Response_to_Reviews

Reviewer comment

Mash: A lot of the results interpretation is based on interpreting the SNP loadings on the specified covariance matrices. This is a secondary use of mash (the primary being the refinement of effect sizes), and while they do interpret these somewhat in the mash paper, these loadings need to be interpreted with caution. If these covariance matrices are similar, mash will somewhat arbitrarily assign weight to any of the matrices because all lead to the same fit to the effect-size data. The mash algorithm doesn't give the full posterior distribution on the loadings so you can't check for posterior correlations there. It seems likely to me that the hypothesis-driven and data-driven covariance matrices are somewhat correlated here, and the correlations may differ between green-up and flowering because of the better correlation between the environmental metrics and flowering than green-up. In the mash paper, they used cross-validation with the likelihood in the testing set as the evaluation metric to compare models. It might be safer to try dropping specific covariance matrices and comparing the model performance in a held-out testing set to evaluate the importance of the hypothesis-derived covariance matrices.

Response

We thank the reviewer for this comment and agree that the selection of covariance matrices to be included in a mash model is a nontrivial question. To clarify our interpretation of the posterior matrix weights provided by mash, we performed an extensive analysis of the performance of mash models when different covariance matrices are included. Specifically, we implemented a model selection approach that uses a greedy algorithm to evaluate the log likelihood of the mash model as additional covariance matrices were included (see Methods lines XXX-YYY and pseudo code below). This is a similar, but more computationally efficient design, than the leave-one-out approach recommended by the reviewer as the runtime of mash increases with the number of covariance matrices included. Additionally, while cross-validation is a powerful approach to model evaluation our limited sample size was prohibitive to the necessary partitioning of individuals included in our analysis.

In practice this greedy algorithm approach offers a fast way to identify the point where redundancy in the addition of a new covariance matrix results in no change to the likelihood of the mash model. Importantly, this removes the potential for arbitrary assignment of weights to the covariance matrices in the model and allows for the selection of matrices that most accurately capture the underlying biology captured by the site-specific effect sizes. We applied this algorithm to only the hypothesis matrices for each phenotype (Figures 1 and 2) using the previously used sets of significantly associated and randomly selected variants in the Gulf subpopulation individuals, Midwest subpopulation individuals, and the two subpopulations combined. The results indicate that only a small number of the original hypothesis matrices (between three and five) are necessary to reach the maximum likelihood model when using the same likelihood ratio test implemented in Urbut et al., indicating that correlation among hypothesis covariance matrices is high. This redundancy was confirmed through our application of the greedy algorithm to a combined set of canonical, data driven, and hypothesis covariance matrices (Figures 3 and 4)

Our primary conclusion from this analysis is in line with what the reviewer posited, that there is extensive redundancy among canonical, data driven, and hypothesis covariance matrices that is leading to biologically informative patterns of covariance between site effect sizes to be arbitrarily distributed among them. We additionally found evidence that while the correlations between data driven and environmentally informed (hypothesis) covariance matrices, there is extensive similarity in the matrices that are included in the maximum likelihood models across both phenotypes and in the analysis of the Gulf subpopulation, Midwest subpopulations, and the combined cohort.

Greedy algorithm for mash model selection.

---

matrices = [all covariance matrices of interest]

most_likely_matrices = [] #a list to store the most likely additional matrix at each iteration

maximum_likelihoods = []#a list to store the maximum likelihood at each iteration

likelihood = -inf

most_likely_matrix = ''

for matrix in matrices:

```
matrix_likelihood = mash(effects, std.errs, matrix)\['likelihood'\] #obtain model likeliho

if matrix_likelihood \> likelihood:

    most_likely_matrix = matrix

    likelihood = matrix_likelihood
```

matrices.remove(most_likely_matrix)

most_likely_matrices.append(most_likely_matrix)

maximum_likelihoods.append(likelihood)

lrt_pvalue = 0

iteration = 2

while lrt_pvalue $<$ 0.05:

```
likelihood = -inf
```

most_likely_matrix = ''

```
for matrix in matrices:

    model_matrices = most_likely_matrices + matrix

    model_likelihood = mash(effects, std.errs, model_matrices)\['likelihood'\]

if matrix_likelihood \> likelihood:

    most_likely_matrix = matrix

    likelihood = matrix_likelihood
```

matrices.remove(most_likely_matrix)

most_likely_matrices.append(most_likely_matrix)

maximum_likelihoods.append(likelihood)

#perform a likelihood ratio test of the most likely model from the previous iteration to the model in this step with one degree of freedom

lrt_pvalue = likelihood_ratio_test(likelihood, maximum_likelihoods[iteration-1], df = 1)

Methods

In order to better understand the behavior of the mash shrinkage algorithm in the presence of correlated covariance matrices, we implemented a "greedy" mash algorithm. Specifically, given n covariance matrices for each phenotype and corresponding effect sizes and standard errors across the k contexts of interest, we fit a set of mash models with each of the n matrices separately. We then identify the model with the maximum log likelihood estimates and select the corresponding matrix, leaving n-1 covariance matrices. We then fit a new set of mash models where the selected covariance matrix is paired with each of the remaining n-1 covariance matrices. The most likely pair of covariance matrices are then selected and the process is repeated for all possible n-2 matrix triplets. This process is repeated until one of two stop conditions are met: (i) all covariance matrices have been added to the mash model, resulting in a final model with all of the original n covariance matrices are included or (ii) a likelihood ratio test with one degree of freedom between the likelihood of the current model and the model from the previous iteration has a p-value > 0.05. The result is a set of log likelihood estimates for each model and the stepwise 'path' of covariance matrices that best explain the observed effect sizes and standard errors observed for the phenotype of interest.

We then applied the greedy algorithm to two sets of covariance matrices. First, we applied it to only the hypothesized covariance matrices for each phenotype as described in Supplementary Table 1. Briefly, the hypothesis covariance matrices included in our analysis of greenup time included: average temperature in the 5 days, 10 days, and 18 days preceding greenup and the

number of cumulative growing degree days (over 12 degrees Celsius) in the 5 days, 10 days, and 18 days preceding greenup. Each of these six covariance matrices were calculated using individuals who were present at both sites. Additionally, each matrix was calculated using all individuals, only individuals from the Gulf subpopulation, and only individuals from the Midwest subpopulation.

For the flowering time phenotype, 12 hypothesized covariance matrices were included in the greedy algorithm. For each phenotype, we used the mash greedy algorithm to identify the most likely covariance matrix at each step using both a set of random independent variants and a set of independent variants that showed the strongest associations with the phenotype across all contexts. The resulting maximum likelihood paths for the selection of covariance matrices as determined by the greedy algorithm for each phenotype, subpopulation, and set of randomly selected or significantly associated variants are shown in Figures 1 and 2.

Our second analyses expanded the set of covariance matrices to all of the matrices included in our analyses of the empirical data. These include: the phenotype specific covariance matrices described above, the data-driven covariance matrices, and the suite of canonical covariance matrices representing context-specific effects, equal effects across contexts, and simple scenarios of heterogeneous effects across contexts. The resulting maximum likelihood paths for the selection of covariance matrices as determined by the greedy algorithm for each phenotype, subpopulation, and set of randomly selected or significantly associated variants are shown in Figures 3 and 4.

Results

We first applied the greedy algorithm to only the set of hypothesis matrices for each of the greenup and flowering phenotypes, respectively. Generally, we observed that only a small number of matrices significantly increased the likelihood of the model as the algorithm progressed. In our analysis of the set of variants that were significantly associated with greenup time in at least one context, 18 hypothesis covariance matrices were included in the initialization of the greedy algorithm. The stop condition, a likelihood ratio test p-value > 0.05 between the most likely model at step n and most likely model at step n+1, was met after only four iterations in both the combined cohort (Figure 1A) as well as the analysis of the Gulf subpopulation alone (Figure 1B). In our analysis of the midwest subpopulation the stop condition was met after only three iterations, see Figure 1C. The relatively small number of necessary iterations, and correspondingly small number of covariance matrices, indicate that the hypothesis covariance matrices were highly correlated with one another and that inclusion

of all the hypothesis covariance matrices introduced redundancy in the model. In the set of randomly selected independent variants, the stop condition was met in the combined cohort after five steps (Figure 1D). The stop condition was met after four and three iterations in the Gulf and Midwest subpopulations, respectively. The overlap between matrices included in the maximum likelihood model prior to reaching the stop condition in the analysis of significantly associated and randomly selected variants was high in the combined cohort; three of four matrices in the analysis of significantly associated variants were also included in the model for the randomly selected variants. Similarly, three of four matrices in the Gulf subpopulation and two of three matrices in the Midwest subpopulation that were included in the model for significantly associated variants were also included in the model for the randomly selected variants.

We then applied the greedy algorithm to the 12 hypothesis covariance matrices for the flowering phenotype. Similar to our analysis of the hypothesis covariance matrices for the greenup phenotype, we observed a similar pattern of correlation in hypothesis covariance matrices for the flowering phenotype. The stop condition for the greedy algorithm was met after five steps when analyzing the combined cohort, the Gulf subpopulation, and the Midwest subpopulation in analysis of variants that were significantly associated in at least one context (Figure 2A-C). When analyzing the set of randomly selected variants the stop condition was met after four steps in the combined cohort, five steps in the Gulf subpopulation, and six steps in the Midwest subpopulation. We also observed that many of the matrices included in the model for the significantly associated variants were also included in the corresponding model for randomly selected variants (Figure 2D-F).

Analysis of the hypothesis covariance matrices alone helped to establish the utility of the greedy algorithm implementation in identifying the maximum likelihood mash model for the observed summary statistics from our analysis of eight sites. However, due to substantial evidence that both canonical and data driven matrices explained a nontrivial proportion of the observed covariance between context specific summary statistics we applied the greedy algorithm with the all three classes of covariance matrices included in the initial set. In our analysis of the greenup phenotype, the stop condition was met after 16 iterations, resulting in a maximum likelihood model that included all 13 canonical covariance matrices, two data driven matrices (total PCA and PC 4), and the hypothesis matrix representing the covariance in average temperature ten days prior to greenup in the combined cohort (Figure 3A). When the Gulf subpopulation was analyzed alone the stop condition was met after five iterations. The five corresponding matrices were composed of four canonical matrices and the hypothesis matrix calculated as the covariance in average temperature five days prior to greenup in only individuals from the Gulf subpopulation (Figure 3B). Finally, the greedy algorithm met the stop condition after 13 iterations when applied to the Midwest subpopulation. Ten of the