***Genome-wide association methodology***

During the previous grant funding period, we gathered datasets for genome-wide association (GWAS) that were very large – dozens of phenotypes collected at ten sites for hundreds of individuals genotyped at tens of millions of SNPs. Commonly used plant genome-wide association software did not perform well with datasets of this size. GAPIT cannot fit the switchgrass SNP file into working memory, while TASSEL takes ~30 minutes of computational time per million SNPs to run a GWAS for a single phenotype. Due to the large sizes of our common garden datasets, we developed a new pipeline, the switchgrassGWAS R package (hereafter {switchgrassGWAS}, to 1) allow fast, less memory intensive GWAS on the *Panicum virgatum* diversity panel, and 2) to analyze the extent to which SNP effects are similar or different for phenotypes measured at different sites.

The goal of {switchgrassGWAS} is to provide clearly documented software resources for researchers using the *Panicum virgatum* diversity panel for GWAS, analyses of pleiotropy, and analyses of GxE. {switchgrassGWAS} depends on the bigsnpr R package (hereafter {bigsnpr}) to run GWAS on single phenotypes at single sites, and depends on the mashr R package (hereafter {mashr}) to combine estimates of SNP effects across sites. {switchgrassGWAS} is a wrapper around these packages that is specific to the *Panicum virgatum* diversity panel. We also included functionalities to find genomic annotations from V5.1 of the *P. virgatum* genome for specific genomic regions and to produce common plots and datasets for GWAS and mash analysis.

{bigsnpr} performs fast statistical analysis of massive SNP arrays encoded as matrices. The package can handle matrices that are too large to fit in memory by memory-mapping these matrices to binary files on disk. {bigsnpr} takes less than 5 seconds of computational time per million SNPs to run a GWAS for a single phenotype; a performance improvement of more than 300x over TASSEL. It also incorporates current gold standards in the human genetics literature for SNP quality control, pruning, and imputation, as well as population structure correction in GWAS.

{mashr} uses Empirical Bayes methods to estimate and test many effects in many conditions. It is a flexible, data-driven method that shares information on patterns of effect size and sign in any dataset where effects can be estimated on a condition-by-condition basis for many conditions (here, phenotypes at different common garden sites) across many units (here, SNPs). It first learns patterns of covariance between SNPs and phenotypes from SNPs without strong effects, then combines these data-driven covariances with the original condition-by-condition results to produce improved effect estimates. In this way, {mashr} shares information between conditions to increase the power to detect shared patterns of effects. We used {mashr} to compare both the magnitude and sign of significant phenotypic effects, using only SNPs with a local false sign rate of 0.05 or less for one or more phenotype. The local false sign rate is analogous to a FDR, but is more conservative, in that it also reflects the uncertainty in the estimation of the sign of the effect (Stephens 2017).

Privé, Florian, et al. [“Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr.”](https://doi.org/10.1093/bioinformatics/bty185) Bioinformatics 34.16 (2018): 2781-2787.

Sarah Urbut, Gao Wang, Peter Carbonetto and Matthew Stephens (2019). [Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions.](https://doi.org/10.1038/s41588-018-0268-8) Nature Genetics **51**, 187-195.

Stephens, M., 2017 False discovery rates: a new deal. Biostatistics (Oxford, England) 18**:** 275-294.