



Morganton Scientific

North Carolina School of Science and Mathematics

Journal of Student STEM Research

An Analysis on Genomic Correlation for Gallstone Susceptibility

Sanjita Srinath¹  

¹North Carolina School of Science and Mathematics 

Abstract

Gallstones, typically benign and harmless hardened deposits of digestive fluids, present in the gallbladder can cause severe painful complications if left untreated and can lead to removal surgery. [1] Quantitative Trait Loci (QTL) analysis can be used to find potential genetic links through the analysis of logarithm of the odds (LOD) scores which can indicate a possible connection between those loci on the mouse chromosome and phenotypic presentation of a trait [2] linked to gallstone susceptibility such as weight, presence, severity and liver weight. Analyses were performed using the R/QTL package on a cohort of mice, in an intercross breed and fed a high-fat diet [3]. Analyses were compared between augmented data sets (to possibly prevent overfitting) and an analysis run on a non-augmented data set.

Keywords QTL Analysis, Gallstone Disease, Genomic Correlation

1. INTRODUCTION

The gallbladder is an organ in the upper right portion of the abdomen, directly below the liver, that releases bile, a fluid that the liver produces, that digests fats. Bile is a solution of cholesterol, bilirubin, and bile salts [4]. There are two primary categories of gallstones, cholesterol and pigment stones. Cholesterol stones form due to a lack of balance of cholesterol, bilirubin, and bile salts in the bile. It can form due to excess bilirubin or cholesterol or a lack of bile salts. The cause of pigment stones is currently unknown, but they tend to develop in patients already suffering from cirrhosis, biliary tract infections, and hereditary blood disorders such as sickle cell anemia [1].

While some traits are very clearly and individually linked to a single particular spot on the genome, most traits are inherently complex and thus, there are multiple locations on the genome that can influence the manifestation of that trait. Gallstone disease and susceptibility fall under that umbrella of traits. Gallstone susceptibility is a multifaceted trait influenced by both genetic and environmental factors and represents a significant health concern with considerable variability in its occurrence among individuals. The development of gallstones is associated with a complex interplay of genetic predisposition and lifestyle factors such as diet and obesity [1].


Quantitative Trait Loci (QTL) analyses can be used to find multiple locations on the genome with high logarithm of the odds (LOD) scores which can indicate a possible correlation or causation between the two. QTL analysis uses statistical methods to link quantitative phenotypic traits to genetic markers on the chromosome to try to genetically explain extremely complex phenotypes [2].

We seek to uncover QTLs that may harbor candidate genes influencing gallstone formation.

The findings from this data-driven approach not only contribute to our understanding of the genetic determinants of gallstone susceptibility but also pave the way for potential insights into personalized preventive strategies and therapeutic interventions.

Published Jun 27, 2024

Correspondence to
Sanjita Srinath
srinath24s@ncssm.edu

Open Access 

Copyright © 2024 Srinath. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license, which enables reusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator.

2. DATA SET(S)

2.1. Setup

In this study, we used the R/qtl package in R [5], developed by Bromen et. al as well as Ritsert Jansen’s MQM method [6] to perform our data analysis. R is a programming language and open-source software environment specifically designed for statistical computing and data analysis. Widely used by statisticians, data scientists, and researchers, R provides a comprehensive suite of tools for data manipulation, statistical modeling, visualization, and the development of custom analytical workflows. R/qtl is an R package designed for conducting Quantitative Trait Locus (QTL) analysis, a statistical method used to identify genetic loci associated with variations in quantitative traits. Developed for genetic mapping studies, R/qtl provides a range of tools for analyzing experimental cross populations, facilitating the detection and characterization of genomic regions influencing complex traits.

2.2. Data collection

Data for this project was obtained from the Mouse Phenome Database at the Jackson Laboratory, a grant-funded resource that provides integrated genomic and phenomic data on behavioral, morphological, and physiological characteristics in mice [7]. The Jackson Lab is an independent non-profit biomedical research lab that primarily conducts genomic research with mice.

Figure 1 shows an in-depth chart, demonstrating visually how the mice are bred for effective analysis. This specific dataset, the Lyons1 data set, looks at plasma lipids and gallstone susceptibility in the F2 progeny of a DBA/2J x CAST/EiJ intercross. [3] There are two primary crosses of mice used for QTL analyses, intercross and backcross. As depicted by Figure 1, an intercross is characterized by two homozygous mice in the parental generation bred to produce heterozygous F1 or first-filial generation children. These F1 mice are then bred together to produce the F2 generation who are then utilized in experiments and studies. An interesting side note is that all of the AA, or homozygous dominant mice in the parental generation are deeply inbred and thus genetically identical, and so are all of the BB, or homozygous recessive mice in the parental generation [8, Chapter 3, Section 2].

Only male mice were utilized in the study. The animals had unrestricted access to both food and water and were housed in a temperature-regulated environment (71.6°F - 73.4°F approximately) which had a 14 hours of light and 10 hours of dark cycle. The animals were initially fed a low-cholesterol diet until the age of 6-8 weeks when they were switched to a lithogenic, high-cholesterol diet. This diet was composed of 15 percent butterfat, 1 percent

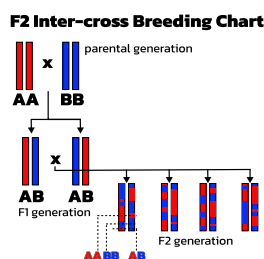


Figure 1. Breeding Chart for Intercross Strains

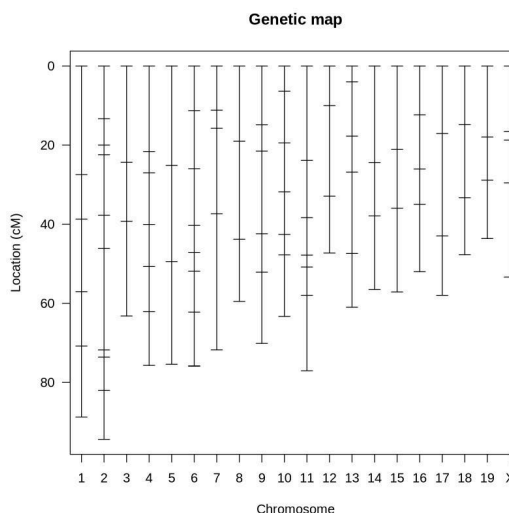


Figure 2. Genetic Map for Markers on Mouse Chromosome

cholesterol, 0.5 percent cholic acid, 2 percent corn oil, 50 percent sucrose, and 20 percent casein. All experimental protocols were approved by the Institutional Animal Care and Use Committees of The Jackson Laboratory and Harvard University [9].

2.3. Data structure

The data used was an F2 intercross with 278 individuals. There were 15 phenotypes and all 15 phenotypes had over 96.8 percent of the individual mice phenotyped. Mice have 20 chromosomes, 19 autosomes, and one sex chromosome, the X chromosome. There are 109 molecular markers in this data and the genetic map is shown above in Figure 2.

There was a 97.4 percent rate of genotyping, meaning this data is extremely complete and this is shown in Figure 3 below. We can see that there is very little missing data which is 2.6 percent of total data missing according to the summary function in R/qtl.

Of the 15 phenotypes presented in the dataset, we chose to focus on four: a score on the severity of the gallstone, the number of gallstones measured, the weight of the gallstones, and the aggregates of the severity of cholesterol monohydrate crystals, which is a key indicator of gallstone development.

3. DATA PREPARATION AND MODELING

3.1. Data Preparation

Prior to running the analyses, we had to prepare the data further. We first completed a pairwise recombination factor plot to take a look at the physical distances between markers on the chromosome and ensure that they are accurate. We first estimate recombination fractions between markers within a genetic cross with the est.rf function. Recombination fractions are crucial in genetic mapping as they indicate the likelihood of genetic crossovers occurring between markers during the formation of gametes. These fractions are fundamental for constructing genetic maps, elucidating the distances between genetic markers, and identifying regions of the genome associated with specific traits through QTL analyses. We then generate a visual representation of the estimated

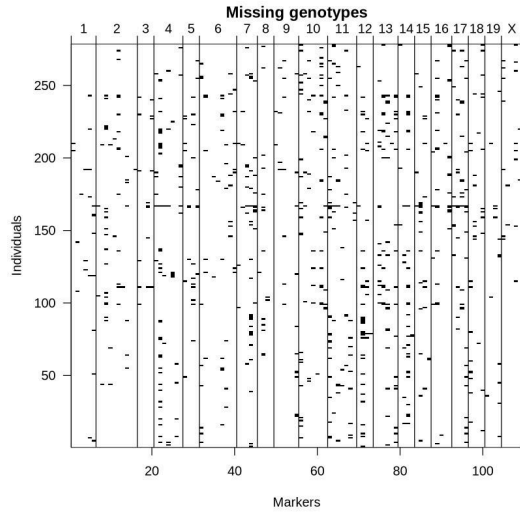


Figure 3. Missing Data Map

recombination fractions which is essential in understanding the genetic linkage and physical distances between markers along the chromosomes, providing researchers with insights into the genetic architecture of traits and facilitating the identification of potential genomic regions influencing complex phenotypes. We then plotted our pairwise recombination scores and LOD scores in Figure 4.

As evidenced by the lack of large red spots and a clean line, this data is clean and alright to use for further analysis.

3.2. Exploratory Data Analysis

After cleaning our data set and ensuring quality control, we then began exploratory data analysis. The first step was using the R/ ggplot2 library to explore trends within the phenotypic data. We used a correlation heat map to identify correlations between phenotypes and the manifestation of gallstones. Correlation heat maps are a visual representation of the coefficient of determination between various factors, or the r-squared value in a color-coded matrix. A value with an absolute value of 1 has a very strong

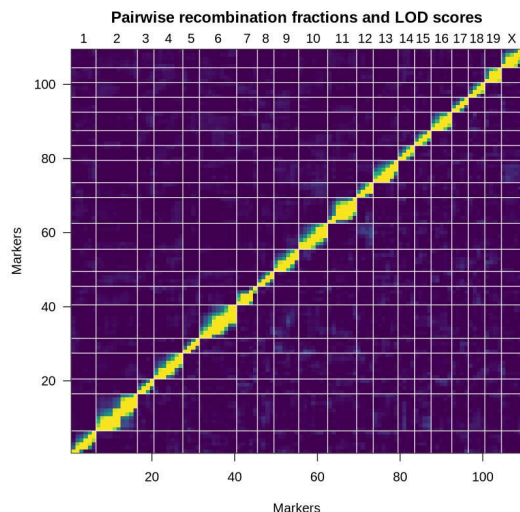


Figure 4. Pairwise Recombination Scores and LOD Scores

correlation, either positive or negative. A value closer to 0 has less correlation and is more random. A positive number correlates to a positive correlation, meaning as the x-value increases, so does the y-value. A negative number correlates to a negative correlation, meaning as the x-value increases, the y-value decreases. This is shown below in Figure 5.

```

      pgm          liver_wt
gallstone_presence gallstone_score
  Min.   :0.0000   Min.   :0.9225   Length:278
  Length:278
  1st Qu.:0.0000   1st Qu.:1.3244
  Class  :character Class :character
  Median :1.0000   Median :1.5524
  Mode   :character Mode  :character
  Mean   :0.7266   Mean   :1.6115
  3rd Qu.:1.0000   3rd Qu.:1.8027
  Max.   :1.0000   Max.   :3.2405
gallstone_count    gallstone_weight
gallstone_sandy    gallstone_solid
  Length:278        Length:278
  Length:278        Length:278
  Class  :character Class :character
  Class  :character Class :character
  Mode   :character Mode  :character
  Mode   :character Mode  :character

      ChMC_agg      ChMC_ind      chol
HDL_log
  Length:278        Length:278
  Length:278        Length:278
  Class  :character Class :character
  Class  :character Class :character
  Mode   :character Mode  :character
  Mode   :character Mode  :character

      nonHDL
  Length:278
  Class  :character
  Mode   :character
  
```

```

      pgm          liver_wt
gallstone_presence gallstone_score
  Min.   :0.0000   Min.   :0.9225
  Min.   :0.0000   Min.   :0.000
  1st Qu.:0.0000   1st Qu.:1.3244   1st
  Qu.:0.0000     1st Qu.:0.000
  Median :1.0000   Median :1.5524
  Median :1.0000   Median :2.000
  Mean   :0.7266   Mean   :1.6115
  
```

Mean :0.6015	Mean :1.288	
3rd Qu.:1.0000	3rd Qu.:1.8027	3rd
Qu.:1.0000	3rd Qu.:2.000	
Max. :1.0000	Max. :3.2405	
Max. :1.0000	Max. :2.000	
		NA's :7
NA's :7		
gallstone_count	gallstone_weight	
gallstone_sandy	gallstone_solid	
Min. : 0.000	Min. :0.0000	Min. :0.000
Min. :0.0000		
1st Qu.: 0.000	1st Qu.:0.0000	1st Qu.:0.000
1st Qu.:0.0000		
Median : 0.000	Median :0.0000	Median :1.000
Median :0.0000		
Mean : 1.129	Mean :0.1508	Mean :1.443
Mean :0.6494		
3rd Qu.: 0.000	3rd Qu.:0.0000	3rd Qu.:3.000
3rd Qu.:0.0000		
Max. :28.000	Max. :3.1600	Max. :4.000
Max. :4.0000		
NA's :7	NA's :9	NA's :7
NA's :7		
ChMC_agg	ChMC_ind	chol
HDL_log		
Min. :0.000	Min. :0.0000	Min. : 67.0
Min. :1.079		
1st Qu.:0.000	1st Qu.:0.0000	1st Qu.:185.0
1st Qu.:1.806		
Median :1.000	Median :0.0000	Median :223.0
Median :1.892		
Mean :1.601	Mean :0.7232	Mean :243.9
Mean :1.901		
3rd Qu.:3.000	3rd Qu.:1.0000	3rd Qu.:285.0
3rd Qu.:1.991		
Max. :4.000	Max. :4.0000	Max. :668.0
Max. :2.455		
NA's :7	NA's :7	NA's :1
NA's :1		
nonHDL		
Min. : 8.0		
1st Qu.:102.0		
Median :137.0		
Mean :158.0		
3rd Qu.:194.5		
Max. :631.0		
NA's :3		

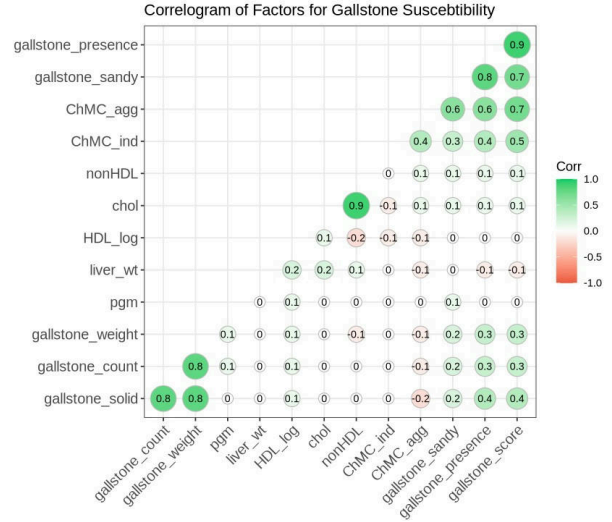


Figure 5. Correlation Heat Map

In this figure, we can see that there is a high positive correlation between multiple factors, particularly between the binary classification of the solidity of gallstones and the number of gallstones, the weight of gallstones and the number of gallstones, the weight of the gallstones and the binary solidity classification, and the presence of gallstones and the severity score.

We also ran a dendrogram heat map to analyze correlations as well as find hierarchical correlations between our phenotypic factors, shown in Figure 6. This map shows both a heat map to show correlations between various factors, similar to our previous graph, but also shows hierarchical relationships between our phenotypic factors. This helps us understand the degree of the relationship between the factors.

3.3. QTL Analyses

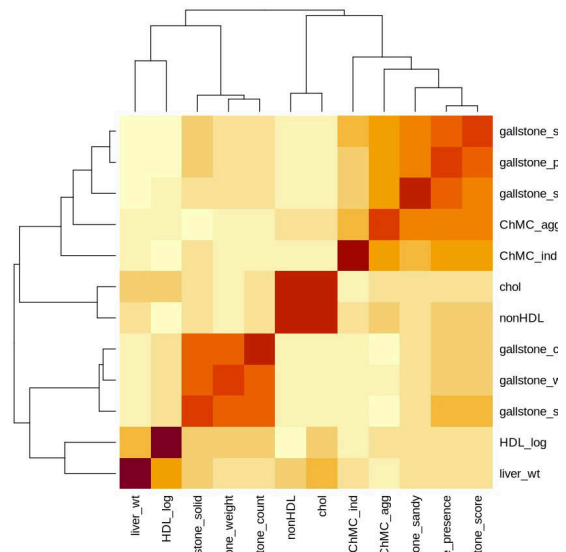


Figure 6. Dendrogram Hierarchical Heat Map

Following this, we can begin the setup for the QTL analysis. We first calculated the genotypic probabilities for individuals in the genetic cross by determining the likelihood of different genetic marker configurations based on specified parameters such as recombination step size, genotyping error probability, and the Haldane map function. We then simulated genotypic data for the markers in the genetic cross, incorporating factors like recombination, genotyping errors, and mapping functions. This is the primary first step we must do prior to running any analyses and determining the locus of interest.

Following this, we completed the same steps for each of our four factors to generate a main scan analysis as well as effect plots for the highest probability locus of interest as determined by LOD scores. We first used the function scanone with a normal model and the “em” method, which is the Expectation-Maximization method which estimates missing genotype probabilities in the genetic mapping analysis. While this is specifically excellent for data sets with missing phenotypic information, we find that it is still a robust method of analysis.

For each method, we then completed a permutation on the scan with 100 permutations to assess the significance of LOD peaks for each phenotype. We then assigned threshold values based on our permutation results with confidence intervals of 95 percent, 90 percent, and 63 percent respectively. After this, we plotted the results onto a main scan plot with colored lines representing our thresholds. We also ran a summary of the scan per phenotype and identified the most probable locus of interest per scan. We then used this location to identify a molecular marker in our data set and run an effect plot. These plots are particularly useful for understanding how genetic variation at specific loci influences the phenotypic variation in a quantitative trait. We can see how homozygous recessive or dominant, or heterozygous affects the manifestation of different phenotypic traits.

Completing this, we decided to explore augmented data to see if results run on augmented data on a total QTL analysis for the data set. To do this, we first created an augmented data set derived from our cross with a minprob of 0.1. This establishes a minimum probability threshold for considering the effects of additional markers or QTLs. This threshold, which influences the augmentation step, allows us to filter out less statistically significant QTL effects, refining the model and focusing on those with higher confidence. The choice of the minprob threshold serves as a key determinant in balancing sensitivity and specificity in the identification of quantitative trait loci, tailoring the analysis to the desired level of statistical rigor. We chose 0.1 because we wanted more statistically significant results rather than a broader overview with less statistically significant results. We then ran a geno.image on both the augmented cross and the original cross and compared the plots. Following this we took a scan of each cross, using the mqm scan for the augmented set and scanone for the original set, and found the peaks on each plot. We then found a molecular marker corresponding to each peak and compared it to each other. Following this, we used that marker we identified earlier as a cofactors and took another mqm scan with the cofactor of D18Mit64, the marker we identified. We then proceeded to plot all three main scans together on the same plot and compared the peaks.

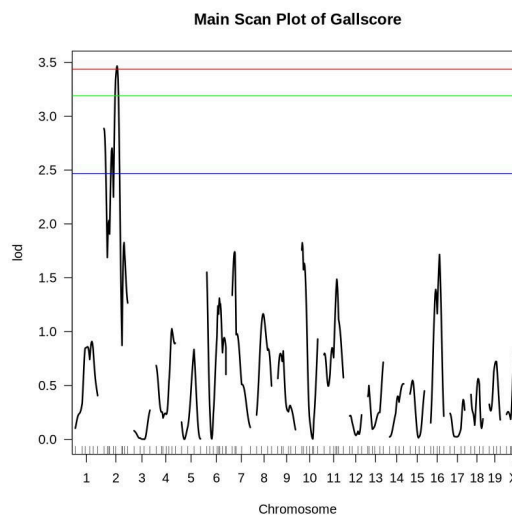


Figure 7. Main Scan for Gall Count

4. RESULTS

4.1. Expectation-Maximization Model

Using the Expectation-Maximization model, we generate 4 separate main scans with threshold lines at 95 percent confidence, 90 percent confidence, and 63 percent respectively.

4.1.1. Gall Count:

As shown below in Figure 7, for a count of gallstones present, we found two loci with a peak over 95 percent confidence.

There was 98 percent confidence in the correlation between c6.loc6 and phenotypic manifestation and 96 percent confidence in the correlation between c8.loc58 and phenotypic manifestation. We then found the correlated molecular markers for those two spots which were D6Mit46 and D8Mit88 respectively. Using those two markers, we then plotted an effect plot as shown below in Figure 8. As we can see in both figures, a homozygous DD genotype at both of these locations can correlate to gallstone susceptibility and a higher amount of gallstones while heterozygous DC and homozygous CC both present lower amounts of gallstones.

4.1.2. Gall Score:

As shown below in Figure 9, for a score on how severe the gallstones were, we found one locus with a peak over or equal to 95 percent confidence.

There was 95 percent confidence of correlation between c2.loc52 and phenotypic manifestation. We then found the correlated molecular markers for this spot which was D2Mit94. Using this marker, we then plotted an effect plot as shown below in Figure 10. As we can see, a homozygous CC genotype at this location can correlate to a lower gallstone severity score while heterozygous DC and homozygous D both present higher scores. It can be said then that high gallstone severity is a dominant trait at this location in the genome.

4.1.3. Gall Weight:

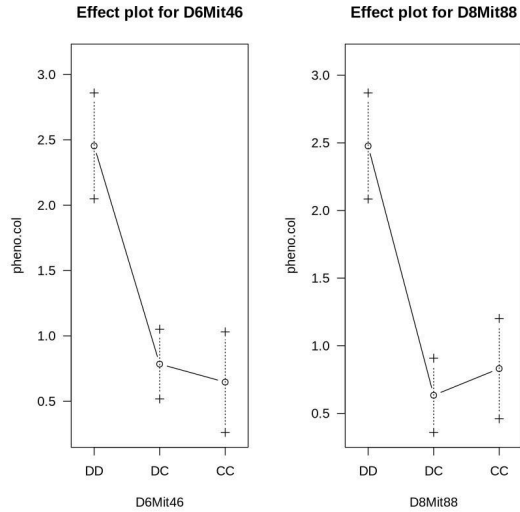


Figure 10. Effect Plot showing Allele vs Phenotypic Presentation

As shown below in Figure 11, for a score on how severe the gallstones were, we found one locus with a peak over or equal to 95 percent confidence.

There was 95 percent confidence in the correlation between c2.loc52 and phenotypic manifestation. We then found the correlated molecular markers for this spot which was D8Mit88. Using this marker, we then plotted an effect plot as shown below in Figure 12. As we can see, a homozygous DD genotype at this location can correlate to a higher gallstone weight score while heterozygous DC and homozygous C both present lower weights. It can be said then that high gallstone weight is a recessive trait at this location in the genome.

4.1.4. Cholesterol Monohydrate Crystals, aggregates:

As shown below in Figure 13, for a score on the cholesterol monohydrate crystals, we found one locus with a peak over or equal to 95 percent confidence.

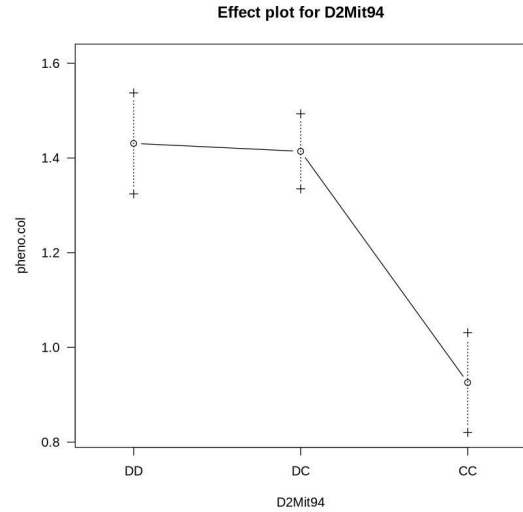


Figure 12. Effect Plot showing Allele vs Phenotypic Presentation

There was 95 percent confidence of correlation between c6.loc56 and phenotypic manifestation. We then found the correlated molecular markers for this spot which was D6Mit62. Using this marker, we then plotted an effect plot as shown below in Figure 14. As we can see, a homozygous DD genotype at this location can correlate to a lower cholesterol monohydrate crystal score while heterozygous DC and homozygous D both present higher scores.

4.2. Augmented QTL Analyses Comparison

After generating an augmented dataset, we then used geno.image to plot both crosses respectively as shown below in Figure 15.

The genotypes CC, DC, and DD are displayed in the colors red, blue, and green, respectively. The white spaces represent missing data. As we can see, the augmented data is filled in much better, and there is no missing data. While running the summary function on the data, we see that nothing has changed in the augmented as compared to this original other than there being much more individuals in the data set (1343 as compared to 278). Additionally, the percent

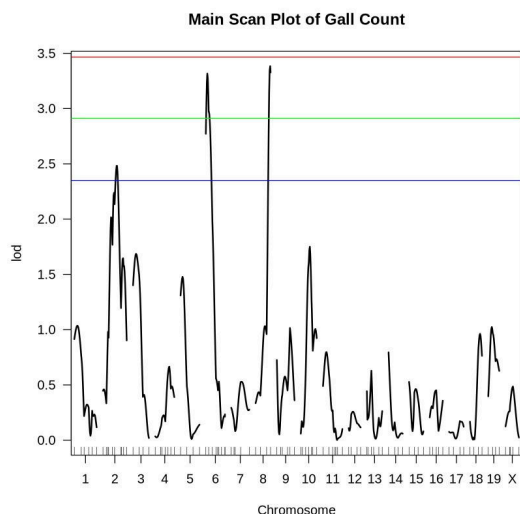


Figure 10. Main Scan for Gall Score

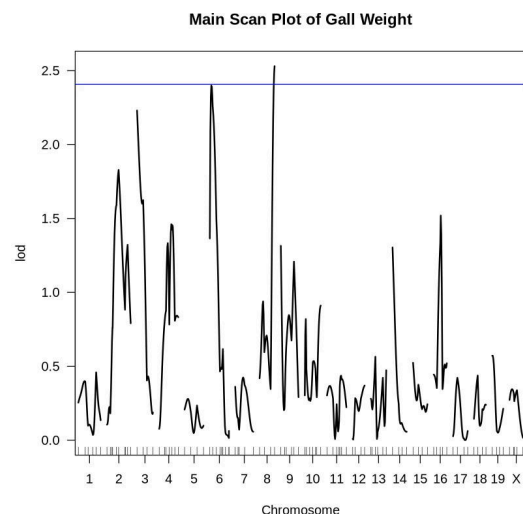


Figure 12. Main Scan for Gall Weight

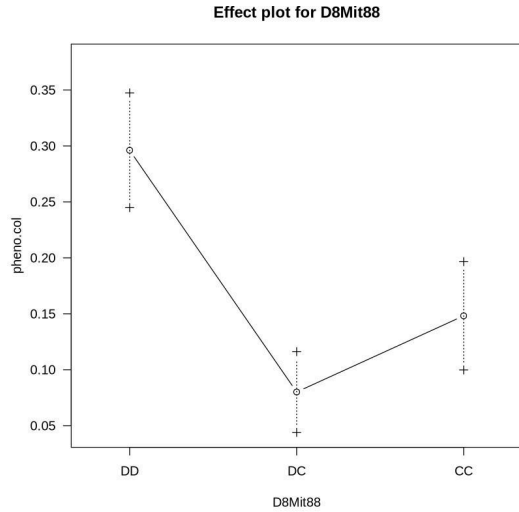


Figure 15. Effect Plot showing Allele vs Phenotypic Presentation

phenotyped remains approximately the same and the percent genotyped jumps up to 100 percent.

Next, we complete an mqmscan on the augmented dataset and a scanone on the original dataset and take a look at the maximum point on both of these. The augmented dataset has a peak at c18.loc5 which is 5 centimorgans on chromosome 18. The original dataset has a peak at c18.loc4 which is 8.46 centimorgans on chromosome 18. We then extract a marker for both of these positions which both are D18Mit64. We can then set that marker as a covariate and analyze for a new peak, with this marker as an additional variable.

We then plotted all three of these plots on the same map with green representing the original data, red representing the augmented data, and blue representing augmented data with the peak as a covariate, as shown in Figure 16 below.

As we can see, the augmented data narrows down the peaks into only a few spots, showing how it counters overfitting due to the small nature of the original data set. As the original data set was

Main Scan Plot of Severity of Cholesterol Monohydrate Crystals, aggrega

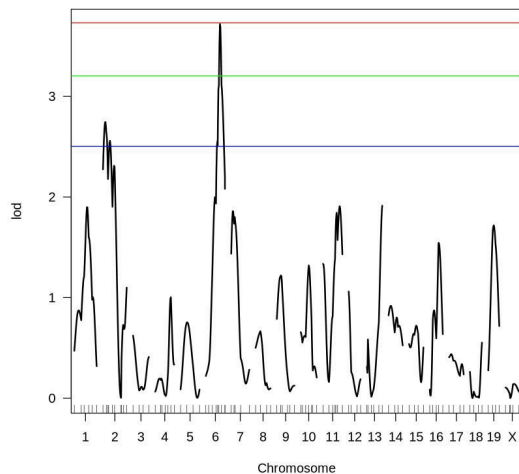


Figure 15. Main Scan for Gall Weight

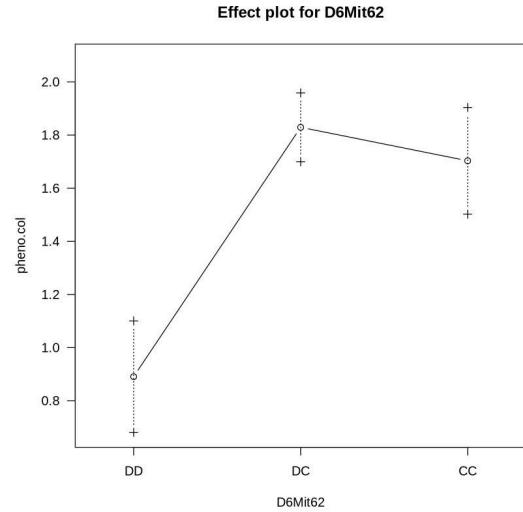


Figure 16. Effect Plot showing Allele vs Phenotypic Presentation

much smaller than the augmented one, we can hypothesize that there are fewer peaks on the augmented model as it eliminates peaks on the original which could be due to overfitting.

5. DISCUSSION

In this study, we identified a great many loci of interest to investigate as shown in the table above.

These results are statistically significant as each of them passes the threshold value of 0.05 meaning that there is a 95 percent confidence rate for a correlation between that loci and the associated genotypes. In targeting this disease from a genomic standpoint, it may be worthy to first target those markers that are dominant for the associated phenotype. These would be D2Mit94 and D6Mit62. Research may be further made into chromosomes 2, 6, 8, and 18 as

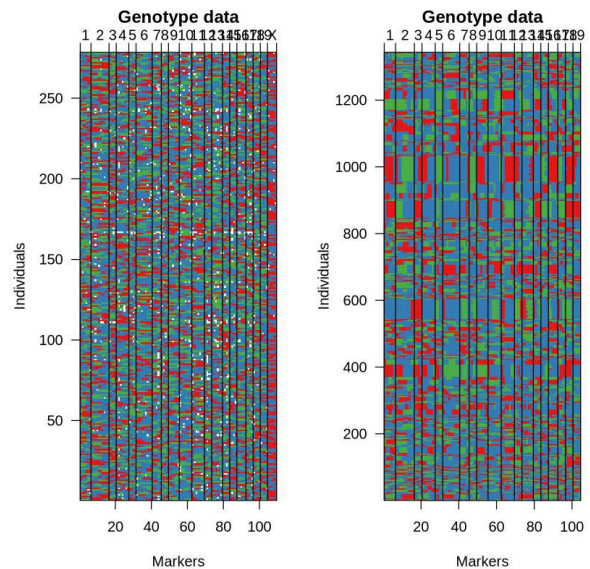


Figure 16. Plot Grid of Original Genotype Data (left) and Augmented Genotype Data (right)

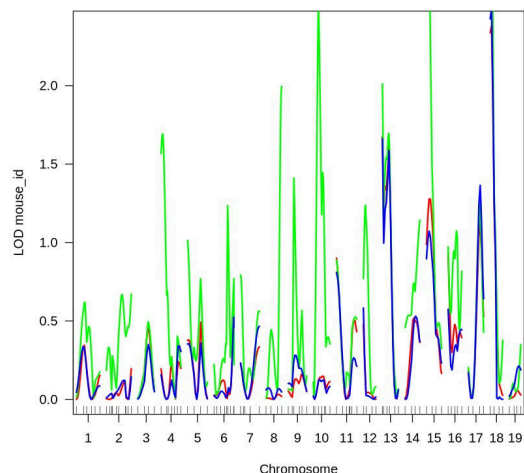


Figure 16. Overlaid Scans on Three Models

those are the most prominent chromosomes which correlate to increased gallstone susceptibility.

6. CONCLUSION

Through our comprehensive QTL analysis exploring the genomic correlation for gallstone susceptibility, we uncovered three chromosomes of interest and 2 molecular markers of interest to target first. We conducted a robust exploration using the R/qtl package, leveraging the power of statistical methods like the Expectation-Maximization model.

Molecular Marker	Dominant/Recessive	Threshold	Phenotype
D6Mit46	Recessive	0.05	High Gall Count
D8Mit88	Recessive	0.05	High Gall Count
D2Mit94	Dominant	0.05	High Gall Score
D8Mit88	Recessive	0.05	High Gall Weight
D6Mit62	Dominant	0.05	High Cholesterol Aggregate Crystal Formation
D18Mit64	Recessive	0.05	High Susceptibility (Augmented Data)
D18Mit64	Recessive	0.05	High Susceptibility (Original Data)

Our findings illuminated several key loci with high logarithm of the odds (LOD) scores, providing significant insights into the genetic underpinnings of gallstone susceptibility. Notably, we identified loci associated with gallstone count, severity, weight, and cholesterol monohydrate crystals. The allelic variations at these loci demonstrated correlations with distinct phenotypic presentations, unraveling the complexity of genetic influences on gallstone-related traits. Furthermore, by employing an augmented dataset and comparing results with the original dataset, we sought to enhance the robustness of our analysis. The augmented data, with its increased sample size, presented a more comprehensive view of the genomic landscape associated with gallstone susceptibility. The overlay of scans from the original and augmented datasets, along with the inclusion of a covariate, provided a nuanced understanding of the genetic factors at play.

Our study contributes to the fundamental understanding of gallstone susceptibility and lays the foundation for personalized preventive strategies and therapeutic interventions. The identified QTLs harbor candidate genes that may play pivotal roles in gallstone formation, paving the way for further targeted research.

ACKNOWLEDGMENTS

Special thanks are extended to Mr. Robert Gotwals of the North Carolina School of Science and Math and the Mouse Phenotype Database at the Jackson Laboratory.

This paper is solely the work of the author. All references are included in the bibliography and are cited appropriately.

The authors declare that they have no competing interests.

The data for this work was obtained from <https://phenome.jax.org/projects/Lyons1>.

REFERENCES

- [1] "Gallstones." [Online]. Available: <https://www.hopkinsmedicine.org/health/conditions-and-diseases/gallstones>
- [2] C. Myles and M. Wayne, "Quantitative trait locus (QTL) analysis," *Nature Education 1 (1)*, vol. 208, 2008.
- [3] Lyons, "Lyons1: Plasma lipids and susceptibility to gallstones in F2 progeny of DBA/2J x CAST/Eij." [Online]. Available: <https://phenome.jax.org/projects/Lyons1>
- [4] "Gallbladder: What Is It, Function, Location & Anatomy." [Online]. Available: <https://my.clevelandclinic.org/health/body/21690-gallbladder>
- [5] K. W. Broman, H. Wu, S. Sen, and G. A. Churchill, "R/qtl: QTL mapping in experimental crosses," *bioinformatics*, vol. 19, no. 7, pp. 889–890, 2003.
- [6] D. Arends, P. Prins, R. C. Jansen, and K. W. Broman, "R/qtl: high-throughput multiple QTL mapping," *Bioinformatics*, vol. 26, no. 23, pp. 2990–2992, 2010.
- [7] M. A. Bogue *et al.*, "Mouse phenome database: curated data repository with interactive multi-population and multi-trait analyses," *Mammalian Genome*, vol. 34, no. 4, pp. 509–519, 2023, doi: 10.1007/s00335-023-10014-3.
- [8] L. M. Silver, *Mouse genetics concepts and applications*. Oxford University Press, 1995.
- [9] M. A. Lyons *et al.*, "Quantitative trait loci that determine lipoprotein cholesterol levels in DBA/2J and CAST/Ei inbred mice1, 2," *Journal of lipid research*, vol. 44, no. 5, pp. 953–967, 2003.