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

30th 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 our manuscript. In this and the previous re-submission, we have expended significant efforts to address *all* the concerns of the three reviewers, to the extent of modifying our algorithm and reprocessing and re-analyzing *hundreds* of datasets.

We are heartened that Reviewers #1 and #3 endorsed our manuscript for publication in *Nature Communications*. However, we are rather surprised by Reviewer #2's continued criticisms. Fundamentally, we feel that the remaining criticisms represents relatively minor sources of bias in these data. Nonetheless, we have tried to do all potential calculations to address the criticism -- involving many month-long computes. We demonstrate in our response and manuscript that the effects he/she claims are largely inconsequential to the results we report in our piece. Moreover, Reviewer #2's comments suggest that there is a universally accepted standard for reporting allelic effects, which is simply not the case and we hope to make this clear in this letter and in the response.

Deleted: or

Now in detail, Reviewer #2 had cited two major concerns in both rounds of reviews: (a) mapping and (b) overdispersion in the datasets.

For (a), the reviewer suggests that we have missed an important source of allelic mappability bias. However, our use of the personal diploid genome is explicitly designed to compensate for the majority of this bias, which is termed the 'reference bias'. The personal genome approach accounts for additional issues beyond the reference bias, leading to better read alignment due to its ability to incorporate variants beyond just SNVs, such as indels and structural variation (as shown by Sudmant et al.) [5]. Various studies have a different take on how to account for these mappability biases (Supplementary Table 1), with many agreeing that the alignment to a personal genome, as we have done here, is a conservative and effective method for mitigating a large amount of potentially confounding bias [1, 2, 3].

Deleted: as also explained in our current response to

Deleted: , the

Deleted: differences in mapping, or 'allelic mapping bias', *includes*

Formatted: Font: Not Italic

Deleted: which we have already accounted for

Moved (insertion) [1]

Formatted: Font color: Auto

Deleted: the use of diploid personal genomes.

Deleted: the bias

Formatted: Font color: Auto

Deleted:

Nonetheless, in this round of revision, we have strived to quantify and compensate for the remaining bias highlighted by Reviewer #2, which we termed the 'ambiguous mapping bias'

[1,4]. We show that this remaining bias has a much smaller effect than the reference bias when using the personal genome approach and does not change the main results of our previous submission. In response to Reviewer #2’s comment we have added this filter in order to correct for this small bias and make our approach fully compliant with what he/she sees as the ‘standard’ for the field. While small, this addition required substantial engineering effort, involving many month-long computes to reprocess all 1,263 datasets.

Deleted: the ambiguous mapping

Deleted: Thus, we interpreted

Deleted: as asking us to add in a small bias correction

Deleted: or

Deleted: standard

Deleted: actually

Deleted: the

Deleted: in a uniform fashion. Moreover, our approach actually exceeds this level of correction since it accounts for additional issues, such as reference bias correction, better read alignment and the ability to incorporate variants beyond just SNVs, e.g. indels (as shown by Sudmant *et al.*)

Moved up [1]: [5].

Deleted: re-

Formatted: Font color: Auto

Deleted: his

For (b), in Reviewer #2’s previous comments, he/she mentioned that “the correct analysis must use *some* strategy to estimate the overdispersion parameter and take it into account when testing for ASE”. Based on just this very general description, we first responded by explaining that there is indeed a range of perspectives and methods to account for the issue of overdispersion [4,6-9] (please also refer to Supplementary Table 1). We then went to great lengths to implement a novel two-step procedure to account for overdispersion in the context of our approach. In response, he/she commented that the previous methods were “mistakes” and that they “got it wrong”. We would like to point out that these methods are some of the *most current* work by *authorities* and *peer-reviewed* by colleagues in the field. More importantly, the key point that we are trying to 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, most recently, Castel *et al.* from *Genome Biology* [1] 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). We have also cited at least five other methods in the response that advocates for a global estimation.

Deleted: While the reviewer stated there are at least three site-specific methods, we

Deleted: three others

Deleted: does instead

Deleted: of overdispersion

Given the plurality of current approaches, the fact that the reviewer has been insisting on his/her points of view suggests his/her prejudice for a particular ‘right’ approach, when there is simply no firm consensus. Furthermore, our current approach has already been extensively discussed and utilized in the ENCODE [10], and the Epigenomics Roadmap consortia. It has also been implemented in the recent *Nature* publication by the 1000 Genomes Project Structural Variants (SV) group [5], which was the reason 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 considerable efforts to modify our manuscript to take into account Reviewer #2’s criticisms while preserving the main themes of our manuscript. We are encouraged by the other two reviewers’ endorsements of our current manuscript and indeed 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.

Deleted: deeply

Deleted: strongly

Yours sincerely,

Mark Gerstein
Albert L. Williams Professor of Biomedical Informatics,
Co-director of the Yale Program in Computational Biology and Bioinformatics
Co-chair of 1000 Genomes Project Consortium Functional Interpretation Group

[1] Castel *et al.* (2015). *Genome Biol.*, 16(1):195, PMID: 26381377
[2] Panousis *et al.* (2014). *Genome Biol.*, 15(9):467, PMID: 25239376
[3] Stevenson *et al.* (2013). *BMC Genomics*, 14:536, PMID: 23919664
[4] van de Geijn *et al.* (2015). *Nat Methods*, doi: 10.1038/nmeth.3582 [epub ahead of print], PMID: 26366987
[5] Sudmant *et al.* (2015). *Nature*, 526(7571):75-81. PMID: 26432246
[6] Sun (2012). *Biometrics*. 68(1):1-11
[7] Mayba *et al.* (2014). *Genome Biology*. 15(8):405
[8] Crowley *et al.* (2015). *Nature Genetics*. 47(4):353-60
[9] Harvey *et al.* (2015). *Bioinformatics*. 31(8):1235-42
[10] Djebali *et al.* (2012). *Nature*, 489(7414):101-8, PMID: 22955620