RESPONSE TO REVIEWERS FOR "ALLELE-SPECIFIC BINDING AND EXPRESSION: A UNIFORM SURVEY OVER THE 1000-GENOMES-PROJECT INDIVIDUALS"

RESPONSE LETTER

Reviewer #1

-- Ref1 – Endorsement for publication --

Reviewer Comment	This reviewer did not have formal comments to the authors as s/he found the revised paper to be satisfactory and endorses publication.
Author Response	We thank the reviewer for his/her thorough examination of our manuscript and endorsing our paper for publication.

Reviewer #2

-- Ref2.1 – General comment --

Reviewer Comment	The authors did not adequately address my two major concerns.
Author Response	We thank the reviewer for the thorough examination of our manuscript. We have provided additional analyses and responses.

-- Ref2.2 – mapping to the personal diploid genome --

Reviewer Comment	My first comment was that mapping bias should be addressed. The authors replied by explaining that they excluded reads that map to more than one location. This is indeed a standard step in more alignment. Yet, the challenge when looking for ASE is not standard. Different alleles may have different mapping probabilities and this must be taken into account. Failing to do so results in a high number of falsely identified ASE. I must admit that it is a bit concerning to me that the authors interpreted my comment as a question regarding their standard alignment approach. In my mind, it points to a deep lack of familiarity with the ASE literature.
Author Response	We would like to thank the reviewer for pointing out the importance of allelic mapping bias, which actually also includes the reference bias. In fact, reference bias has been widely regarded as the main source of allelic mapping bias, since the more standard alignment procedure is actually the alignment of reads to the human reference genome, not to the personal genomes [1-5]. Many publications have specifically cited the use of the personal

Deleted: ,3,4,

genomes as a rigorous but computationally intensive procedure to correct for reference bias [1,3-5]. Thus, we are acutely aware of this primary issue in mapping bias, and have chosen to focus specifically on rectifying the reference bias by aligning reads to their corresponding diploid personal genome.

There is currently no single solution to totally eliminate allelic mapping bias [1]. Hence, while a small proportion of the mapping bias will still exist, we do expect the majority of the allelic bias to be accounted for, or at least alleviated, in the form of the reference bias by the use of the personal genomes.

Nonetheless, in this revision, we have further examined another bias, which we termed the 'ambiguous mapping bias'. In the context of the personal genome, this can occur due to sequence homology in other regions (new Figure 1 in the manuscript), as described also by previous studies [1,5,6]. To date, the primary strategy to manage this bias has been via simulation of uniquely mapped reads and the identification and removal of sites in which >5% of the total number of reads exhibit such ambiguous mapping bias [1,5,7-9]. However, we found that site removal can be overly stringent, since many of these implicated sites are still detected as allele-specific under the beta-binomial test if we remove only the reads with ambiguous mapping bias (new Supplementary Table 5 in the manuscript). Hence, we adopted the ambiguous-read-removal strategy. Even though it is computationally more expensive, it provides the double advantage of being able to remove potential false positives and yet still able to retain those that are strongly allele-specific. Interestingly, while we were working on this submission, van de Geijn *et al.* published in *Nature Methods* a tool that also similarly removes reads, instead of sites, in order to account for allelic mapping bias [6].

Note also that we integrated the ambiguous-read-removal approach with our personal genome construction. This allows us to take into account serious biases that none of the other allelic methods accounted for before, such as the occurrence of SVs and indels, and quality control via the removal of SNVs within CNV regions. As discussed in the 1000 Genomes SV paper (of which this is formally a companion), not taking into account larger genomic variants can cause great differences in RNA-seq read alignments and allelic imbalance [10].

Even upon integrating the ambiguous-read-removal strategy into our pipeline and re-implementation, we found very slight changes to our main results. While the effects are small, we had to

Deleted: ,4,

Deleted: within this small proportion of allelic mapping bias

Deleted: ,8,

Deleted: (since we need to filter and re-process the original read pile),

Deleted: So far, we have reprocessed all the datasets and analyses carefully twice, with each round taking approximately 3 months.

	reprocess all 1,280 datasets and re-perform all downstream analyses carefully in order to consistently and properly incorporate this into AlleleDB, expending several months of re-computes. We hope we have satisfied the reviewer by carefully implementing and accounting for not one, but two, types of allelic mapping bias, in the context of diploid personal genomes. Finally, we have modified the manuscript by including results of the additional analyses in the supplementary materials, 'Discussion' section and details of the new AlleleDB pipeline in the 'Results' and 'Methods' sections. [1] Castel *et al.* (2015). *Genome Biol*., 16(1):195 [2] Degner *et al.* (2009) *Bioinformatics.* 25(24) [3] Satya *et al.* (2012) *Nucleic Acids Res*. 40(16):e127 [4] Stevenson *et al.* (2013) *BMC Genomics*. 14:536 [5] Panousis *et al.* (2014). *Genome Biol.*, 15(9):467 [6] van de Geijn *et al.* (2015). *Nat Methods*, doi: 10.1038/nmeth.3582 [epub ahead of print] [7] Kilpinen *et al.* (2013). *Science*, 342(6159):744-7 [8] Lappalainen *et al.* (2013). *Nature,* 501(7468):506-11 [9] The GTEx Consortium (2015). *Science*, 348(6235):648-60 [10] Sudmant *et al.* (2015). *Nature*, 526(7571):75:81
Excerpt From Revised Manuscript	Please refer to Supplementary Tables 1, 3 and 5 and their corresponding legends. Please also refer to the 'Results' section under 'AlleleDB Workflow'. *"…The third module filters reads that exhibit a bias we term 'ambiguous mapping bias' (AMB). This bias occurs at a locus when reads containing one allele are preferred, not because of better alignments, but because of sequence homology in the region overlapping the other allele, with another location in the genome. As a result, reads with the other allele align ambiguously to multiple locations and are consequently removed, resulting in an erroneous allelic imbalance at that locus (Figure 1). This module detects reads that exhibit AMB via simulations. Briefly, for each original uniquely mapped read that overlap at least one heterozygous SNV on one parental genome (we call 'O read'), we simulate reads that represent all possible haplotypes of that 'O' read (we call 'S' reads). We then align the 'S' reads to the other parental genome. 'O' reads with 'S' reads that map to multiple locations are filtered from the aligned reads obtained in Step 2b (see Figure 1 and 'Methods')."* We also devote an entire section in 'Methods' under the heading 'Accounting for ambiguous mapping bias (AMB)'. *"We term this 'ambiguous mapping bias' (AMB), because reads from one allele might align ambiguously to multiple locations, resulting in reads with the other allele being unduly favored (Figure 1).[19,38,35] Several strategies have been implemented in dealing with the ambiguous mapping bias (Supplementary Table 1)… Finally, we identify and filter the 'O' reads which give rise to 'S' reads that align to multiple loci in the other parental genome or if they do not map back to the same location – 'O' reads with AMB. We also exclude and 'O' reads in which neither of the alleles of the overlapping*

Deleted: main

Deleted: the

Deleted: genome. Additionally, our approach is already conservative, with multiple additional filters in place, such as quality control via the removal of highly over-dispersed datasets and using the beta-binomial test with an FDR of 5% for all datasets

Deleted: improved

Deleted: for ambiguous mapping bias

Deleted: a discussion in the

Deleted: Dixon *et al.* (2015). *Science*, 518(7539):331-6¶ [11] Rozowsky *et al.* (2011). *Mol Syst Biol*., 7:522¶ [12]

Formatted: Font color: Red

Deleted: and 'Methods' section under 'Accounting for ambiguous mapping bias'

Deleted: ¶
*"…**(3)***

Deleted: *preferentially map to one allele over the other due to sequence homology (Figure 1), which*

Formatted: Font color: Auto

Deleted: .

Deleted: *occur*

Deleted: *maps to multiple locations and are thus removed,*

Deleted:

Deleted: *worse alignment*

Deleted: *ambiguous alignment. For a*

Deleted: *('original*

Deleted: *read, even though we found that most original reads overlap only 1 heterozygous SNV (typically >90%; Supplementary Table 3).*

Deleted: *simulated*

Deleted: *Original*

Deleted: *simulated*

Deleted: *or do not map back to the same location on the other parental genome are removed. (Figure 1). We subsequently re-align the*

Deleted: *read pile to the diploid personal genome*

Deleted: Please refer to the 'Discussion'

Deleted: for more description.¶
¶
"The second allelic

Formatted: Font: 12 pt, Not Italic

Deleted: *stems from loci with sequence homology.*

Deleted: ,

Deleted: *). … We also show that ambiguous mapping bias*

...

	SNVs matches the nucleotide on the corresponding read, as they suggest sequencing errors." There is also a paragraph in the 'Discussion' section. *"The second allelic mapping bias stems from loci with sequence homology, or 'ambiguous mapping bias' (AMB). Our implementation of a read-removal strategy has the dual advantage of removing false positives and yet retaining robust allele-specific SNVs, as compared to the more stringent site-removal strategy. Interestingly, this strategy has also been employed very recently by van de Geijn et al.[38] Besides allelic differences, ambiguous mapping is also highly dependent on the length of the read, as shown by Degner et al., with the bias decreases with increasing read length.[19] We envision that AMB will be further alleviated by long read technologies being employed in functional assays."*

Deleted: *also*

Deleted: *. that*

Deleted: *ambiguous mapping bias*

-- Ref2.3 – Over-dispersion –

Reviewer Comment	My second major concern was regarding the binomial test to identify ASE. The authors begin their response by citing other papers that used such a test. I am not sure what it the argument presented here, especially since the authors proceed by acknowledging over-dispersion in their data. So, yes, other paper got it wrong in the past, but this is hardly a reason to perpetuate this mistake. As for their revised approach, estimating a global over-dispersion parameter is not effective. Removing some loci because of 'too much' over-dispersion is ad hoc and was not justified. But more importantly, there are at least 3 published methods now to identify ASE using models that estimate site-specific over-dispersion, account for mapping bias, and report p values based on permutation. Why not use one of those published methods?
Author Response	While we thank the reviewer for his/her comment, we want to clarify that the purpose of the references is not to make any claims on the 'correctness' of the methods, but to point to the broader reality that there is currently a diversity of methods in the field, where there is no firm consensus on the 'right' approach. The fact that these publications are recent and peer-reviewed at influential journals indicates the plurality of the methods accepted by the community, each with their own advantages and limitations. For example, van de Geijn *et al.* [1] is a very recent publication in *Nature Methods* that presented a software, which performs alignment to the human reference genome, accounts for mapping bias and uses the beta-binomial test to account for an individual-specific (not site-specific) global over-dispersion. However, it is not able to take into account indels and larger structural variants, which can be accommodated by the construction of personal genomes. Moreover, the estimation of a global over-dispersion has also been employed extensively in many recent and peer-reviewed software that detect allele-specific expression [1-5].

	Additionally, our revised approach estimates over-dispersion at two levels. An over-dispersion parameter is estimated for each dataset to remove *entire datasets* (not loci) that are deemed too over-dispersed and that might result in higher number of false positives. After which, for each sample (for RNA-seq and each sample and transcription factor, TF, for ChIP-seq experiments), we pool the datasets and estimate the individual-specific global over-dispersion (for each sample for RNA-seq and also each sample and transcription factor for ChIP-seq) and apply this estimation to the beta-binomial test for each site in that individual (or TF). Hence, in this manner, the estimation of the over-dispersion can accommodate user-defined site-specific estimation of over-dispersion if necessary. Our R code is provided on our website for modifications and more customized analyses by the user. We further point out that our two-step serial procedure is novel. By removing datasets that are too over-dispersed at the outset, this first step serves as a quality control to homogenize the pooling of datasets before the second overdispersion calculation. This fits very well into our pipeline as it facilitates the harmonization and uniform processing of large amounts of data and alleviates an ascertainment bias in which more positives might stem from these highly over-dispersed datasets if they are not removed. Hence, we have retained our estimation and use of a global over-dispersion for detecting allele-specific variants. [1] van de Geijn *et al.* (2015). *Nat Methods*, doi: 10.1038/nmeth.3582 [epub ahead of print] [2] Sun (2012). *Biometrics*. 68(1):1-11 [3] Mayba *et al.* (2014). *Genome Biology.* 15(8):405 [4] Crowley *et al.* (2015). *Nature Genetics.* 47(4):353-60 [5] Harvey *et al.* (2015). *Bioinformatics*. 31(8):1235-42
Excerpt From Revised Manuscript	

Reviewer #3

-- Ref3.1 – Endorsement for publication --

Reviewer Comment	The manuscript is much improved and the authors have sufficiently addressed the majority of my concerns. I have the following minor comments:
Author Response	We thank the reviewer for the thorough examination of the manuscript and we are pleased that the reviewer finds our improved manuscript satisfactory.

-- Ref3.2 – Include additional references --

Reviewer Comment	1) Imprinting discussion should reference recent imprinting paper from GTEx. Lappalainen in Genome Research. 2) Heritability analyses of ASE should reference Li, AJHG, 2014.
Author Response	We have included the references in the respective sections of the manuscript.
Excerpt From Revised Manuscript	Please refer to the 'Discussion' section and also the 'Results' section under "ASB and ASE Inheritance analyses using CEU trio". "It could also be a result of other epigenetic effects such as genomic imprinting where no variants are causal.[41]", where reference 41 is by the GTEx consortium and Baran *et al.* published in *Genome Research.* "The CEU trio is a well-studied family and with multiple ChIP-seq studies performed on different TFs. Previous studies have also presented allele-specific inheritance.[10,15,21]", where reference 21 is by Li *et al.* published in *American Journal of Human Genetics*.