MDSC 403: Computation for Bioinformatics

Final Project – FALL 2019

Rachel Wong

UCID: 30001531

**Introduction**

The goal of this project was to successfully analyze three given files, a genotype file containing each sample and its nucleotide base at a specific position, a phenotype file containing the sample and its associated phenotypic measure, and a gene model file containing the characteristics of each position on the chromosomes. This research goal will be achieved by determining if there is an association between genotype and phenotype, and what that association is.

**Hypothesis**

I hypothesize that there exists an association between genotype and phenotype, where the nucleotide bases from each sample at each position, correlates to that sample’s overall phenotypic measure. I can further hypothesize that significant positions can be found that are more significant in affecting its associated phenotypic measure. This would be the result of single nucleotide-polymorphisms (SNPs) where the significant positions that are affected, affect the overall phenotypic measure more. As well, this would be feasible since it follows that most of the genome is non-coding, and therefore there are significant positions that affect the phenotypic measure more than other positions.

Alternate hypothesis: There is an association between genotype and phenotype, and the

significant positions are found in coding regions.

Null hypothesis: There is no association between genotype and phenotype.

**Objectives**

1. Show the association between genotype and phenotype

2. Identify significant positions

3. Identify characteristics of the significant positions and conclude on the association

**Methods and Rationale for Pipeline**

Please see attached code for more details. The general pipeline for this analysis is as follows:

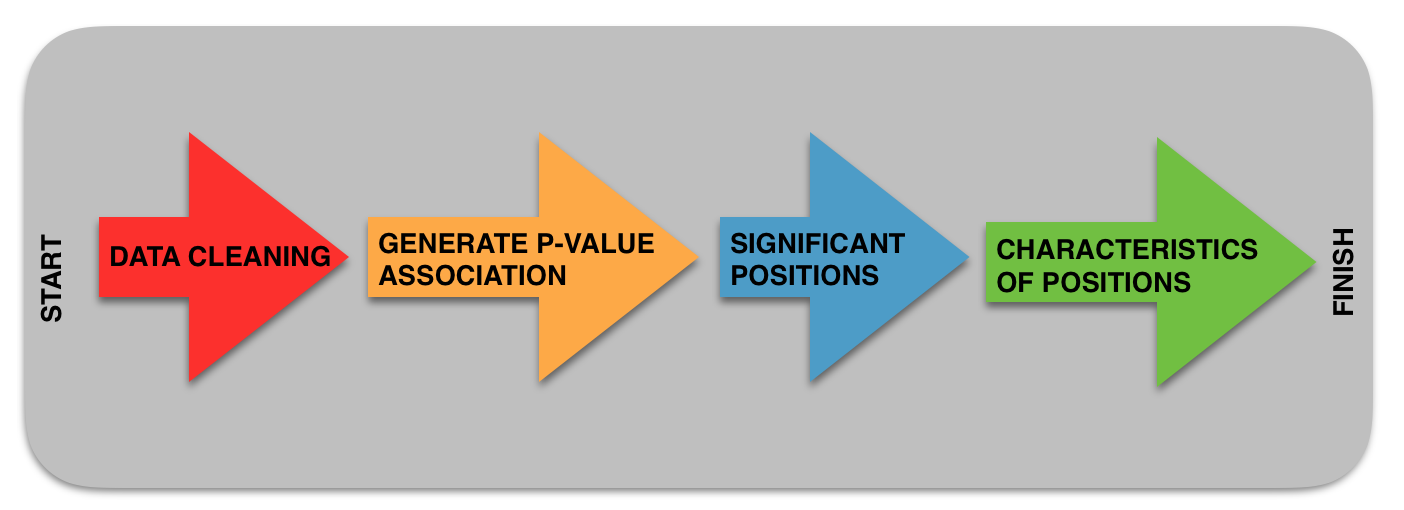


Figure 1: General illustration of the pipeline used from start of the analysis to finish.

This pipeline has 4 main steps as described in more detail with rationale below:

1. Data Cleaning

The first step in the pipeline is to first clean and organize the data. This step was done in three different parts. Part 1 ensured that the phenotype data matched the genotype data. This was done by adding an “X” before each sample ID in the phenotypes file since the sample IDs in the genotype file all began with an “X”. Part 2 was to remove any samples containing N/A for the phenotypic measure. This was to ensure that only samples with a phenotypic measure were analyzed, since samples without a phenotypic measure would be useless to this analysis of association between genotypes and phenotype. Part 3 was to sort the phenotype file in the same order as the genotype file so that the samples were in the same order. This was important for when we traversed through each file to analyze the samples and their associated information.

The main rationale for the data cleaning was to organize the data so that analysis was more efficient and effective.

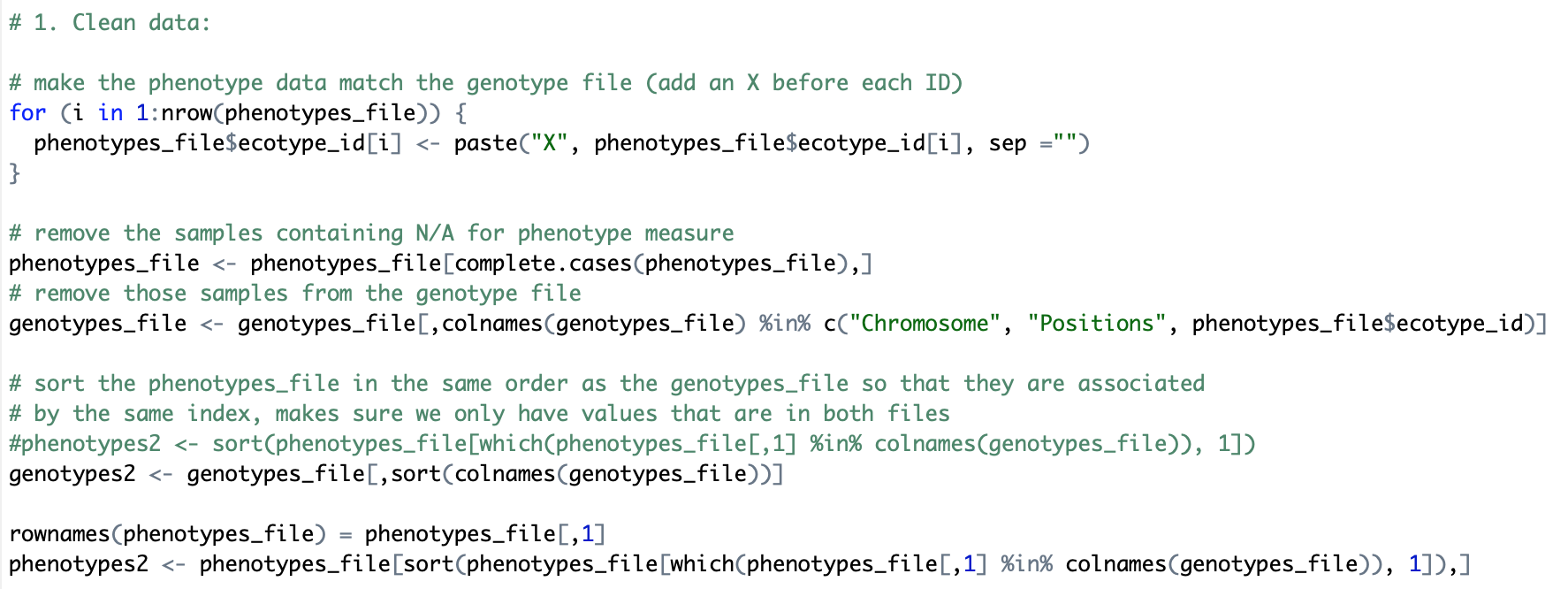
****

Figure 2: Code snippet of the data cleaning step.

2. Generate P-Value Association

This second step in the pipeline was to generate p-values from a linear regression of the prediction of the phenotypic measure, Y, to each X component, the nucleotide for each position for each sample ID. This step was extremely difficult and resulted in a program running for over 18 hours to generate p-values to be further used in analysis (step 3, Manhattan Plot). The final analysis for this step generated p-values by taking the transpose of the genotype file and running a linear regression for each position with the phenotypic measure for each sample.

The main rationale for this step was that it was necessary to analyze the prediction of each position to its phenotypic measure

Multiple failed attempts were done to try to generate these p-values including:

1. Try generating p-values of the difference between mutant and wild-type samples, but this did not work since we would have had to “create” our own wild-type.
2. Use for loops to compare the phenotypic measures with each genotypic file position for each sample, but this nested for loop did not work.
3. Try transposing the phenotype file to analyze it row by row but the for loops from above still did not work.
4. Tried transposing the phenotype file and then merging it with the genotype file so that we could use “.,” to compare all other variables form the phenotypic variable for the linear regression, but “.,” only works for fields and not rows.
5. Tried to transpose the genotype file instead of the phenotype file, or just transposing the merged file created in d, but Rstudio crashed every time I tried to do this.
6. Tried to split the analysis into subgroups to analyze but Rstudio still crashed.

The pipeline eventually worked by taking the transpose of the genotypes file and running a linear regression for each position with the phenotypic measure for each sample using a for loop to go position by position. However, this code took over 18 hours to run.

I did not know how to write the code for this step to get the final results, and my RStudio continued to crash with all of my attempts at working with the large genotypes file. Brett sent me his final p-values results which I used for the next steps in my pipeline (1).

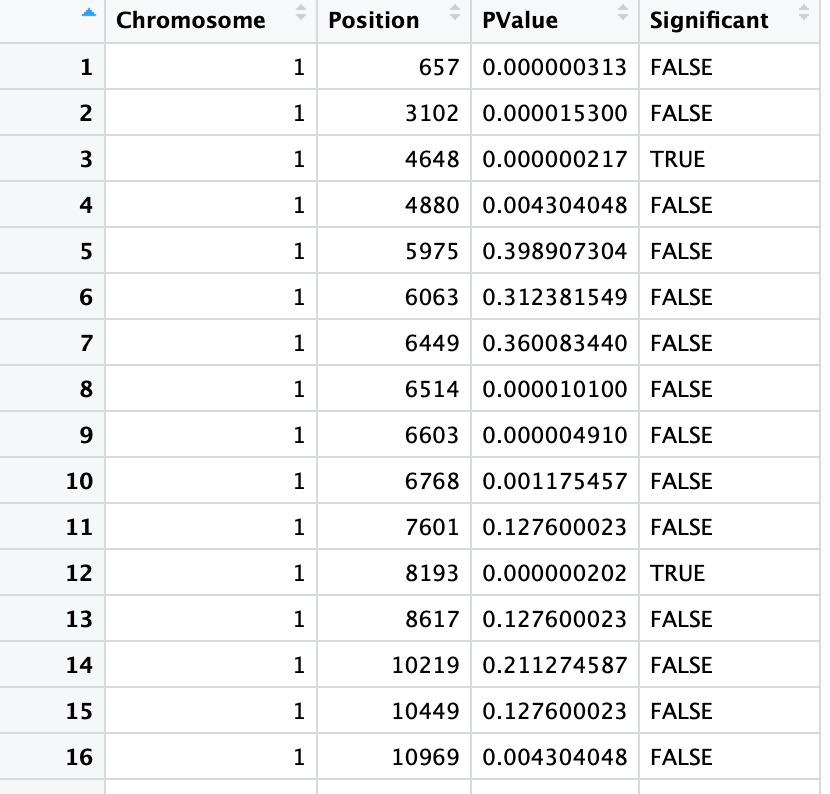
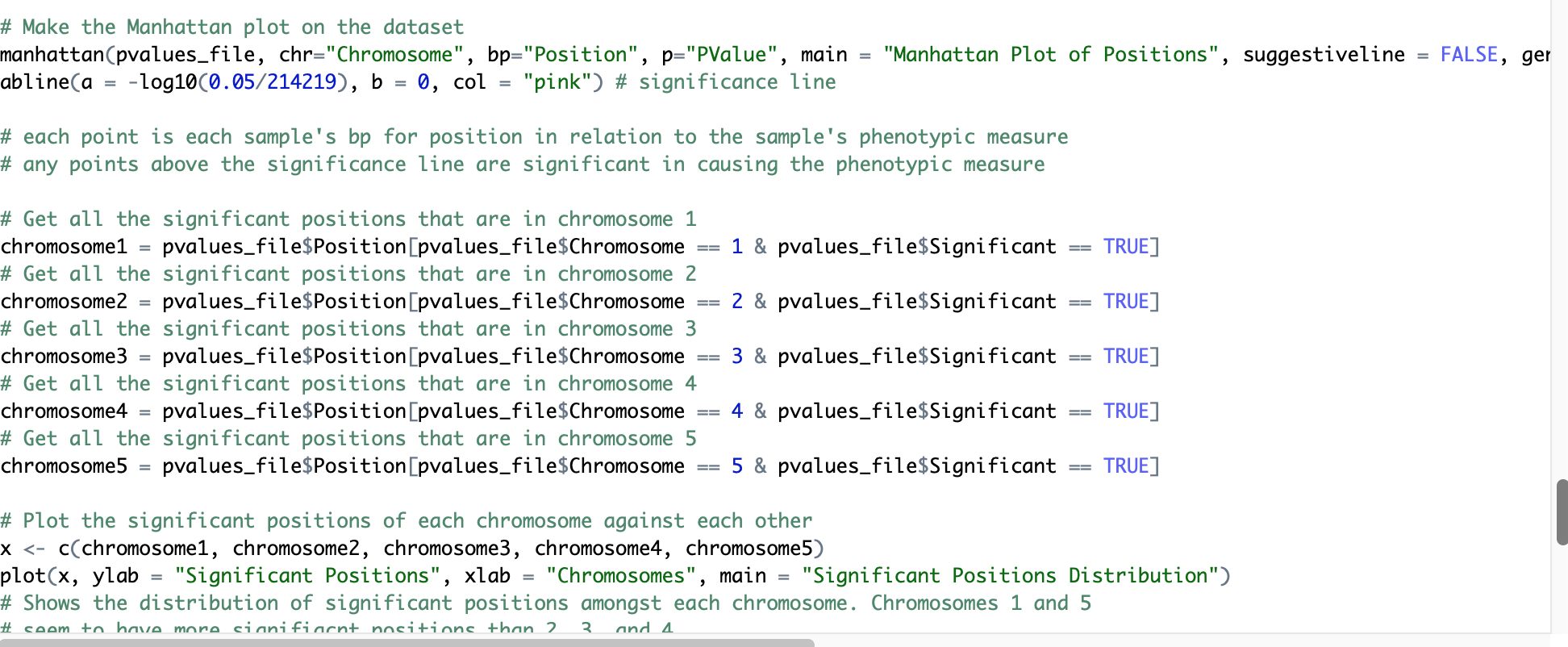


Figure 3: Final results containing p-values and the significance of the positions. Code for generating this is taken from Brett, and some of my failed attempts at generating this are in my attached code (1).

3. Significant Positions

The third step in my pipeline was to determine significant positions. This was done with an alpha value of 0.05 and plotting the p-values and positions on a Manhattan plot (using the library qqman). I also identified the chromosomes where the significant positions were found and plotted them to show the distribution of significant positions among the 5 chromosomes. A qqplot was also generated from the Manhattan plot information to compare the distribution of the p-value with an expected distribution of chance.

The main rationale for this step was to identify the significant positions, which are significant in resulting in the associated phenotypic measure. These significant positions are then used in downstream analysis to determine the association between genotype and phenotype, and why certain positions are more significant than others.



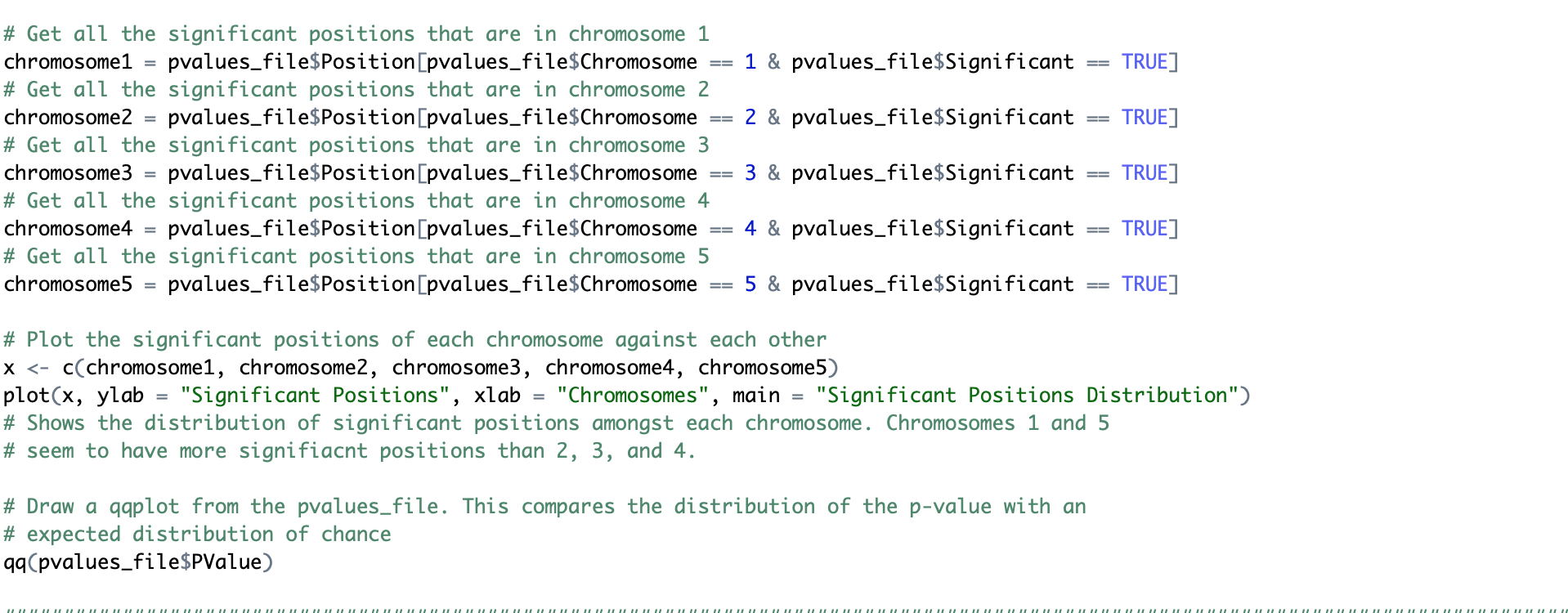


Figure 4: Code snippet showing the generation of the plots for step 3 and determination of the significant positions and their associated chromosome distribution.

4. Characteristics of Positions

The fourth step in my pipeline was to characterize the significant positions with the gene model file. The gene model file provides descriptions of each position, for example, if it is coding vs. noncoding.

The main rationale for this part is that this would be the step that may provide an explanation as to why the significant positions are significant. For example, if a significant position was in a coding region, the explanation may be that regions of coding DNA have more effect on phenotypic measure.

I was unable to write a proper code for this step of the pipeline. I wanted to write a code that would compare each significant position to the ranges given in the gene model file. If the significant position was found in a specific range, it would add that specific position as a value to the key of column 3 (V3). For example, if position 1 was an exon, the code would add the value 1 to the key exon. After this, I would be able to identify the number of significant positions that were exons vs. CDS vs. etc.

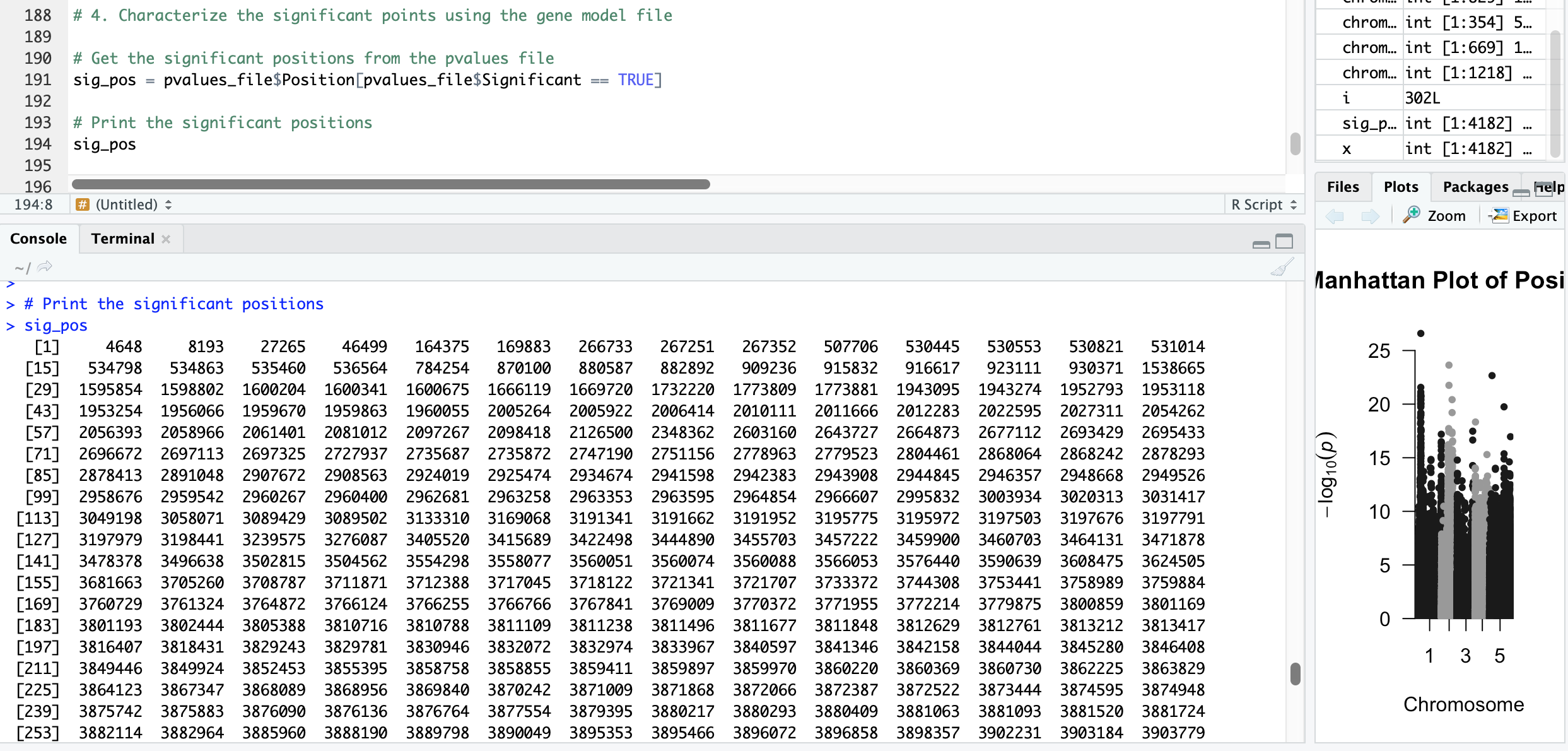


Figure 5: Code snippet showing the significant positions determined.

**Main Findings**

The main findings are summarized in these plots below (Manhattan, Chromosome Distribution, qqplot) as well as the example analysis of step 4 that I would have continued but did not know how to write a code for it.

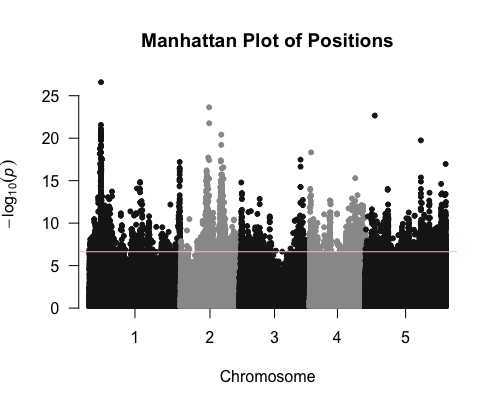


Figure 6: Manhattan plot of positions generated from step 3. Each point is each sample’s nucleotide base for the position in relation to the sample’s phenotypic measure. The pink line indicates the line of significance, and each point above the line are significant in affecting the phenotypic measure.

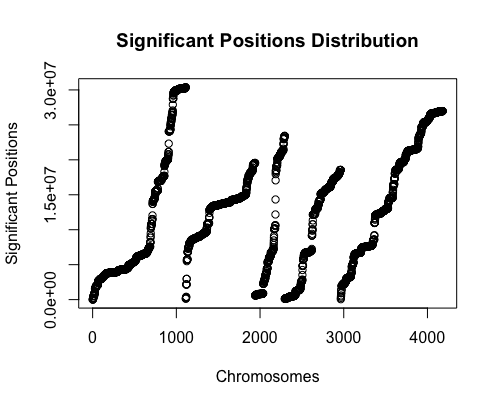


Figure 7: Significant positions distribution. Plot of the distribution of significant positions among the 5 chromosomes. Each index is one chromosome, and the points are each a significant position. Chromosome 1 and 5 seem to have more significant positions than 2, 3, and 4.

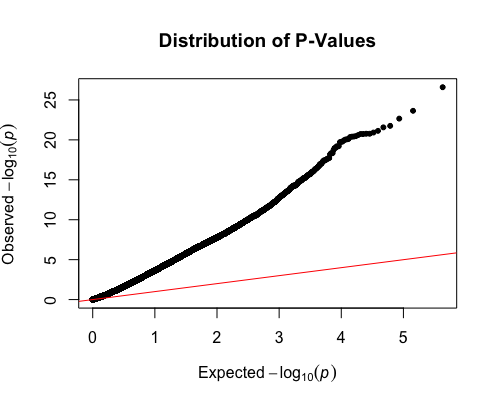


Figure 8: Distributions of p-Values. A qqplot comparing the distribution of the p-values with an expected distribution by chance (red line).

In step 4 of the pipeline to compare the significant positions to their characteristics in the gene model file, I found examples such as:



Figure 9: Examples of characteristics of the significant positions.

**Conclusion**

In conclusion, objective 1 of showing the association between genotype and phenotype can be seen in the generated Manhattan plot (Figure 6). Significant positions from objective 2 is also completed in the Manhattan plot where the points above the significance line are significant positions. These significant positions were characterized based on chromosome distribution (Figure 7), showing that there are more significant positions in chromosomes 1 and 5 than in 2, 3, and 4. The Manhattan plot also identified that these significant positions tend to cluster. As well, the ggplot comparing the distribution of p-values identifies that these results are significant and not due to chance (Figure 8). We can accept the alternative hypothesis, and reject the null.

From the next step of analyses (determining characteristics of significant positions using the gene model file), we can infer that most of the significant positions come from coding DNA. This supports my hypothesis that the phenotype is determined by coding DNA, and that regions of non-coding DNA or “Junk DNA” are not as significant in affecting phenotypic measure. Therefore, significant positions affecting phenotypic measure are likely to be coding regions. This also makes sense in that most of the genome is non-coding, and therefore we have a much smaller number of actual significant positions that affect the phenotype.

As well, from my research, TAIR10 is *Arabidopsis thaliana (thale cress)* (2)*.* TAIR10 has been identified to have many annotated non-coding regions (3). These regions included microRNA, long intergenic RNA, small nucleolar RNA, natural antisense transcript, small nuclear RNA, and small RNA (3). Regions of coding DNA includes exon regions, CDS, mRNA, gene, and protein. As seen in figure 9, the significant positions are all in regions of coding DNA. For position 164375, it is associated with protein coding gene known as AT1G01450. The online TAIR database shows this information for AT1G01450 (4):

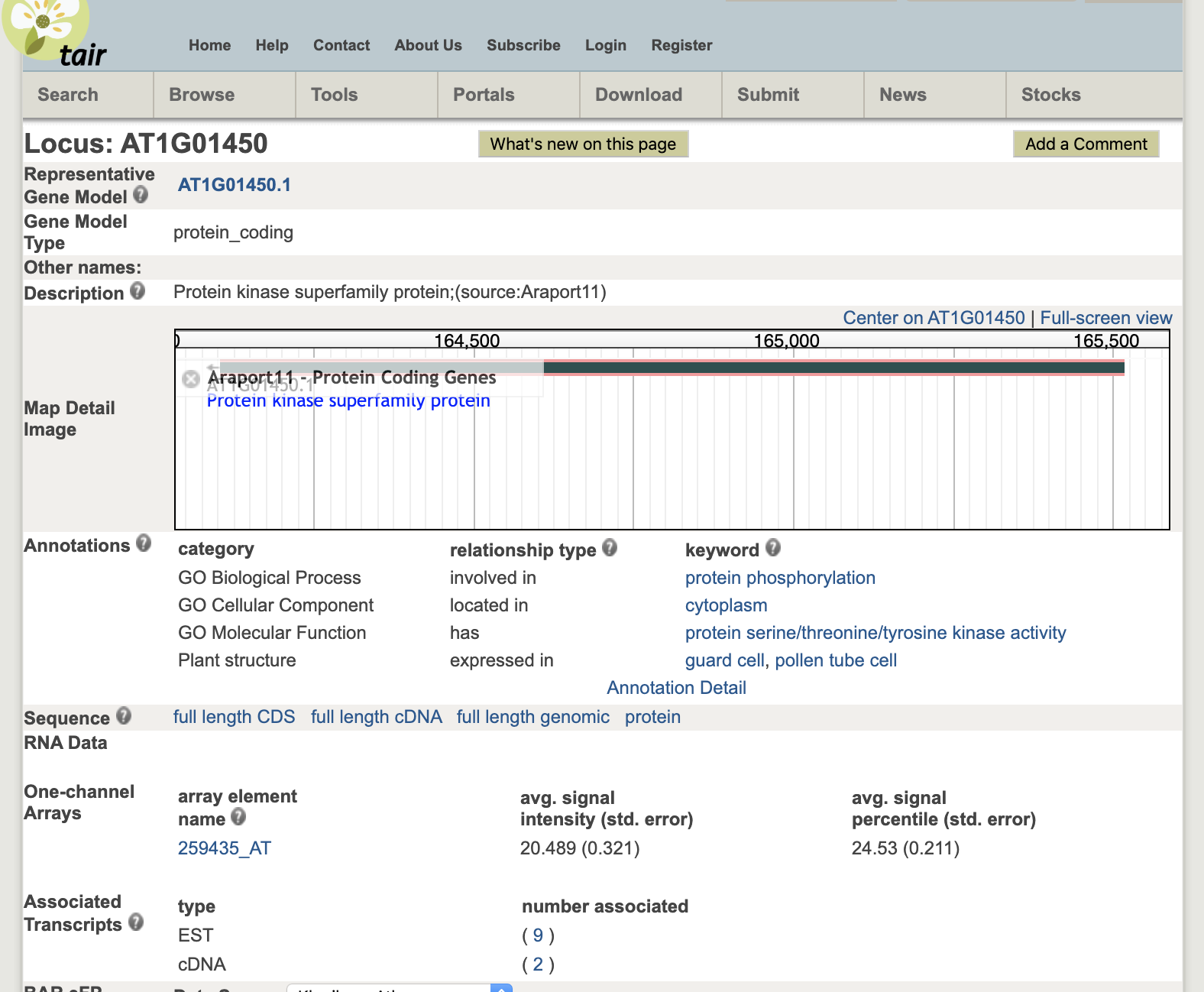


Figure 10: TAIR database for AT1G01450 found from significant position 164375. The database identifies this gene as from the protein kinase superfamily protein (4).

Further analysis could be done to identify more genes of significance, and further characterize each significant position identified. This further analyses should further support my hypothesis.

**Reflection and Bonus**

This project could be further analyzed more deeply as mentioned above. The methods used in this project required a lot of computing power and space since the files were so large, particularly the genotype file, which resulted in RStudio crashing multiple times during the different analysis methods I tested. I learned that there are multiple different ways to analyze the association between genotype and phenotype, and between files in R in general. My approach was simple to understand, and more difficult for R because of all the computing power it took. I think if I knew more coding, there would be more efficient ways to analyze.

The original approach I wanted to take was to determine wild-type and mutants, where the wild-type would be the most common nucleotide at a specific position and every sample that had a base that was not the majority base, would be considered mutant. For example, if 18/20 samples had a T at a specific position, than T would be the wild-type and A, C, and G would be mutant. From this, I would try to do a t-test between phenotype measures of the wild-types and the mutants. This would result in p-values representing how likely the phenotypic measure differs between wild-type and mutant. My Manhattan plot would then be plotting the difference between wild-type and mutant, and high points would mean that the mutants had very different phenotypic measures than the average of wild-types. This approach would have followed the assumption that the phenotypic measures followed a normal distribution, which could be supported by the central limit theorem. However, I was unable to generate code for this analysis which would have been likely much more efficient.

**References**

1. Code for generating p-value association and p-values file from Brett.
2. <https://www.ncbi.nlm.nih.gov/assembly/GCF_000001735.4/>
3. Cheng CY, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. Araport11 a complete reannotation of the Arabidopsis thaliana reference genome. Plant J. 2017;9:789–804. doi: 10.1111/tpj.13415.
4. <https://www.arabidopsis.org/servlets/TairObject?id=29120&type=locus>