Editor (Comments to the Author):

The manuscript has been reviewed by three experts in the field. All referees raised substantial questions and provided many comments. The authors may need to review more QTL mapping papers in MAGIC populations to see how other people handle multiple parents. I believe that all QTL mapping studies for multiple parent populations use parent specific effects (the IBD approach). The way the authors presented their work appears that the IBD approach is their unique invention. As a statistical method paper, the presentation of the work is some how ad hoc. The authors need to provide in more detail the statistical models so that readers can understand what your are doing. Also, simulation studies are required to demonstrate the advantages of the new method over existing methods, as suggested by one reviewer. You may decide not to do simulations if you can provide theoretical proof that the new method is more powerful or more precise than existing models. All comments from the reviewers should be carefully addressed if you decide to resubmit the manuscript back to the journal.

Dear editor, thanks for your comments and suggestions.

We followed your advice and compared the models used by different researchers in NAM and MAGIC populations. We came to the conclusion that we cannot share your opinion that all QTL mapping studies for multiple parent populations use parent specific effects. For instance, Bandillo et al. (2013) only use an IBS genotype matrix without modelling parent-specific effects in a Rice MAGIC population. The same applies to Mackay et al. (2014) who also did only include IBS data and have not attempted to model founder probabilities in a MAGIC wheat population. Furthermore there are some papers dealing with both methods and comparing them (Sannemann et al., 2015 and Milner et al., 2015). They observed only minor differences between both methods, however they favour the haplotype approaches. Methods that predominantly model parent-specific effects are mainly used in the US NAM population of maize (Buckler et al., 2009 and further papers from this group). Our own group, however, only used an IBS genotype matrix so far.

Therefore, you see that it is not uniform how to model genotypes in multi-parental populations. We hope that our manuscript can contribute to ignite more discussion about the topic.

We did not intend to sell the IBD method as our invention, we rather wanted to highlight that it is the first time that we use this matrix in our HEB-25 population. We rephrased respective parts in the manuscript to avoid readers getting this false impression.

Following your advice, we conducted simulation studies and added their findings to the manuscript to discuss our methods in more detail and show their limitations and advantages.

Referee #1 (Remarks to the Author):

This paper presents results of analysis of multiparent mapping populations in barley using three different association models. The proposed IBD approach has some appealing features, notably its simplicity, but I don't think it can be defended as biologically realistic. A little work is needed to make it a realistic model. The results are surprising in the sense that the IBD method produces generally reasonable results, but I can't tell how much that is due to the specific nature of the barley introgression populations used (where the allele effects are likely to be consistent across different wild donor accessions) vs. the general utility of the IBD model.

You are right; from a biological point of view the model is very unrealistic. What 'IBD-M' represents is more or less a classical linkage mapping approach, conducted across the whole NAM population. We do not propose this model as our favourite; we just included it as comparison and as a step on the way from 'IBS-M' to 'IBD-MxF'. However, we understand that this might be confusing for readers and also other reviewers criticised this model. Therefore, we decided to remove model 'IBD-M' from the paper.

I think the following changes to the analysis are necessary:

1. Imputation of missing data. I assume the rate of missing data is low from the SNP array, but it would be good to say exactly what it is. I can't understand why the authors would choose the mean value imputation method (developed for unordered markers where essentially no information is available to impute), when they have a nearly perfect situation of known pedigree and a good linkage map. The missing data should be imputed using the known parental genotypes and the flanking marker information.

We added information about the amount of missing data points (only 0.6% of the SNP data) in the manuscript. We think in agreement with other studies for this low percentage the mean imputation approach would be sufficient. We nevertheless re-imputed the matrix based on flanking markers and re-ran all analyses for this matrix.

2. The IBD calling method. What was done seems like an ad hoc approximation done mainly for ease of computation. But it will clearly do poorly in progenies and regions where recombination occurred. I can envision cases where recombinations occur inside the 5 cM interval around a non-informative SNP, but the method proposed will 'fill in' the non-informative SNPs as donor parent genotypes, and the recombination breakpoint will incorrectly appear outside the imputed interval, between it and the next informative marker indicating a different IBD status. It could also do some strange things in intervals where the recombination breakpoint can be observed based on informative markers, but the non-informative markers will have some intermediate score (non-integer), which might suggest a recombination (which is real), followed closely by another recombination (as you move to a non-informative marker with an intermediate score) and yet a third recombination (as you move back to an informative marker). There seems no good reason to use this method when you have a good linkage map, instead you should just take each family one at a time and 'impute' the non-informative markers within the family using the linkage information. Then put all the families back together in one data set. This should not be hard. This also impacts the MxF method.

As already mentioned for the IBS matrix we also added information about missing data points in the IBD matrix (44.9%) in the manuscript. The reason to use the 5cM interval was to reduce the risk of assigning wrong genotypes if the marker order is wrong. But we agree

with you that in case of recombination our method cannot reliably determine recombination breakpoints. We therefore re-imputed the IBD matrix based on your suggestion to use flanking markers and re-ran all analyses for this matrix.

I am also skeptical that this model will perform well more generally. The authors discuss this point, which is appreciated, indicating that in this particular population, the allele effects from different wild donors will tend to have the same direction. But this is not likely to be generally true where multiparent designs sample diversity from within the domesticated species. I think this model is in general not a good idea.

As mentioned above, we decided to remove model 'IBD-M' from the paper.

3. The method of cumulating biallelic SNP effects into QTL cluster effects. I like this general idea, and this is the real heart of the paper, so it would be good to make this a robust method. Unfortunately, again there is a very ad hoc flavor to how it's done. The authors choose a more or less arbitrary interval length and then just sum up the allele effects in the interval. It seems like this could be improved with a simple adjustment. Something like summing up the products of (1-2r)a for each SNP in the interval where r is the recombination frequency between the SNP and the interval center, an a is the SNP's allele effect.

Thanks for your encouraging words and your suggestion. Regarding the ad hoc flavour that you mentioned we can state that originally we checked different interval lengths and 26 cM reflected the IBD-MxF effects best. We substantiated this finding to the mean introgression size (26.3 cM), which alsmost perfectly corresponds to our interval length. However, we understand your concerns and altered the interval length from 2-40 cM to check its impact on the precision of the cumulation method as part of our simulation study (where we can compare to the truly modelled effect). We observed that cumulation precision was constantly increased up to > 0.90 at 22 cM (see Supplementary Figure S5), where a plateau was reached and no further significant improvement could be observed. Furthermore we used the mean difference of cumulated and truly simulated effects as another measure to quantify the precision of the method. The mean difference was minimised at 26 cM. Therefore we chose to keep this interval (also the cumulation precision is slightly increased compared to 22 cM). In a further step we also tried the method you suggested, including recombination frequency to fine-tune the cumulated effect. For this purpose we defined r as the distance in cM between the SNP and the interval center, divided by 100. However, we observed a slight decrease in cumulation precision rather than any improvement, independent of the simulated scenario. A possible reason might be that the model selection process already outputs the best SNP effect estimates combination, since all SNPs are modelled at the same time. Therefore we assume that indirectly the effect estimates might already be adjusted for their recombination rate.

4. The introduction and discussion seem overly enthusiastic about the utility of the cumulation method and also too negative about the MxF method. The correlations between the cumulated SNP effects near the QTL peaks and the MxF estimates (or the haplotype effect estimates in the MAGIC population) are not that high. We don't know which model is more correct, but they are certainly not really in accordance with each other. Also the MxF model actually has the best prediction ability on average across traits, although the reader would never get this idea from reading the text. The fact that the IBS-M method picks out more significant markers but has mostly worse prediction accuracy is probably a good indication that the false positive rate is too high. The authors briefly mention the

overfitting, but their overall conclusion doesn't seem to worry about it much. If the authors are correct that grain color is simply inherited, the large number of significant IBS-M SNP tests is a clear indication that something is wrong.

We followed your suggestion and adjusted respective parts in the discussion, especially by hinting to the high rate of false positives obtained by IBS-M. However, we still favour IBS-M due to several aspects. The main aspect might be that the applied model is as simple as possible and nevertheless has a high correlation to the real effects, as exemplified in our simulation study. It can be applied without the need of further information about the donors and their relationship. The high rate of false positives can be counterbalanced, for instance by applying cross-validations. By doing so you get the detection rate as a clear indicator of robust SNPs that are detected with high reproducibility (Figure 2 and Suppl. Fig. S4). Based on cumulation of SNP effects of those regions we obtain the high cumulation precision in the simulation studies (Suppl. Fig. S5), although the prediction ability based on all SNPs (including false positives) is low (Table 3).

One other detail that should be clarified: the authors state that the prediction accuracy is based on predictions from marker effects. But there is a complication in the multiparent designs, that is that the population main effect is also in the model. Was that effect ignored in the prediction? It will tend to inflate how good all the methods are if it is included. Another option is to estimate the accuracy within families and average or estimate the accuracy of within-family deviations. Or just clarify what was done and explain in disscusion what the effect of including or excluding the family mean effect is.

Indeed, we included the population main effect also in the prediction. Based on your suggestion and the other reviewers' comments we decided to exclude the population main effect from all models to avoid inflating the predictions and to conform it to the model for the MAGIC population containing no population main effect.

Finally, I am confused about the statement that 'modelling only interaction effects causes a loss in degrees of freedom...'. The number of df in the numerator of the F-test increases for the MxF model compared to a two-state model like IBS. So, I don't know what the authors mean. If you are talking about the df in the denominator, usually the loss of a few df out of maybe thousands in these large designs is minimal.

Thank you for your comment. You are absolutely right. We talked about the df in the denominator and made an error in reasoning about its impact. Therefore we removed this thought.

Referee #2 (Remarks to the Author):

This manuscript explores various ways to consider genotypic data for analyzing the parental contribution effects in NAM populations. Their main aims were to derive IBD calls from IBS data, test whether IBD calls were superior to IBS, develop a novel approach to modeling parent-specific QTL effects without IBD data or modeling haplotypes, and use the approach in a MAGIC population.

As the authors indicated there are various ways to define identify by descent, but I don't agree with its use here. The IBD matrix described effectively eliminates the haplotype differences among parents in the HEB population. Then, when they model an interaction factor, any shared haplotypes are obscured as well. This seems to be the opposite of a haplotype analysis in which the true (or likely) identity by descent is estimated. The interest in this technique would be greater if there was a haplotype analysis for QTL mapping in the HEB population. The authors do not sufficiently explain the novelty or benefit of using this technique over a haplotype technique. While there may be error or variability in determining shared haplotypes among the HEB founders, it seems like that may be more informative than removing the haplotype variability among them.

We agree with you that our way of interpreting IBD obscures shared haplotypes. There are many ways to define IBD and the question is always how far should we trace it back? We follow a very traditional and clear approach by referring IBD to the parents of our population. This way we can easily distinguish cultivated (i.e. Barke) and wild alleles and estimate their differences (to clarify our way of interpreting IBD in a NAM population we changed respective parts in the introduction). We think that we are nevertheless able to distinguish also different sub-groups within the wild alleles by comparing their MxF effects. Interestingly, the proposed cumulation method also considers allelic SNP combinations within a specific genomic region (similar to what is done when modelling haplotypes) to obtain family-specific QTL effects.

Regarding the novelty and benefits of this technique, we highlight its simplicity, low computational load combined with the fact that the definition of haplotypes follows a more functional approach, based on their estimated effects rather than genetic similarity. This way, for instance, it has been possible to distinguish haplotype effects for two parents of the MAGIC population, where R/mpMap failed.

M&M

The authors need to include a description of the experimental design they used for collecting the phenotype data and (maybe I did not see it) the location (e.g., database, website) for accessing the data.

We briefly explain the experimental set-up now and hint on detailed information in Maurer et al. 2016. All data will be made available in Dryad. Unfortunately, Dryad only offers the free upload of raw data and additional files after acceptance in Heredity. The link to access the data will be included in the final version.

P11L189 Mean imputation results in poorer performance than other methods. Imputation from flanking marker and parental calls would be more appropriate. If these mean imputed markers are included in the calculations of the IBD matrix, this would amplify the error in the IBD calculations surrounding any missing data.

We followed your suggestion and used flanking marker imputation instead of mean imputation in the revised version.

P16L277 A 26cM interval seems too large to cumulate marker effects for the purpose of estimating QTL effects. The mean introgression size does not seem like an appropriate measure. The size of the population should have allowed for a higher mapping resolution, and therefore a smaller interval over which to cumulate effects. Presumably LD blocks would be smaller than the mean introgression size. Sannemann et al 2015 used a 5cM interval to calculate effect sizes. 20cM interval seems excessively large, and would probably be cumulating across multiple QTL.

In our simulation studies we included also one part concerning the optimum interval size to be chosen for cumulation of effects. We altered the interval from 2-40 cM to see differences in the cumulation precision by correlating with the truly simulated effect. At approximately 22 cM in the optimum case of no noise we observed the highest precision (see Figure S5). Furthermore we used the mean difference of cumulated and truly simulated effects as another measure to quantify the precision of the method. The mean difference was minimised at 26 cM. Therefore we chose to keep this interval (also the cumulation precision is slightly increased compared to 22 cM).

Regarding the interval size in the MAGIC population we based our decision to use 20 cM on the conservative approach of interpreting LD by considering the LD of unlinked loci in the whole population (Breseghello and Sorrells, 2006). Based on this LD fell below the population-specific threshold of 0.021 at approximately 20 cM.

Besides on LD we believe that the optimum interval size depends also on the diversity of parents used in the population (the more diverse the smaller the interval) and on the number and density of SNP markers across the genome (the more markers available the smaller the interval).

More discussion about using family main effect in GWAS models for HEB-25, why not kinship? Why not IBS + MxF? How much different is the model without a family main effect?

We followed the suggestions of Würschum et al. (2012) to select Model-B (containing a population main effect) for joint linkage association mapping studies. In their publication the authors hint on the benefits of this model and why it is superior over models containing kinship and population structure factors. We made the same observations in our NAM population, so we decided to base our previous publications on Model-B. For the reason of continuity we also based this manuscript on Model-B. However, we agree with you and another reviewer to remove the population main effect for the sake of comparability with the model for the MAGIC population. By the way, the power of Model-A (which is the one without population main effect) was comparable to Model-B in Würschum et al. 2012.

We also ran the IBS + MxF model. However, this model was very memory-demanding and had serious problems with overfitting. Another reason not to include it in this manuscript was that it is hard to interpret if you have SNPs segregating in only a subset of families and

in addition also family-specific effect estimates within those differentially segregating SNPs.

Why do marker selection instead of including all markers for predictive ability?

We wanted to keep the models as simple and computationally efficient as possible. Therefore, we did not want to touch the wide field of genomic prediction. However, we note that in Maurer et al. (2015) we already applied whole genome prediction models (RR-BLULP and Bayes-C-Pi) and highlighted their potential. Maybe in future a promising approach could combine whole genome prediction and the proposed cumulation of adjacent marker effects.

It seems like it might be problematic to compare effectiveness of two different techniques, when one (NAM) includes a family effect and the other (MAGIC) does not.

As mentioned before, we removed the population main effect for the sake of comparability with the model for the MAGIC population.

Results

It is unclear why the correlation of the average cumulated QTL effects and the average IBD MxF effect estimate for each QTL is a relevant measure of appropriateness of the method.

We chose this comparison since we assumed that the IBD-MxF effect is the most realistic estimate that we have. However, actually we don't know the truth and the IBD-MxF effects or haplotype effects might be overestimated due to smaller sub-groups. So we share your concerns about using this as a parameter to evaluate the appropriateness of the method. Therefore, in our revised version of the manuscript we investigated cumulation precision in the simulation studies, where we know the truly simulated parent-specific QTL effects. This way we observed high accordanceof cumulated and true effects in an optimistic simulation without any background noise ($r > 0.9$) and even in a pessimistic approach with high background noise ($r > 0.5$), see Suppl. Figure 5. In addition to cumulation precision, which is based on correlation, we added the mean difference of cumulated and truly simulated effect as another parameter to evaluate the appropriateness of the method in the simulation studies.

Discussion

The discussion of QTL detection focuses on the fact that fewer markers were detected in IBD-M model than IBS-M. The explanation given is the elimination of rare alleles, but the IBD model also may be eliminating the variability in alleles with moderate frequency, which also are likely to be affected when "their effect is mixed with those of all other families." Please explain more about what is outlined by Ogut et al (2015) regarding small effect allele detection improving with IBD modeling.

Thanks for your suggested explanation. It is indeed another good fact to consider. However, as we decided to completely remove model IBD-M from the paper we also removed this part of the manuscript. Since the work of Ogut *et al.* is also relevant for

interpretation of IBD-MxF we added one sentence explaining more about the work of Ogut *et al*. in the respective section.

The discussion of prediction using IBS vs IBD "the estimation of QTL effects is therefore derived from a more solid base and better applicable to the prediction set." It is unclear what you mean by "more solid base." I would interpret this finding to be due to the presumption that there are large effect differences between cultivated and wild barley, which can be captured well by a relatively small number of markers. Therefore, the IBD models are simpler and therefore better predictors, they explain the larger wild vs cultivated variation. Because the IBS markers include variation among the wild barley parents, these markers are likely to be explaining smaller effect alleles with more error.

What we mean with "more solid base" is that the training set contains a higher amount of lines carrying the respective wild allele, therefore effect estimation is more reliable due to higher number of observations. Furthermore, the chance of having the allele also in the test set is higher if alleles are segregating in the whole population (as in our IBD matrix). However, as we decided to remove model 'IBD-M' from the entire paper also the section you mentioned was removed.

Why wasn't a haplotype based method compared with the other QTL mapping methods in the HEB-25 population?

The aim of the study was not to conduct a comparison of different models in multi-parental populations. Our intention was rather to check whether an IBD-based method, modelling family-specific QTL effects, was able to outperform our existing IBS-based method from former publications. In the course of the work we realized that there is no clear advantage for the IBD-based method. We attributed this to the fact that the simple IBS-M model indirectly already models family-specific effects, expressed as several tightly linked SNPs with different effect estimates at the same time.

Table 2: Why are the CVs for the MAGIC population so low as compared to the NAM population?

The CVs were so low because they had been calculated for the parental allelic means given in Table 3 in Sannemann *et al.,* 2015, whereas the CVs for the NAM population were calculated for the QTL effects. For reasons of uniformity we changed the calculation of the CVs for the MAGIC population by considering the allelic effects instead of the allelic means of each QTL. Thank you for this hint.

Figure S3: Are these figures from a specific family, or averaged across HEB families?

These figures present the results of model IBS-M, which does not distinguish families. So, the values presented are the averages of marker effects across 100 cross-validations. We changed title and description of the figure to avoid misinterpretations.

Minor

P3L35-37: Unclear phrasing, “there is scope left whether…”

We rephrased this passage.

P3L45: More information about how the IBS vs IBD “differed” in abstract.

We added more information about the performance regarding number of detected QTL and prediction ability.

Were twin markers reevaluated during the subsampling?

No, twin markers were completely removed from the dataset at an early stage and have not been considered in any analysis.

P22L397 “wheat from the chaff” unclear use of analogy

We rephrased this section.

Referee #3 (Remarks to the Author):

The manuscript describes several methods for mapping QTL based on modeling the underlying genotypes as IBD versus IBS. The authors describe the methods and implement them on a barley NAM population and a barley MAGIC population. The authors also compare their method to existing methods.

On the positive side, the manuscript is clearly written, the authors map several different traits and compare the number of QTL discovered by each method. However, if the purpose of the manuscript is to argue that the proposed methods are an improvement upon existing methods, then simulations and a comparison with other methods would be of more use in evaluating the utility of these methods. The authors could use the existing NAM and MAGIC genotype data to simulate QTL with known allele effects and locations and summarize the results. The authors could then present estimates of power and Type I error rate, which is lacking in the current manuscript. Including the experimental data in the current manuscript could be an important final section in such a manuscript. Perhaps there is a way to use the cross-validation data to estimate the power to detect QTL with varying effect sizes and estimate the false positive rate.

The authors' observation that the IBD-MxF model may be detecting a set of closely linked SNPs with differing allele effects is interesting. The estimation of parent allele effects in the Pflugs Intensiv and Criewener 403 lines is also interesting since the R/mpMap failed here. This is an important are and the authors have made a significant contribution to the field.

Thanks for your comments and suggestions! We added simulations to the revised version of the manuscript and base our discussion on the results obtained therein. We added information about QTL detection power and false positive rate of the models and discuss their impact on the usability of the models. Our purpose was not to compare existing methods to figure out which performs best. Including it would have gone beyond the scope of the paper. Therefore we rather focus on our models and the emergence of the idea of cumulating parent-specific QTL effects. It would be interesting to see in upcoming investigations how it performs compared to several other published methods.