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

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

Abstract

Analysis of genome structural variations (SV) at breakpoint resolution is fundamental to understanding the mutational processes generating them. Previous analyses of SV breakpoints were limited by the size distribution of the analyzed sets and/or confidence in breakpoint resolution. Here we described the discovery and analysis of the largest to date and representative compendium of 8,943 confident breakpoints of deletions relative to the reference genome in 1,092 samples sequenced by the 1000 Genomes Project. Using sequence features at breakpoints we classified the deletions into likely mechanisms of origin: non-allelic homologous recombination (NAHR), transposable element insertion (TEI), and non-homologous (NH) mechanisms. SVs in each class exhibit pronounced and significant increases in SNP and indel densities around their breakpoints. This is likely explained by relaxed selection acting on those regions as their evolutionary conservation is also reduced. Density of all substitution types increases close to TEI and NH breakpoints. However, for NAHR breakpoints, we observed both an increase and a decrease in densities of different substitutions, e.g., an increase in C to T and a depletion of C to A substitution densities, associated with an increase in CpG di-nucleotide motifs. Classical NAHR mechanism postulates replication as a requirement for generating a SV. However, based on the association of NAHR SVs in our dataset with active genomic regions, an open chromatin state, dynamically methylated regions and early replication timing we hypothesize that some of these events originated in cells prior to replication. NH SVs with extra sequences at a junction often have identifiable template sites for the sequence, which are located at a rather well defined distance of 2-7 kbps from the breakpoints and replicate later than regions of breakpoints. This may be consistent with the existing hypothesis that a collapsed replication fork at a breakpoint has to wait for the unwinding of proximal DNA to switch template and restore replication.

Alexej Abyzov 4/11/14 5:23 PM
Deleted: ty

Alexej Abyzov 4/11/14 7:19 PM
Deleted: transitions

Alexej Abyzov 4/11/14 7:19 PM
Deleted: transversion

Alexej Abyzov 4/11/14 6:27 PM
Deleted: Furthermore

Alexej Abyzov 4/11/14 7:07 PM
Deleted: suggest

Results

Deriving the confident set of breakpoints

We performed comprehensive discovery of deletions {REF PHASE1}, targeted breakpoint assembly {REF TIGRA-SV}, breakpoint mapping with two pipelines {REF AGE CROSSMATCH}, stringent filtering **(Fig. 1A)**, and experimental validation (see **Methods**). Due to the often-inconsistent results aligning contigs and mapping breakpoints by different aligners and processing pipelines we applied stringent filtering to ensure the physical continuity of flanking and inserted (if any) sequences at breakpoints. For filtering we utilized unmapped reads and an empirical null model **(Fig. 1B)**. Briefly, the model used inner sequences adjacent to deletion breakpoints to construct junctions simulating random sequences, i.e., null sequence junctions. Note, this model imitates biologically relevant sequence homologies around breakpoints. We realigned unmapped reads to real and null junctions and optimized criteria to consider a read supporting a junction by interrogating alignments to null junctions, as such alignments represent random noise. Next, guided by validation, we performed ad-hoc filtering of deletions to reduce systematic false positives arising as a result of using split-read information during calling, assembly, and filtering. In particular, we did not include deletions having the breakpoint signature of variable tandem repeats in the final set. For validation we performed PCR amplification across breakpoints and tested for difference in intensity values for SNP probes across individual with and without deletions – Rank Sum (IRS) test {REF PILOT} (see **Methods**). The final set consisted of 8,943 deletion breakpoints with consistent FDR estimates from PCR and IRS validations, i.e., of 6.8% and 6.4% for deletion existence from PCR and IRS, respectively, and 13.7% for deletion presence with correct breakpoints from PCR. We have further confirmed XXX% of the breakpoint sequences with OMNI SNP genotyping array, and 39% of breakpoint sequences in trios with high coverage and long read data (Table S1). YRI and CEU trios were sequenced with recent sequencing technology and PCR-free library preparation to roughly 60X coverage of each individual genome. The medium insert size of fragment library was 400 bps with read length of 250 bps, meaning that for most fragments read 3-ends overlap for a considerable number of bases. Using such overlaps we reconstructed fragment sequences, realigned them around the breakpoints in our set and compared the resulting alignments with the breakpoints (see **Methods**).

Overall, these breakpoints are of higher quality than those derived in the pilot phase of the 1000 Genomes Project {REF PILOT} and are more representative in their length distribution than those used recently by the project {REF PHASE1}, as it was limited to large non-repetitive events that could be well genotyped (**Fig. 1C**). By using BREAKSEQ software {REF BREAKSEQ}, we performed further classification of the likely mechanisms of origin of the deletions using sequence signatures at breakpoints from the following classes: non-allelic homologous recombination (NAHR), transposable element insertions (TEI), and non-homologous (NH) events. Note, our set consists of deletions relative to the reference genome but it does contain bona fide insertions relative to ancestral start, such as transposable elements {REF BREAKSEQ}. The final set consisted of 13% NAHR, 25% TEIs, and 61% NH deletions. It should be noted that NAHR and TEI events are more difficult to discover as having repeats between and at breakpoints, thus, this set is still likely to under represent those events.

Variant co-aggregation with deletion breakpoints

To analyze the association of variants with deletion breakpoints we aggregated SNPs and indels found in the same group of individuals around the breakpoints. In order to reduce the contamination of our analysis with false positive calls we only used variants that reside in the confident sites as defined by the mask of the 1000 Genomes Project {REF} and calculated densities 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. 2A and S1**). Interestingly, the increase in indel density was highest for NAHR breakpoints (**Fig. S1**). Note, the scale of increases is vast relative to the 450-650 bp insert size of sequencing libraries. Therefore, the observed increases are not caused by false calls due to compromised mapping around SV breakpoints. Analysis of sequence conservation around breakpoints revealed that the

Alexej Abyzov 4/12/14 11:31 PM
Deleted:

Alexej Abyzov 4/11/14 5:25 PM
Deleted: discovery

Alexej Abyzov 4/12/14 7:00 PM
Moved down [1]:

Alexej Abyzov 4/12/14 7:00 PM
Moved (insertion) [1]

Alexej Abyzov 4/12/14 7:00 PM
Deleted: the resulting set

Alexej Abyzov 4/12/14 7:06 PM
Deleted: as validation exercises

Alexej Abyzov 4/12/14 7:07 PM
Deleted: We further

PHASE 3?.

Alexej Abyzov 4/12/14 11:31 PM
Deleted:

Alexej Abyzov 4/12/14 7:24 PM
Deleted: the final set

Alexej Abyzov 4/12/14 7:24 PM
Deleted: of

Alexej Abyzov 4/11/14 5:31 PM
Deleted: deletions

Alexej Abyzov 4/11/14 5:51 PM
Deleted: such

Alexej Abyzov 4/12/14 8:33 PM
Deleted: significantly larger than

Alexej Abyzov 4/11/14 5:36 PM
Deleted: is

Alexej Abyzov 4/11/14 5:33 PM
Deleted: likely to be

Alexej Abyzov 4/11/14 5:36 PM
Deleted: , and rather more

MORE Y GOOD

CLEE COMMENTS PNAS + NATH

increases are likely to be explained by the co-occurrence of different variants in genomic regions experiencing reduced selection. This is evident by the aggregated conservation score decreasing around breakpoints in conjunction with an increase in SNP densities. Aside from overall SNP density, the densities of all individual substitution types also increases close to NH and TEI breakpoint **(Table S2)**. However, this is not the case for NAHR breakpoints, for which C to T substitutions are enriched while T to A and C to A are depleted (**Fig. 2B; Table S1**). Further analysis, performed by removing CpG di-nucleoties from consideration, revealed that the increase in C to T substitutions is due to the enrichment of CpG motif exclusively around NAHR breakpoints, but not around NH or TEI breakpoints (**Fig. 2B**). The motif is known to be highly mutable and, particularly, for C to T substitutions when methylated. Thus, this analysis reveals the potential association of NAHR breakpoints with regions of methylation. We directly tested for methylation levels around breakpoints.

Association of breakpoints with methylation, chromatin states and active regions

When aggregating methylation levels from H1 Stem Cell Line {REF Shantao} we did not see convincing increase in methylation around breakpoints of either class (Fig. S?). We further tested for an association of breakpoints with Dynamically Methylated Regions (DMR) {REF doi:10.1038/nature12433}. Interestingly, NAHR breakpoints were much stronger associated with DMRs than NHs (enrichment and p-value) and TEI (enrichment and p-value) breakpoints.

Next, we used two states of the chromosome's interactome as defined by Hi-C experiment {REF HI-C} and roughly corresponding to open and closed chromatin, to investigate any correlation of breakpoints with open and active DNA chromatin. We tested for the occurrence of breakpoints in genomic bins of 1 Mbps assigned to either state. To determine the significance of our findings we fixed relative rearrangements of chromatin states but randomized their positions relative to breakpoints (see **Methods**). We observed that NH 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 active chromatin marks and this observation is also confirmed with the new set of breakpoints derived in this study (**Fig. S2**). 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.

Micro-insertion at breakpoint deletions and relation with replication timing

Multiple studies have reported the existence of micro-inserted sequences at deletion breakpoints {REF}. In our dataset we observed 2,391 (27%) deletions with micro-insertions ranging in length from 1 to 96 bps with majority of less than 10 bps in length (**Fig. 4A**). Those are likely to arise from technical ambiguity in breakpoint reporting when there is SNP and/or indel close to breakpoints. We, therefore, performed the following analyses for micro-insertions longer than 10 bps.

In agreement with previous finding {REF Conrad Kidd}, micro-insertions were observed almost exclusively (83%) for NH events. Replication based mechanisms were suggested to generate deletions with micro insertions that are copies of some sequence in the genome {REF Lupski}. To test for this possibility we semi-manually determined the likely genomic origin, i.e., template site, of 133 (37%) inserted sequences of which 114 were 20 bps or longer, constituting 42% 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. We categorized template sites as those: i) within a deletion, total of 49; ii) outside of a deletion but on the same chromosome, total of 52 (39%); and iii) on a different chromosome, total of 25 (19%). Seven template sites spanning breakpoints and were excluded from analysis.

It was previously observed that NH events typically have few bases of homology around their breakpoints and template site {REF BREAKSEQ, KIDD, CONRAD, LUPSKI}. We do confirm this observation (**Fig. 4B and S3A**) for blunt deletions and those 108 template sites located on the same chromosome as the corresponding deletion. However, no sequence micro-homology around breakpoints was apparent for deletions having template site on a different chromosome (**Fig. S3**).

NOT CLEAR

Alexej Abyzov 4/12/14 8:27 PM
Deleted: overal

Alexej Abyzov 4/12/14 8:27 PM
Deleted: i

Alexej Abyzov 4/12/14 8:27 PM
Deleted: es

Alexej Abyzov 4/12/14 8:27 PM
Deleted: density

Alexej Abyzov 4/11/14 5:37 PM
Deleted: transitions

Alexej Abyzov 4/11/14 5:52 PM
Deleted: i.e.,

Alexej Abyzov 4/11/14 5:52 PM
Comment [1]: Shantao

Alexej Abyzov 4/11/14 5:52 PM
Comment [2]: Shantao methylation plots

Alexej Abyzov 4/11/14 5:52 PM
Comment [3]: Shantao

Alexej Abyzov 4/11/14 5:52 PM
Comment [4]: Shantao

Alexej Abyzov 4/12/14 8:36 PM
Deleted: W

Alexej Abyzov 4/12/14 8:41 PM
Deleted: we simulated noise by circularly permuting

Alexej Abyzov 4/12/14 8:44 PM
Deleted: along the genome; preserving their relative arrangement but

Alexej Abyzov 4/12/14 8:44 PM
Deleted: ing

Alexej Abyzov 3/7/14 4:19 PM
Comment [5]: Arif

Alexej Abyzov 4/12/14 8:49 PM
Deleted: to be explained by existence of base mismatches and/or indels close to deletion breakpoints in the aligned contig. Mismatches and indels are penalized and including them in the alignment decreases the overall alignment score, while aligning few bases between the mismatch/indel and breakpoints cannot compensate for the alignment score decrease. As such, an aligner chooses not to align those few bases and report as micro-insertion. Given our alignment parameters (see **Methods**) it is likely that micro-insertions shorter than 10 bps arise due to such effect. An enrichment of point mutations close to deletion breakpoints has been previously described {REF LUPSKI} and was also observed the same in this study on a larger scale (**Fig. 1**).

Alexej Abyzov 4/11/14 5:45 PM
Deleted: spann

The distribution of the nearest distance between template site and either of the breakpoints revealed relative preferred arrangement **(Fig. 4C)**. The template site was located either within 10-100 bps (proximal site) or in the range from 2 to 7 kbps (distal site) of one of the breakpoints. Existence of such characteristic distances may signify the mechanism(s) leading to generation of micro-insertions.

It was previously noted {REF Koren} that breakpoints of deletions generated by different mechanism show different association with replication time. We confirm those observations: NAHR deletions are, typically, associated with early replicating regions, HN with later ones, while TEIs show no significant relation to replication time. Furthermore, template sites outside deletions, typically, replicate later (**Fig. 4D**) than breakpoint regions (p-value < 0.03 by binomial test). However, that was not the case for template sites within deletions or on a different chromosome (**Fig. S3**). The former can be easily rationalized, as replication time can be determined only on a large scale and is, typically, the same within entire deleted region. The reason for the latter, however, is not clear. But, note, the distinct relation of such sites with replication time may stress the earlier observation from sequence micro-homology analysis, that deletions with template site on the same and different chromosomes are, in fact, different (**Fig. S3**).

Discussion

- We provide large, less biased, and high quality dataset of breakpoints
- They aggregate with SNPs and indels. Perhaps, expected (the indel paper uses much smaller scale, about 5% increase in +/- 2kb regions)

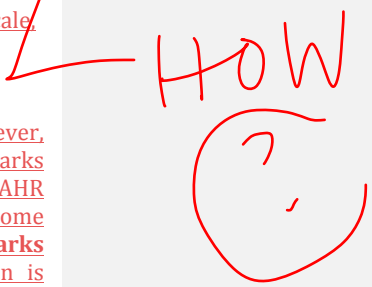
As for other classes, NAHR breakpoints were associated with less conserved regions (**Fig. 2**). However, and different from other classes, they were also associated with open chromatin and active histone marks (**Fig. 3**), i.e., with likely functional regions. This poses a paradox. More than that, classical NAHR mechanism postulates DNA replication as a requirement for generating a SV during chromosome segregations. However, no defined structure of DNA exists at such stage {REF Mirny} **histone marks during replication?**, and therefore, no association of NAHR breakpoints with open chromatin is expected. Association with early replication timing is also stunning. By the time of chromosome segregation DNA replication is completed and replication time should not play a role. And, finally, we found no association of NAHR breakpoints with recombination hotspots, as postulated by classical mechanism. We, thought, that sequence content could be the reason for all these "unexpectancies" but open/active regions of chromatin are no different in repeat content and segmental duplications than closed regions and thus are not prone for NAHR generation due sequence content.

We, therefore, hypothesize that the observed and classifieds as likely NAHR events happen in a cell before replication. Open/active chromatin is more likely to have single stranded DNA (ssDNA) than the closed/inactive chromatin. Such, ssDNA can serve as a template in double strand repair pathway. Negative selection purifies generated NAHR SVs resulting in them being associated with less conserved regions. Association of NAHR breakpoint with enhancer and with dynamically methylated regions (DMR) supports this observation, as DMR and enhancers are likely to mark genomic regions that are not ubiquitously functional in human tissues. Furthermore, possible association with methylation (through mutational signature of methylated regions, **Fig. 1**) suggests that regions with observed NAHR could be non-functional in the cell of origin. Association with early replication timing in our hypothesis is transient through open chromatin, which replicates first.

Alexej Abyzov 4/12/14 11:33 PM
Deleted:

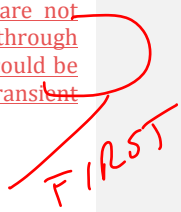
Alexej Abyzov 4/12/14 8:59 PM
Deleted: Interesting that proximal template sites typically occur within the deleted sequence and, perhaps, can be explained by co-occurrence of two indels, discovered as a one deletion, or as deletion with a nearby indel. In other words, micro-insertion is genomic sequence between two proximal variants. However, the other peak (around 2-7 kbps) in the distribution could

Alexej Abyzov 4/12/14 8:56 PM
Moved down [2]: We hypothesize that the distance to template site could be related to DNA packing in a cell or to the length of DNA to wrap around the replication bubble. To investigate this further we compared replication times of breakpoints and template sites.



Alexej Abyzov 4/12/14 9:51 PM
Comment [6]: Shantao to confirm.

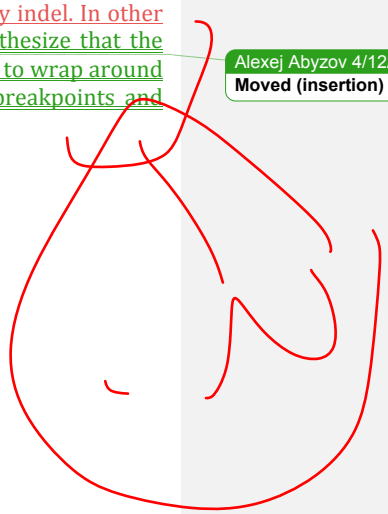
Alexej Abyzov 4/12/14 10:22 PM
Comment [7]: Shantao's to do



- Methylation correlates with meiotic recombination. Use CpG SNPs as surrogates {REF PMID: 19158364}. They corrected the C>T on CpG by the number of CpG and total SNPs
- DMR association with NAHR suggests that NAHR might be more likely to cause diseases?
- Insight into template switching from mapping inserted sequence and correlation timing.
-

Proximal template sites often occur within the deletion breakpoint. Such cases, perhaps, can be explained by co-occurrence of two indels, discovered as a one deletion, or as deletion with a nearby indel. In other words, micro-insertion is genomic sequence between two proximal variants. We hypothesize that the distance to template site could be related to DNA packing in a cell or to the length of DNA to wrap around the replication bubble. To investigate this further we compared replication times of breakpoints and template sites.

Alexej Abyzov 4/12/14 8:56 PM
Moved (insertion) [2]



- We did not find significant association between NH and ASR
- **Recombination motif….Recombination hot spot?**

Methods

Discovery and merging

Deletions discovered by five CNV callers {REF} were merged with the set of breakpoints discovered in 180 pilot samples of the 1000 Genomes Project {REF}. The merged set contained 113,649 deletion calls. For each call we collected read pairs around its boundaries in samples where deletion was discovered and assemble them with TIGRA-SV {REF} into contigs spanning breakpoints. The contigs were aligned to the deleted regions with CROSSMATCH {REF} and AGE {REF} to identify deletion breakpoints (see below). This way we inferred 36,237 breakpoints, of which 17,947 (50%) breakpoints were exactly the same by two approaches, 9,537 (26%) breakpoints were different by the two approaches, and 8,753 (24%) were uniquely inferred by either one of the approached. In cases when there were different breakpoints inferred from the two alignments we chose breakpoints from AGE alignments, as AGE method was specifically designed to align contigs with structural variations. Given large disagreement between the two approaches we further filtered breakpoints by aligning unmapped reads to sequence junctions of the deletions (see below and **Fig. 1**). Based on PCR validation, we further performed ad-hoc filtering of deletions to reduce systematic false positives arising from using synonymous split-read (SR) approaches: deletion calling by SR, breakpoint derivation from assembly (which is SR based), and filtering from read mapping to junction (which is SR like). To summarize, all filtering steps were: i) removing breakpoints not passing criteria for support by mapped reads to their junction (see below); ii) removing deletions classified as VNTR, as their breakpoints are in very repetitive regions; iii) removing breakpoints found by only split-read calling approaches Delly, Pindel, and assembled in the pilot, as, in case of mistake, assembly/filtering is likely to repeat it; iv) removing deletions with breakpoints inferred from only CROSSMATCH alignments; v) removing deletions called by only one method with breakpoints inferred from only AGE alignments. The first three filters were the most effective in removing false positive calls (**Fig. S4**). 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. FDR for deletion breakpoints includes mistakes when deletion is not present but also includes cases when the breakpoint is incorrectly determined (**Fig. S4**).

Defining breakpoints from CROSSMATCH alignments

Alexej Abyzov 3/22/14 10:28 PM
Comment [8]: Ken

Defining breakpoints from AGE alignments

Assembled by TIGRA-SV contigs of at least 100 bps in length were aligned to the corresponding predicted deleted region extended by 2 kbps downstream and upstream. AGE was run with option '-indel –match=1 –mismatch=-10 –go=-10 –ge=-1', specifying: that contigs are expected to have insertion/deletion, that score for base match is 1, that mismatch penalty is -10, that gap opening penalty is -10, and that gap extension penalty is -1. Alignments consistent with the predicted deletion were selected to identify deletion breakpoints. The consistency was defined by the following criteria: i) at least 90% of bases in a contig are aligned; ii) there must be at lest 98% of identical bases in entire alignment; iii) there should be at least 97% identical bases in alignment of each flank, i.e., downstream or upstream from the deletion; iv) each flank must have at least 30 base pairs aligned; v) regions between breakpoints must have 50% reciprocal length overlap with the predicted deletion bounds; vi) breakpoints should be within 200 bps of the predicted deletion bounds; vii) alternative alignments, if any, must satisfy all of the conditions above. In case of multiple contigs alignments satisfying the above condition, the one for contig with highest coverage, as per assembly, was chosen to define breakpoints.

Filtering breakpoints by mapping unmapped reads to breakpoint junctions

Most of the reads utilized in assembly were from 30 to 70 bps in length, i.e., rather short. This fact complicates assembly and makes it rather prone to mistakes, particularly, in repetitive region, for which deletion breakpoints are enriched. Therefore, to ensure physical (rather than artificial, as a results of assembly error) continuity of flanking and inserted (if any) sequences at breakpoints we performed breakpoint filtering by utilizing unmapped reds. For each derived deletion breakpoint we constructed breakpoint junction sequence by joining 100 bps downstream with 100 bps upstream of the breakpoints. Micro-insertion (if exists) was inserted in the middle. The set of all 36,237 junctions sequences from 200 to 298 bps in length comprised junction library. Unmapped reads were mapped to the junction library using Bowtie 0.12.7 {REF} with the options '--best --strata -v 3 -m 1', requiring to make ungapped alignment with at most 3 mismatches and report unique alignments only. Prior to mapping, and the same way it was done by BWA {REF} during alignments preparation by the 1000 Genomes Project, the reads were trimmed at low quality 3'-end up to the average base-quality of 15. Reads mapping with less than 3% of mismatches of their length and having aligned bases in downstream and upstream flanking sequences were retained as supporting the junction they aligned to. We chose a particular cut off *d* on the number/fraction of bases aligned to each flank to choose reads supporting breakpoints. Breakpoints that had supporting reads from two different individuals passed the filter. This requirement ensures that breakpoints passing the filter are for heritable germline deletions, as singletons could be of somatic origin.

In total, we attempted realigning 15.8 billion reads to the junction library. Given the large number of realigned reads and large size of the junction library, some of the read mappings could be by chance. To discriminate between real and random mappings we developed an empirical null model (**Fig. 1C**). The model is based on imitating the junction library with semi-random sequence, thereby, creating null junction library and mapping unaligned reads to that null library. Such mapping will represent a random noise and can be used for optimization of values of *d*. The library is generated from inner sequences of deleted regions (**Fig. 1C**). Such an approach is advantageous in that it allows preserving genomic (e.g., nucleotide content and sequence homology at breakpoints) and data features (e.g., read coverage) associated with the loci of breakpoints.

We realigned all unmapped reads to the null junction library and varied the values of *d* to find the cut off at which the number of null junction passing the filter will be <5% of the number of real junctions passing the filter at the same cut off (**Fig. S5**), i.e., we aim at <5% in-silico FDR. This criterion let us to setting value of *d* at 13 bps. The empirical null model allowed us to stratify the precision of breakpoint by various categories. For example, and as expected, we observed that breakpoints found by only one approach (either AGE or CROSSMATCH based) have higher in-silico FDR. The order of breakpoints of different classes by corresponding in-silico FDR was (from lowest to highest): NH, TEI, NAHR, and VNTR. This is also expected, as in the same order breakpoints of different classes have progressively more repeats around their breakpoints.

To summarize, the developed empirical model captures essential biological features of breakpoints, not biased by using data loci different from the breakpoints, and allows translating random mapping into estimated FDR. We suggest that such empirical model can be used to estimate FDR of genotyping known breakpoints from sequencing data. When, however, applied to breakpoint filtering/validation one should keep in mind that the approach may not account for systematic false positives arising during structural variant calling by split-read method(s), as was observed in our analysis (see above).

PCR and IRS validations

Comparing with OMNI genotypes

A set of 11,472 breakpoints derived in the pilot of the 1000 Genomes Project was tested on a custom SNP array designed by ILLUMINA and named OMNI 2.5s array. The pairs of probes were designed such that one probe would hybridize to the reference allele and the other one to breakpoint sequence, i.e., to the alternative allele. The probes were different in only one nucleotide to mimic probes for SNP genotyping.

Alexej Abyzov 3/26/14 7:40 PM
Comment [9]: Matt

Accordingly, all the downstream hybridization signal processing was performed with standard software for SNP array analysis.
Probe design, hybridization in 431 individuals, and genotyping quality control resulted in confident array derived genotypes for 4,385 (38%) breakpoints. XXX of our confident breakpoints were in this set (**Table S1**) and 292 individuals were both sequenced by the 1000 Genomes Project and genotyped by this array. Comparison of samples genotyped as having a deletion by array and by mapping reads to sequence junctions, as we did for filtering breakpoints, revealed that individuals with deletion genotype by read mapping represent almost a perfect subset of those genotyped by arrays (**Fig. S6**). This is easy to rationalize by noting that individuals in the 1000 Genomes Project were sequenced at shallow 4-8X coverage, thus not likely to have many reads covering breakpoint sequences, particularly, in the case of heterozygous deletion. Furthermore, requirement for reads mapped to deletion sequence junction to extend at least 13 bps across the junction in each direction further reduces the number of reads that we consider supporting deletions.

Confirmation of breakpoints in high coverage trios

Alexej Abyzov 4/4/14 5:23 PM
Comment [10]: Daniel

[Breakpoint validation was performed OR We validated XXX% of breakpoints] in trios using HiSeq 2500 long read high coverage data from the Broad Institute.
The reads in the HiSeq 2500 data were 250 bp in length, with the majority of pairs having insert size of ~400 bp. This implies that the reads in most read pairs significantly overlapped in sequence, typically with length ~100 bp. In our validation method, we merged overlapping read pairs to construct longer reads, where the identity of bases in the overlapping sequence were selected from the read in the pair with the higher quality score. Therefore, these longer reads contained more reliable sequences than either of the reads in the original pair alone, and were easier to align [TO??] by virtue of their increased lengths.

The 8,943 breakpoints in our set were genotyped in trios by CNVnator, and 4,385 breakpoints had genotypes consistent with deletions in trios (<1.5 or <0.5, for diploid and haploid regions, respectively). Read pairs in the HiSeq 2500 data with coordinates in the vicinity of these regions were merged to form longer reads as described above, where the length of a given pair's overlapping sequence was estimated assuming a binomial distribution for the number of mismatched bases in the overlap (p_mismatch=0.75, n=length of the overlapping sequence). We selected overlap lengths that minimized the p-value under this assumption (i.e., given k mismatches in a overlap of length n, the probability that at most k mismatches would occur by chance). We only considered merged read pairs that exhibited a low ratio of number of mismatches to overlap length (<0.2) and low p-values (<1e-10) to increase the reliability of the overlap estimates. Our method, when tested using these criteria on NA12878 read pairs with known overlaps, yielded correct overlap estimates 99.9483% of the time (6057303 of 6060436 read pairs).
We then determined if these longer reads demonstrated support for the 4,385 breakpoints that had genotypes consistent with deletions in the trios. We first used AGE alignment to select longer reads that contained sequences consistent with deletions. [HOW DID find_support.pl DETERMINE SUPPORT?]. We considered a given breakpoint to be validated by the reads if the plurality of the reads constructed in the vicinity of that breakpoint had both coordinates and microinsertion lengths that matched those of the breakpoint. [WHAT IS RELEVANT DATA TO INCLUDE?]

Aggregation calculation

Intersection with open/closed chromatin

Alexej Abyzov 4/4/14 3:04 PM
Comment [11]: Shantao

We used the Hi-C data generated on human lymphoblastoid cell line (GM06990) (REF PMID: 19815776). The bin size is 1Mb and we counted the breakpoints of different mechanisms fall into each bins. Then we calculated the percentage of breakpoints with positive eigenvectors, that is, in an open chromatin state. The number is compared with an empirical distribution generated by circular permutation to calculate p-value. The circular permutation is made by joining the end of bin array with the beginning to make it

circular. Then we rotate this circular array to every possible position and calculate the percentage of active chromatin state each time.

After dropping the original position and two adjacent positions, distributions of NH and NAHR passed the Shapiro–Wilk test of normality. Therefore we used normal distribution to determine their p-value. For STEI, it does not show a significant change in the empirical distribution and it does not pass the normality test.

(Do you want me to use purely empirical prob. distribution or normal approximation (normality test + Z-score give NH and NAHR a much lower p-value. But STEI does not pass the normality test. Is this statistically sound?)

To determine the significance of our findings we simulated noise by circularly permuting chromatin states along the genome; preserving their relative arrangement but randomizing their position relative to breakpoints (see **Methods**).

Double-strand break analysis

We used an aphidicolin generated double-stand break map of HeLa cell line (REF PMID: 23503052). We counted the number of breakpoints fall into the bins provided in the original dataset. We compared the number the number with the empirical distribution generated by circular permutation.

Intersection with enhancers

Mapping template sites

Majority of micro-insertions are less than 10 bps in length (**Fig. 4A**). Those are likely to be explained by existence of base mismatches and/or indels close to deletion breakpoints in the aligned contig. Mismatches and indels are penalized and including them in the alignment decreases the overall alignment score, while aligning few bases between the mismatch/indel and breakpoints cannot compensate for the alignment score decrease. As such, an aligner chooses not to align those few bases and report as micro-insertion. Given our alignment parameters (see **Methods**) it is likely that micro-insertions shorter than 10 bps arise due to such effect. An enrichment of point mutations close to deletion breakpoints has been previously described {REF LUPSKI} and was also observed the same in this study on a larger scale (**Fig. 1**). We, therefore, performed the following analyses for micro-insertions longer than 10 bps.

Replication time analysis