Table S1. Breakpoint coordinates, breakpoint annotation, and additional information.
Additional file.

Table S2. Comparison of SV annotations and SV with micro-insertions in different sets.

	# of people	# of variants > 100 bp	NH	NAHR	TEI	VNTR	# with MI	# with MI > 10 bp
Lam et al.[1]	14	1,961	45%	28%	21%	5%	0	0
Kidd et al.[2]	17	1,054	52%	26%	19%	3%	160	82
Conrad et al.[3]	3	324	70-80%*	10-15%*	0%	10-15%*	103	41
Pang et al.[4]*	1	7,330	13%	8%	24%	55%	unknown	0
This study	1,092	8,709	61%	13%	25%	0%	2,391	635

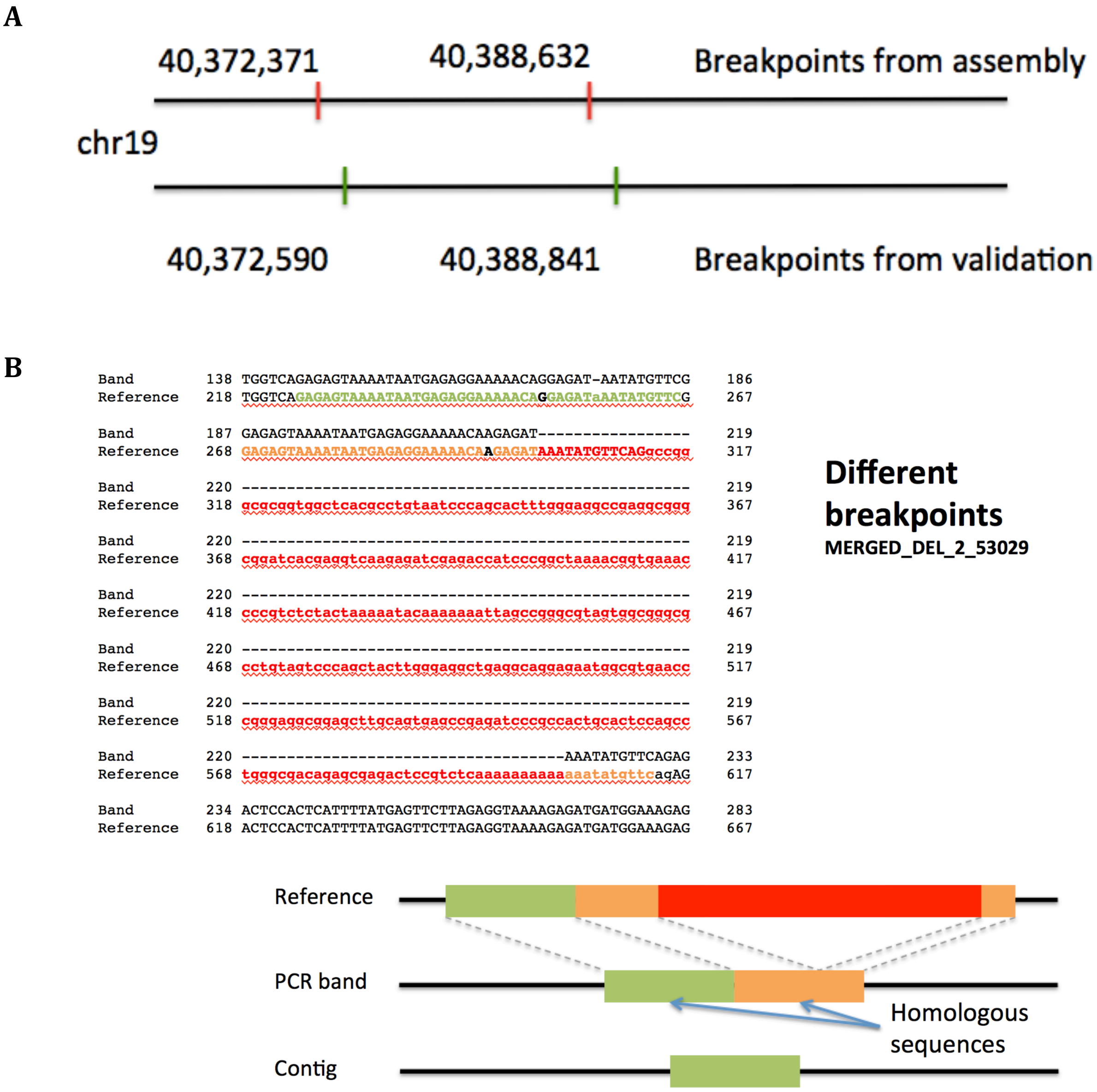
* Including calls that are not at breakpoint resolution, i.e., from **Fig. 3** in Pang et al.[4] and Table 1 in Conrad et al.[3]

Table S3. Statistics on constructing new breakpoint library for BreakSeq2 and ascertaining BreakSeq2 performance.
Additional file.

Table S4. Testing for SNP/indel enrichment around SV breakpoints. Distributions of normalized SNP/indel densities around breakpoints and at large distance were tested by t-test. Regions around breakpoints were defined as a 200 kbps region upstream of the 5'-breakpoints and a 200 kbps region downstream of 3'-breakpoint. Regions at distance were defined between 1 Mbps to 800 kbps upstream of the 5'-breakpoints and between 800 kbps and 1 Mbps downstream of 3'-breakpoint. Regions were divided into bins of 40 kbps in length. Bonferroni correction was applied given that we did 42 tests: 21 for SNPs and 21 for indels.

Breakpoint type	SNP/indel type	Raw p-value	Bonferroni corrected p-value, * -- significant
NH	All	5.80x10^{-7}	**2.44x10^{-5*}**
	C>A	2.48x10^{-7}	**1.04x10^{-5*}**
	C>G	5.91x10^{-7}	**2.48x10^{-5*}**
	C>T	1.51x10^{-6}	**6.35x10^{-5*}**
	T>A	1.58x10^{-7}	**6.63x10^{-6*}**
	T>C	4.79x10^{-7}	**2.01x10^{-5*}**
	T>G	8.12x10^{-9}	**3.41x10^{-7*}**
TEI	All	6.48x10^{-4}	**2.72x10^{-2*}**
	C>A	1.64x10^{-3}	6.90x10^{-2}
	C>G	8.86x10^{-3}	3.72x10^{-1}
	C>T	2.00x10^{-3}	8.40x10^{-2}
	T>A	1.15x10^{-4}	**4.83x10^{-3*}**
	T>C	1.56x10^{-3}	6.55x10^{-2}
	T>G	8.92x10^{-4}	**3.75x10^{-2*}**
NAHR	All	6.23x10^{-5}	**2.62x10^{-3*}**
	C>A	1.64x10^{-4}	**6.89x10^{-3*}**
	C>G	3.74x10^{-1}	1
	C>T	6.82x10^{-6}	**2.86x10^{-4*}**
	T>A	8.35x10^{-6}	**3.51x10^{-4*}**
	T>C	2.08x10^{-1}	1
	T>G	1.23x10^{-1}	1

Table S5. Location of micro-insertions.
Additional file.



```
Band        138 TGGTCAGAGAGTAAAATAATGAGAGGAAAAACAGGAGAT-AATATGTTCG    186
Reference   218 TGGTCAGAGAGTAAAATAATGAGAGGAAAAACAGGAGATaAATATGTTCG    267

Band        187 GAGAGTAAAATAATGAGAGGAAAAACAAGAGAT-----------------    219
Reference   268 GAGAGTAAAATAATGAGAGGAAAAACAAGAGATAAATATGTTCAGgccgg    317

Band        220 --------------------------------------------------    219
Reference   318 gcgcggtggctcacgcctgtaatcccagcactttgggaggccgaggcggg    367

Band        220 --------------------------------------------------    219
Reference   368 cggatcacgaggtcaagagatcgagaccatcccggctaaaacggtgaaac    417

Band        220 --------------------------------------------------    219
Reference   418 cccgtctctactaaaaatacaaaaaaattagccgggcgtagtggcgggcg    467

Band        220 --------------------------------------------------    219
Reference   468 cctgtagtcccagctacttgggaggctgaggcaggagaatggcgtgaacc    517

Band        220 --------------------------------------------------    219
Reference   518 cgggaggcggagcttgcagtgagccgagatcccgccactgcactccagcc    567

Band        220 ------------------------------------AAATATGTTCAGAG    233
Reference   568 tgggcgacagagcgagactccgtctcaaaaaaaaaaaaatatgttcagAG    617

Band        234 ACTCCACTCATTTTATGAGTTCTTAGAGGTAAAAGAGATGATGGAAAGAG    283
Reference   618 ACTCCACTCATTTTATGAGTTCTTAGAGGTAAAAGAGATGATGGAAAGAG    667
```

Figure S1. Examples of discrepancies in predicted and validated breakpoint coordinates. A) Most frequently, predicted breakpoints were shifted relative to those derived from validation excesses. One such example is depicted. All such cases were removed by post validation filteres. B) In one case (chr9:35803108-35803461) assembly collapsed tandem repeat around breakpoints resulting in shorter contig and overestimated breakpoints.

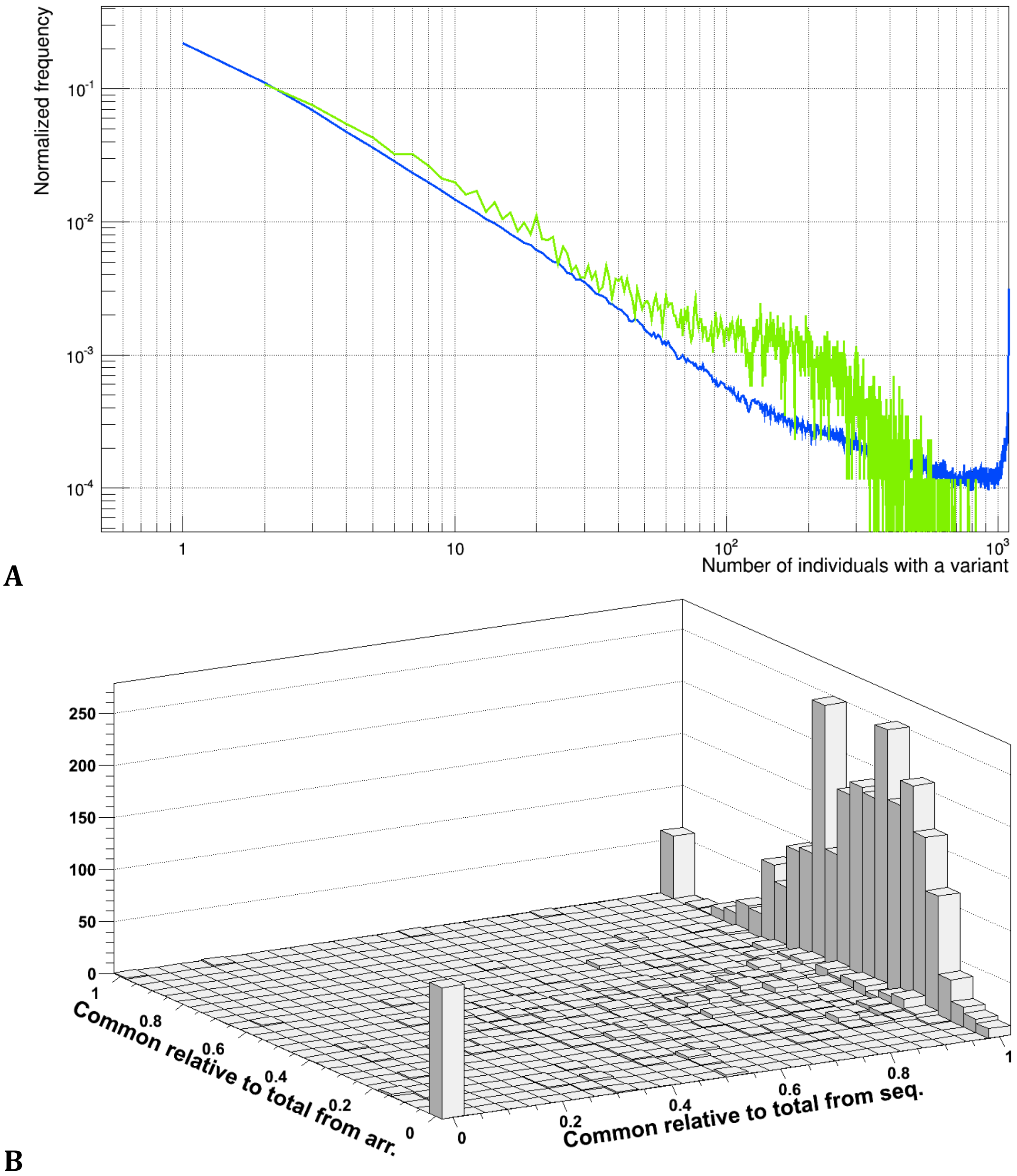


Figure S2. Variant sample frequency spectrum. A) Frequency for bi-allelic SNPs[5] is in blue, while frequency for deletions in the this study is in green. SNPs and deletions were discovered from the same 1,092 individuals. B) Correlation of samples genotyped as having deletion from OMNI SNP genotyping array (y-axis) and from mapping reads to sequences of breakpoint junctions (x-axis) in 292 samples. Values on x/y axis is the fraction of samples with deletion common between the two ways of genotyping divided by the number of samples genotypes as having deletion by read mapping/by OMNI SNP array. Number of deletions with such fractions is on z-axis. Genotyping from sequencing, i.e., from mapping reads to sequence of deletion junctions, underestimates frequency of deletion by roughly 60%. Therefore, the true frequency spectrum for deletion in A, is shifter right.

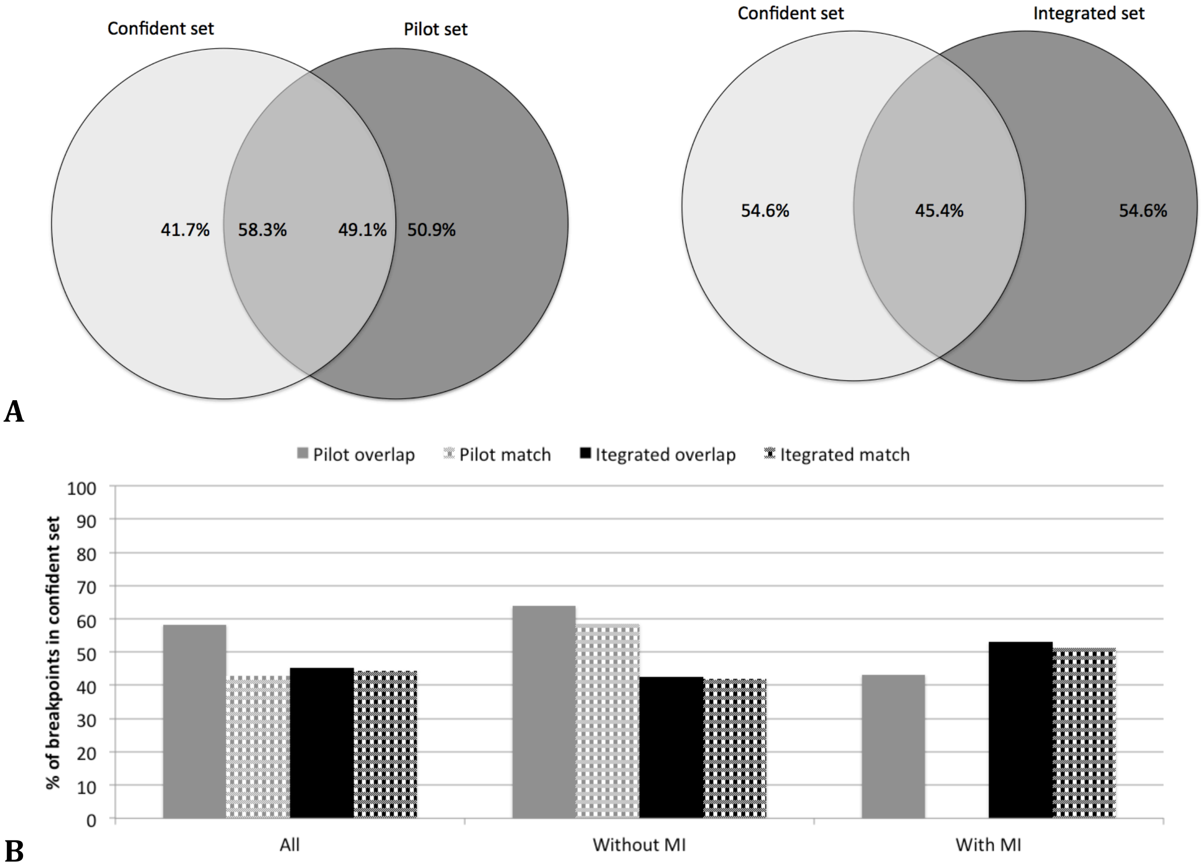


Figure S3. Comparison of confident set of breakpoints with pilot and integrated sets of breakpoints. A) Fraction of SV breakpoints that overlap 50% reciprocally between sets. B) Breakdown of confident breakpoints with/without MIs by 50% reciprocal overlap and exact match to breakpoints in pilot/integrated sets. For exact match we only matched start and end of the breakpoints. Overlap is higher with pilot set, but there is marked difference in the fraction of overlapping and exactly matching breakpoints to pilot set. The difference is particularly drastic for breakpoints with MIs, demonstrating that pilot set was particularly limited in resolving MIs. The difference is minor when comparing to integrated set. However, integrated set represented deletion breakpoint in a narrow size range (**Fig. 1C**).

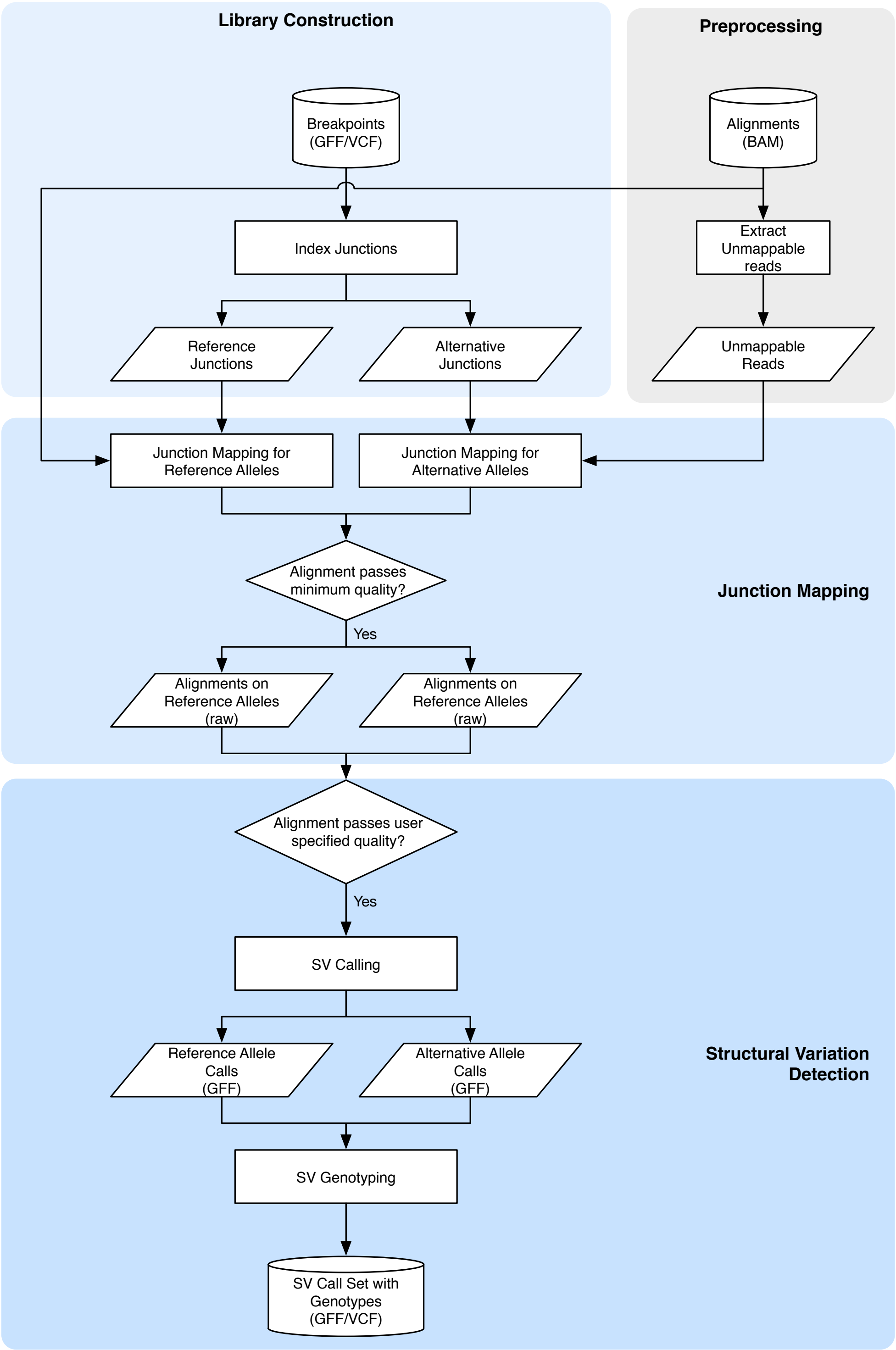

Library Construction
Preprocessing
Breakpoints (GFF/VCF)
Alignments (BAM)
Index Junctions
Extract Unmappable reads
Reference Junctions
Alternative Junctions
Unmappable Reads
Junction Mapping for Reference Alleles
Junction Mapping for Alternative Alleles
Alignment passes minimum quality?
Yes
Junction Mapping
Alignments on Reference Alleles (raw)
Alignments on Reference Alleles (raw)
Alignment passes user specified quality?
Yes
SV Calling
Reference Allele Calls (GFF)
Alternative Allele Calls (GFF)
Structural Variation Detection
SV Genotyping
SV Call Set with Genotypes (GFF/VCF)

Figure S4. BreakSeq2 workflow. Reads which are unmapped, soft-clipped or badly mapped are considered unmappable against the reference genome. Further filtering based on the mapping quality and the edit distances of the alignment against the reference genome is done to narrow down the list of unmappable reads. Alignments of the unmappable reads against the breakpoint library is used to gather evidence for SVs. The more comprehensive the breakpoint library is, the better BreakSeq2 is expected to perform.

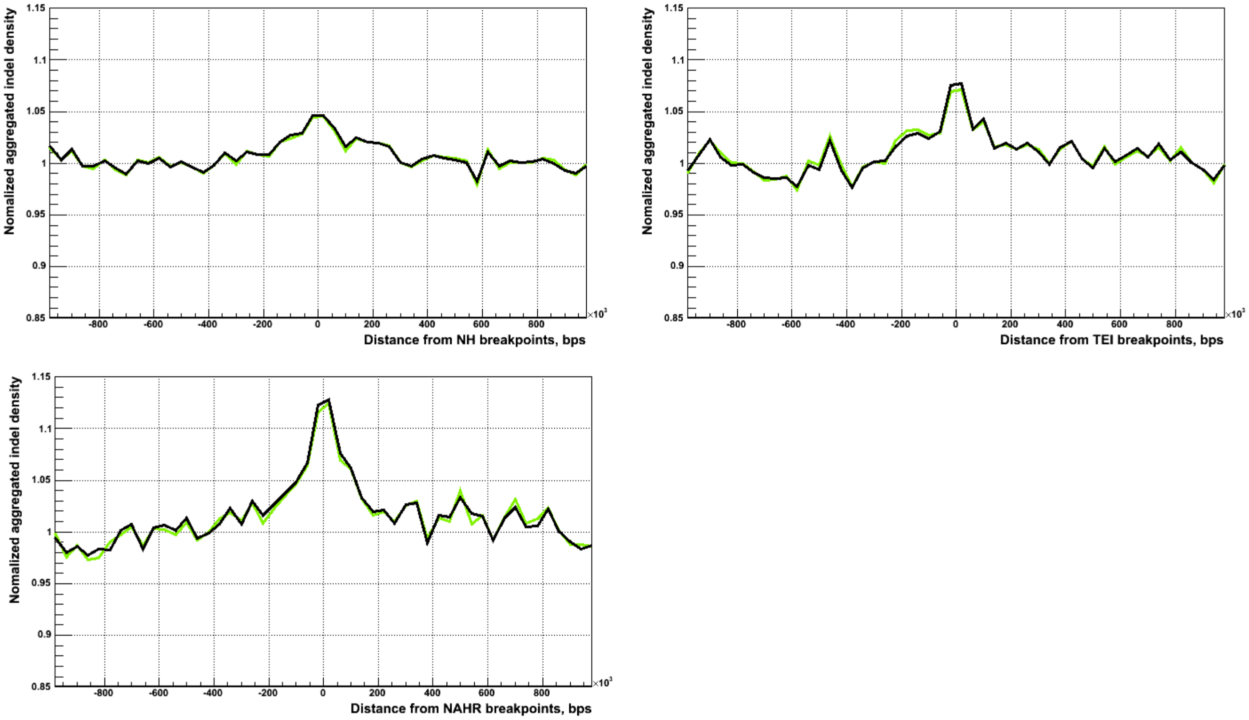


Figure S5. Indel aggregation around deletion breakpoints. Aggregation for indels of 1-6 bps in length is in black; aggregation for indels of 1 bp in length is in green.

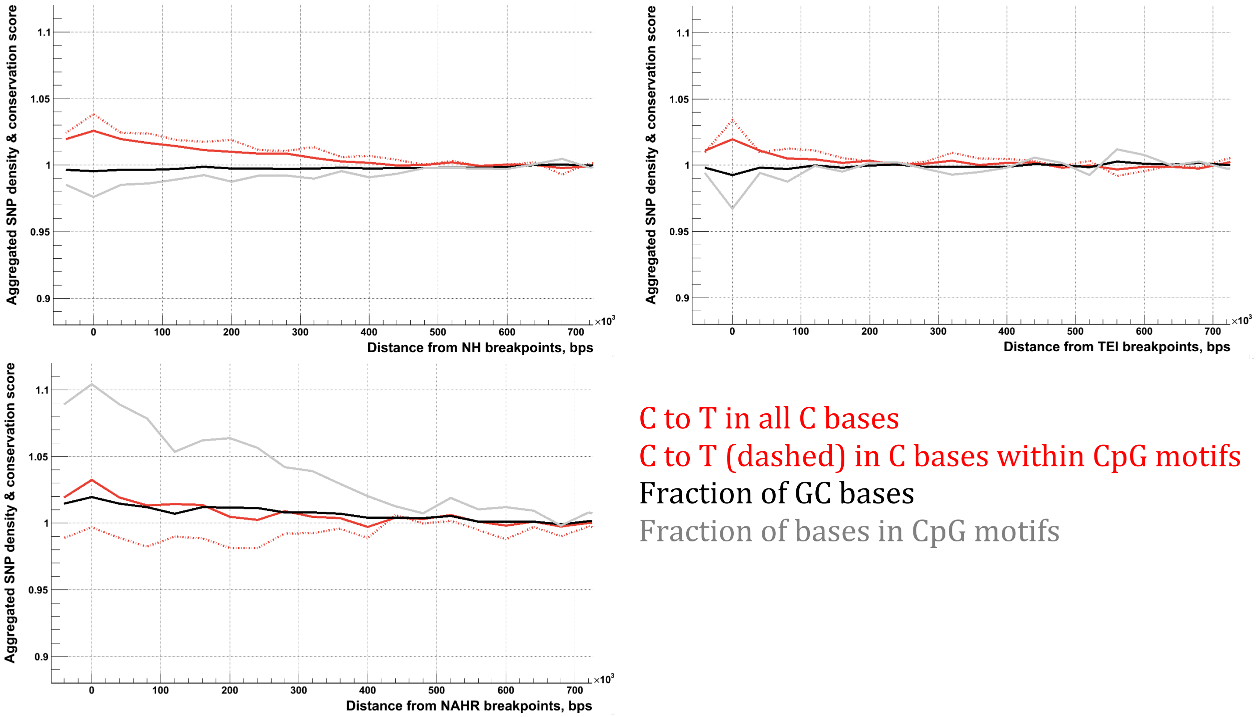


Figure S6. C to T mutations, GC content and CpG contents around breakpoints of different classes. All curves are normalized to unity at tails. Only unmasked bases, i.e. those where the 1000 Genomes Project can do confident SNP calling, were used in the analysis. NAHR breakpoints show very different distributions from the breakpoints of other classes. They do show increase in GC and CpG content while NH and TEI do not. Frequency of C to T substitutions also decreases in CpG motifs around NAHR while increases around NH and TEI. The latter may imply association of NAHR with regions of lower methylation and association of NH and TEI with regions of higher methylation.

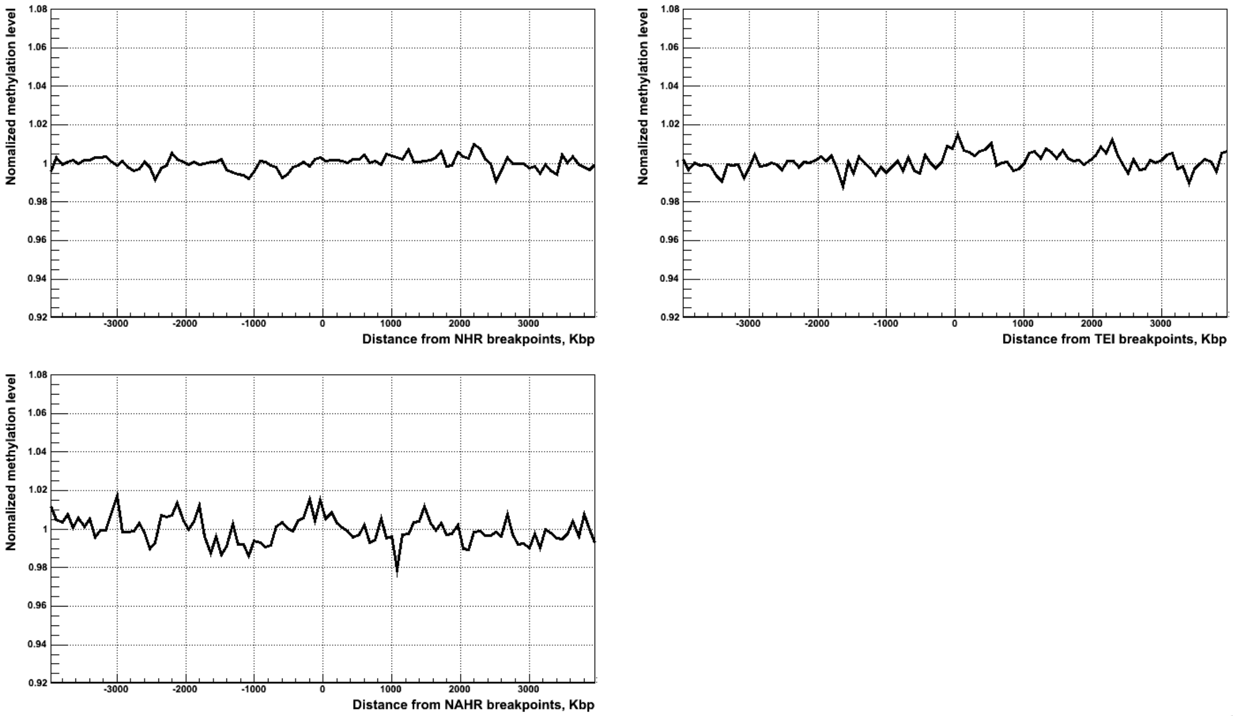


Figure S7. Methylation levels in H1ESC cell line around breakpoints of different classes. There is no noticeable change in methylation level around breakpoints of either class. On a smaller scale we do observed increase in methylation level in the regions of about 1 kbp around breakpoints of each type (data not shown). Though, this could be technical artifact, as breakpoints generally have higher repeat content and all calculated values, including methylation level, will be prone to mistakes in such regions. For instance, SNPs densities in unmasked sites showed sharp increase in such proximity to breakpoints.

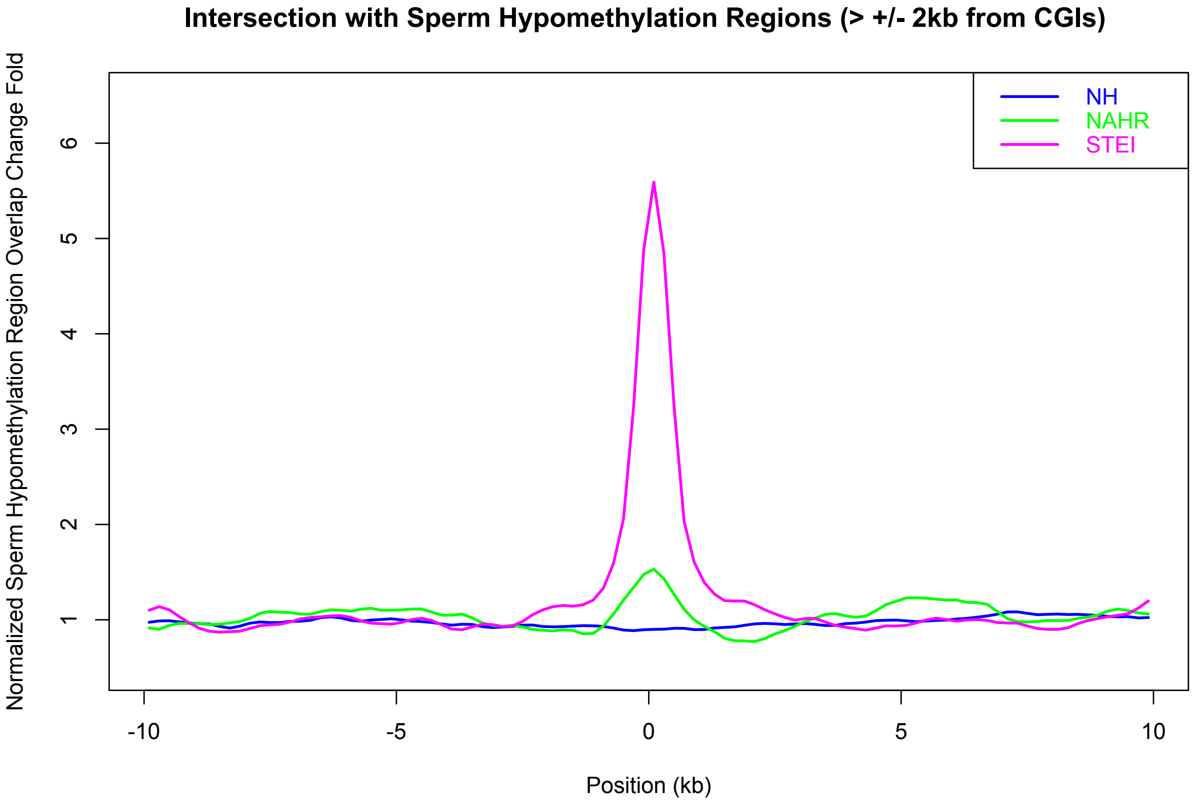


Figure S8. Overlap of breakpoints with hypomethylated regions in sperm. Only regions outside of 2 kbp windows of CpG islands were considered. Coordiantes of CpG islands were downloaded from http://epigraph.mpi-inf.mpg.de/download/CpG_islands_revisited/

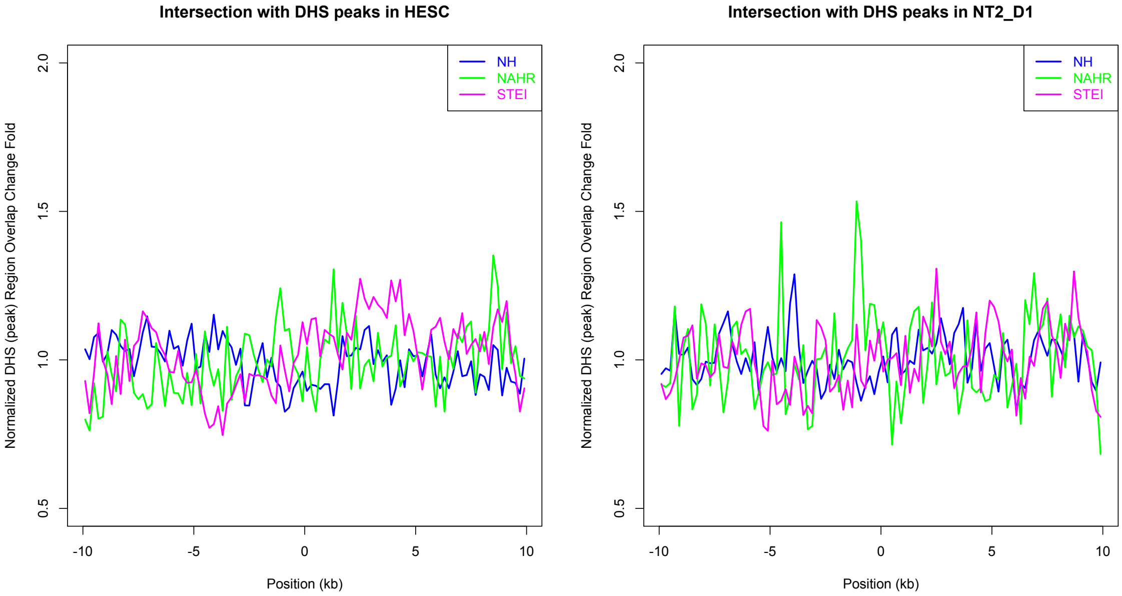


Figure S9. Aggregation of DNAse peak around breakpoints of different classes. Data for hESC (human embryonic stem cell) and NT2 (testis cancer, metastasis site on lung) cell lines have been utilized. Arguably, NH breakpoints exhibit depletion for DNAse sites while NAHR breakpoints exhibit enrichment.

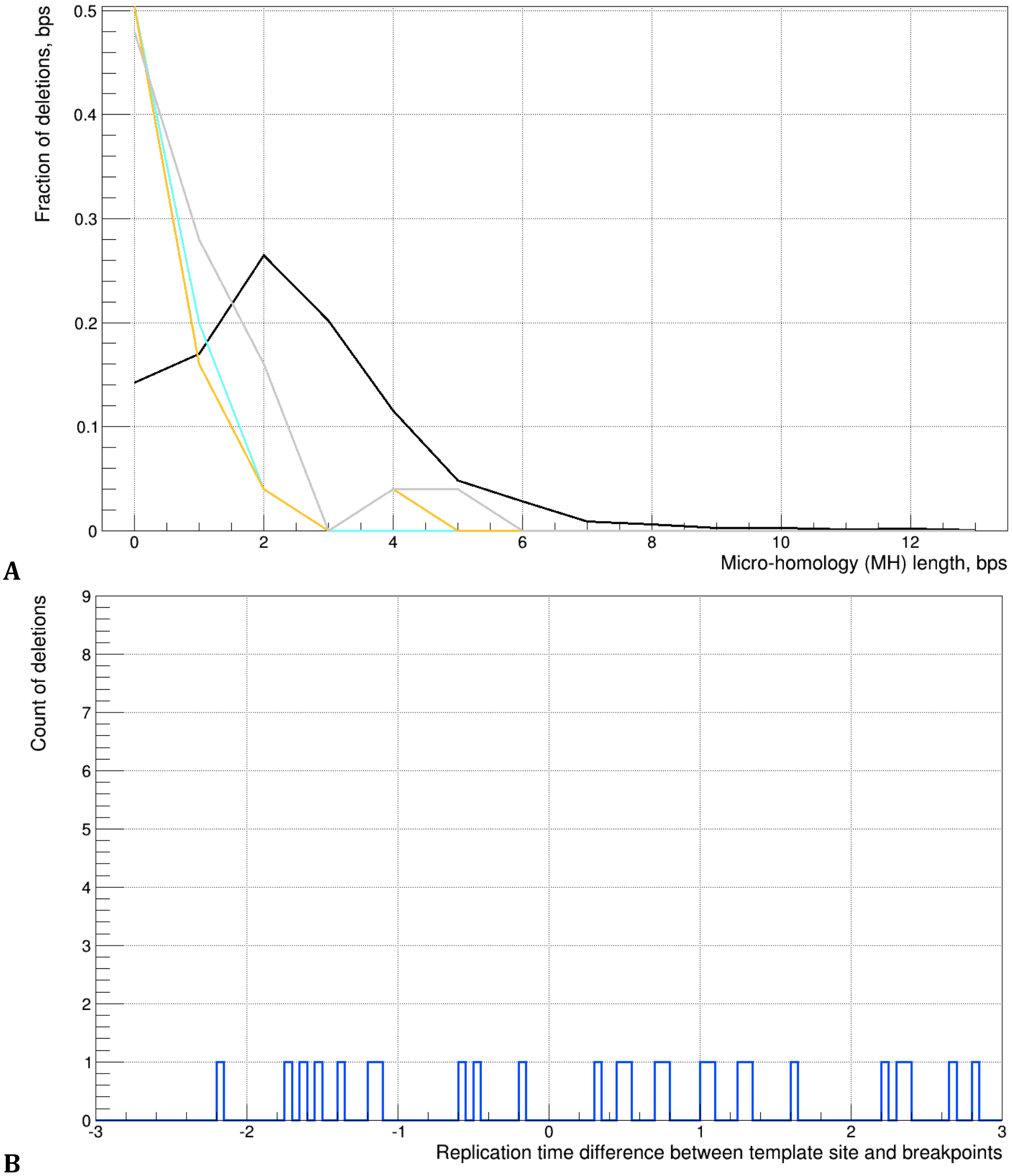


Figure S10. Analysis of micro-insertions (MI) with template sites on different chromosome. A) Length of micro-homology (MH) at deletion junction is not different from random (see also **Fig. 4**). For deletions with MIs and identified template site, MHs are calculated for 5'-ends/3'-ends of the deletion and the template site. B) The difference in replication time between template site and breakpoints does not reveal significant later or earlier replication time of template sites.

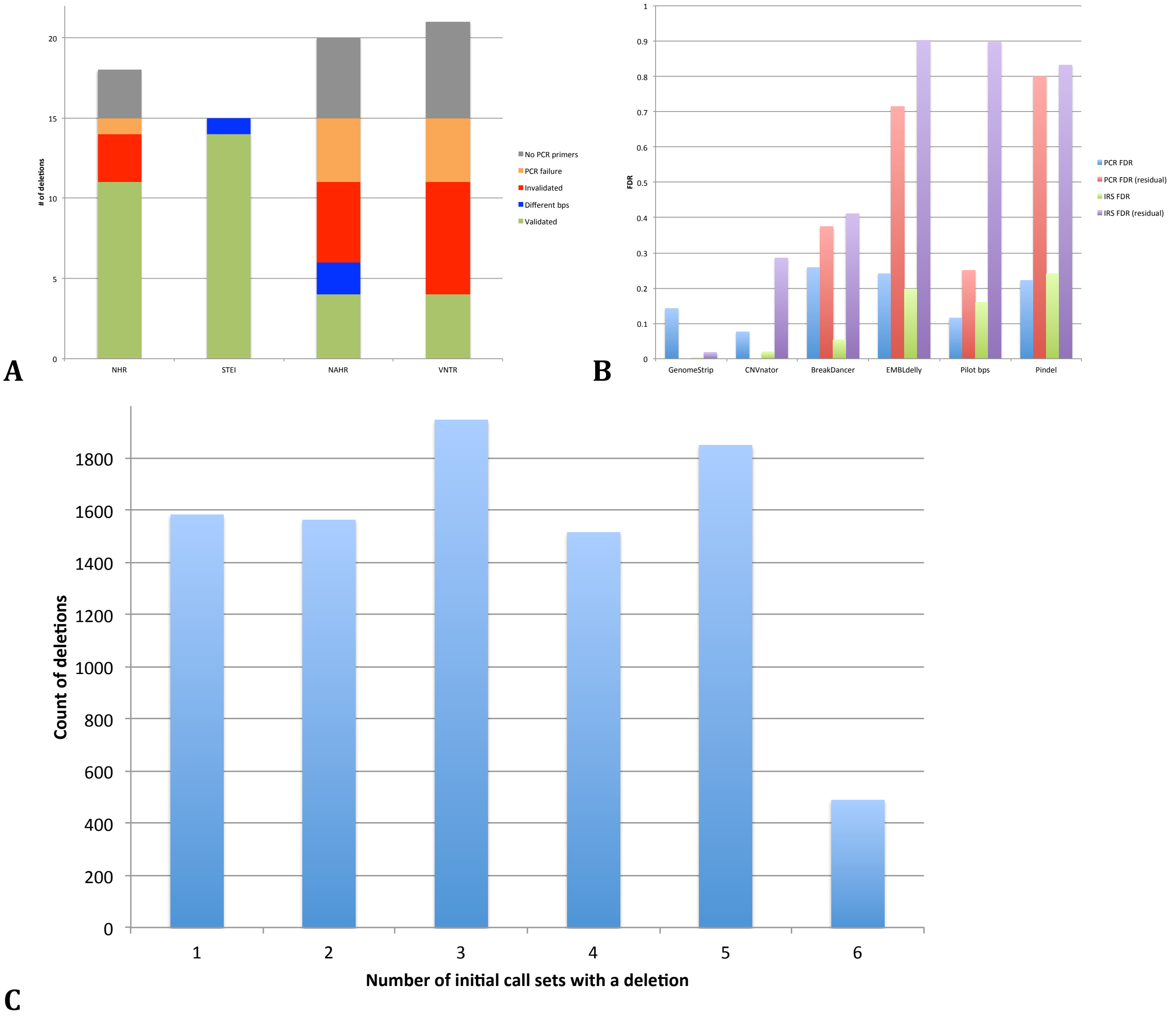


Figure S11. Validation results before final deletion filtering. A) Breakdown by classification mechanisms. Deletions classified as Variable Number of Tandem Repeats – VNTR do not validate well as their breakpoints are in very repetitive sequences; B) Breakdown by calling method. Methods discovering deletions from split-read analysis (Delly, Pindel, and assembly in the pilot) have overall high FDR and very high residual FDR. C) Breakdown of deletions in the final set by presence in the initial call sets.

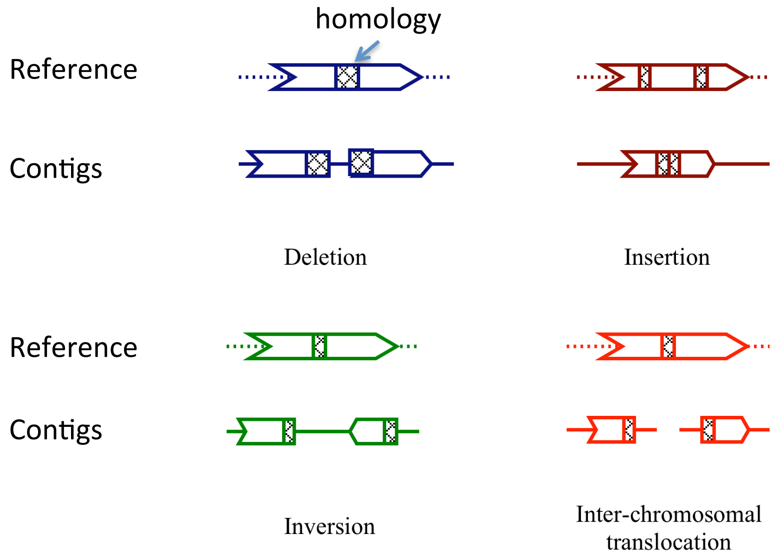


Figure S12. Examples of CROSSMATCH alignments to derive breakpoints of structural variations.

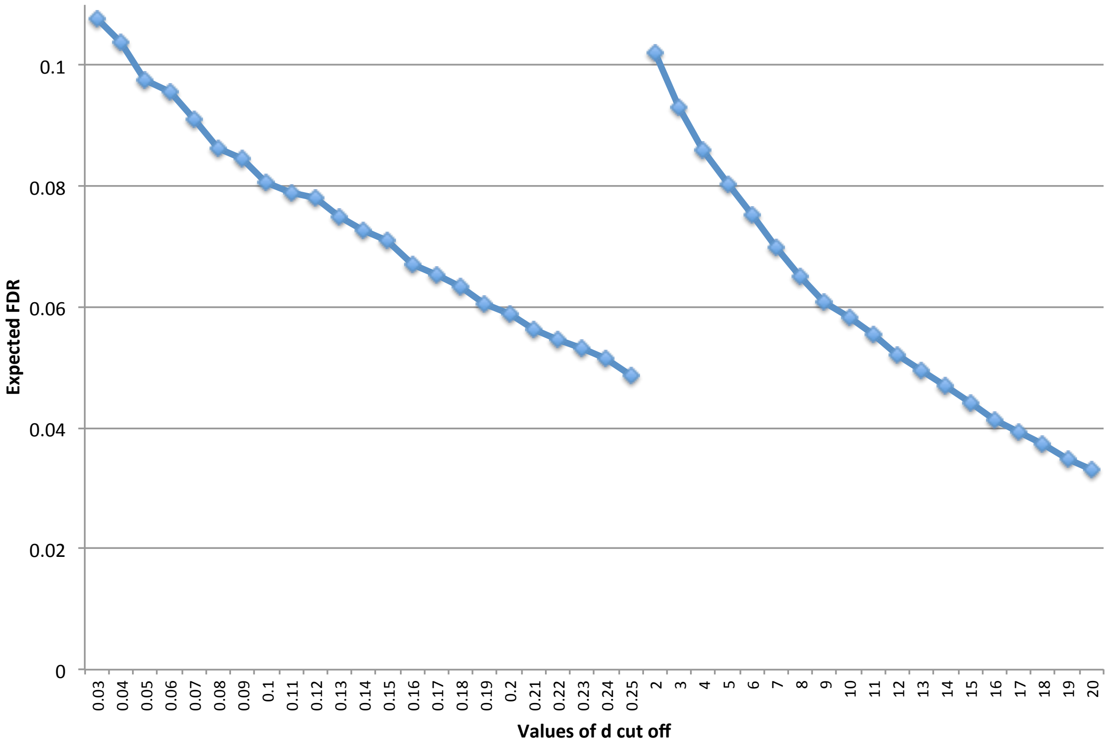


Figure S13. *In-silico* FDR for breakpoint support with the values of *d*. When realigning all unmapped reads to the null junction library and varied the value of *d* to compare the number of null junction passing the filter with the number of real junctions passing the filter. The cutoff *d* is defined as the number/fraction of bases aligned to each flank for deciding, which reads supported breakpoints. Left curve represents the results when *d* is calculated as a fraction of read length. Right curve represents the results when *d* considered in number of bases.

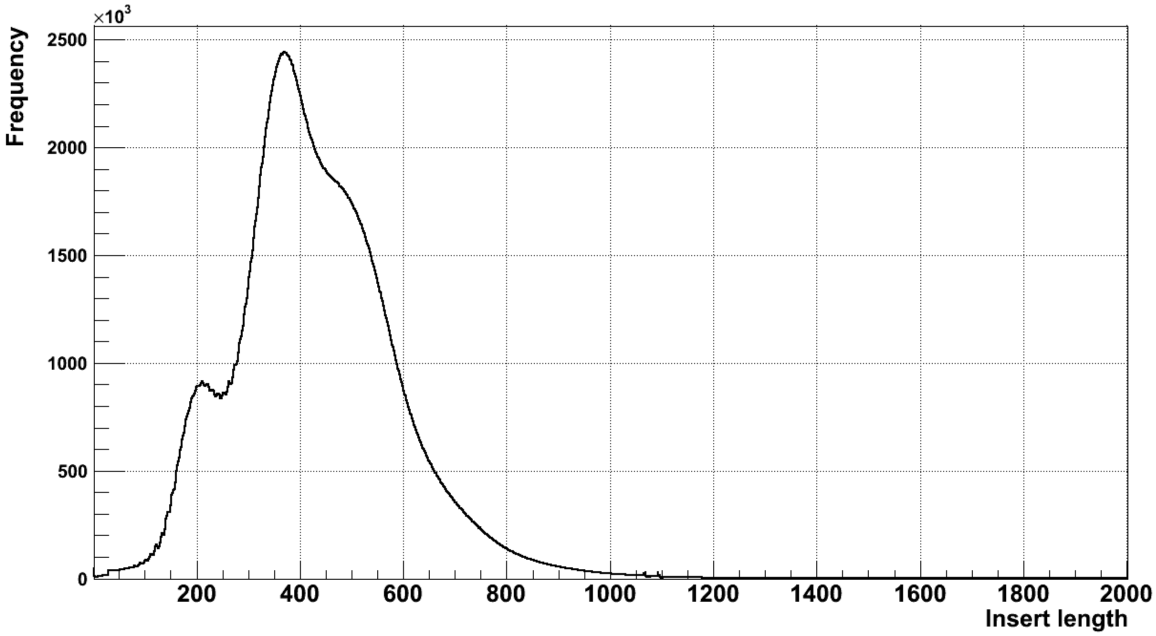


Figure S14. Distribution of insert lengths of read pairs in NA12878 Illumina HiSeq 2500 high-coverage data with 250 bp reads. The majority of read pairs significantly overlap at 3'-ends.

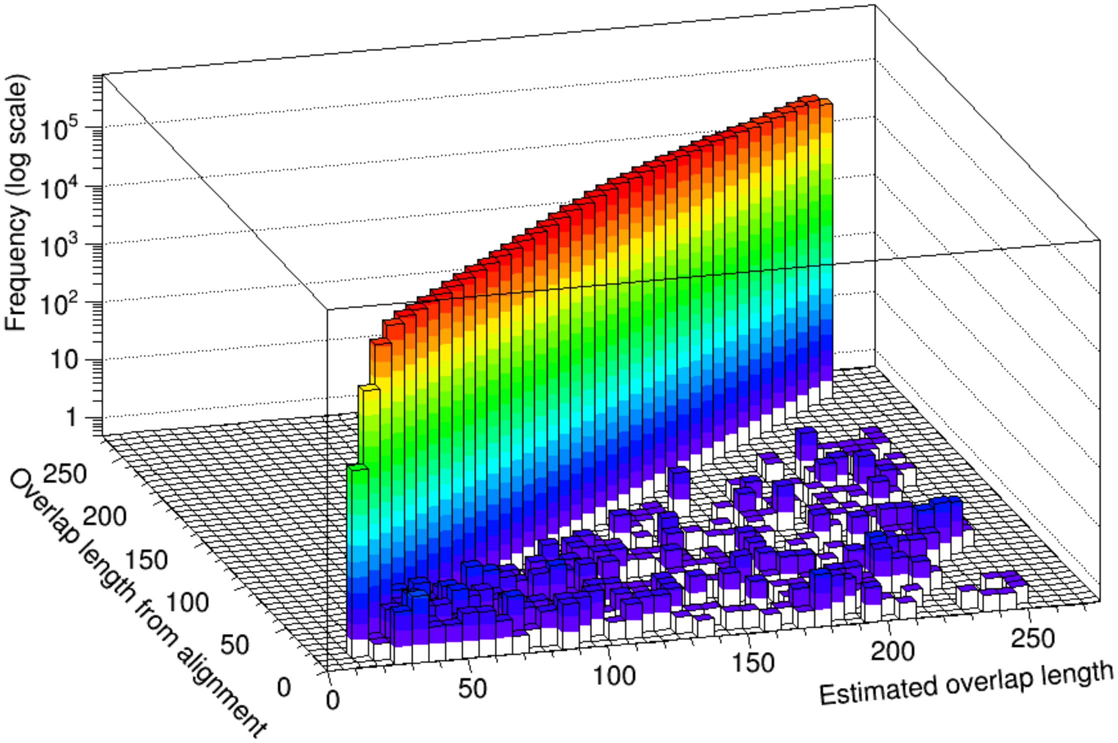


Figure S15. Comparison of overlap for reads in the same pair. Two estimates are: i) from sliding reads' 3'-ends against each other; ii) and from independent alignment of reads to the reference genome.

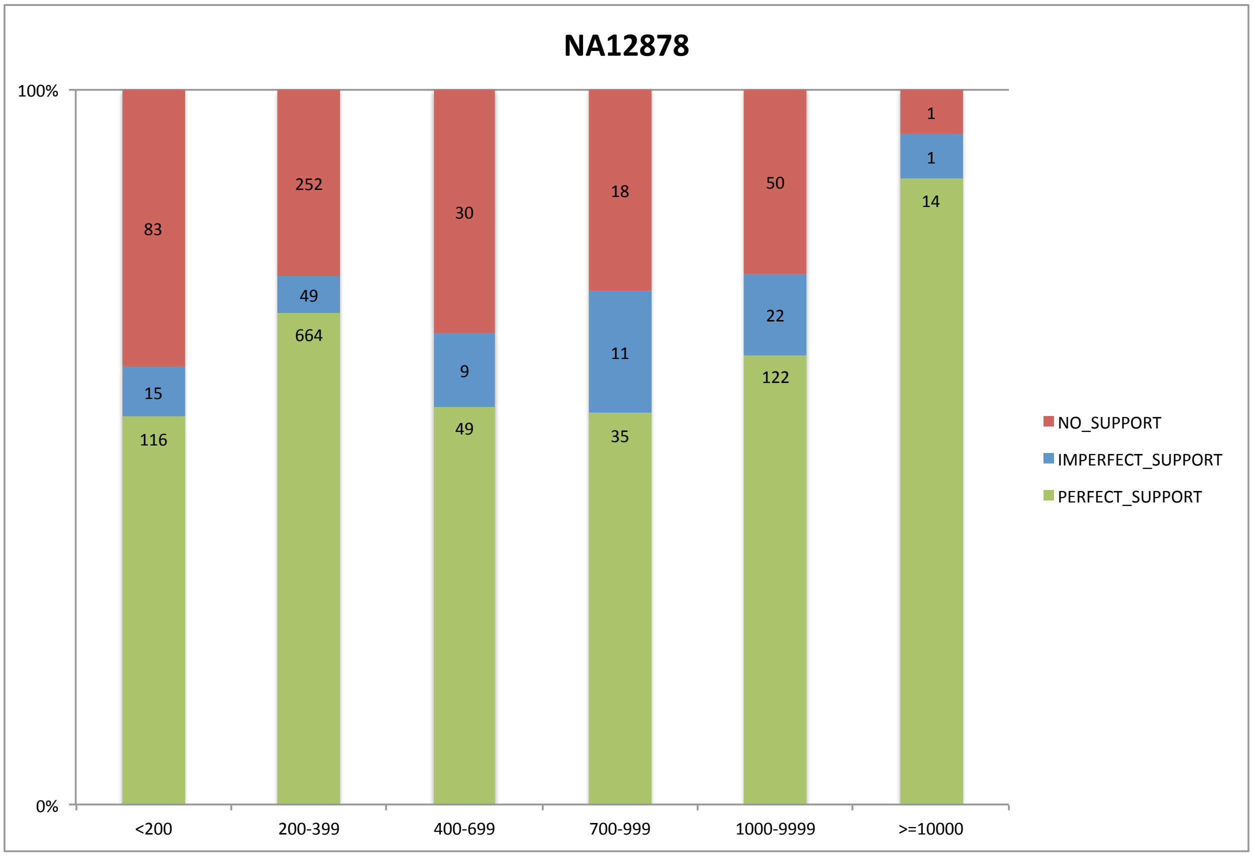


Figure S16. Confirmation rate for deletion breakpoints in high coverage individuals. Confirmation across different deletion lengths in NA12878 sample is shown. The confirmation rate decreases with size, reflecting possible genotyping error for small deletions.

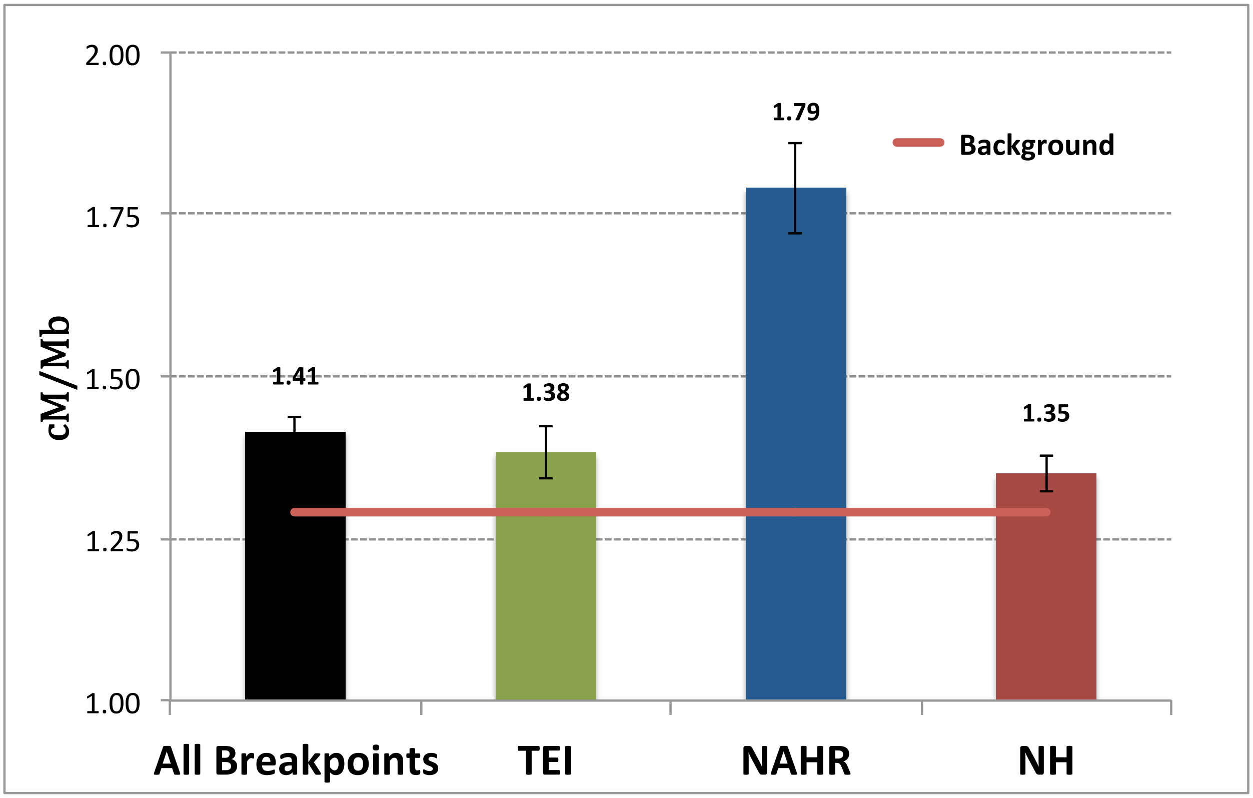


Figure S17. Association of breakpoints of different classes with recombination rates across genome.

Supplementary reference

1. Lam, H. Y. K. *et al.* Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat Biotechnol* **28,** 47–55 (2010).
2. Kidd, J. M. *et al.* A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell* **143,** 837–847 (2010).
3. Conrad, D. F. *et al.* Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. *Nat Genet* **42,** 385–391 (2010).
4. Pang, A. W. C., Migita, O., MacDonald, J. R., Feuk, L. & Scherer, S. W. Mechanisms of formation of structural variation in a fully sequenced human genome. *Hum. Mutat.* **34,** 345–354 (2013).
5. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).