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 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.

Reviewer #2 had cited two major concerns in both rounds of reviews: (a) accounting for differences in mapping between alleles and (b) overdispersion in the datasets.

For (a), as explained in our current response to the reviewer, the allelic differences in mapping, or 'allelic mapping bias', *includes* the reference bias, which we have already accounted for by the construction and use of the diploid personal genomes. Moreover, at least three other publications from peer-reviewed journals such as *Nature*, *Science* and *PLoS Genetics* regarded the reference bias as the **major source** of allelic mapping bias [1, 2, 7]. More importantly, various studies have a different take on how to account for the bias (please see Supplementary Table 1 of our revised manuscript), with many agreeing that using the personal genome is one of the most rigorous ways [3, 5, 6]. Therefore, there is no "*only*" solution to this problem, as suggested by the reviewer. In fact, there is no single solution to totally eliminate this issue [3]. Nonetheless, in this round of revision, we have accounted for the next major bias, which we termed 'ambiguous mapping bias' [4]. We also show that it has a smaller effect and does not change the main results of our previous submission. For the second time, we have taken another 3 months to reprocess all the datasets and analyses. We hope that we have satisfied the reviewer in this aspect by accounting for not one, but two, main types of allelic mapping bias.

For (b), in his previous comments, he mentioned that "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

Deleted: <*object*>

Deleted: of over-dispersion in the ChIP-seq and RNA-seq datasets and (b)

Deleted: mappability

Deleted: the

Moved (insertion) [1]

Deleted: Specfically, for (a), he mentioned that to account for over-dispersion

actually a wide range of methods. We then went to great lengths to implement a novel two-step procedure to account for overdispersion in the context of our approach, taking approximately 3 months to reprocess all data and analyses.

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 did not account for overdispersion. Ding *et al.* [2] neither explicitly correct for allelic mapping bias nor account for overdispersion. While we were revising our manuscript, we have also become aware of two more publications. 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 overdispersion into account. Van de Geijn *et al.* from *Nature Methods* [4] introduced a new allele-specific detection tool that takes into account overdispersion 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 again insisting on his/her points of view suggests his/her prejudice for a particular 'right' approach, when there is simply no firm consensus.

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 (SV) group [8]. That is why we initially submitted this manuscript as a companion to the 1000 Genomes paper, as the methods were extensively used by the consortium, particularly in the SV and Functional Interpretation groups.

We have made significant efforts to improve our manuscript and incorporate all the reviewers' comments, to the extent of spending months 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 consideration of this cover letter when making your decision.

Yours sincerely,

Mark Gerstein
Albert L. Williams Professor of Biomedical Informatics,
Co-director of the Yale Program in Computational Biology and Bioinformatics

Deleted: craft and

Deleted: over-dispersion

Deleted: 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

Deleted: published methods

Deleted: identify ASE using models that estimate site-specific over-dispersion" and we should use one of them.

Deleted: ¶
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.¶

Deleted: performed their alignment on the human reference genome and allele-specific detection using binomial tests, not accounting

Deleted: over-dispersion. On the other hand,

Deleted: performed their alignment on the human reference genome and allele-specific detection using binomial tests, but did *not*

Deleted: explicitly.

Deleted: , which adopted different approaches to allele- ...

Deleted: over-dispersion

Deleted: over-dispersion

Deleted: For (b), in the first round of reviews, he mentior ...

Moved up [1]: Therefore, there is no "*only*" solution to

Deleted: Nonetheless, in this round of revision, we have, ...

Deleted: group [faa]. In particular, the personal genome ...

Deleted: Currently, there is a plethora of approaches ...

Deleted: (over a year now)

Deleted: please do consider

Moved down [2]: Co-chair of 1000 Genomes

Deleted: and Member of the 1000 Genomes ...

Deleted: Molecular Biophysics & Biochemist ...

Formatted: Indent: Left: 0.5", First line: 0.5"

Co-chair of 1000 Genomes Project Consortium Functional Interpretation Group

Moved (insertion) [2]

[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
[8] Sudmant *et al.* (2015). *Nature*, 526(7571):75-81. PMID: 26432246