RESPONSE LETTER

Referee 1.1 – General positive comment

Reviewer comment	The study by Abyzov, Gerstein and colleagues describes an ambitious effort to identify and characterize structural variant (SV) breakpoints from a large population (N=1,092) sequenced as part of the 1000 Genomes Project. Analysis of 8,943 breakpoints yielded significant insights into three known mutational mechanisms (NH, NAHR and TEI) and certain factors intrinsic to the genome that predispose to structural mutation.
Response	We thank the reviewer for the thorough evaluation of our manuscript.

Referee 1.2 – Replication/division (Minor #1)

Reviewer comment	"We hypothesize that NAHR deletions occur without replication in embryonic and germline cells." Don't all events that occur at meiosis (which is when most NAHR events are thought to occur) occur in germline cells without replication?
Response	We are not entirely certain about the critique here. We think the referee emphasizes that NAHR events are believed to occur during chromosome segregation, i.e., after DNA replication. We agree. What we meant is that NAHR deletions can also occur without cell division. We will clarify this in the manuscript.
Excerpt from manuscript	*We hypothesize that since replication is largely devoid of chromatin structure, some of the NAHR deletions occur without DNA replication and cell division, in embryonic and germline cells, and then are passed on to the offspring.*

Referee 1.3 – Region predisposition to mutations (Minor #2)

Reviewer comment	The results of this study suggest some intriguing commonalities between SNPs and SVs that should be discussed. It has been shown that regions of open chromatin (nucleosome free DNA) are associated with higher rates of nucleotide substitution (Michaelson, Sebat Cell 2012). This study and Michaelson et al suggest that the genomic features that predispose to certain classes of SV also predispose to certain classes of nucleotide substitution (e.g. NAHR correlates with C->T, NH may correlate with most other mitotic SNV events. This interepretation is consistent with Fig 2B).
Response	Good suggestion. We now discuss this possibility in the text.
Excerpt from manuscript	*... Alternatively, one can suggest that genomic features (e.g., nucleosome free DNA) that predispose to certain classes of SNPs may also predispose to certain classes of SVs. ... The classical NAHR mechanism postulates meiotic cell division as a requirement for generating a germline SV. This implies certain associations that we did observe in our study. In particular, NAHR breakpoints were associated with higher recombination rates, with higher GC content and with higher density of CpG motifs.*

Alexej Abyzov 1/27/2015 7:17 PM

Deleted: *one can suggest that genomic features, like nucleosome occupancy[35] and replication timing[34], that predispose to certain classes of SV may also predispose to certain classes of nucleotide substitution. We did not observe association of breakpoint in any class with DNAse hypersensitive sites. Also, differential association of breakpoints in each class with replication timing (i.e., NAHR with early, NH with late, and no association for TEI) and association of SNP with late replication, make it unlikely that replication timing explains co-occurrence of SNPs and deletions. Though, replication timing can be a contributing factor to association of NH breakpoints and SNPs ...*

Referee 1.4 – Variant co-occurrence from genomic features (Minor #3)

Reviewer comment	Likewise, the correlation of SV breakpoints with SNPs may not be driven entirely by selection (as is suggested in the paper). The correlation may also reflect that they are associated with common genomic features. No?
Response	We agree and this follows from the analysis in our manuscript. We clarified and explicitly state this.
Excerpt from manuscript	*... Alternatively, one can suggest that genomic features (e.g., nucleosome free DNA) that predispose to certain classes of SNPs may also predispose to certain classes of SVs. While indeed we see such associations (e.g., for NAHR breakpoints), however, except for reduced conservation, we did not find any other feature that would be universally associated with breakpoints in each class.*

Alexej Abyzov 1/27/2015 7:28 PM

Deleted: *... 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 **S5**). 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 ... one can suggest that genomic features, like nucleosome occupancy[35] and replication timing[34], that predispose to certain classes of SV may also predispose to certain classes of nucleotide substitution. We did not observe association of breakpoint in any class with DNAse hypersensitive sites. Also, differential association of breakpoints in each class with replication timing (i.e., NAHR with early, NH with late, and no association for TEI) and association of SNP with late replication, make it unlikely that replication timing explains co-occurrence of SNPs and deletions. Though, replication timing can be a contributing factor to association of NH breakpoints and SNPs ...*

Referee 1.5 – Smaller scale and nucleosome occupancy (Minor #4)

Reviewer comment	The comparison of breakpoints and chromatin states was performed on relatively coarse (kilobase and Megabase) scales. Some features (e.g. nucleosome occupancy) vary on smaller scales (100 bp). SV breakpoints

	might show different patterns for fine and course features. Based what we know about nucleotide substitutions, one might predict that, while NH breakpoints are depleted for active chromatin marks and are somewhat correlated with closed chromatin... on a fine a scale NH breakpoints might still be associated with exposed (nucleosome free) DNA. Are they?
Response	We thank the reviewer for this useful suggestion. We have aggregated at smaller scale DNAse signal for hESC cells and nucleosome occupancy for NA12878 cell line around breakpoints of different classes. No clear enrichment of depletion of the signal around NH breakpoints was observed. However, we observed association of TEI breakpoints with nucleosome free DNA, and association of NAHR breakpoints with accessible DNA.
Excerpt from manuscript	*We correlated our breakpoint with DNAse hypersensitive sites and with nucleosome free DNA (Fig. S9). DNAse data revealed association of NAHR breakpoints with accessible DNA at a kilobase range. Nucleosome occupancy data further uncovered preference of NAHR breakpoints to reside in nucleosome fee regions. Analysis of the both data types revealed no association with NH breakpoints, but depletion of TEI breakpoints in nucleosome occupied and DNAse accessible regions.*

Alexej Abyzov 1/27/2015 7:32 PM
Deleted: and NT2 lines

Alexej Abyzov 1/27/2015 7:33 PM
Deleted: nucleosome

Alexej Abyzov 1/27/2015 7:34 PM
Deleted: *TO BE COMPLETED WHEN TEXT IS FINALIZED.*

Referee 2.1 – General positive comment

Reviewer comment	In this paper the authors identify, classify, and analyze 8,943 breakpoints associated with deletions in 1,092 samples sequenced by the 1000 Genomes Project. The study is well designed and well written.
Response	We thank the reviewer for the thorough evaluation of our manuscript.

Referee 2.2 – Reanalysis of previously published dataset

Reviewer comment	The paper attempted to characterize variant formation mechanisms(NAHR, NH, and TEI) through breakpoint mapping. This has been done in numerous papers before. I would appear to be just a further characterization of SV from the 1000 Genomes Project data, which have already been investigated in other studies … The authors mentioned the existence of other papers that did similar analysis using the same data. The improvement compared to the other published papers on 1000G SVs, is basically the high resolution breakpoint mapping that improved the prediction of variant formation mechanism, although they failed to show that. The paper is a resource based paper using previously published datasets. Integrating methylation, HI-C and histone marks data and offers no identifiable or significant message. The findings although interesting, have mostly been reported already in other studies.
Response	In our analysis we **did not** use a previously published breakpoint set. Rather, we derived a new set. The pilot phase of the 1000 Genomes Project used very early next-gen data with very short reads. Such short reads limited our ability to resolve breakpoints and particularly micro-insertion at breakpoint junction (**Fig. S3**). Phase 1 derived an intermediate set of breakpoints, but it was limited by the requirement to have deletions with genotypes, which in turn allowed for only large and non-repetitive CNVs in the set. The set used in this study addresses the shortcomings of the two previous sets. Specifically, we filtered candidate breakpoints by mapping read to their sequences thereby ensuring the sequences' continuity (including micro-insertion). And this was done for deletions across entire size range. The overlap between confident set and pilot/integrated sets was roughly 50% (**Fig. S3**). We would also like to emphasize that there were no publications of breakpoint analysis that would match the scale of the 1000 Genome Project. And while previous studies were biased towards analysis of SV spectrum in one or just a few individuals, our study is unique in that we analyzed breakpoints in the unprecedented large population. However, we do see that we did not succeed in getting this point across to referees, even though **Fig. 1C** was specifically made to address such questions. We now revised the text, added supplementary **Fig. S3** and did our best to clarify this. Below we provide our response to specific comments. Also, while we recapitulated some previously known reported results, we also reported multiple novel findings. We point out that the referee did not question our novel analysis of co-association of breakpoints with SNPs and the relation of breakpoints and template sites with replication timing.
Excerpt from manuscript	*Overall, these breakpoints are of higher quality than those derived in the pilot phase of the 1000 Genomes Project[17] (**Fig. S3**) and are more representative in their length distribution than those used recently in the following phase[21] (**Fig. 1C**), as the latter set was limited to large non-repetitive events that could be well-genotyped across the analyzed populations. A large fraction of the dataset, 3,739 (42%), were deletions of at least a thousand bases in length. This set was also significantly larger than those analyzed previously[14-16,24,25] (**Table S2**).*

Referee 2.3 – How new breakpoint set is different?

Reviewer comment	The authors suggest that the breakpoints are of much higher quality than those derived in the pilot phase, without evidence to back this up. They also don't differentiate how much different this refined set is compared to the data presented in the phase 1 dataset. After comparing a random selection of examples from the data in Table S1 vs. entries in 1000G phase 1, it was observed that in many cases the breakpoints are identical. How often were the original breakpoints improved? … Also, the authors suggest it is the largest collection to date, although it would appear as if more variants were analyzed in the Pang et. al paper (Human Mutation, 2012) and in the Mills 2011 paper (reference 17 where > 10,000 validated).

Response	See response to comment 2.2. Also, we now provide comparisons of the datasets in **Fig. S3**. Note, Mills et al. 2011 data set is the pilot set. Venter's genome variant set (Pang et al.) has indeed close to a million variants but the majority of the variants are indels, with breakpoints resolved for less than 7,000 of SVs larger than 100 bp, and for less than 1,000 of SVs larger than 1 kbp, while we present almost 9 thousands breakpoints longer than 100 bps, and almost 4,000 longer than 1 kbp. Additionally, only 1,516 variants from Venter's genome had 50% reciprocal overlap with deletions in our set.
Excerpt from manuscript	*This set was also significantly larger (when counting variants larger than 100 bps) than those analyzed previously (Table S2).*

Referee 2.4 – Accuracy of breakpoints

Reviewer comment	It would be interesting to have a measure of breakpoint accuracy for those which were validated by PCR. What were the confounding factors where the predicted and confirmed breakpoints were different? ... For the deletions confirmed by PCR, but where the breakpoints were not accurate, what accounts for the inaccuracy? Are there any specific reasons for breakpoint inaccuracy, especially considering that they avoid smaller deletions and those in repeats?
Response	We now elaborate on this. The major confounding factor is repeats. Misassembly of breakpoint sequences typically results in derived breakpoints being shifted relative to the real ones.
Excerpt from manuscript	*Precision was confounded by repeats around breakpoints. Typically, we observed shift between breakpoint coordinates from assembly and validation, but in one case we observed that assembly collapsed repeats (**Fig. S1**).*

Referee 2.5 – Data fall out

Reviewer comment	In general the authors avoid difficult regions including limiting the dataset to variants greater than 1kb, and only identifying micro-insertions that are larger than 10bp. In both cases, the majority of the data fall outside the categories analyzed, resulting in a much smaller and partial dataset.
Response	We believe it is a misunderstanding. It is apparent that roughly half of our data set consists of CNVs smaller that 1 kbp (**Fig. 1C** lower panel). We also found micro-insertion (MIs) ranging in length from 1 to 96 bps. For downstream analysis we considered MIs longer than 10 bps for two reasons: (i) short MIs could be the result of misinterpreting SNPs or indels close to breakpoints; (ii) short MIs are very likely to be mapped to multiple places in genome, and such mapping is not informative for the analysis of template sites. Even about half of longer MIs could not be mapped to the genome unambiguously (original Table S3, now Table S5).
Excerpt from manuscript	*A large fraction of the dataset, 3,739 (42%), were deletions of at least a thousand bases in length ... In our dataset we observed 2,391 (27%) deletions with micro-insertions ranging in length from 1 to 96 bps with the majority being less than 10 bps in length (**Fig. 4A**).*

Referee 2.6 – Mostly deletions

Reviewer comment	The study only characterizes deletions (with a small handful of bona fide insertions from the TEI category), without any information on duplications or inversions which would be interesting as previous reports have shown significantly different patterns based on the variant type.
Response	In this case we are limited by the generated call sets that contained only deletions. We now revise the text and title to make clear that the conducted analysis is for deletions only.

Referee 2.7 – Subcategorization

Reviewer	With such a large number of variants with nucleotide resolution breakpoints, why did the authors only

comment	investigate 3 broad classes? Although mentioned, the authors did not attempt to subcategorize variants from the NH processes (FoSTeS, MMBIR, NHEJ).
Response	In fact, in the original manuscript we were able to subcategorize some events in NH class, e.g. those for which we were able to find template sites of micro-insertions (original Table S3). To elaborate on this, we now provide rough estimates of the proportions of deletions likely generated by template switching mechanisms in the NH class, along with examples of deletions likely created by other mechanisms, such as retrotransposition-mediated deletions and deletions generated through recombination across right arms of two oppositely oriented *Alus.*
Excerpt from manuscript	*Large fraction of NH deletions (58%) had evidence of being generated though template-switching mechanisms, i.e., contained at least 2 bp identity around breakpoints or MI longer than 10. The remaining NH deletions are likely to arise through NHEJ. MI of one deletion (chr1:200,258,970-200,259,149) consisted of the sequence of 3'-end of Alu element and 21 bp long poly-A tail, thus, is likely templated from RNA of Alu element[26]. We also identified a deletion (chr17:1654955-1655422) generated with breakpoint signature of recombination across right arms of two oppositely oriented Alus[27]. Overall, deletions in this set were generated though variety of mutational mechanisms.*

Referee 2.8 – Comparison with previously published results?

Reviewer comment	How do the results here compare to previous analyses (references 14-17). Aside from listing the number of sv breakpoints in the introduction, it would be helpful to have a comparison of the breakdown in various classes compared to previous studies to show how this study is an improvement.
Response	We now provide such comparison in Table S2. Table demonstrates that we collected the largest count of NAHR, NH and TEI events out of all studies. We intentionally eliminated VNTR, as short read technologies do not allow confident breakpoint resolution for such SVs, and validation with PCR is rarely possible. The largest collection of breakpoints for these three different mutational mechanisms allowed us to conduct analyses that were not performed previously: co-aggregation of SVs with SNPs/indels, relation of SVs with open chromatin and histone marks, and analysis of micro-insertion template sites.
Excerpt from manuscript	*This set was also significantly larger than those analyzed previously[14-16,24,25] (**Table S2**).*

Referee 2.9 – Confirmation rate

Reviewer comment	Why were only 28% of the breakpoints confirmed with the array, and 39% of breakpoint sequences in trios. Seems like a low confirmation rate. It isn't clear if this is a fraction of all variants tested by each approach or a fraction of the entire set of deletions? If the latter is true, how often did each approach fail to validate the breakpoints?
Response	We realize that this was not clearly explained. Confirmation was done in a limited number of samples, while the denominator was the count of all breakpoints in all individuals. We fix now provide the requested numbers.
Excerpt from manuscript	*Using read depth approach, we genotyped 4,384 variants from the set as deletions in two trios sequenced to high coverage by long reads. Using these data as supporting evidence we confirmed 3,034 breakpoint sequences (34% of entire set) and, after minimizing confounding factor, calculated yet another FDR estimate of 18% for deletion presence with precise of breakpoints (**Table S1** and **Methods**).*

Referee 2.10 – Type (Minor)

Reviewer comment	Page 3: We used these two additional date (replace with "data") sources.
Response	Thanks. We fixed it.

Referee 2.11 – Figure improvement (Minor)

Reviewer comment	Figure 3B should be placed before Figure 3A
Response	Fixed it.

Referee 3.1 – General positive comment

Reviewer comment	This is a comprehensive study of deletion breakpoints in the 1000 genomes project samples based on a combined analysis using 5 different software packages for identifying indels, and then local alignment of these regions to identify the breakpoint sequences.
Response	We thank the reviewer for the thorough evaluation of our manuscript.

Referee 3.2 – General critical comment

Reviewer comment	Overall there is a mix of novel and previously reported findings presented. In several places results shown clearly recapitulate previous observations and not novel. In other places I had some major concerns with either the methods or the conclusions that were drawn from the data, and I found some of the approaches used inadequate to support the presented results/conclusions. ... Overall while of interest, I thought the manuscript has many weaknesses that need improvement.
Response	We did recapitulate some previously know reported results, as this is a standard scientific practice. Below we provide responses to specific comments, which we hope, will clear some confusion and highlight novel findings.

Referee 3.3 – Additional analyses

Reviewer comment	I would like to see the authors present much more data on the breakpoint dataset on which the entire study is based, with particular emphasis and clearer explanation of the deletions in terms of allele frequency, genomic location, and how many were called by the 5 different approaches used. This will allow the reader to gain insight into the results and analysis shown which is currently lacking.
Response	We thanks the referee for this comment and now report the additional analyses requested.
Excerpt from manuscript	*As expected, we find exponentially more of less frequent deletions, with roughly 54% genotyped in less than 2% of studied individuals (**Fig. S2**). Using OMNI genotyping arrays we estimated that our breakpoint genotyping while being very precise misses roughly 60% of samples; the results of shallow 4-8X sequencing. Additionally, due to stringent criteria for breakpoint support, breakpoints of rare deletions are less likely to be confirmed by read mapping. As a consequence, frequency spectrum of deletions in the set was shifted toward more common events as compared to the SNP set discovered from the same data (**Fig. S2**) ... Around 16% of deletions were present in only one initial call sets merged (**Fig. S11**), stressing that the majority of deletion sites were detected by multiple algorithms.*

Referee 3.4 – Mostly deletions

Reviewer comment	This study looks only at deletions. This is not a problem, but the results might be different if other types of SV were studied. As a result, I think the title should perhaps be revised to make it clear that this is specifically a study of deletions only, and not SVs in general.
Response	In this case we are limited by the generated call sets that contained only deletions. We now revise the text and title to make clear that the conducted analysis is for deletions only.

Referee 3.5 – More info on the dataset

Reviewer comment	More details needed on deletion calls to ascertain the quality of this dataset. MAFs, are they heritable/show Mendelian inconsistencies in trios, or fit with HWE? What fraction were unique to single individuals, and how does the frequency spectrum compare with SNPs?
Response	Shallow 4-8X sequencing and stringent criteria for deletion breakpoint support make determination of complete (i.e., with knowledge of heterozygous and homozygous states) deletion genotypes across population and extremely challenging task. We, therefore, report on deletion frequency spectrum per individuals and compare it with the one for SNPs. Compare to SNPs our deletion set is biased towards more common deletions. There are no signletons in our set, as we required supporting reads from two individuals to exclude somatic variants (that are unlikely to have exact same breakpoints in unrelated individuals) and reduce experimental errors (that are less likely to be the same in two independent

	samples). There were no trios in the 1,092 individuals. For two trios that we used for validation, the Mendelian inconsistency rate of CNVnator genotypes was ~7%.
Excerpt from manuscript	*As expected, we find exponentially more of less frequent deletions, with roughly 54% genotyped in less than 2% of studied individuals (**Fig. S2**). Using OMNI genotyping arrays we estimated that our breakpoint genotyping while being very precise misses roughly 60% of samples; the results of shallow 4-8X sequencing. Additionally, due to stringent criteria for breakpoint support, breakpoints of rare deletions are less likely to be confirmed by read mapping. As a consequence, frequency spectrum of deletions in the set was shifted toward more common events as compared to the SNP set discovered from the same data (**Fig. S2**) ... Around 16% of deletions were present in only one initial call sets merged (**Fig. S11**), stressing that the majority of deletion sites were detected by multiple algorithms.*

Referee 3.6 – Purifying selection

Reviewer comment	Further to this, in the discussion the authors state the purifying selection likely underlies the distribution of deletions they observe in the genome. this is why the allele frequency spectrum is important to know. It is already well documented that indels and other damaging variants show purifying selection, and tend to be rarer than less deleterious variants. As such, it naturally follows that one would expect them to be enriched in less conserved regions of the genome
Response	We'd like to point out that along with purifying selection we discussed other factors that underlie the distribution of deletions (e.g., open/closed chromatin, CpG motifs, etc.). We specifically mentioned purifying selection in relation to all breakpoint classes being enriched in less conserved genomic regions. But we do agree with the logic outlined by the reviewer. We actually used it but in the reverse direction, i.e., enrichment of deletion in less conserved regions suggests purifying selection acting on them. In response to the reviewer's comment, we now provide frequency of deletions in the population (**Fig. S2**) in comparison to SNPs' one. Frequency spectrum of deletions in our set is shifted towards more common events, which is consistent with deletions being enriched in less conserved regions due to purifying selection.
Excerpt from manuscript	*frequency spectrum of deletions in the set was shifted toward more common events as compared to the SNP set discovered from the same data (**Fig. S2)***

Referee 3.7 – Confirmation rate

Reviewer comment	Page 3: The authors state: We used these two additional date sources as supporting evidence, and confirmed 28% of the breakpoint sequences with the array, and 39% of breakpoint sequences in the two trios." This sentence is rather ambiguous in its meaning. One way to interpret this is that 72% of breakpoints with array and 61% of sequences with the trios did NOT validate, which would suggest a high false positive rate. Or do instead the authors mean that they could only look at 28% and 39% of the breakpoints they describe because the platform used did not cover the remainder of predictions? If this latter case is what they mean, then the sentence is not really that useful, as simply to tell the reader what fraction of sites overlapped with a SNP array is not very informative. If this is the case, what I think the sentence should say is what was the validation rate of the sites that could be investigated. Note there is also a typo here, which I think should read "data"
Response	We revised the text to clarify this. Please also see response to comment 2.9.

Referee 3.8 – Reduced selection vs co-occurrence

Reviewer comment	Page 4, the authors state that the likely explanation for the co-occurrence of deletions and SNPs is reduced selection. An alternative (or complimentary) explanation might be that both SNPs and indels co-occur in regions of late replication and/or those with higher recombination frequency. In fact this association has been reported before, see PMID: 23176822, and should be acknowledged clearly. Can the authors perform an analysis of these factors, which might better explain their observation over that currently discussed?
Response	Agree. This is an alternative/complementary explanation. We now acknowledge it. Association of NAHR and C to T SNP was described in the original submission. We also now reason in the text that replication timing is not the major factor for co-occurrence of SNPs

	and breakpoints.
Excerpt from manuscript	*... one can suggest that genomic features (e.g., nucleosome free DNA) that predispose to certain classes of SNPs may also predispose to certain classes of SVs. While indeed we see such associations (e.g., for NAHR breakpoints), however, except for reduced conservation, we did not find any other feature that would be universally associated with breakpoints in each class.*

Referee 3.9 – Methylation analysis

Reviewer comment	"We next searched for an association of deleted regions with hypomethylated regions in sperm as compared to H1ESC26. A strong association was observed for TEI and NAHR breakpoints (Fig. 3B)." There is a strong inherent confounder here for both of these associations. As the authors point out, transposable elements (TEs) are constitutively hypomethylated in sperm, and thus anything that looks at TEs compared to the rest of the genome will always get an answer that says "enriched for hypomethylation", Similarly for NAHR breakpoints, as the authors point out in the preceding section, NAHR events are highly enriched for CpGs, which tend to correspond to CpG islands, most of which are also constitutively unmethylated in sperm (Ref 27). I think the authors need to consider this confounder carefully, rather than leading the reader to conclude that this is maybe a causal relationship, which the current dataset does nothing to prove
Response	Hypomethylation around TEI variants is expected, as these are transposable elements. Thus, in our view, the observed association is not confounded rather consistent with general knowledge. Association of NAHR with hypomethylation is a novel finding. We thank the referee for suggesting an alternative explanation of our methylation analysis. We now performed additional analyses to see whether CpG islands can explain decreased level of methylation around NAHR breakpoints. We excluded genomic regions around CpG islands from analysis but did not see significant change in intersection of NAHR and hypomethylated regions (Fig. Sx). This is easy to rationalized as only 4.6% of NAHR breakpoints were within 2kbp of CpG islands. In fact, Molaro et al., Cell, 2011, also noticed that overlap between hypomethylated regions and CpG islands in small, 24-27% (Fig. 1B in Molaro et al.). Furthermore, just to clarify, we suggested a possible causal relationship of open/active chromatin and NAHR breakpoints based on a few lines of evidence. Methylation analysis was one of such evidence but not the only one.
Excerpt from manuscript	*Such an enrichment for TEI (mostly Alus) could reflect the long-standing observation of demethylation of Alu elements in sperm[34]. Alternatively, the enrichment could reflect a preference of transposon integration complexes for hypomethylated DNA, as has been observed in somatic TEIs in cancer genomes[35]. Similar enrichment for NAHR deletions is consistent with the reduced C to T substitution densities in CpG regions around the deletions' breakpoints. This observation is not confounded by CpG islands, most of which are also constitutively unmethylated in sperm (**Fig. S8**).*

Referee 3.10 – NAHR and recombination

Reviewer comment	Page 6: The authors state "Similarly, we observed a strong correlation of recombination rate with NAHR Breakpoints in closed chromatin (Pearson Coefficient 0.94)". I'm rather troubled by this result. A correlation of 0.94 of recombination rate with NAHR breaks in closed chromatin implies that nearly all NAHR sites reported are explained by recombination rate. How was this even calculated? Breakpoints are surely a discrete trait (presence/absence), so how does one perform a Pearson correlation?
Response	We divide genome into bins and correlate number of breakpoints (i.e., breakpoint density) and average recombination rate within each bin. We agree that high correlation does imply that that nearly all NAHR site are explained by recombination rate. But this is observed only for closed chromatin. In other this is a conditional correlation. For open chromatin it is not the case (**Fig. 3B**), and this is one of the lines of evidence for our hypothesis that NAHR occurs without cell division.

Referee 3.11 – Proximal and distal

Reviewer comment	I have some serious issues with the section "Micro-insertions at breakpoint deletions and their relation to replication timing". For Fig 4C, I am having a lot of trouble believing that this is a real observation. Although we as geneticists have constructed maps of chromosomes where we number bases from the tip of the p-arm, through the centromere, down to the end of the q-arm of each chromosome, our classifications of what is therefore "proximal" and "distal" to my knowledge has little relevance to biological processes of DNA replication and rearrangement that are being discussed here as they occur in cells. Mammalian DNA replicates via mutliple origins per chromosome that proceed along each chromosome that to my knowledge are largely unrelated to the arbitrary definition of what is a "proximal" or "distal" direction. What is the rationale therefore that relative orientation should even be relevant here? Is this difference between proximal and distal shown in Fig 4C really significant?
Response	Our understanding is that the referee thinks that “proximal” and “distal” relates to distance from centromeres. We meant to refer to the distance of the template site from the closest deletion breakpoint. We believe this confusion is purely because of the terminology, and we now clarify this by calling template sites as “adjacent” and “distant”. Provided that the confusion is resolved, we believe the question about the difference between proximal/adjacent and distal/distant template sites is not relevant. But it is significant that the distribution of templates site relative to breakpoints is bimodal.
Excerpt from manuscript	*The template site was typically located either between 20 to 60 bps (adjacent site) or between 2 to 6 kbps (distant site) of one of the breakpoints.*

Referee 3.12 – Resolution of replication timing measurements

Reviewer comment	Second, I also am very troubled by the data shown in Fig 4D. How was replication time determined at this level of resolution necessary for this test? Fig 4C shows the vast majority of templates are <10kb from the breaks, and often <1kb. The study of Koren et al only reports replication timing in 100kb intervals, so I do not see that the relative resolution of this dataset to the template sites is in any way meaningful. The authors even allude to this in the same paragraph. Also as they point out, its not even a fair question to if templates within the deletion have a different replication time to the deletion itself, so why state that it was not significant in the way that is done, contrasting it with templates that occur outside of the deletion region? I find it very misleading to state therefore that "the same effect was not significant for template sites within deletions", as this is not a reasonable question to even ask.
Response	We believe that this is a misunderstanding. Data by Koren et al. (ref 30, PMID:23176822) are of about 1 kb resolution. The file with normalized replication timing (http://genepath.med.harvard.edu/mccarroll/datasets.html) contains almost 2.4M genomic intervals, which corresponds to roughly 1.2 Kbp average interval size. Also, here is the quote from the Koren et al. “*We defined varying-size, equal-coverage chromosomal windows as segments covered by 200 reads in the G1 fraction and counted S phase reads in the same windows. The average size of these segments was ~2 Kb.*” This allowed us to conduct and report a novel analysis of the association of micro-insertions with replication timing.

Referee 3.13 – NAHR and open chromatin

Reviewer comment	Discussion. The authors state there is a paradox in the association of NAHR with open chromatin, as NAHR occurs at a point where no transcription occurs. However, this ignores the fact that NAHR is associated with higher CpG content, which itself is a correlate of promoter/regulator regions. Also many gene families arose by duplication and are polymorphic in copy number via NAHR, thus setting up a further link between NAHR and transcription which is potentially relevant here. I think this conclusion is rather naive and not well supported by the data
Response	We respect the referee’s deep thinking about this point, as it is one of the important points in our manuscript. However, we stated that from the classical view of germline NAHR occurring during cell division one would not expect association with open chromatin. And the paradox is that we do see such an association. Regarding the comment of association

	with CpG: it is consistent with both the classical view of NAHR and the one hypothesized here (i.e., NAHR without cell division), as CpG content is associated with both recombination hotspots and promoter regions. The link between gene duplication families and NAHR that the referee suggests does not extend to the association of NAHR with open chromatin, as it lacks the comparison between genic/open and intergenic/closed regions. We therefore do not see how our statement is naïve and we made it based on several lines of evidence: direct comparison with chromatin state, association comparison with active chromatin marks, correlation with recombination rate, and association with early replication timing.
Excerpt from manuscript	*The classical NAHR mechanism postulates meiotic cell division as a requirement for generating a germline SV. This implies certain associations that we did observe in our study. In particular, NAHR breakpoints were associated with higher recombination rates, with higher GC content and with higher density of CpG motifs. However, and unlike other classes, they were also associated with open chromatin, higher DNA accessibility and active histone marks in mitotically dividing cells. This poses a paradox. No defined structure of DNA exists at the time of chromosome segregation[39] and histone marks are gone[40], thus, no association of breakpoints with open/active chromatin is expected. In fact, as a result of purifying selection one might expect an inverse relation of breakpoints with open chromatin and active histone marks, such as in the case of NH breakpoints. Neither recombination rate nor the fraction of bases in segmental duplications or in repeats explain these associations for NAHR breakpoints. 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.*

Referee 3.14 – Association of NAHR with de-methylation

Reviewer comment	Discussion. The authors state "Additionally, we found two lines of evidence associating NAHR breakpoints with de-methylation: lower frequency of C to T SNPs in CpG motifs and an enrichment with de-methylated regions in sperm" and "We, thus, argue that the observed association of NAHR breakpoint with de-methylation..... is real...". Again, here there are major confounders due to the unusual methylation landscape of sperm in reaching this conclusion that are not considered properly. It may be true that there are less C>T SNPs and enrichment for demethylated regions in regions of NAHR, but correlation is distinct from causation, and there is nothing shown in this manuscript that shows causation. More importantly, the authors own analysis of actual methylation levels around breakpoints (Fig S3) shows absolutely no evidence of association of hypomethylation with indels. The legend to Fig S2 even states this quite clearly "There is no noticeable change in methylation level around breakpoints of either class." As such I am not sure why the authors state the opposite here in the Discussion, and I think it is wrong to say that the data supports this association, as it will tend to perpetuate the false conclusions of Li et al. In my opinion this conclusion is not well supported by the data presented here and should be removed.
Response	We agree that that correlation is distinct from causation. Our point was that the association is not due to technical artifacts. And yes we didn't see deletion breakpoint association with methylation in hESC but we did see association of NAHR breakpoints with hypomethylation in sperm. Later, when combining multiple associations/lines of evidence we hypothesize a relationship between open chromatin and NAHR breakpoints. We now clarify the text of the manuscript and report on additional analyses, which, as we understand, were suggested in comment 3.9.
Excerpt from manuscript	*DNA methylation levels from H1ESC line showed no change close to breakpoints of all classes (**Fig. S7**). We next searched for an association of breakpoints with hypomethylated regions in sperm as compared to H1ESC[33]. A strong association was observed for TEI and NAHR breakpoints (**Fig. 3A**). In particular, the TEI breakpoints were five times and NAHR breakpoints were over 50% more likely to reside in hypomethylated regions than expected by chance (both p-values < $2x10^{-4}$). Such an enrichment for TEI (mostly Alus) could reflect the long-standing observation of demethylation of Alu elements in sperm[34]. Alternatively, the enrichment could reflect a preference of transposon integration complexes for hypomethylated DNA, as has been observed in somatic TEIs in cancer genomes[35]. Similar enrichment for NAHR deletions is consistent with the reduced C to T substitution densities in CpG regions around the deletions' breakpoints. This observation is not confounded by CpG islands, most of which are also constitutively unmethylated in sperm (**Fig. S8**).*

Referee 3.15 – Validation by read depth

Reviewer comment	Methods, deletion validation by read depth. Validation by read depth should vary depending on size of the deletion. Small deletions will be less likely to validate, while large ones should be easily detected. This section says only 34% of breakpoints were validated, which is quite a low rate. And this is after choosing only those

	that the read depth supported the presence of a deletion. What fraction of calls did not even show any evidence of a deletion? Can the authors give more information? If they focus on larger deletions, is the validation rate better? When a breakpoint does "not validate" what exactly does that mean? No deletion was seen by read depth, or just the boundaries of it appeared different? Overall I would like to see clearer evidence of the quality of the calls that form the dataset presented here.
Response	The reviewer is absolutely correct. We now provided description of such analysis. Indeed, for smaller deletions confirmation rate was lower. To overcome confounding effect that reviewer is mentioning, like possible misgenotyping by read depth, we used large deletion genotyped in at least 3 high coverage samples, to provide yet another FDR estimate for our breakpoint set.
Excerpt from manuscript	*Our ability to confirm breakpoints was confounded by incorrect genotypes (i.e., deletion not present in a sample but genotyped as such), as we observed a lower confirmation rate for smaller deletions (**Fig. S16**). An additional confounding factor was the limited ability to construct long reads, because the 3'-ends had high sequencing error, and reliable overlap for paired reads could not be found. In particular, less than 30% of considered pairs of reads have an identifiable overlap. Therefore, unconfirmed breakpoints could be categorized as: i) just false breakpoints; ii) true breakpoints but with incorrect genotype; or iii) true breakpoints with correct genotype, but with no constructed long reads covering the junction. To estimate FDR of the set we minimized the number of breakpoints in the latter two categories by considering deletions larger than 10 kbp and genotyped in at least 3 individuals. This resulted in a FDR estimate of 18% for deletion presence with correct breakpoints.*

Referee 3.16 – Methylation normalization

Reviewer comment	Methods: Authors state the methylation levels were normalized to number of CpG in each bin and normalized to in each interval. Why was this done, and how? I do not see that one can easily normalize methylation levels in this way, or why one would even want to. I am concerned that doing so would introduce artifacts in the data
Response	Normalization by the count of CpG bases is necessary to account for different sequence content in each bin. Furthermore, in the original manuscript we normalized the aggregation signal to have a genomic average of one. Such normalization only scales the signal, without changing its shape. But we agree it is not essential in our case and we now display methylation signals for non-normalized data.

Referee 3.17 – Grammar (Minor)

Reviewer comment	Page 5: " But similar effect for NAHR deletions" is poor grammar. Revise.
Response	We revised it.

Referee 3.18 – Figure improvement (Minor)

Reviewer comment	Legend to Fig4. what is MN an abbreviation for?
Response	Fixed it.