Dear Editor,

We would like to thank both reviewers for their insightful comments. In almost all cases, we have revised the manuscript in accordance with their recommendations. Our responses to the reviewers’ comments are below.

We would also like to ask for a waiver of publishing charges. This work is unfunded. A letter from our head of department is attached.

Responses to Reviewer 1’s General Comments

*1). I think it is misleading to say that there is no multiple testing correction problem associated with multi-locus models without proof. Indeed, my intuition behind this matter is that as the number of markers increase, it is going to become more difficult for a multi-locus model to identify a quantitative trait nucleotide associated with a trait. 5his [sic] is why e.g., entry/exit permutation thresholds for stepwise model selection get more stringent as the number of markers increase. While multiple testing is not as straightforward in multi-locus models as it is in single locus models (e.g., do a Bonferroni or Benjamini-Hochberg correction across 1 million single-locus tests fitted at each genome-wide marker), I argue that it is still there. Simply put, the more markers that are available to be put into a final multi-locus model, the more stringent the criterion needs to be to make sure that the correct model is selected. Bogdan, Ghosh, and Doerge (Genetics, 2004 DOI: 10.1534/genetics.103.021683) does a great job of describing this, and how, e.g., a modified Bayesian Information Criterion (similar to the EBIC discussed in this manuscript) for stepwise model selection can translated to a multiple testing correction problem.*

Response – Manuscript revised: A thought provoking comment. The multiple testing issues comes from the number of hypothesis tests being performed. In the case of Eagle, only a few tests are performed in building the “best” model. In the case of locus-by-locus (single) locus methods, a large number of tests are performed because each locus is tested separately. We have revised the text to make this point (Lines 105 – 117 in the revised manuscript). While it is true that it becomes more difficult to find true associations as the number of SNPs increases, this is not a multiple testing issue. This is concerned with signal verse noise. The signal (true association) needs to be stronger in order for it to be disentangled from the background noise (SNPs not in association with the trait). We have also removed the statement, line 66 in the original paper, that “multiple testing is not an issue” for multi-locus methods. For some multi-locus methods, multiple testing is still an issue.

*2). First part of comment. If I am understanding equations (2) and (3) correctly, is it true that the only way that this differs from the MLMM is that an identity matrix is being used to model the variance-covariance between individuals instead of, e.g., an additive genetic relatedness matrix? That is, if we were to replace this identity matrix with an additive genetic relatedness matrix, would we get exactly the same results as the MLMM?*

Response: In equation (2), the random effect you are referring to is a-S and yes, we assume its (co)variance structure follows a simple identity matrix. However, a-S does not contain individual effects like in MLMM. It contains SNP effects. Also, in equation (2), these random SNP effects are multiplied by the model matrix M-S. This matrix contains the SNP genotypes. This then means the (co)variance structure for our genetic model is M-S σa2 I Mt-S = σa2M-S Mt-S . This is discussed in the reference Verbyla *et al.* 2007 that we cite at the beginning of the section. Consequently, the (co)variance matrix used in Eagle for the genetic component of the model is very similar to MLMM. The biggest differences lie in how we identify SNP to enter the model , our use of dimension reduction to reduce computation, and how we implement Eagle in R. Eagle is significantly faster, computationally, than MLM. These differences are discussed in lines 217 to 234 of the revised manuscript.

*2). Second part of comment. I suggest that the authors analyze real marker and phenotype data from, e.g., a crop diversity panel with a high amount of familial relatedness to show that Eagle sufficiently controls for false-positive associations arising from familial relatedness and population structure.*

Response: Fortunately, this has already been done by Arunas Verbyla (co-author) through a series of methodological publications in the plant space (Verbyla *et al.* 2012 TAG, Verbyla *et al.* 2012 Genetic Research, Verbyla *et al.* 2014 G3, Verbyla *et al.* 2014 TAG,). These publications evaluate the statistical behaviour (including false-positive rate) of the methodology upon which Eagle is based.

*3.) In terms of computational efficiency. Although I am not suggesting that the following is mandatory, I think it would be great (if feasible given time constraints) if the simulation study were to be revised so that i.) the sample size is held constant while the number of markers increased and ii.) the number of markers was held constant while the sample size increases. This would allow a rigorous evaluation of computational performance as a function of sample size and number of markers.*

Response: We did consider this in the early design phase of the simulation study. We opted for a set of scenarios that typify association studies in a range of species. However, in retrospect, what you have suggested would have been the better option. Unfortunately, the computer cluster upon which our study was based has since been decommissioned. To do what you suggest would mean several months of work as we would need to begin the simulation study from scratch. For this reason, we have chosen not to do this here. However, we are developing, currently, a GPU version of Eagle. As part of writing this work up, we will be performing a simulation study to demonstrate how performance scales with sample size and SNP number. Here, our simulation study will be designed as you have suggested.

Responses to Reviewer 1’s comments made while reading the manuscript

*Line 38: Cite the statistical methodology, not the software.*

Response: [Manuscript revised]. We have replaced the two citations to software papers with methodological references.

*Line 52: This wording is a bit sloppy. I suggest describing this in terms of controlling the experimentwise versus the comparisonwise type I error rate.*

Response: [Manuscript revised]. We have restructured and rewritten most of this paragraph (lines 42 – 58 in the revised manuscript).

*Line 56: What about the Benjamini-Hochberg (1995) procedure that controls the FDR, or Bonferroni? I see this approach being widely used in the crop GWAS community.*

Response: [Manuscript revised]. We have removed the statement that “there is still no well accepted way of correcting for multiple testing in the context of genome-wide association mapping”.

*Line 67: I disagree with this. Please provide a citation.*

Response: [Manuscript revised]. We’ve changed the wording and we have provided references as requested. Lines 67-70 in the revised manuscript.

*Line 73: Why? This needs to be cited.*

Response: [Manuscript revised]. We haven’t added a reference but we have changed the wording so that we are now explaining what we mean by “results are immediately interpretable”. Lines 76 & 77 in the revised manuscript.

*Line 96: Why are there more autosomal SNPs than the total number of SNPs across the genome?*

Response [Manuscript revised]: This is the case. There are 353,967 autosomal SNPs and 359,599 total number of SNPs. However, we had to look twice at the numbers as well. The numbers were formatted slightly differently which we think was causing the initial confusion. The numbers are now formatted the same. Line 99 in the revised manuscript

*Line 104: Please elaborate more on model selection versus variable selection, and please provide a citation.*

Response: [Manuscript revised]. We’ve removed the distinction between model selection and variable selection. We’ve also added citations for model selection. Finally, we have reworded the paragraph. Lines 105 – 117 in the revised paper.

*Line 108: In LD with what?*

Response: [Manuscript revised]. We’ve change the wording to be “in strongest linkage disequilibrium with the genes influencing a trait”. Lines 116 and 117 in the revised paper

*Line 148: Change residual maximum likelihood to restricted maximum likelihood.*

Response: [Manuscript revised]. Made the change. Line 157 in the revised paper.

*Line 194: Citations are needed for all of these approaches.*

Response: [Manuscript revised]. The citations had drifted to further down in the paper. We have moved them back to where they belong. Lines 203 – 205 in the revised paper.

*Line 200: Please add a sentence that compares and contrasts the ridge versus LASSO penalty with respect to the identifying SNPs associated with traits in a multi-locus model.*

Response: We weren’t sure what was being asked here.

*Line 203: This, right over here is multiple testing.*

Response: This has been addressed above in our response to comment 1.

*Line 208: MLMM has a backwards step in addition to a forward step.*

Response [Manuscript revised]: True. We have amended the manuscript. Line 218 in revised paper.

*Line 212: Please reword these sentences. Both the score statistic and the test used in MLMM are, fundamentally hypothesis tests, and both should testing H0: no association between a given marker and trait.*

Response: [Manuscript revised] We have reworded the paragraph to make it clear that there is no testing taking place here. The score statistic and p-value is being used only to identify the next best SNP to include in the model. Lines 221 – 234 in revised paper.

*Line 252: Please define a replicate within the context of marker and phenotypic data. What is the experimental unit?*

Response: [Manuscript revised] We found our Stability Selection section to be poorly worded. We rewrote this section. Lines 288 - 322 in revised paper.

*Line 334: Please add some more description about these data. Where were these data generated from? What species, what is the number of chromosomes? What is the size of the genome?*

Response: [Manuscript revised] All of these questions have been answered earlier in Section 2.4. We now reference this section here Line 371 in revised paper.

*Line 348: What was the effect size?*

Response: [Manuscript revised] Effect sizes have been added. Lines 380 – 381 in revised manuscript.

*Line 354: What is a chunk?*

Response: [Manuscript revised] Definition has been added. Lines 386 & 387 in revised manuscript.

*Line 393: Why 40 kb? Please justify and cite previous work, if applicable.*

Response: [Manuscript revised]. Done. Justified and cited. Lines 427-429.

*Line 456: Why were not the other multi-locus approaches tested in the mouse data analysis?*

Response: We felt that we had performed a detailed comparison of the multi-locus methods in our simulation study. In the mouse data analysis, we instead wanted to highlight the benefits of using Eagle over the method-of-choice single-locus approach used in Nicod et al., 2016.

*Line 475: Please cite the permutation approach.*

Response: [Manuscript revised] Citation added. Line 511 of revised paper.

*Line 489: False positive rate or false discovery rate?*

Response: [Manuscript revised] We used false discovery rate wrongly. It should be false positive rate. We have replaced false discovery rate with false positive rate throughout the paper.

*Line 494: Please show some Manhattan plots.*

Response: [Manuscript revised] Manhattan plots have been added to Supplementary Materials and referenced in the paper. Line 528 & 529 of revised paper.

*Line 529: How do you know that these are 100% true? Was a follow-up biological study conducted where the causal mutations were identified?*

Response: [Manuscript revised] We know because these extra results were seen in the original study. They are not seen here with our single-locus analyses because we are using (by design) less informative SNP genotype data instead of the SNP dosages as used in the original study. This was discussed in the first paragraph of Section 3.5. To make this clear, we have added an extra sentence. Line 563 – 566 of revised paper.

Responses to Reviewer 2’s Comments

*An incontrovertible contribution of this manuscript is a computationally efficient and well-documented implementation of a multi-locus association method. Lack of such implementations is perhaps one contributor to the lack of interest in multi-locus methods the authors note in their introduction. However, I am doubtful that, as presented, Eagle is a substantial methodological contribution. The authors heavily cite three previous papers (Verbyla et al. 2007, 2012, and 2014) in describing Eagle. Comparing Eagle to those papers suggests that the only real differences are in the software implementation, the use of marker genotypes instead of intervals, and the use of the extended BIC to evaluate potential covariates.*

Response: “I am doubtful that, as presented, Eagle is a substantial methodological contribution”. The methods are not brand new. However, the fact that Eagle is based on well-established methods is actually one of Eagle’s strengths. The methods are tried-and-true. The points of methodological innovation in this paper are 1. recognising the value of a method that has been locked away in the plant qtl mapping space 2. transforming the methodology into a more usable and implementable form and 3. identifying a “better” model choice mechanism (i.e. extBIC). We’ve been open about the origin of the methodological developments. We would like to leave the tone of the paper unchanged.

*2. The authors classify Eagle and MLMM as model selection (Table S1) and all other methods as variable selection. How do the authors distinguish between model and variable selection? The authors’ conception of model selection seems to be entwined with the use of information theoretic criteria denoted as “threshold free” methods. However, MLMM is not strictly threshold free: The use of the modified Bonferroni criterion instead of extended BIC imposes a Bonferroni-corrected significance cutoff on all SNP covariates included as fixed effects. The authors note that the results are similar using either cutoff.*

Response: [Manuscript Revised]: Agreed. The distinction between model selection and variable selection is a blurred one. We’ve removed this distinction from the paper. We have revised Table S1 and changed the wording of lines 102 – 111. We have also removed reference to Eagle being “threshold free”. Changes made: Table S1 revised. New wording lines 105 – 117 in revised paper.

Responses to Reviewer 2’s Suggestions for Improvement:

*1. The most directly comparable method seems to be MLMM which uses both forward and backward selection to determine a final model. Why have the authors chosen to only use forward selection in Eagle?*

Response: We found no benefit in including a backward selection step. Also, it increases computation.

*2. The authors have declined to compare Eagle against multi-locus methods that are more recent than MLMM, specifically, FarmCPU (*[*https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005767*](https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005767)*) and BLINK (*[*https://academic.oup.com/gigascience/article/8/2/giy154/5238723*](https://academic.oup.com/gigascience/article/8/2/giy154/5238723)*). Is there a particular reason for not comparing Eagle to either or both of these methods?*

Response: [Manuscript revised] There was a significant time gap between the running of the simulation study and the writing of the paper due to a change in job definition for the lead author. We have addressed this issue by running FarmCPU on the mouse data. We are unable to include FarmCPU in the simulation study as the computing cluster upon which all the runtimes are based has been decommissioned. We present the results from FarmCPU and differences with Eagle in the discussion (Lines 600-619 in revised manuscript).