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

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

Abstract

Structural variants (SVs) affect more bases in the human genome than those arising from point mutations (i.e. SNPs) While entire SVs can be thousands of bases in length, only the few bases around their breakpoints hold the most crucial information about their formation mechanism, which mostly fall into three classes non-allelic homologous recombination (NAHR), transposable element insertion, and non-homologous (NH) mechanisms. Here we identify, determine the formation mechanism, and analyze 8,943 associated with deletions in 1,092 samples sequenced by the 1000 Genomes Project and make them available as a public resource, the largest breakpoint collection to date. Overall, SV breakpoints have more nearby SNPs (and indels) than the genomic average. This effect is true over large (megabase) genomic scales and is likely explained by relaxed selection acting on breakpoint regions since it also correlated with reduced cross-species conservation. NAHR breakpoints had distinct SNP aggregation profile explained by occurrence of nearby CpGs, which, in turn, are linked with methylation. Moreover, by correlating with Hi-C data and histone marks, we find that these breakpoints are associated with open chromatin. We hypothesize that since replication is mostly devoid of chromatin structure, some of the NAHR deletions occur in early embryonic and germ cells and then passed on through the germline. We do not find any associated correlations with NH breakpoints. However, we do find that they are often coupled with micro-insertions (MIs). We identified template sites for over one hundred of such MIs. Surprisingly, these are located at two characteristic distances from the breakpoint and tend to be replicated later. This findings are consistent with a template switching mechanism, suggesting particular spatial and temporal configurations for DNA during some NH events.

Alexej Abyzov 5/17/14 10:31 PM
Deleted: arrising

Alexej Abyzov 5/17/14 10:37 PM
Deleted: are associated with a slightly different

Alexej Abyzov 5/17/14 10:38 PM
Deleted: mutational

Alexej Abyzov 5/17/14 10:37 PM
Deleted: associated with

Alexej Abyzov 5/17/14 10:37 PM
Deleted: associated

Alexej Abyzov 5/17/14 10:44 PM
Deleted: "

Alexej Abyzov 5/17/14 10:44 PM
Deleted: " and early replication, and this finding is borne out by many sources of evidence: methylation and histone-mark profiles, Hi-C data and replication timing information

Alexej Abyzov 5/17/14 10:44 PM
Deleted: many

Alexej Abyzov 5/17/14 10:44 PM
Deleted: events are

Alexej Abyzov 5/17/14 10:33 PM
Deleted: with

Alexej Abyzov 5/17/14 10:49 PM
Deleted: also

Alexej Abyzov 5/17/14 10:49 PM
Deleted: is

Introduction

Genome structural variations involving hundreds and thousands of bases are common during evolution and are widespread in the human genome {REF}. The comparably larger fraction (3) of the human genome affected by SVs than SNPs implies they may have greater, or at least similar, consequences for phenotypic variation and evolution as SNPs (1, 2). Not surprisingly, SVs can cause and have been associated with numerous diseases (4, 5)(6, 7)(4, 8)(9, 10).

SV occurrence and existence is a complex phenomenon that is not completely understood. SVs, as other variants, are genetic imprints of mutational process in cells. The sequence content of SVs can carry important information about their origin, but bases around their breakpoints hold the most crucial remaining details of SV genesis. Long homologies around breakpoints suggest that an SV originated as a result of Non-Allelic Homologous Recombination (NAHR), short homologies with high content of mobile element within SV regions suggest the origin by transposable element insertions (TEI), while none or little homology (NH) at breakpoints can indicate the origin by template switching mechanism during replication or from non-homologous end-joining (NHEJ) event. Mistakes in breakpoint resolution of just several bases can lead to misclassification of mutational signatures and compromise downstream analysis. Thus, studying SVs at breakpoint resolution is fundamental to understanding the mutational mechanisms generating them.

Few systematic genome wide studies of SV breakpoints have been carried out to date {Lam:2010cp} {REF Kidd, BreakSeq, Conrad, Pilot}. While these studies confirmed existing knowledge about the mechanisms generating SVs and provided new insights, they also demonstrated that complete ascertainment of breakpoints for most SVs in the human population and detailed understanding of the mutational process generating them is a goal yet to be achieved. Three studies by Lam et al. {REF}, Conrad et al. {REF}, and Kidd et al., {REF} analyzed 1,961, 324, and 1,054 SV breakpoints respectively, in 14, 3, and 17 individuals respectively. The majority of SVs analyzed in those studies were larger than 1 kbps. Analysis of genomes from 180 individuals in the pilot phase of the 1000 Genomes Project {REF} revealed that there are at least an order of magnitude more SVs present in the human population, a significant fraction, if not most, of which are smaller than 1 kbps. The challenge of precise breakpoint identification from inexpensive short-read sequencing was also realized {REF Pilot 18 Koreans}.

Along with advance in breakpoint ascertainment multiple studies aiming at deciphering genome function have been conducted that generated wealth of functional genomic data. For example, ENCODE project {REF} released data on chromatin marks, methylation, DNAse hypersensitive sites, and transcription binding sites in multiple cell lineages and tissues {REF}. These data allow studying SV breakpoints in the genome functional content. Here we describe the discovery and analysis of large set of 8,943 high confidence deletion breakpoints from 1,092 individuals sequenced in the phase 1 of the 1000 Genomes Project {REF}. We put special emphasis on the derivation of our set of high precision breakpoints and provide it as a valuable resource. The subsequent downstream analysis, including correlation of the breakpoints' with functional genomic data, reveals important details of their mutation mechanisms and the genomic characteristics associated with them.

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**). 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 that this model imitates biologically relevant sequence homologies around breakpoints. We realigned unmapped reads to real and null junctions and optimized criteria for considering whether a read supports a junction by interrogating alignments to null junctions, as such alignments reflect random noise. For validation we performed PCR amplification across breakpoints and tested for differences in intensity values for SNP

Alexej Abyzov 5/16/14 5:29 PM
Deleted: but

Alexej Abyzov 5/16/14 5:29 PM
Deleted:

Alexej Abyzov 5/16/14 5:57 PM
Deleted: Finally, our

Alexej Abyzov 5/16/14 6:24 PM
Deleted: Because the use of different aligners and processing pipelines for mapping breakpoints yielded inconsistent results, we applied stringent filtering to ensure the physical continuity of flanking and inserted (if any) sequences at breakpoints.

Alexej Abyzov 5/16/14 6:25 PM
Deleted: Next, guided by validation, we performed ad-hoc filtering of deletions to reduce systematic false positives arising from the use of split-read information during calling, assembly, and filtering. In particular, we did not include deletions with the breakpoint signature of variable tandem repeats in the final set.

probes across individuals 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 (6.8%) and IRS (6.4%) validations for deletion existence, and 13.7% for deletion presence with correct breakpoints from PCR. We have further confirmed 28% of the breakpoint sequences with the OMNI SNP genotyping array, and 39% of breakpoint sequences in trios with high coverage and long read data (Table S1 and **Methods**).

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. B) Null model for breakpoint filtering. C) 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 biased toward large non-repetitive deletions for the purpose of reliable genotyping, resulting in the strong underrepresentation of mobile element insertions.

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**). We further classified the deletions by likely mechanisms of origin using sequence signatures at breakpoints from the following classes {REF BREAKSEQ}: non-allelic homologous recombination (NAHR), transposable element insertions (TEI), and non-homologous (NH) events. Note that our set 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.

We provide this dataset as a public resource (Table S1) with complete information about breakpoint coordinates, mechanism classification, and sequence of micro-insertions (MI) at breakpoints, if applicable. The resource can be used in various study settings including SV genotyping by mapping reads to breakpoint sequences.

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. 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 kbp regions around breakpoints of each class (**Fig. 2A and S1**). One might suggest that false SNPs calls as a result of read mismapping around breakpoints could cause the observed increase, as reads spanning the SV junctions are often misaligned. However, the scale of increases is large relative to the 450-650 bp insert size of sequencing libraries and, therefore, cannot be artificial. Analysis of sequence conservation around

Alexej Abyzov 5/16/14 6:27 PM
Deleted: By using BREAKSEQ software {REF BREAKSEQ},

Alexej Abyzov 5/16/14 6:27 PM
Deleted: w

Alexej Abyzov 5/16/14 6:28 PM
Moved down [1]: It should be noted that NAHR and TEI events are more difficult to discover as they contain repeats between and at breakpoints, and so are likely to be underrepresented in this set.

Alexej Abyzov 5/16/14 6:48 PM
Formatted: Indent: First line: 0"

Alexej Abyzov 5/16/14 6:42 PM
Deleted: compromised

Alexej Abyzov 5/16/14 6:42 PM
Deleted: artifacts

Alexej Abyzov 5/16/14 6:41 PM
Deleted: .

Alexej Abyzov 5/16/14 6:41 PM
Deleted: T

Alexej Abyzov 5/16/14 6:41 PM
Deleted: we assert that the observed increases are

Figure 2. Co-aggregation of SNPs with deletion breakpoints found in the analyzed samples. A) Normalized by nucleotide frequency and background SNP densities increased while conservation decreased in 400 kbp regions around breakpoints of each class. B) Densities increase for substitutions of all types around NH and TEI breakpoints but this is not the case for NAHR breakpoints. C to A substitution showed the most pronounced increase close to TEI and NH but is depleted around NAHR breakpoints. Increase of C to T substitutions around NAHR breakpoints is driven by SNPs in CpG motifs as evident from red bars. Furthermore, is solely due to enrichment of CpG motifs (**Fig. S2**). This is consistent with common knowledge that NAHR events are associated with sites of recombination.

breakpoints revealed that the 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 increase close to NH and TEI breakpoints **(Table S2)**. However, this is not the case for NAHR breakpoints, for which C to A and T to A are depleted while C to T substitutions are enriched (**Fig. 2B; Table S1**). We hypothesized, that the observed differences in SNP aggregation can be explained by the sequence and motif content around breakpoints of each class and/or different selection pressure acting on substitutions of each type. Indeed, further analysis, performed by removing CpG di-nucleotides from consideration, revealed that the increase in C to T substitutions is due to the enrichment of the CpG motif exclusively around NAHR breakpoints, but not around NH or TEI breakpoints (**Fig. 2B** and **S2**). This is expected, as it is known, that the motif itself, C to T mutations within it, and NAHR breakpoints are all associated with recombination hot-spots {REF doi: 10.1101/gr.086181.108 doi:10.1101/gr.1970304}. Indeed, NAHR breakpoints in our set were strongly associated with higher recombination rates (enrichment of 1.4 with *p-value* < 10^{-3}), and no significant association for breakpoints of other classes. However, unexpectedly, density of C to T substitutions in CpG motifs decreased close to NAHR breakpoints (**Fig. S2**). Since such substitutions are methylation-associated we directly tested for methylation levels around breakpoints.

Alexej Abyzov 5/16/14 6:46 PM
Deleted: Actually, C to A substitution showed the most pronounced increase close to TEI and NH breakpoints. Perhaps

Alexej Abyzov 5/16/14 6:49 PM
Deleted: d

Alexej Abyzov 5/16/14 7:08 PM
Deleted: S

Alexej Abyzov 5/16/14 7:08 PM
Deleted: deletions

Alexej Abyzov 5/16/14 7:08 PM
Deleted: S

Alexej Abyzov 5/16/14 7:08 PM
Deleted: deletions

Alexej Abyzov 5/16/14 7:08 PM
Deleted: depletions

Alexej Abyzov 5/16/14 7:09 PM
Deleted: the genomic average

Alexej Abyzov 5/16/14 6:51 PM
Deleted: =XXX and p-value=XXX accordingly

Association of breakpoints with methylation, chromatin states and active regions

Methylation levels from H1ESC line showed no change close to breakpoints of all classes (**Fig. S3**). We next searched for an association of deleted regions with hypomethylated regions in sperm as compared to H1ESC {REF PMID: 21925323}. Strong association was observed for TEI and NAHR breakpoints (**Fig. 3B**). In particular, the TEI breakpoints were five times and NAHR breakpoints were over two times more likely to reside in hypomethylated regions than expected by chance (both *p-values* < $2x10^{-4}$). Demethylation of transposable elements in sperm was known for a while. But similar effect for NAHR deletions may explain their correlation with reduced C to T substitutions density in CpG. {REF Lupski about hypomethylation~SVs in sperm? PMID: 22615578}

Figure 3. Relation of breakpoints of each class to chromatin states, histone marks and methylation. A) Breakpoint co-occurrence with chromatin states, defined by corresponding eigenvector, of Hi-C data (upper panel). The genome wide co-occurrence is ordered by the value of the eigenvector (lower panel). Curves were smoothed using sliding window of 3,000 bins. NAHR breakpoints are associated with open chromatin. This association cannot be explained by higher content of segments duplications (SDs) or recombination rate (RR). B) Overlap of breakpoints with hypomethylated regions in sperm. NAHR and STEI breakpoints show strong association. Increase in methylation level in H1ESC in 2 kbp region was observed for breakpoints of all classes (**Fig. S3**). C) Association with histone marks. NH breakpoints are associated with depletion of active marks (all but red lines). TEI breakpoints show no associations with any marks. NAHR breakpoints are depleted for repressive H3K27me3 mark, and are enriched for all active marks (all but red lines).

Next, we used two states of the chromatin'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 100 kbps assigned to either state. To determine the significance of our findings we fixed relative arrangements of chromatin states and fixed relative arrangements breakpoints but randomized positions of the states and breakpoints with respect to each other (see **Methods**). We observed (**Fig. 3A**) that NH and TEI breakpoints are depleted for open chromatin, while NAHR breakpoints are enriched (p-value < 10^{-4}). Segmental duplications (SDs) are known to mediate NAHR. We indeed (**Fig. 3A**) saw positive correlation (Pearson coefficient 0.85) between NAHR and SDs but only in the closed chromatin, while in the open chromatin we observed anti-correlation (Pearson coefficient -0.32). Similarly, we observed strong correlation of recombination rate with NAHR breakpoints in closed chromatin (Pearson coefficient 0.94) but significantly weaker correlation in the open one (Pearson coefficient 0.28). This suggests two conditions for generating deletions by NAHR. We further analyzed an association of NAHR breakpoints with 10 chromatin marks (**Fig. 3C**). The three classes of breakpoint showed very different associations. NH breakpoints were depleted for all active marks and also for H3K9me3 repressive mark. TEI breakpoints showed weak depletion of active marks. However, NAHR breakpoints were different. Not only the density of all active marks increases close to NAHR breakpoints but also density of repressive H3K27me3 mark decreases. As active marks are linked to open chromatin, these observations corroborate the association of NAHR with open chromatin.

Alexej Abyzov 5/17/14 9:03 PM
Deleted: The

Alexej Abyzov 5/17/14 9:03 PM
Deleted: with open chromatin is further corroborated by their association

Alexej Abyzov 5/17/14 9:12 PM
Deleted: active

Micro-insertions at breakpoint deletions and their relation to replication timing

One of the suggested mechanisms for NH deletion generation is non-homologous end joining of DNA ends during double stranded breaks. We found no significant association of breakpoints of any type with regions fragile for double stranded breaks {REF Crosetto}.

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 the majority less than 10 bps in length (**Fig. 4A**). Those are likely to arise from technical ambiguities in breakpoint reporting when there are SNPs or indels close to breakpoints. We therefore performed the following analyses for micro-insertions longer than 10 bps.

As in previous studies {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., the 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, which were 49 (37%) in total; ii) outside of a deletion but on the same chromosome, totaling 52 (39%); and iii) on a different chromosome, with a total of 25 (19%). Seven template sites spanned breakpoints and were excluded from analysis.

Figure 4. Analysis of micro-insertions (MI) at deletion junctions. A) Most MIs are up to 10 bps in length and likely to arise from technical ambiguities in breakpoint reporting when there are SNPs or indels close to breakpoints. Larger MIs are typically found for NH deletions. B) Length of micro-homology (MH) at deletion junction. For deletions with MIs and identified template site, MHs are calculated for 5'-ends/3'-ends of the deletion and the template site (panel insert). Both ends show MN longer than expected by chance and similar to the distribution for blunt deletions. C) The distribution of the nearest distance from template site breakpoints in log10 space. The distribution is almost symmetrical and exhibits distinct peaks in the ranges 10-30 bps (proximal sites) and 2-6 kbps (distal sites). D) The difference in replication time between template site and breakpoints reveals later replication of template sites. For template sites outside the deletion the effect is significant (p-value < 0.03 by binomial test). The effect is even more significant (p-value < 0.01) when excluding difference of up to 0.01 as such small values are comparable to measurement error.

It was previously observed that NH events typically have few bases of homology around their breakpoints and template sites {REF BREAKSEQ, KIDD, CONRAD, LUPSKI}. We do confirm this observation (**Fig. 4B and S3A**) for blunt deletions and those 101 template sites located on the same chromosome as the corresponding deletion. However, no sequence micro-homology around breakpoints was apparent for deletions having template sites on different chromosomes (**Fig. S4A**). The distribution of the nearest distance between template site and either of the breakpoints revealed preferred relative arrangement **(Fig. 4C)**. The template site was typically located either within 10-30 bps (proximal site) or in the range from 2 to 6 kbps (distal site) of one of the breakpoints. The existence of such characteristic distances may signify the mechanism(s) leading to the generation of micro-insertions.

It was previously noted {REF Koren} that breakpoints of deletions generated by different mechanisms are associated differently with replication time. We confirm those observations: NAHR deletions are typically associated with early replicating regions, NH 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

Alexej Abyzov 5/16/14 6:57 PM
Deleted: so

Alexej Abyzov 5/16/14 7:04 PM
Deleted: 10

Alexej Abyzov 5/16/14 7:17 PM
Deleted: 8

Alexej Abyzov 5/16/14 8:38 PM
Deleted: 3

Alexej Abyzov 5/16/14 7:02 PM
Deleted: 10

for template sites within deletions. This 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. Similarly we did not see preference for later/earlier replication time for template site on different chromosome (**Fig. S4B**).

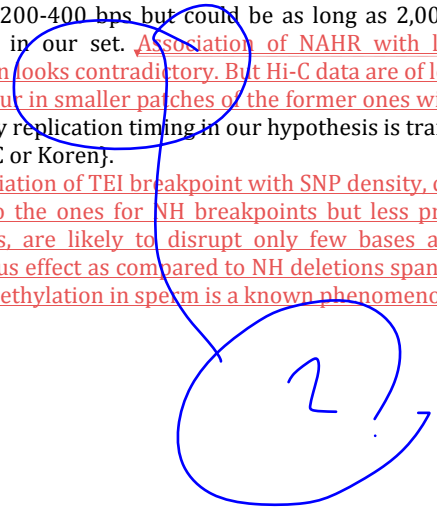
Discussion

In this study we derived a large set of germline deletion breakpoints. This set represents deletions across broad scale of length, with high quality of breakpoint sequences, and across three likely mechanism of origin thereby allowing us to classify breakpoints into 3 classes: NH, NAHR, and TEI. It should be noted that NAHR and TEI events are more difficult to discover as they contain repeats between and at breakpoints, and so are likely to be underrepresented in this set. Further analysis revealed that common feature for breakpoints of all classes is the association with evolutionary less conserved genomic regions, spanning to hundred of kilobases downstream and upstream of the breakpoints. This is likely due to purifying selection acting on SVs. While associations with other measures: CpG motif density, various types of nucleotide substitutions, histone marks, open chromatin and methylation, were different.

Classical NAHR mechanism postulates meiotic cell division as a requirement for generating a germline SV. This implies certain associations, which we did observe in our study. In particular, NAHR breakpoints were associated with higher recombination rates, with higher GC content, higher density of CpG motifs, and with methylation-linked mutations. However, and different from other classes, they were also associated with open chromatin and active histone marks in mitotically dividing cells. This poses a paradox. No defined structure of DNA exists at the time of chromosome segregation {REF Mirny} and histone marks are gone {http://dx.doi.org/10.1016/j.cell.2012.06.046}, thus, no association of breakpoints with open/active chromatin is expected. In fact, as a result of purifying selection one might expect inverse relation of breakpoints with open chromatin and active histone marks, like in case of NH breakpoints. Neither recombination rate no fraction of bases in segmental duplications explain these association for NAHR breakpoints. Furthermore, the association of NAHR with early replication timing is also stunning. By the time of chromosome segregation DNA replication is complete and replication time should not play a role.

We therefore hypothesize that a fraction of classified as likely NAHR events, occur in embryonic and germ cells without replication. Open/active chromatin unpacks DNA that is easy to melt and likely to contain single stranded DNA as a result of transcriptional activity. Such DNA can serve as a template in double strand break repair pathway for breaks in homologous region(s) that are close in space and thereby likely to be from the same chromosome {REF Hi-C}. In fact, intramolecular NAHR, which is homologous recombination between regions of the same continuous chromosome, has been previously suggested {REF Hurles S. Cohen, D. Segal, Cytogenet. Genome Res. 124, 327 (2009)}, and the consequence of such an event would be generation of a deletion and an extra-chromosomal circular DNA (eccDNA). eccDNA was recently extensively analyzed {REF CIR DNA} in mouse somatic tissues and human cancer cell-lines. The striking observation was that eccDNA were enriched in CpGs and exons, supporting the suggestion that unpacked DNA is a requirement for eccDNA generation. Length of eccDNA circles was typically 200-400 bps but could be as long as 2,000 bps, consistent with median of 418 bp for NAHR deletions in our set. Association of NAHR with less conserved genome and also with open/active chromatin looks contradictory. But Hi-C data are of low resolution (100 kbps) and it is feasible that NAHR could occur in smaller patches of the former ones within the larger regions of the latter ones. Association with early replication timing in our hypothesis is transient through open chromatin, which replicates first {REF Hi-C or Koren}.

Association of TEI breakpoint with SNP density, conservation, open chromatin and histone marks was similar to the ones for NH breakpoints but less pronounced. We think it is due to TEIs, as bona fide insertions, are likely to disrupt only few bases around insertion site and, thus, less likely to have deleterious effect as compared to NH deletions spanning from hundreds to million bases. TEI association with demethylation in sperm is a known phenomenon that we confirmed in our study.



Alexej Abyzov 5/16/14 10:43 PM
Deleted: or on different chromosomes (**Fig. S3**).

Alexej Abyzov 5/16/14 10:43 PM
Deleted: e

Alexej Abyzov 5/16/14 10:53 PM
Formatted: Font:Bold

Alexej Abyzov 5/16/14 10:45 PM
Moved down [2]: The reason for the latter, however, is not clear. But note that the distinct relation of such sites with replication time may stress the earlier observation from sequence micro-homology analysis, that deletions with template sites on the same and different chromosomes created by different mechanisms (**Fig. S3**). It is also possible that template sites on different chromosomes arise from mis-mapping the sequence of micro-insertion.

Alexej Abyzov 5/16/14 6:28 PM
Moved (insertion) [1]

Alexej Abyzov 5/17/14 9:22 PM
Deleted: that

Alexej Abyzov 5/17/14 9:22 PM
Deleted: probably

Alexej Abyzov 5/17/14 9:24 PM
Deleted: associated

Alexej Abyzov 5/17/14 9:30 PM
Deleted: NAHR

Alexej Abyzov 5/17/14 9:35 PM
Deleted: all

Alexej Abyzov 5/17/14 9:35 PM
Deleted: “expectancies”

Alexej Abyzov 5/17/14 10:24 PM
Deleted: a

Alexej Abyzov 5/17/14 10:24 PM
Deleted: before

Alexej Abyzov 5/17/14 10:29 PM
Deleted: However, most of sites of eccDNA origin lacked long homologies around breakpoints, and would not be classified as NAHR events. A potential explanation could be that homologous recombination is the dominant double stranded break repair pathway in embryonic stem cells, where germline variations are likely to arise, as compared to somatic tissues {REF PMID: 20446816}, where eccDNA was analyzed.

Our analysis also provided insight into the mechanism(s) of generating deletions in the NH class. Such deletion thought to originate from NHEJ and template switching mechanisms during replication. The latter predict {REF Lupski} that a replication fork can accidentally switch sites of template DNA during DNA duplication. Switching sites skips some genome sequences, thereby generating deletions, or re-replicates the same sequence, thereby generating duplications. Generation micro-insertions at breakpoints occurs when switching happens more than once. We found that template sites for MIs have sequence micro-homology at breakpoints, are located at two characteristic distances from breakpoints (10-30 bps – proximal and 2-6 kbps – distal) and replicate later than the regions of breakpoints. In about half of cases template sites were within breakpoints of corresponding deletions. One might explain such cases by the co-occurrence of two deletions (or deletion and an indel), generated in different individuals (possibly by different mechanisms) and eventually integrated on the same allele and discovered as a single deletion. In other words, it might be suggested that micro-insertions are genomic sequences between two adjacent variants. We think that such an explanation does not apply to most cases. The distributions of nearest distance to breakpoints for sites within and outside breakpoints are very similar and both have the same two characteristic distances. This suggests that deletions with MI template sites within and outside breakpoints are generated by the same mechanism, e.g., by template switching. As template sites from outside deletions could not be explained by variant co-occurrence, we argue that template sites within deletions could not be explained by variant co-occurrence either.

We also observed that template site at different chromosome do not have sequence micro-homology at breakpoints and are not associated with later or earlier replication time. This may imply that deletions with template sites on the same and different chromosomes created by different mechanisms. It is also possible that template sites on different chromosomes arise from mis-mapping the sequence of micro-insertion.

We further hypothesize that the distance to template site could be related to DNA packing in a cell during replication. For example, larger characteristic distances could reflect the length of DNA when wrapped with one loop around the replication bubble to bring a template site close to collapsed or stalled replication fork. Later replication times of template sites suggest that it would still be in the form of a double helix and when dissociated, perhaps by another replication bubble, could provide template sequence for the template switching by the collapsed or stalled fork.

Alexej Abyzov 5/17/14 10:22 PM
Deleted: T

Alexej Abyzov 5/17/14 10:22 PM
Deleted: predict

Alexej Abyzov 5/16/14 7:04 PM
Deleted: 10

Alexej Abyzov 5/16/14 10:45 PM
Moved (insertion) [2]

Alexej Abyzov 5/16/14 10:48 PM
Deleted: The reason for the latter, however, is not clear. But note that the distinct relation of such sites with replication time may stress the earlier observation from sequence micro-homology analysis,

Alexej Abyzov 5/16/14 10:49 PM
Deleted: (**Fig. S3**)

Alexej Abyzov 5/16/14 10:49 PM
Deleted:

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 the deletion was discovered and assembled 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 the two approaches, 9,537 (26%) breakpoints were different by the two approaches, and 8,753 (24%) were uniquely inferred by either one of the approaches. In cases where the two approaches inferred different breakpoints, we chose breakpoints from AGE alignments, as the AGE method was specifically designed to align contigs with structural variations. Given the 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 performed an additional filtering of deletions to reduce systematic false positives arising from the use of 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 only found by split-read calling approaches Delly, Pindel, and assembled in the pilot (the reason is that in case of mistake by a discovery methods, assembly/filtering is likely to repeat it, because relaying on SR approach) 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. S5**). The final set consisted of 8,943 deletion breakpoints with consistent FDR estimates from PCR (6.8%) and IRS (6.4%) tests for deletion presence, and 13.7% for deletion with correct breakpoints from PCR. FDR for deletion breakpoints includes mistakes when deletion is not present but also includes cases in which the breakpoint is incorrectly determined (**Fig. S5**).

Alexej Abyzov 5/16/14 6:24 PM
Deleted: further

Alexej Abyzov 5/16/14 6:23 PM
Deleted: ad-hoc

Alexej Abyzov 5/16/14 10:53 PM
Deleted: 4

Alexej Abyzov 5/16/14 10:53 PM
Deleted: 4

Defining breakpoints from CROSSMATCH alignments

For a contig assembled from an intra-chromosomal variant in the genomic interval [a,b], we prepared a local reference sequence excised from [a-w,b+w], with w=500 bp by default. For a contig assembled from an inter-chromosomal rearrangement, we prepared two local reference sequences from [a-w, a+w] of chromosome c1 and from [b-w, b+w] of chromosome c2, respectively. We mapped each contig assembled by TIGRA to the corresponding reference sequences using CROSSMATCH. In the default setting, we used the following CROSSMATCH parameters: -bandwidth 20 -minmatch 20 -minscore 25 -penalty -10 -discrep_lists -tags -gap_init -10 -gap_ext -1. We removed contigs that had more than 2 hits to the reference and ignored alignments that had substitution rates greater than 0.5%. If a contig differs substantially from the reference, CROSSMATCH returns multiple local alignments, together with a set of statistics describing the quality of the alignments. A glocal alignment (combination of local and global alignment) was constructed from these local alignments {PMID: 12855437}. We used that alignment as the basis for reporting the existence of breakpoints and details about the type, size, orientation, and location of the breakpoints (**Fig. S6**). For example, the glocal alignment that supports a deletion breakpoint contains two local 1-monotonic alignments to the reference {PMID: 22563071}. The gap between the end position of the first alignment and the start position of the second alignment corresponds to the size of the deletion, while the bases shared by both alignments correspond to breakpoint homology.

Alexej Abyzov 5/16/14 10:53 PM
Deleted: 4

Alexej Abyzov 5/16/14 10:53 PM
Deleted: 5

Defining breakpoints from AGE alignments

Contigs assembled by TIGRA-SV 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', which specifies that contigs or reference region are expected to have large insertions/deletions; that the score for base match is 1; that the mismatch penalty is -10; that the gap opening penalty is -10; and that the 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 contig alignments satisfying the above condition, the one with the contig of 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 regions, for which deletion breakpoints are enriched. Therefore, to ensure physical (rather than artificial, as a result of assembly error) continuity of flanking and inserted (if any) sequences at breakpoints we performed breakpoint filtering by utilizing unmapped reads. For each derived deletion breakpoint we constructed a breakpoint junction sequence by joining 100 bps downstream with 100 bps upstream of the breakpoints. The micro-insertion (if present) was inserted in the middle. The set of all 36,237 junctions sequences from 200 to 298 bps in length comprised the 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 that ungapped alignments are made with at most 3 mismatches and that only unique alignments are reported. Prior to mapping, and in the same way it was done by BWA {REF} during alignment 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 considered in potential support of the junction they aligned to. We chose a particular cutoff *d* on the number/fraction of bases aligned to each flank for deciding, which reads supported 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 the 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 sequences, thereby creating a null junction library, and mapping unaligned reads to that null library. Such a mapping will represent random noise and can be used for optimizing 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 cutoff at which the number of null junction passing the filter was <5% of the number of real junctions passing the filter at the same cutoff (**Fig. S7**), i.e., we aimed for <5% *in-silico* FDR. This criterion led to setting the 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

Alexej Abyzov 5/16/14 10:54 PM
Deleted: 6

breakpoints of different classes by corresponding *in-silico* FDR was (from lowest to highest): NH, TEI, NAHR, and VNTR. This is also expected, as breakpoints of different classes have progressively more repeats around their breakpoints in the same order.

To summarize, the developed an empirical model that captures essential biological features of breakpoints, that is not biased because it uses data loci different from the breakpoints, and that allows translating random mapping into estimated FDR. We suggest that such empirical models can be used to estimate FDR of genotyping known breakpoints from sequencing data. However, when it is 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

We selected 15-22 deletions of each class for PCR validation. Deletions were selected randomly but required to be genotyped in at least two samples out of 319 for which we had DNA available. Here we relied on genotyping by mapping reads to deletion breakpoint sequence junctions. For the selected deletions we designed primers with Primer3 such that the primers would amplify the breakpoint sequence. For each deletion we ran PCR in at least one sample genotyped as having it and sequenced resulting band with Sanger technique. In case the deletion was not confirmed we ran PCR in another sample genotyped as having it.

IRS – Intensity Rank Sum – validation was described elsewhere {REF SV Pilot}. Briefly, the validation considers intensities of SNP probes within deleted regions and correlates it with deletion genotypes across samples. It is expected, that for such SNPs, samples with deletion will have lower intensity values than samples without the deletion. Rank sum tests are performed to access the statistical significance of correlations. IRS only tests the validity of deletion sites and does not provide validation of breakpoints. Results of performing this exercises are summarized in (**Fig. S5**).

Alexej Abyzov 5/16/14 10:54 PM
Deleted: 4

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 the breakpoint sequence, i.e., to the alternative allele. The probes were different in only one nucleotide to mimic probes for SNP genotyping. 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. 2,483 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 to those done by mapping reads to sequence junctions, as we did for filtering breakpoints, revealed that individuals with deletion genotypes by read mapping represent almost a perfect subset of those genotyped by arrays (**Fig. S8**). This is easy to rationalize by noting that individuals in the 1000 Genomes Project were sequenced at a shallow 4-8X coverage, and thus not likely to have many reads covering breakpoint sequences, particularly in the case of heterozygous deletions. Furthermore, the requirement that reads mapped to deletion sequence junction must extend at least 13 bps across the junction in each direction further reduces the number of reads that we consider supporting deletions.

Alexej Abyzov 5/16/14 6:53 PM
Deleted: XXX

Alexej Abyzov 5/16/14 10:56 PM
Deleted: 7

Confirmation of breakpoints in high coverage trios

Breakpoint validation was performed using data for two trios sequenced with HiSeq 2500 at 60X coverage with 250 bp reads. The 8,943 deletions in our confident set were genotyped in trios by CNVnator {REF}, and when genotypes suggested the presence of deletions (estimated copy-number less than 1.5 or less 0.5, for diploid and haploid regions, respectively), corresponding breakpoints were selected for further investigation. Read pairs with coordinates in the 2 kbp vicinity of these regions were

extracted from bam files, and each was tested for an overlap at 3'-ends. If a suitable overlap was detected, the reads were merged into a long continuous (gapless) genomic fragment (see next section).

Using AGE {REF}, we generated split-fragment alignments of such fragments around breakpoints and searched for breakpoint support the same way we did for contig alignment with AGE (see above). We considered breakpoints with such supporting reads as at least partially confirmed. We considered a given breakpoint to have perfect support if read alignments had breakpoint coordinates and micro-inserted sequences (if any) that matched exactly. A breakpoint was considered validated if the majority of split-fragment alignments consistent with the deletion matched the breakpoints perfectly. We confirmed 3,034 (34%) breakpoint sequences perfectly, and for 423 (4.7%) more we observed slight differences in the sequence at breakpoints.

Constructing long genomic fragments

The reads in the HiSeq 2500 data were 250 bp in length, with average insert size of ~400 bp (**Fig. S9**). This means that the reads in most read pairs significantly (50 bp or more) overlapped in sequence at 3'-ends. In our validation method, we merged overlapping read pairs to construct long genomic fragments. To merge a given pair aligned near deletion breakpoints, we needed to estimate the length of its overlapping sequence. We slid the 3'-ends of each read in a pair against each other starting from an overlap of 1 base and continuing up to 250 bases. For a given overlap of length n, we assumed a binomial distribution for the number of mismatches. 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 with the uniform probability for each mismatch of $p_{mismatch} = 0.75$. We only considered merged read pairs that had mismatch counts less than 20% of overlap length, and *p-values* smaller than 10^{-10}. We tested our approach by independently aligning overlapping reads and comparing these overlaps to those from alignment. Consistent overlaps were observed for 99.95% of read pairs (**Fig. S10**). Pairs of reads with identified overlaps were merged into genomic fragments, and bases in overlapping sequences were chosen by taking the base with the higher quality score at positions of mismatches. These genomic fragments were from 250 to 480 bps in length and of higher sequencing quality than either of the reads in the original pair alone.

Alexej Abyzov 5/16/14 10:54 PM
Deleted: 8

Alexej Abyzov 5/16/14 10:54 PM
Deleted: 9

Aggregation calculation

Almost 40 million of SNPs and indels found by the 1000 Genomes Project {REF} in the same group of individuals were aggregated around the breakpoints of each class. To reduce the contamination of our analysis with false positive calls, we only used SNPs and indels that reside in the confident sites as defined by the mask derived by the project. This reduced the number of variants by 25%. SNP density was calculated with respect to the number of such sites. Densities of substitutions at C and G bases were calculated with respect to the number of not masked C and G sites. Densities of substitutions at A and T bases were calculated with respect to the number of not masked A and T sites. Each aggregated density was then normalized to yield density of one in the interval [±500 kbps, ±1 Mbps].

Histone mark data generated by ENCODE {REF} project were used for the aggregation analysis. We utilized contained normalized histone signals provided by the project. Aggregated signal in each bin was normalized with respect to number available bases, i.e., undetermined bases of the reference genome were excluded from the aggregation. Each aggregated signal was then normalized to yield signal of one in the interval [±2 Mbps, ±4 Mbps].

We utilized methylation data generated with bi-sulfide sequencing by {REF Shantao}. The data were provided for only those CpG sites where confident methylation level estimation could be made, which is about 95% of all CpG sites. Aggregated methylation levels were then normalized to the number of CpG site in each bin and then normalized to yield level of one in the intervals [±2 Mbps, ±4 Mbps].

Intersection with open/closed chromatin

We used the Hi-C data generated on human lymphoblastoid cell line (GM06990) (REF PMID: 19815776). In that study, chromatin states were defined from chromatin interaction matrix eigenvectors

that correspond to chromatin states. The matrix was calculated for consecutive non-overlapping genomic bins of 100 kbs in length: negative values represent closed chromatin states, and positive ones represent open status. There were total 29,195 bins with non-zero eigenvector values. We assigned each breakpoint with an eigenvector value by finding the bins they belong to. NAHR breakpoints have higher eigenvector values, indicating a more open chromatin state. Meanwhile, NH and TEI breakpoints show lower values (**Fig. 3A**). To test this hypothesis, we utilized a rank sum test with restricted permutation. Rank sum was defined by the summation of ranks of the eigenvalues of certain breakpoint subtypes. Then the observed rank sum was compared with an empirical distribution generated by circular permutation. That is, we joined the end of the whole genome bin array with the beginning to make it circular, and rotated this circular array to every possible position. We calculated the rank sum for each position. This forms an empirical distribution for the null hypothesis. The p-values are corrected by Bonferroni method for multiple testing, i.e., testing for three sets of breakpoints: NH, TEI, and NAHR.

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. Then we compared the counts with the empirical distribution generated by circular permutation generated the same way as for comparison with open/closed chromatin (see above).

Mapping template sites

The majority of micro-insertions are less than 10 bps in length (**Fig. 4A**). These are likely to be explained by the existence of base mismatches 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 reports them as a 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 in this study on a larger scale (**Fig. 1**). We therefore performed the following analyses for micro-insertions longer than 10 bps.

We first uniquely mapped MIs with up to one mismatch to the reference genome using Bowtie {REF} with the following options '-n 0 -l 5 -r --best --strata -v 0 -m 1'. Next, not mapped MIs of at least 20 bases in length were aligned to the reference genome by Blat {REF}. We then manually examined alignments and selected only one such that: i) MI is aligned almost full length with few mismatches and/or short indels, ii) the alignment has much better alignment score than other alignments. In total we mapped 133 template sites, of which 66 were mapped manually (**Table S3**).

Replication time analysis

We utilized data by Koren et al. {REF Koren}, which had average replication timing from 3 experiments. Using the data we identified replication time to each breakpoint and template site. Difference in replication time can be calculated relative to each breakpoint. We use the difference that is smaller in absolute value.

Calculating association with recombination rates

Recombination rate data was derived from the Rutgers third generation genetic map. We used the sex-averaged genetic positions, ignoring the X and Y chromosomes. Genetic positions were divided by the difference in adjacent physical positions in the map in order to obtain values in terms of centimorgans per base pair (cM/Bp). Linear interpolation was performed to obtain recombination rate values for each base of each chromosome. Significance values we obtained by conducting a circular permutation experiment in the same fashion as for intersection with open/closed chromatin (see above).