MCB585

Calvin Perkins

Final Report

11/30/23

**Performing a genome-wide association study (GWAS) to validate a lab dataset.**

My post-doc mentor, Dr. Cedar Warman, collected phenotype data from a tomato diversity panel of 220 accessions and four different species of tomato. We hope to identify genes correlated with reproductive heat tolerance traits by using the diversity panel dataset in a GWAS. Before exploring new traits, the dataset must be validated with a positive control. To validate our dataset, I consulted a paper in which a similar diversity panel dataset was collected (Mata-Nicolás 2020). Mata-Nicolás et al. performed an SNP-based GWAS on numerous traits measured from their own tomato diversity panel. One of the traits they collected was tomato locule number. Tomato locules are chambers within a tomato fruit that develop to hold the seeds. I chose to validate our dataset with the tomato locule number trait as it is a reproductive trait, similar to the trait we are ultimately interested in researching. To make sure our data was usable for GWAS, the significant loci identified in my GWAS studies should overlap with the GWAS output from Mata-Nicolás 2020.

I used a k-mers based GWAS approach to identify genomic sequences of k-length that correlated with the locule number phenotype. To run a k-mers GWAS, I used the kGWASflow pipeline established in Corut & Wallace 2023. kGWASflow combines multiple different processes using a Snakemake workflow organizing structure. I implemented the kGWASflow pipeline on to the University of Arizona HPC so that it could manage the computationally intensive processes. To test whether my HPC setup was successful, I ran the pipeline on a test dataset provided by the workflow. Once I successfully reproduced the results, I was confident that it would work on my own dataset. kGWASflow takes user created .tsv and .pheno input files which contain sample, phenotype, FASTQ, and path information. I assembled the required input files within Rstudio, and pulled the FASTQ files containing paired-end sequencing reads on to the HPC using the nf-core pipeline: fetchngs. Once the inputs had been created, I modified the kGWASflow provided configuration file to work within my working directory on the HPC. I initiated the kGWASflow workflow using a slurm job that allocated 64 cpus per task and 400GB memory. Once initiated, kGWASflow moved the provided FASTQ files through a preprocessing step which performs quality control and generates a multiQC report used as an input for k-mers GWAS. kGWASflow uses an adapted version of k-mers GWAS designed in Voichek et al. 2020 which counts and filters k-mers. These k-mers are used to create a table which assigns k-mers to different samples. The k-mers table is then used to generate a kinship matrix which determines relatedness between samples. A permutation based 5% family-wise error-rate threshold is determined with the matrix, and the k-mers surpassing this threshold are considered statistically significant (Corut 2023). The information generated by k-mers GWAS is then compiled into a results output .bam file.

The output of my kGWASflow run on locule number had a p-value threshold of ~2.5e-10, creating a table of 3393 significant k-mers. I imported the kGWASflow output .bam file into Rstudio to analyze and plot the data. I generated a Manhattan plot using the ggplot() function to display the significant k-mers mapped to their chromosomal position on the *Solanum lycopersicum* reference genome (Figure 1). There were significant k-mer hits present on each chromosome, with a high enrichment at the end of chromosome 2. When cross referencing my k-mer hits with the significant SNPs identified in Mata-Nicolás 2020, there was overlap on chromosome 2 and 11 for the locule number phenotype. On chromosome 2, there was high overlap within the 45,000,000 to 48,000,000 base pair range. When mapping the significant k-mers to the *Solanum lycopersicum* reference genome with the Sol genomics network, a significant k-mer at position 47,629,125 on chromosome 2, associated to the coding sequence of a gene (Solyc02g087100). This gene encodes for the lipid metabolizing enzyme, alpha-dioxygenase (Figure 2). Another significant k-mer mapped to position 52,907,500 on chromosome 11 which encodes a SafE family sulfite exporter (Solyc11g071310). These results aligned with significant SNPs present in Mata-Nicolás 2020 and serve as a validated positive control for our dataset. With a validated dataset, we can now feel confident performing a GWAS on phenotypes that have not been previously analyzed.

I encountered many errors during this process. Some of the major errors were related to the specific packages loaded into the individual snakemake environments. For example, on each run, the multiqc step failed because the package was out of date. To fix this, I had to manually update the multiqc package in an interactive environment before resubmitting the slurm job. Another important consideration that came up during this process was the type of input genomes that could be used in the pipeline. kGWASflow is designed to only work with pair-end reads and cannot support both pair-end and long read FASTQ files. The pipeline also had issues generating plots for the *Solanum lycopersicum* inputs because it appears to be configured for E. coli. My next steps will be to run kGWASflow on the uncharacterized phenotype of pollen tube burst. After generating a table of significant k-mers pollen tube burst, I will use this to search for genes with enriched significant k-mers for follow-up characterization studies.

A graph of dna sequence

Description automatically generated with medium confidence

**Figure 1 - Manhattan plot for significant k-mers correlated to tomato locule number.** Plotting GWAS hits on each chromosome of the ITAG4.0 Solanum lycopersicum reference genome. Significance threshold was calculated within the kGWASflow pipeline. K-mers outside the significance threshold were excluded for clarity.

A screenshot of a computer

Description automatically generated

**Figure 2 - Sol Genomics JBrowse screenshot of alpha-dioxygenase gene position.** Significant k-mers identified from my kGWASflow analysis overlapped with SNPs identified in Mata-Nicolás et al. 2020 which mapped to the position of a gene encoding alpha-dioxygenase (Solyc02g087100.1.1).

**References**

<https://github.com/nf-core/fetchngs/tree/1.5>

**Corut, A.K. and Wallace, J.G. (2023).** kGWASflow: a modular, flexible, and reproducible Snakemake workflow for k-mers-based GWAS. G3: Genes, Genomes, Genetics: jkad246.

**Mata-Nicolás, E., Montero-Pau, J., Gimeno-Paez, E., Garcia-Carpintero, V., Ziarsolo, P., Menda, N., Mueller, L.A., Blanca, J., Cañizares, J., Van Der Knaap, E., and Díez, M.J. (2020).** Exploiting the diversity of tomato: the development of a phenotypically and genetically detailed germplasm collection. Hortic Res 7: 66.

**Voichek, Y. and Weigel, D. (2020).** Identifying genetic variants underlying phenotypic variation in plants without complete genomes. Nat Genet 52: 534–540.