Yale University

Bass Building, Rm 432A
266 Whitney Avenue
PO Box 208114
New Haven, CT 06520-8114

203 432 6105
360 838 7861 (fax)
mark@gersteinlab.org

21st November 2015

Deleted: <*object*>

Nature Communications
75, Varick Street
Fl 9, New York
NY, 10013-1917
USA

Dear Dr. Cho,

Thank you for the invitation to revise and resubmit the manuscript. We have worked very hard to make sure we address *all* the concerns of the three reviewers, to the extent of reprocessing *all* the datasets and downstream analyses for each round of submission. We are therefore heartened that Reviewers #1 and #3 find our responses satisfactory and have endorsed our manuscript for publication in *Nature Communications*. However, we are rather surprised by Reviewer #2's comments.

Deleted: We are heartened that reviewers

Deleted: reviewer

Reviewer #2 had cited two major concerns in both rounds of reviews: (a) accounting of over-dispersion in the ChIP-seq and RNA-seq datasets and (b) differences in mappability between the alleles.

Deleted: The

Specfically, for (a), he mentioned that to account for over-dispersion "the correct analysis must use *some* strategy to estimate the over-dispersion parameter and take it into account when testing for ASE". Based on just this very general description, we responded by first explaining that there is actually a wide range of methods. We then went to great lengths to craft and implement a novel two-step procedure to account for over-dispersion in the context of our approach, where we estimate over-dispersion twice, on a per-dataset and per-individual basis. The reviewer responded by saying that the previous methods "got it wrong" and that our approach removes only "some loci because of too much over-dispersion" and is "not effective". He also mentioned "there are at least 3 published methods to identify ASE using models that estimate site-specific over-dispersion" and we should use one of them.

First, his/her interpretation of our approach is not correct. We do not remove loci because of too much over-dispersion, instead we remove *entire datasets* because they are highly over-dispersed and will lead to the detection of more false positives if included in our database. While we showed with actual results in Figure 2 that individual over-dispersed datasets can lead to a higher number of detected 'positives', he/she made a very general statement that our approach is ineffective, without pointing to any specific study, tool or method. We have provided in our current response *5* other tools (some very recent) that use, advocate or include the calculation of global and individual-specific over-dispersion in their allele-specific variant detection.

In response to his comment that the previous methods were "mistakes" and that they "got it wrong", we would like to emphasize that the publications that we cited in our responses are a selection of the most current work performed by authorities in the field and peer-reviewed by colleagues in the community. The key point that we are trying make is not to show the 'correctness' of these methods, but to point to the broader reality that there is currently a diversity of methods in the community. For example, while the GTEx consortium [1] did attempt to correct for allelic mapping bias, they performed their alignment on the human reference genome and allele-specific detection using binomial tests, not accounting for over-dispersion. On the other hand, Ding *et al.* [2] performed their alignment on the human reference genome and allele-specific detection using binomial tests, but did *not* correct for allelic mapping bias explicitly. While we were revising our manuscript, we have also become aware of two more publications, which adopted different approaches to allele-specific variant detection. Castel *et al.* from *Genome Biology* [3] describes a new tool in the GATK software package and discussed the best practices for allele-specific analyses that do *not* take over-dispersion into account. Van de Geijn *et al.* from *Nature Methods* [4] introduced a new allele-specific detection tool that takes into account over-dispersion on a per-individual basis (similar to our pipeline; not site-specific as suggested by Reviewer #2). Given the plurality of current approaches, the fact that the reviewer is insisting on his/her points of view suggests his/her prejudice for a particular 'right' approach, when there is simply no firm consensus.

Deleted: main

Deleted: at present

For (b), in the first round of reviews, he mentioned that "the personal genome indeed eliminates the reference bias but does not eliminate the error associated with differences in mappability between the two alleles" and "the only solution to date has been to map each allele separately and only retain reads that map uniquely at each allele". He/She is implying that the reference bias is mutually exclusive of the allelic mapping bias and is suggesting that there is a one 'true' protocol that everyone in the field follows. Firstly, we want to clarify that the allelic differences in mapping, or 'allelic mapping bias', is a generic description to depict differential probability in the alignment of reads to the different alleles along a heterozygous locus. This, *includes* the reference bias, which mainly occurs because most allele-specific studies in the field use the human reference genome for alignment, hence the allele-specific SNV detection has been shown to favor the reference allele. There has been at least three other publications from peer-reviewed journals such as *Science*, *Nature* and *PLoS Genetics* that regarded the reference bias as the **major source** of allelic mapping bias. More importantly, various studies may have a different take on how to account for the bias, with many agreeing that using the personal genome is one of the most rigorous but also computationally intensive ways to manage the reference bias [3, 5, 6]. Therefore, there is no "*only*" solution to this problem, as suggested by the reviewer. Nonetheless, in this round of revision, we have, again, gone the extra mile to placate the reviewer by accounting for another potential but less significant source of allelic mapping bias, we termed it 'ambiguous read mapping', and reprocessing *all* the datasets for the second time.

In our endeavor to mine the wealth of existing datasets, we have come to appreciate and acknowledge this diversity, and thus have advocated for the need to uniformly process the datasets. Our allele-specific detection approach is technically reasonable. Our use of the personal genomes has already been cited by many previous publications in the field as a more rigorous way of alleviating allelic mapping bias [3, 5, 6]. Furthermore, our current approach has already been extensively discussed and ultimately utilized in the ENCODE, Epigenomics Roadmap and 1000 Genomes Project consortia. The ENCODE consortium has utilized an earlier version of our approach in its 2012 publication [7]. It is currently being used by the Epigenomics Roadmap consortium in their allele-specific analyses. It has also been implemented in the recent peer-reviewed *Nature* publication by the 1000 Genomes Project Structural Variants group [faa]. In particular, the personal genome construction was shown to be especially useful in structural variant analyses since it is able to incorporate indels and structural variants; the other allele-specific methods are only limited to single nucleotide variants.

Deleted: The

Deleted: in an earlier version of our approach

Deleted: .

Currently, there is a plethora of approaches developed to address various concerns. We have made significant efforts to improve our manuscript and incorporate all the reviewers' comments, to the extent of spending months (over a year now) reprocessing all the datasets in each revision, while preserving the main themes of our manuscript. However, we fear Reviewer #2's insistence on his/her single approach in performing allele-specific detection when there are multiple ways. Nonetheless, we are deeply encouraged by the other two reviewers' firm endorsements of our current manuscript and indeed strongly believe that our approach and resource will generate considerable interest in the community. Hence, we do hope to seek your understanding and please do consider this cover letter when making your decision.

Deleted: We have worked very hard to address *all* the concerns from *all* three reviewers. In fact, we have gone to the extent of reprocessing *all* the datasets and downstreams analyses for each round of submission. In addition, we implemented novel ways to uniformly process datasets in a high throughput manner. Specifically, we devised a novel serial two-step procedure to account for over-dispersion on both a per-dataset and per-individual basis. In this round of submission, we have also taken into account allelic mapping bias in the context of a diploid personal genome. ¶
¶
We agree that allele-specific analyses are challenging. Hence

Deleted: , with corresponding pros and cons,

Deleted: Reviewer #2 has given some reasonable suggestions, thus we have

Deleted: in trying to

Deleted: his/her, and

Formatted: Font: Not Italic

Deleted: other

Deleted: an

Deleted: take into consideration

Yours sincerely,

Mark Gerstein
Co-chair of 1000 Genomes Project Consortium Functional
Interpretation Group and Member of the 1000 Genomes
Project Consortium Structural Variation Group
Albert L. Williams Professor of Biomedical Informatics,
Molecular Biophysics & Biochemistry,
and Computer Science,
Co-director of the Yale Program in Computational Biology and Bioinformatics

[1] The GTEx Consortium (2015). *Science*, 348(6235):648-60, PMID: 25954001
[2] Ding *et al.* (2014). *PLoS Genet.*, 10(11):e1004798, PMID: 25411781
[3] Castel *et al.* (2015). *Genome Biol.*, 16(1):195, PMID: 26381377
[4] van de Geijn *et al.* (2015). *Nat Methods*, doi: 10.1038/nmeth.3582 [epub ahead of print], PMID: 26366987
[5] Panousis *et al.* (2014). *Genome Biol.*, 15(9):467, PMID: 25239376
[6] Stevenson *et al.* (2013). *BMC Genomics*, 14:536, PMID: 23919664
[7] Djebali *et al.* (2012). *Nature*, 489(7414):101-8, PMID: 22955620

Formatted: Font: Not Bold