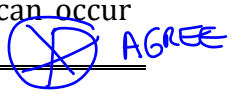
RESPONSE LETTER

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	Our statement was not entirely clear. What we meant is that NAHR deletions can occur without cell division. We will clarify it in the manuscript.



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 will discuss this possibility in the text.

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 analysis in our manuscript. We will clarify and explicitly state this.
Excerpt from manuscript	*Increase of C to T substitutions around NAHR breakpoints is driven by SNPs in CpG motifs as evident from red bars. Furthermore, this 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.*

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 will perform such analysis and answer the question.

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	We believe the referee is confused. In our analysis we **did not** use previously published breakpoint set rather derived a new set. While we recapitulated some previously know reported results we also reported multiple novel finding. Below we provide response to specific comment. Though, we should point out that the referee did not question our novel analysis of co-association of breakpoints with SNPs and relation of breakpoints and template sites with replication timing.

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). … 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 will provide detailed comparison of datasets. Few quick notes: (i) 1000G pilot breakpoint set was derived from short 30bp reads, which does not allow to find longer micro-insertions; (ii) published 1000G phase 1 breakpoint set and the set used in this study have only ~50% overlap; (iii) majority of Pang et al. breakpoints breakpoints are for indels, with less than 8,000 variants longer than 100 bps, while we present almost 9,000 breakpoints longer than 100 bps (Fig. 1C lower panel).

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	Due to high accuracy of our breakpoint dataset we only had 3 deletions for which breakpoints sequence from assembly and from PCR were different. Such limited statistics does not allow make confident claims about what are the confounding factors for assembly. But we do agree that detailed reporting discordance during validation is a value to scientific

	commuting, so we will describe the 3 deletions in the manuscript. In fact we reported one case in Figure S5.

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 mis-understanding. Apparently, roughly half of our data set are 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 that 10 bps for two reasons: (i) short MIs could be the result of mis-interpreting SNPs or indels close to breakpoints; (ii) short MIs are likely to be mapped in multiple place in genome, and such mapping is not informative for analysis of template sites. Even about half of longer MIs could not be mapped to genome unambiguously (Table S3).
Excerpt from manuscript	*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	We can revise text/title to make clear that the analysis was conducted is for deletions only.

Referee 2.7 – Subcategorization

Reviewer comment	With such a large number of variants with nucleotide resolution breakpoints, why did the authors only investigate 3 broad classes? Although mentioned, the authors did not attempt to subcategorize variants from the NH processes (FoSTeS, MMBIR, NHEJ).
Response	In fact we were able to subcategorize some events in NH class, i.e. those for which we were able to find template sites of micro-insertion (Table S3). These are likely to arise from template switching events. Complete sub-categorization is challenging because: (i) suggested FoSTeS and MMBIR mechanisms have the same signature and can generally be classified as template switching mechanism; (ii) NHEJ and template-switching mechanism have similar signature at breakpoint; (iii) these two mechanism have also similar signature with other possible CNV formation mechanism, like retrotransposition mediated deletions (Vogt et al., Genome Biol. 2014 Jun 2;15(6):R80.). We will provide rough estimate of the proportion of each possible mechanism in NH class, along with examples for each mechanism.

Referee 2.8 – 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 limited number of samples, while the denominator was the count of all breakpoints in all individuals. We will clarify and provide requested numbers.

Referee 2.10 – Type (Minor)

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

Referee 2.11 – Figure improvement (Minor)

Reviewer comment	Figure 3B should be placed before Figure 3A
Response	Will do.

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. 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. 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 response to specific comment, which we hope, will clear some confusion and highlight novel findings. We can also conduct and report additional analysis requested by the reviewer.

Referee 3.3 – 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	We can revise text/title to make clear that the analysis was conducted is for deletions only.

Referee 3.4 – 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	We will describe the requested details. See also response to comment 2.3.

Referee 3.5 – 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 will report on this.

Referee 3.6 – 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 realize this was not clearly explained. Confirmation was done in limited number of samples, while the denominator was the count of all breakpoints in all individuals. This is the latter case described by the referee. We will clarify and provide requested numbers.

Referee 3.7 – 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/complimentary explanation. We will acknowledge it.

Referee 3.8 – 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	We suggested possible causal relationship of open/active chromatin and NAHR breakpoints based on a few lines of evidence. And methylation analysis was one of such evidence but not the only one. But we thank the referee for suggesting an alternative explanation of our methylation analysis. We will perform additional analyses to see whether CpG islands can explain decreased level of methylation around NAHR breakpoints.

Referee 3.9 – 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 also noticed typo in the manuscript text; this is Spearman (not Pearson) correlation coefficient. Agree, high correlation does imply that that nearly all NAHR site are explained by recombination rate. But this is observed only for closed chromatin. For open chromatin this is not the case, and this is one of the novel finding.

Referee 3.10 – 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. While we meant to refer to distance of template site from closest deletion breakpoint. We believe this confusion is purely because of the terminology, and we will clarify this by calling template sites as "adjacent" and "distant". Provided that the confusion is revolved, we believe the question about the difference between proximal/adjacent and distal/distant template sites is not relevant. <u>But it is significant that distribution of templates site relative to breakpoints is bimodal.</u>

Referee 3.11 – 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 the reviewer is confused. Koren et al. (ref 30, PMID:23176822) are of about kb resolution. The file (http://genepath.med.harvard.edu/mccarroll/datasets.html) with normalized replication timing contains almost 2.4 M genomic intervals, which corresponds to roughly 1.2 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 3.12 – 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 referee's deep thinking about it, as it is one of the major points in our manuscript. However, we stated that from 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 classical and hypothesized here (i.e., without cell division) NAHR. Suggested by the referee link between gene duplication families and NAHR does not project to association of NAHR with open chromatin, as it misses comparison with intergenic regions/closed chromatin. We, therefore, don't see how our statement is naïve and we made it based on few lines of evidence: direct comparison with chromatin state, association comparison with active chromatin marks, correlation with recombination rate, 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, higher density of CpG motifs, and with methylation-linked mutations. However, and unlike 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[31] and histone marks are gone[32], 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.*

Referee 3.13 – 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. We did not say and did not meant that de-methylation is the cause of deletions by NAHR. By real we meant that association is not due to technical artifacts. Later, when combining multiple associations/lines of evidence we hypothesize of causal relationship between open chromatin and NAHR breakpoints. We will clarify the text of the manuscript and report on additional analysis, which, as we understand, was suggested in comment 3.8

Referee 3.14 – 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	We will report the requested info. See also our response to comment 2.8.

Referee 3.15 – 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	The idea behind such normalization is to have asymptotic methylation level at 1. As asymptotic value corresponds to average across genome, we effectively normalized

	methylation to have average values at 1.

Referee 3.16 – Grammar (Minor)

Reviewer comment	Page 5: " But similar effect for NAHR deletions" is poor grammar. Revise.
Response	Will do.

Referee 3.17 – Figure improvement (Minor)

Reviewer comment	Legend to Fig4. what is MN an abbreviation for?
Response	Will correct.