Identification of Signficant SNPs associated with T-Cell Receptor Diversity

John Lin, Case Western Reserve University

April 05, 2019

Contents

Abstract

There is significant interest in advancing personalized medicine with the ultimate goal of creating highly tailored treatments for individual patients. By aggregating phenotypic data with genomic data, one can draw novel insights into how genetic processes regulate traits and diseases. In particular, single nucleotide polymorphisms (SNPs) contribute to genetic variation resulting in different phenotypes between individuals. One such trait that could be studied further is one's T-cell receptor (TCR) repertoire. Diverse TCR repertoires are associated with strong adaptive immune systems. However, what SNPs are significant predictors in TCR repertoire is not well studied. Below discusses recent work in using advanced filtering techniques to select potentially significant SNPs and initial approaches developed to model TCR diversity given a set of genotypes.

Introduction

Please refer to Report 1 for earlier background information and findings.

Motivation

Two prevous studies were conducted on a cohort of chronic kidney disease (CKD) patients from MetroHealth Medical Center Main Campus' nephrology clinics in Cleveland, OH (see Bailey et al. 2018; Crawford et al. 2018). From March 2016 to July 2017, 134 biospecimens were collected from consenting patients as part of the MetroHealth/Institute for Computational Biology Pilot study (MIPS) (Bailey et al. 2018). After surveying these patients, 62% indicated return of research results specific to their data was important. DNA was extracted from these samples and genotyped by Illumina's MultiEthnic Genotyping Array (MEGA) BeadChip. In parallel, Crawford et al. (2018) selected 15 of these samples ranging in CKD status to have T-cells sequenced by immunoSeq Adaptive Biotechnologies (Biotechnologies 2017). Crawford et al. (2018) note that there was some correlation between TCR diversity and CKD status. Although, the small sample size greatly limited the power their their study, the findings illustrate the potential for new applications of genomic data and investigation of disease processes. In this project, due to the small sample size, it is not a true Genome Wide Association Study (GWAS). Instead, the aim is to find potential signals correlated with TCR diversity and possibly model TCR diversity as a function of those signals.

In this report, further reduction of the number of predictors (in this case, SNPs) through various filtering techniques common to GWAS will be used. Additionally, initial modeling approaches using ridge regression and lasso regression are presented.

The Additive Model

The additive model is a common genetic model to measure the importance of SNPs in regards to a specific phenotype. In this model, independent tests of simple linear regression are performed for every SNP. The minor allele (allele 1, A1) is by default considered to be of significance in the different genotypes compared to the major allele (allele 2, A2). Therefore, the quantitative values of 0, 1, and 2 correspond to no-presence-of-A1, one-allele-is-A1, and both-alleles-are-A1, respectively. This is demonstrated in the below table, using an example of the TT, TC, and CC genotypes, where T is the minor allele.

Genotype	Coding (A1/A2)	Dosage Value (0/1/2)
TT	A1 A1	2
TC	A1 A2	1
CC	A2 A2	0

Using the immunoSeq data for 1 SNP across the 15 patients, we can fit a linear regression by the least squares method producing the plot below (see Figure 1). It should be noted, the above is for demonstration purposes only. The linear regressions and p-values will be automated by `plink`.

```
snp <- "rs6945601"
plinkPedSingle = plinkPed[,c(snp,"phenotype")]
lm.fit <- lm(phenotype ~ rs6945601, data = plinkPedSingle)
jpeg("linearRegressionExample.jpg")
ggplot(data = plinkPedSingle) +
  geom_point(mapping = aes(x = rs6945601, y = phenotype), color = "blue") +
  geom_line(aes(x = rs6945601, y = predict(lm.fit)), color = "red") +
  xlab(paste("snp", snp)) + ylab("TCR Productive Clonality") +
  ggtitle("SNP Genotype vs TCR Productive Clonality")
```

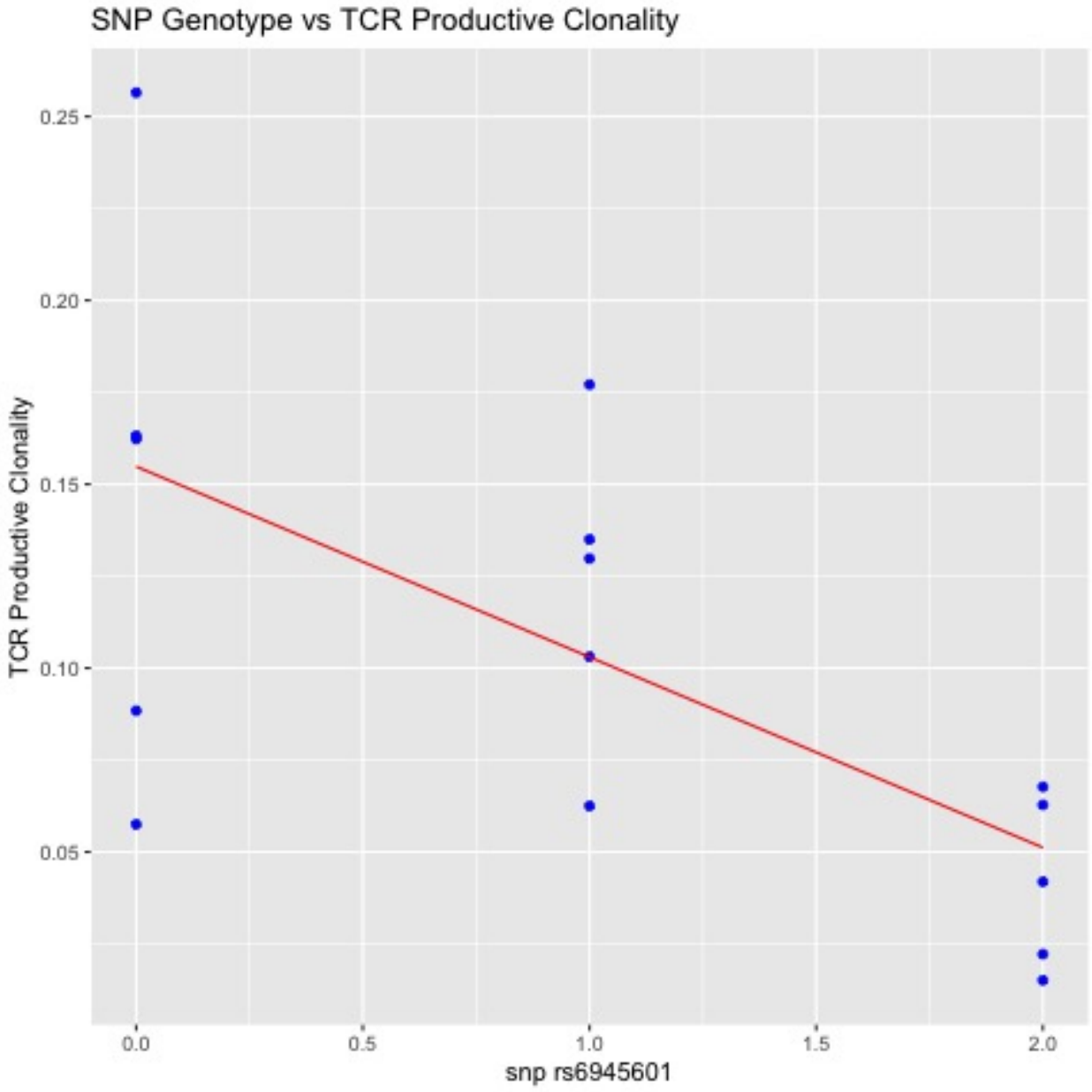

Figure 1: Linear Regression Example

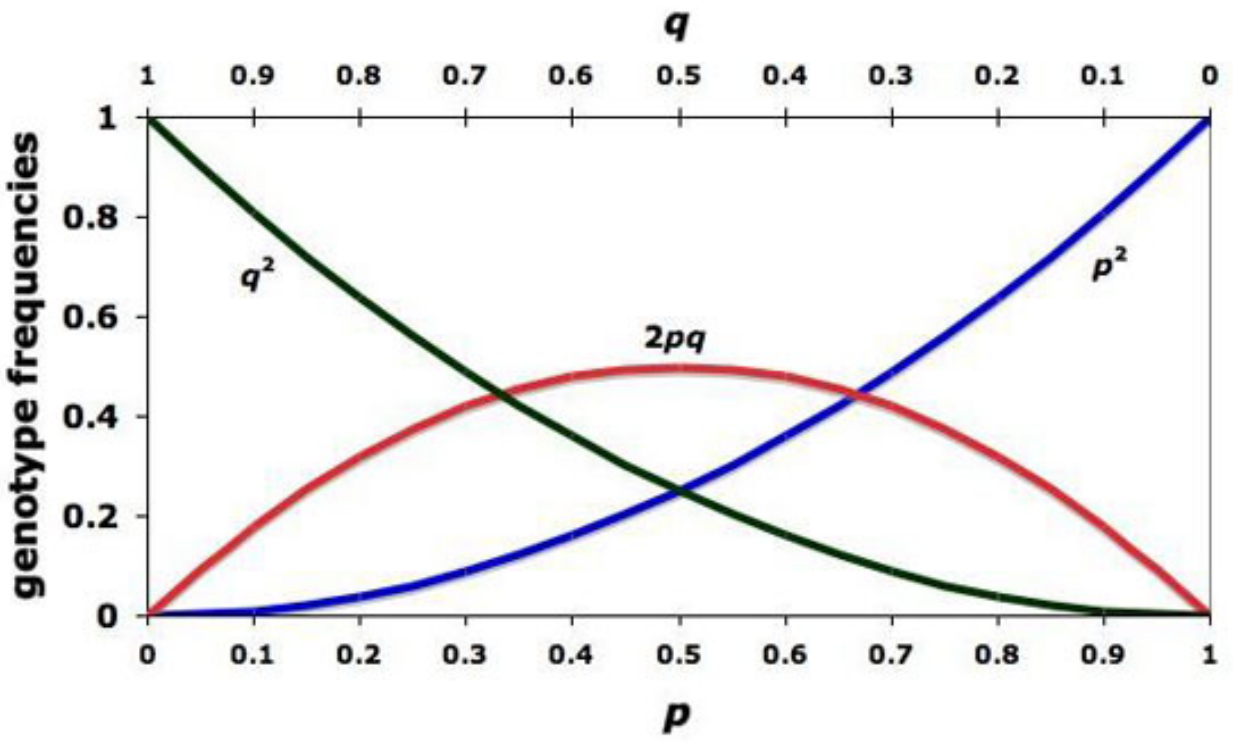

Figure 2: Distribution of Genotype Frequencies in HWE

Filtering: Minor Allele Frequency (MAF)

MAF filtering is performed in order to avoid false positive significant p values. The simple linear regressions performed for each SNP above assume each SNP is drawn from the same distributions. However, allele frequencies vary for each SNP. Therefore, p values derived from lower MAFs have less power and may indicate false positives. Most GWAS filter for MAF $< 10\%$ (Tabangin, Woo, and Martin 2009). Using an MAF threshold of 10% in a population of 15 individuals requires each allele be at least present 3 times (for 1 SNP, 15 people –> 30 alleles –> $30(.10) = 3$).

Filtering: Hardy-Weinberg Equilibrium (HWE)

Hardy-Weinberg equilibrium employs the following:

- Natural selection is not active on the specific locus
- Migration/mutation do not affect allele frequencies
- Population size is infinite
- Random mating
- Allele frequencies do not change between generations.
- p represents the allele 1 frequency and q represents the allele 2 frequency. Therefore, the proportions of different genotypes equate to

$$p^2 + 2pq + q^2 = 1$$

  Here, p^2 represents genotype AA, $2pq$ Aa, and q^2 aa. This results in the distributions in Figure 2.

Therefore, any deviations from Hardy-Weinberg can be calculated through a chi-square test if the number of individuals associated with each genotype is known.

See (Andrews 2010).

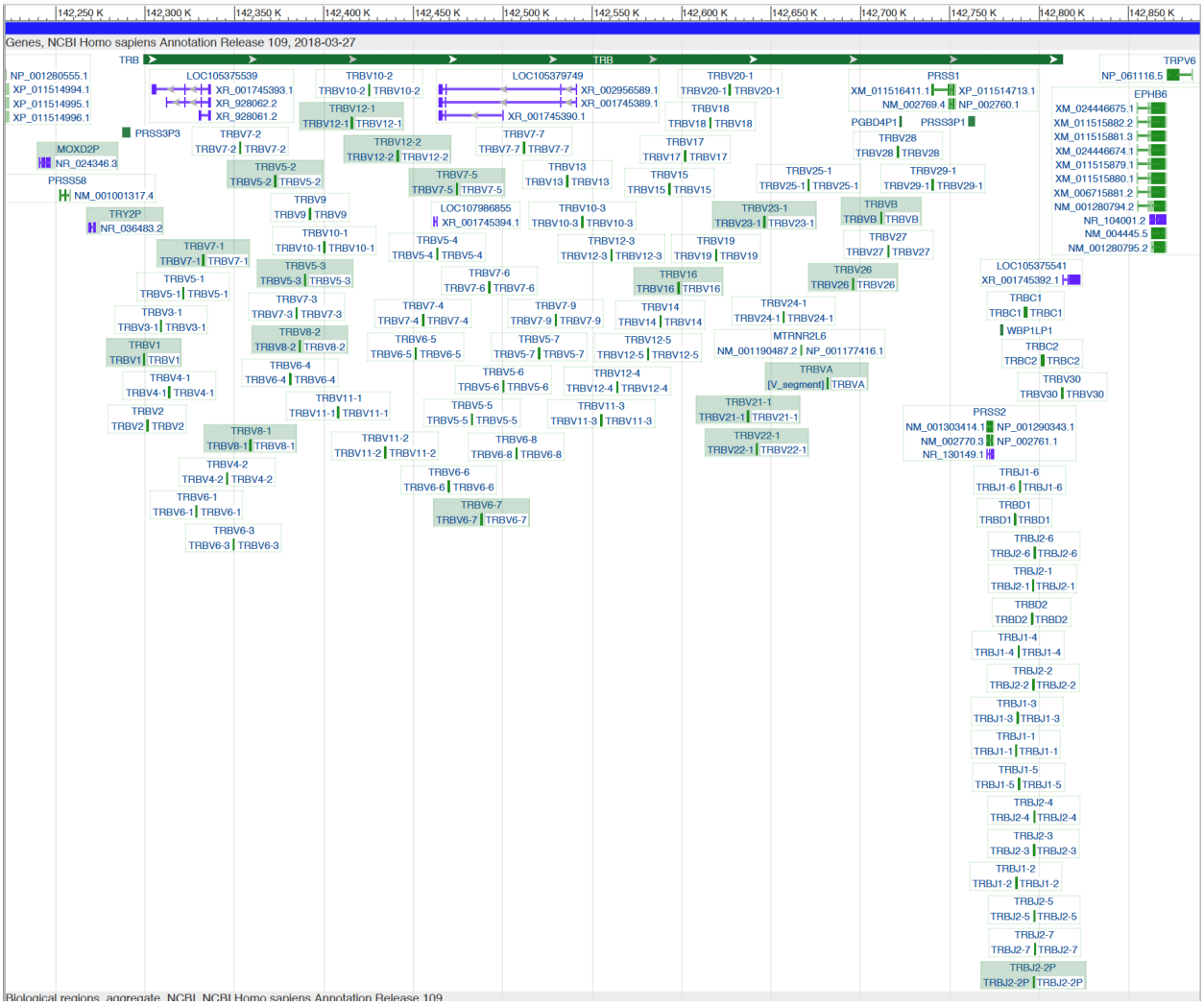

Figure 3: TRB Locus

Filtering: T-Cell Receptor Beta Locus (TRB)

From NCBI (2019) and as shown in Figure 3, the TRB locus resides on chromosome 7 between 141998851 and 142510972 Mb. This is from the GRCh37.p13 build of the human genome. As shown in Figure 3, there are a variety of TRB genes in this region, accounting for the VDJ genes for which T-Cell somatic recombination is notable. In order to catch possible edge cases, the window was expanded by 50,000 bp on each side, resulting in **chromosome 7 positions between 141948851 and 142560972 Mb**. Ultimately, these positions will be used for further SNP filtering.

High Dimensional Considerations

As the number of predictors greatly outnumbers the number of observations, there are some special considerations for these high dimension cases (see James et al. 2013):

- Least squares regression should not be used, since this leads to overfitting.
- The test MSE increases since variance increases.
- R^2, C_p, AIC, and BIC are not appropriate to use in high dimension settings.

Dimension reduction methods such as ridge regression and lasso regression can be used to reduce the variance. It is important to select an appropriate tuning parameter (in this case λ) in order to shrink the coefficients associated with the given predictors. Additionally, selecting predictors that are truly associated with the response can decrease test MSE and decrease variance leading to a better model. Because of this, ridge regression and lasso regression will initially be used to model productive clonality (TCR diversity) as a function of a set of genotypes (SNPs). Future work will involve model refinement.

Data Science Methods

Exploratory Data Analysis and Initial Modeling

Productive Clonality

Normal distribution of productive clonality amongst the 15 samples was checked. A histogram and density plot were constructed as shown below. It appears there is fairly normal distribution of productive clonality, given the low number of samples. See Figures 4 and 5.

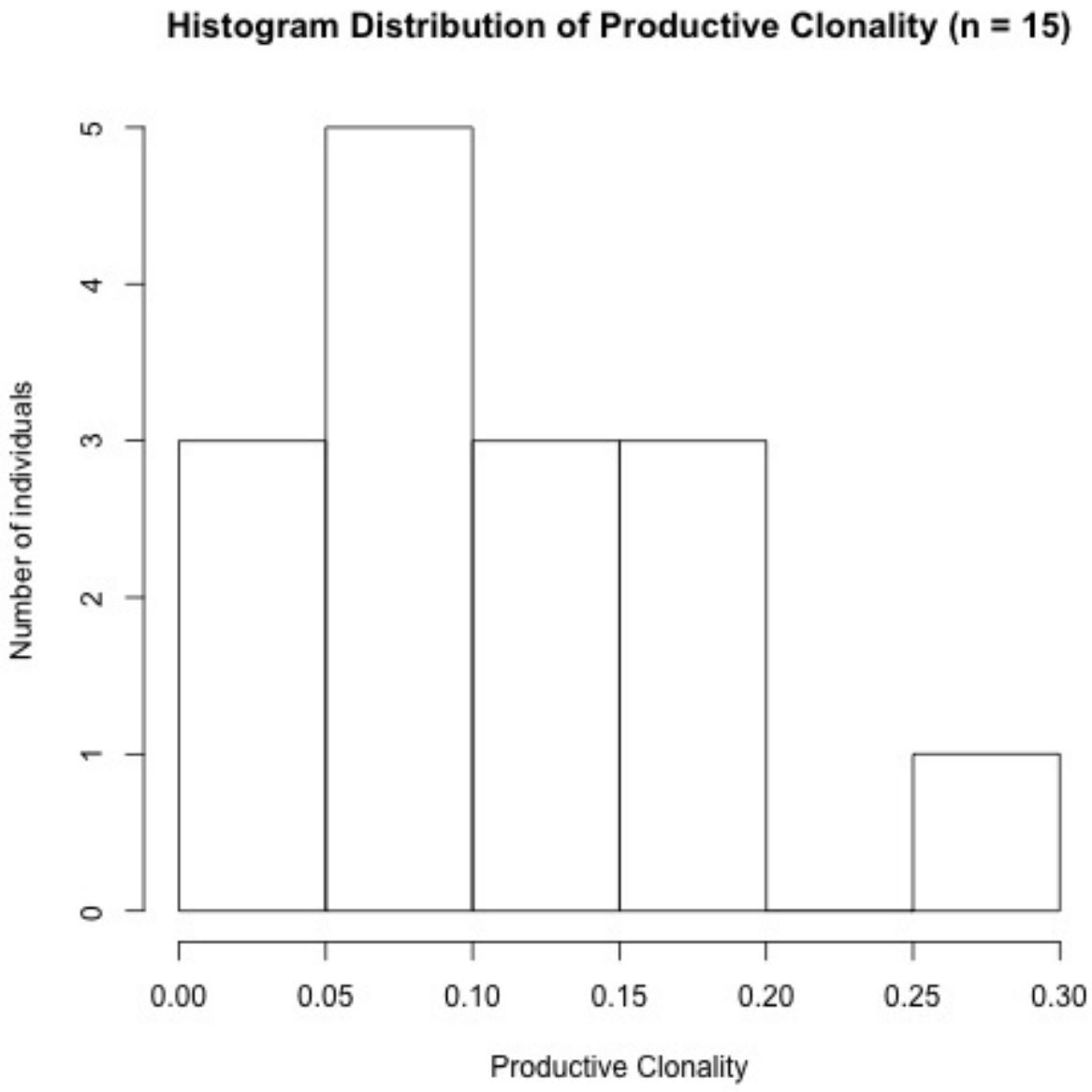

Figure 4: TCR Histogram

Additionally, it should be noted the mean of the productive clonality is 0.1030. Recall that productive clonality ranges from 0 (diverse TCR, strong immune system) to 1 (not diverse, weak immune system. The spread of the data is indicated in Figure 6.

```
# Descriptive Statistics
summary(as.numeric(as.character(tcrEmrPheno$Productive.Clonality)))
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 0.0151  0.0600  0.0884  0.1030  0.1487  0.2565
```

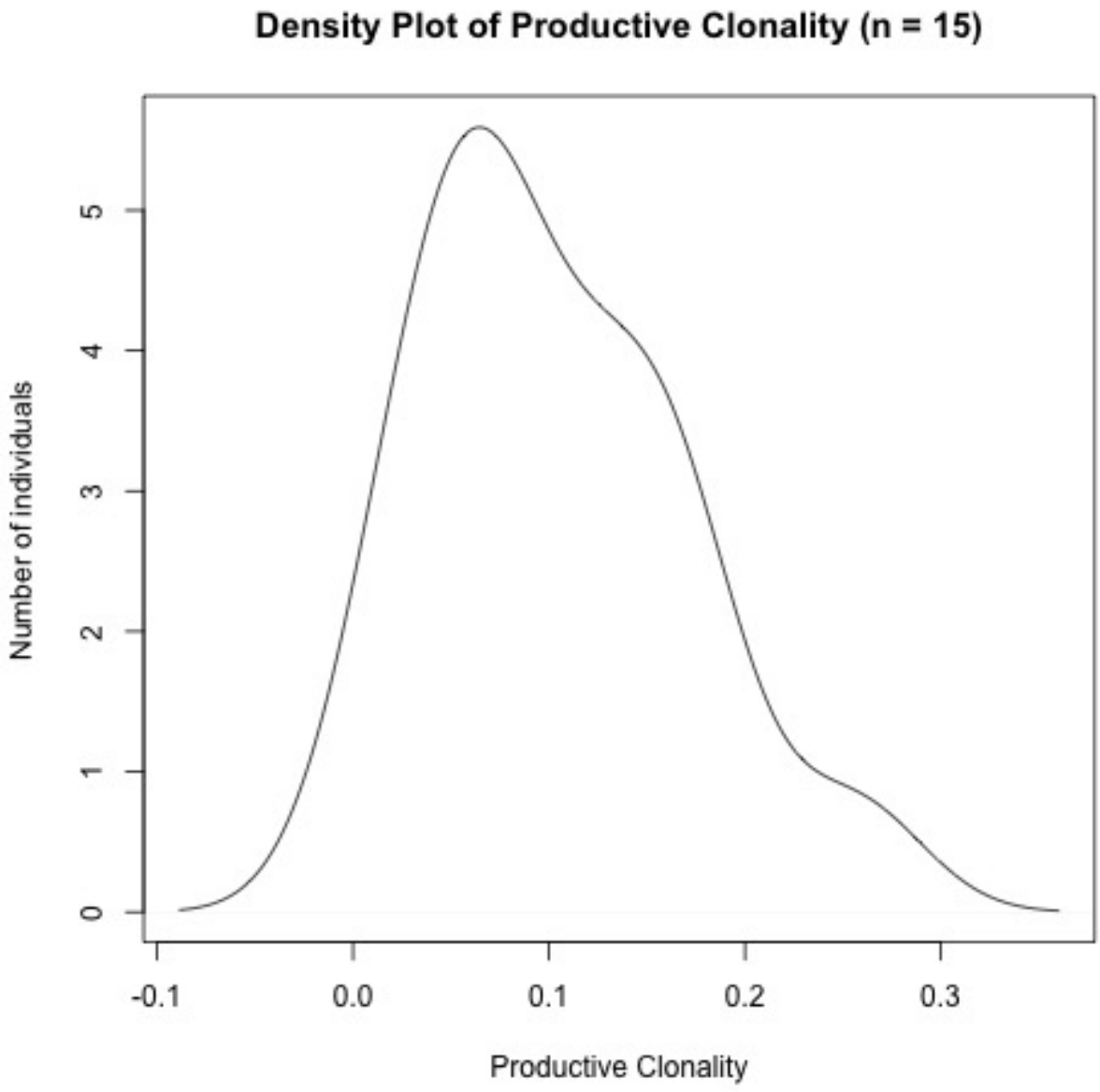

Figure 5: TCR Density

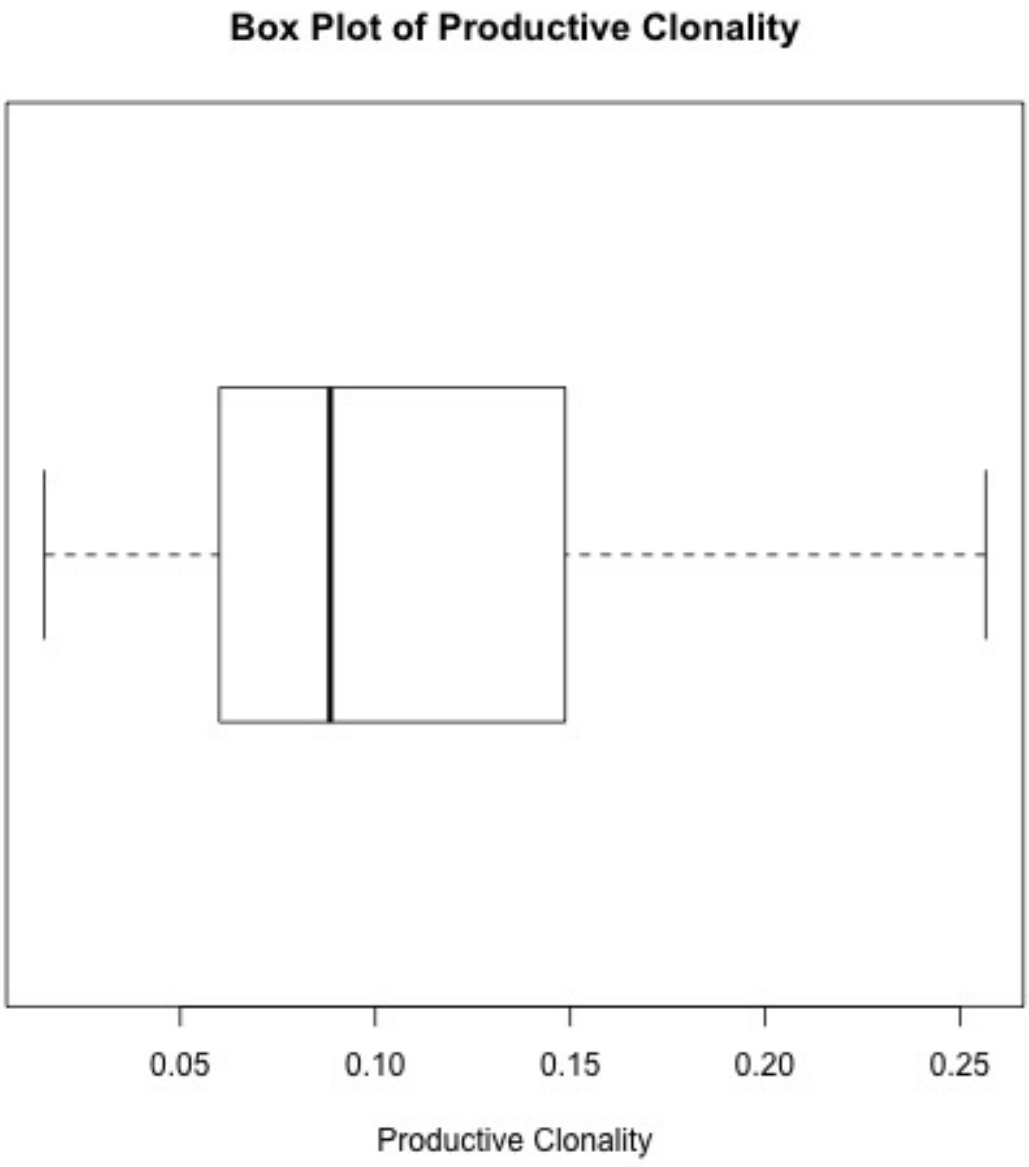

Figure 6: Box Plot of Productive Clonality

Manhattan Plots - SNPs vs Significance

Significant SNPs were identified via `plink`. The `plink` outputs were fed into `R` to generate Manhattan plots, operating under the additive model as discussed above. These visualized the distribution of SNPs as well as preliminary significance levels through `plink` and `R`. The plot (along with its Q-Q plot, Figures 7 and 8) below shows distribution of SNPs along chromosome 7 (the chromosome of interest) as well as the p-values ($-log_{10}(p)$). Higher $-log_{10}(p)$ values indicate greater significance. Additionally, a blue line was arbitrarily plotted at $-log_{10}(p) = 1.30103$ which corresponds to $p = 0.05$. The plot shows a variety of SNPs crossing this threshold. From the raw data, there were 97332 SNPs on chromosome 7. 63571 were removed due to MAF filtering (MAF = 0.10) and 1 was removed due to HWE filtering (HWE = 0.0001). 33760 remained resulting in the plot below.

```
/storage/software/plink --bfile /storage/mips/MIPS_Updated.2019-02-21/data/MIPS_SexCorrected --pheno ..
```

```
manhattan(plinkLinearAll, suggestiveline = -log10(0.05))
```

```
qq(plinkLinearAll$P)
```

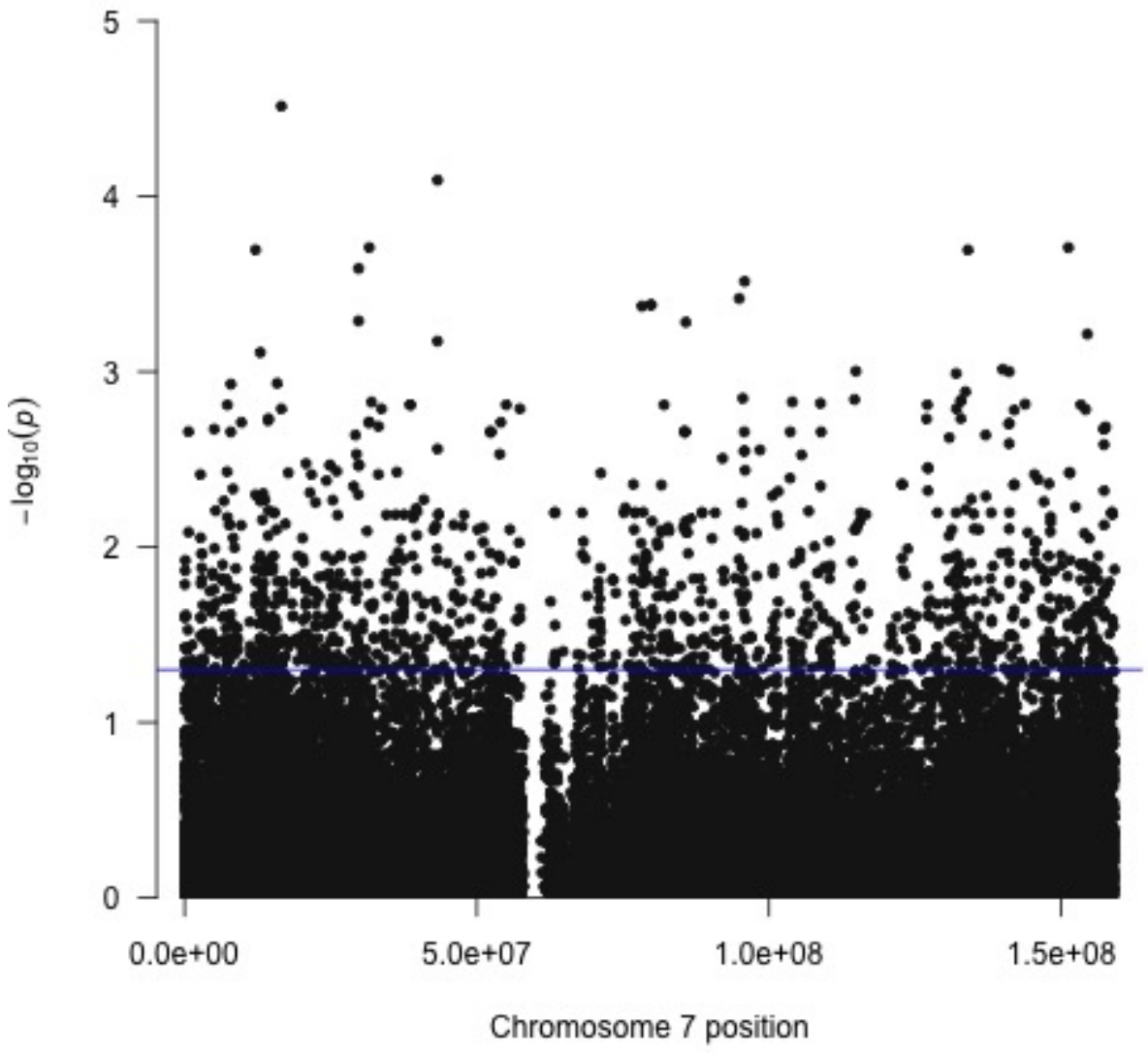

Figure 7: Manhattan Chr 7 All

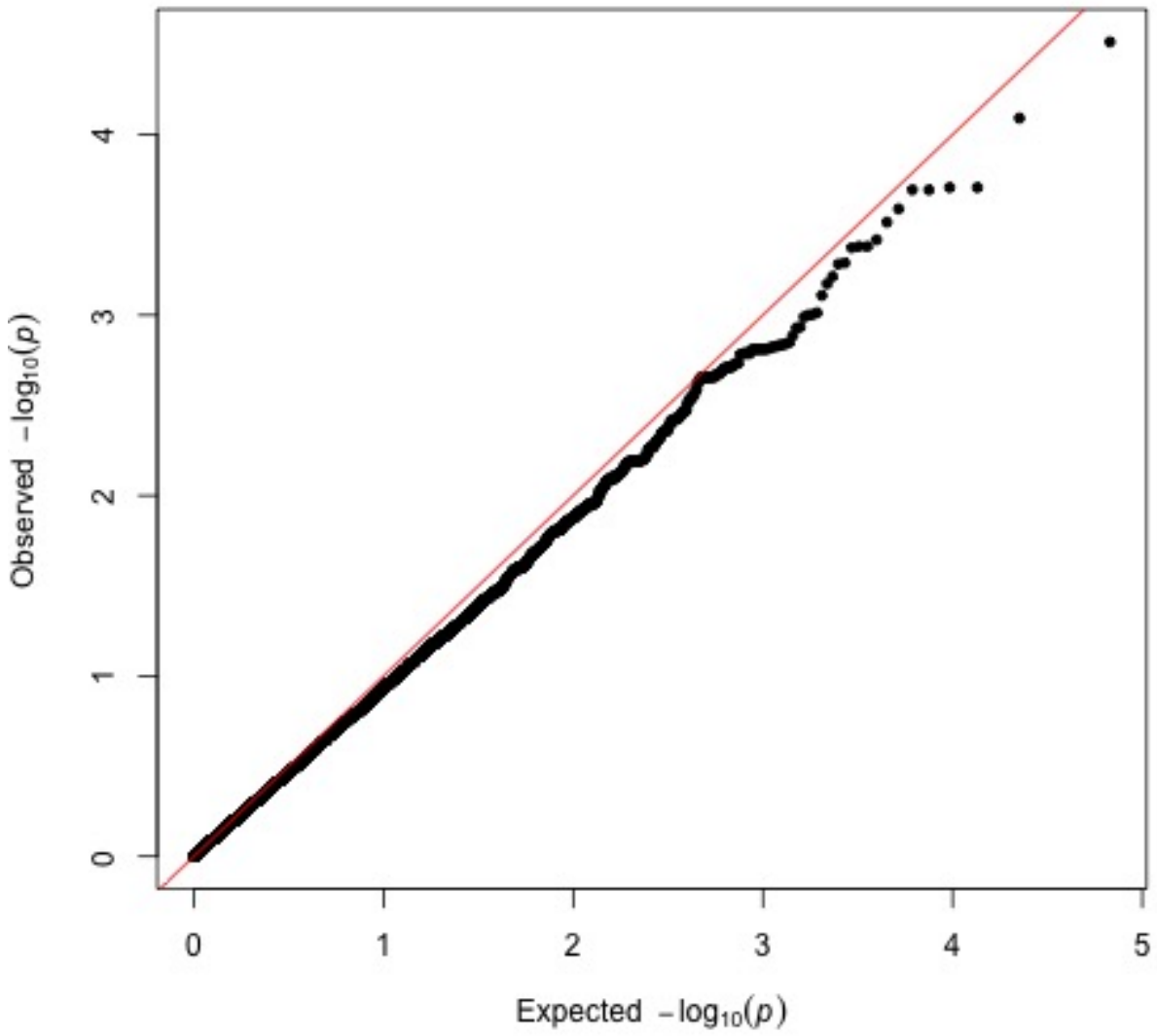

Figure 8: Q-Q Plot All Chr 7

In order to zoom onto the specific region of interest (the TRB locus), one used the coordinates noted above. 654 SNPs reside in this window. 495 SNPs were removed due to MAF filtering (MAF = 0.10) and 1 removed due to HWE filtering (HWE = 0.0001). This resulted in 159 SNPs shown in the below Manhattan plot along with its corresponding Q-Q plot (Figures 9 and 10).

```
/storage/software/plink --bfile /storage/mips/MIPS_Updated.2019-02-21/data/MIPS_SexCorrected --pheno ../
```

```
manhattan(plinkLinear, xlim = c(141948851, 142560972),
          suggestiveline = -log10(0.05), annotatePval = 0.05,
          highlight = as.character(plinkLinear$SNP[which(plinkLinear$P < 0.05)]))
```

```
qq(plinkLinear$P)
```

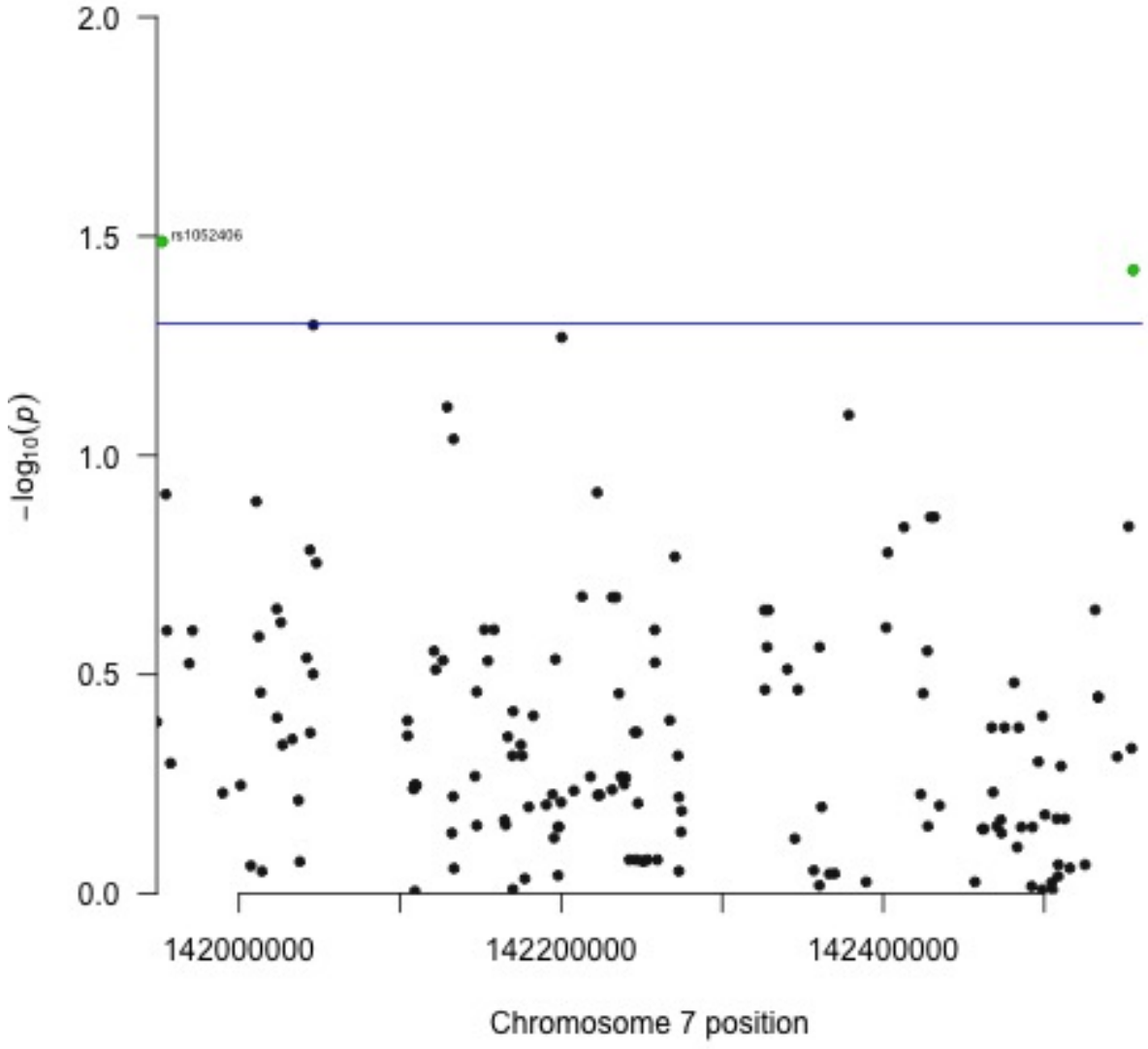

Figure 9: Manhattan TRB

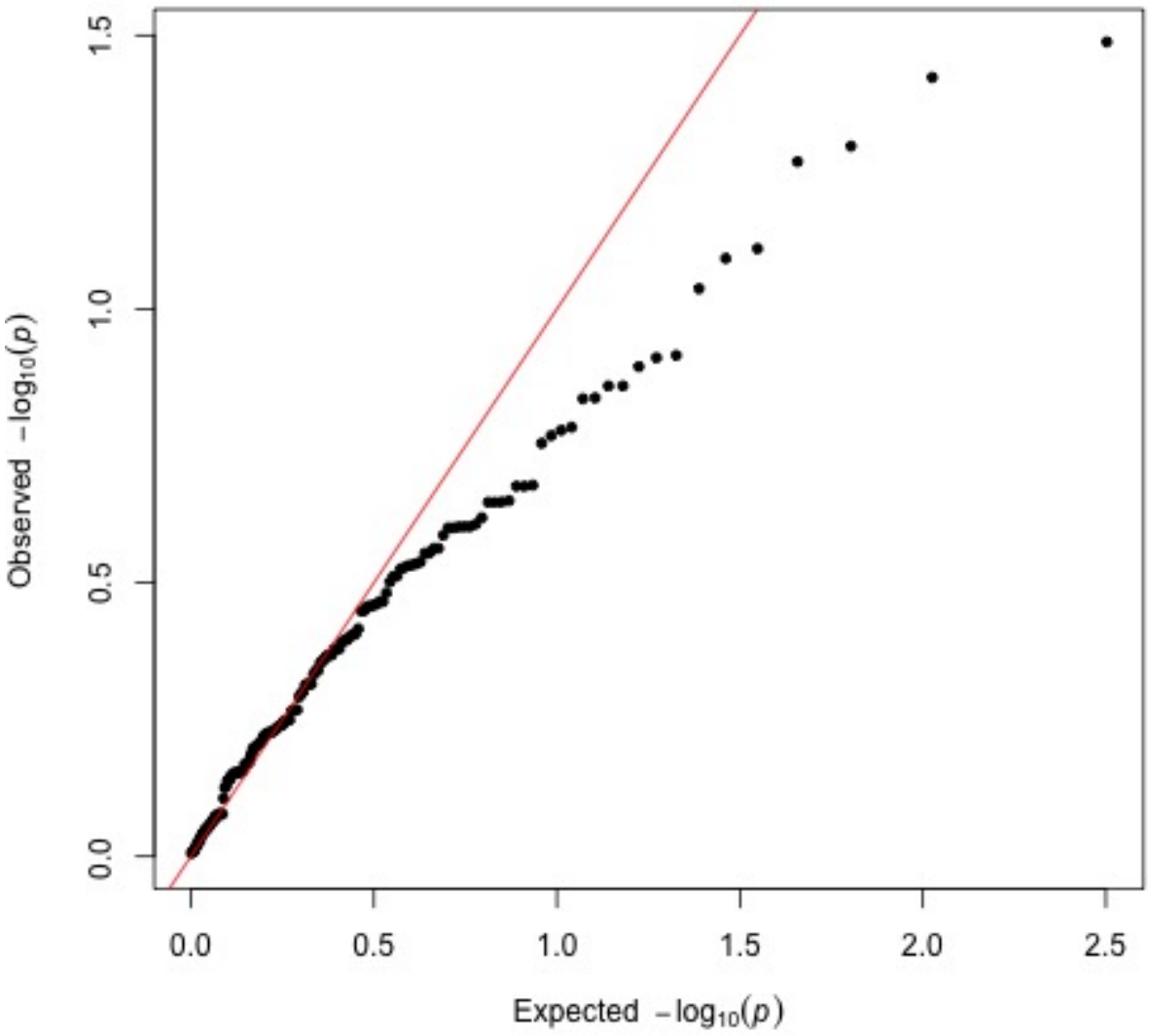

Figure 10: Q-Q Plot TRB

Two notable SNPs are highlighted in green since they cross the significance threshold. These SNPs are:

1. rs1052406
2. rs1009848

Additionally, the Q-Q Plot above shows that there might be a true association between some variants and productive clonality since the points trail off the $X = Y$ line, as is commonly noted in GWAS.

Dimension Reduction Methods: Ridge Regression and Lasso Regression

As noted above, this high dimension setting ($p >> n$) necessitates the use of dimension reduction methods in order to select notable features. Following the filtering as noted above and identification of possible significant SNPs (for subsequent comparison), the set of SNPs within the TRB window amongst the 15 individuals along with their respective productive clonality metrics were modeled through ridge regression and lasso using the `glmnet` package. The `plinkPedJoin` data frame was generated and used as shown below.

	FID	IID	AGE	SEX	phenotype	rs2960763	rs1052406	rs983539	rs1894317	rs7786497	rs9640366	rs57018991	rs114872369	r
1	MIPS001	MIPS001	62	0	0.0625	2	2	1	0	2	2	0	2	
2	MIPS002	MIPS002	59	0	0.0151	2	2	2	2	2	0	2	2	
3	MIPS011	MIPS011	52	1	0.0677	1	2	2	2	1	1	2	2	
4	MIPS012	MIPS012	77	0	0.0222	2	1	1	1	2	2	1	2	
5	MIPS013	MIPS013	67	1	0.2565	1	1	2	2	1	2	2	2	
6	MIPS014	MIPS014	62	1	0.1623	2	1	2	2	2	1	2	2	
7	MIPS015	MIPS015	75	0	0.0575	1	2	2	1	1	2	1	2	
8	MIPS003	MIPS003	41	0	0.0419	2	1	2	2	2	2	2	2	
9	MIPS004	MIPS004	56	0	0.1632	2	1	2	1	2	2	1	2	
10	MIPS005	MIPS005	69	0	0.1298	1	0	2	2	2	2	2	1	
11	MIPS006	MIPS006	62	0	0.1031	2	1	2	2	2	2	2	1	
12	MIPS007	MIPS007	62	0	0.1350	2	1	2	2	2	2	2	1	
13	MIPS008	MIPS008	75	1	0.0628	2	2	1	1	2	2	1	2	
14	MIPS009	MIPS009	65	1	0.1771	2	0	2	2	2	2	2	2	
15	MIPS010	MIPS010	42	1	0.0884	2	1	2	2	2	1	2	2	

Please refer to **The Data Book** for more information on the specific variables in the data structure above.

The above data frame was modeled through `glmnet` using different alpha values (0 for ridge regression and 1 for lasso). 100 different, random lambda values were generate ranging from 10^10 to 10^-2 to use in the regression models. Leave-one-out-cross-validation (LOOCV) was used to select the best lambda values associated with each type of regression model. LOOCV was used as the validation method due to the small sample size.

```
# lambda values to test
grid <- 10^seq(10, -2, length = 100)

# Ridge regression, alpha = 0
ridge.fit <- glmnet(obs, resp, alpha = 0, lambda = grid)

...

# LOOCV since nfolds = number of observations
cv.ridge <- cv.glmnet(obs, resp, alpha = 0, nfolds = 15, grouped = FALSE)

...
# Lasso, alpha = 1
lasso.fit <- glmnet(obs, resp, alpha = 1, lambda = grid)
...
# LOOCV since nfolds = number of observations
cv.lasso <- cv.glmnet(obs, resp, alpha = 1, nfolds = 15, grouped = FALSE)
```

Results

Refinement of the Data Set

As shown above, the initial number of SNPs on chromosome 7 was 97332. After updating for TRB coordinates, filtering for MAF (MAF = 0.10) and HWE (HWE = 0.0001), the number of SNPs drops dramatically to 159 SNPs. 3 SNPs were subsequently removed due to genotyping missingness in some individuals, resulting in the number of SNP predictors, $p = 156$ (amongst 15 individuals, $n = 15$)

Significant SNPs and their linear regressions

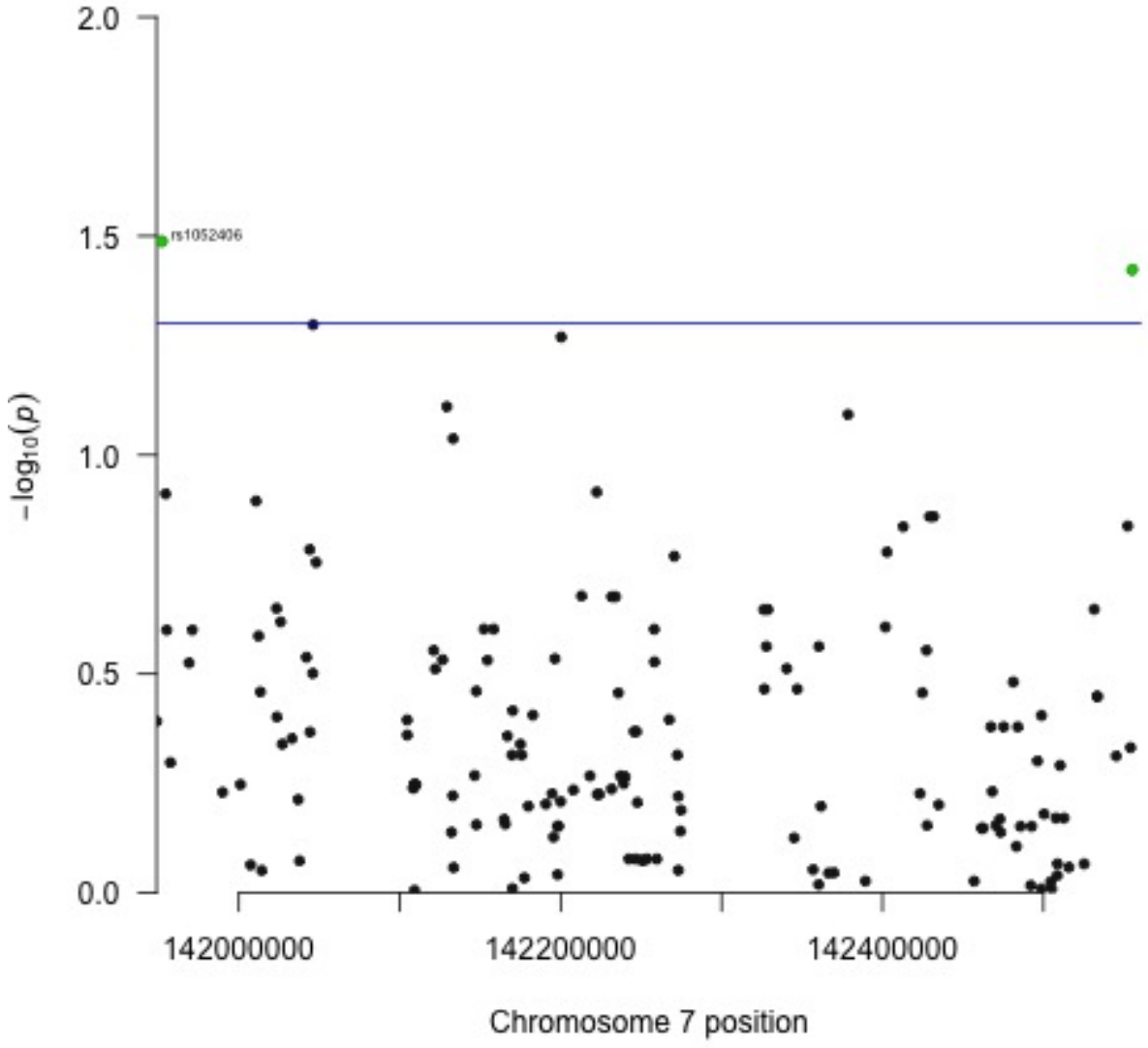

Figure 11: Manhattan TRB

	CHR	SNP	BP	A1	TEST	NMISS	BETA	STAT	P
2	7	rs1052406	141952110	T	ADD	15	0.0547400	2.39300	0.03251
159	7	rs1009848	142555251	C	ADD	15	-0.0564700	-2.31300	0.03775

Figure 12: plink Output of Significant SNPs in TRB window

As noted in the EDA above, Figure 11 shows these 2 significant SNPs (highlighted in green) with the output from `plink` in Figure 12.

These SNPs are on the edges of the window defined.

Initial Modeling Evaluations: Ridge vs Lasso

Figure 13 shows values of coefficients related to SNPs as lambda changes with ridge regression. All SNPs are used in this model.

```
plot(ridge.fit, label = TRUE)
```

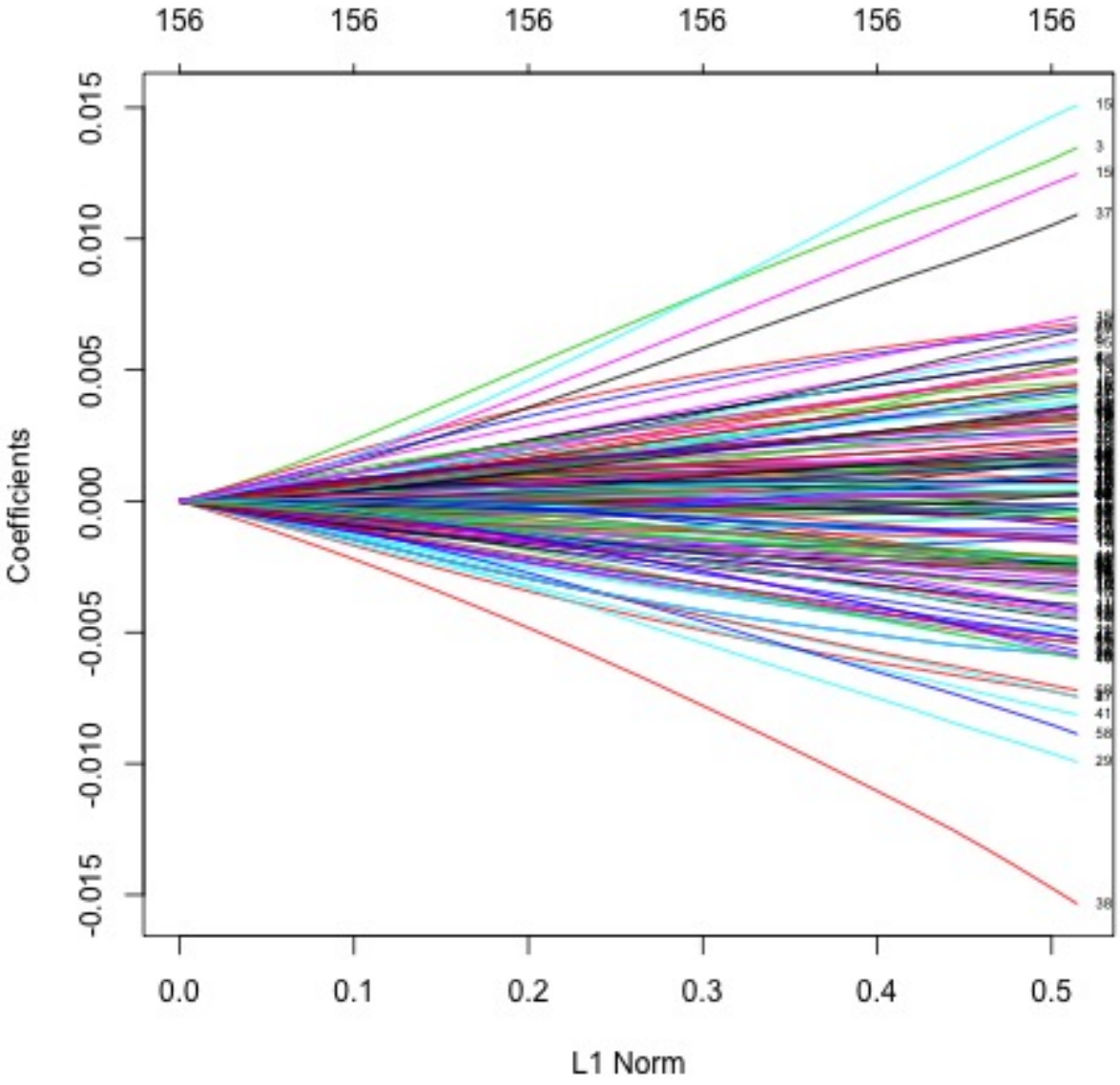

Figure 13: Ridge Coefficients

```r
plot(cv.ridge)
```

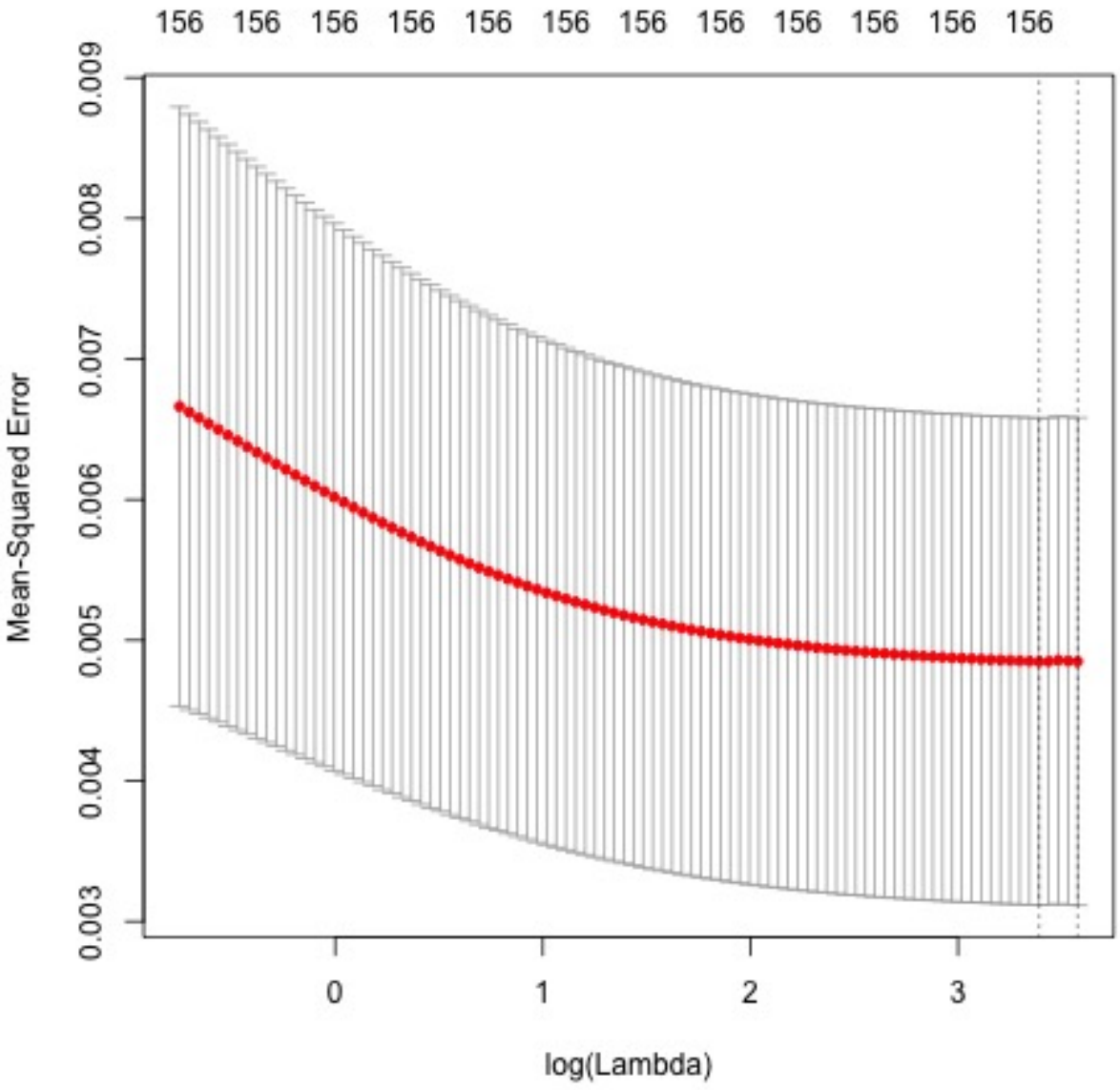

Figure 14: Ridge LOOCV

Figure 14 shows the results of LOOCV with ridge regression. The best value of lambda in this model is 29.68392, determined from LOOCV.

```
cv.ridge$lambda.min
# 29.68392
```

One should note, the confidence intervals for the coefficients are quite large, indicating weak confidence in the coefficient values obtained.

The values of the coefficients are greatly diminished ranging from -0.01353497 to 0.08316423 in this model with this lambda value. The coefficients are shown below.

```
predict(ridge.fit, type = "coefficients", s = bestLambda)
# (Intercept)      0.0831642288
# rs2960763       -0.0022503971
# rs1052406       -0.0069068121
# rs983539         0.0122147470
# rs1894317        0.0050636154
# rs7786497       -0.0030319377
# rs9640366        0.0055895997
# rs57018991       0.0050739072
# rs114872369     -0.0050839163
# rs56095598      -0.0021901136
# rs7357130        0.0016386777
# rs4726517       -0.0047615570
# rs2960771        0.0027608587
# rs361401         0.0013032186
# rs965274         0.0007451834
# rs17162994       0.0043427294
# rs13232310      -0.0001189795
# rs2960760        0.0012539590
# rs867882340     -0.0006434213
# rs2156936       -0.0004098765
# rs6951080        0.0006408840
# rs10243101       0.0013127907
# rs2855836       -0.0010786486
# rs2011381       -0.0053363080
# JHU_7.142044277  0.0024258613
# kgp9647465       0.0021437952
# JHU_7.142046090  0.0064136074
# JHU_7.142048057  0.0039969771
# rs361433        -0.0044584092
# JHU_7.142104666 -0.0089882814
# JHU_7.142108486  0.0030969482
# JHU_7.142108564  0.0031429527
# rs6949980       -0.0021933605
# JHU_7.142109203  0.0015365131
# JHU_7.142109973 -0.0021415539
# rs139219715      0.0035081909
# rs6958838       -0.0024360443
# rs2226967        0.0097586708
# rs56260489      -0.0135349727
# rs111223175     -0.0003961989
# JHU_7.142132963 -0.0006935546
# JHU_7.142133123 -0.0074621998
# rs2367199       -0.0026032797
```

```
# rs7778566        -0.0005676552
# JHU_7.142147534   0.0012916953
# JHU_7.142147606 -0.0002844838
# rs2011310        -0.0048437078
# rs2734166        -0.0067980139
# rs2734165        -0.0048494181
# rs2855961         0.0012387044
# rs55641081        0.0047576330
# rs2011726        -0.0022817469
# rs2255150        -0.0025616804
# rs111221772      -0.0018847738
# rs2855958         0.0029675260
# JHU_7.142175005   0.0013491462
# rs2734153        -0.0025354739
# rs1882723         0.0046839109
# rs2734190        -0.0078829870
# rs79317402       -0.0050182387
# rs17232          -0.0035616415
# rs10260565       -0.0001464879
# rs4726528         0.0016066723
# rs2040366         0.0015689781
# JHU_7.142198061   0.0001848580
# rs17276          -0.0021018853
# rs6976636         0.0001609631
# rs73546368        0.0058678516
# rs11768792       -0.0066262122
# JHU_7.142207944   0.0028676604
# rs145823082       0.0062383772
# rs73544570        0.0003629281
# rs6961143         0.0015681003
# rs6959460        -0.0005600366
# rs361467         -0.0005678424
# rs17249           0.0008498656
# rs17248          -0.0056282504
# rs111221709      -0.0056169664
# rs2367191         0.0006894592
# rs2734118         0.0003677104
# rs11768398       -0.0039051798
# rs17209           0.0037086074
# rs2734112         0.0002485240
# rs17304           0.0015592569
# rs2855907         0.0002709589
# JHU_7.142247124   0.0015720473
# rs361358          0.0009467531
# rs361437         -0.0003294631
# rs361436         -0.0003381843
# rs2734077         0.0002668957
# rs61432115        0.0033874573
# rs2855896        -0.0023473821
# rs55693295        0.0002628172
# rs2213187         0.0026019709
# rs144142335      -0.0015052152
# rs2854546         0.0054564864
```

```
# rs361489         -0.0051379887
# rs361488          0.0000592324
# rs2854538         0.0019568269
# rs2854536        -0.0020454538
# rs6464507         0.0007294790
# rs17260           0.0008436647
# rs13239736       -0.0029907876
# rs6974518         0.0007570969
# rs10215447        0.0021862277
# rs76491280       -0.0053896594
# rs6943492         0.0008128059
# rs2078176         0.0038785892
# JHU_7.142360311   0.0017720363
# rs2013987        -0.0030183860
# JHU_7.142366506 -0.0019496296
# rs6975391        -0.0025167364
# rs17243          -0.0047553146
# rs6968949         0.0003983333
# rs6946770        -0.0038241971
# JHU_7.142402769 -0.0036866787
# rs117890692      -0.0049800978
# rs17251           0.0001937589
# rs4726571        -0.0036748879
# rs556093563       0.0034645922
# rs78648534        0.0030122364
# rs10241932        0.0029320471
# rs374512173       0.0029062041
# rs6959895         0.0013935910
# rs10273639        0.0038351406
# rs1811091        -0.0028281665
# rs1985888        -0.0027922178
# JHU_7.142467438   0.0030689814
# rs1969595         0.0045642918
# rs73742423        0.0029917742
# rs111946127      -0.0013700902
# rs34118966        0.0030279268
# exm2245131        0.0046376047
# JHU_7.142483223   0.0017777363
# JHU_7.142484097   0.0029998289
# JHU_7.142485771 -0.0005543025
# rs2367486         0.0012984734
# rs867750397      -0.0005503117
# rs12539089       -0.0031021700
# rs1799887         0.0032218739
# rs1042955        -0.0013201148
# rs11327          -0.0019760532
# JHU_7.142504533 -0.0050640038
# rs3134906         0.0006713413
# rs57055421       -0.0023603611
# rs114143077      -0.0041047327
# rs361462         -0.0013471428
# rs7805607        -0.0023111457
# rs17163745       -0.0012087708
```

```
# rs10268022      -0.0040121815
# rs17722379       0.0112578216
# rs6965992        0.0040717127
# rs6966006        0.0040564396
# rs2063993       -0.0032277199
# rs7789029        0.0025602531
# rs8177107        0.0136449685
# rs1009848        0.0064349913
```

Figure 15 (similar to Figure 13) shows that lasso results in a great reduction in the number of SNPs used.

```
plot(lasso.fit, label = TRUE)
```

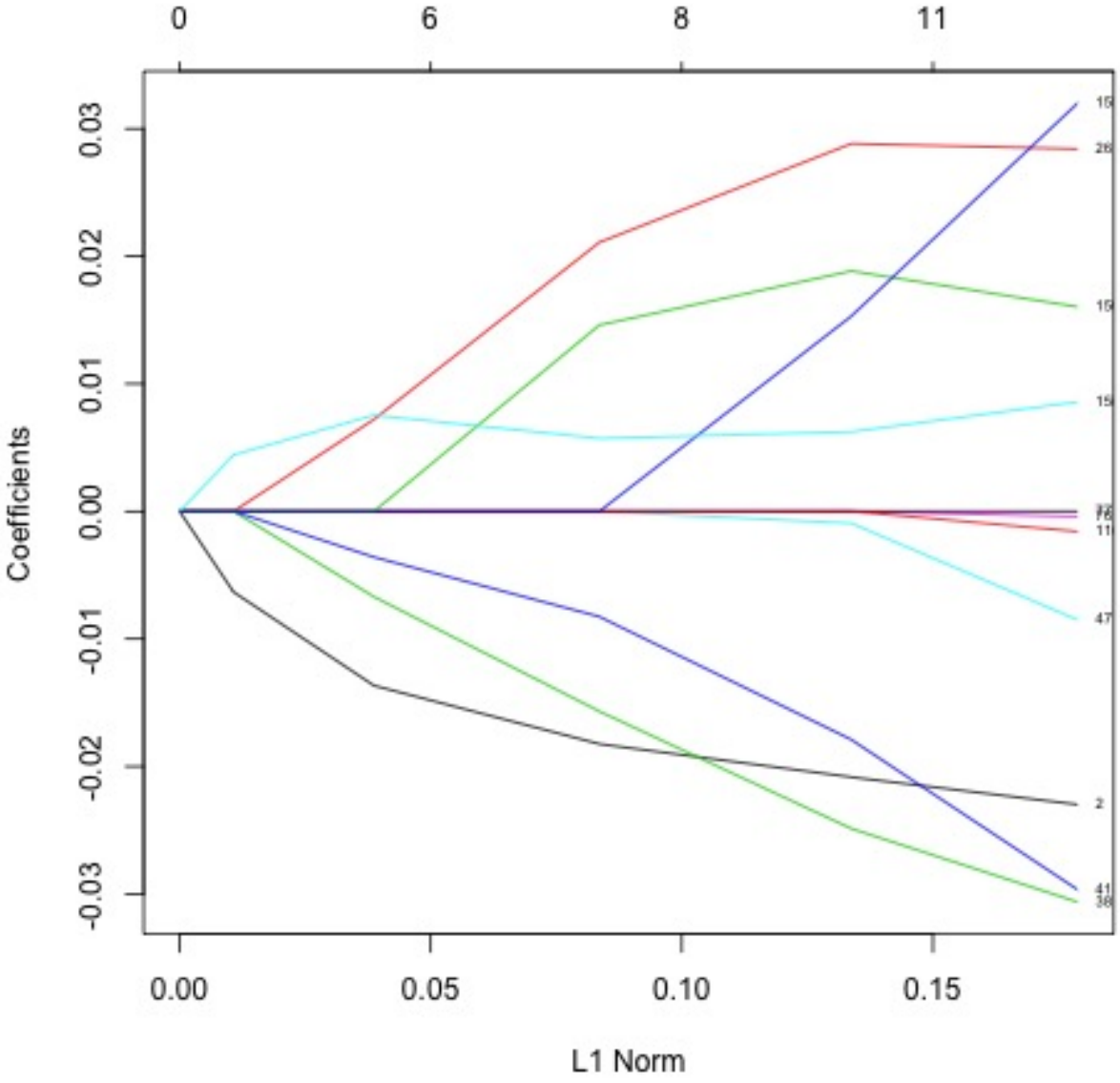

Figure 15: Lasso Coefficients

When performing LOOCV with lasso (see Figure 16), the best value of lambda was 0.03575439 and two SNPs were identified in the model. Their coefficient values are presented below.

```
bestLambda <-  cv.lasso$lambda.min
# 0.03575439
```

SNP	Coefficient
rs1052406	-0.002969940
rs1009848	0.002077136

These two SNPs are consistent with the SNPs identified in Figure 11. It should be noted that the confidence intervals in this plot are also quite large.

```
plot(cv.lasso, label = TRUE)
```

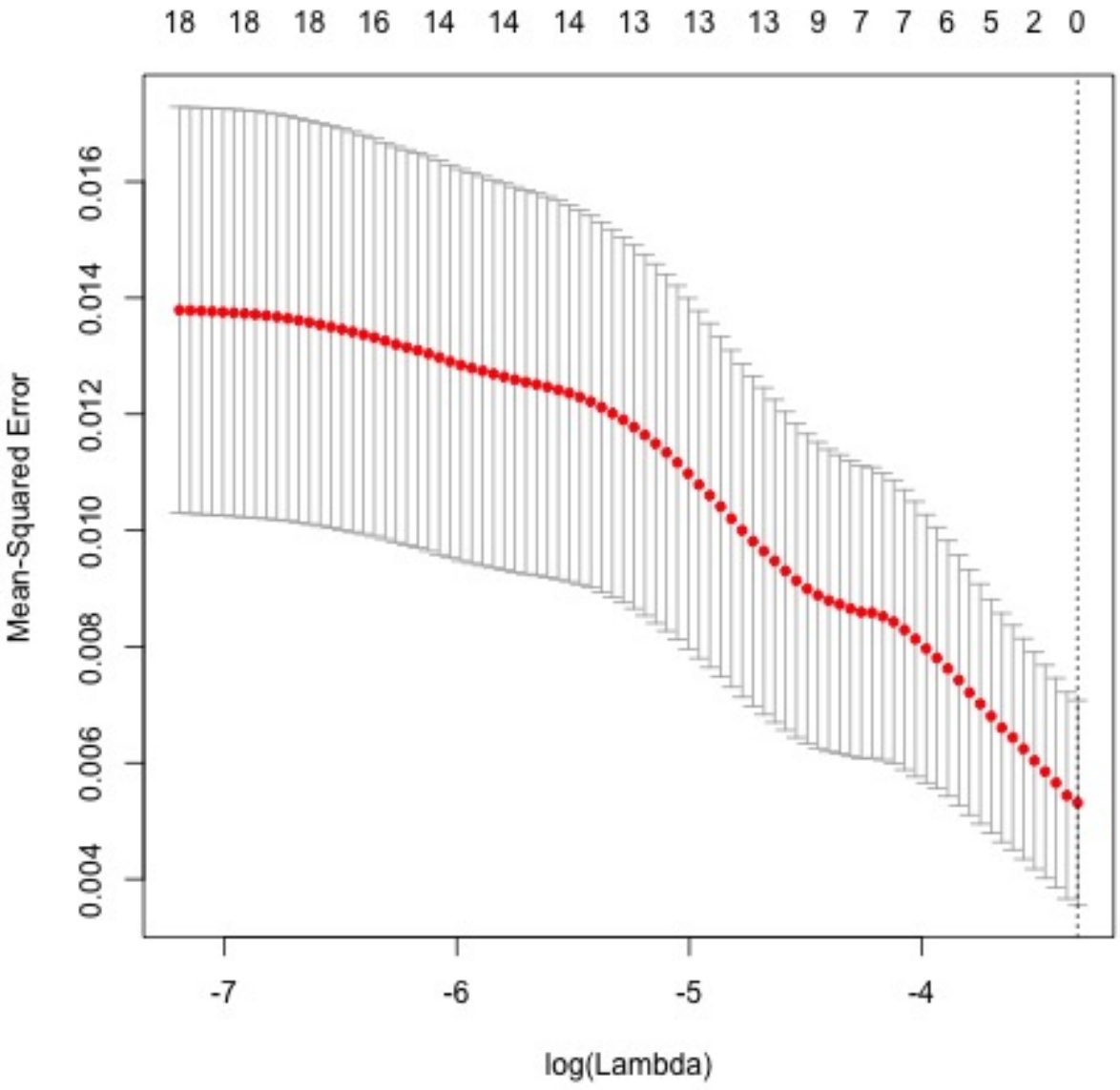

Figure 16: Lasso LOOCV

The Data Book

File Type	Variable	Units	Description
plinkPedJoin	FID	N/A	The Family ID (same as IID in this case)
plinkPedJoin	IID	N/A	The Individual ID
plinkPedJoin	AGE	years	The age of the patient at the time of specimen collection
plinkPedJoin	SEX	N/A	1: Male, 2: Female, 0: Unknown
plinkPedJoin	phenotype	N/A	Productive Clonality (range is 0 to 1)
plinkPedJoin	rsXXX	N/A	Dosage value for given SNP under additive model (0, 1, or 2)
BED	N/A (binary)	N/A	Contains genotype data, if readable, each row is a patient and each column is a SNP with genotype values in cells
FAM	FID	N/A	The Family ID (same as IID in this case)
FAM	IID	N/A	The Individual ID
FAM	PID	N/A	The Patient ID (N/A in this case, all 0)
FAM	MID	N/A	The Maternal ID (N/A in this case, all 0)
FAM	Sex	N/A	1: Male, 2: Female, 0: Unknown
FAM	Phenotype	N/A	-9 is for phenotype missing otherwise productive clonality
BIM	CHR	N/A	The chromosome number of SNP location
BIM	ID	N/A	SNP identifier
BIM	GD	centimorgans	genetic distance
BIM	position	base pairs	genetic location on chromosome
BIM	allele1	N/A	nucleotide of minor allele
BIM	allele2	N/A	nucleotide of major allele
.assoc.linear	CHR	N/A	The chromosome number of SNP location
.assoc.linear	SNP	N/A	SNP identifier
.assoc.linear	BP	base pairs	genetic location on chromosome
.assoc.linear	A1	N/A	nucleotide of minor allele
.assoc.linear	CHR	N/A	The chromosome number of SNP location
.assoc.linear	TEST	N/A	type of test done (additive, dominance, genotype 2df)
.assoc.linear	NMISS	N/A	number of nonmissing obs
.assoc.linear	BETA	N/A	the least square coefficient
.assoc.linear	STAT	N/A	t-statistic
.assoc.linear	P	N/A	the p-value (significance)

Discussion

Conclusion thus far

The small sample size of this study should be acknowledged, but also the exploratory rather than predictive aims should be emphasized. From the above, the TRB locus window, MAF and HWE filtering work together to reduce the number of SNPs significantly to 156 SNPs to be used in subsequent analyses. Initial modeling attempts through ridge regression and lasso regression produced two models, neither of which are particularly strong given the large confidence intervals associated with generated coefficient values. Optimal lambda values (tuning parameter) for each model were obtained through LOOCV due to the small sample size of this study (n = 15). Interestingly, the two SNPs identified via Figure 11 (rs1052406 and rs1009848) are also the two SNPs identified to be relevant predictors via lasso regression. Current challenges are lack of power and unable to produce a test set due to the small sample size.

Next Steps and Improvements

Future work will involve:

1. Investigation of other Methods for Dimension Reduction. Primarily PCA techniques and elastic-net methods will be used.
2. Refinement of Model Evaluation. Information criterions will be produced to more accurately and objectively assess models.
3. Additional covariates such as age, sex, and genetic ancestry should be included to increase power, since it is know that immune system weakens with age and these are typical covariates in GWAS.
4. The SNPs noted above should be investigated for biological signficance.

Acknowledgements

- I would like to thank Dr. Dana Crawford and Dr. William Bush for their continued guidance throughout this project.
- Additionally, I would like to thank Tyler Kinzy for his support and instruction.

References

Andrews, Christine A. 2010. "The Hardy-Weinberg Principlee." https://www.nature.com/scitable/knowledge/library/the-hardy-weinberg-principle-13235724.

Bailey, J.N.C., D.C. Crawford, A. Goldenberg, A. Slaven, J. Pencak, M. Schachere, W.S. Bush, J.R. Sedor, and J.F. O'Toole. 2018. "Willingness to participate in a national precision medicine cohort: Attitudes of chronic kidney disease patients at a Cleveland public hospital." *Journal of Personalized Medicine* 8 (3): 1–11. https://doi.org/10.3390/jpm8030021.

Biotechnologies, Adaptive. 2017. "Understanding the immunoSEQ Assay: From Inquiry to Insights (2019-01-31)." http://adaptivebiotech.com/wp-content/uploads/2019/01/Understanding-the-immunoSEQ-Assay-From-Inquiry-to-Insigl pdf.

Crawford, Dana C, Jessica N Cooke Bailey, Kristy Miskimen, Penelope Miron, Jacob L Mccauley, John R Sedor, John F O Toole, et al. 2018. "Somatic T-cell Receptor Diversity in a Chronic Kidney Disease Patient Population Linked to Electronic Health Records Institute for Computational Biology , Departments of 2 Population and Quantitative Health Sciences and 3 Genetics and Genome Sciences , Ca." *AMIA Jt Summits Transl Sci Proc.* 2017: 63–71.

James, Robert Gareth, Witten Daniela, Hastie Trevor, and Tibshirani. 2013. *An Introduction to Statistical Learning.* 7th ed. New York: Springer.

NCBI. 2019. "TRB T cell receptor beta locus [Homo sapiens (human)]." https://www.ncbi.nlm.nih.gov/gene/6957.

Tabangin, Meredith E, Jessica G Woo, and Lisa J Martin. 2009. "The effect of minor allele frequency on the likelihood of obtaining false positives." *BMC Proceedings* 3 (S7): 5–8. https://doi.org/10.1186/1753-6561-3-s7-s41.