the actual error rates were

generally very low (data not shown). The distribution of Q scores in erroneous reads paralleled the Q scores for all reads with a peak at Q=27 (Supplementary Fig. 15).

Scoring System for Mismatch Base Positions

We designed a scoring function for variant based on the 454 Q values and the frequency of observation of a given substitutions. First, we adjusted the Q score of the variant bases to Q' as follows:

$$Q' = 27, \quad Q > 27,$$

$$Q' = Q + 10 \text{ for } Q < 15.$$

A distance penalty was added to Q' scores for bases within 100 bases of the end of the read.

$$Q^* = Q' - 0.1278\,(100 - d)$$

$$Q^* = Q', \; d > 100.$$

where d was distance of the mismatch from the end of the read. At each variant position, we summed the Q* of the identical variant bases.

$$S_V = \Sigma\, Q^*$$

Filtering Criteria for SNPs

Variant score (S_V): Mismatch base positions found among the 454 reads were scored as described above. The positions could be divided into those that coincided with SNPs in dbSNP build 126 ("known SNPs") and those that didn't ("novel SNPs"). The preponderance of known SNPs are judged to be bona fide variation, as described in the text. The scores of known SNPs were used to set a cut-off of $S_V = 28$ for a SNP position to be considered valid. The ratio of variant bases to total bases covering a given position was required to be 0.2.

Homopolymer runs: It is well known that single base insertions or deletions are generated as a artefact in homopolymer nucleotide runs in 454 sequencing. On the non-autosomal portion of the X chromosome, any variant that appeared to be a heterozygote was assumed to be an error. We investigated the association of heterozygous and homozygous variants on the X with homopolymer runs. We found in proximity to longer homopolymer runs an increase in the number of heterozygous variants, which suggested homopolymer runs not only contributed to insertion or deletion errors but also sometimes to substitution errors (Supplementary Fig. 10). Therefore we analyzed all novel SNPs, especially those occurring in tandem or spaced apart by only one base, and found a strong association (Supplementary Fig. 11) with longer homopolymer runs. Known SNPs were associated with homopolymer runs of >5 bp less than 3% of the time, whereas novel SNPs were associated 33% of the time. Therefore novel variants were removed if they were associated with a homopolymer run >5 bp within a 13 bp window centred on the SNP, or if the homopolymer run =5 bp and $S_V \leq 54$.

Coverage and the Detection of Heterozygotes

The sensitivity of detecting heterozygous SNPs by whole genome shotgun sequencing is limited by the depth-coverage. If the WGS average depth-coverage is *C*, the coverage *k* for each base of the genome approximately follows the Poisson distribution[22]:

$$f(k \mid C) = \frac{C^k \bullet e^{-C}}{k!}$$

For each heterozygous SNP site of the diploid genome, if the number of reads covering it is *K*, the number of reads *i* from one of the two alleles follows the binomial distribution:

$$f(i \mid K, 0.5) = \frac{K!}{i! \bullet (K-i)!} \bullet 0.5^K$$

Assuming the requirement of calling a heterozygous SNP is to observe at least two 454-reads from both alleles at the SNP site, the sensitivity of detecting heterozygous SNPs is computed based on the two statistical distributions described above (Supplementary Fig. 4). The value from 7X random sequencing is 75%. To achieve 99% sensitivity, >12X depth of coverage is required.

We used binomial probabilities to make genotype calls at each SNP position. Given the bases present at a variant position, we defined the null hypothesis as follows: among all the bases covering one SNP position, if one of them is the same as the reference, the SNP is heterozygous. If true, the number of variant reads follows a binomial distribution. From this we determined the probability (p) for each SNP position. If p <=0.01, we reject the null hypothesis, and declare the SNP a homozygote.

Sensitivity and Specificity of SNP Discovery

We estimated false positive and false negative rates for genotypes associated with the known SNPs by comparison to the Affymetrix genotyping array. For the genotype calls from 454 sequencing (see Supplementary Table 2, column 3 "454 Sequence Outcome") rows B, F and I, which include 8,268 genotypes, were assumed to be false positive homozygous calls. Rows E and J, including 2,311 genotypes, were assumed to be false positive heterozygous calls. The total number of homozygous variant genotypes, called by sequencing was 124,070; heterozygotes added to 105,013. The specificity for heterozygous and homozygous genotypes was calculated as the fraction of the total that were false positive. The sensitivity was the fraction of Affymetrix genotypes the sequencing correctly called. The specificity was calculated as 0.978 for heterozygotes,

0.933 for homozygotes and sensitivity was 0.758 for heterozygotes, 0.951 for homozygotes.

Transition and transversion ratios (Supplementary Fig. 3) were used to estimate the false positive rate for the novel SNPs as follows: The empirical ratio of SNP base changes representing nucleotide transitions to those representing nucleotide transversion among known SNPs was 2 to 1 (67% of the base changes are transitions). Since the known SNPs were shown to be very accurate we assumed this to be the true ratio for this genome. Novel SNPs, with 61% of base changes as transitions, deviate from this ratio due to the presence of random error wherein only 1/3 of the changes are nucleotide transitions. Given these assumptions, we estimated false positive rate of novel SNP as follows: denote x as the proportion of true novel SNPs, y as the proportion of false positive novel SNPs, then $x + y = 1$, and further $2x/3 + y/3 = 0.61$. Therefore $x = 0.83$ and $y = 0.17$, i.e., the false positive rate was approximately 17%.

Given the specificity of the novel SNPs, estimated to be 0.83, with the assumption that Affymetrix SNPs are representative of those found in the human genomes we estimate the total number of SNPs in the subject's genome to be approximately 3.7 million (Supplementary Table 3).

Discordance between Genotyping by SNP array and Sequencing

At 26.9% (31,709) of the heterozygous genotyping positions where the second allele was not found by 454 sequencing, the sequence exhibited either the reference (17.8%) or the variant allele as determined by genotyping (5.6%) The asymmetry in the detection of the single reference allele and single variant allele was attributable to the fact that the filters were employed on the data from variant positions, not on reference alleles (see p. 17 "Filtering Criteria for SNPs") effectively lowering the sequence coverage for variants.

A variety of other discordant outcomes of 454 sequence variation are seen at very low rates in Supplementary Table 2. These include a) observation of an alternative allele, that is, a variant that is different from that given in dbSNP, 88 cases (0.02% of all markers); b) observation of a homozygote of the opposing dbSNP allele manifest by genotyping (0.5% of all markers); c) observation of a heterozygote, involving correct dbSNP alleles, where genotyping manifest a homozygous variant allele (0.24% of all markers). These anomalous results comprise 0.75% of all genotyping calls, but the majority of them involve the known variant, therefore it's likely that the DNA sequencing result is correct for some portion of these cases.

Filtering Criteria for Insertions and Deletions

Insertions and deletions (indels) were often associated with short tandem repeats (STR) sequences. Variation in placement of the gap within the repeat to achieve an optimal alignment can vary from read to read causing ambiguity in the precise location of the mutational event. Within separately aligned reads indels, indels in close proximity to one another often represent the same event. The spacing between nearest neighbour deletions, scanned consecutively across the chromosomes is evaluated in Supplementary Fig. 12 as a cumulative proportion. The graph manifests two phases: a rapid accumulation between 1 and 20 bp, followed by a slower accumulation from 20 to over 400 bp. We attribute the rapidly accumulating phase to identical insertions or deletions that were ambiguously placed within STRs, and proceeded to cluster indels in close proximity to one another and record them as a single event. To accomplish this, the genomic coordinates of all deletions and all indels were sorted by start position on each chromosome. A given deletion or insertion N, is grouped with the previous event N-1, if a) ratio of the smaller to larger is 0.8; b) the distance between the start coordinates, is the larger of 15 bp or 7 times the indel size. To be considered a valid indel, a given insertion or deletion had to be supported by at least two reads and the ratio of variant to reference was greater than 0.25.

The read set also harboured over 12.5 million one base indels representing the upper bound of the overall error rate (0.08%; 12.5M/15.3B aligned bases). As expected, given the systematic errors associated with homopolymeric runs in pyrosequencing, 10.4 million of the indels were associated with homopolymeric runs 2-20 bases in length. Furthermore, of the 2.1 million indels not associated with homopolymer runs, 0.13 million were supported by two or more reads. Errors in homopolymer length measurement were a systematic source of indel errors for this technology when employing single reads and decrease with multiple reads[3].

Characterization of No-hit Reads

After the initial mapping there were 1,499,855 reads that failed to find a match in the reference genome by BLAT (see p. 14 "454 Reads and Criteria for Mapping" and Supplementary Table 1). A small set of 10 reads, greater than 200 bp, with no known repeats were compared to Genbank, NT database, using BLAST. Six of the reads matched with >99% identity to phase 3 human fosmids sequenced as part of a human genome gap filling project underway at the University of Washington Genome Center (see http://www.genome.washington.edu/UWGC/Projects/index.cfm?PID=164&ST=2). The fosmids that were matched by the 454 reads mapped to telomeres on chromosomes such that the location of the 454 reads clearly mapped within the telomere to sequence not present in the reference genome (Supplementary Fig. 8).

Assembly and Analysis of No-hit Reads

The 1.5 million no-hit reads were augmented by addition of 2,186,821 reads with low quality matches to the reference, i.e., those that failed the mapping filter due to >4 mismatches or >4 indels (see p. 14 "454 Reads and Criteria for Mapping") for sequence assembly. The low-quality part of the reads (see Supplementary Figs. 13 and 14) was trimmed 50 bp from the 3' end of a read and, further, discarded from assembly any read

without a minimal length of 50 bp. The assembly of the remaining 2.55 million trimmed reads followed the standard ATLAS-WGS procedure[18,19]. Briefly, we detect overlaps using the Atlas Overlapper allowing a maximum discrepancy no more than 5% within the overlap. We removed 445,000 highly repetitive reads, and discarded ambiguous overlaps at potential repeat boundaries. This resulted in approximately one million reads clustered into bins with a minimum of 2 reads per bin. The reads in each overlap bin were and assembled with Phrap.

A total of 169,643 contigs resulted with a total size of 48 Mb and N50 size of 296 bp. We masked repeats in the contigs using RepeatMasker and required that a contig contain at least 100 contiguous bases to further process. 110,353 contigs remained after repeat masking, spanning 29 Mb with a N50 size of 267 bp; 1294 contigs were longer than 1000 bp and the longest was 10,724.

We used megablast with 1e-30 expect cut-off to search a human mRNA sequence database of over 40,000 sequences (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/bigZips/mrna.fa.gz) with these contigs. Significant matches were found to 417 mRNA, which had no map coordinates on build 36 (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/all_mrna.txt.gz). We selected from this set 104 mRNA that matched 886 contigs with >96% identity, and that were covered across greater than 40% of their length. The contigs matching these mRNA sequences ranged in size from 296 to 5121 bp. The 104 mRNA sequences were compared to the reference genome using BLAT to confirm they did not have a cognate gene on the reference genome. From this set, 33 of the mRNA sequences had no hit; 27 had a partial hit, but the genomic match and the contig match did not overlap on the mRNA sequence. The latter may represent small deletions of exons in the reference genome. The remaining 44 mRNA matched the reference genome at a location on the mRNA partially overlapping the contig.

A more sensitive search for related gene sequences was conducted using BLASTX to compare the contigs to the Genbank NR protein database. We limited the search to the set of 1279 contigs greater than 1000 bp in length. Sixty matches with expect < 10^{-10} were saved.

PCR Amplification and Sanger Sequencing

Amplification primers were designed to target regions, but also incorporated at the 5' ends sequence complementary to forward and reverse sequencing primers. PCR targets were amplified using the HotStarTaq Master Mix kit (Qiagen #203446) according to manufacturer's instructions with some modifications. PCR products were generated using 10 ηg of DNA for both the experimental and cell line reference sample, 2.8 μl of the 2X Hot Star Taq Master Mix, 3.2 pmol of each primer and sterile H_20 for a total reaction volume of 8 μl. Reactions were cycled following the Qiagen protocol with longer annealing and extension times of 45 sec for 40 cycles of amplification. Excess primers and dNTPs were removed from the PCR reaction by treatment with 5ul of 1:10 dilution of Exosap-IT (USB #78202). The PCR products were incubated with Exosap-IT at 37°C for 15 min and then inactivated by heating at 80°C for 15 min. Samples were then diluted with 22 μl of 1 X TE (10 mM Tris pH 8.0; 0.2 mM EDTA) to a concentration of approximately 20-40 ηg/μl in preparation for cycle sequencing. Sanger reactions were generated using Applied Biosystems BigDye Terminator v3.1 at 1/64th dilution, 4 pmol primer, 40 ηg of PCR product and standard cycling conditions. Reactions were purified by ethanol precipitation and dried under vacuum. The reactions were resuspended in 20 μl of 0.1 mM EDTA and sequenced on an Applied Biosystems 3730xl DNA Analyzer using the RapidSeq36 run module. Base calls were determined using the 3XX base caller software provided by Applied Biosystems.

Comparative Genome Hybridization

Agilent 244K microarray. The Agilent 244K Whole Human Genome Oligo Microarray Kit (Agilent Technologies, Inc, CA, USA) contains 238,459 formatted 60-mer oligonucleotides, representing a compiled view of the human genome at an average resolution of 9 kilobases. The procedures for DNA digestion, labelling, and hybridization were performed according to the manufacturer's instructions, with some modifications. Briefly, we digested 2 μg of DNA from experimental and reference samples with *Alu*I (10 units) and *Rsa*I (10 units) (Promega) at 37°C for 2 hours. The labelling reaction was performed with Bioprime CGH labelling Module (Invitrogen) at 37°C for 2 hours in the presence of Cy5-dCTP, for the experimental sample, or Cy3-dCTP (PerkinElmer), for the reference sample (standard reference Caucasian male, Kleberg Cytogenetics Laboratory, Baylor College of Medicine). This CGH experiment is referred to as Agilent_1. Experimental and reference targets for each hybridization were purified, pooled and incubated with human Cot-1 DNA (Invitrogen) and Blocking Agent (Agilent Technologies, Inc.). The sample was applied to the array by using an Agilent microarray hybridization chamber, and hybridization was carried out for 40 hours at 65°C in a rotating oven (Agilent Technologies, Inc.). After washing, the slides were scanned into image files using an Axon microarray scanner (GenePix 4000B from Axon Instruments, Union City, CA, USA) and the data were analyzed using the CGH Analytics software (Agilent Technologies, Inc.). The experiment was repeated using a second reference DNA, Coriell number NA10851, and is referred to as Agilent_2.

Nimblegen HD2 microarray. We tested a second sample of experimental DNA was co-annealed with reference DNA, Coriell number NA10851, to a NimbleChip HD2 Array (NimbleGen Systems, Inc, WI, USA) in collaboration with NimbleGen Systems. The HD2 array has 2.1 million probes, each between 50 and 75 bases, with a reported resolution of ~5 kb. Fluorescent images generated using a 5 micron scanner. Log2 ratios were analyzed for copy number variation using NimbleGen SignalMap software.

The Agilent_1 gain and loss events are the results to which the other two experiments, Agilent_2 and NimbleGen, and the DNA sequence data were compared (Fig. 1 c, and Supplementary Table 4). Agilent_2 and NimbleGen, using the same reference DNA sample, had excellent concordance although they were not perfectly concordant (data not shown), which is expected given the differences in oligonucleotide probes represented on the two platforms. Results from Agilent_1 and Agilent_2 were as different from each other as Agilent_1 was from the DNA sequencing coverage (see main text and Supplementary Table 4).

Affymetrix Gene Chip 500

Duplicate genomic DNA samples from lymphocytes, 250 ηg each, were annealed to the Affymetrix 250K *Nsp*I and 250K *Sty*I arrays according to the manufacturer's protocol. Briefly, PCR was performed in quadruplicate for each enzymatic preparation and DNA sample concentrations and lengths were checked using the NanoDrop ND-1000 and gel analysis. Arrays were hybridized in an Affymetrix Hybridization Oven 640 specifically calibrated to the 49C hybridization temperature and scanned using the GeneChip Scanner 3000 with autoloader and 7G upgrade and GeneChip Operating System v. 1.4. Genotypes were determined using the Bayesian Robust Linear Model with Mahalanobis distance classifier (BRLMM) algorithm provided by Affymetrix in the GeneChip Genotyping Analysis software v. 4.0.

LEGENDS to SUPPLEMENTARY FIGURES

Supplementary Figure 1. Data processing summary for Watson whole genome shotgun reads. Over 106.5 million 454-reads were compared to the reference genome (NCBI Build 36) using BLAT. The genomic location with the best match score was taken as the approximate map location for each read. Based on a set of filter rules, 93 million reads with BLAT hits were categorized as mapped unambiguously. The local alignment program cross_match was then applied between the reads and local reference sequence to produce refined alignment. The cross_match results were used to harvest SNPs as well as insertions and deletions. SNPs were separated from 454 sequencing error using a series of filters based on quality score, frequency of observation, and empirical error patterns of 454 sequencing. Indels were also filtered for frequency of observation, and indel positions were clustered to merge events in relative alignment offset due to the presence of short tandem repeats. All SNP and indel events were compared to dbSNP to identify which were known (in dbSNP) and which were novel. SNPs and indels were further classified by their impact on gene and protein structure. Missense SNPs were analyzed by Polyphen to obtain a prediction on the impact of the predicted amino acid change of protein function. Approximately 290 SNPs that were predicted to be damaging were tested in GO for overrepresented gene groupings.

The reads that were not mapped by BLAT or mapped ambiguously were assembled using the Atlas-WGS assembler. The assembled contigs were filtered to remove human repeats (Repeat_masker) and the remaining contigs were compared to human mRNA genes and the NT database to find novel genes and novel human genomic sequences.

Supplementary Figure 2. 454-read coverage on chromosome 1. (p arm) Coverage across chromosome 1 was calculated in consecutive 5K windows. Gray bar corresponds to missing sequence at the centromere. Red dots are regions of 3 or more consecutive windows with coverage less than 3. Regions with long runs of 0 coverage (coloured red) corresponded to highly repetitive or high identity segmental duplications in the reference sequence. Reads were available for these regions but they did not pass the mapping criteria and were removed so they would not interfere with the analysis of sequence variation.

Supplementary Figure 3. The spectrum of base changes observed in subject's SNPs. For known SNPs, the number of transitions (A<->G or T<->C changes) was about 67% of the total base changes. For novel SNPs, the number of transitions was approximately 61% of the total base changes. Red bars, known SNPs right axis; blue bars, novel SNPs, left axis.

Supplementary Figure 4. Detection of heterozygotes as a function of WGS coverage. The theoretical yield of heterozygotes as a function of coverage was calculated as described in “Filtering Criteria for SNPs”, p 17, above.

Supplementary Figure 5. Segment of exon 12 of *SGEF* gene. A 30 bp region of *SGEF* is rendered schematically in the UCSC Genome Browser cantered on the putative 4 bp frame-shift mutation in the subject's genome. The red side-bar adjacent to the “human mRNAs from Genbank” track, which harboured the deletion of GTCA at position 155440983-86. The blue side bar is adjacent to vertebrate genomic sequence exhibiting the deletion of a tandem repeat of the previous GTCA at position 155440987-89. A “Mammalian Gene Collection” mRNA, green side bar, did not bear the deletion.

Supplementary Figure 6. Correspondence of sequence coverage with comparative array hybridization data. A. Heterozygous loss defined by Agilent 244K aCGH. Coverage is measured in 5kb windows (see Fig. 1 main text and Supplementary Fig. 2). Top track ("Loss"), black bar from 55,124,00 to 55,207,000 depicts region predicted by comparative array hybridization to be a heterozygous loss. Track **a**, Agilent aCGH log2 data; track **b** is Nimblegen 2.1 aCGH log2 data; track **c** is 454 sequencing coverage; track **d**, genes in the region; track **f** repeats from RepeatMasker. There were no segmental duplications in this region (track **e**). This event was predicted to delete 4 odorant receptors, OR4C11, OR4P4, OR4S2 and OR4C6. See also Supplementary Table 4, CNV region 11. B. Top track ("Loss"), black bar from 73,071,000 to 73,114,000 depicts a region predicted by comparative array hybridization to be a heterozygous gain. Tracks **a**-**d** are as in part A. Track **e**, segmental duplications are orange, >99% similar, dark yellow >98% similar and light to dark grey, 90-98% similar. This gain event was predicted to create an extra promoter for the PDXDC1 gene in the duplicated homolog. Note the Nimblegen track does not appear to corroborate the gain. Also note the apparent absence of probe data to the left of the predicted region. This may be due to probe failures within the segmental duplications.

Supplementary Figure 7. Repeat classes in subject's "No-hit" reads. Blue bars, Repeats in reads that failed to match the reference sequence (No-hit) blue bars; Repeats in reads that match the reference sequence, red bars.

Supplementary Figure 8. No-hit reads match fosmids extending into the sequencing gaps at the telomeres in the reference. BLAST Non-repetitive "No-hit" reads to the "nt" database of Genbank. Some hit fosmid sequences of finished clones that are not placed on the tiling path. The black box with a

Genbank accession in each graph is the location of the fosmid end hit on the genome, A. AC174047.2, and B. AC174440.2. The red arrow indicates the direction the fosmid sequence goes (extending distally in both cases) and the blue box is the conceptual location of the subject's “No-hit” read on the fosmid. The arrows are not drawn to scale.

Supplementary Figure 9. Characteristics of sequencing reads from the 454 FLX instrument and their matches to the genome. Data from 234 sequencing runs were filtered in two stages to produce two sets: Tier 1 and Tier 2, see p. 12 above, “Image and Signal Processing.” A. Read lengths; B. Number of hits to the genome; C. Alternate hits, that is, matches that scored within 1% of top scoring hit; D. Fraction of read matched to genome; E. Number of mismatches in best hit; F. Number of indels in best hit.

Supplementary Figure 10. Association of homopolymer runs with errors on X chromosome. SNPs appearing as heterozygotes on the X chromosomes were considered to be errors. In this figure the so-called heterozygotes were seen to be disproportionately associated with long homopolymers (5 bp or longer) compared to the homozygous, suggesting long homopolymers may trigger substitution events.

Supplementary Figure 11. Association of known and novel SNPs with homopolymer runs. Novel SNPs were seen to be disproportionately associated with homopolymer runs. Only 3% of known SNPs were found in proximity to homopolymer runs >5 bp in length whereas 33% of novel SNPs were associated with homopolymer runs >5 bp.

Supplementary Figure 12. The distance between adjacent deletions (blue) and insertions (red) as a cumulative proportion.

Supplementary Figure 13. Accumulation of errors at each base position in a set of *S. aureus* reads. More than >300,000 reads from an *S. aureus* clone were aligned to a reference genome. All mismatches in this data set were assumed to be error. The indel rate increase dramatically toward the end of the read in parallel to the mismatch base rate. The indel rate is 5-7 times greater.

Supplementary Figure 14. Error rate as a function of distance to the 3' end of a read. The error rate (blue diamonds, left axis) was transformed into a $\log_{10}$ score (red diamonds, right axis).

Supplementary Figure 15. Error frequency and Q score. Distribution of Q scores in 454 reads. Q scores were evaluated in *S. aureus* reads mapped to a high quality reference genome (see p. 16 "Errors in 454 Reads"). Red line, distribution of all Q scores; bars, distribution of Q scores at erroneous bases.

SUPPLEMENTARY TABLES

Supplementary Table 1. Summary of 454 read mapping

Stage	Tier 1[a]	Tier 2[a]	Total
Input	78,563,567	29,618,639	108,182,206
Matched[b]	78,030,003	28,694,007	106,724,010
Duplicate[c]	2,186,821	41,655	2,228,476
Nomatch	533,568	966,287	1,499,855
Pass Filter[d]	67,563,754	25,616,765	93,180,519

a, Tier 1 reads pass the primary filter for sequence quality and length from the base calling step (see Supplementary Methods p. 12 "Image and signal processing"); Tier 2 reads pass the quality filter but not the length filter (see Supplementary Fig. 9A for length distribution of 454 FLX reads).

b, Reads that failed to find a match to the reference genome under BLAT conditions used for mapping the reads.

c, Reads originating from two or more beads and a single DNA fragment in the same emulsion droplet. Recognized as two reads within a sequencing run that have the same start coordinate and lengths within 2 bp.

d. Reads whose matching alignment passed the quality criteria as described in the Supplementary Methods.

Supplementary Table 2. Microarray validation unabridged of 454 SNPs

Affymetrix Genotype[a]	Affymetrix Tally	454 Sequence Outcome[b]	454 Tally	Percent Agreement
Homo Ref	254753	A. Homo Ref	253348	99.4
		B. Homo Var	578	
Homo Var	104547	C. Homo Var	99387	95.1
		D. Homo Ref	2016	
		E. Hetero[c]	2266	
		F. Homo Alt[d]	43	
Hetero	135413	G. Hetero	102702	75.8
		H. Homo Ref	24062	(17.8)[e]
		I. Homo Var	7647	(5.65)[e]
		J. Hetero Alt[f]	45	

a, Homo Ref, homozygous for reference allele; Homo Var, homozygous for the variant allele; Hetero, heterozygous. Affymetrix counts for each genotype are given in the second column.

b, The genotype based on the alleles observed in 454 reads at each position of an Affymetrix SNP.

c, Hetero, these positions had two alleles, which were the dbSNP alleles, even though the Affymetrix call had been homozygous.

d, Homo Alt, mean the site was homozygous but the allele present was not the one in dbSNP. The sites heterozygous according to Affymetrix were sometimes called as homozygous by 454 sequencing. This is due mainly to fluctuation in coverage (see Fig. 1 b main text).

e, Numbers in parenthesis indicate positions that Affymetrix genotyping called heterozygous, but sequencing correctly identified only one of the two alleles.

f, Hetero Alt, SNPs that were called heterozygous but the variant alleles did not match those listed dbSNP.

Supplementary Table 3. Estimation of subject's genome-wide SNP tally[a]

	Observed		Predicted[b]		Total Genome
	Known	Novel	Known	Novel	
Het	1.35	0.51	1.74	0.55	2.29
Homo	1.36	0.10	1.34	0.09	1.42
Total	2.71	0.61	3.08	0.64	3.72

a, All numbers are in millions.

b, Predicted SNP tallies were calculated by multiplying the observed tallies by the specificity and dividing by the sensitivity (see text "Sensitivity and Specificity of SNP Discovery", p. 19 above).

Supplementary Table 4. Copy number variation regions in subject's genome[a]

Region	Cyto	Start	End	Length (bp)	Agilent_1[b]	Agilent_2[b]	Nimblegen[c]	454Seq Cov[d]	Seg Dup Cov[e]	Tandem Dup Gene Family[f]	Genes	Annotation
1	chr1: p36.13	16,820,714	17,130,584	309,870	-	-	-	+/-	100%	No	CROCC	Ciliary rootlet coiled-coil. Deletion of 5' exon; Involved in centrosome organization and biogenesis. Complex gene with evidence for ~12 transcripts and several proteins with no sequence overlap.
2	chr1: p36.11	25,482,036	25,536,138	54,102	+	+	+	0	100%	Yes	RHD	Rh blood group D, deleted at high frequency.
3	chr1: q21.3	150,823,073	150,852,905	29,832	--	--	--	--	0%	Yes	LCE3C, LCE3B	Epiderminal differentiation factors, upregulated in response to environmental factors (UV radiation). Breakpoint mapped, this study Chr1:150,822,165 150,854,364 (see Figure 4).
4	chr1: q31.3	195,005,520	195,165,287	159,767	+	+	+	+	67%	Yes	CFHR3, CFHR1, CFHR4	Compliment factor H-related. Binds specific bacterial proteins, deletion increases risk of atypical hemolytic uremic syndrome; no information available on the effect of duplications.
5	chr1: q44	246,794,522	246,875,051	80,529	-	-	-	-	50%	Yes	OR2T34, OR2T10, OR2T11, OR2T35	Members of odorant receptor family 2.
6	chr2: q37.3	242,602,220	242,701,038	98,818	-	-	-	-	67%	No		
7	chr3: q26.1	163,997,228	164,101,835	104,607	+	+	+	-0	0%	No		
8	chr4: q13.2	69,057,735	69,165,872	108,137	+	+	+ (3)	+	50%	Yes	UGT2B17	UDP-glucuronosyltransferase 2 polypeptide B17 (EC 2.4.1.17), deletion polymorphism is associated with a reduced rate of NNAL detoxification in vivo and may increase individual susceptibility to tobacco-related cancers & prostate cancer.
9	chr4: q13.2	70,188,454	70,299,604	111,150	-	0	0	-	99%	Yes	UGT2B28	UDP-glucuronosyltransferase 2 polypeptide B28, expressed in liver, probable role in catabolizing androgen and estrogen.
10	chr5: p15.33	745,859	826,103	80,244	-	0	0	-	100%	No	TPPP	Tubulin polymerization promoting protein. Delete 5' exon. Plays role stabilization of physiological microtubular ultrastructures and cell survival. Suggested to play a pro-aggregatory role in the common neurodegenerative disorders hall-marked by alpha-synuclein aggregates.
11	chr6: p21.32	32,586,131	32,629,907	43,776	+	0	0	--	10%	Yes	HLA-DRB5, HLA-DRB1	Major histocompatibility complex II, DR beta 5, DR beta 1.
12	chr7: p14.1	38,281,765	38,318,969	37,204	-	-?	0	-	5%	No	TARP promoter	T-cell receptor gamma alternate reading frame protein. This gene is the unrearranged T-cell receptor gamma and is expressed in non-lymphoid tissues. Complex locus with ~20 variant transcripts and several proteins with no sequence overlap.
13	chr9: q33.2	123,371,378	123,421,714	50,336	-	-	-	-	0%	No	DAB2IP	Disabled homolog 2 interacting protein. A Ras GTPase-activating protein (GAP) that acts as a tumor suppressor gene and is inactivated by methylation in prostate and breast cancers.
14	chr11: q11	55,124,730	55,207,364	82,634	-	-	-	-	0%	Yes	OR4C11, OR4P4, OR4S2, OR4C6	Members of odorant receptor family 4.
15	chr14: q11.2	21,635,428	22,012,322	376,894	-	-	-	-	0%	No		

16	chr14: q24.3	73,071,404	73,114,182	42,778	-	0	0	--	99%	Yes	ACOT1, ACOT2, HEATR4	Acyl-CoA thioesterase 1 & 2. Involved in very long chain fatty acid metabolism. HEATR4 (3' exon). HEAT repeats (37 47 aa, related to ARM repeats) are thought to be protein-protein interaction surfaces; proteins containing these structures are involved in intracellular transport. Complex locus producing >10 variant transcripts; several proteins with no sequence overlap. HEAT=**h**untingtin, **E**F3, PP2**A** and yeast P13-kinase **T**OR1.
17	chr14: q32.33	105,082,760	105,286,523	203,763	-	0	0	- (2)	100%	No		
18	chr15: q11.2	18,810,051	19,465,418	655,367	-	-	- (7)	+/-	95%	No	POTE15, LOC283755	Protein expressed in **p**rostate, **o**vary, **t**estis, and **p**lacenta 15. Harbors a protein-protein interaction Ankyrin domain. LOC283755 is a complex locus producing >25 variant transcripts and multiple proteins with no sequence overlap. This deletion is on the proximal boundary of the Prader-Willi/Angleman syndrome region (15q11-q13) adjacent to the centromere.
19	chr16: p13.11	14,956,177	14,997,419	41,242	+	0	0	+	90%	No	PDXDC1	Pyridoxal-dependent decarboxylase domain containing 1. Group II decarboxylase family. Complex locus with ~25 variant transcripts and several proteins with no sequence overlap.
20	chr16: p11.2	31,958,973	33,539,082	1,580,109	+	+	+ (6)	+	80%	No	TP53TG3, LOC729355	TP53 target gene 3. inducible by TP53 in colon cancer cell line SW480. Normal expression predominantly in testis. Speculation: plays role in p53-mediated signaling.
21	chr17: q21.31	41,521,544	41,645,038	123,494	+	+	+ (3)	+	0%	No	KIAA1267	Full length cDNA from human brain cDNA library. 5' exon. Found in many other tissues; complex transcription ~16 variant transcripts and several proteins with no sequence overlap.
22	chr22: q11.23	22,677,959	22,725,353	47,394	+	0	+ (2)	-	30%	Yes	GSTT1	Glutathione S-transferase theta 1, polymorphisms affecting the activity of the gene may be associated with the risk of developing chronic severe ethanol liver damage; associations for cancers reported: lung cancer short term survival, stomach cancer, adul
23	chr22: q13.1	37,689,058	37,715,431	26,373	-	-	-	--	20%	Yes	APOBEC3B APOBEC3A	Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3B, 3A (3' exon). Related to C to U RNA-editing cytidine deaminases, suggested to play role in growth or cell cycle control. Several reports that gene plays a role in inate immunity against specific lentiviruses, intracisternal A-particles and inhibits L1 retrotransposition by a DNA deamination-independent mechanism. Breakpoint mapped by Kidd et al. '07 PLOS Genetics. Build 36, chr22:37688577-37718162.

a, Copy number regions defined by log2 ratio thresholds for Agilent_1 (Agilent 244K aCGH chip). Green shading indicates the Agilent_1 result was supported by one or more of the other platforms: Agilent_2, Nimblegen or 454 sequence coverage.

b, Variation type by Agilent_1 and Agilent_2 chip, '+' is gain, '-' is loss, '--' homozygous deletion. '?' indicates the regions overlapped but were not identical. The two chips differ in the reference DNA to which the subject's DNA was compared. Agilent_2 and Nimblegen HD2 used identical reference DNAs (see Methods).

c, Variation type by NimbleChip HD2. '+', '-' as in footnote b; 0, no variation; The nimblegen data resolved some Agilent regions into multiple events indicated by numbers in parentheses.

d, Variation type by 454 sequence coverage measured over 5kb windows. '+','-','0' as in footnote c. '+/-' both gain and loss within regions.

e, Percent of CNV region occupied by segmental duplications.

f, Presence of tandemly duplicated gene families in CNV region.

g, Genes that lie within CNV region or across the predicted breakpoints.

Supplementary Table 5. Mutations matching HGMD markers with possible disease association

HGMD_Acc	Chr	Coordinate	HUGO Symbol	Gene	Cyto	Phenotype	Zygosity
CM981315	1	11777063	MTHFR	5,10-methylenetetrahydrofolate reductase (NADPH)	1p36.3	Neural tube defect, additional risk	Het
CM041751	1	70677388	CTH	Cystathionase (cystathionine gamma-lyase)	1p31.1	Homocysteine levels	Homo
CM015072	1	94316822	ABCA4	ATP-binding cassette, sub-family A (ABC1), member 4 (Stargardt disease 1, ABCR)	1p22	Stargardt disease	Het
CM003809	2	38155681	CYP1B1	Cytochrome P450, subfamily 1 (dioxin inducible), polypeptide 1	2p21	Breast or lung cancer	Homo
CM004824	2	38155894	CYP1B1	Cytochrome P450, subfamily 1 (dioxin inducible), polypeptide 1	2p21	Higher catalytic activity	Homo
CM031355	3	38620424	SCN5A	Sodium channel, voltage gated, type V, alpha polypeptide	3p	Phenotype modifier	Homo
CM941277	3	172214994	SLC2A2	Solute carrier family 2 (facilitated glucose transporter), member 2 (GLUT2)	3q	Diabetes, NIDDM	Het
CM013814	5	147460220	SPINK5	Serine protease inhibitor, Kazal type 5	5q32	Atopy, maternally inherited	Homo
CM971591	8	31144196	WRN	Werner syndrome	8p	Thromboembolic disease	Het
CM962423	8	143993541	CYP11B2	Cytochrome P450, subfamily XIB (steroid 11-beta-hydroxylase), polypeptide 2	8q	Low renin hypertension	Homo
CM972826	12	46559162	VDR	Vitamin D receptor	12q	Higher bone mineral density,	Het
CM001349	12	119901033	TCF1	Transcription factor 1, hepatocyte nuclear factor 1	12q24.3	Insulin resistance	Het
CM024106	19	1348443	GAMT	Guanidinoacetate N-methyltransferase	19p13.3	Arginine:glycine amidinotransferase deficiency	Het
CM030471	19	18041413	IL12RB1	Interleukin 12 receptor, beta 1	19p13.1	Tuberculosis, susceptibility to	Het
CM030470	19	18041451	IL12RB1	Interleukin 12 receptor, beta 1	19p13.1	Tuberculosis, susceptibility to	Het
CM002115	19	46550761	TGFB1	Transforming growth factor, beta 1	19q13.1	Osteoporosis	Het
CM004814	19	50546759	ERCC2	Excision repair cross-complementing rodent repair deficiency, complementation group 2 (XPD)	19q13	Improved DNA repair	Het
CM015299	19	50559099	ERCC2	Excision repair cross-complementing rodent repair deficiency, complementation group 2 (XPD)	19q13	Increased response to UV	Het
CM890104	20	4628251	PRNP	Prion protein	20p	Gerstmann-Straeussler syndrome	Het
CM023931	20	56912202	GNAS	GNAS complex locus	20q13	Essential hypertension	Het

Supplementary Table 6. mRNA sequences matching novel contigs

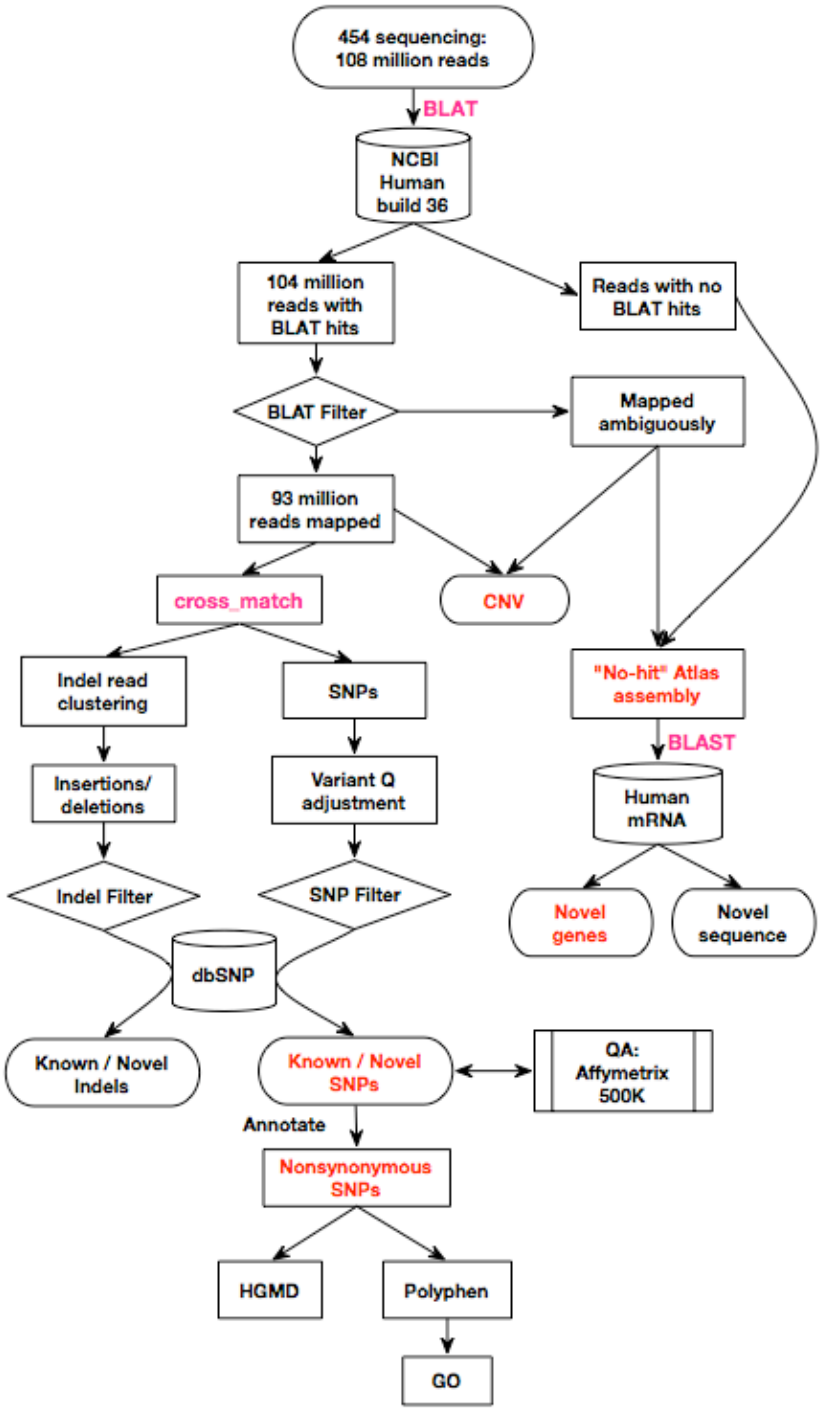
Accession	Title	cDNA tissue	ORF integrity[a]	Protein Accno	Function	Clone date
AF007190.1	SIB 227C intestinal mucin (MUC3) mRNA, partial cds	intestinal mucosa	1 ORF	AAC02268.1	integral membrane protein	1997
AF007191.1	SIB 276 intestinal mucin (MUC3) mRNA, partial cds	intestinal mucosa	1 ORF	AAC02269.1	integral membrane protein	1997
Z34280.1	MUC5AC mRNA for mucin (partial)	tracheobronchial mucosa	2 ORF	CAA84034.1[b]	oligomeric gel-forming mucin	1994
Z34281.1	MUC5AC mRNA for mucin (partial)	gastric mucosa	2 ORF	CAA84035.1	oligomeric gel-forming mucin	1994
AF038190.1	clone 23582 mRNA sequence	infant brain	pieces	none	unknown	1997
AJ001481.1	DUX1 protein	rhabdomyosarcoma, TE671	1 ORF	CAA04776.1	transcription factor, 2 homeo domain	1998
AK022843.1	FLJ12781 fis, clone NT2RP2001861	teratocarcinoma, NT2, 2 wks post retinoic acid induction	1 ORF	BAB14267.1	unknown	2004
AK021843.1	FLJ11781 fis, clone HEMBA1005963	whole embryo	1 ORF	BAB13908.1	unknown	2004
AK054704.1	FLJ30142 fis, clone BRACE2000183	cerebellum	pieces	none	unknown	2001
AK057223.1	FLJ32661 fis, clone TESTI1000055	testis	1 ORF	BAB71386.1	weakly similar to homeobox protein SIX1	2001
AK092540.1	FLJ35221 fis, clone PROST2000761	prostate	pieces	none	unknown	2002
AK092650.1	FLJ35331 fis, clone PROST2014659	prostate	1 ORF	BAC03936.1	unknown	2002
AK097509.1	FLJ40190 fis, clone TESTI2019228	testis	1 ORF	BAC05080.1	unknown	2002
AK097979.1	FLJ40660 fis, clone THYMU2019686	thymus	pieces	none	unknown	2002
AK123447.1	FLJ41453 fis, clone BRSTN2011211	subthalamic nucleus	pieces	none	unknown	2002
AK123640.1	FLJ41646 fis, clone FEBRA2024019	fetal brain	pieces	none	unknown	2003
AK125773.1	FLJ43785 fis, clone TESTI2052211	testis	1 ORF	BAC86282.1	unknown	2003
AK127140.1	FLJ45197 fis, clone BRCAN2002892, moderately similar to Ras-related protein Rab-7	caudate nucleus	pieces	none	unknown	2003
AK128667.1	FLJ46827 fis, clone UTERU2016464	uterus	pieces	none	unknown	2003
AL162012.1	from clone DKFZp761N09121; partial cds	amygdala	1 ORF	CAB82364.1	nonclathrin coat protein gamma2-COP	2000
AL162042.1	from clone DKFZp761L1212	amygdala	pieces	none	unknown	2004
AL390137.1	from clone DKFZp547C074	fetal brain	pieces	none	unknown	2004
AL390147.1	from clone DKFZp547D065	fetal brain	1 ORF	CAB99089.2	unknown	2004
AX747653.1	Sequence 1178	Patent EP1308459	pieces	none	unknown	2003
AY094596.1	Ras-related protein Rab-7 (RAB7) mRNA, complete cds	unknown	1 ORF	AAM22519.1	ras oncogene family	2002
BC007382.1	RAB7B, member RAS oncogene family, mRNA (cDNA clone IMAGE:3658826), complete cds	endometrial adenocarcinoma	1 ORF	AAH07382.1	nucleotide binding, small GTPase mediated signal transduction	2002
BC017092.1	RAB7B, member RAS oncogene family, mRNA (cDNA clone MGC:9726 IMAGE:3851998), complete cds	colon adenosarcoma	1 ORF	AAH17092.1	nucleotide binding, small GTPase mediated signal transduction	2002
BC031926.1	hypothetical MGC50722, mRNA (cDNA clone IMAGE:4508951), partial cds	testis embryonal carcinoma	1 ORF	AAH31926.1	possibly related to a rat centrosome associated protein	2002
BC042107.1	cDNA clone MGC:50722 IMAGE:5170879, complete cds	medullla	1 ORF	AAH42107.1	similar to rat centrosome-associated protein 350	2002
BC048003.1	IMAGE clone:5314428	hypothalmus	pieces	none	unknown	2003
BC087853.1	Gene family with sequence similarity 20, member C, mRNA (cDNA clone IMAGE:30375185), partial cds	White Matter pool: 5 brain tissues-- femoral artery, olfactory tract, optic tract, cerebellar white matter, cerebral white matter	1 ORF	AAH87853.1	DUF domain share among several proteins, function unknown. Homology with a mouse dentin matrix protein.	2002
BX647503.1	from clone DKFZp686H1297	retina	pieces	none	unknown	2003
CS267002.1	sequence 694 from Patent WO2005118806	Patent WO2005118806	pieces	none	unknown	2005

a, 1 or 2 ORF, single large open reading frame; pieces means no single continuous ORF--suspicious for pseudogene or sequencing errors
b, CAA84034.1, appears to be an incorrectly annotated ORF

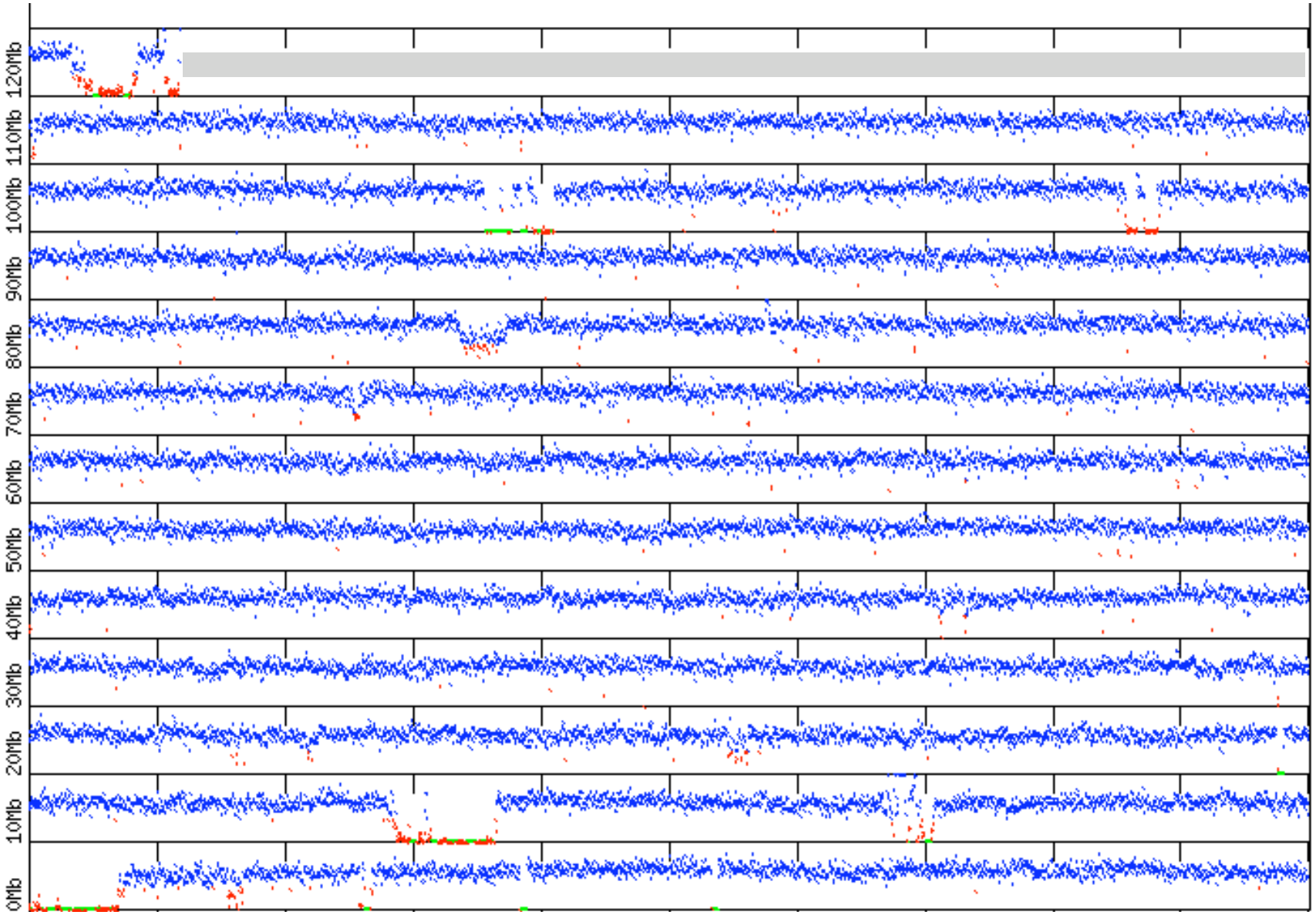
Supplementary Table 7. Genes related to novel contigs

Nucleotide_ID	Protein_ID	Species	Gene	DEFINITION
NM_033178.2	NP_149418	Homo sapiens	DUX4	double homeobox 4
AK074131.1	BAB84957	Homo sapiens		FLJ00204 protein
AY204751.1	AAP13576	Homo sapiens	ABCA13	ABC A13
NM_020223.1	NP_064608	Homo sapiens		family with sequence similarity 20, member C
AK127601.1	BAC87052	Homo sapiens		unnamed protein product
XM_373859.2	XP_373859	Homo sapiens		hypothetical protein
XM_372335.1	XP_372335	Homo sapiens		similar to double homeobox protein
XM_370593.1	XP_370593	Homo sapiens		similar to double homeobox protein
XM_374852.1	XP_374852	Homo sapiens		similar to double homeobox protein
XM_378973.1	XP_378973	Homo sapiens		hypothetical protein XP_378973
NM_203348.1	NP_976223	Homo sapiens		hypothetical protein LOC399693
AJ006205.1	CAB43492	Homo sapiens	MUC-B1	MUC-B1
AF113616.1	AAF13032	Homo sapiens	MUC3	intestinal mucin 3
AK002086.1	BAA92078	Homo sapiens		unnamed protein product
XM_001134429	XP_001134429.1	Homo sapiens	MUC5AC	S56015 gastric mucin MUC5AC
T51869(EST)	T51869	Homo sapiens		hypothetical protein DKFZp547C074.1 (fragment)
U50040.1	AAC50453	Homo sapiens	SIP-110	signaling inositol polyphosphate 5 phosphatase
NM_001733	NP_001724.3	Homo sapiens	C1R	1102166A compliment C1r, activated form
XM_354799.1	XP_354799	Mus musculus		similar to axonemal dynein heavy chain 7
XM_358044.1	XP_358044	Mus musculus		similar to Ac1147
NM_176893.2	NP_795712	Mus musculus	Mink1	misshapen-like kinase 1 isoform 2
NM_028746.2	NP_083022	Mus musculus	Slc7a13	aspartate/glutamate transporter 1; solute carrier family 7, (cationic amino acid transporter, y+ system) member 13
XM_149659.3	XP_149659	Mus musculus		RIKEN cDNA 1110030H18
XM_344083.1	XP_344084	Rattus norvegicus		similar to Forkhead box protein L1
XM_344435.1	XP_344436	Rattus norvegicus		similar to repetin
NM_133309.2	NP_579843	Rattus norvegicus	Capn8	calpain 8
CAAE01015033.1	CAG11277	Tetraodon nigroviridis		unnamed protein product
CAAE01014989.1	CAG07279	Tetraodon nigroviridis		unnamed protein product
CAAE01014581.1	CAF99752	Tetraodon nigroviridis		unnamed protein product
BC070720.1	AAH70720	Xenopus laevis	Mcm9	mini-chromosome maintenance deficient 9
AY450930.1	AAS07044	Chlamydomonas reinhardtii	SAG1	plus agglutinin
AJ437620.1	CAD27181	Streptococcus agalactiae	fbsA	fibrinogen-binding protein
M59166.1	Q01133	Calliactis parasitica		Antho-RFamide neuropeptides precursor

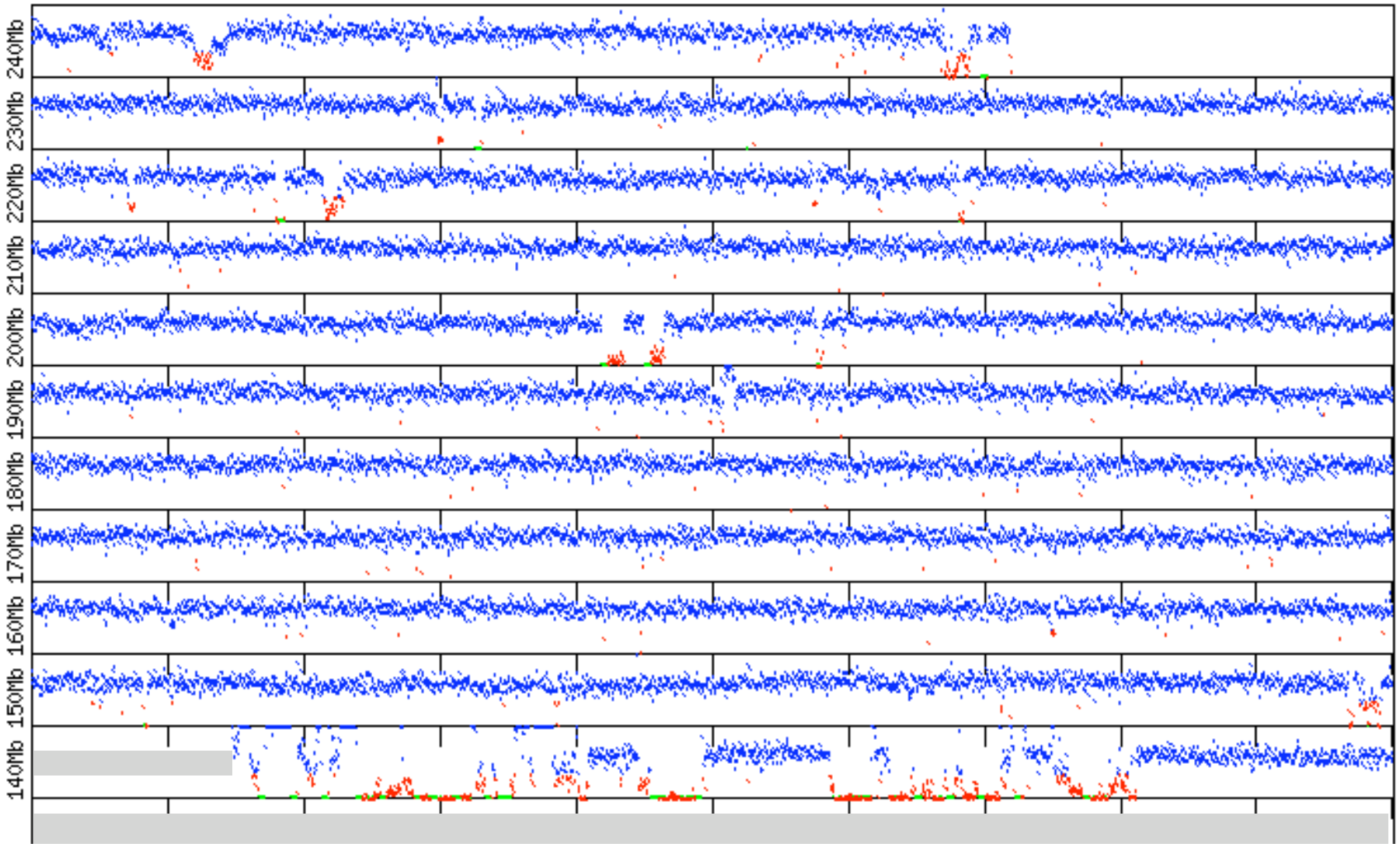
Supplementary Figure 1.



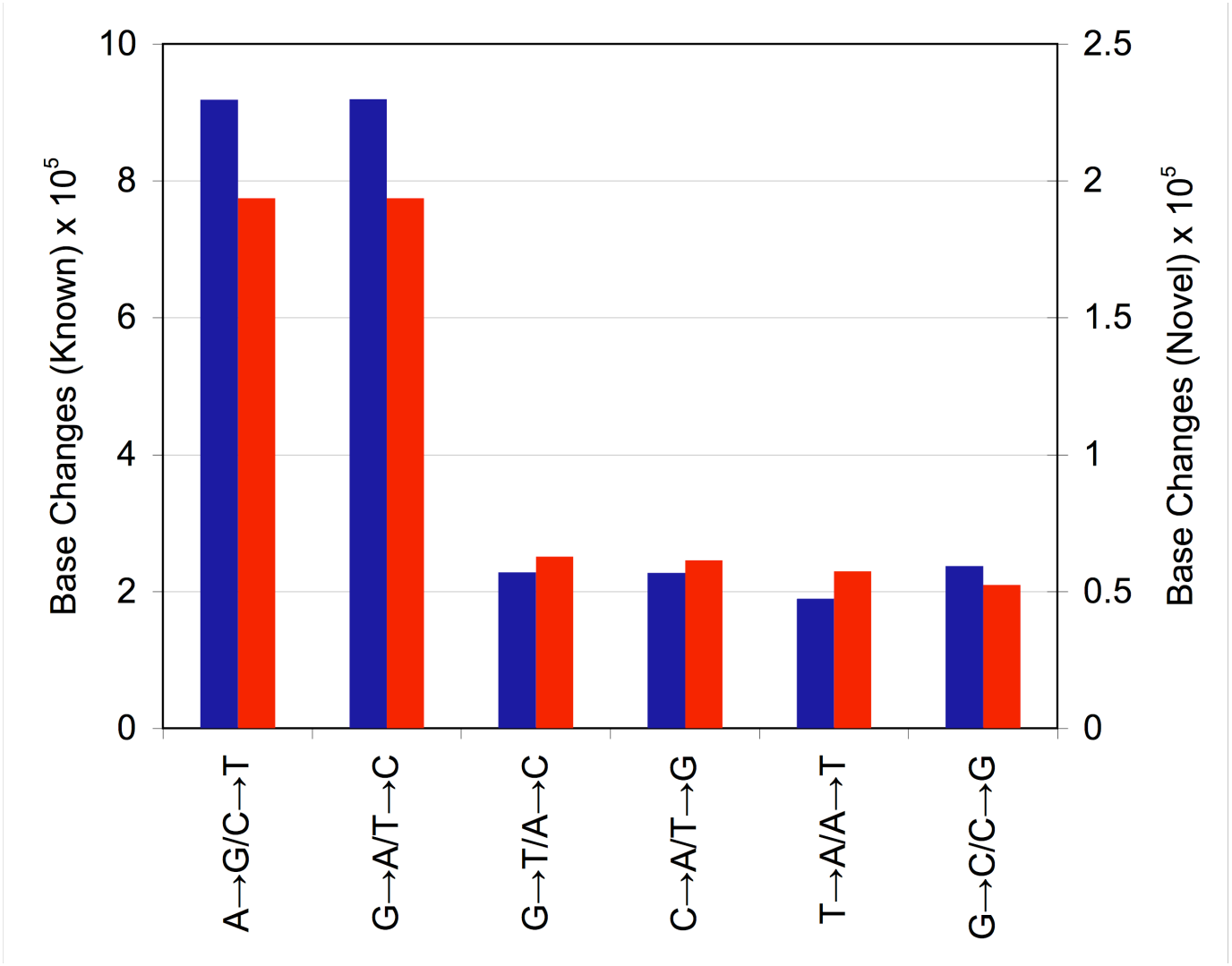
Supplementary Figure 2 a.



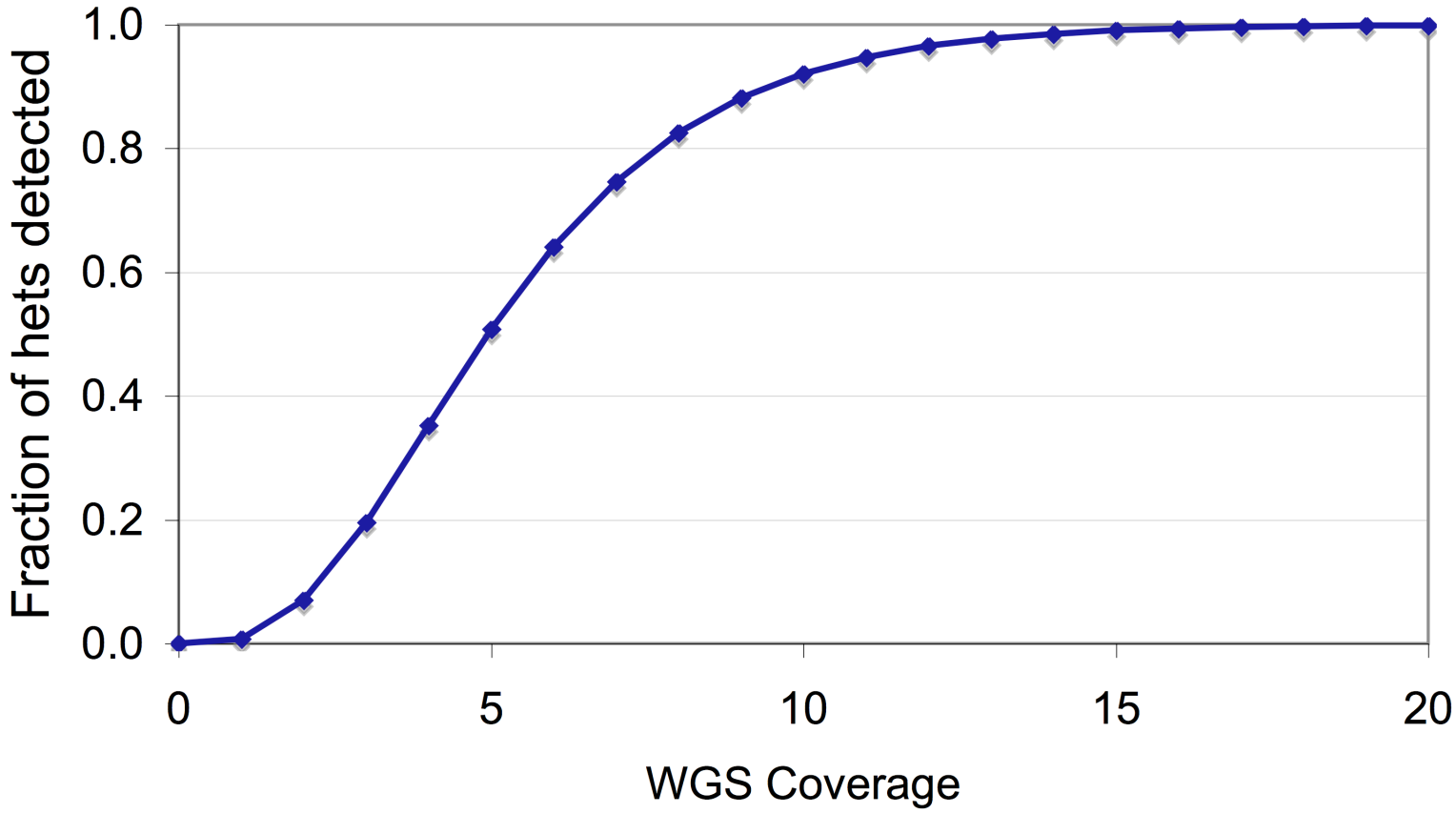
Supplementary Figure 2 b.



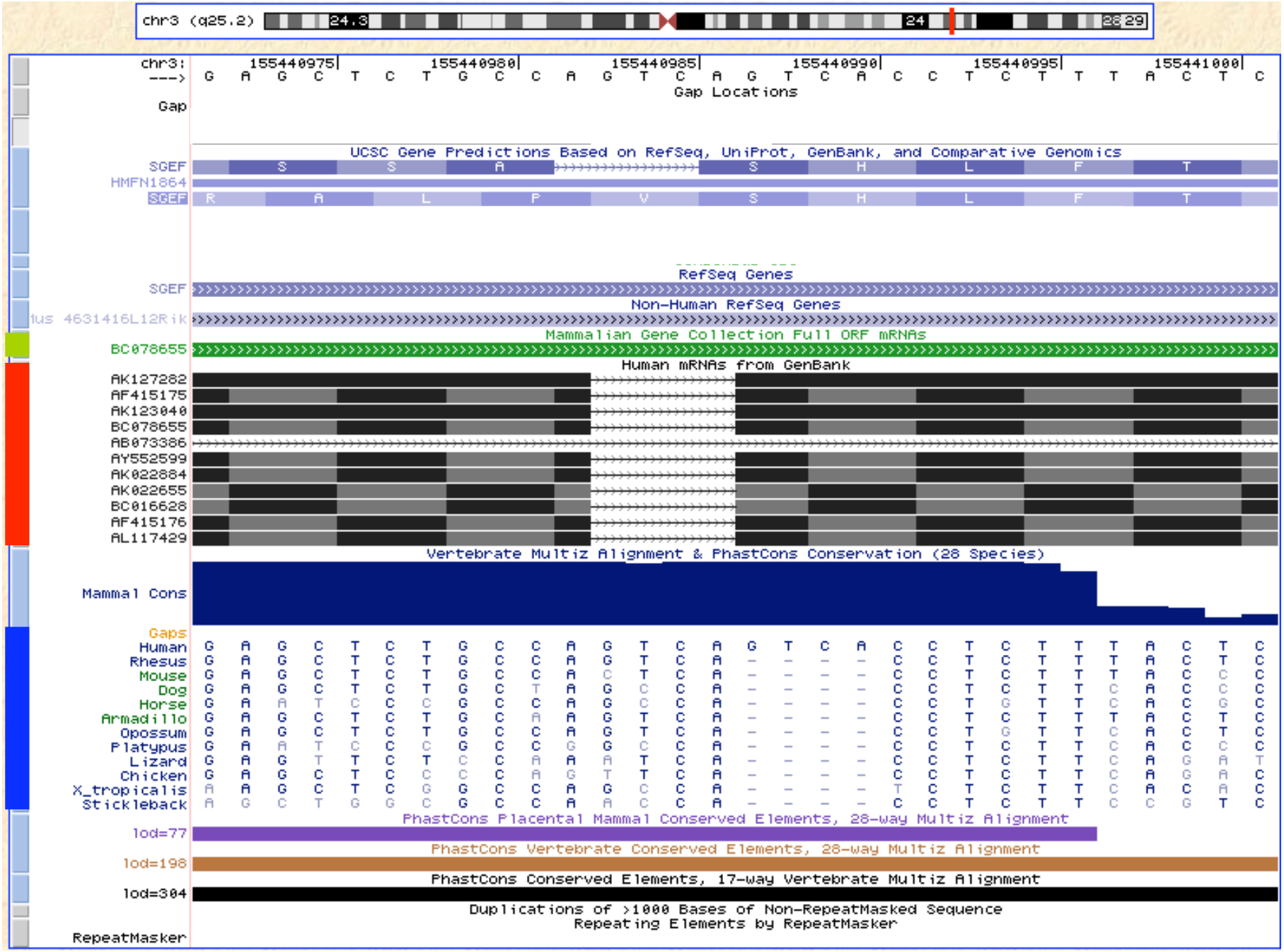
Supplementary Figure 3.



Supplementary Figure 4.

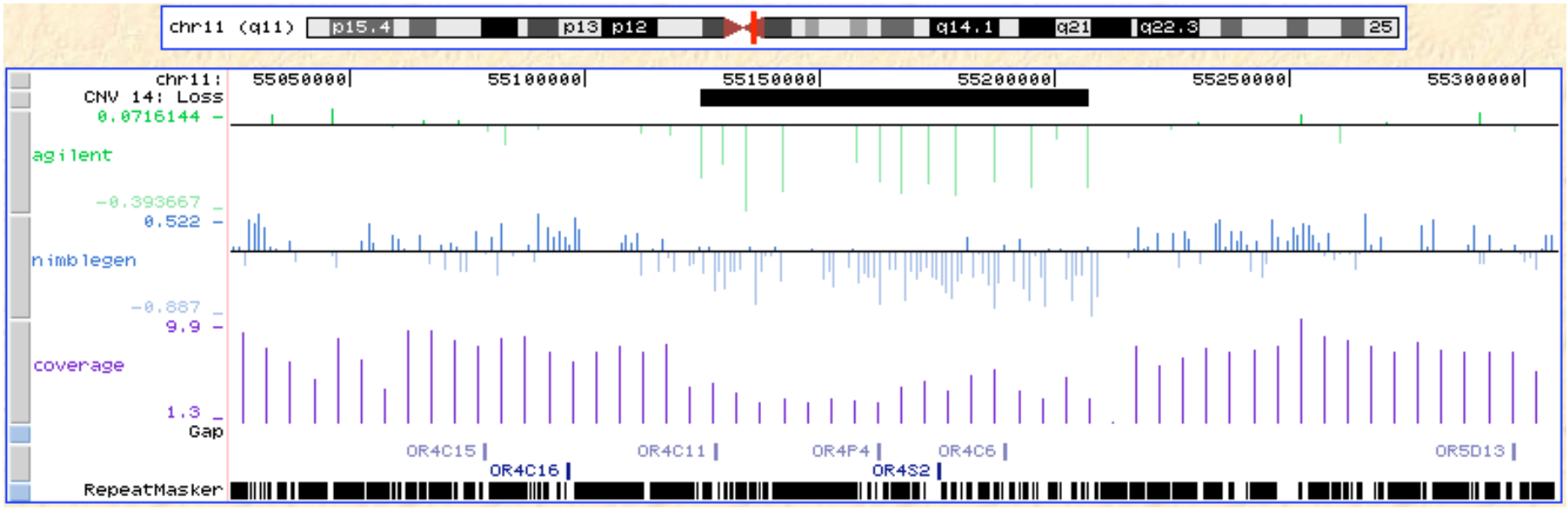


Supplementary Figure 5.



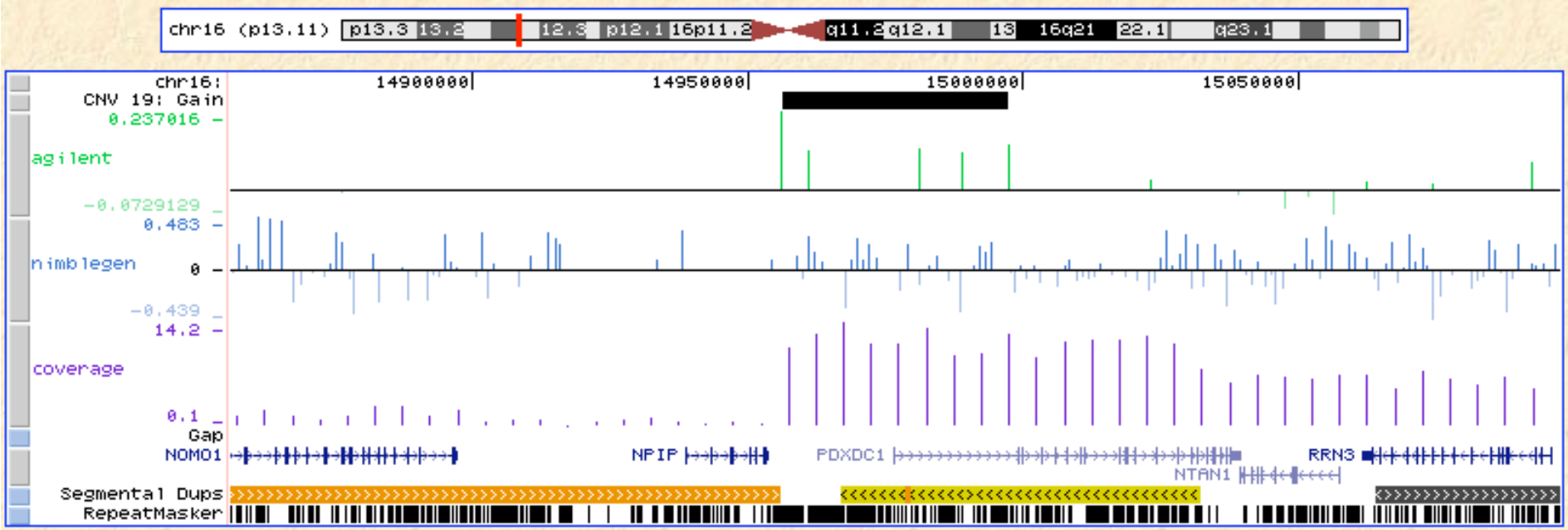
Supplementary Figure 6 a.

Heterozygous loss



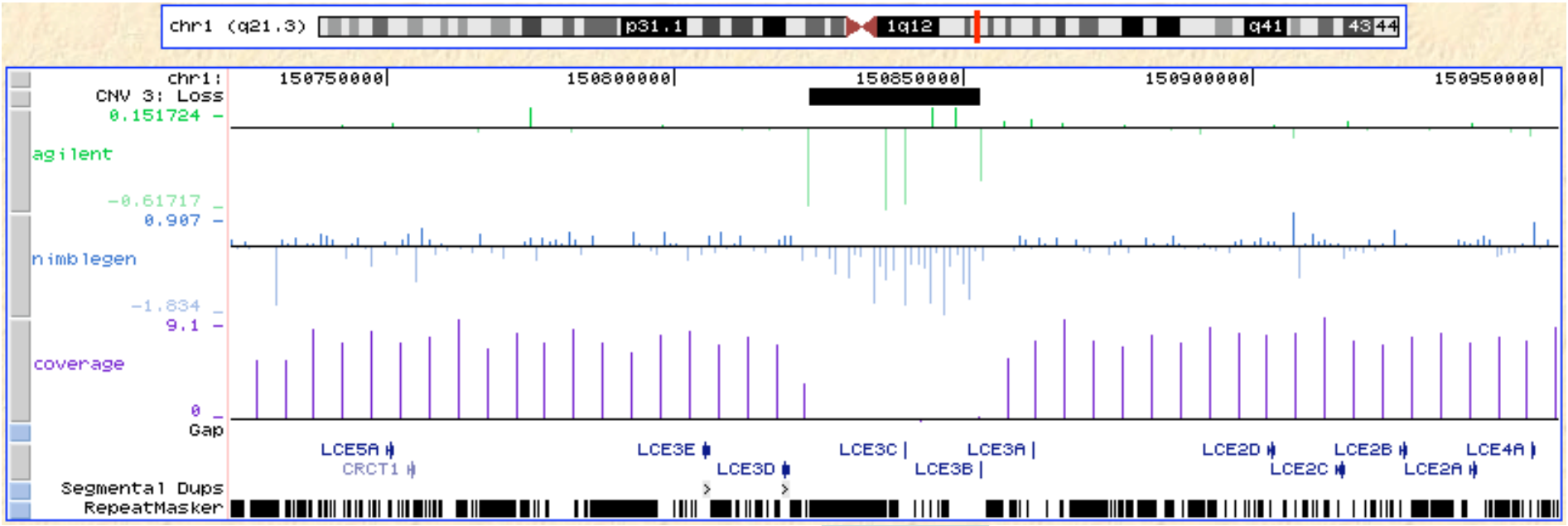
Supplementary Figure 6 b.

Heterozygous gain



Supplementary Figure 6 c.

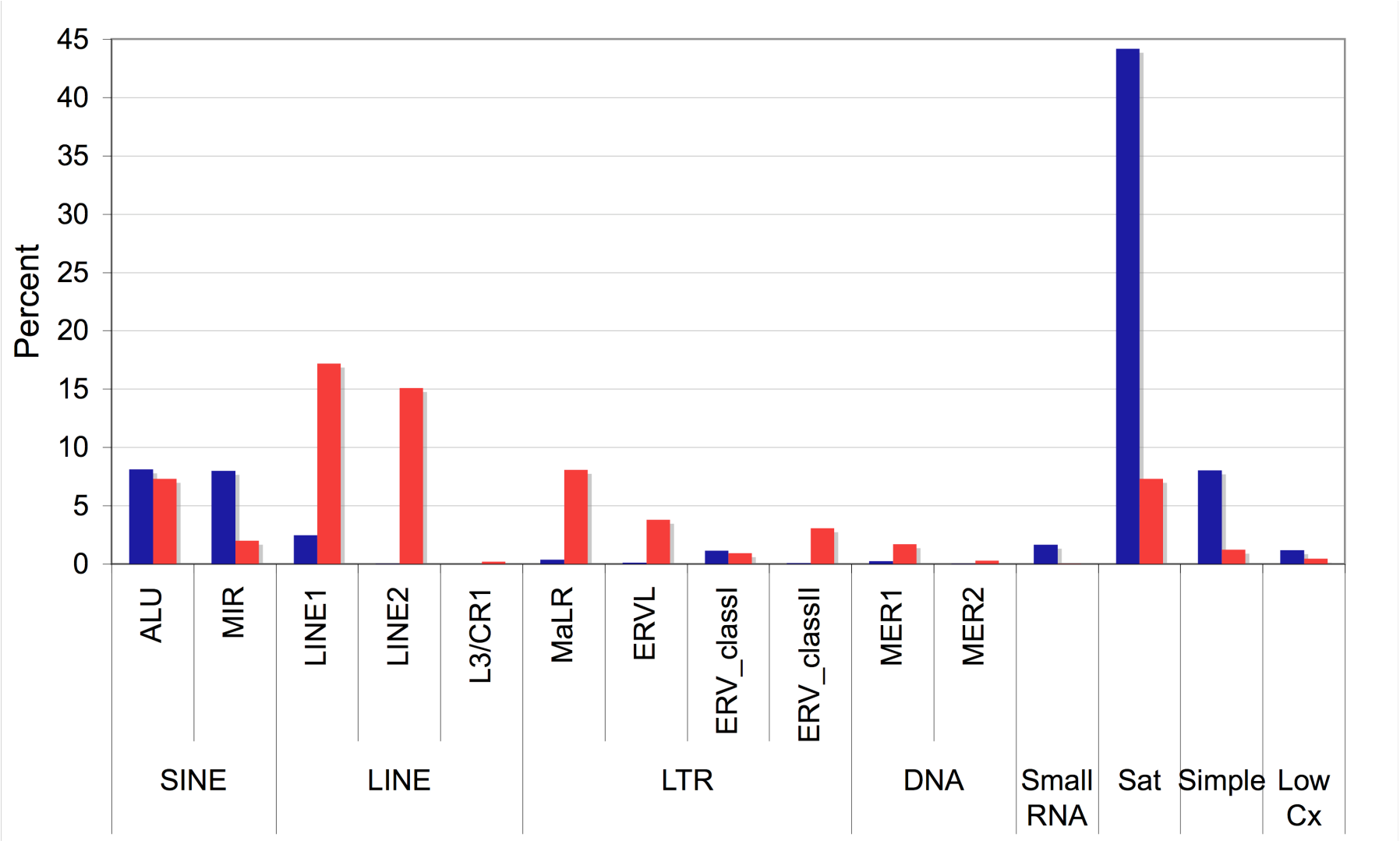
Homozygous loss



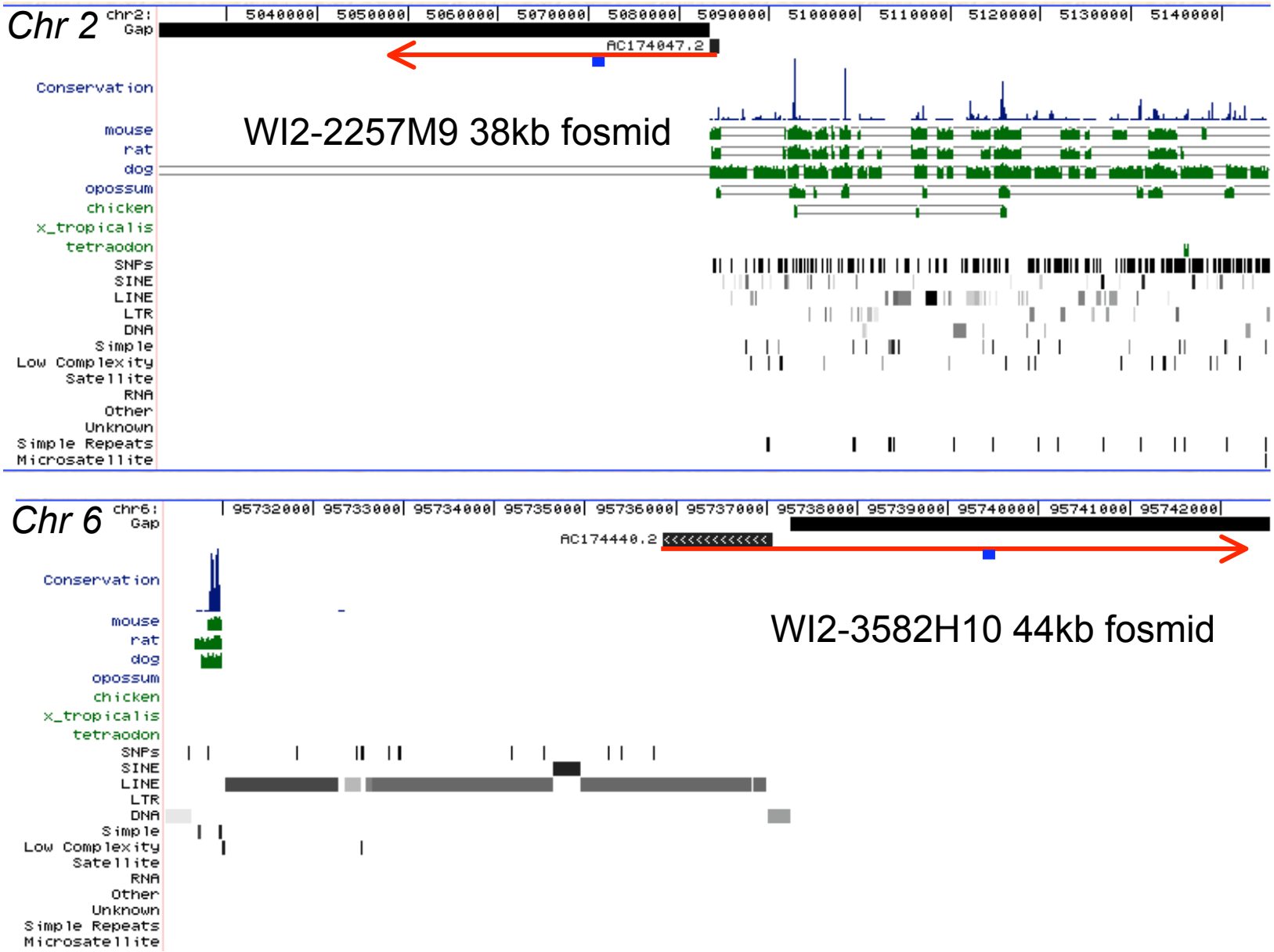
Nonhomologous end joining

...CACAGT-GCAGAACATTTGGAGCTGCTCCATGC**AG**catccctgggatgcaaggctggttcaac
...CACAGT-GCAGAACATTTGGAGCTGCTCCATGC**AG**catccctgggatgcaaggctggttcaacaca...
...CACAGTAGCAGAACATTTGGAGCTGCTCCATGC**AG**catccctggg
...CACAGT-GCAGAACATTTGGAGCTGCTCCATGC**AG**catccctggg
...CACAGT-GCAGAACATTTGGAGCTGCTCCATGC**AG**catccctgggatgcaaggctggttcaacaca...
TGC**AG**catccctgggatgcaaggctggttcaacaca...
| <---L1P4 element -->

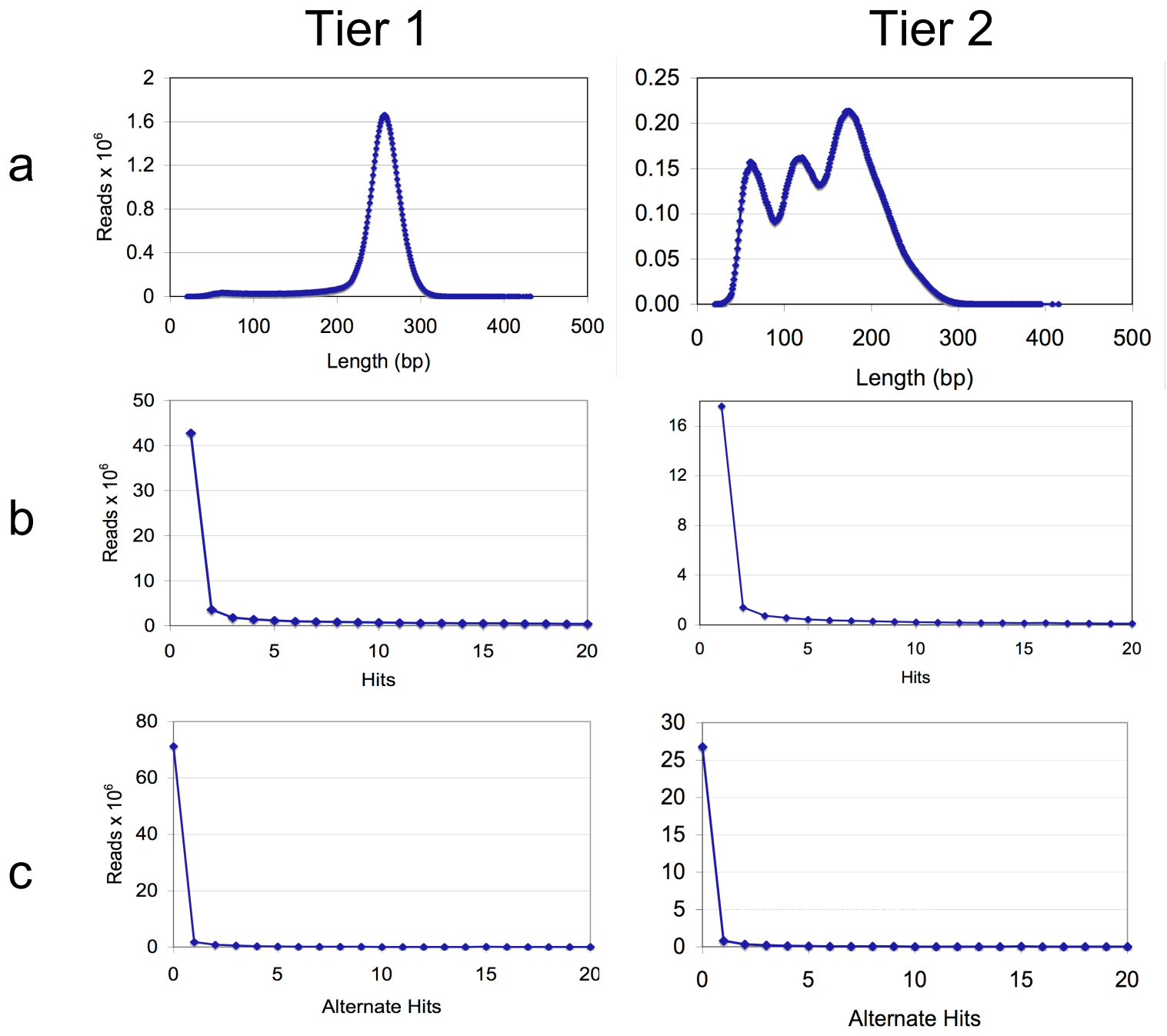
Supplementary Figure 7.



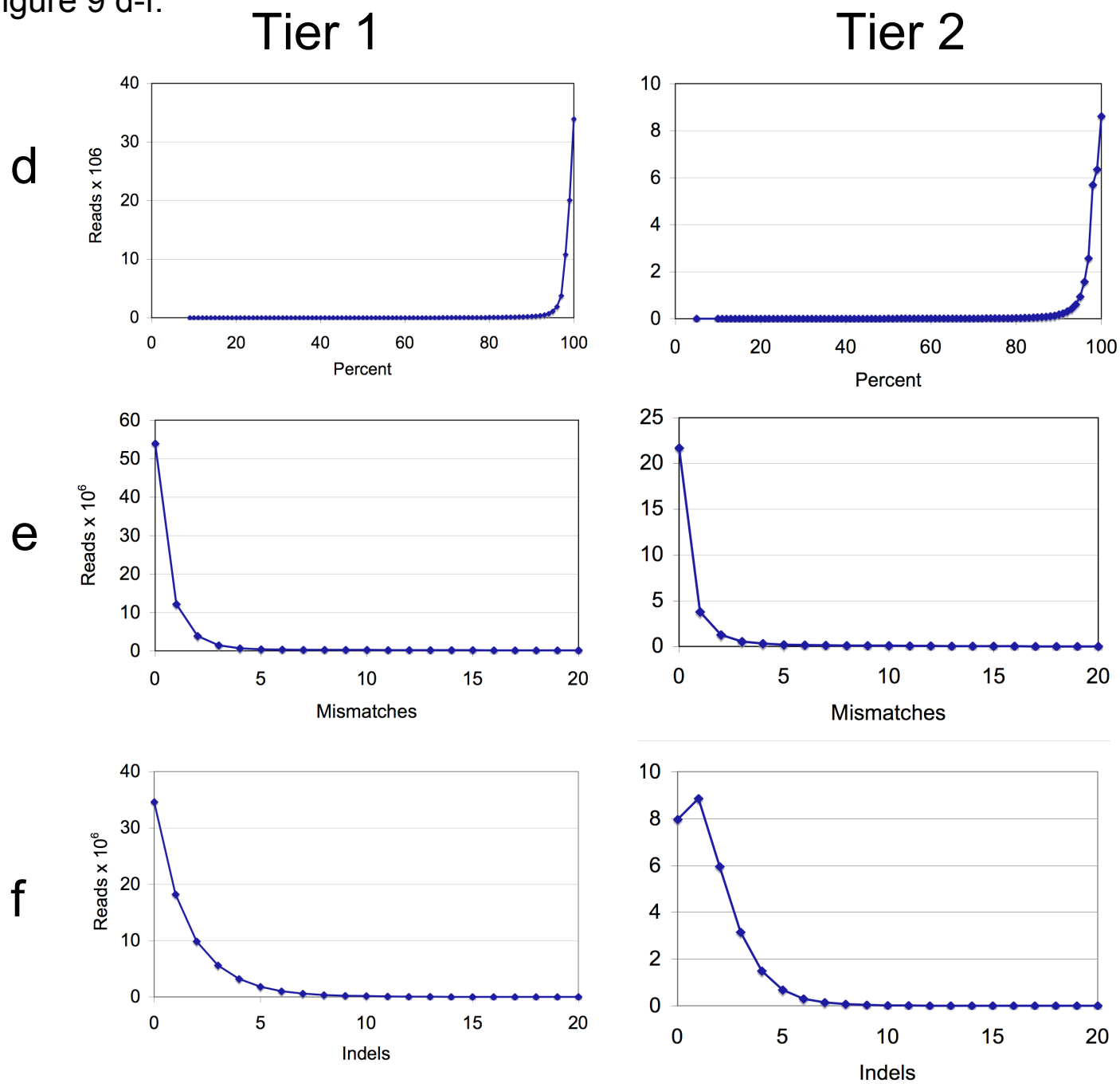
Supplementary Figure 8.



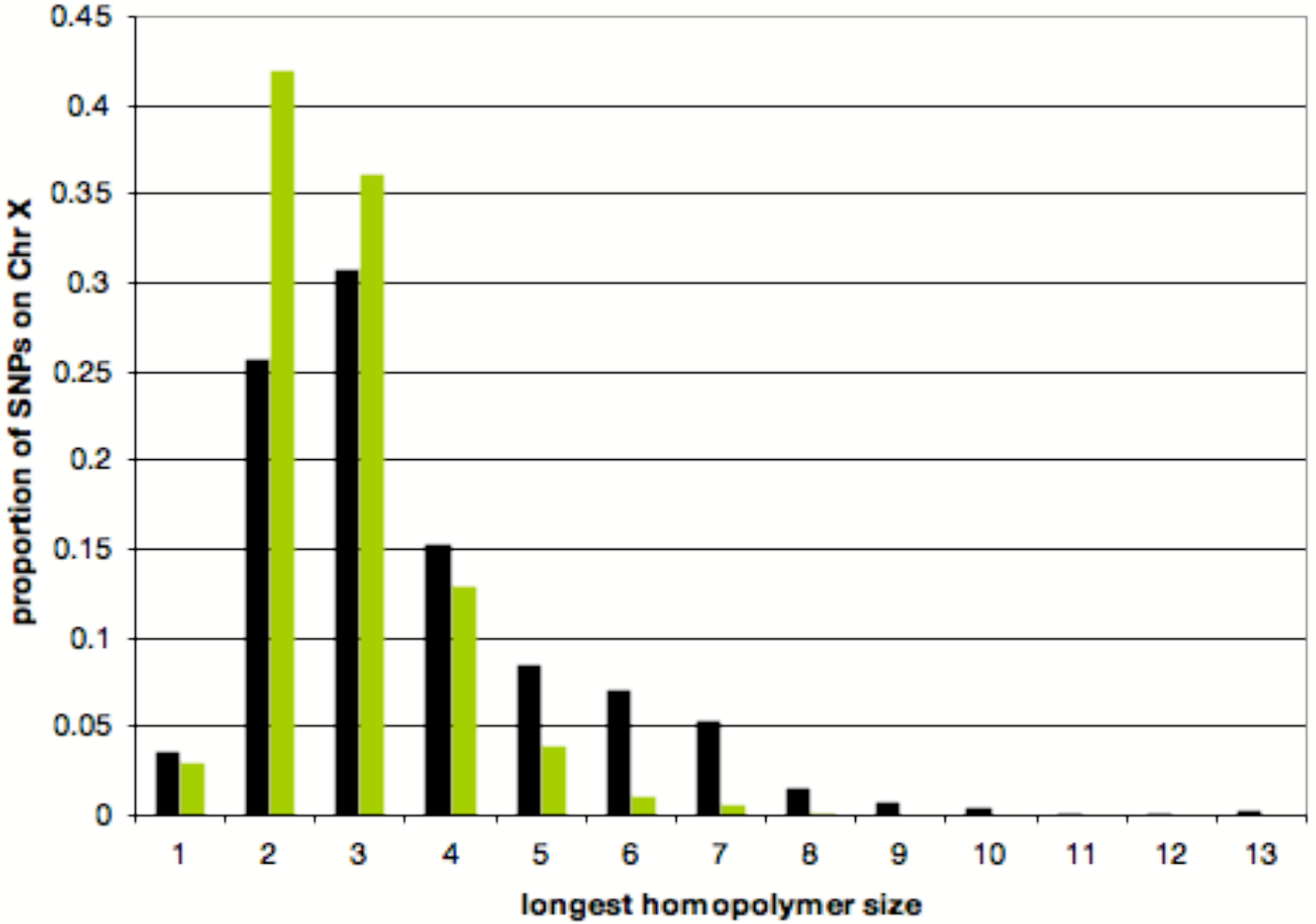
Supplementary Figure 9 a-c.



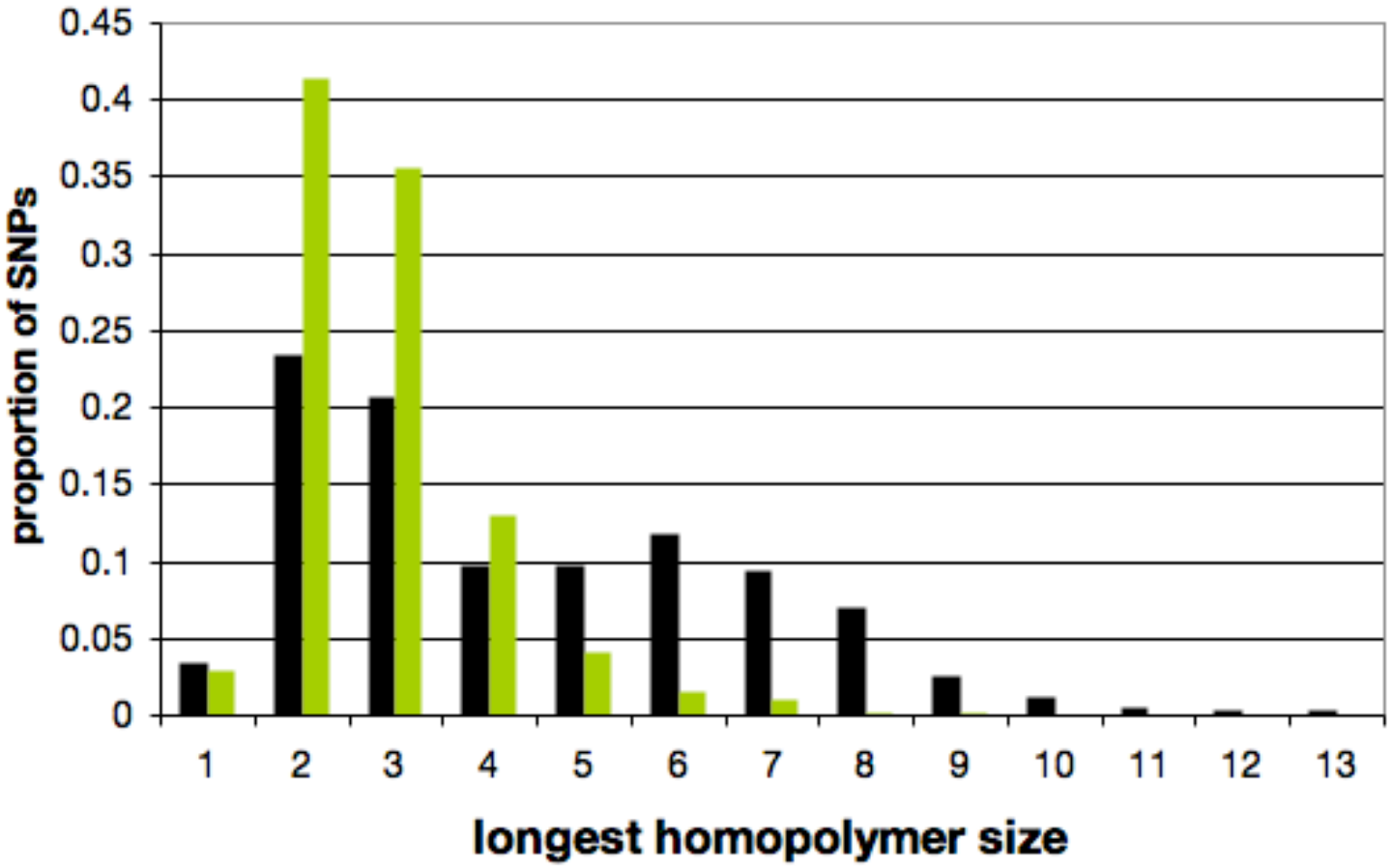
Supplementary Figure 9 d-f.



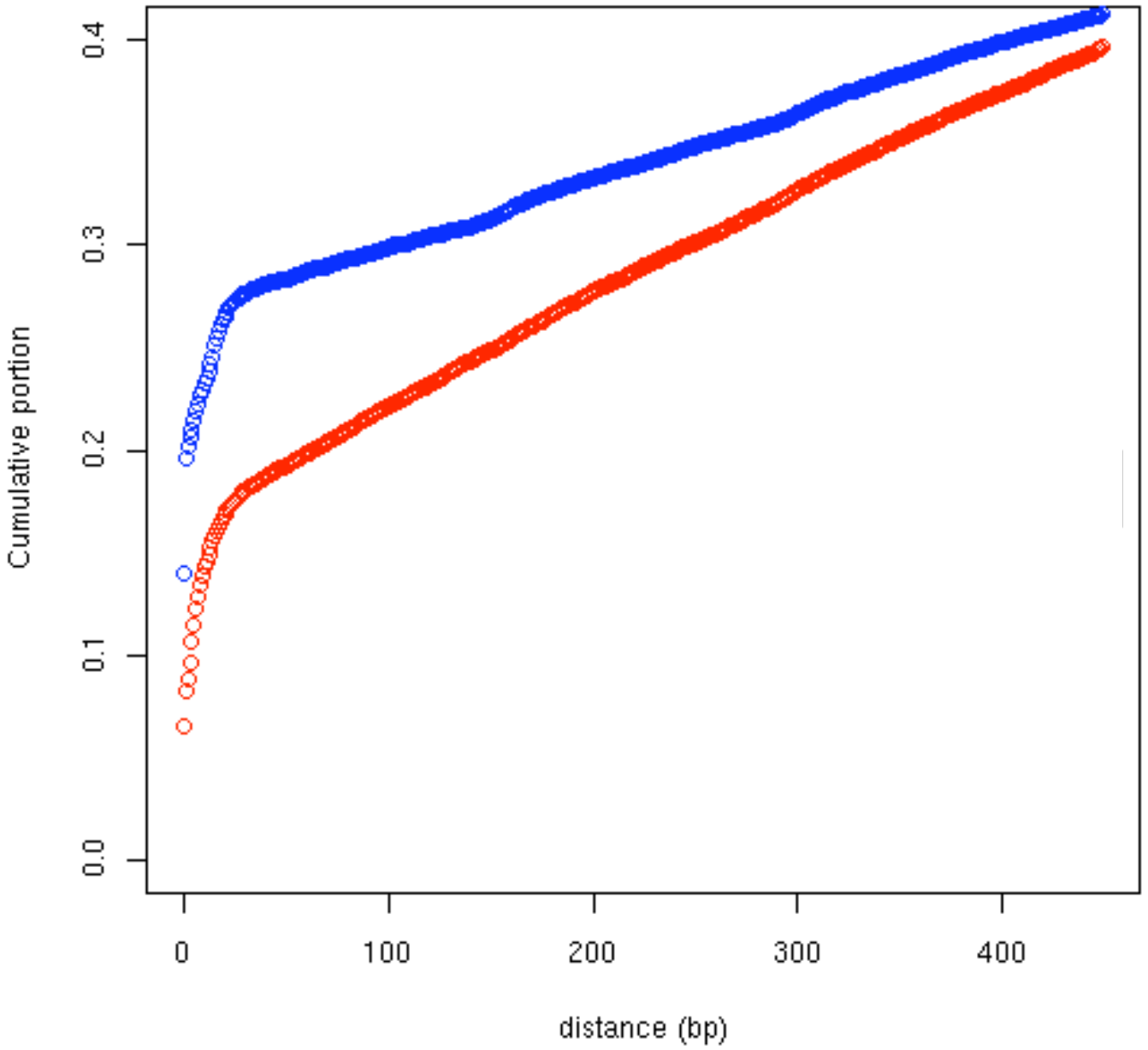
Supplementary Figure 10.



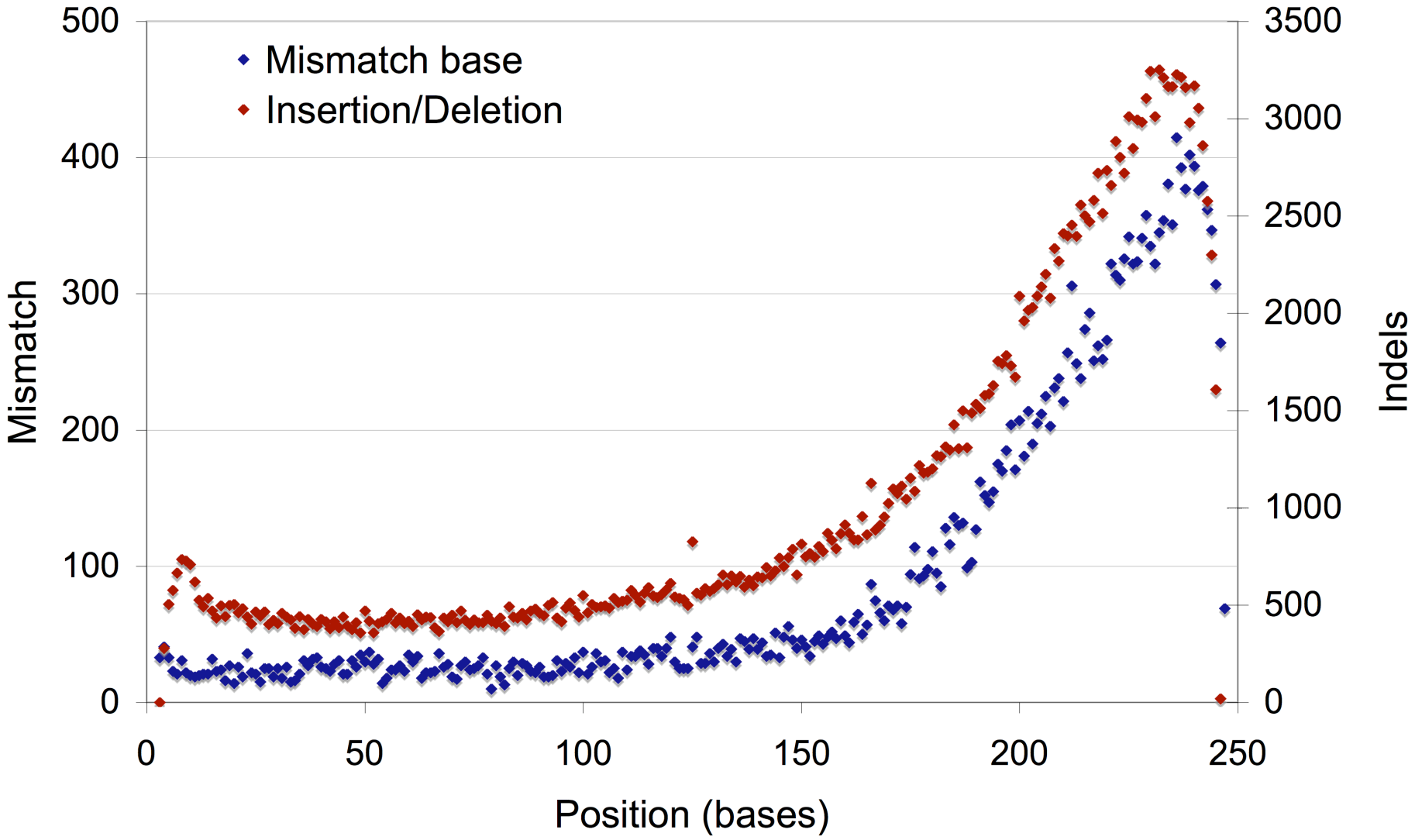
Supplementary Figure 11.



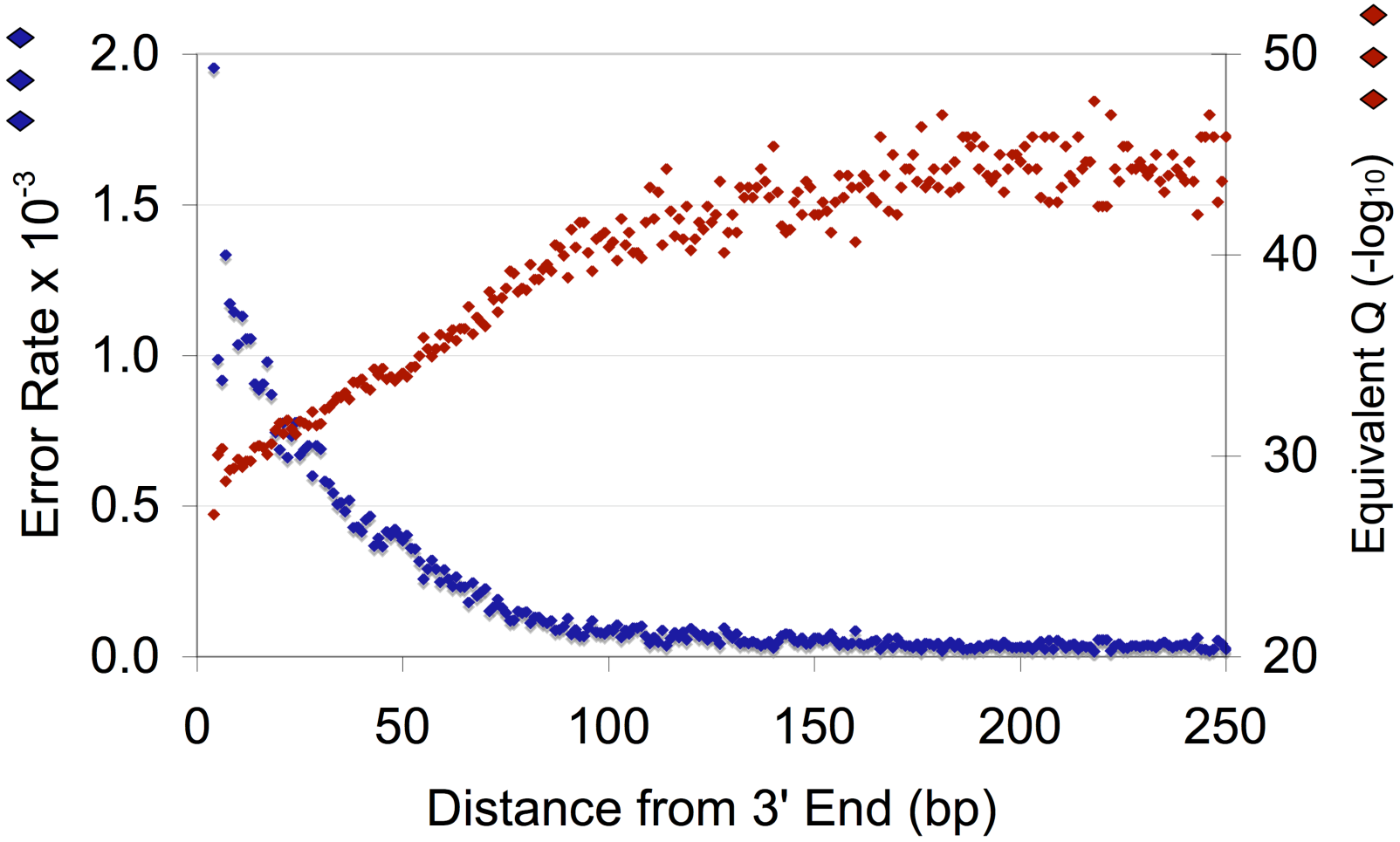
Supplementary Figure 12.



Supplementary Figure 13.



Supplementary Figure 14.



Supplementary Figure 15.