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

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 are 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.

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 main 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 at present 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.

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. The use of the personal genomes in an earlier version of our approach has been cited by many previous publications in

Deleted: <*object*>

Deleted: 27th August

Formatted: Font color: Red

Formatted: Font color: Red

Deleted: .

Formatted: Font: Italic

Deleted: where

Deleted: these datasets. Our allele-specific detection

the field as a more rigorous way of alleviating allelic mapping [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. 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.

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, there is a plethora of approaches, with corresponding pros and cons, developed to address various concerns. Reviewer #2 has given some reasonable suggestions, thus we have made significant efforts in trying to incorporate his/her, and *all* the other reviewers' comments, while preserving the main themes of our manuscript. However, we fear an 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 do take into consideration this cover letter when making your decision.

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

Deleted: 3
Deleted: also
Deleted: utilized
Formatted: Font color: Red
Formatted: Font color: Red
Deleted: Moreover, our approach is used
Deleted: analyses of
Deleted: Specifically
Deleted: is
Deleted: The manuscript from the Structural Variant Group of the 1000 Genomes Project consortium has just been recently peer-reviewed and accepted by *Nature*.
Deleted: Furthermore, building a personal genome not only reduces
Formatted: Font: Italic
Deleted: reference bias as mentioned by reviewer #2, but ...
Deleted: show in our new
Deleted: and responses, it is
Deleted: less affected by the type of
Deleted: that was highlighted
Deleted: Degner *et al* [4] and van de Geijn *et al* [5].
Deleted: Therefore
Deleted: made
Deleted: . While we have also tried to add new analyses ...
Deleted: review to address specifically reviewer #2's ...
Deleted: ,
Deleted: ¶ ...
Formatted: Underline
Deleted: ¶ ...
Deleted: .
Deleted: , Z.
Deleted: .
Moved (insertion) [1]
Deleted: .
Deleted: ¶ ...
Moved up [1]: (2015).
Deleted: *bioRxiv*. doi: http://dx.doi.org/10.1101/011221¶ ...
Formatted: Font: Times New Roman, Bold
Formatted: Normal