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 point out that the reference bias is not a separate issue from the allelic mapping bias, which is the generic term to describe differential mapping probabilities of the alleles; the allelic mapping bias *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

Deleted: General positive comment

Deleted: agree with the reviewer

Deleted: is still an issue, mostly because allelic bias cannot be totally eradicated with current methods [1]. The two main types

Deleted: that are most widely discussed in the field are

Formatted: Font: Italic

Deleted: and mapping bias arising from sequence homology with other genomic locations [2]. ¶
¶
Reference

Deleted: , in fact,

personal genomes [1,2,3,4,6]. Many publications have specifically cited the use of the personal genomes as a rigorous but computationally intensive procedure to correct for reference bias [1,3,4,5,6]. A recent study by Panousis *et al.* found that the bias towards the reference allele contributes to the main bulk of the overall mapping bias in allele-specific expression [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 to reads from each individual's assay to their corresponding diploid personal genome.

While a small proportion of the mapping bias do still exist, we 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. This small proportion of allelic mapping bias can occur due to situations where short reads that carry one allele may map perfectly to a reference genome but reads with the other allele map to multiple loci (due to sequence homology in other regions) (Figure 1) as described also by previous studies [1,5,6]. We termed this 'ambiguous mapping bias'. To date, the primary way to manage this bias has been the identification and removal of sites in which >5% of the total number of reads exhibit such ambiguous mapping bias [1,5,7,8,9,10]. There is currently no single 'solution' to perfectly eliminate all allelic mapping bias [1].

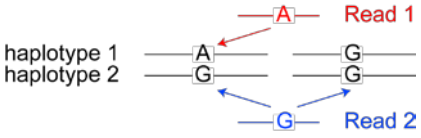

Figure 1. Adapted from van de Geijn et al. showing allelic mapping bias in a personal genome due to sequence homology in other locations. *Here, Read 1 uniquely maps to the haplotype 1, but Read 2 with the alternate allele maps to multiple locations in the other haplotype, and is therefore removed.*

Here, we investigated the effect of the ambiguous mapping bias on the detection of allele-specific SNVs, in the context of the diploid personal genome.

We chose two representative RNA-seq and two ChIP-seq datasets (from NA12878) for our ambiguous mapping bias analyses with personal genome alignments. We found that only a small proportion of SNVs (2-4%) associated with allele-specific expression (ASE) had an allelic bias >5%. On the other hand, there is a higher proportion of SNVs associated with allele-specific binding (ASB) that exhibit >5% allelic mapping bias (19-21

Deleted: Many publications have specifically cited the use of personal genomes as a rigorous but computationally intensive procedure to correct for reference bias [1,3,4,5,6].

Deleted: a diploid personal genome. Nonetheless, we undertook this endeavor, to not only construct diploid personal genomes for all 382 individuals, but also created tools for the personal genome construction.

Deleted: While

Deleted: , we agree with the reviewer that a

Deleted: the

Deleted: still exists. This is especially the case in

Deleted: (multi)

Deleted: Most studies have examined

Deleted: allelic bias due to sequence homology in

Deleted: context of the human reference genome. The

Deleted: solution

Deleted: date

Formatted: Font: Not Bold

Deleted: ,

Deleted: allelic

Deleted: However, we note that this can be overly stringent, because it potentially removes a considerable number of sites that might still be allele-specific even after **removing reads** with

Deleted: , especially at sites with many reads.

Deleted: We investigated the effect of the allelic mapping bias (due to sequence homology) and the two removing strategies on the detection of allele-specific SNVs, in the context of the diploid personal genome. Briefly, for each individual, we **(1)** first align the reads to the two 'reference' haplotypes, each with their own sets of SNVs and indels.

Moved down [1]: For each haplotype, we **(2)** retain only those reads that uniquely mapped to regions with heterozygous SNVs, and then artificially create the same reads but with a single allele change at the heterozygous SNV position.

Deleted: (3) We then map these simulated reads to the other haplotype. For those simulated reads that align to multiple loci in the other haplotype, **(4)** we filter their original reads from the read pool and conduct another remapping and counting with the beta-binomial test to detect allele-specific SNVs. At this juncture, we cautiously note that a read can map to more than one heterozygous SNV, and they can also affect allelic ...

Deleted: allelic

Deleted: In line with previous studies, we

Deleted: %) (Table 2). ¶
¶ ...

%). Also, we further examined the set of SNVs that showed >5% allelic mapping bias and found that if we remove only the reads that exhibit allelic mapping bias, many of them are still detected as allele-specific under the beta-binomial test (Supplementary Table in mannuscript). Together, these imply that **removing by sites** can be overly stringent, because it potentially removes a considerable number of sites that might still be allele-specific even after **removing reads** with mapping bias, especially at sites with many reads. As a result, we decided on removing only reads that exhibit such a bias from the original pool of reads. This is computationally more expensive since we need to re-process the original read pile, but this strategy effectively removes potential false positives, and retains 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 [6].

Hence, in our revision, we carefully implemented an 'ambiguous-read-removal' strategy. Briefly, for each individual, we **(1)** first align the reads to the two parental haplotypes, each with their own sets of SNVs and indels. For each haplotype, we **(2)** retain only those reads that uniquely mapped to regions with heterozygous SNVs, and then artificially create the same reads but with a single allele change at the heterozygous SNV position. If the read overlaps multiple heterozygous SNVs, all possible haplotypes are generated. **(3)** We then map these simulated reads to the other haplotype. **(4)** For those simulated reads that align to multiple loci in the other haplotype, we filter their original reads from the read pool before read counting and detecting allele-specific SNVs with the beta-binomial test.

Here, we have accounted for two main types of allelic mapping bias in the context of the diploid personal genome. Additionally, our approach is already conservative, with multiple filters in place, such as removing highly over-dispersed datasets and using the beta-binomial test with an FDR of 5% for RNA-seq and 10% for ChIP-seq datasets. The personal genome is also able to handle various mapping artefacts not easily handled by using only the reference genome. Particularly, with the ability to incorporate larger variants beyond single nucleotide variants (such as indels), the personal genome serves as a more representative genome, as demonstrated by a much better alignment of unique reads [11,12]. We also envision that this ambiguous mapping bias will be further alleviated by longer reads being used in ChIP-seq and RNA-seq datasets in the near future.

Deleted: ; for example, 5 out of 11

Formatted: Font: Bold

Deleted: with >5% allelic bias (CTCF ChIP-seq dataset 2) and 4 out of 10 AS SNVs (RNA-seq dataset 1) were

Deleted: considered

Deleted: (Table 2). ¶
¶

Deleted: only

Deleted: and then re-align the filtered read pool to both haplotypes.

Deleted: at the same time,

Deleted:

Moved (insertion) [1]

Deleted: allelic

Deleted: with

	[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] Dixon *et al.* (2015). *Science*, 518(7539):331-6 [11] Rozowsky *et al.* (2011). *Mol Syst Biol*., 7:522 [12] Sudmant *et al.* (2015). *Nature*, 526(7571):75:81 We have included new sections in the 'Results', 'Discussion' and 'Methods' section about our new module on allelic mapping bias.
Excerpt From Revised Manuscript	

Deleted: addition

-- 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. In particular, we have utilized our approach in the 1000 Genomes Structural Variant group, whose manuscript has recently been peer-reviewed and published by *Nature.* 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]. Our revised approach estimates over-dispersion at two levels. An over-dispersion 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 and is introduced to homogenize the pooling of datasets, by removing datasets that are too over-dispersed at the outset. 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	

Deleted: those

Deleted: originate

Deleted: 20132

Formatted: Font: Not Italic

Reviewer #3

-- Ref3.1 – Endorsement for publication --

Deleted: General positive comment

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.[35]", where reference 35 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*.