www.sciencemag.org/cgi/content/full/1149504/DC1

# Supporting Online Material for

## Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome

Jan O. Korbel,[1,2]* Alexander Eckehart Urban,[3]* Jason P. Affourtit,[4]* Brian Godwin,[4] Fabian Grubert,[5] Jan Fredrik Simons,[4] Philip M. Kim,[1] Dean Palejev,[5] Nicholas J. Carriero,[6] Lei Du,[4] Bruce E. Taillon,[4] Zhoutao Chen,[4] Andrea Tanzer,[7,8,9] A. C. Eugenia Saunders,[3] Jianxiang Chi,[10] Fengtang Yang,[10] Nigel P. Carter,[10] Matthew E. Hurles,[10] Sherman M. Weissman,[5] Timothy T. Harkins,[11] Mark B. Gerstein,[1,6,12] Michael Egholm,[4]† Michael Snyder[1,3]†

*To whom correspondence should be addressed. E-mail: megholm@454.com (M.E.) or michael.snyder@yale.edu (M.S.)

Published 28 September 2007 on *Science* Express
DOI: 10.1126/science.1149504

**This PDF file includes**



**Other Supporting Online Material for this manuscript includes the following:**
(available at www.sciencemag.org/cgi/content/full/1149504/DC1)

**Table S1.** List of predicted SVs, available as a separate Excel file.

# Supporting Online Material

## Materials and Methods

**Initial sample preparation and sequencing of paired-ends.** SVs were analyzed in DNA from cell lines of individuals NA15510 and NA18505; these cell lines have been studied in one or more previous SV/CNV studies [e.g. (*S1, S2*)]. We expect that the rate of genomic changes in these cell lines is low for two reasons: (*i*) SVs detected by our approach are frequently shared across individuals. (*ii*) It was recently estimated that less than 0.5% of deletion events in cell line DNA from the HapMap study (*S3*) collection (NA18505 is from this set) are due to somatic changes (*S1*).

Paired end sequences were determined with the following steps: (*i*) 5 micrograms of intact genomic DNA was hydrodynamically sheared (HydroShear - Genomic Solutions, Ann Arbor, MI) and purified with AMPure™ SPRI beads (Agencourt, Beverly, MA) to yield DNA fragments of the desired size (~3kb); (*ii*) after protection of EcoRI restriction enzyme cleavage sites through methylation, a biotinylated hairpin linker was ligated to the ends of the genomic DNA fragment; *(iii)* the fragments were digested with EcoRI and subsequently circularized by ligation of compatible adaptor ends, *(iv)* the circularized DNA was randomly fragmented by nebulization, and *(v)* DNA fragments containing paired ends were isolated by streptavidin-affinity purification with the biotinylated linker. These steps were followed by *(vi)* ligation of adaptors providing for subsequent amplification to increase library yield, and *(vii)* subsequent 454 Sequencing (454 Life Sciences/Roche Diagnostics, Branford, CT). We estimated the rate at which chimerical constructs (artifacts from the ligation reaction) are formed to be <2% on the basis of BLAST (*S4*) based sequence alignments.

**Computational mapping of paired-ends to the reference genome.** A computational analysis pipeline for massive data processing (run over 200,000 cpu hours on up to 440 processors; this included parameter optimization) was developed to map and compare paired-end reads to the human reference genome (assembly from March 2006; National Center for Biotechnology Information (NCBI) build 36). Overall, sequenced reads had a median length of 265 bp (mean=258 bp), spanned the 44 bp 'linker'-sequence in 65% of the reads, and yielded a median tag size (i.e., *end*) of 106 bp (mean=109 bp; standard

deviation=54.8 bp) on either side of the linker. After recognition and removal of the linker sequence 'GTTGGAACCGAAAGGGTTTGAATTCAAACCCTTTCGGTTCCAAC' (aligned at minimum sequence identity of 90%) the end sequences were separately matched to the genome. First, ends were aligned to the reference genome with Megablast (*S5*) (parameters: '-p 80 –s 11 -W 11'). Each reported genomic target region was subsequently extended by 250 bp on either side, and ends optimally realigned to the respective hit region with highly sensitive, albeit much slower, Smith-Waterman sequence alignment (*S6*). 78% of paired ends (fragments with recognizable linker sequence) passed the initial alignment procedures and yielded at least one Megablast hit to the genome for each end. We then analyzed the distribution of paired-end spans through mapping paired-ends onto the reference genome, with best-hits (**Fig. 1A**). Cutoffs were defined to distinguish paired-end spans falling into the usually observed, expected range (i.e. concordant paired-ends) from discordant paired-ends which were used as indicators for deletions and simple insertions (the latter spans fall into the tails of the distribution, i.e., beyond the defined cutoffs). In order to optimally represent the size distribution we derived cutoffs empirically for each experimental batch by (*i*) first removing all paired-end spans >10 kb from the list (a span which we considered to occur by chance only in the case of SV, or chimera formation during circularization), and (*ii*) determining the upper cutoff (*D*) and lower cutoff (*I*) to be the 0.00135 quantiles. Assuming that our data were normally distributed, these quantiles corresponded to approximately 3 standard deviations from the mean (*S2*) . Mean cutoffs across experimental batches were *I*=741 bp and *D*=6810 bp for NA18505, and 633 bp and 6482 bp for NA15510. (After removing the left tail of the distribution (values <500bp), we also modeled the log of the resulting distribution as a normal mixture, yielding very similar cutoff points.) The average span of paired-ends was 3083 bp (NA1850) and 3064 bp (NA15510).

**Computational prediction of SVs.** We further developed an algorithm for calling and fine-mapping SV. For each pair of ends with matching regions in the reference genome, we initially discarded all but the 30 best-scoring hits to the genome, and subsequently determined the best placement: i.e., the end sequence matches were combined with the

goal to identify most plausible paired-end alignments with an optimized form of the placement algorithm as described (*S2*) . Specifically, we awarded ends with highest sequence identity when aligned to the reference genome (+1), and with the longest sequence alignment (+1). Furthermore, scores were awarded for ends mapping to the genome with allelic levels of sequence identity (≥99.5%; +1). Finally, to avoid SV misassignment because of closely related sequences in the genome, we penalized cases in which end matches '*A*' had close (but not identical) matches '*B*' in the reference genome that when combined to a paired-end resulted in a concordant pair (-2). For this purpose we assumed match '*A*' is *not* nearly identical with match '*B*', if "length of (match) '*A*'" – 2 ≥"length of (match) '*B*'" and "sequence identity '*A*'" – 2% ≥ "sequence identity '*B*'". We further penalized pairs for which ends matched in different orientation (-2). With this procedure, we determined best placements for 63% of all paired-ends with recognizable linker sequence (8,549,989 for NA18505 and 4,224,311 for NA15510), out of which 80% yielded ≥97% sequence identity matches (for both ends) when aligned with the reference genome.

The best placements of paired-ends were used for identifying several different categories of SV: (*i*) deletions (size $s_d \geq 3$ kb) were identified from two or more overlapping discordant paired-ends with paired-end span >cutoff *D* (with the condition that both putative breakpoints are spanned); (*ii*) simple insertions (3 kb > $s_{si}$ > 2 kb) were identified from two or more overlapping discordant paired-ends with paired-end span <cutoff *I*; (*iii*) mated insertions were identified from two unpaired SVs that lie in nearby (i.e. 6 kb) genomic regions and had ≥2 paired-ends linking to a common, distant genomic region <100 kb (see **Fig. 1B**; we are thus most confident in size assignments of SVs <100 kb); mated insertions may involve tandem duplications or events related to transpositions. (*iv*) Inversions were called when ≥2 paired-ends matched different strands (consistent with an inversion). (*v*) Unmated insertions were predicted from ≥2 paired-ends that support a rearrangement of a genomic region in which loci change relative order without changing the relative orientation (i.e., the strand). (These events are similar to mated insertions; however, unmated insertions have only one assigned breakpoint.) In each case we required at least two paired-ends to support a predicted SV. Furthermore, at least one paired-end had to match the human reference genome at sequence identity

≥97% (high-stringency match, assessed for both ends. In addition, ends were required to yield best-scoring sequence alignments genome-wide to their respective region as assessed by Blat (*S7*). We also performed the following redundancy filtering steps: (i) in 5 instances a mated insertion and a simple insertion evidently corresponded to the same SV event (the 5 simple insertions were removed); (ii) in 6 instances a breakpoint of an unmated insertion was consistent with one of the breakpoints of a simple insertion event (the 6 unmated insertions were removed). Alignment qualities were nearly identical for normally mapped pairs and those that detected different SVs, e.g. for both NA15510 and NA18505 the respective sequence identities of matches to the reference genome differed by less than 0.5% on average for high-stringency matches; the remaining minor difference is likely due to a slight increase in variation in regions affected by SV. Lastly, in samples sequenced at (nearly) full coverage, regions where discordant paired-ends overlapped with concordant ones could have been used as evidence for heterozygosity; this analysis predicted heterozygous events for 80% of NA18505 SVs and thus 20% homozygous events.

Overall our analysis identified >400 SVs in NA15510 and >800 in NA18505. In the early stages of our work, a few additional SVs were identified with less stringent *PEM* scoring criteria. We have included a sequence-confirmed SV and a FISH confirmed SV in Table S1; the former case was included in our breakpoint analysis. Both cases were not used in our other analyses.

**Array Comparative Genome Hybridization (array-CGH).** A set of 8 oligonucleotide microarray chips was synthesized and hybridized by NimbleGen (NimbleGen Systems Inc., Madison, WI) to test for CNVs genome wide at moderate resolution. Each chip contained 385,000 oligonucleotides of length 50-75 b covering the genomic sequence approximately uniformly, with most repetitive regions (such as repeat-masked regions; www.repeatmasker.org) under-represented. The arrays were probed with fluorescently labeled genomic DNA from NA15510 (Cy5) and NA18505 (Cy3) and normalized with Bioconductor (www.bioconductor.org) as described (*S8*). *PEM*-identified SVs were considered to be validated by array-CGH with the following criteria: SVs which *PEM* predicted to be shared among NA15510 and NA18505 were initially excluded.

Furthermore, we excluded SVs covered by less than 10 or more than 1000 microarray probes, as these measurements were more likely to be affected by statistical and experimental bias. With these criteria, 31 (65%) out of 48 *PEM* predicted deletions in NA15510 showed signals with significant *P*-values ($P<0.05$; Mann-Whitney U test (*S9*); 29 of these passed the very stringent Bonferroni correction). Furthermore, when the analysis was restricted to SVs <100kb, for which size assignment is most reliable, 25 (69%) out of the remaining 36 regions were successfully validated (in 23 of these, *P*-values were robust to Bonferroni correction). An even more stringent protocol that also eliminates SV indels with partially intersecting genomic coordinates (if the latter are shared between NA15510 and NA18505) revealed an even higher validation success rate of 78% (i.e. 14 out of the 18 remaining regions tested validated; 13 of these were robust to Bonferroni correction). (Note that we could not use array-CGH to validate SV indels in NA18505, as we currently did not sequence NA15510 deeply enough and thus are missing many shared SVs.)

**PCR analysis**. PCR primers were designed for predicted SVs with Primer3 (*S10*) (parameters: Tm=65ºC; Tmin=62 ºC; Tmax=68 ºC; optimum length=25bp; min-length=22bp; max-length=30bp; primers matching to a human repeat-library (human_mispriming_lib), available from the Primer3 website, were excluded) to generate amplicons spanning the breakpoint-junction-sequences of predicted structural variants. PCR was carried out with JumpStart™ REDAccuTaq® LA DNA Polymerase (Sigma-Aldrich Inc., St. Louis, MO) on PTC-225 DNA Engine Tetrad™ Cycler (Bio-Rad, formerly MJ Research, Hercules, CA) in a 25 µl or 50 µl reaction volume and with 10 or 20 ng of genomic DNA as template. The following program was used: Initial denaturation at 94°C for 30 sec, followed by a 3-Step-Touchdown: 1. (94°C 5 sec, 68°C 30 sec, 68°C 6 min), 2. (94°C 5 sec, 66°C 30 sec, 68°C 6 min), 3. (94°C 5 sec, 64°C 30 sec, 68°C 6 min); followed by an additional cycle of 68°C 30 min. Fragments up to 8 kb in size were visualized by gel electrophoresis and scored.

When initially estimating the overall validation rate for PCR, we tested 40 randomly picked SVs for which at least one and up to 5 primers were designed according to the *Primer3* (*S10*) parameter settings indicated above. Of the 40 SVs that were tested,

33 yielded a single, clear PCR band at the expected size range in at least 1 reaction (scored as positive), one did not yield any band (scored as negative), and 6 yielded smears or multiple bands (probably because the SVs were located in repetitive regions) and were thus regarded as non-interpretable.

To identify SVs likely to be heterozygous PCR was also used to identify sequence present in the reference genome that is also predicted to be disrupted by a SV. We focused on SVs with identified breakpoint junctions (Table S1), and tested for the presence of the reference allele in 41 SVs from NA18505 and 30 SVs from NA15510. 5 primer pairs were tested for each SV, we inferred homozygosity in 15% of SVs in NA18505 (6 out of 41), and 23% in NA15510 (7 out of 30), all of which never revealed bands indicative of heterozygosity. Those numbers are in close agreement with the computational analyses presented above and measures in (*S2*).

In order to increase the number of validated SVs we further carried out >800 PCR experiments in a one-pass fashion, i.e. with only one primer pair per predicted SV: in 58% of the experiments the SV was validated (6% could not be scored, mostly due to smears, and were not considered for calculating the success rate). It is widely assumed that SVs/CNVs are inherited in a Mendelian fashion [e.g. (*S11, S1*)]. We therefore analyzed Mendelian patterns of inheritance for 5 *PEM*-identified SVs that could be monitored in PCR reactions enabling detection of both the SV and reference alleles simultaneously: 9 meioses (covering 5 SVs) were analyzed in parent-offspring trios [with individuals of an African family (Y005) and members of a European family (CEPH/UTAH pedigree 1420)]; in all cases the observed band patterns were consistent with Mendelian segregation.

**Sequencing of breakpoints**. In order to sequence breakpoint junctions, PCR fragments were extracted either by gel-purification or gel-extraction with Millipore Ultrafree®-DA centrifugal filter devices (Millipore Corp., Bedford, MA) or by bead-purification from the reaction mixture with Agencourt® AMPure® (Agencourt Biocience Corporation, Beverly, MA). Amplified fragment pools (50 – 150 fragments each) were randomly sheared by nebulization, converted to blunt-ends, and adaptors ligated with the GS DNA Library Preparation kit according to the manufacturer's protocols (454 Life Sciences,

Branford, CT; Roche Diagnostics, Indianapolis, IN). The resulting single stranded DNA shotgun libraries were then sequenced with 454 Sequencing. Both the resulting reads (median length=250bp) and contigs generated by 454's *de novo* assembler Newbler (see software user manual, 454 Life Sciences and Roche Diagnostics) were scanned for the respective SV-breakpoints with BLAST (*S4*) alignment against the human reference genome; we required best-hits to the genome for both portions of a read/contig matching on either side of a candidate breakpoint junction. Alternatively, if unassembled reads were used for breakpoint identification, we required at least two reads to support a breakpoint.

To assess the quality of the breakpoint calls we initially sequenced the breakpoint junctions of 14 randomly chosen SVs also represented in the Celera assembly (R27c), requiring the absence of ambiguous base calls (represented as N's) within 500 bp of the breakpoint observed in the Celera assembly. In all 14 cases the SV breakpoint junction obtained by us matched to the same genomic site evident from the Celera sequence. Minor differences were sometimes observed (typically 2 bp or less) which were attributable to SNPs, low complexity sequences and microhomologies at the junctions; such minor differences do not affect our breakpoints classification described in the text.

We finally compared SV breakpoint coordinates obtained from sequencing and assembly comparison to the PEM-predicted breakpoint coordinates (making use of the fact that partially overlapping paired-ends usually improve the resolution of provisional breakpoint assignments by PEM) and determined a mean resolution of 644 bp for initial breakpoint calls.

**Fiber-FISH**. DNA from cell lines of NA15510 and NA18505 grown in RPMI1640 medium enriched with 15% fetal calf serum was used to prepare extended chromatin fibers. Approximately 2-3 ml of cell suspension was centrifuged at 1200 rpm for 5 min. The cell pellets were washed twice with PBS and diluted to a final concentration of approximately 2-3 × 106 /ml. 10 µl cell suspension were spread over a 1 $cm^2$ area on the upper part of a polylysine-coated slide (Sigma) and left to dry at room temperature for approximately 30 min. The air-dried slides were then fitted into a Cadenza coverslip and clamped in a nearly vertical position with a bent metal rack. 150 µl of freshly made lysis

solution (containing 5 parts 70mM NaOH, 2 parts absolute ethanol) was applied to the gap at the top of the microscopic slide and Cadenza coverslip assembly. As soon as the lysis solution level dropped below the frosted edge of the microscopic slide, 150 µl of 96% ethanol was added. The slide was allowed to drain until the meniscus stopped falling (approximately 30 s) and the slide was carefully lifted off by pulling its top back from the Cadenza coverslip. The slides were then air-dried and treated with a 3:1 acetic acid/ethanol fixative for 5 min, then dehydrated in an ethanol series (70%, 90%, 100%). Finally, the slides were treated with 0.01% pepsin (Sigma) at 37 ºC for 5 min and dehydrated in the ethanol series again. DNA from the fosmid clones selected for fiber-FISH were labeled with either Digoxigenin-11-dUTP or Fluorescein-12-dUTP (Roche) with the WGA2 Kit (Sigma). For hybridization approximately 100 ng of each digoxigenin- and fluorescein-labeled probes were used. FISH was carried out following the previously published protocols (*S12, S13*). Digoxigenin-labeled probes were visualized by monoclonal mouse anti-dig antibody (Sigma) and Texas Red-conjugated goat anti-mouse IgG (Invitrogen, Carlsbad, CA). Fluorescein labeled probe was detected with Alexa 488-conjugated rabbit anti- fluorescein IgG and Alexa 488-conjugated donkey anti-rabbit IgG (Invitrogen). After detection, slides were mounted with mounting solution containing 4′,6-diamidino-2-phenylindole (DAPI, Vector Labs, Orton Southgate, UK). Images were captured and processed with the SmartCapture software (Digital Scientific, Cambridge, UK). Four inversions were analyzed and three confirmed (one of the latter was confirmed for both NA15510 and NA18505). The fourth was predicted to be very small (4 kb in size) and could not be definitively determined.

**Comparison of SVs to the Celera assembly.** A number of predicted SVs (deletions, insertions and inversions) were confirmed by comparing the respective region of interest to the Celera assembly (R27c) of the human genome. For each predicted SV, 500bp fragments flanking the predicted breakpoints were extracted from the human reference genome assembly (ncbi36) and concatenated. The combined 1000 bp fragment was then searched against the Celera assembly with Blat (*S7*). Non-overlapping best matches to the Celera assembly were parsed with custom Perl scripts (www.perl.org; available upon request from the authors) followed by manual analysis, and the *PEM* identified SV

assumed as confirmed if supported by the span and orientation of the parsed best-scoring Blat matches. We automatically removed many instances where sequences matched imperfectly (i.e., 236 instances, in which less than 90% of the 1000 bp sequence matched the Celera assembly), and 17 instances where parts of the respective region of interest in the Celera genome were annotated as gaps. This reduced the number of possible validations.

**Comparison of identified SVs across samples and surveys**. The overlap of SVs identified by *PEM* and CNVs reported in the Database of Genomic Variants (DGV) and in the recent large-scale analysis of CNVs carried out by Redon *et al.* (*S1*) was calculated by intersecting SV indels identified by *PEM* to previously reported CNVs with the available SV and CNV coordinates. We compared our SVs with CNVs/SVs represented in DGV (*i*) using all *PEM* SVs, and also (*ii*) focusing on *PEM* SVs at the size range 50-500kb. When addressing the overlap of *PEM* SVs and Tuzun *et al.* (*S2*) SVs, we applied more stringent criteria: i.e. we identified instances where paired ends supporting a predicted *PEM* SV spanned the corresponding region of an SV predicted in Tuzun *et al*. (*S2*) at the same locus, with the same size, and SV-type – taking into account a conservative estimate for the expected resolution of both approaches in determining breakpoint junctions; i.e. 3kb for *PEM*, 40kb for fosmid-paired-end sequencing (while deletions of 8 kb are detected by the latter method as recently reported (*S2*), the approximate breakpoint precision [term defined in (*S14*)] is up to 40 kb both for identifying deletions (*S8*) and inversions). Note that for technical reasons, the size ranges in which simple insertions are identified do not overlap between approaches, and thus insertions were not included when comparing *PEM* and fosmid-paired end sequencing. SVs shared between NA18505 and NA15510 were determined with a similar (stringent) approach: i.e. we identified instances where paired-ends supporting a SV in NA15510 spanned a predicted SV (same locus, size, and type) in NA18505, and extrapolated with the expected coverage for NA18505 (93% of SV events) to estimate the fraction of SVs shared between NA15510 and NA18505.

**Gene Ontology Analysis.** We used GOToolBox (http://crfb.univ-mrs.fr/GOToolBox/index.php) to evaluate the enrichment/depletion of gene functional categories among protein coding genes intersecting with SVs. In particular, we analyzed genes intersecting with SVs by their "Gene Ontology (GO) functional classes" (GO Biological Processes; www.geneontology.org), and after correcting for multiple testing (Bonferroni correction; hypergeometric test) we found several significant relationships when analyzing a broad/inclusive GO category (i.e. GO terms at 'annotation level' 3), consistent with previous findings (*S9, S15-18, S1, S19, S20, S2*). In particular, genes involved in organismal physiological processes (including, e.g., immunity, and cell-cell signaling) are enriched amongst genes associating with SVs, while genes involved in cellular physiological processes (such as, cell metabolism) are depleted. Furthermore, when analyzing annotations more specifically (i.e., using GO level 6), we found proteins that are likely to be involved in interactions with the environment such as those involved in immune response (*P*=9e-18), sensory perception of smell (*P*=0.001), and sensory perception of chemical stimuli (*P*=0.003) are frequently affected by SVs. Retrovirus and transposition related proteins were also suggested to be affected by SV (both are combined in GO term '0006313'; level 8; *P*=6e-11); this may be due to their role in the formation of many SV events.

**Breakpoint analysis: association with various repeat elements**.

We initially analyzed the genomic elements that intersect with breakpoint junctions. In particular, we analyzed segmental duplications (SDs; obtained from the UCSC genome browser (http://genome.ucsc.edu; data set termed: (hg18.genomicSuperDups))) and medium- to high-frequency repetitive elements identified by repeatmasker (obtained from the UCSC genome browser): LTRs, L1/LINEs, L2/LINEs, *Alu*/SINEs. To test for significant enrichment (or depletion), we calculated approximate *P*-values by carrying out permutation tests using 10,000 randomized trials. For evaluating enrichment of repeat elements, the locations of SVs with inferred breakpoint junctions were randomized (global enrichment analysis; genomic locations were randomly picked from ascertainable regions of the genome, defined as genomic positions spanned by a paired-end (i.e. *best placement*); during the randomization process SV sizes were kept unchanged). Using this

protocol we observed both SDs and L1 elements to be significantly enriched (SD: 2.6-fold, $P<0.0001$ from permutations; L1: 1.25-fold; $P<0.01$), and found L2 elements to be significantly depleted (4.4-fold depletion, $P<0.0001$). To control for potential biases in sequence composition and context, we also determined local enrichment *P*-values by randomizing the SV locations in a window ±50 kb around their original location followed by evaluation of the overlap of repeat elements in the randomized location. After applying this correction, SDs and L1 elements were found to be only slightly enriched (SD: 1.3-fold, $P<0.05$; L1: 1.03-fold, non-significant); while L2 elements were still significantly depleted (3.8-fold depletion, $P<0.01$). *Alu* elements were neither found to be significantly enriched/depleted in the global nor the local enrichment analysis.

For the repeat analysis in **Fig. 2** (which uses chromosomal ideograms obtained with permission from the University of Washington Department of Medicine/Pathology; http://www.pathology.washington.edu), we analyzed the genomic features of breakpoint regions (i.e., the vicinity of predicted breakpoints) in 3 kb windows. In particular, we mapped SVs onto chromosomal bands (i.e. the ideograms) and analyzed the overlap of SVs with SDs (obtained from http://humanparalogy.gs.washington.edu) and medium- to high-frequency repetitive elements retrieved from the USCS Genome browser (http://genome.ucsc.edu): satellite repeats, LINEs, LTRs. For each SV in **Fig. 2** we indicate by color the most abundant element, determined from the total number of occurrences of repeat elements in the respective breakpoint regions. Note that the number of breakpoints and respective genomic contexts that were analyzed per event differ between event classes: e.g., three locations are relevant for a mated insertion, consisting of the target region in the reference genome in which a sequence presumably was inserted, as well as the start- and end-points of the sequence predicted to be inserted into the former region (which is usually located elsewhere in the genome); for a deletion event, both breakpoints flanking the deletion event in the reference genome were analyzed; for simple insertions, one breakpoint was analyzed (i.e. the target region where sequence presumably was inserted in); in the case of inversions and unmated insertions, both of the respective predicted breakpoints (expected within ~3kb of the matched ends, respectively) were analyzed.

Breakpoints junction sequences were analyzed both by computational analyses and by manual inspection, and the plausible mechanism of origin for >90% of the SVs was deduced. Repeat elements and other sequence features were retrieved from the USCS Genome browser (http://genome.ucsc.edu), and breakpoint junctions analyzed by manual inspection of BLAT and BLASTN sequence alignments (with ±100 bp of flanking sequence from both breakpoints). NHEJ was inferred when one or both breakpoints of an SV resided in unique sequence, or when repeat elements were present at the junctions but BLASTN analysis indicated no sequence identity at the breakpoint besides expected microhomologies of ~5 bp or less that are frequently associated with NHEJ (*S21*). Insertions of 1 up to several bases directly at the inferred junction, another well-known hallmark of NHEJ (*S21*), were frequently observed (Table S1). We also observed SVs that formed through NAHR; those had homologous sequences at both breakpoint junctions. The presence of extended regions of sequence similarity (≥50 bp) was confirmed in all these cases by both BLASTN and manual inspection. Retrotransposition events were inferred for SV indels by the presence of a L1 or SVA element. As expected, polyA stretches (*S22*) were present at the 3'-breakpoint junction of all suspected L1 and SVA retrotransposition events and elements were flanked by duplicated target DNA sequence (8-19bp). DNA transposition events were not observed (*S21, S23*), but one instance of a human endogenous retroviral insertion was evident in both the NA15510 sample and the Celera sequence. This element contained a 6 bp duplication of target sequence (**Fig. 5**) but lacked a polyA stretch, as expected (*S23*). The USCS Genome browser mammalian conservation track was inspected manually in order to support assignments of SV indels, in particular retrotransposition events, to the classes above (e.g., recent LINE insertions typically cause a break in the mammalian conservation track, as they are not present in other primate sequences).

**Relating sequencing coverage to the expected portion of SVs identified.** We used the Poisson approximation to the binomial distribution [i.e., expanding equations previously given in (*S24*)] for relating coverage of SV-identification to sequencing coverage. Note that we determined coverage on the basis of effectively matched paired-ends (i.e. such with best-placement; other reads were not considered for this calculation). Given a

sequencing coverage λ = total physical span of optimally placed paired ends falling into the usually observed, expected range of paired-ends / size of diploid euchromatic genome, and the number of observations *k*, the probability *P* of covering a certain genomic element *k* times is (*S24*):

$$P(k;\lambda) = \frac{e^{-\lambda}\lambda^{k}}{k!} \qquad \text{Eq. 1}$$

For simplicity we equated 'covering a genomic element' with 'detecting an SV'. Taking into account the requirement of evidence from ≥2 paired-ends per predicted SV, we initially calculated the probability of missing an SV as $P(k<2;\lambda) = P(0;\lambda) + P(1;\lambda)$. Thus, $P(k\geq 2;\lambda) = 1 - P(k<2;\lambda)$, and hence $P(k\geq 2;\lambda) = 1-(e^{-\lambda} + \lambda e^{-\lambda}) = 1-(1+\lambda)e^{-\lambda}$, we estimated that for sample NA18505, 93% of all SVs within the detection range of *PEM* are identified (λ = 4.3x coverage; i.e. $P(k\geq 2; 4.3) = 1-(1+4.3)\,e^{-4.3} = 0.93$). Furthermore, for NA15510 (λ = 2.1x), 62% of all SVs were expected to be identified by the approach.

**Detection of SNPs associated with breakpoints**. It has been found that a portion of CNVs are in linkage disequilibrium with SNPs catalogued by the international HapMap project (*S3*), and that SNPs may be used to reliably predict the presence of these nearby ('linked') CNV/SV by association (*S1*). Contigs assembled from sequenced amplicon pools can be mined for SNPs directly adjacent to an SV breakpoint. By alignment of sequences to the reference genome, we identified 344 putative SNPs that are within 3 kb of their respective breakpoints (**Table S4**); of these 183 (53%) had been described previously (http://www.ncbi.nlm.nih.gov/projects/SNP). SNPs derived by *PEM* may serve as useful predictors for nearby SVs.

In order to identify SNPs, sequence reads generated from amplicons were mapped against reference amplicons (i.e. regions near the predicted SVs) derived from human genome build 36 with the software 454 Mapper (see software user manual, 454 Life Sciences and Roche Diagnostics). The mapper aligns reads to its unique reference position and reports consensus sequence as well as variations. All homozygous and heterozygous variations (with a frequency cutoff of 40% or above) were considered candidate SNP positions and retained for further analysis. All candidate SNP positions

were mapped against dbSNP 126 (http://www.ncbi.nlm.nih.gov/projects/SNP) in order to identify potentially novel SNPs.

**Considerations for applying *PEM* in different genomic regions**. Highly repetitive regions and recent segmental duplications (SDs) may in some instances be problematical for a paired-end approach; repetitive regions are less likely to be identified through best placements of reads, which may limit the ability of *PEM* to identify SVs adjacent to or within repetitive elements/duplicated regions. However, the fact that a high fraction of SVs previously detected by fosmid paired end sequencing (*S2*) have been successfully identified by *PEM* (see main text) indicates that this effect is small. At the level of breakpoint junctions, repetitive and/or low-complexity sequences may hamper the sequencing and unambiguous assignment of fine-mapped breakpoints identified through *PEM*. Nevertheless, the fact that we infer similar portions of alternative SV formation events using (*i*) DNA sequence generated by 454 Sequencing and (*ii*) regions from the Celera assembly indicates that this bias should be relatively minor.

**Data retrieval**. Fine-mapped coordinates of SVs are available from **Table S1** and from the Database of Genomic Variants (http://projects.tcag.ca/variation); accession numbers are available from **Table S5**, and at http://sv.gersteinlab.org/.

**Supplemental References**


S1. R. Redon, et al. (2006) *Nature*, **444**, 444-54.
S2. E. Tuzun, et al. (2005) *Nat Genet*, **37**, 727-32.
S3. D. Altshuler, et al. (2005) *Nature*, **437**, 1299-320.
S4. S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990) *J Mol Biol*, **215**, 403-10.
S5. Z. Zhang, S. Schwartz, L. Wagner and W. Miller (2000) *J Comput Biol*, **7**, 203-14.
S6. T. F. Smith and M. S. Waterman (1981) *J Mol Biol*, **147**, 195-7.
S7. W. J. Kent (2002) *Genome Res*, **12**, 656-64.
S8. A. E. Urban, et al. (2006) *Proc Natl Acad Sci U S A*, **103**, 4534-9.
S9. D. F. Conrad, T. D. Andrews, N. P. Carter, M. E. Hurles and J. K. Pritchard (2006) *Nat Genet*, **38**, 75-81.
S10. S. *Rozen*, H. Skaletsky, in *Bioinformatics Methods and Protocols: Methods in Molecular Biology* K. S, M. S, Eds. (Humana Press, Totowa, NJ, 2000) pp. 365-386.
S11. T. L. Newman, et al. (2006) *Hum Mol Genet*, **15**, 1159-67.
S12. J. X. Chi, et al. (2005) *Chromosoma*, **114**, 167-72.
S13. S. M. Gribble, et al. (2004) *Chromosome Res*, **12**, 35-43.
S14. B. P. Coe, et al. (2007) *Genomics*, **89**, 647-53.
S15. D. A. Hinds, A. P. Kloek, M. Jen, X. Chen and K. A. Frazer (2006) *Nat Genet*, **38**, 82-5.
S16. A. J. Iafrate, et al. (2004) *Nat Genet*, **36**, 949-51.
S17. D. P. Locke, et al. (2006) *Am J Hum Genet*, **79**, 275-90.
S18. S. McCarroll, et al. (2006) *Nat Genet*, **38**, 86-92.
S19. J. Sebat, et al. (2004) *Science*, **305**, 525-8.
S20. A. J. Sharp, et al. (2005) *Am J Hum Genet*, **77**, 78-88.
S21. E. V. Linardopoulou, et al. (2005) *Nature*, **437**, 94-100.
S22. V. P. Belancio, M. Whelton and P. Deininger (2007) *Gene*, **390**, 98-107.
S23. R. E. Mills, E. A. Bennett, R. C. Iskow and S. E. Devine (2007) *Trends Genet*, **23**, 183-91.
S24. E. S. Lander and M. S. Waterman (1988) *Genomics*, **2**, 231-9.

## Supplemental Tables

**Table S1.** List of predicted SVs, available as a separate Excel file.

**Table S2.** Predicted SV classes - breakdown.

| Individual | Deletions | Insertions (simple/ mated/ unmated) | Insertion (simple) | Insertions (mated) | Insertions (unmated) | Inversions | Total SVs |
|---|---|---|---|---|---|---|---|
| NA15510 | 303 | 119 | 14 | 24 | 81 | 50 | 472 |
| NA18505 | 550 | 203 | 25 | 58 | 120 | 72 | 825 |
| Total | 853 | 322 | 39 | 82 | 201 | 122 | 1297 |

**Table S3. List of NA15510 internal IDs for SVs shared (in common) with NA18505**

| | | | | | |
|---|---|---|---|---|---|
| 3 | 202 | 318 | 451 | 592 | 859 |
| 4 | 203 | 320 | 465 | 598 | 860 |
| 6 | 204 | 327 | 467 | 602 | 861 |
| 8 | 209 | 334 | 468 | 606 | 862 |
| 9 | 211 | 335 | 474 | 612 | |
| 14 | 212 | 345 | 478 | 643 | |
| 26 | 215 | 347 | 485 | 678 | |
| 28 | 217 | 352 | 490 | 682 | |
| 31 | 219 | 362 | 494 | 691 | |
| 34 | 225 | 363 | 500 | 701 | |
| 35 | 232 | 364 | 506 | 703 | |
| 42 | 237 | 369 | 513 | 705 | |
| 43 | 241 | 370 | 515 | 714 | |
| 47 | 243 | 371 | 517 | 732 | |
| 56 | 253 | 373 | 523 | 735 | |
| 67 | 254 | 378 | 524 | 751 | |
| 76 | 255 | 380 | 527 | 760 | |
| 89 | 257 | 381 | 528 | 811 | |
| 93 | 258 | 382 | 530 | 831 | |
| 99 | 260 | 386 | 534 | 832 | |
| 102 | 261 | 387 | 535 | 839 | |
| 103 | 267 | 389 | 536 | 840 | |
| 108 | 269 | 393 | 539 | 841 | |
| 110 | 275 | 394 | 541 | 842 | |
| 118 | 277 | 397 | 544 | 843 | |
| 135 | 281 | 399 | 545 | 844 | |
| 138 | 282 | 401 | 550 | 845 | |
| 142 | 283 | 403 | 556 | 846 | |
| 148 | 286 | 408 | 559 | 847 | |
| 160 | 289 | 409 | 561 | 848 | |
| 169 | 290 | 415 | 564 | 849 | |
| 174 | 292 | 425 | 565 | 850 | |
| 181 | 297 | 427 | 568 | 851 | |
| 182 | 298 | 432 | 572 | 852 | |
| 183 | 302 | 434 | 576 | 853 | |
| 191 | 306 | 435 | 578 | 854 | |
| 194 | 311 | 437 | 586 | 856 | |
| 199 | 313 | 439 | 587 | 857 | |
| 200 | 316 | 443 | 590 | 858 | |

**Table S4**. SNP calls

| Chr | Coordinate | Base found | Base in b36 | SNP rsID | SNP genotypes |
|---|---|---|---|---|---|
| chr21 | 10130371 | C | T | rs4913777 | C/T |
| chr7 | 6891005 | A | G | rs4720693 | A/G |
| chr2 | 165726463 | C | T | rs4667789 | C/T |
| chr2 | 165726592 | A | T | rs4667791 | A/T |
| chr2 | 165726600 | C | T | rs4667792 | C/T |
| chr1 | 208794218 | C | T | rs4845052 | C/T |
| chr10 | 5274966 | A | C | rs4880720 | A/C |
| chr6 | 57539011 | T | A | rs5007797 | T/A |
| chr8 | 6978982 | C | T | rs4596677 | C/T |
| chr5 | 46311810 | C | A | rs4975958 | C/A |
| chr2 | 165726350 | G | C | rs11885920 | G/C |
| chr8 | 120223306 | C | T | rs11777125 | C/T |
| chr4 | 70501620 | T | C | rs11946584 | T/C |
| chr4 | 70501691 | C | T | rs11940905 | C/T |
| chr4 | 116153353 | - | A | rs11435788 | -/A |
| chr2 | 165721727 | T | C | rs12463840 | T/C |
| chr18 | 63109788 | A | G | rs12605618 | A/G |
| chr10 | 77925015 | T | G | rs12219798 | T/G |
| chr4 | 81114115 | G | A | rs12645327 | G/A |
| chr2 | 165720877 | G | A | rs13011371 | G/A |
| chr10 | 5677688 | C | T | rs12777138 | C/T |
| chr15 | 18849612 | T | C | rs28540688 | T/C |
| chr19 | 46046146 | G | A | rs28399443 | G/A |
| chr2 | 19630270 | G | A | rs6740759 | G/A |
| chr2 | 19630510 | A | G | rs6731115 | A/G |
| chr20 | 14719472 | C | G | rs6074799 | C/G |
| chr6 | 49047065 | C | T | rs6458640 | C/T |
| chr15 | 18840026 | A | T | rs6599973 | A/T |
| chr15 | 18840044 | C | T | rs6599974 | C/T |
| chr15 | 18840658 | A | G | rs6422229 | A/G |
| chr15 | 18849952 | A | C | rs6599977 | A/C |
| chr15 | 18850520 | C | T | rs6599978 | C/T |
| chr4 | 70508377 | G | A | rs6826237 | G/A |
| chr4 | 70508448 | T | G | rs6814603 | T/G |
| chr4 | 70508450 | C | A | rs6832784 | C/A |
| chr4 | 70508758 | G | A | rs6831951 | G/A |
| chr4 | 70509917 | G | A | rs6839067 | G/A |
| chr6 | 57539416 | - | TAC | rs33940047 | -/CTA |
| chr6 | 57539439 | - | TCA | rs33913327 | -/ATC |
| chr4 | 81106392 | - | GACA | rs33968007 | -/AGAC |
| chr4 | 190629948 | A | G | rs28814911 | A/G |
| chr18 | 46122972 | C | T | rs7233302 | C/T |
| chr18 | 63108997 | G | A | rs7235162 | G/A |
| chr11 | 103771597 | G | T | rs7102522 | G/T |
| chr3 | 74229657 | G | A | rs7427517 | G/A |

| | | | | | |
|---|---|---|---|---|---|
| chr15 | 18840178 | G | A | rs7402668 | G/A |
| chr8 | 13696528 | G | C | rs7017452 | G/C |
| chr4 | 49013822 | A | T | rs7377877 | A/T |
| chr4 | 49013874 | A | C | rs7377882 | A/C |
| chr4 | 70508686 | A | G | rs6858314 | A/G |
| chr7 | 6889588 | G | C | rs10270059 | G/C |
| chr7 | 6890016 | T | A | rs10263100 | T/A |
| chr7 | 6891221 | G | CC | rs10255784 | G/C |
| chr7 | 38366375 | A | T | rs10225471 | A/T |
| chr7 | 61486136 | C | T | rs10281866 | C/T |
| chr7 | 61486225 | A | G | rs10228846 | A/G |
| chr7 | 61486399 | C | G | rs10282232 | C/G |
| chr11 | 104798530 | A | C | rs10750724 | A/C |
| chr4 | 81113730 | T | C | rs10003491 | T/C |
| chr4 | 81106993 | A | G | rs11098964 | A/G |
| chr4 | 81107064 | C | T | rs11098965 | C/T |
| chr21 | 10117139 | C | A | rs461063 | C/A |
| chr21 | 10117188 | T | G | rs466171 | T/G |
| chr21 | 10117238 | T | C | rs411818 | T/C |
| chr21 | 10117285 | T | C | rs412034 | T/C |
| chr1 | 183081178 | T | C | rs593486 | T/C |
| chr20 | 4391634 | G | A | rs297676 | G/A |
| chr20 | 4392986 | A | G | rs167223 | A/G |
| chr11 | 58393937 | A | G | rs567460 | A/G |
| chr5 | 151436036 | A | C | rs154696 | A/C |
| chr5 | 151436183 | C | T | rs160037 | C/T |
| chr5 | 151442784 | G | C | rs787124 | G/C |
| chr21 | 20721261 | T | C | rs1028278 | T/C |
| chr11 | 5766922 | C | T | rs1453432 | C/T |
| chr6 | 74648613 | T | C | rs1370439 | T/C |
| chr9 | 137358716 | A | G | rs1111083 | A/G |
| chr21 | 10130868 | G | A | rs1752237 | G/A |
| chr21 | 20721816 | C | T | rs2187021 | C/T |
| chr21 | 20767362 | C | T | rs1786401 | C/T |
| chr21 | 20767378 | G | A | rs1735803 | G/A |
| chr21 | 20767387 | T | G | rs1735802 | T/G |
| chr21 | 20767481 | G | A | rs1735800 | G/A |
| chr21 | 20767680 | A | G | rs1735799 | A/G |
| chr21 | 20768129 | A | C | rs1735925 | A/C |
| chr21 | 20768420 | G | C | rs1735924 | G/C |
| chr21 | 20768435 | T | C | rs1735923 | T/C |
| chr7 | 61486104 | T | C | rs1823978 | T/C |
| chr7 | 61495581 | G | A | rs1840511 | G/A |
| chr10 | 5677656 | G | A | rs2380195 | G/A |
| chr10 | 5677761 | A | G | rs2380196 | A/G |
| chr10 | 5677778 | A | G | rs2380197 | A/G |
| chr10 | 77924718 | T | C | rs1907324 | T/C |
| chr13 | 80713640 | A | G | rs1937489 | A/G |

| | | | | | |
|---|---|---|---|---|---|
| chr13 | 80713866 | A | G | rs1904258 | A/G |
| chr13 | 80714011 | C | T | rs1904257 | C/T |
| chr3 | 89601022 | A | G | rs1912965 | A/G |
| chr15 | 18839522 | T | A | rs2062576 | T/A |
| chr15 | 18839594 | A | G | rs2062574 | A/G |
| chr15 | 18850611 | C | G | rs1846741 | C/G |
| chr15 | 18850917 | T | C | rs1827248 | T/C |
| chr15 | 18851079 | C | A | rs1988128 | C/A |
| chr8 | 126671312 | C | G | rs2124038 | C/G |
| chr4 | 20986195 | A | G | rs1994983 | A/G |
| chr4 | 70509359 | G | A | rs1897441 | G/A |
| chr4 | 70509387 | G | A | rs2217587 | G/A |
| chr5 | 40014655 | A | G | rs1876166 | A/G |
| chr6 | 57539192 | T | C | rs3857619 | T/C |
| chr6 | 57539425 | C | A | rs3996812 | C/A |
| chr8 | 6978994 | A | G | rs4397427 | A/G |
| chr4 | 49013774 | A | C | rs4311769 | A/T/C |
| chr4 | 49013993 | C | A | rs4022027 | C/A |
| chr4 | 49015400 | G | C | rs4440192 | G/C |
| chr4 | 173671444 | A | G | rs3104245 | A/G |
| chr21 | 10130908 | T | C | rs2479478 | T/C |
| chr21 | 20767429 | G | T | rs2776098 | G/T |
| chr7 | 38349483 | G | A | rs2975073 | G/A |
| chr7 | 38349490 | C | T | rs2975072 | C/T |
| chr7 | 38350632 | A | CC | rs2534582 | A/C |
| chr7 | 38350651 | A | G | rs2534583 | A/G |
| chr7 | 38352463 | T | C | rs2534587 | T/C |
| chr1 | 143802718 | A | G | rs2590154 | A/G |
| chr1 | 143802751 | C | T | rs2794072 | C/T |
| chr1 | 143802921 | C | G | rs2590153 | C/G |
| chr1 | 143803166 | G | C | rs2762756 | G/C |
| chr1 | 143803181 | T | C | rs2762755 | T/C |
| chr1 | 143803270 | G | A | rs2794071 | G/A |
| chr1 | 143803289 | T | C | rs2762753 | T/C |
| chr1 | 143803951 | G | A | rs2794070 | G/A |
| chr1 | 143803961 | G | A | rs2794069 | G/A |
| chr1 | 143804001 | A | G | rs2590151 | A/G |
| chr1 | 143804086 | T | A | rs2596316 | T/A |
| chr1 | 143804167 | C | A | rs2596315 | C/A |
| chr10 | 77932188 | A | G | rs2579759 | A/G |
| chr10 | 77932208 | T | C | rs2579758 | T/C |
| chr10 | 77932330 | A | C | rs2637237 | A/C |
| chr6 | 57409819 | G | A | rs2397293 | G/A |
| chr6 | 57409902 | G | C | rs2397294 | G/C |
| chr6 | 57539171 | A | G | rs2397937 | A/G |
| chr6 | 57539216 | C | T | rs2397938 | C/T |
| chr6 | 57539413 | C | A | rs2397535 | C/A |
| chrX | 35545873 | G | A | rs2878512 | G/A |

| | | | | | |
|---|---|---|---|---|---|
| chr4 | 39792 | A | G | rs2859211 | A/G |
| chr4 | 116146071 | T | G | rs2583521 | T/G |
| chr4 | 116147122 | A | G | rs2620423 | A/G |
| chr17 | 49513177 | A | G | rs8070658 | A/G |
| chr17 | 49523656 | G | A | rs8075842 | G/A |
| chr1 | 141732572 | A | G | rs9282950 | A/G |
| chr18 | 46123212 | G | T | rs8089629 | G/T |
| chr6 | 57409273 | C | T | rs7751464 | C/T |
| chr4 | 40198 | C | T | rs7685192 | C/T |
| chr4 | 49014226 | A | C | rs9291384 | A/C |
| chr4 | 70508131 | C | T | rs7670238 | C/T |
| chr4 | 70508142 | G | T | rs7668967 | G/T |
| chr4 | 81114041 | C | G | rs7673310 | C/G |
| chr5 | 46305650 | G | A | rs8185213 | G/A |
| chr17 | 65966415 | C | T | rs9904480 | C/T |
| chr6 | 49038824 | A | T | rs9395432 | A/T |
| chr6 | 57409620 | T | A | rs9382736 | T/A |
| chr15 | 18841341 | C | T | rs9744615 | C/T |
| chr1 | 141731586 | C | T | ENSSNP35434 | C/T |
| chr1 | 141731606 | T | C | ENSSNP35435 | T/C |
| chr1 | 141731610 | A | T | ENSSNP35436 | A/T |
| chr1 | 141731670 | T | C | ENSSNP35442 | T/C |
| chr1 | 141731898 | T | C | ENSSNP35449 | T/C |
| chr1 | 141749353 | G | A | ENSSNP35771 | G/A |
| chr1 | 141749424 | A | T | ENSSNP35772 | A/T |
| chr1 | 141749486 | G | A | ENSSNP35775 | G/A |
| chr1 | 141749516 | G | A | ENSSNP35776 | G/A |
| chr15 | 18840989 | C | A | ENSSNP1066221 | C/A |
| chr15 | 18849536 | G | A | ENSSNP1066244 | G/A |
| chr15 | 18849545 | C | T | ENSSNP1066245 | C/T |
| chr15 | 18849599 | T | A | ENSSNP1066250 | T/A |
| chr15 | 18849607 | T | G | ENSSNP1066251 | T/G |
| chr15 | 18849618 | T | C | ENSSNP1066253 | T/C |
| chr15 | 18849627 | G | A | ENSSNP1066254 | G/A |
| chr21 | 10130511 | G | C | ENSSNP1938763 | G/C |
| chr21 | 10130519 | T | G | ENSSNP1938764 | T/G |
| chr21 | 10130521 | A | C | ENSSNP1938765 | A/C |
| chr21 | 10130568 | G | A | ENSSNP1938769 | G/A |
| chr21 | 10130608 | T | C | ENSSNP1938770 | T/C |
| chr21 | 10130725 | A | G | ENSSNP1938779 | A/G |
| chr21 | 10130780 | G | C | ENSSNP1938781 | G/C |
| chr21 | 10130838 | T | G | ENSSNP1938782 | T/G |
| chr1 | 150821139 | G | A | | |
| chr1 | 150854406 | C | T | | |
| chr1 | 208788820 | C | T | | |
| chr10 | 5627028 | T | C | | |
| chr10 | 77931338 | G | A | | |
| chr11 | 101070812 | T | C | | |

| | | | | | |
|---|---|---|---|---|---|
| chr11 | 101071969 | A | C | | |
| chr11 | 101072011 | T | G | | |
| chr11 | 101072014 | G | T | | |
| chr11 | 101072021 | C | T | | |
| chr11 | 101072042 | G | A | | |
| chr11 | 101072070 | C | G | | |
| chr11 | 101079962 | T | G | | |
| chr12 | 68884318 | T | G | | |
| chr14 | 84365087 | A | G | | |
| chr15 | 18841496 | A | G | | |
| chr15 | 18841519 | C | T | | |
| chr15 | 18841526 | G | A | | |
| chr15 | 18841535 | G | AT | | |
| chr15 | 18841546 | T | C | | |
| chr15 | 18841556 | C | T | | |
| chr15 | 18849649 | A | G | | |
| chr17 | 15729015 | G | C | | |
| chr17 | 63285250 | - | AT | | |
| chr17 | 63285268 | G | A | | |
| chr17 | 63285269 | G | A | | |
| chr17 | 63285283 | C | T | | |
| chr18 | 46122601 | C | A | | |
| chr18 | 46122922 | A | G | | |
| chr18 | 46122957 | T | C | | |
| chr18 | 46130388 | T | C | | |
| chr18 | 46130571 | G | C | | |
| chr2 | 165721589 | C | T | | |
| chr2 | 165721598 | G | A | | |
| chr2 | 165726134 | G | A | | |
| chr2 | 19631065 | T | C | | |
| chr2 | 4766028 | G | T | | |
| chr20 | 4397612 | G | A | | |
| chr21 | 10118152 | C | T | | |
| chr21 | 10118166 | A | G | | |
| chr21 | 10118168 | C | T | | |
| chr21 | 10118182 | A | G | | |
| chr21 | 10118227 | C | T | | |
| chr21 | 10118243 | C | T | | |
| chr21 | 10118356 | C | T | | |
| chr21 | 10130561 | G | A | | |
| chr21 | 10130566 | G | T | | |
| chr21 | 10130817 | C | T | | |
| chr21 | 10130827 | A | C | | |
| chr21 | 10130986 | A | C | | |
| chr3 | 56582771 | - | T | | |
| chr3 | 68830548 | T | AAAAAAAAAAA | | |
| chr3 | 68830553 | T | A | | |
| chr3 | 89590876 | A | G | | |

| | | | | | |
|---|---|---|---|---|---|
| chr3 | 89601252 | A | G | | |
| chr4 | 173670539 | C | T | | |
| chr4 | 173670544 | A | C | | |
| chr4 | 49013798 | - | AA | | |
| chr4 | 49013904 | G | T | | |
| chr4 | 49014051 | G | A | | |
| chr4 | 49014164 | T | G | | |
| chr4 | 49014948 | C | T | | |
| chr4 | 49014963 | T | A | | |
| chr4 | 49014981 | G | A | | |
| chr4 | 49015192 | A | C | | |
| chr4 | 49015194 | G | A | | |
| chr4 | 49015252 | C | T | | |
| chr4 | 49015300 | T | C | | |
| chr4 | 49015416 | A | C | | |
| chr4 | 49015425 | G | T | | |
| chr4 | 49015452 | T | A | | |
| chr4 | 49015522 | C | A | | |
| chr4 | 49015529 | T | C | | |
| chr4 | 81111528 | A | G | | |
| chr4 | 81111545 | C | T | | |
| chr4 | 81111558 | G | A | | |
| chr5 | 46305720 | G | C | | |
| chr5 | 46308085 | T | C | | |
| chr5 | 46310009 | A | G | | |
| chr5 | 46310013 | G | A | | |
| chr5 | 46310017 | G | A | | |
| chr5 | 46310132 | C | T | | |
| chr5 | 46310626 | G | A | | |
| chr5 | 46310633 | C | T | | |
| chr5 | 46310661 | A | T | | |
| chr5 | 46310718 | A | C | | |
| chr5 | 46310730 | C | G | | |
| chr5 | 46310750 | C | T | | |
| chr5 | 57358343 | A | G | | |
| chr5 | 57358718 | C | A | | |
| chr5 | 57358911 | C | T | | |
| chr5 | 57358942 | G | A | | |
| chr5 | 57359784 | C | G | | |
| chr5 | 57370650 | A | G | | |
| chr5 | 57370736 | A | G | | |
| chr5 | 57370804 | C | T | | |
| chr6 | 49046749 | G | A | | |
| chr6 | 49046816 | A | C | | |
| chr6 | 49047044 | C | G | | |
| chr6 | 49047238 | A | G | | |
| chr6 | 57539274 | T | C | | |
| chr7 | 38348669 | C | T | | |

| | | | | | |
|---|---|---|---|---|---|
| chr7 | 38348759 | G | T | | |
| chr7 | 38348865 | T | A | | |
| chr7 | 38349166 | T | C | | |
| chr7 | 38349211 | A | G | | |
| chr7 | 38349240 | G | A | | |
| chr7 | 38349257 | C | T | | |
| chr7 | 38349275 | G | A | | |
| chr7 | 38349321 | A | G | | |
| chr7 | 38349341 | C | T | | |
| chr7 | 38349342 | A | G | | |
| chr7 | 38349364 | C | T | | |
| chr7 | 38349370 | T | A | | |
| chr7 | 38349372 | T | A | | |
| chr7 | 38349373 | T | C | | |
| chr7 | 38349382 | G | A | | |
| chr7 | 38349576 | C | G | | |
| chr7 | 38349620 | A | C | | |
| chr7 | 38349647 | T | A | | |
| chr7 | 38349657 | A | G | | |
| chr7 | 38350023 | A | G | | |
| chr7 | 38350059 | A | G | | |
| chr7 | 38350631 | G | CC | | |
| chr7 | 38350678 | C | G | | |
| chr7 | 38350689 | A | G | | |
| chr7 | 38366282 | T | A | | |
| chr7 | 38367589 | C | T | | |
| chr7 | 38367603 | C | G | | |
| chr7 | 38367607 | C | T | | |
| chr7 | 38367616 | T | A | | |
| chr7 | 38367644 | A | G | | |
| chr7 | 61485543 | A | G | | |
| chr7 | 61485548 | G | C | | |
| chr7 | 61485585 | C | G | | |
| chr7 | 61485604 | C | A | | |
| chr7 | 61485641 | C | T | | |
| chr7 | 61485642 | A | T | | |
| chr7 | 61485644 | A | G | | |
| chr7 | 61495164 | C | G | | |
| chr7 | 6866672 | T | C | | |
| chr7 | 6889428 | C | T | | |
| chr7 | 6889438 | T | A | | |
| chr7 | 6889440 | T | G | | |
| chr7 | 6889442 | T | G | | |
| chr7 | 6889477 | A | G | | |
| chr7 | 6889496 | C | A | | |
| chr7 | 6889522 | A | G | | |
| chr7 | 6889535 | G | A | | |
| chr7 | 6889544 | T | C | | |

| | | | | | |
|---|---|---|---|---|---|
| chr7 | 6889546 | G | T | | |
| chr7 | 6889595 | G | A | | |
| chr7 | 6889703 | A | T | | |
| chr7 | 6890112 | G | T | | |
| chr7 | 6890118 | C | T | | |
| chr7 | 6890126 | T | G | | |
| chr7 | 6890146 | A | C | | |
| chr7 | 6890621 | C | T | | |
| chr7 | 6891225 | T | CC | | |
| chr8 | 120222844 | G | T | | |
| chr8 | 6978939 | T | C | | |

**Table S5**. Accession numbers.

| Accession | Data Type | Sample(s) |
|---|---|---|
| | | |
| GSE9002 | Array-CGH (microarray) | NA15510 vs. NA18505 |
| SRA000197 | PEM paired-ends (DNA) | NA15510 |
| SRA000198 | PEM paired-ends (DNA) | NA15510 |
| SRA000199 | PEM paired-ends (DNA) | NA18505 |
| SRA000200 | PEM paired-ends (DNA) | NA18505 |
| SRA000201 | PEM paired-ends (DNA) | NA18505 |
| SRA000202 | PEM paired-ends (DNA) | NA18505 |
| SRA000203 | PEM paired-ends (DNA) | NA18505 |
| SRA000204 | Amplicon pool sequences (DNA) | NA15510 |
| SRA000205 | Amplicon pool sequences (DNA) | NA18505 |