Table S2. 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

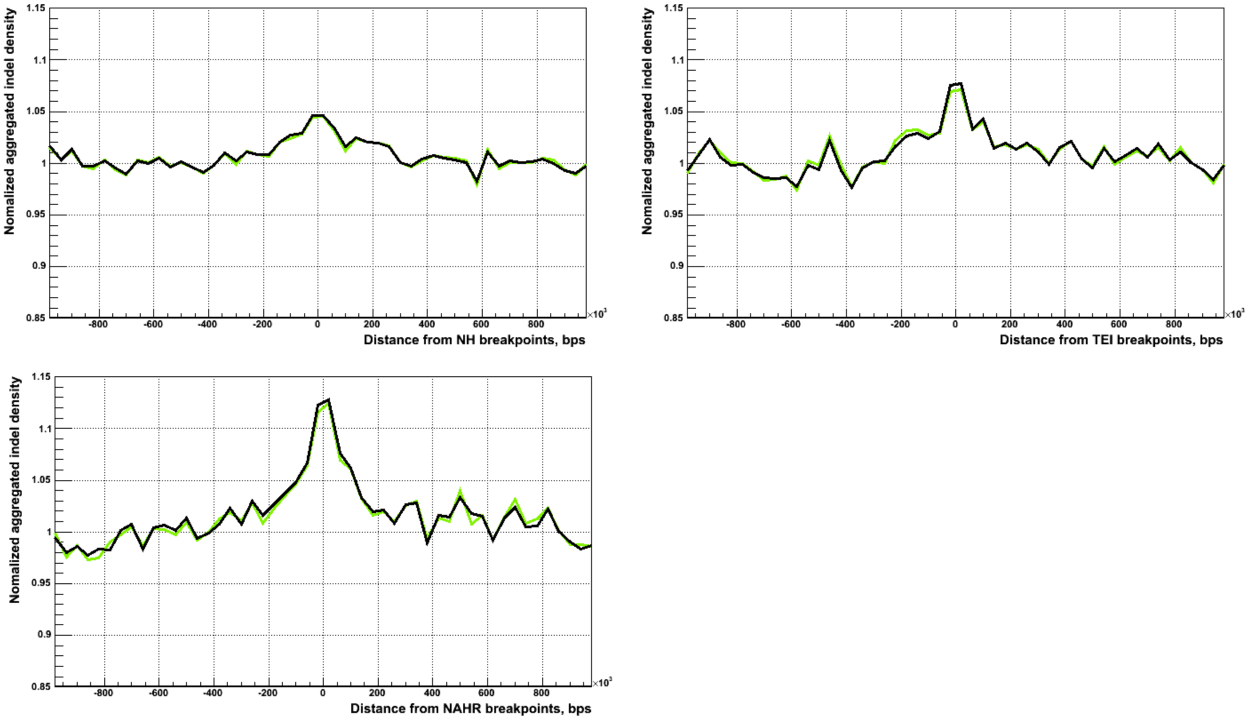


Figure S1. Indel aggregation around deletion breakpoints. Aggregation for indels of 1-6 bps in length is in black; aggregation for indels of 1 bp in length is in green.

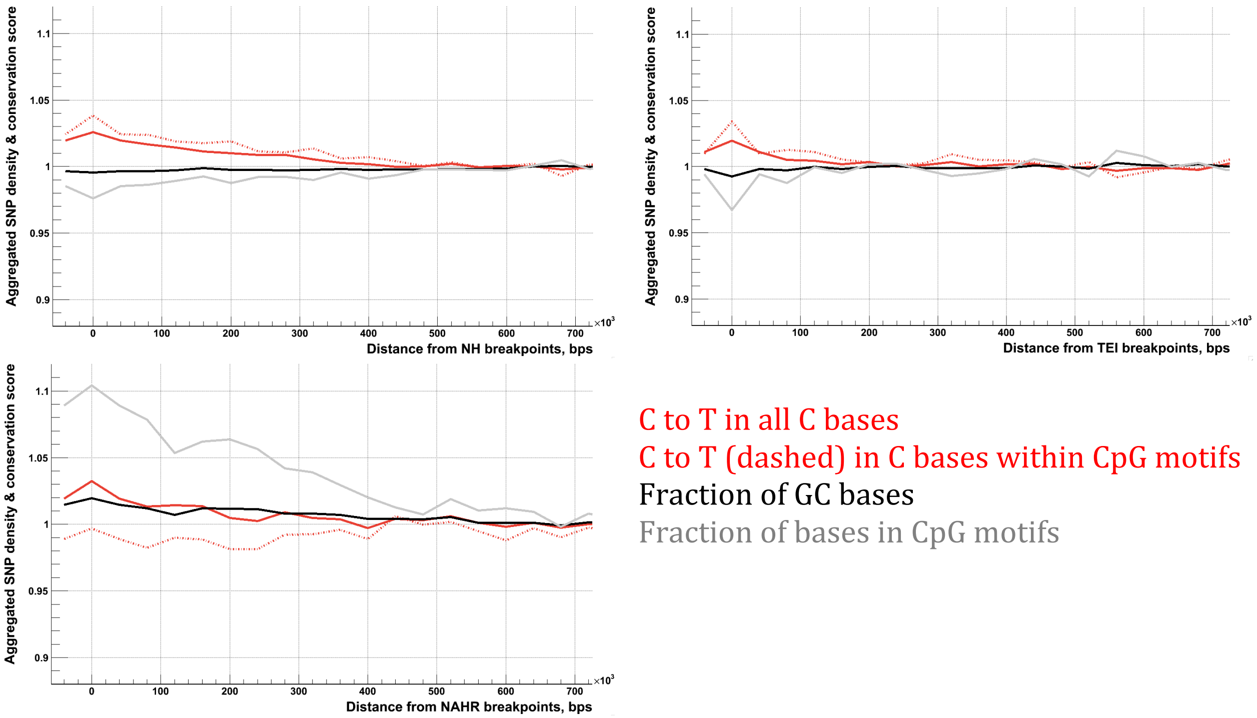


Figure S2. C to T mutations, GC content and CpG contents around breakpoints of different classes. All curves are normalized to unity at tails. Only unmasked bases, i.e. those where the 1000 Genomes Project can do confident SNP calling, were used in the analysis. NAHR breakpoints show very different distributions 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.

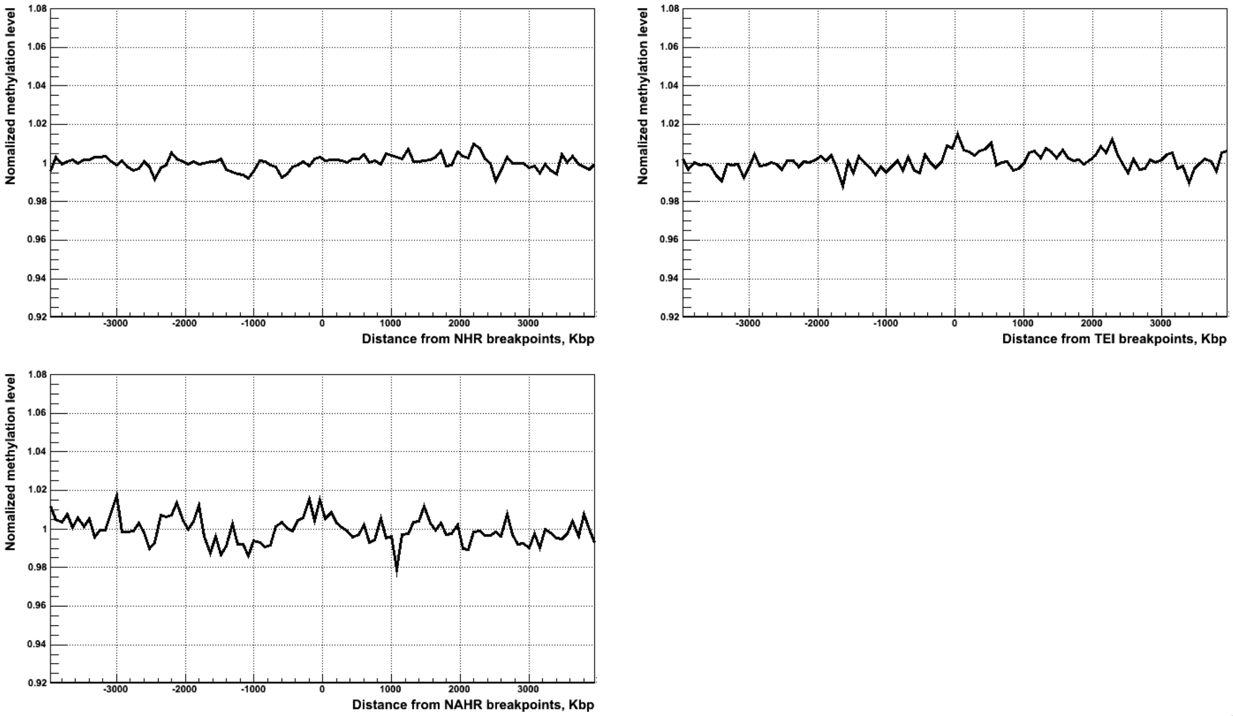


Figure S3. Methylation levels in H1ESC cell line around breakpoints of different classes. There is no noticeable change in methylation level around breakpoints of either class. On a smaller scale we do observed increase in methylation level in the regions of about 1 kbp around breakpoints of each type (data not shown). Though, this could 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.

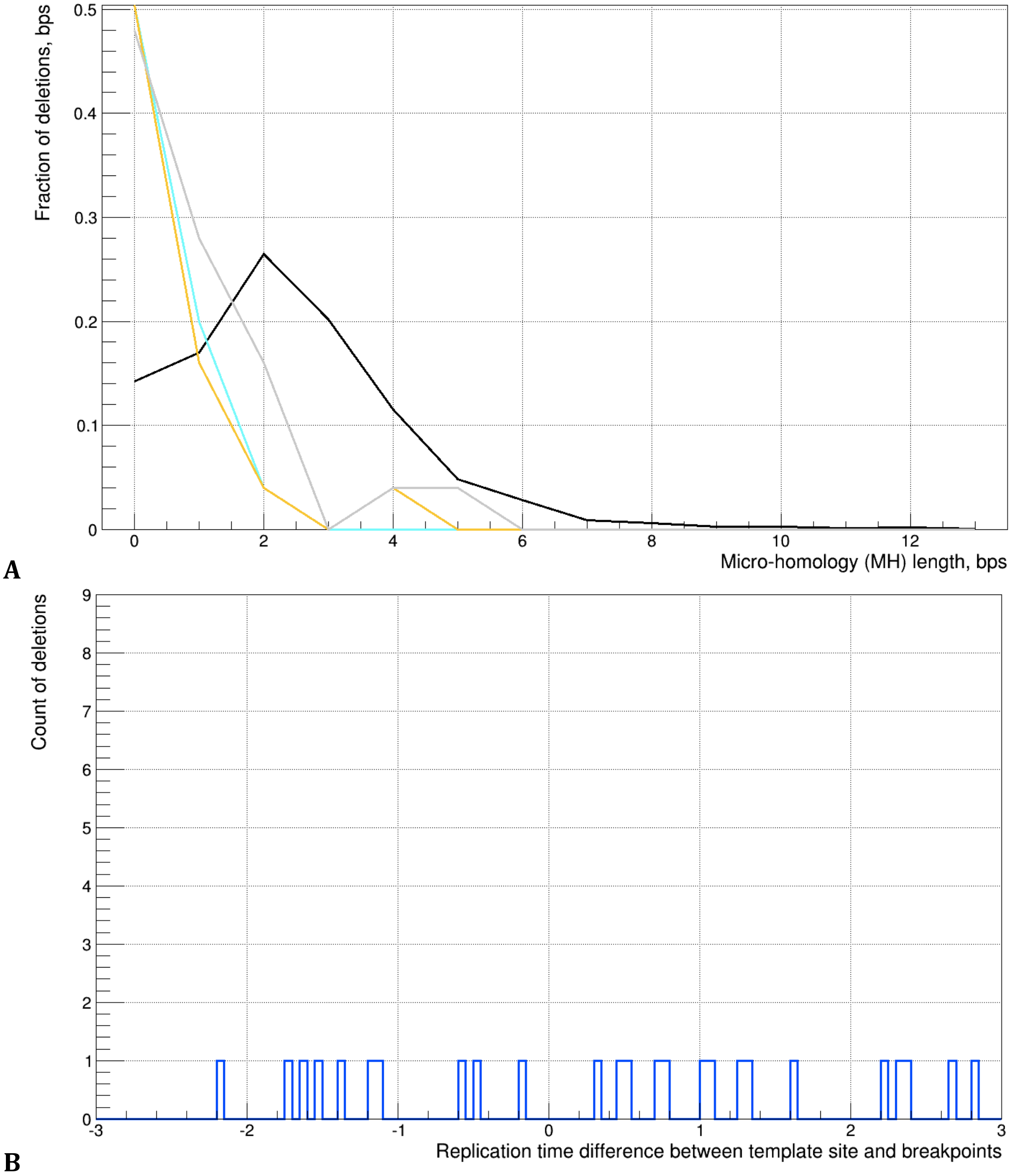
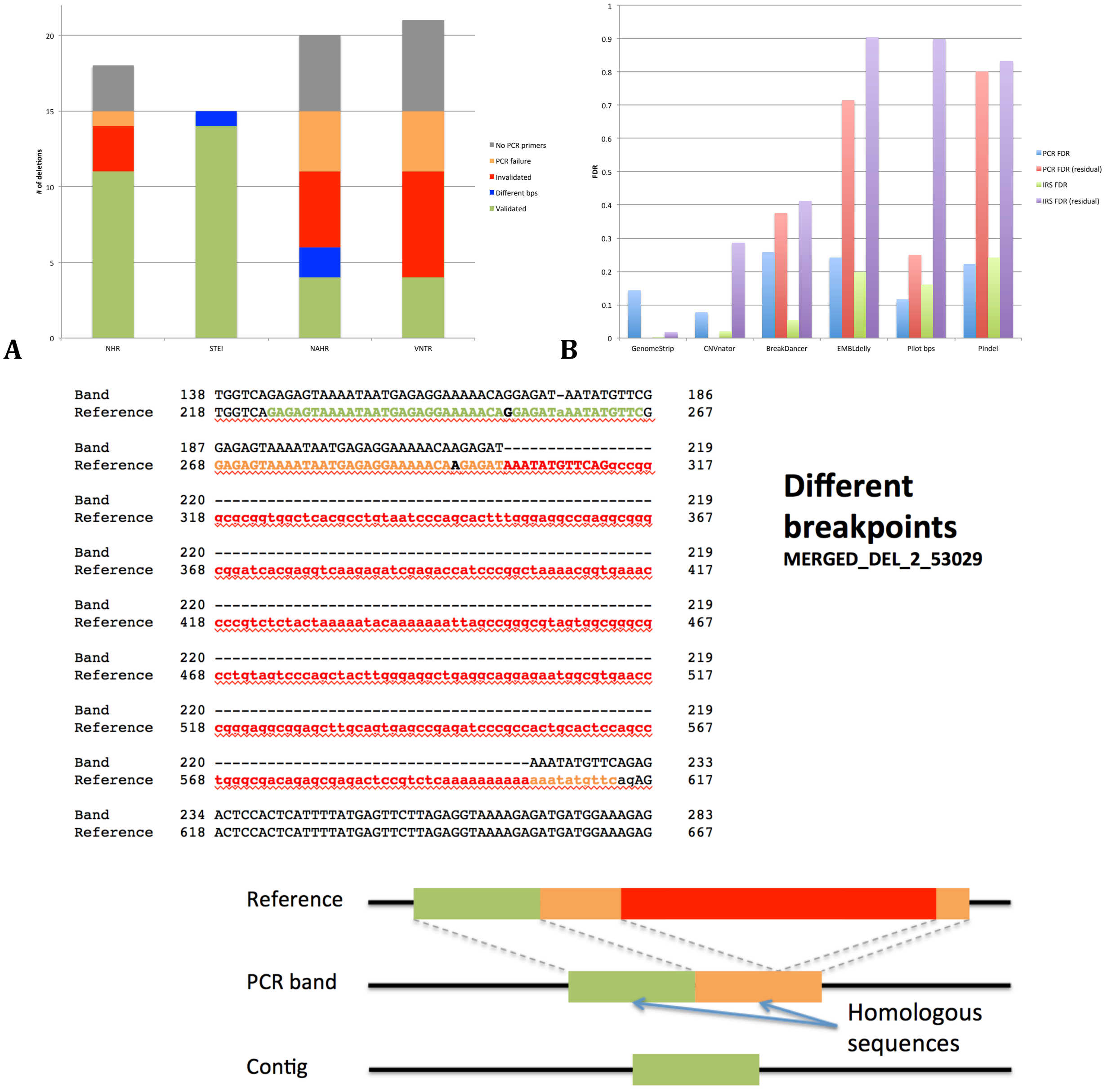


Figure S4. Analysis of micro-insertions (MI) with template sites on different chromosome. A) Length of micro-homology (MH) at deletion junction is not different from random (see also **Fig. 4**). For deletions with MIs and identified template site, MHs are calculated for 5'-ends/3'-ends of the deletion 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.



```
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 cccgtctctactaaaaatacaaaaaattagccgggcgtagtggcgggcg    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
```

Different breakpoints

MERGED_DEL_2_53029

C

Figure S5. 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) Example of different breakpoint from assembly and band sequencing.

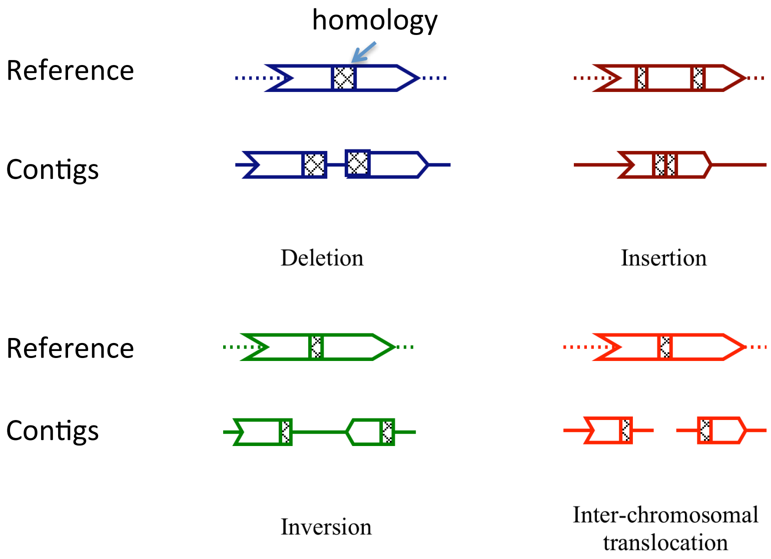


Figure S6. Examples of CROSSMATCH alignments to derive breakpoints of structural variations.

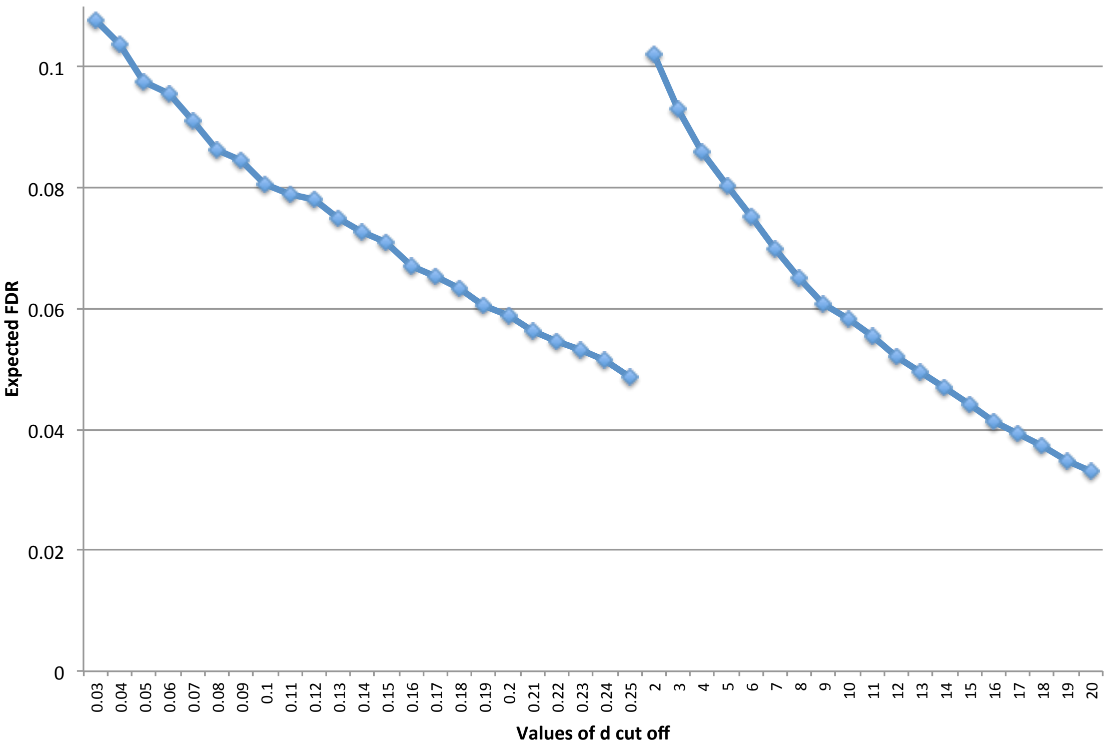


Figure S7. *In-silico* FDR for breakpoint support with the values of *d*. When realigning all unmapped reads to the null junction library and varied the value of *d* to compare the number of null junction passing the filter with the number of real junctions passing the filter. The cutoff *d* is defined 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.

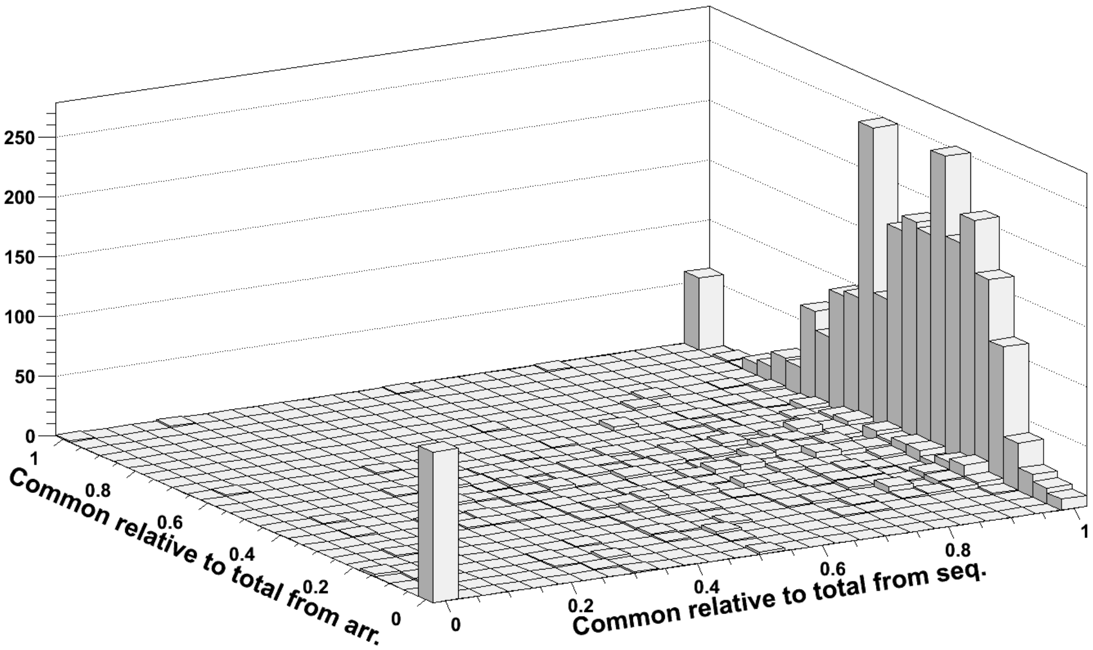


Figure S8. Correlation of samples genotyped as having deletion from OMNI SNP genotyping array (y-axis) and from mapping reads to sequences of breakpoint junctions (x-axis). Values on x/y axis is the fraction of samples with deletion common between the two ways of genotypes divided by the number of samples genotypes as having deletion by read mapping/by OMNI SNP array. Number of deletions with such fractions is on z-axi.

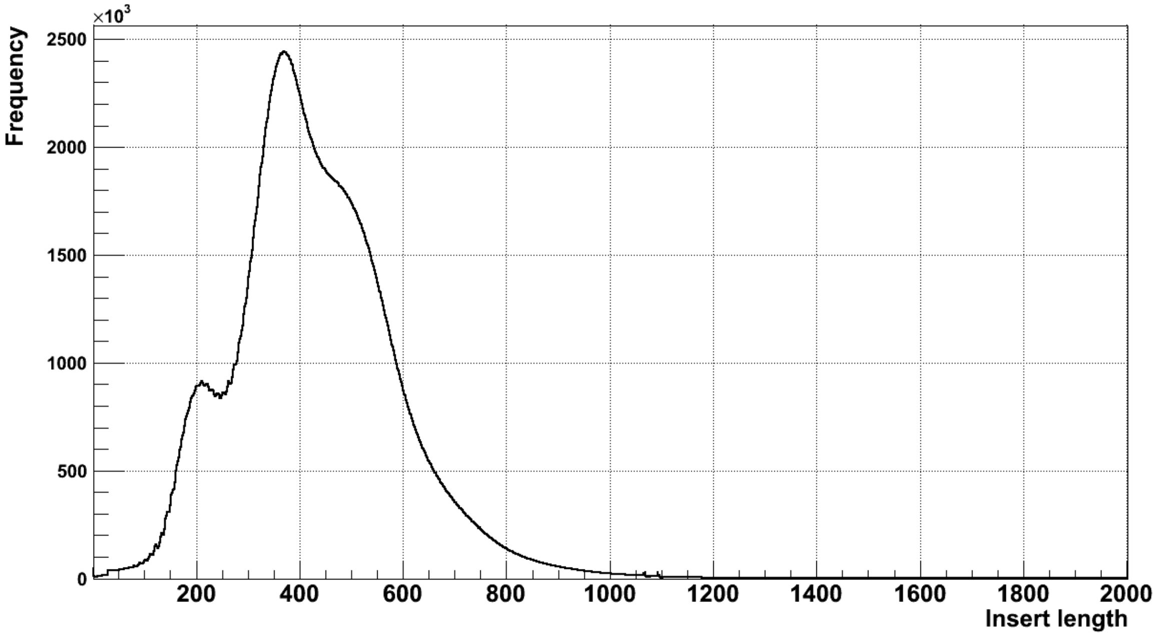


Figure S9. Distribution of insert lengths of read pairs in NA12878 Illumina HiSeq 2500 high-coverage data with 250 bp reads. The majority of read pairs significantly overlap at 3'-ends.

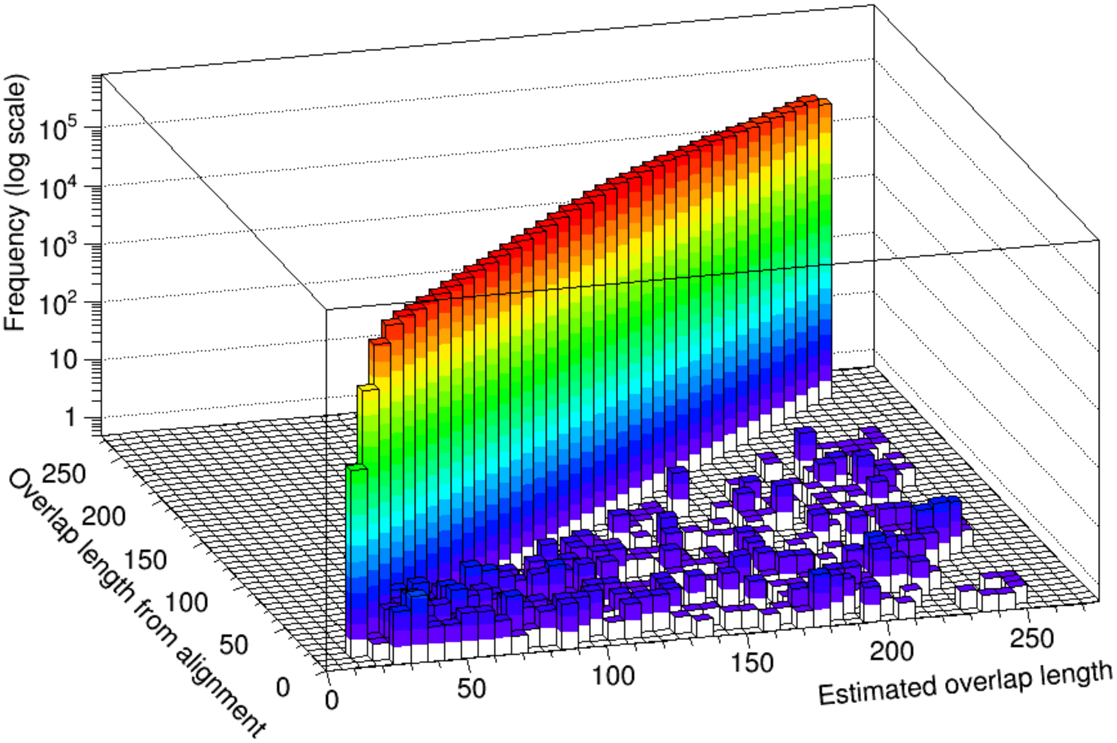


Figure S10. Comparison of overlap for reads in the same pair. Two estimates are: i) from sliding reads' 3'-ends against each other; ii) and from independent alignment of reads to the reference genome.

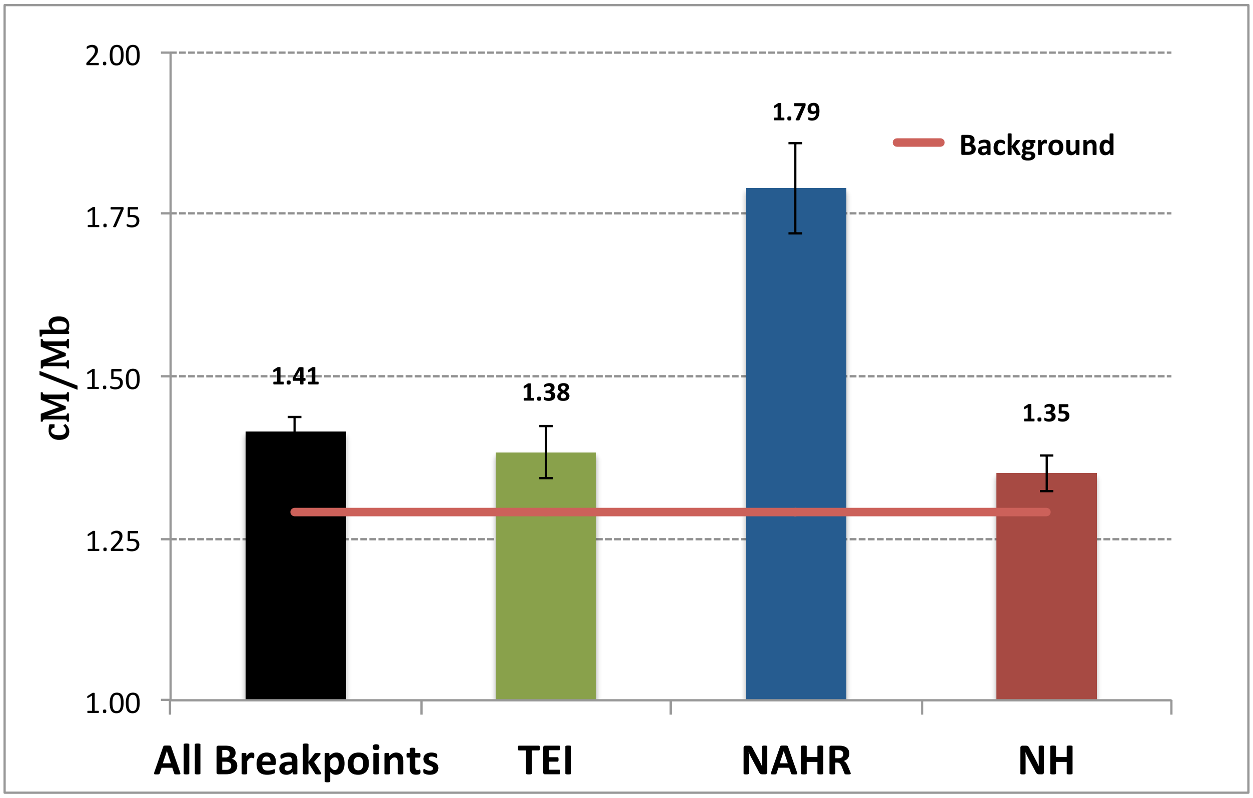


Figure S11. Association of breakpoints of different classes with recombination rates across genome.