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

RESPONSE LETTER

Reviewer #1

-- Ref1 – General positive comment --

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 agree with the reviewer that allelic mapping bias can be an issue, and it has first been mentioned in Degner *et al* [1]. We are aware of the allelic bias. We believe that it is accounted for, or at least largely alleviated, by the construction of two parental genomes. Here, we performed additional analyses to show that allelic bias only affects a small proportion of our results. We attribute this to our approach being already conservative, such as

Deleted: We

Deleted: this

Deleted: , demonstrating that

Deleted: is

	filtering highly over-dispersed datasets and using the beta-binomial with an FDR of 5% or 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 much better alignment of unique reads. Together, these conservative thresholds, filtering steps, the accommodation of larger variants and not using the reference genome are able to detect allele-specific SNVs with already a low number of false positives. Moreover, there is indeed still a discussion in the community on how to handle these issue. For example, while Kasowski *et al* [2] and Ding *et al.* [3] accounted for several other biases, both did not account for allelic bias, the former using personal genomes while the latter used the reference genome. [1] Degner *et al.* (2009) *Bioinformatics.* 25(24) [2] Kasowski, M. et al. (2013). Science. 342(6159):750-2 [3] Ding, Z. *et al.* (2014). *PLoS Genet.* 10(11):e1004798
Excerpt From Revised Manuscript	

Deleted: alleviates this type of allelic 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, 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] presented a software that perform alignment to the human reference genome, accounts for allelic bias and allele-specific detection using the beta-binomial test to account for a 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 accepted by *Nature.* Our revised approach estimates over-dispersion at two levels. An over-dispersion is estimated for each individual dataset to remove *entire datasets* that are deemed too over-dispersed and 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 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. While the estimation of a global over-dispersion has also been employed extensively in many recent software that detects allele-specific expression [1-5], we point out that our two-step serial procedure is novel and homogenizes the pooling by removing datasets that are too over-dispersed in the first place. The two-step procedure additionally facilitates our uniform processing of large amounts of data and alleviates an ascertainment bias in which more positives might originate 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). *bioRxiv*. doi: http://dx.doi.org/10.1101/011221 [2] Sun (20132). *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	

Formatted: Font: Not Italic

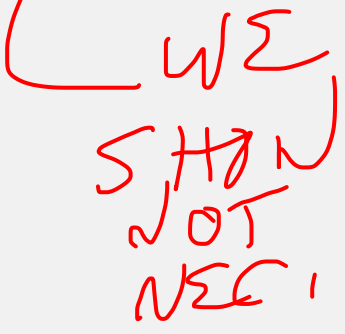
Deleted: Also, our

Deleted: individual (

Deleted: latter step have

Deleted: also

Deleted: Perhaps we were not sufficiently clear, we have amended the manuscript to better reflect this.



Revised Manuscript	

Reviewer #3

-- Ref3.1 – 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*.