RESPONSE LETTER

-- Ref1.1 – Statement rephrase --

Reviewer Comment	The authors present a computational approach to discover novel retroduplications from exome sequencing and whole genome sequencing data, by utilizing information of both exon-exon junctions and discordant locations of paired-end reads. Based on this approach, the authors performed comprehensive discovery and analysis of retroduplications in individuals from the 1000 Genomes Project (reflecting 26 human populations). They detect numerous parent genes having novel retroduplications absent in the reference genome, including such that differentiate between populations, and show that phylogenetic trees of human populations can be derived solely based on retroduplication variation. Novel retroduplications frequently arise from genes with relatively high expression level. Several novel intragenic insertion events are highlighted that warrant follow-up studies focusing on their potential functional impact. This study has been very thoroughly performed, and it adds an additional category of structural variation to the released Phase 3 variant categories. I regard it as an important contribution and would recommend its publication. A few comments that should help the authors in preparing their revised manuscript are below. • The authors hypothesize that the retroduplications common to several population groups were inherited from a common ancestor, while exclusive retroduplications emerged after population divergence. Here I feel the wording should be rephrased (toned down), as it is not clear whether individuals from these populations have been sampled deeply enough (e.g. to "we hypothesize that many of the exclusive retroduplications emerged after population divergence.")
Author Response	We thank the review for endorsement of publication. We agree with the reviewer about revising language, and have edited the text accordingly.
Excerpt From Revised Manuscript	Line 167 in the main text

-- Ref1.2 – Comparison between retroduplications and SNPs --

Reviewer Comment	• Along the same lines, it would be very nice to see a comparison of retroduplications with other forms of

	genetic variation. Are there proportionally more private retroduplication variants in the 1000 Genomes Project dataset than seen for other variant classes (e.g. SNPs)?
Author Response	We thank the reviewer for highlighting this point. It is a great suggestion to compare with other forms of genetic variation. In the revision, we performed comparison between retroduplications and SNPs.
Excerpt From Revised Manuscript	

Zhang Yan 3/20/2017 4:08 PM
Comment [1]: Files available. Will update the result in a day or two.

-- Ref1.3 – Example of CAV3/SLMO2 expression alteration --

Reviewer Comment	• The description of intragenic insertion events in potential disease genes (i.e. CAV3 / SLMO2) is indeed interesting. Is there evidence from Geuvadis data (or other expression/eQTL datasets that have been established in 1000 Genomes populations) that these intragenic retroduplication insertion events are associated with expression alteration of CAV3 / SLMO2 in carriers.
Author Response	We thank the reviewer for highlighting this point.
Excerpt From Revised Manuscript	

Zhang Yan 2/27/2017 11:49 PM
Comment [2]: To do
ALEX says: Can we use GTEx in some form?

Shantao 3/20/2017 4:17 PM
Comment [3R2]: Would take too long. Also I did some exploration on the GTEx public data, the isoform (the one retrodup inserts) doesn't seem to express in the GTEx cohort.
Can you quickly check Geuvadis data?

Zhang Yan 2/27/2017 3:57 PM
Comment [4]: Higher sensitivity

-- Ref2.1 – Method comparison & code release --

Reviewer Comment	This is a bioinformatically and statistically sound analysis of retrotransposition in the extended 1000 Genomes data set (n=2535). It uses an exon-exon junction strategy using exome sequencing data, controlling the false discovery rate using decoy exon junction libraries, and identifies insertion sites based on discordant paired-end reads, using shallow-sequenced whole genome sequencing data pooled by population. The subject of the analysis has been explored in the past by various publications, and the authors indicate that the novelty of the current study resides in a "large number of calls" unique to their callset due to newly enrolled populations from Phase 3 data, and "the higher sensitivity" of their methods. It would be important that the authors substantiate their claims in detail, to further support the publication of their approach. In addition, it would be central to publication in this technical journal that the authors release a package to run their pipeline in other datasets –particularly now that there will be increasing access to large numbers of

Zhang Yan 2/27/2017 11:40 PM
Comment [5]: Elaborate the comparison -- STL/ANN, we have the Venn Diagram in the suppl. – somewhat in detail. Do you have any calculation of the sensitivity while doing simulation for calculating FDR?

	whole genome sequences. There is limited use of the mutation/SNV that may characterize the retrotranposed gene. This is a question of clinical relevance: the aligners will map the duplicate reads to the parental gene and generate false image of mosaicism or heterozygous calls. Exploring this aspect would be a significant contribution to understanding the impact of the observations to routine genome annotation.
Author Response	We thank the reviewer for these suggestions. a) Higher sensitivity [[Alex: In my previous paper retrodups were discovered per population because coverage was really shallow. In my previous publication there were less than 1 retrodup per person when using population data. Here we are discovering significantly more per person, and counts are comparable (?) to those when using high coverage data. You can actually check how many of those found from high coverage analysis are found by population study and by this approach. High coverage calls for 2 individuals are given in my previous paper. Similar arguments can probably be made with respect to other previous publications.]] STL: In Alex's previous publication, his method was applied to two deep sequenced trios. In each trio, the pipeline identified 13(CEU)/11(YRI) novel retrodups, the numbers are comparable to what we found in our study (median: six per individual). We would expect the sensitivity of our exome sequencing method lie between low coverage and high coverage WGS. Indeed, in the YRI trio, using just WES, we identified seven unique retroduplications, which is slightly less than 11. Last, in the work of Abyzov et al., many these retroduplications were verified both computationally and experimentally. Furthermore, in six retroduplications we identified by exon-junction methods applied to WXS in the CEU child (NA12878), four of them were previous identified by Abyzov et al., using the WGS high coverage dataset (SKA3, TDG, CBX3 and AP3S1). In YRI trio, 5/7 overlapped with the set reported by Abyzov et al., (TMEM5, CBX3, ATP9B, MFF and AP3S1) but not

Zhang Yan 2/27/2017 11:53 PM
Comment [6]: Mark suggests running simulation. To do.
Alex: Yes, I think the reviewer suggests exploring this aspect. Since he/she does not specify how are free to anything reasonable. Could we look at our reads that are supporting calls and see whether we have positions such that there is a mismatch supported by at least two reads? Or we can say that we are not going to do this because one need to know entire sequencing of retrodup, which in our case can only be hypothesized around splice-junctions.

Shantao 3/20/2017 4:31 PM
Formatted: Font:(Default) Arial

Shantao 3/20/2017 4:32 PM
Formatted: List Paragraph, Numbered + Level: 1 + Numbering Style: a, b, c, ... + Start at: 1 + Alignment: Left + Aligned at: 0.25" + Indent at: 0.5"

Shantao 3/20/2017 4:32 PM
Formatted: List Paragraph

Shantao 3/20/2017 5:20 PM
Formatted: Highlight

Shantao 3/20/2017 4:54 PM
Comment [7]: STL2YZ: Why is NA12891 and NA12892 missing? They are the parents of the CEU trio.

Shantao 3/20/2017 5:37 PM
Comment [8]: But not ENSG00000004455 and ENSG00000092199. Yet ENSG00000004455appeared in Alex's set. (S3 table) and ENSG00000092199 has multiple pseudogenes in ref.

(PABPC1 and ENSG00000004455(AGAIN!)).

b) Code

We have wrapped up the code and will publish it with the manuscript (**link).**

c) Retroduplication erodes SNPs calling

The reviewer raised an excellent point about occult retroduplications affecting SNP calling. In the revision, we built a simple model for read generating and SNP calling (SXX Fig). We demonstrated retroduplications, when carrying alternative alleles, could have disastrous impacts on SNP calling. The intuition of our reviewer is correct. If the retroduplications carry an alternative allele and the parent gene is ref./ref., correct genotyping is almost impossible. Even when the parent gene also carries one alternative allele, the genotyping correct rate is barely around 55%-60%.

Interestingly, we found as the sequencing depth increases, calling performance deteriorates, regardless of genotypes. Namely, if the retroduplication carry reference alleles, it actually slightly boosts the genotyping when sequencing is shallow. And when the retroduplication has the alternative alleles, the higher the sequencing depth is, the worse SNP calling performs.

Of course, we must admit the SNP calling in real world in more complicated and perhaps has better performance. However, our simple model nonetheless revealed an unambiguous eroding effect of retroduplication to correct SNP genotyping. Our results raised a critical point about the role of retroduplications in WXS data processing, especially in the era of ultra-deep sequencing. Our method is able to help researchers find potential duplication events in genes and take corresponding caution and actions in SNP callings.

In the revised manuscript, we added a new section to explain our new results and reflect the discussions above.

Excerpt From	

Shantao 3/20/2017 5:21 PM
Comment [9]: We can use this for the validation request by review 3.

Shantao 3/20/2017 5:20 PM
Formatted: Highlight

Shantao 3/20/2017 4:31 PM
Formatted: List Paragraph, Numbered + Level: 1 + Numbering Style: a, b, c, ... + Start at: 1 + Alignment: Left + Aligned at: 0.25" + Indent at: 0.5"

Shantao 3/20/2017 4:31 PM
Formatted: List Paragraph

Shantao 3/20/2017 4:31 PM
Formatted: List Paragraph, Numbered + Level: 1 + Numbering Style: a, b, c, ... + Start at: 1 + Alignment: Left + Aligned at: 0.25" + Indent at: 0.5"

Shantao 3/20/2017 4:31 PM
Formatted: Font:(Default) Arial

Zhang Yan 2/27/2017 10:43 PM
Comment [10]: tar or github

Shantao 3/20/2017 4:31 PM
Deleted: We have wrapped up the code and will publish it with the manuscript (**S8 File**). ... [1]

Zhang Yan 3/19/2017 11:11 PM
Comment [11]: Detailed model added to supplement?

Shantao 3/10/2017 11:55 PM
Deleted:

Shantao 3/20/2017 4:30 PM
Comment [12]: Mark, what do you think?

Shantao 3/20/2017 4:32 PM
Deleted: ... [2]

Revised Manuscript	

-- Ref2.2.1 – Retroduplication deletions --

Reviewer Comment	The discussion on the retroduplication deletions that are absent in the individuals but present in the reference genome is unclear. Are the 68 such deletion events universally absent in Hg19, or are there unique or population specific. Are these events mapping to reported "deletions" in 1000 Genomes?
Author Response	We thank the reviewer for pointing this out. We have modified the text to clarity it. These 68 regions are present in Hg19 as processed pseudogenes (i.e. non-coding retroduplications), but reported as deletions in 1000 Genomes Phase 3. Yes, these events overlap with reported "deletions" in 1000 Genomes Phase 3.
Excerpt From Revised Manuscript	Line 146 in the main text

Zhang Yan 2/27/2017 4:01 PM
Comment [13]: Check it later

-- Ref2.2.2 – Statement rephrase --

Reviewer Comment	The comment "We did not see variable retroduplication significantly associated with its parent gene expression" is unclear and needs clarification.
Author Response	The reviewer made a good point. We have also modified the text to clarify it. Specifically, for each population, we test whether having novel retroduplication(s) is associated with the parent gene's expression level. We did not see any significant association from this analysis. i.e. No retroduplication event was identified as an eQTL.
Excerpt From Revised Manuscript	Line 191 in the main text

-- Ref2.2.3 – Text rephrase --

Reviewer Comment	In line 238: "This observation is consistent with our knowledge that in normal individuals" – please change the text to refer to "general population"
Author Response	We thank the reviewer for the suggestion. The change has been made.
Excerpt From Revised Manuscript	Line 241 in the main text

-- Ref2.2.4 – Segmental duplication --

Reviewer Comment	Fig 3C – How to interpret the enrichment in category "segmental duplication"? – As stated in the text, one of

	the criteria to differentiate DNA from RNA-mediated duplicates is by distant location from the parental gene.
Author Response	Segmental duplication came from DNA duplication mechanisms, and can be either tandem or interspersed. The general trend is retroduplicated copies tend to reside more interspersed over the genome than duplicated copies [1]. Kim et al. [citation] have found retroduplications enriched around areas of segmental duplications, because the repeats generated by retroduplication can cause NAHR which is the cause of DNA-mediated segmental duplication. This is consistent with our observation.
Excerpt From Revised Manuscript	

Zhang Yan 2/27/2017 11:01 PM
Comment [14]: Double check

-- Ref3.1 – Code release --

Zhang Yan 2/27/2017 11:02 PM
Comment [15]: Ref2.1 as well

Reviewer Comment	Zhang, Li, and colleagues have reported on the diversity of retroduplications in over 2,500 whole exomes, making this study the largest yet. The authors also show that the distribution of retroduplications is not random and find that parent genes of retroduplications tend to be highly expressed. Interestingly, the authors exhibit that retroduplications can be used as population markers, recapitulating human super-population phylogeny and confirming theories of human migration. Additionally, the authors test for enrichment of noncoding elements and show that retroduplications are typically inserted in regions of open chromatin. Finally the authors provide a suggestive genetic etiology for long QT syndrome via the SLMO2 insertion into the last intron of CAV3 in people with African ancestry. Altogether, Zhang, Li, and colleagues have presented a comprehensive analysis of retroduplications in a population representative of global variation, providing an excellent guide for subsequent analyses. However, before recommending this paper for publication, I believe the authors should address the following minor revisions: 1) In the lines of the open-source nature of PLoS publications, providing the libraries of exon-exon junction and decoys for public use might encourage others to compare their methods to the authors. Additionally if it's feasible, the authors should consider sharing their software; however, the methods are clear enough in this respect. Providing these resources might spur others to

	implicate retroduplications in human diseases, since this class of structural variantion tends to be overlooked.
Author Response	We thank the reviewer for the suggestions. We will provide the code in the supplement (**S8 File**), Readers can use it to produce exon-exon junctions and decoys, rebuild the calling pipeline and even make their own adjustment. We are glad to see the potential that users can apply the scripts to both general and disease genomes, and both human and non-human genomes. [[Shantao: we will do an URL link for code downloading]]
Excerpt From Revised Manuscript	

Zhang Yan 2/27/2017 11:03 PM
Comment [16]: tar or github

-- Ref3.2 – FDR in real data --

Reviewer Comment	2) Previously Abyzov et al. Genome Research 2013 reported a 18% FDR for retroduplication calling using a combination of methods, including PCR, in high coverage whole genomes. The authors' estimate of 2% for high coverage WES based on statistical assumptions is promising, but it would be interesting to report a FDR based on high-coverage whole genomes. 1000 Genomes provides high coverage WGS for a handful of phase 3 samples, as well as extremely deep sequenced (>70x) high quality, PCR-free WGS from the HGSV consortium, which an analysis of a single chromosome would suffice and be in accordance of data use policies. I understand that WGS has it's own inherit biases, but for interested researchers a "real world" estimate of the FDR of the authors' methods would be beneficial, since WES datasets of varying complex diseases are publicly available.
Author Response	The reviewer made a great suggestion. [[MBG mentioned NA19240 YRI trio]] [[Shantao: we will need all HGSV samples]]
Excerpt From Revised Manuscript	

-- Ref3.3 – Length of insertion --

Reviewer Comment	3) Is it possible to accurately estimate the lengths of the novel insertions? If so, could the authors provide a summary on the distribution of insertion lengths? If not, explain why it's not possible to estimate the length.
Author Response	The reviewer raised a reasonable point. However, to accurately estimate the length, one needs to get the entire sequence of the novel insertions. This is often difficult since most of the reads

Zhang Yan 3/19/2017 11:21 PM
Deleted: y

Zhang Yan 3/19/2017 11:22 PM
Deleted: because

Zhang Yan 3/19/2017 11:22 PM
Deleted: inserted

	originated from the insertions will be mapped back to the parent gene. We are only able to pull out a very small portion of reads that span the exon-exon junctions and insertion points, which are not sufficient for accurate length estimation. Furthermore, this can become more complicated when co-insertion of L1 elements happen. Last, one might think read depth increase in the parent gene regions can give hints on insertion length. However, Abyzov et al., (REF) reported higher read depth in exons (WGS data) than introns is only seen in 65% of the proposed retroduplications. Moreover, WXS does not cover introns that are used as control and read depth method does not give boundaries shape enough for precise length estimate. [[Shantao: Alexej, I have a new idea about inferring the length if the insertion is inversed; could we suck out all the reads with wrong orientations and do read depth or even assembly to infer the inserted sequence? I guess we can also show a few examples here…just we cannot do this for every retroduplication]]
Excerpt From Revised Manuscript	

Zhang Yan 3/19/2017 11:26 PM
Comment [17]: ?

-- Ref3.4 – Sequencing data coverage and read lengths --

Reviewer Comment	4) Could the authors explicitly mention the mean coverages and read lengths of the WES and low coverage WGS in the main text? I might have missed it though.
Author Response	The 1000 Genomes Project paper describes "All individuals were sequenced using both whole-genome sequencing (mean depth 7.4x) and targeted exome sequencing (mean depth 65.7x)." [2] We have also added this information in the text. The read length is 76bp for exome sequencing and 100bp for whole-genome sequencing.
Excerpt From Revised Manuscript	Line 272 in the main text

Zhang Yan 3/19/2017 11:28 PM
Comment [18]: Could you help check this, Shantao?

Shantao 3/20/2017 4:19 PM
Deleted: I think reads were 75-100, but it will be easy for Shantao to check if he has bam files.

-- Ref3.5 – More description on disease-associated insertions --

Reviewer Comment	5) The authors state that they "have detected several interesting intragenic insertion events". Could the authors briefly expand on this, highlighting the total number found and the average per genome? Such an estimate would be beneficial as a comparison for researchers interested in implicating retroduplications in human disease (e.g. affected samples have higher burden).
Author Response	We thank the reviewer for highlighting this point. We have identified a total of 12 retroduplication insertion events related to disease. We report all these disease-associated events and

Zhang Yan 2/27/2017 4:12 PM
Comment [19]: To Shantao, do you have this number?

Shantao 3/11/2017 12:16 AM
Comment [20]: This is basically the supplement table…
I will do a five-minutes quite stat. summary of it. But again, without genotyping, our sensitivity is low.

	affected populations in S3 Table. We highlighted the numbers in the text.
Excerpt From Revised Manuscript	Line 257 in the main text

-- Ref3.6 – Language improvement --

Reviewer Comment	6) The language of the main text is difficult to follow. Before this paper goes to publication, it would be more inviting for readers if the authors address this.
Author Response	We thank the reviewer's suggestion. We have thoroughly refined the language in this revision.
Excerpt From Revised Manuscript	

Zhang Yan 2/27/2017 11:10 PM
Comment [21]: WORKING ON THIS

References:

1. Sisu C, Pei B, Leng J, Frankish A, Zhang Y, Balasubramanian S, et al. Comparative analysis of pseudogenes across three phyla. Proc Natl Acad Sci U S A. 2014;111: 13361–6. doi:10.1073/pnas.1407293111
2. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Nature. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2015;526: 68–74. doi:10.1038/nature15393