Analysis of genome structural variation breakpoints from 1,092 humans revealed details of mutation mechanisms

Shantao Li, Daniel Rhee Kim, Arif Harmanci, Adrian Stuetz, Matthew Hurles, Charles Lee, Jan Korbel, Ken Chen, Alexej Abyzov, Mark Gerstein

Abstract

We have discovered, validated, and analyzed a representative set of 8,943 breakpoints of deletions relative to the reference genome in 1,092 samples sequenced by the 1000 Genomes Project. Using sequence feature at breakpoints we characterized the deletions into likely mechanisms of origin: non-allele homologous recombination (NAHR), transposable element insertion (TEI), and non-homologous (NH) mechanisms. Deletions in each class exhibit pronounced and significant increase in the normalized SNP and indel density around their breakpoints and this is likely to be explained by relaxed selection acting on those regions as their evolutionary conservation is also reduced. Density of all different substitutions increased close to TEI and NH breakpoints. However, for NAHR breakpoints, we observed both increase and decrease for densities of different mutations, e.g., increase for C>T and depletion for C>A densities associated with increase in CpG di-nucleotide motifs. Furthermore, association of NAHR event with active genomic regions, open chromatin, and early replication timing suggests that large fraction of these events is of pre-replication origin. For NH breakpoints containing extra sequence at a junction and identifiable template location for the extra sequence, we observed distinct profile of the template origin and its later replication timing relative to breakpoints. These observations are consistent with NH breakpoints being generated by template switching mechanism during replication.

Results

Deriving the confident set of breakpoints

We performed comprehensive deletions discovery {REF PHASE1}, targeted breakpoint assembly {REF TIGRA-SV}, careful breakpoint mapping {REF AGE CROSSMATCH}, stringent filtering **(Fig. 1)**, and experimental validation (see **Methods**). For filtering we utilized unmapped reads and empirical null model **(Fig. 1C)**. Briefly, the model used inner sequences adjacent to deletion breakpoints to construct junctions simulating random sequence, i.e., null sequence junctions. Note, this model imitates sequence homologies around breakpoints. We realigned unmapped reads to real and null junctions and optimized criteria to consider a reads supporting a junction by interrogating alignments to null junctions, as such alignments represent random noise. Alignment of read to real junctions ensures continuity of flanking and inserted (if any) sequences at breakpoints. For the resulting set we performed Intensity Rank Sum (IRS) test {REF PILOT} and PCR amplification across breakpoints as a validation exercise. We further performed ad-hoc filtering of deletions to reduce systematic false positives arising during calling, assembly, and filtering from read mis-mapping. In particular, we did not include deletions having breakpoint signature of variable tandem repeats in the final set. The final set consisted of 8,943 deletion breakpoints with consistent FDR estimates from PCR and IRS tests, i.e., of 6.8% for deletion presence from PCR, 13.7% for deletion presence with correct breakpoints from PCR, and 6.4% for deletion presence from IRS test. We have further confirmed XXX% of the breakpoint sequences with OMNI SNP genotyping array, and YYY% of breakpoint sequences in trios with long read high coverage data (Table S1).

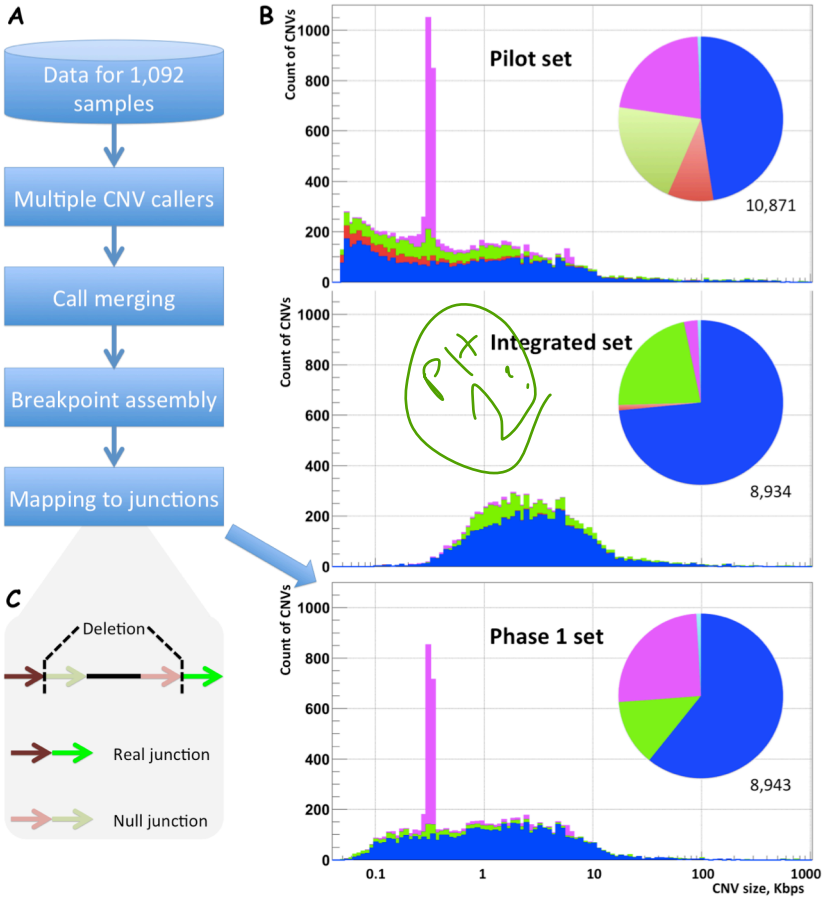


Figure 1. Deriving confident set of breakpoints. A) Conceptual steps for the derivation. Breakpoints from local target assembly are filtered by mapping reads to putative junctions. C) Null model for breakpoint filtering. B) Comparison of different breakpoint sets. Note, the pilot set {REF} was included in the derivation as one of the call sets. Integrated set {REF} was bias towards large non-repetitive deletions for the purpose of reliable genotyping, resulting in mobile element insertions being strongly under represented.

Overall, these breakpoints are of higher quality than those derived in the pilot phase of the 1000 Genome Project {REF PILOT} and is of better representation then the one used recently by the project {REF PHASE1}, as it was limited to large non-repetitive events that could be well genotyped. By using BREAKSEQ software {REF BREAKSEQ}, we further performed classification of the deletions by the likely mechanisms of their origin using sequence signatures at breakpoints: non-alleles homologues recombination (NAHR), transposable element insertions (TEI), non-homologous (NH) events. Note, our set consists of deletions relative to the reference genome but the final set does contain bona fide

insertions of transposable elements {REF BREAKSEQ}. The final set contained 13% of NAHR deletions, 25% of TEIs, and 61% of NH deletions. It should be noted that NAHR and TEI events are more difficult to discover as having repeats at breakpoints and in deleted regions, thus, our set is still likely to still under represent those events.

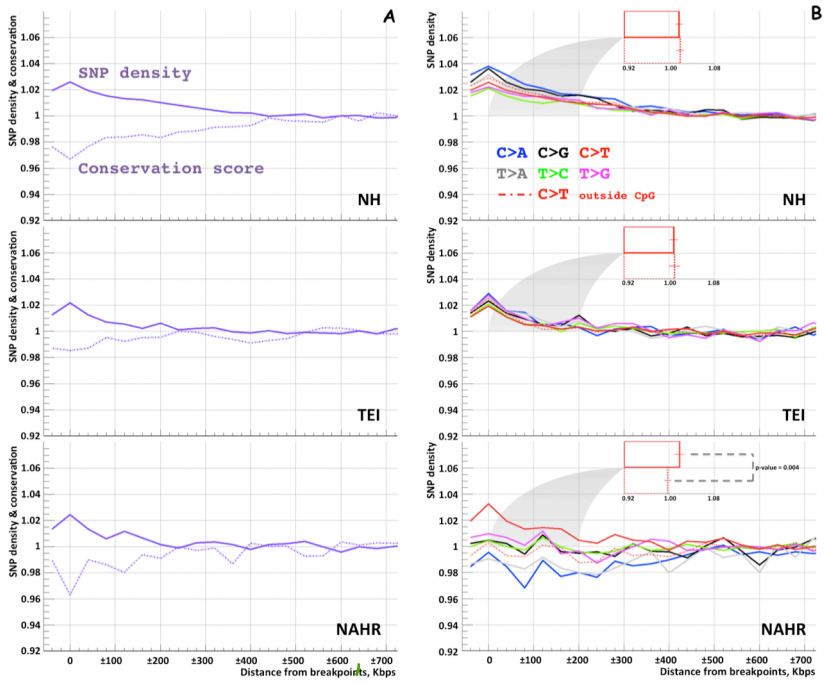


Figure 2. Co-aggregation of SNPs and deletion breakpoints found in the analyzed samples. A) Normalized SNP densities increased while conservation decreased in 400 kbps regions around breakpoints of each class. B) Densities increase for substitutions of all types around NH and TEI breakpoints. Increase of C>T substitutions around NAHR breakpoints is explained by enrichment of CpG motifs.

Variant co-aggregation with deletion breakpoints

To analyze association of variants with deletion breakpoints we aggregated SNPs and indels found in the same population around the breakpoints. We used variants that are only in the confident sites as defined by mask of the 1000 Genomes Project {REF} and calculated density with respect to the number of such sites. Normalized densities (see **Methods**) of both SNPs and indels increased in 400 kbps regions around breakpoints of each class **(Fig. 1A and S?)**. This is likely to be explained by co-occurrence of different variants in genomic regions with reduced selection, as aggregated conservation score around breakpoints decreases on par with the increase in SNP density. Besides total SNP density, densities of individual substitution types also increase close to NH and TEI breakpoint **(Table S?)**. However, it is not the case for NAHR breakpoints, for which C>T substitutions are enriched while T>A and C>A are depleted **(Fig. 1B; Table S?)**. Further analysis revealed that increase in C>T is due to enrichment of CpG motif exclusively around NAHR breakpoints, i.e., not around NH or TEI breakpoints. These motif is known to be highly mutable and, particularly, for C>T substitutions when methylated. Thus, this analysis revealed potential association of NAHR with regions of methylation.

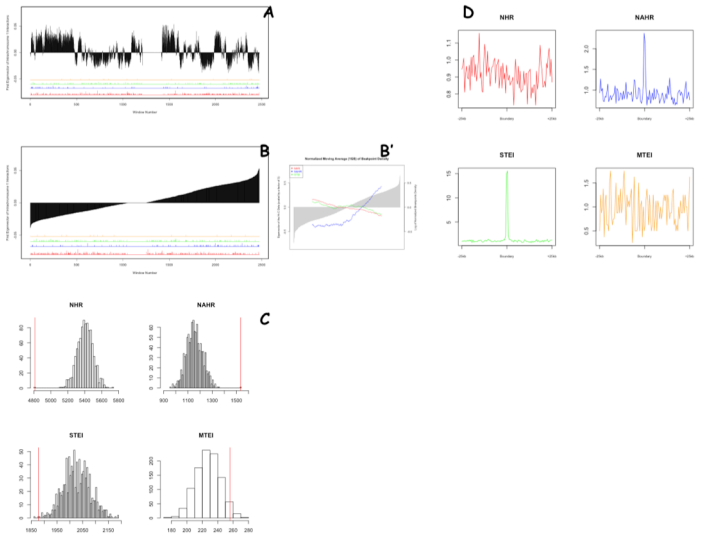


Figure 3.

Association of breakpoints with chromatin states and active regions

We used two state of chromosome interactome as defined by Hi-C experiment {REF HI-C} and roughly corresponding to packed/unpacked chromatin, to investigate for any correlation of breakpoints with DNA open and active chromatin. We tested for the occurrence of breakpoints in genomic bins of XXX bps assigned to either state. To determine the significance of our findings we circularly permuted breakpoints

along the genome, thus preserving their relative arrangement but randomizing their position relative to bins, to simulate random occurrences (see **Methods**). We observed that both NH and TEI breakpoints are depleted for open chromatin while NAHR breakpoints are enriched **(Fig. 3)**. We had previously observed {REF FIG} that NAHR breakpoints are associated with chromatin marks and this observation is confirmed with the new set of breakpoints derived in this study **(Fig. S?)**. Similarly, previously observed {REF FIG} enrichment of NAHR with enhancers was replicated in this study (p-value=PPP) on a larger set of YYY enhancers {REF FUNSEQ} (see **Methods**).

Change in expression of nearby genes? Arif's results.

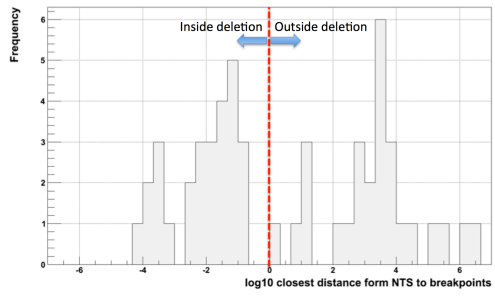
To elaborate on these results we analyzed the association of DNA methylation {REF} with breakpoints of each class, and strong association was observed for TEI and NAHR breakpoints **(Fig. 3D)**. In particular, the methylation was 15 times higher (p-value=PPP) than background around TEI breakpoints and 2.5 times higher (p-value=PPP) than background around NAHR breakpoints. Methylation of transposable elements is expected, as this a way for a cell to silence their activity {REF}. The potential relevance methylation of NAHR breakpoints will be discussed below.

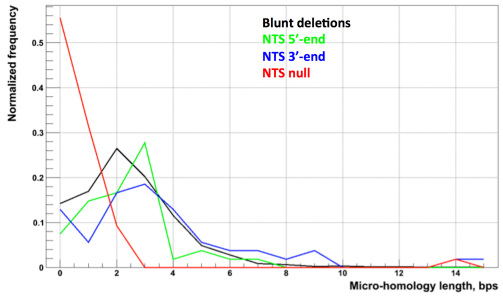
Micro-insertion at breakpoint deletions and relation with replication timing

Multiple studies have reported existence of micro-inserted sequences at deletion breakpoint. In our dataset we observed 2391 (27%) deletions with micro-insertions ranging in length from 1 to 96 bps with majority of only few bps in length **(Fig. 4A)**. Replication based mechanisms were suggested to generate deletions with micro insertions that are copies of some sequence in the genome. To test for this possibility we semi-manually determined the likely genomic origin, i.e., template site, of 133 inserted sequences of which 132 were 15 bps or longer, 30% of all micro-insertions of such length. Other micro-insertions did not map to the reference genome, mapped only partially, or mapped to multiple locations. In agreement with previous finding {REF Conrad Kidd} mapped micro-insertions were observed almost exclusively (83%) for NH events and their template site was frequently, in 108 or 81% of cases, was located on the same chromosome as the deletion. The distribution of the nearest distance between template site and either of the breakpoints revealed their relative preferred arrangement **(Fig. 4B)**. The template site was located either within 100 bps or in the range from 1 to 10 kbps of one of the breakpoints. Interesting that proximal template sites typically occur within the deleted sequence and, perhaps, can be explained by co-occurrence of two indels (detected as a one deletion) or indel with a deletion. In other words, micro-insert is genomic sequence between two proximal variants. However, the other peak in the distribution could signify details of the mechanisms leading to generation of micro-insertions. We hypothesize that this length could be related to DNA packing in the cell or to the length of DNA to wrap around the replication bubble. To investigate this further we compared replication timings of breakpoints and template sites. Describe micro-homology **(Fig. 4C)**.

It was previously noted {REF Koren} that breakpoints of deletions generated by different mechanism different association with replication timing. We confirm those observations: NAHR deletions are typically occur during early replication, HN events tend to occur at later replication while TEIs show now relation to replication.

A Micro insertion distribution





B

C. Correlation with replication timing

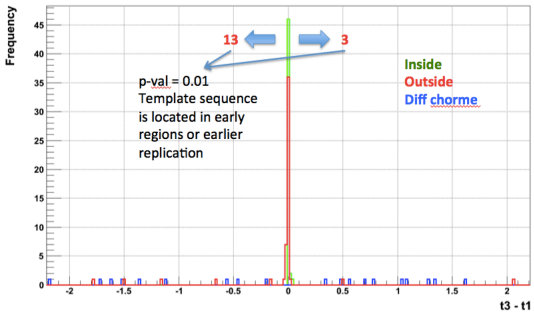


Figure 4.

Discussion

- We provide large, less biased, and high quality dataset of breakpoints
- They aggregate with SNPs and indels. Perhaps, expected.
- Hypothesis about ssDNA in relation to NAHR event and C>T mutations
- Insight into template switching from mapping inserted sequence and correlation timing.

We have advanced science beyond imaginable.

Methods

Discovery and merging

Deletions discovered by five CNV callers {REF} were merged with the set of breakpoints discovered in about 180 pilot samples of the 1000 Genomes Project {REF}. For the resulting set we assembled TIGRA-SV {REF},
OMINI 2.5s overlap and genotype concordance (genotype from OMNI vs genotype from mapping to junctions, this is additional prove of the approach by mapping to junctions)
Comparison with Pilot and Integrated

Aggregation calculation

Intersection with open/closed chromatin

Circular breakpoint permutation to calculate p-value

Intersection with enhancers

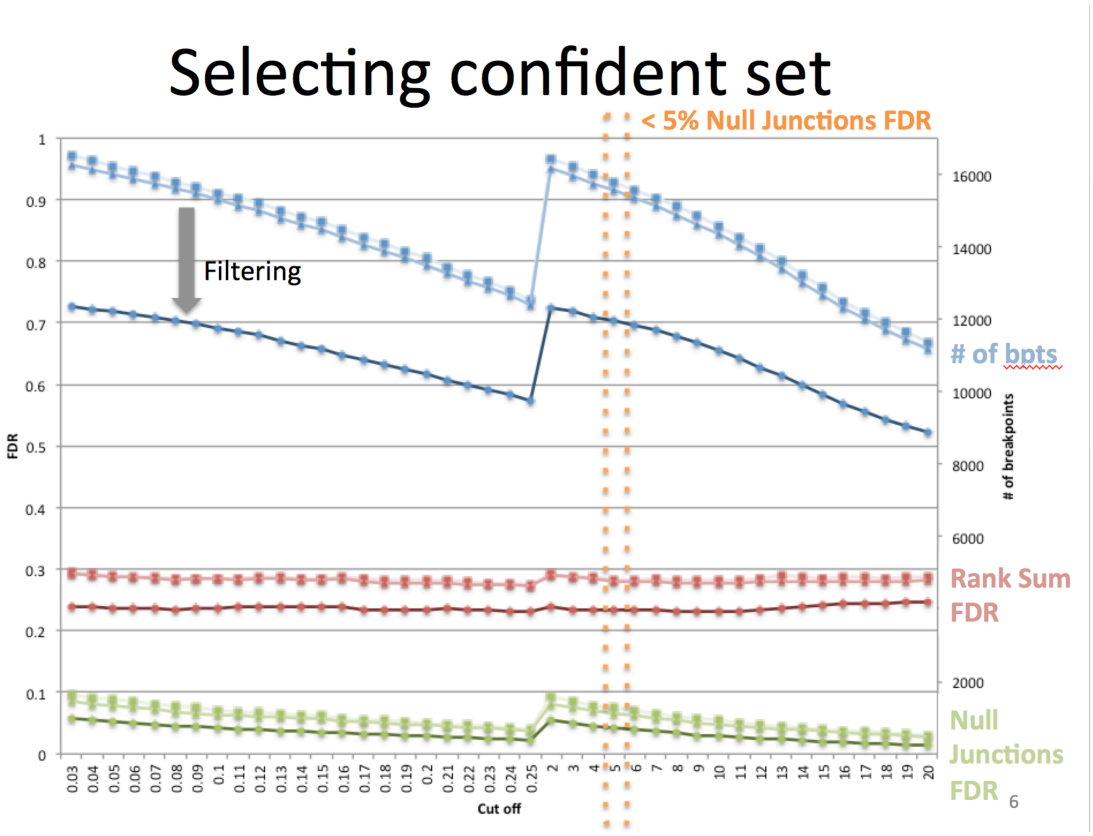


Figure S1

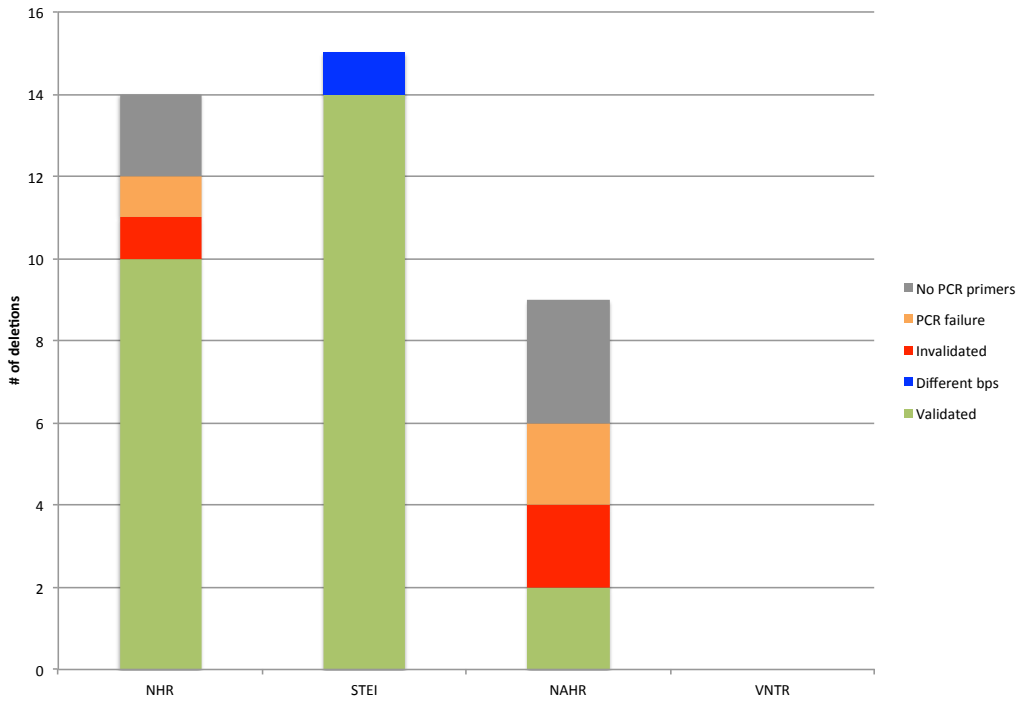


Figure S2

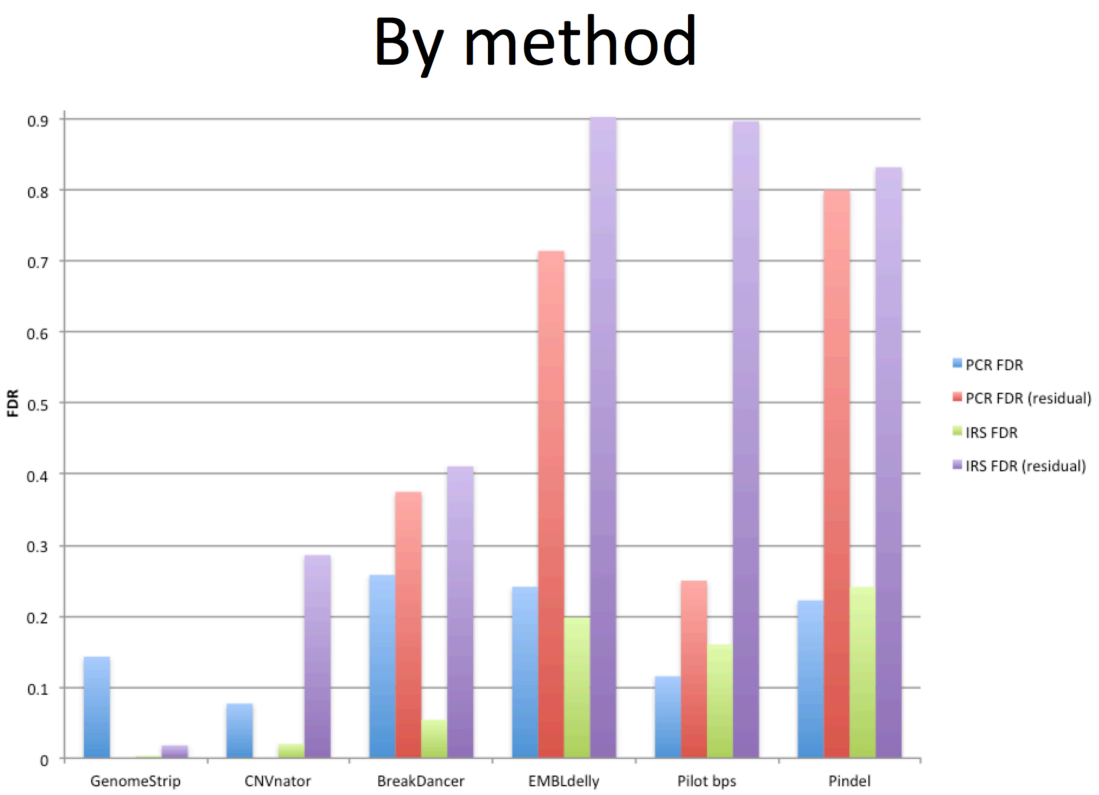


Figure S3

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

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

Band       220 ------------------------------------------------   219
Reference  318 gcgcagtggctcacacctgtaatcccagcactttgggaggccgaggcggg   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
```

Different breakpoints
MERGED_DEL_2_53029

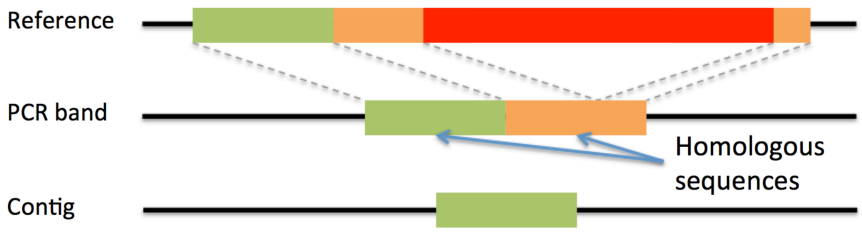


Figure S4

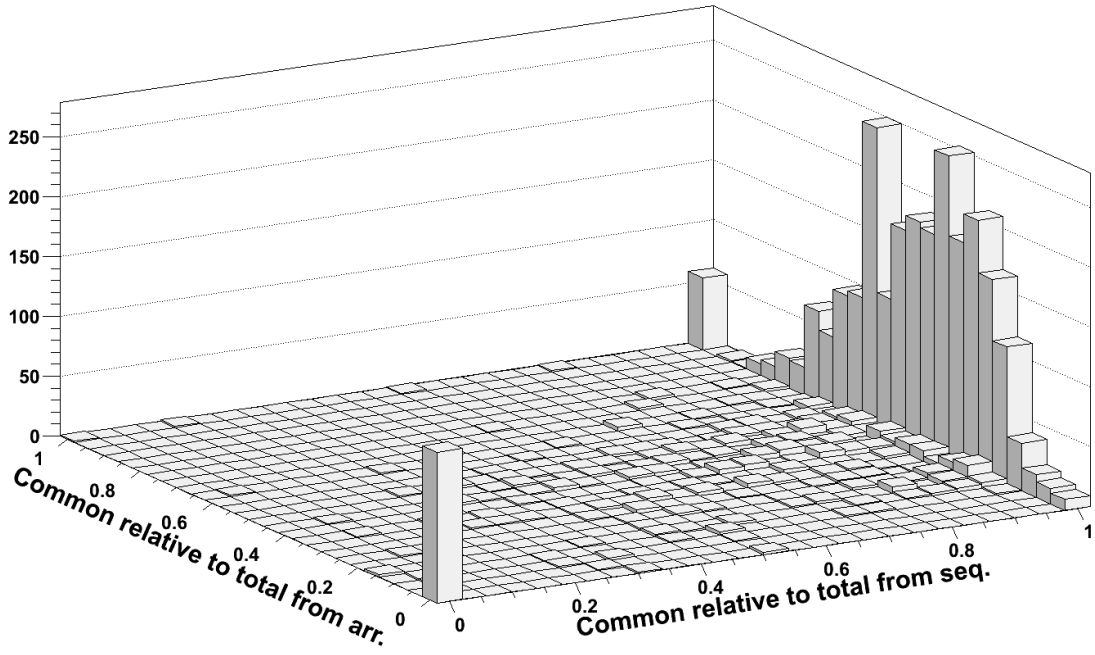


Figure S5

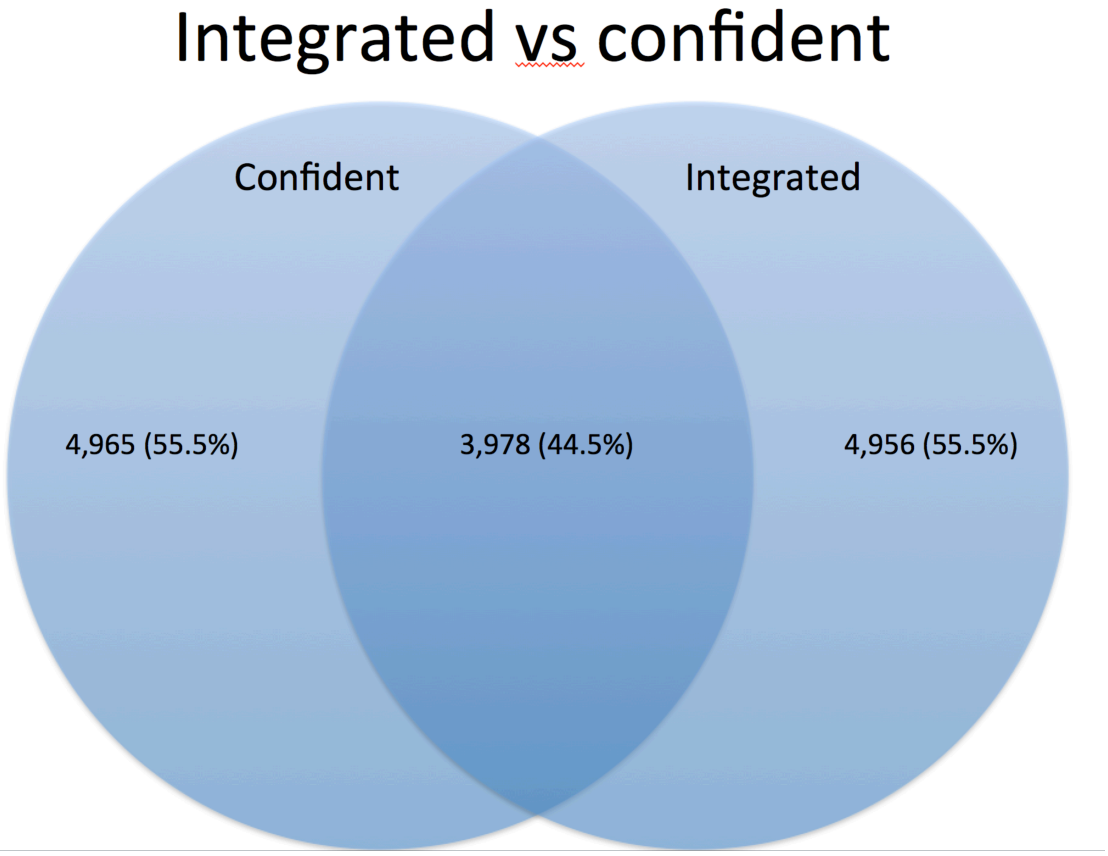


Figure S6. Make it 3-way Venn diagram with Pilot.

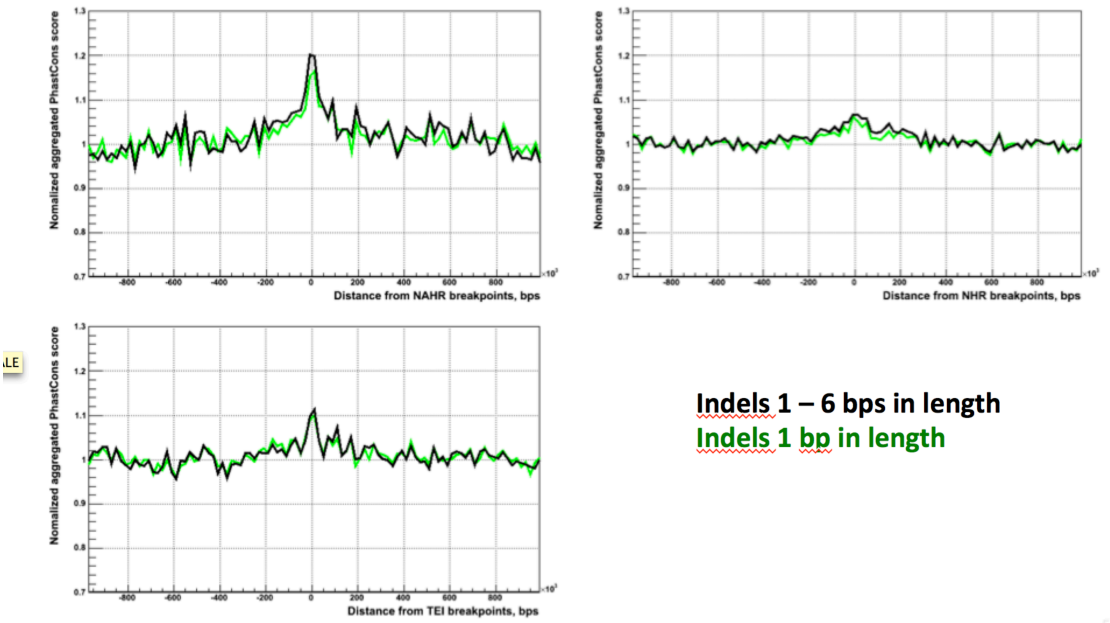


Figure S7.

SNP aggregation at small scale

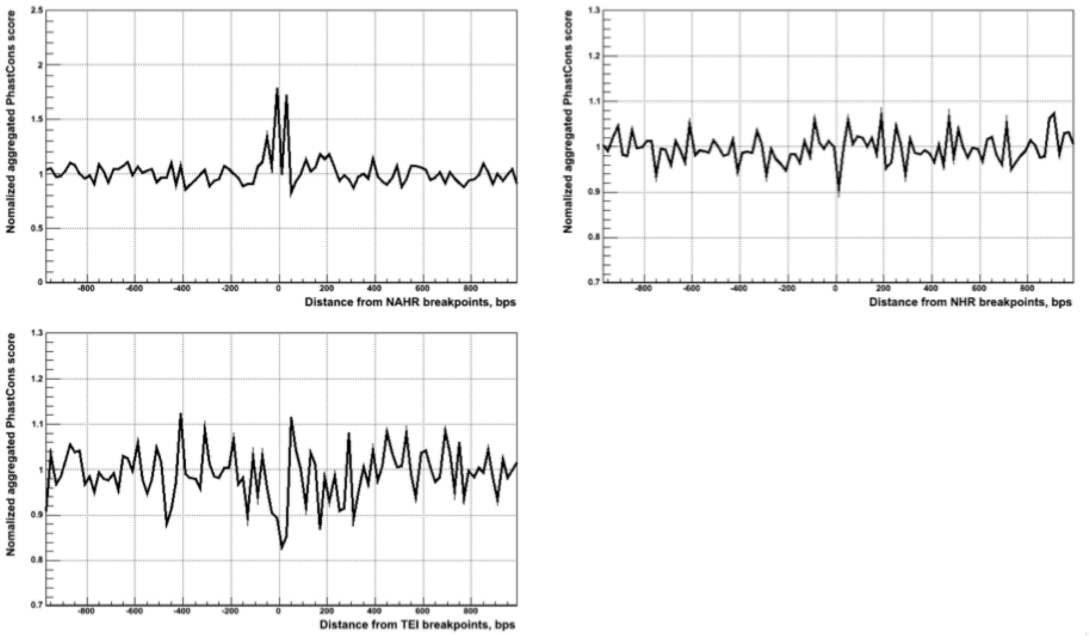


Figure S8.

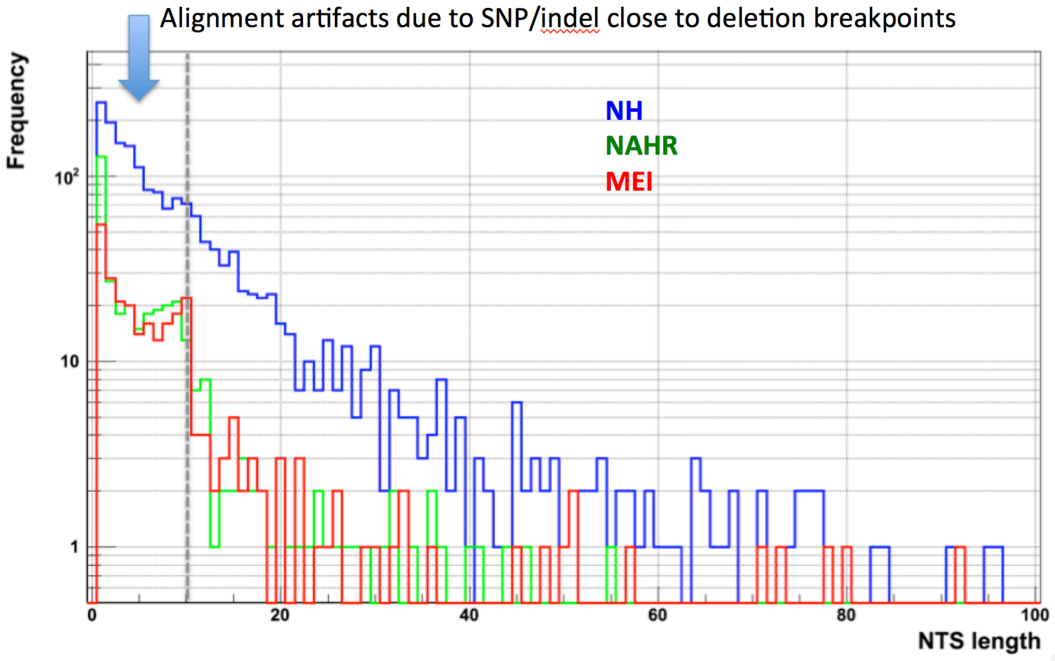


Figure S9, Work on redefining breakpoints and removing accumulation at < 10.