Dear Dr. Daetwyler and Dr. Long,

We were pleased with the useful feedback given by reviewers to our manuscript “Genetic Association**s in Four Decades of Multi-Environment Trials Reveal Agronomic Trait Evolution in Common Bean”. We have addressed all the editorial issues suggested by the three reviewers, and included in the attached cover letter are point-by-point responses to these issues.**

**We seriously considered the analysis changes suggested by reviewers one and three, and in fact redid the entirety of our analysis using the suggested model changes. However, for reasons we elaborate on below, we chose to keep our original BLUP model, as the GxE models that were possible given our sparse dataset resulted in a large loss of power to detect significant associations but did not substantially alter either the rank order of the association results, or the mash analysis of shared SNP effects.**

**Overall, we are grateful that these reviewer criticisms have improved the manuscript, which we present here as a resubmission to Genetics. In the replies below, all line numbers refer to the word document with tracked changes. Looking forward to hearing from you.**

Sincerely,

Alice MacQueen

GENETICS/2019/302613   
**Genetic Associations in Four Decades of Multi-Environment Trials Reveal Agronomic Trait Evolution in Common Bean**   
  
Dear Dr. MacQueen:   
  
Three experts in the field have reviewed your manuscript, and I have read it as well, and I am pleased to tell you that it is potentially suitable for publication in GENETICS. However, the reviewers have suggested additional analyses and have provided in-depth comments and concerns that need to be addressed in a revised manuscript. You can read their reviews at the end of this email.   
  
If you submit a revised manuscript please include a response to each of the reviewers' comments. I will highlight several here, but please consider all comments made by the reviewers. R1 suggest investigating the use of factor analytic models to generate the BLUPs for downstream analysis. My own experience has also been that FA models may be more appropriate in data structure such as yours. R2 ask that you discuss why a key regions was not detected in your analyses. R3 suggest more work is needed to be able to claim that pleiotropy and not linkage has been observed. They also suggest, as the use of Mash is emphasized in the manuscript, that you may want to compare Mash to another method with similar aims.   
  
I look forward to receiving a revised manuscript. I expect it could be submitted within 90 days, but please let me know if you think you will need more time to complete the revision. A revised manuscript will be considered a resubmission, and may be sent out for review.   
  
Follow this link to submit the revised manuscript: <https://genetics.msubmit.net/cgi-bin/main.plex?el=A3NR5CLP7A7JPq4I7A9ftdmmOpcr6vwXLMGsQygHL0xQZ>   
  
If you have questions about the reviews or this message, please contact me.   
  
  
Sincerely,   
  
Hans Daetwyler   
Associate Editor   
GENETICS   
  
  
Anthony Long   
Senior Editor   
GENETICS

Reviewer #1 (Comments for the Authors (Required)):   
  
The main objective of the manuscript is to identify genomic regions regulating multiple traits in common bean. To identify these regions, the authors apply the mash method, proposed by a recent paper by Urbut et al. (2019, Nat. Genetics) to the genotypic BLUPs (genetic main effects) for multiple traits obtained across environments. By calculating the genotypic main effects across environments, the authors circumvent the issue of high imbalance in the multi-environment data. That allows them to concentrate on the multi-trait analysis. The main question addressed in this paper is relevant to many crops and although the paper implements already existing methodology, they are the first ones (to my knowledge) to implement this method in crop data coming from historical multi-environment data. The manuscript is well written and the supporting data and code available in GitHub are clear and complete. That certainly increases the impact of the paper, as they facilitate the implementation in other crops. However, I do have a few comments:

- My main concern is about the calculation of the BLUPs for the genotypic main effect. The authors used rrBLUP to fit a model with locations, years and their interactions as fixed and with genotype main effects as random. The variance-covariance between genotypes was modelled with a compressed Kinship matrix calculated with GAPIT. The model that was used looks rather incomplete to me. I suggest to model the genotype by environment interaction explicitly (with a factor analytic structure, for example) and then produce the BLUPs for the genotypic main effects. Or, at least, to decompose the GxE into GxL, GxY and GxLxY and then produce the predictions for the genotypic main effect. That would also probably improve the heritability estimates, which are rather low now as part of the GxE variance is accounted as residual variance.

*This reviewer and reviewer three bring up good points about model specification for this dataset. We made the model changes that these reviewers suggested and redid the entirety of our analysis using BLUPs from the new models. However, for the reasons we elaborate on below, we decided to keep our original model, as the model changes that were possible given our sparse GxE matrices resulted in model overparameterization, aliasing of CDBN entry coefficients, and a subsequent loss of power to detect significant associations.*

*We decided to decompose the GxE matrix into GxL and GxY, and not use factor analysis, because our GxE matrices were so sparse that we could only do factor analysis on 9 GxL matrices. We thought that using 9 factor analytic models, 10 models with GxE decomposed into GxL and GxY, and 3 models without GxE would be confusing and complicate our downstream comparisons of SNP effects. As discussed in the paper, the GxE matrices for each phenotype were extremely sparse – the genetic correlation matrices constructed from GxLxY data never have more than 15% of cells with correlations. Since most genotypes were only grown in sequential years, for 1-4 years, the GxY matrix was even sparser – no genetic correlation matrix for any phenotype from GxY data had more than 6% of cells with correlations. Given that factor analysis requires matrices without missing values, we judged that these matrices did not have enough complete cases for factor analysis. We are unaware of methods to model such sparse matrices. However, 9 phenotypes had between 40% and 99% of genetic correlations constructed from GxL data: seed yield, seed weight, days to maturity, days to flowering, seedfill duration, biomass, plant height, harvest index, and lodging. For these nine phenotypes, we could have used factor analysis to produce covariates for the GxL matrix, but chose not to in the interests of model consistency.*

*We incorporated both reviewer one and reviewer three’s comments (that effects involving year and location should be fitted as random effects) to fit a model for phenotype in ASReml with CDBN Germplasm Entry as a fixed effect, and with four random effects: Location, Year nested within Location, Location nested within CDBN Germplasm Entry, and Year nested within CDBN Germplasm entry. We could fit this model for 19 of 21 CDBN-derived phenotypes.*

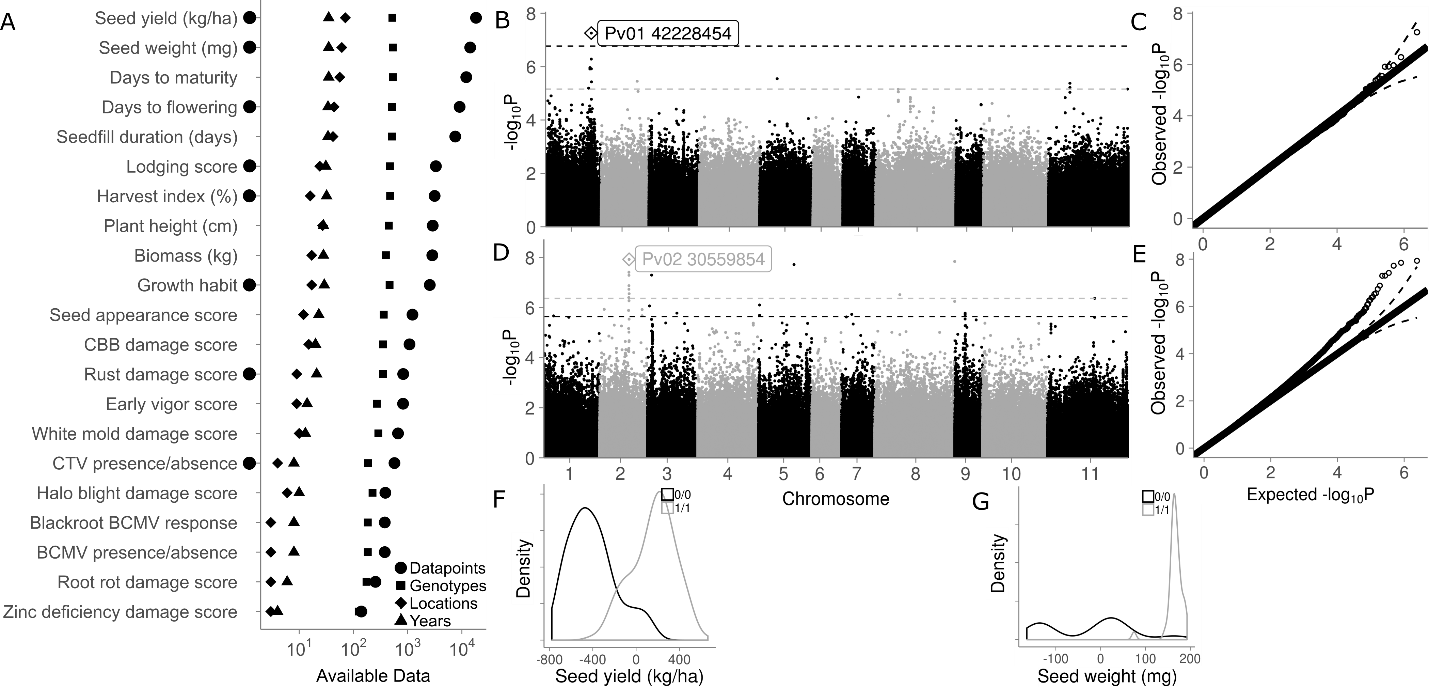
*As the reviewer suggests, this change does markedly improve the heritability estimates, as measured by narrow-sense heritability calculated from GAPIT. Below is a table of Vg, Ve, and h2 for each model, the original model in rrBLUP and the new model in ASReml:*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Phenotype | ASReml\_Vg | ASReml\_Ve | ASReml\_h2 | Sig | rrBLUP\_V\_g | rrBLUP\_V\_e | rrBLUP\_h2 | Sig |
| Seed yield (kg/ha) | 8.00E+04 | 3.99E+04 | 0.667 | Y | 5.32E+04 | 2.22E+05 | 0.193 | Y |
| Seed weight (mg) | 3560.03 | 937.94 | 0.791 | Y | 2340 | 1076 | 0.685 | Y |
| Days to maturity | 4.82 | 6.65 | 0.420 | N | 12.3 | 18.3 | 0.402 | N |
| Days to flowering | 5.71 | 0.17 | 0.971 | N | 4.68 | 6.11 | 0.434 | Y |
| Seedfill duration (days) | 3.27 | 4.35 | 0.429 | N | 5.584 | 17.812 | 0.239 | N |
| Lodging score | 0.33 | 0.11 | 0.750 | Y | 0.244 | 0.466 | 0.344 | Y |
| Harvest index (%) | 14.82 | 17.09 | 0.464 | N | 13.5 | 31.5 | 0.299 | Y |
| Plant height (cm) | 25.38 | 7.69 | 0.767 | N | 13.8 | 42.5 | 0.245 | N |
| Biomass | 2.11E+05 | 2.04E+05 | 0.508 | N | 2.46E+05 | 6.72E+05 | 0.268 | N |
| Growth habit | 0.25 | 0 | 1.000 | Y | 0.160 | 0.154 | 0.509 | Y |
| Seed appearance score | 0 | 0.15 | 0.000 | N | 0.017 | 0.241 | 0.067 | N |
| CBB damage score | 0.38 | 0.99 | 0.277 | N | 0.282 | 1.127 | 0.200 | N |
| Rust damage score | 4.44 | 0 | 1.000 | Y | 3.201 | 1.957 | 0.621 | Y |
| Early vigor score | 0.43 | 0.35 | 0.551 | N | 0.064 | 0.747 | 0.078 | N |
| White mold damage score | 0.59 | 0.54 | 0.522 | N | 0.143 | 0.631 | 0.185 | N |
| CTV presence/absence | 0.03 | 0.1 | 0.231 | N | 0.025 | 0.132 | 0.157 | Y |
| Halo blight damage score | 0.27 | 0.89 | 0.233 | N | 0.176 | 0.708 | 0.199 | N |
| BCMV presence/absence | 0.01 | 0.08 | 0.111 | N | 0.016 | 0.097 | 0.145 | N |
| BCMV blackroot response | 0.13 | 0.12 | 0.520 | N | 0.049 | 0.086 | 0.363 | N |
| Root rot damage score | 0.07 | 0 | 1.000 | N | 0.455 | 3.193 | 0.125 | N |
| Zinc deficiency damage score | 2.83 | 0.39 | 0.879 | N | 3.831 | 1.395 | 0.733 | N |
| First Year in the CDBN | 79.15 | 23.49 | 0.771 | N | 79.15 | 23.49 | 0.771 | N |

*However, we noticed a large drop in statistical power to detect associations (~1/3 of our power) as a result of these model changes. Our original models gave 8 phenotypes with associations above a 10% FDR, and this model gave five (though the top SNPs for the other traits still retained some of the highest p-values). Instead of 33 associations above the 10% FDR threshold, we found 22. Phenotypes now required even more datapoints (on average, 7800 instead of 6500) and higher heritability (h2, on average, of 84.2% instead of 40.5%) to have associations above the 10% FDR threshold. This power drop was likely caused by the aliasing of coefficient estimates for many CDBN germplasm entries, aliasing that was caused by overparameterization of random effects in the model relative to the amount of data that exists for each phenotype.*

*The variance explained by GxE terms of these models, and particularly the GxY terms, was very small. We are certain that this would not be the case for full rank GxE models for the same germplasm; however, that is not the dataset we have. Here are the tables of fixed and random effects for three phenotypes. The GxE terms are Taxa:Year!Taxa and Taxa:Location\_code!Taxa, and the variance explained by each of these is in the Var % column. The variance explained by the GxE terms for most phenotypes fell between 0 and 1%.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Seed Yield Fixed Effects | Df | Sum of Sq | Wald statistic | Pr(Chisq) |  |  |
| (Intercept) | 1 | 812.6 | 812.6 | < 2.2e-16 | \*\*\* |  |
| Taxa | 326 | 4063.3 | 4063.3 | < 2.2e-16 | \*\*\* |  |
| residual (MS) |  | 1 |  |  |  |  |
|  |  |  |  |  |  |  |
| Seed Yield Random Effects | component | std.error | z.ratio bo | und % | ch | Var % |
| Location\_code!Location\_code | 335894 | 78579 | 4.3 | P | 0 | 34% |
| Location\_code:Year!Location\_code | 424338 | 26808 | 15.8 | P | 0 | 43% |
| Taxa:Year!Taxa | 3993 | 944 | 4.2 | P | 0 | 0% |
| Taxa:Location\_code!Taxa | 60299 | 2984 | 20.2 | P | 0 | 9% |
| units!units | 160946 | 2757 | 58.4 | P | 0 | 16% |
| units!R | 1 | NA | NA | F | 0 |  |
|  |  |  |  |  |  |  |
| Days to flowering Fixed Effects | Df | Sum of Sq | Wald statistic | Pr(Chisq) |  |  |
| (Intercept) | 1 | 3476.8 | 3476.8 | < 2.2e-16 | \*\*\* |  |
| Taxa | 320 | 5128 | 5128 | < 2.2e-16 | \*\*\* |  |
| residual (MS) |  | 1 |  |  |  |  |
|  |  |  |  |  |  |  |
| Days to flowering random effects | component | std.error | z.ratio | bound | %ch | Var % |
| Location\_code!Location\_code | 23.581862 | 6.2668663 | 3.762943 | P | 0 | 46% |
| Location\_code:Year!Location\_code | 21.3901562 | 1.87775137 | 11.391368 | P | 0 | 42% |
| Taxa:Year!Taxa | 0.1136479 | 0.04740993 | 2.397133 | P | 0 | 0% |
| Taxa:Location\_code!Taxa | 1.0977744 | 0.11221085 | 9.78314 | P | 0 | 4% |
| units!units | 4.9142913 | 0.12366747 | 39.737945 | P | 0 | 10% |
| units!R | 1 | NA | NA | F | 0 |  |
|  |  |  |  |  |  |  |
| Days to maturity Fixed Effects | Df | Sum of Sq | Wald statistic | Pr(Chisq) |  |  |
| (Intercept) | 1 | 6980.7 | 6980.7 | < 2.2e-16 | \*\*\* |  |
| Taxa | 324 | 3364.2 | 3364.2 | < 2.2e-16 | \*\*\* |  |
| residual (MS) |  | 1 |  |  |  |  |
|  |  |  |  |  |  |  |
| Days to maturity random effects | component | std.error | z.ratio | bound | %ch | Var % |
| Location\_code!Location\_code | 58.4310788 | 14.3335957 | 4.076512 | P | 0 | 44% |
| Location\_code:Year!Location\_code | 55.1262554 | 4.228945 | 13.035463 | P | 0 | 42% |
| Taxa:Year!Taxa | 0.9184755 | 0.1538849 | 5.968589 | P | 0 | 1% |
| Taxa:Location\_code!Taxa | 3.8566817 | 0.2905915 | 13.271834 | P | 0 | 5% |
| units!units | 13.8033056 | 0.3050514 | 45.249111 | P | 0 | 10% |
| units!R | 1 | NA | NA | F | 0 |  |

*****This model change caused little change in our major results and most significant associations. The association above the FDR for seed yield, three of the top four associations for seed weight, one of the two signals above the FDR for lodging, the top four associations above FDR for growth habit, and the top association for rust above the FDR were all unchanged. The associations we saw above the FDR for curly top virus damage score, harvest index, and days to flowering were also present in the top ten associations by p-value for these phenotypes. Essentially, the mash results do not change* - *the vast majority of significant SNPs still cluster within two large regions on Pv01 (Fig. 3b, Table S5), from 5.3Mb to 20.5Mb, and from 34.1Mb to 45.4Mb. These associations encompass two regions of intermediate gene density surrounding the centromere on Chromosome 1, but do not include the gene-rich edges of this chromosome (5 and 6Mb in size, respectively). The identity of the top SNPs in these regions does change, but our former top SNPs still have high -log10(Bayes Factor) values. Finally, the patterns of allelic effects, both in terms of breeding value correlations and mash effect correlations, do not change markedly. Below are comparisons of two GWAS and the mash results between the original model and the ASReml model. To compare these results in detail to our previous results, compare the tracked changes to the original text in the document “METG\_Genetics\_ASReml.docx” between lines 293 and 608.*

***Seed yield and seed weight Manhattans and QQ plots for the original model.***

***A screenshot of a cell phone

Description automatically generated***

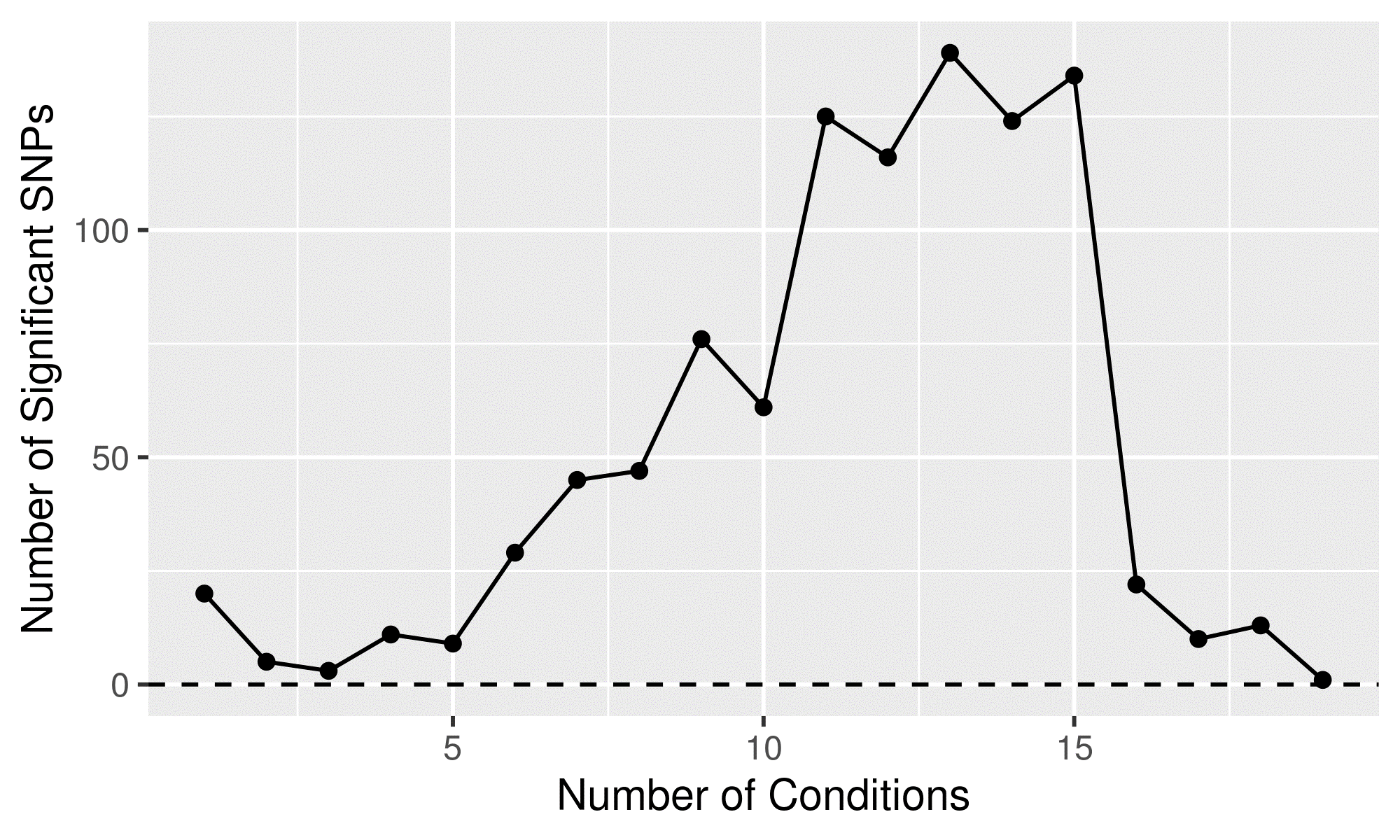
***Seed yield Manhattan & QQ plot for the new model in ASReml.***

***Seed weight Manhattan & QQ plot for the new model using ASReml.***

*A close up of a map

Description automatically generated*

***A close up of a logo

Description automatically generated***A screenshot of a cell phone

Description automatically generated***Original mash results***

***Mash results using associations from the new models using ASReml.***

*Essentially, the choice of model specification here boils down to: do we believe that the phenotypes we model, derived from those in the CDBN, are under genetic control? If so, we should work to maximize our power to detect genetic associations, so long as changing the underlying model does not have a major impact on the underlying associations we detect. The more robust associations are with respect to model specification, the more confident we can be that the associations are real.*

*We think these phenotypes are under genetic control, and so we keep our previous analysis, although we now mention this alternate analysis and host the results on Github (lines 298-304).*

Minor comments   
- Line 273: the address should be <https://github.com/Alice-MacQueen/CDBNgenomics/tree/master/data-raw>, instead of <https://github.com/Alice-MacQueen/CDBNgenomics/tree/master/analysis-paper>

*Edited.*

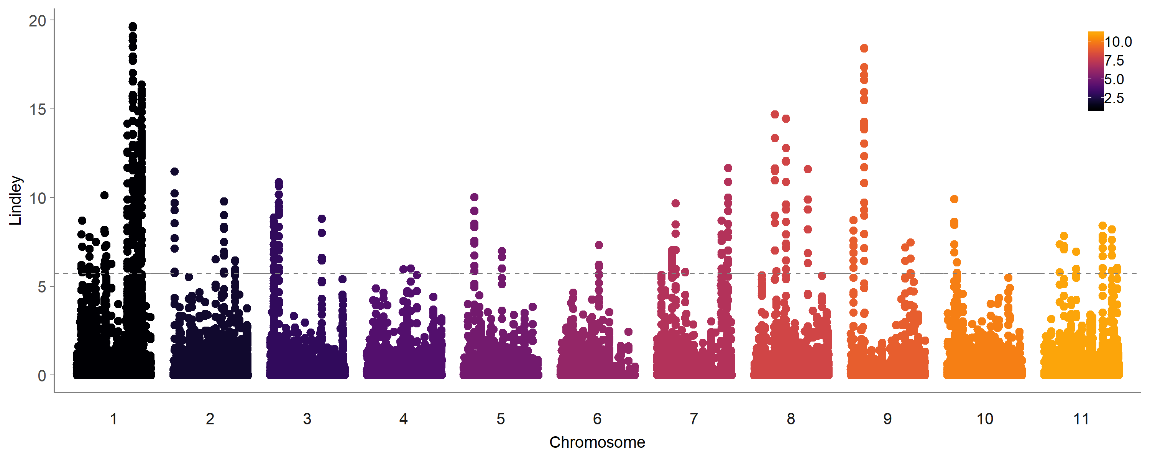
- Figure 3. The point estimates in sub-figure D and E appear very small. Please use the same size as in sub-figure F

*Edited to standardize the size of the point used to show the point estimate.*  
  
Reviewer #2 (Comments for the Authors (Required)):   
  
Review of manuscript GENETICS/2019/302613, "Genetic Associations in Four Decades of Multi-Environment Trials Reveal Agronomic Trait Evolution in Common Bean "   
Authors: Alice H. MacQueen, Jeffrey W. White, Rian Lee, Juan M. Osorno, Jeremy Schmutz, Phillip N. Miklas, Jim Myers, Phillip E. McClean, Thomas E. Juenger   
  
The authors mine phenotypic data collected from bean trials conducted over 70 years across 10-20 locations in North America to detect significant genetic effects that define genomic regions associated with traits of agronomic importance. Much of the findings align with genomic regions previously identified in QTL and GWAS studies. One of the surprising results were the genomic regions not detected by the MET analysis. The fundamental reason to conduct the CDBN trials was to measure performance (yield) of new bean lines across locations and years. The MET analysis detected genomic regions associated with yield on Pv01 adjacent to the determinacy allele which dramatically affects yield but the analysis did not detect a critical region on Pv03 (~38.27Mb) that has been shown to be associated with yield in 10 independent QTL and GWAS studies (Kelly, 2018). Interestingly, a candidate gene for days to maturity was detected in this same region of Pv03 (36.8Mb) and maturity is a major factor in plant yield. This brings into question the value of the MET analysis as a tool to detect genomic regions associated with the major economic driver in bean breeding.

*We anticipated we might have seen this region, but we didn’t: here are two possible reasons for why that might be.*

*One possibility is that the additional noise inherent in a GWAS on data from multiple locations and multiple years – even though partly accounted for in our model – gave us very low power to detect associations of minor, or even moderate effect (phenotypic variation explained by the SNP of, say, 5 – 15%. With two exceptions, studies in Kelly 2018 had the R2 of the most significant association on Pv03 of less than 0.15/15%. In addition, the additive effect in these studies of this locus on yield ranged between 0.7 kg/ha and 219 kg/ha, with the average effect 110-140 kg/ha. Our only significant association was correlated with a difference in seed yield of 104 kg/ha. If this region had an effect of less than 104 kg/ha, as it did in five out of 14 studies (36%), we would not have detected an effect of this region.*

*When we used the local score approach, which has increased power to detect associations of smaller effect (Bonhomme et al 2019), we detected a significant region for seed yield on Pv03 at 36.988 Mb (and a very strong hit for days to maturity only 55kb away), confirming that this region does have an effect on seed yield in this panel, but that our original approach did not have power to detect it.*

***Manhattan for seed yield using the local score approach and ξ = 2, which maximizes the power to detect associations of small effect.***

*A close up of a logo

Description automatically generated****Manhattan for days to maturity using the local score approach and ξ = 2, showing a major association on Pv03.***

*There are many caveats to an approach like the one we used, and thus we considered some overlaps with previous studies necessary to demonstrate the usefulness of this method, but no particular single overlap or lack of overlap, even a robust one of moderate effect, either sufficient or damning evidence for or against the efficacy of this method. We now mention that we detect this region for days to maturity, but not seed yield, and cite the paper.*

Lines 516 and 707: The authors state an alternate allele of Ur-3 first 'appeared' in pinto Sierra and Starlight great northern. 'CDBN until 1988, when it appeared in the pinto Sierra and the great northern Starlight". This sentence needs to be rewritten as new resistant alleles do NOT just 'appear' in new varieties they are bred into the variety over generations of selection. Replace 'appeared' with 'was first observed'

*True. Edited.*  
  
Line 522: The determinacy factor (growth habit) in navy beans was first reported on Pv07 by Kolkman and Kelly (2003). No actual physical mapping was possible in common bean in that era but the same genomic region was confirmed later by Kwak et al. (2012). All the factors listed as associated with the alternate allele on Pv07 (line 601) would be associated with determinacy.

*We add this citation, line 503.*  
  
Lines 523-524: The authors state: "Phvul.001221100, recently identified as the photoperiod sensitivity locus Ppd, or PHYTOCHROME A3 (WELLER et al. 2019)". The association of days to flower and the Ppd gene and the candidate phytochrome A3 gene was first reported by Kamfwa et al. (2015) on Pv01 at 48.3Mb. Please acknowledge the first report.   
*Citation added, line 536.*

Line 545: 'early on' replace with 'in the 1980s'

*Edited.*

Line 686: suggest adding '........delayed flowering to prevent pod set on lower nodes (Vandermark et al. 2014)'.

*Edited.*  
  
[Citation: Kelly, J.D. 2018. Developing improved varieties of common bean, 2:3-17. In: Achieving sustainable cultivation of grain legumes. Vol. 2: Improving cultivation of particular grain legumes (eds. Sivasankar S., Bergvinson D., Gaur P., Agrawal S.K., Beebe S., Tamò M.) Burleigh Dodds Science Publishing, Cambridge UK, pp. 376.]   
  
[Citation. Kolkman, J.M. and J.D. Kelly. 2003. QTL conferring resistance and avoidance to white mold (Sclerotinia sclerotiorum) in common bean (Phaseolus vulgaris). Crop Sci 43:539-548.]   
  
  
  
Reviewer #3 (Comments for the Authors (Required)):   
  
This manuscript describes a genome-wide association study of agronomic traits and yield components in common bean using several decades of breeder phenotyping in a cooperative trial system (the CDBN). The study uses single trait GWAS and a multi-trait method (mash) to identify associations. It argues that the colocalization of association signals reflect shifts in the ideotypes over several decades of breeding. This is an interesting issue, of interest from both a plant breeding and evolutionary genetics perspective. The common bean nurseries provide a good system to investigate it.   
  
Strengths: The study rigorously develops a genotype-phenotype data set including (i) the digitization and curation of multi-decade agronomic phenotype resource and (ii) genotyping of unarchived entries from CDBN to generate a more complete data set. These analyses are well described in the text and the supplemental material, which should facilitate reuse of this resource. The main hypothesis of the study, that shifts in bean ideotypes in the past several decades have led to selection on pleiotropic loci, is potentially interesting to a wide range of geneticists.   
  
Weaknesses: The main hypothesis of the study is not consistently addressed throughout the manuscript. Instead much of the manuscript is exploratory (see objectives in line 166-168). The hypothesis could be further tested by genome scans. The prediction would be a significant enrichment of outlier SNPs within the mash peak regions. Evidence of pleiotropy from the single trait GWAS is also not clear. A systematic analysis of colocalization (compared to an appropriate null model) would be needed to show evidence of pleiotropy.

Specific points to address:   
  
Line 147-149: "Selection for a common bean crop ideotype ... is known to have led to pleiotropic effects...". This claim is not supported by existing evidence. A small phenotypic association study is cited to support the claim (Soltani et al. 2016), but this study does not distinguish pleiotropy vs. linkage. Instead this claim appears to be the main hypothesis to be tested in the current study. (Line 712-713 makes the same claim: "Selection for the common bean ideotype is known to have led to pleiotropic effects")

*From these comments, it appears that we should better clarify the main message of the paper. We wanted to test whether, when a genetic component could be added to MET trials, genetics on MET trials might reveal similar information to balanced GWAS – if we could recover stable genetic effects from an unbalanced MET trial. We also wanted to see whether this kind of analysis might be used to uncover novel results – and certainly, looking at linked or pleiotropic effects using this wealth of phenotypic data is a novel undertaking, at least for this species. We also wanted to create a resource for other researchers to use this dataset or to apply these techniques to other MET datasets. We now emphasize this in the discussion/ conclusion (lines 792-794).*

*However, we are neither methods developers nor able to decisively show that multiple phenotypic effects falling in close genomic proximity are due to pleiotropy. Neither our study nor the previous study we cite provide conclusive evidence that pleiotropy, rather than linked effects, are producing these patterns of genotype-phenotype association. We are of the opinion that colocalization is not sufficient to distinguish between linkage and pleiotropy. Instead, we think that genetic manipulation is necessary to distinguish between pleiotropy and linkage. This kind of genetic manipulation is outside of the scope of this paper. In the paper, we are careful to say that these shared effects are due to ‘pleiotropy or linked effects’.*

*We now expand the mash results and discussion to emphasize that when mash detects a 15Mb and a 11Mb region as having significant effects on multiple phenotypes, that almost certainly indicates linked effects on multiple phenotypes (though it does not preclude pleiotropy also occurring at a locus or loci within this large region). (line 607-609; lines 690 – 699)*

Line 159-161: The text highlights the novelty and value of using mash for crops, here and elsewhere in the manuscript. If this is an important point to make, then this method should be compared to other methods, e.g. MTMM (Korte et al. 2012 Nat Gen) or similar. Otherwise it doesn't seem relevant to emphasize this point.

*This is not the point of our paper – we are not methods developers, and we’re not astute enough in the methods development and comparison approaches to be able to do this comprehensively or well. We now stress that we use this exciting new method to determine if we can use this kind of analysis to uncover novel results. (lines 746-747).*

Line 272: For location and year the variance is the parameter of interest so a random effect seems more appropriate.

*We made the model changes that this reviewer, as well as reviewer one, suggested and redid the entirety of our analysis using values from the new models. However, for the reasons we elaborate on in our response to reviewer one, we decided to keep our original model, as the model changes that were possible given our sparse GxE matrices resulted in model overparameterization, aliasing of CDBN entry coefficients, and a subsequent loss of power to detect significant associations.*

Line 280-282: I know it's common to use the term "control" to describe the population structure (Q) and kinship terms (K), but it's not clear that they "control" anything and I'd argue it's more appropriate to say the Q and K terms "account" for genetic background effects (Vilhjálmsson and Nordborg 2013 Nat Rev Genet).

*Agreed. Edited.*  
  
Line 286, 307-309: Explain these distance cutoffs relative to the extent of LD in the germplasm. Overall, the lack of a linkage disequilibrium analysis in the study makes it difficult to interpret the distance cutoffs (e.g. 20 or 200kb) used to identify the candidate genes.

*We agree that distance cutoffs are arbitrary and certainly meaningless without knowledge of LD. We had relied on region sizes determined in previous work that incorporated LD analysis on an overlapping panel of bean germplasm (Moghaddam et al). We now mention this work as justification for our interval sizes (line 321).*

*During our revisions, we became aware of a new method to better account for LD structure effects on p-values for specific genomic regions, and the localization of effects in the genome, known as the local score approach (Bonhomme et al 2019). This approach uses the cumulative association signals caused by LD between SNPs in a genomic region containing a causal variant to improve GWAS resolution and increase the power to detect QTL of minor effect. The authors suggest using a local score approach with ξ = 3 to improve GWAS resolution when mapping GWAS associations of large effect. Therefore, we chose to select candidate genes for phenotypes with p-values above a Benjamini-Hochberg false discovery rate (FDR) threshold of 0.1 using a local score approach with ξ = 3 (Bonhomme et al 2019). In the paper, we now mention the gene or genes closest to the maximum Lindley score from the local score approach. In cases where a large genomic region has significant Lindley scores, we include all candidate genes in the region with significant Lindley scores.*   
  
Line 320-324: This text is repeated from the introduction.

*Edited.*  
  
Line 364-386: This first section of the Results has no results, but a mixture introductory and discussion points. These points should be covered elsewhere if needed, but not in Results.

*Edited. We chose to keep these points together and moved them to the Methods under “Germplasm: CDBN Breeding Strategies, Diversity Panel and Single Nucleotide Polymorphism Dataset” (starting line 229).*  
  
Line 398: "used to determine the narrow-sense heritability, h2". The Material and Methods don't describe the method. Since the previous sentence mentions kinship, maybe a SNP-based estimate of h2 was used? Provide details.

*Details are added, see Lines 303-305*.  
  
Line 825: "h2 is narrow sense heritability, defined as Vg / (Vg + Ve)". This is the definition of broad-sense heritability, not narrow-sense heritability (which would have Va, the additive heritability in the numerator).

*Edited.*