Final Report: Identification of Signficant SNPs associated with T-Cell Receptor Diversity

John Lin, Case Western Reserve University

May 02, 2019

Contents

Abstract **2**

Introduction **3**
Motivation . . . 3
SNPs and Genotypes . . . 3
TCR Diversity . . . 4
The Additive Model . . . 5
SNP Filtering: T-Cell Receptor Beta Locus (TRB) . . . 7
SNP Filtering: Minor Allele Frequency (MAF) . . . 8
SNP Filtering: Hardy-Weinberg Equilibrium (HWE) . . . 8
Population Stratification . . . 10
High Dimensional Considerations . . . 10

Data Science Methods **11**
Cleaning and Exploratory Data Analysis . . . 11
Modeling . . . 12
Productive Clonality Modeling . . . 12
Genetic Ancestry and Population Stratification . . . 14

Results **16**
Cleaning and Exploratory Data Analysis - Results . . . 16
Modeling - Results . . . 32
Initial Modeling . . . 32
Addition of Age and Sex Covariates . . . 35
ADMIXTURE and plink - Predicting Genetic Ancestry/Population Stratification . . . 37
Final Modeling and Comparisons . . . 42

Discussion **50**
Conclusion . . . 50
Limitations . . . 51

The Data Book **51**

R Packages **52**

Acknowledgements **53**

References **53**

Part a) Define Question • Background on the research area and critical issues • Define the question • Define the ideal data set • Determine what data you can access • Define critical capabilities and identify packages you will draw upon • Obtain the data, define you target data structure • Clean and tidy the data

Part b) Cleaning and EDA • Write you databook, defining variables, units and data structures • Data visualization and exploratory data analysis • Observations of trends and functional forms • Power transformations

• Validate with reference to domain knowledge • Evaluate the types of Modeling Approaches to take

Part c) Modeling and Statistical Learning • Types of modeling to try • Statistical prediction/modeling • Model selection • Cross-validation, Predictive R2 • Interpret results • Challenge results

Part d) Present your final models and learnings • Present your results • Present reproducible code • Comparison to other modeling approaches in the literature

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, potentially 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 of TCR repertoires is not well studied. Preliminary results presented in this report show that SNP rs1052406 may be a significant predictor of TCR. While attempts to model TCR based on sex, age, specific SNPs and genetic ancestry were made, a finalized model could not be defined. Despite this, throughout the process of exploring the data, multiple data science methods were used.

Please note this work is in conjunction with Case Western Reserve University and is operating under the CC-BY-SA 4.0 License, currently. The code used to perform the data analysis is located on GitHub here. Also, the exploratory rather than predictive aims of the below should be emphasized.

The following open source tools were used throughout the course of this project:

1) `plink`, a whole genome analysis toolkit (Purcell et al. (2007))

2) `snpfip`, a program to reverse SNP orientations (Cole and Stovner (n.d.))

3) `ADMIXTURE`, a program to predict genetic ancestriess (DH, J, and K (2009))

Introduction

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 J. Bailey et al. 2018; D. C. 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) (J. 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, D. C. Crawford et al. (2018) selected 15 of these samples ranging in CKD status to have T-cells sequenced by immunoSeq Adaptive Biotechnologies (Biotechnologies 2017). D. C. Crawford et al. (2018) note that there was some correlation between TCR diversity and CKD status. The primary aim of this study is to find potential signals correlated with TCR diversity and secondarily 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 (genome wide association studies) will be used. Additionally, initial modeling approaches using ridge regression and lasso regression are presented.

SNPs and Genotypes

In normal individuals, there are 22 chromosomes (autosomes) and 1 sex chromosome. Humans are diploid organisms, meaning each chromosome contains 2 alleles. One source of genetic variation between humans are SNPs, which are single-base mutations at a given position in the genome (see Figure 1). A SNP must be present at least 1% in the population to be considered a SNP. A single SNP can confer 3 different genetic states or genotypes, according to the nucleotides on the major and minor alleles. For instance, if a given SNP's minor allele is associated with nucleotide A and the major allele with nucleotide G, the 3 different genotypes would be AA, AG, or GG. Depending on the specific genetic transcription and regulation, one specific genotype may be correlated with an altered phenotype than the others.

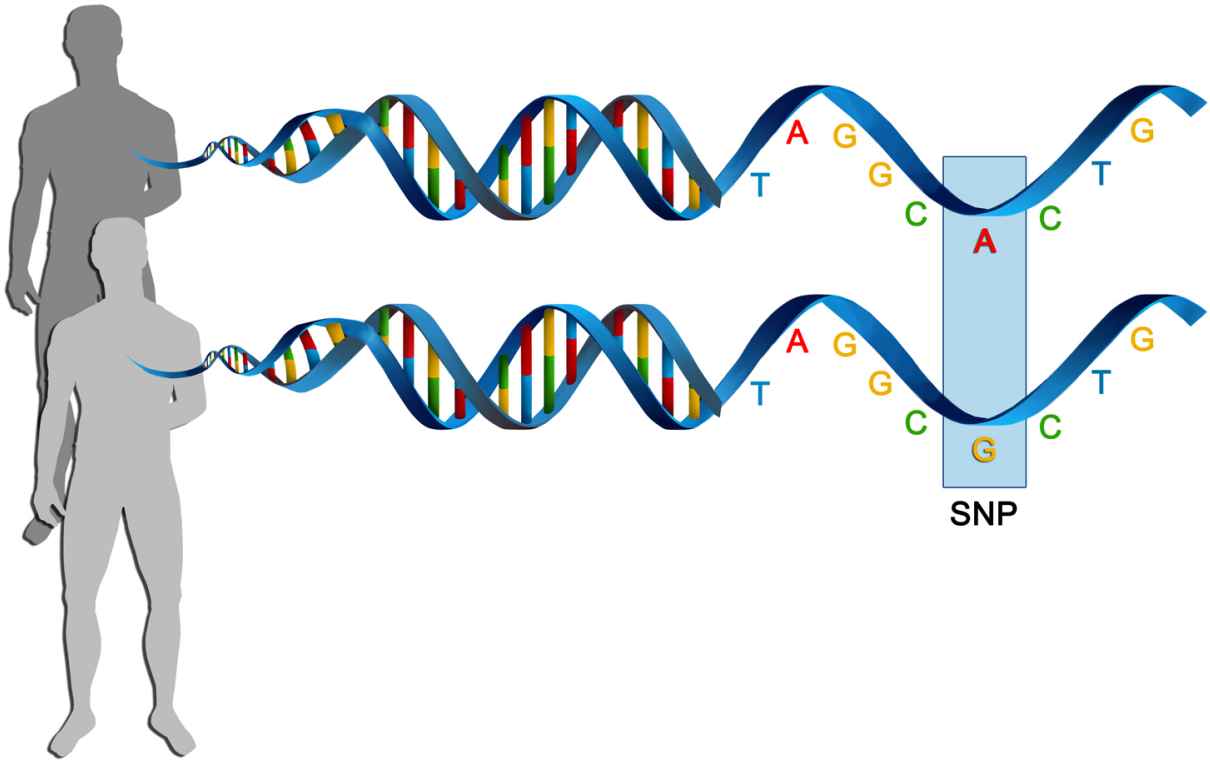

Figure 1: SNP Diagram from (M. 2016)

TCR Diversity

TCR diversity plays a strong role in the adaptive immune system. The adaptive immune system defends the human body against specific pathogens primarily through lymphocytes (B-cells and T-cells). In the context of T-cells, recognition of antigens is mediated through TCR binding to antigens presented by major hisotcompatibility complex class I molecules (MHC1) (D. C. Crawford et al. 2018). The TCR is a transmembrane protein complex that specifies what pathogens are recognized. α and β chains compose the overall structure of 95% of TCRs. In particular, the β chain's CDR3 (Complementarity Determining Region 3) has, historically, played a large role in specifying the TCR's overall structure, and, hence, specificity. At the genetic level, this is determined by a process unique to lymphocytes, known as somatic recombination. Somatic recombination involves a shuffling of V (variable), D (diversity), and J (joining) domains. This V(D)J recombination results in a diverse population of TCRs. This is more elegantly depicted in Figure 2, below.

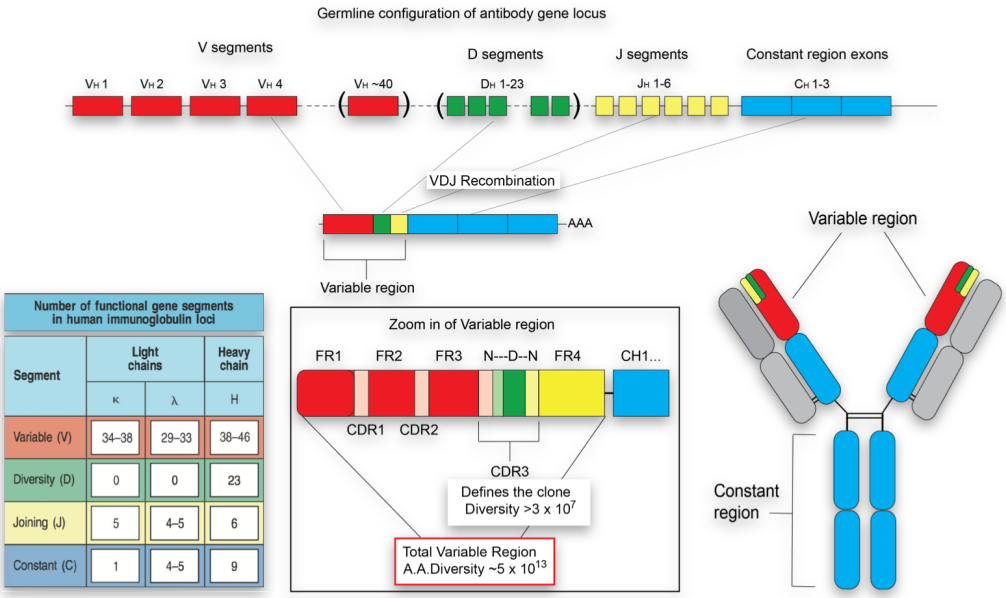

Figure 2: TCR Somatic Recombination from (Zurich 2019)

Therefore, somatic recombination confers the adaptive immune systems ability to defend against a variety of different pathogens. In normal, healthy patients, the TCR repertoire is polyclonal and there are about 10^{13} unique TCR nucleotide sequences. TCR diversity can be measured by productive clonality, which is a Shannon entropy-based measure of clonality. It is calculated as:

$$ProductiveClonality = 1 - \frac{Entropy}{log_2(NumberProductiveUniqueRearrangements)}$$

where entropy is correlated with clone frequency. Values range from 0 (diverse) to 1 (not diverse). A more diverse TCR population is associated with a healthier immune system.

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 3). 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) +
```

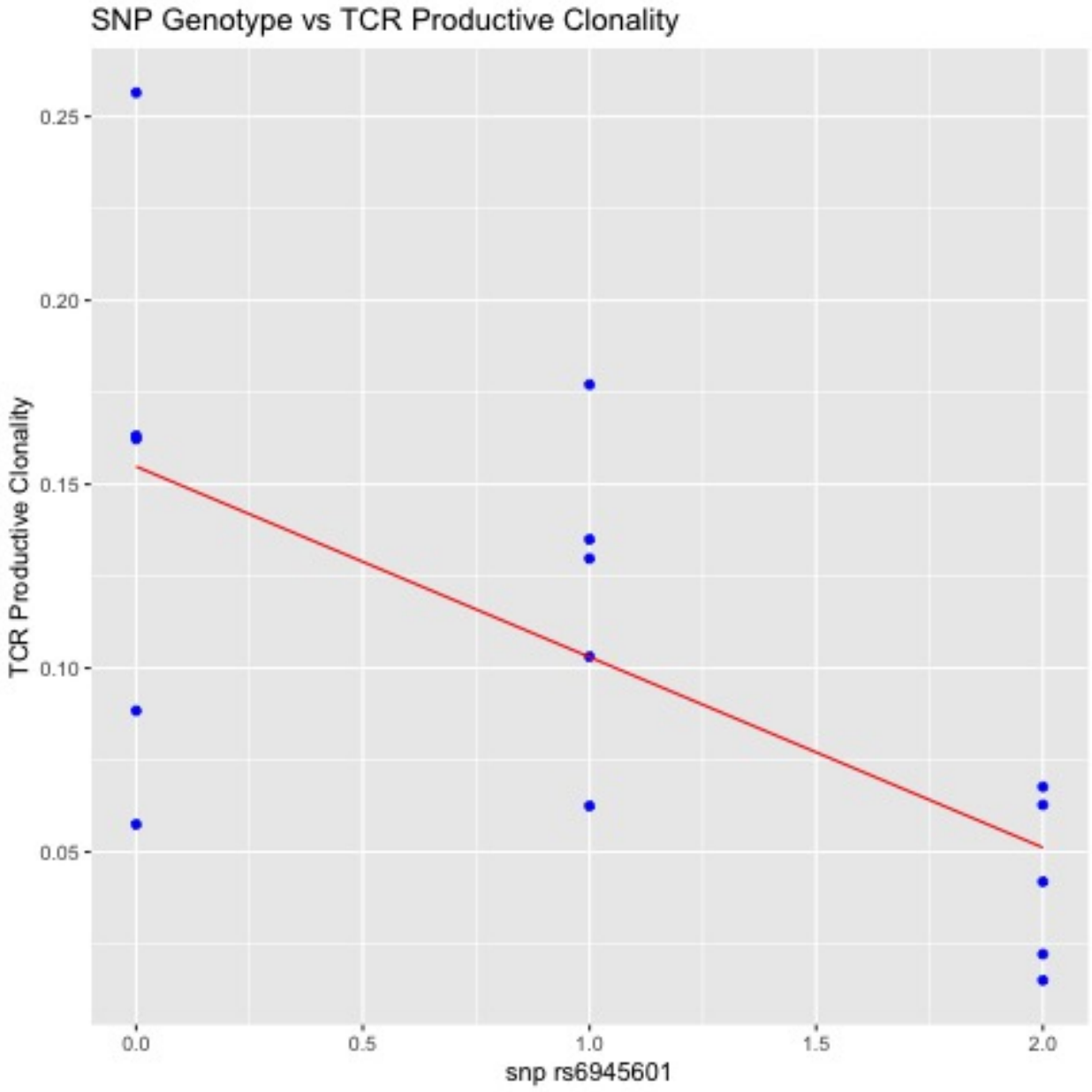

Figure 3: Linear Regression Example

SNP Filtering: T-Cell Receptor Beta Locus (TRB)

In preparation for analysis, there are a variety of different SNP filtering techniques to employ. As a first filtering pass, a genetic window must be defined. 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 4, 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.

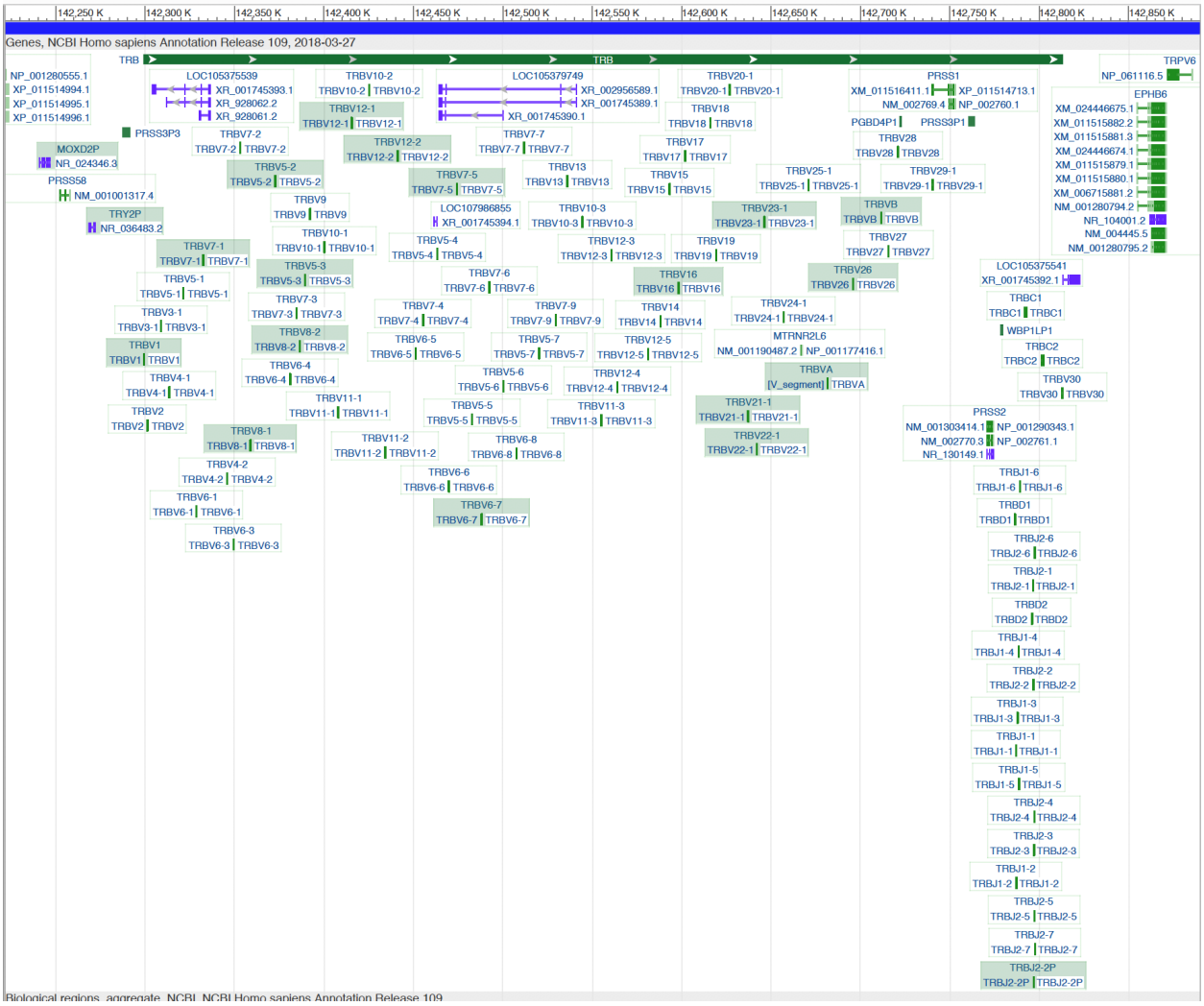

Figure 4: TRB Locus

SNP 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. Some 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 –> $30(.10) = 3$).

SNP 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 5.

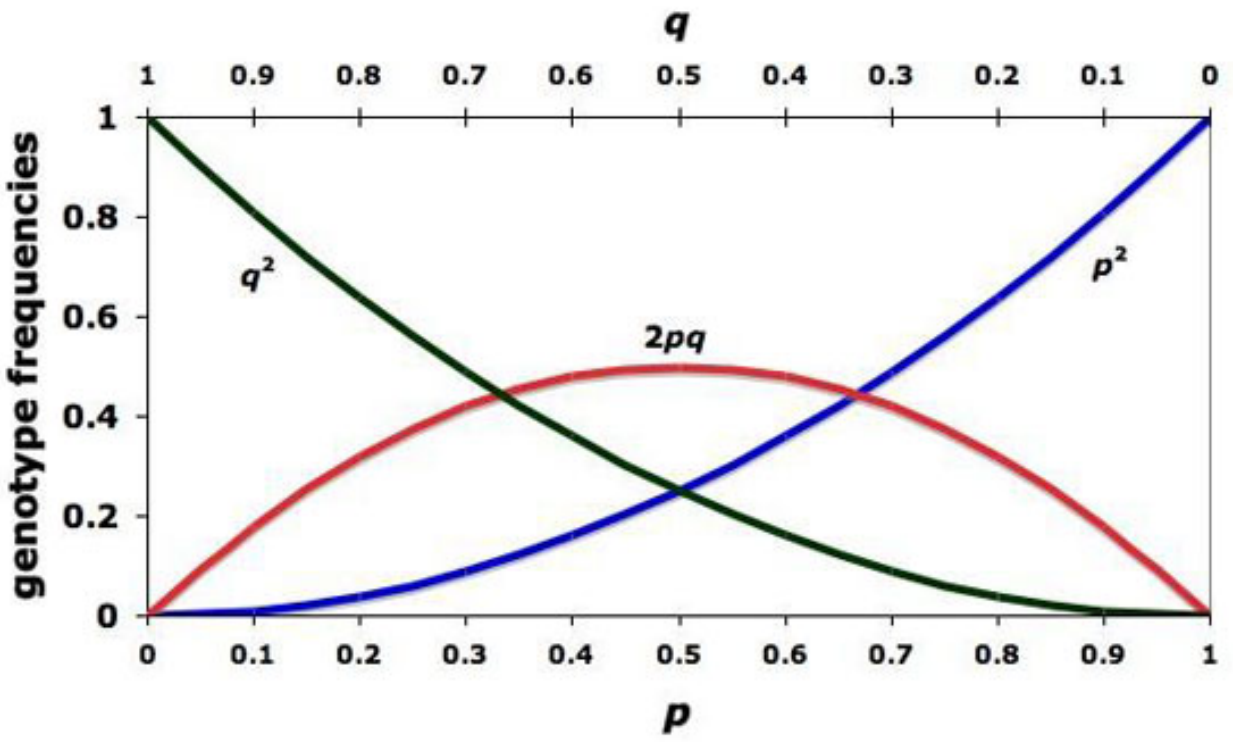

Figure 5: Distribution of Genotype Frequencies in HWE

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).

Population Stratification

Typically, population substructure may contribute to observed disease/trait association (W. S. Bush and Moore (2012)). In order to mitigate the effects of different ancestries contributing to observed effects, ancestries can be measured through common open-source programs like `STRUCTURE`, `EIGENSTRAT`, and `ADMIXTURE`. These measures of ancestry can be incorporated as covariates into a given model in order to account for population stratification.

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).

Data Science Methods

Cleaning and Exploratory Data Analysis

Kluiver et al. (2018) details some common steps in cleaning data in GWAS. The primary aim is to remove confounding variables as well as possible factors relating to sample contamination.

1. Filter SNPs/patients by missingness
2. Sex check
3. Filter by Minor Allele Frequency (MAF)
4. Filter deviations from Hardy-Weinberg Equilibrium (HWE)
5. Filter by heterozygosity
6. Relatedness
7. Account for population stratification

The data was drawn from a cohort of 129 samples. Germline DNA was genotyped on an Illumina MEGA chip. Further processing was done using GenomeStudio (an Illumina software) as well as `plink`. In regards to TCR diversity, 15 of the 129 samples' T-Cells were immuno-sequenced by immunoSeq.

It should be noted that steps 1, 2 and 6 (above) were performed, before data acquisition, by Tyler Kinzy, a research associate in Dr. Jessica Cooke Bailey's research lab at Case Western Reserve University. Genotype data was compared with self-reported sexes.

Initial views of the data were performed using `plink` and `R` to view SNPs along chromosome 7 as well as distributions of productive clonality. Preliminary Manhattan plots were constructed in this phase of the data processing. An example of this code is shown below, where an MAF of 10% and HWE significance level of 0.0001 were the only SNP filters applied to the 15 samples genotype data.

```
/storage/software/plink --bfile
/storage/mips/MIPS_Updated.2019-02-21/data/MIPS_SexCorrected --pheno
../phenotypes/tcrEmrPheno.txt --pheno-name Productive.Clonality --chr 7
--from-bp 141948851 --to-bp 142560972  --maf 0.1 --hwe 0.0001 --linear
--make-bed --out $outputFolder/plinkFiltering/plink5/window --freq counts
--prune
```

```
# Zoomed manhattan with 50,000 bp window on each side
# Read in linear regression results
plinkLinear <- scp_download(session, "/storage/mips/MIPS_Updated.2019-02-21/jxl2059/plinkResults/plinkF
                            plink5/window.assoc.linear", to = "/Users/linjo/GoogleDrive/CaseWesternUniv
plinkLinear <- read.table("/Users/linjo/GoogleDrive/CaseWesternUniversity/
                          tcr-project/data/window.assoc.linear", header = TRUE)
# Remove where snps with P values of NA
plinkLinear <- na.omit(plinkLinear, col = "P")
# Adjust p-values
# p.adjust(plinkLinear$P, method = "bonferroni")
# Generate and save Manhattan plot
jpeg('figures/manhattanZoomed.jpg')
# coordinates for expanded window are 141948851, 142560972
manhattan(plinkLinear, xlim = c(141948851, 142560972),
          suggestiveline = -log10(0.05), annotatePval = 0.05,
          highlight = as.character(plinkLinear$SNP[which(plinkLinear$P <
                                                           0.05)]))
dev.off()
```

As noted in the code above, further refinement of the genetic window correponding to the TRB locus was done in order to reduce the number of SNPs under consideration. Age and sex were also used as covariates in the independent tests for significance of the various SNPs within the TRB locus window.

	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	

Figure 6: plinkPedJoin Data Frame

```
/storage/software/plink --bfile $outputFolder/plinkFiltering/plink5/window
--covar ../phenotypes/tcrEmrPheno.txt  --covar-name SEX,AGE --linear
--make-bed --out $outputFolder/plinkFiltering/plink7/windowCovar
```

Modeling

Productive Clonality Modeling

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 and identification of possible significant SNPs (for subsequent model 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 in `R`. The `plinkPedJoin` data frame was generated and used as shown below in Figure 6.

	phenotype	rs1052406	rs1009848	AGE	SEX	CEU	YRI	PC1	PC2
1	0.0625	2	0	62	0	0.126387	0.873613	0.0995221	0.006113780
2	0.0151	2	0	59	0	0.066900	0.933100	0.1110640	0.002224610
3	0.0419	1	0	41	0	0.164354	0.835646	0.0895487	-0.009758330
4	0.1632	1	0	56	0	0.125628	0.874372	0.0995574	-0.007273350
5	0.1298	0	1	69	0	0.233257	0.766743	0.0764006	-0.008443290
6	0.1031	1	0	62	0	0.120050	0.879950	0.1004450	0.005226350
7	0.1350	1	1	62	0	0.153696	0.846304	0.0930201	-0.003788090
8	0.0628	2	0	75	1	0.382668	0.617332	0.0429192	-0.021396500
9	0.1771	0	1	65	1	0.999990	0.000010	-0.0933614	-0.037770400
10	0.0884	1	0	42	1	0.226394	0.773606	0.0760322	0.000616360
11	0.0677	2	1	52	1	0.997197	0.002803	-0.0940656	-0.041911900
12	0.0222	1	1	77	0	0.999990	0.000010	-0.0927035	-0.016703400
13	0.2565	1	2	67	1	0.782960	0.217040	-0.0468152	0.059513100
14	0.1623	1	0	62	1	0.086026	0.913974	0.1066220	-0.017042900
15	0.0575	2	0	75	0	0.119132	0.880868	0.1017390	0.000568422

Figure 7: Comprehensive Data Frame for Productive Clonality Modeling

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 generated 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)
```

Age, sex, genetic ancestry and principle components were eventually added as covariates in the lasso regressions, above. Utlimately, comparisons between a lasso-admixture model and lasso-pca model were made, primarily using the extracted data and `R`. The aggregated data frame below (coded as `obs` in `R`), shows all the predictors and observations used for this portion of the analysis. See Figure 7.

Therefore, the comprehensive list of response and covariates used in this model is below:

- productive clonality
- rs1052506 genotypes
- rs10009848 genotypes
- Age of patient
- Sex of patient
- Proportion of European ancestry (CEU)
- Proportion of African ancestry (YRI)
- First principle component from PCA
- Second principle component from PCA

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

The lasso-admixture and lasso-pca models were compared by calculating MSE (mean squared error, on the training data due to small sample size) and AIC (Akaike information criterion). `R` code excerpts to perform those calculations are shown below.

```
# Evaluation
# MSE
lasso.pred <- predict(lasso.fit ,s = bestLambda, newx = obs)
mean((lasso.pred - resp)^2)
# [1] 0.003959658
# AIC
tLL <- lasso.fit.best$nulldev - deviance(lasso.fit.best)
k <- lasso.fit.best$df
n <- lasso.fit.best$nobs
AICc <- -tLL + 2 * k + 2 * k * (k + 1) / (n - k - 1)
```

Genetic Ancestry and Population Stratification

Genetic ancestry proportions for the larger cohort of 129 samples were imputed using `ADMIXTURE`. In order to do this, the data was reprocessed and merged with reference populations. Referential genotypes were acquired from the 1000 Genomes Project. The data was input into `plink` to expand the genetic window to chromosomes 1-22 as well as consider only bi-allelic SNPS (SNPS with only 2 states).

```
/storage/software/plink --bfile /storage/mips/MIPS_Updated.2019-02-21/data/
MIPS_SexCorrected --pheno ../phenotypes/tcrEmrPheno.txt --pheno-name
Productive.Clonality --chr 1-22 --maf 0.1 --hwe 0.0001 --biallelic-only strict
--make-bed --out $outputFolder/plinkFiltering/plink8/allWindow --freq counts
```

This produced 540147 SNPs from the sample data.

snpFlip was used to reverse SNP strand orientations in the reference genomes, so that they could be aligned with SNPs in sample data, later.

```
# processing for 129 samples on chr 1-22
snpflip -f ./human_g1k_v37.fasta -b ./plink8/allWindow.bim -o
./snpFlipResults/allWindow_flip
```

After excluding SNPs with ambiguous strand orientations, filtering for bi-allelic SNPs, and filtering for autosomes from the reference populations, the data was merged with the sample data through `plink`. The reference contained 207 individuals of CEU and YRI ancestries (representing 81260710 SNPs).

```
/storage/software/plink -bfile $outputFolder/plinkFiltering/plink12/
allWindowNoDup_flipRef3 -bmerge $referenceFolder/CEU-YRI_unrelFinal2
--snps-only 'just-acgt' -make-bed
--out $outputFolder/plinkFiltering/plink13/allWindowNoDup_flipRef4
```

$$L(Q,F) = \sum_i \sum_j \left\{ g_{ij} \ln \left[\sum_k q_{ik} f_{kj} \right] + (2 - g_{ij}) \ln \left[\sum_k q_{ik} (1 - f_{kj}) \right] \right\}$$

Figure 8: Log-Likelihood

Following this, an MAF filter of 5%, a genotyping rate of 2% and pruning for LD were applied, leaving 62674 SNPS amongst 336 individuals.

```
/storage/software/plink -bfile $outputFolder/plinkFiltering/plink13/
allWindowNoDup_flipRef4 --maf 0.05 --geno 0.02 --indep-pairwise 50 10 0.1
--make-bed --out $outputFolder/plinkFiltering/plink14/
allWindowNoDup_flipRef5

/storage/software/plink -bfile $outputFolder/plinkFiltering/plink14/
allWindowNoDup_flipRef5 --exclude  $outputFolder/plinkFiltering/plink14/
allWindowNoDup_flipRef5.prune.out --make-bed --out $outputFolder/plinkFiltering/
plink15/allWindowNoDup_flipRefPruned
```

Assuming 2 distinct groups (K) in this data set, `ADMIXTURE` was used to process the data and predict genetic ancestries. `ADMIXTURE` uses a global-ancestry (proportion of ancestries over entire genome) and model-based (ancestry coefficients used as parameters of a statistical model) approach. It uses Bayesian statistics and a Monte Carlo Markov Chain model. The goal is to estimate q_{ik} (kth population's contribution fraction to ith person's genome) and f_{jk} (the frequency for allele 1 of jth SNP in kth population) by maximizing the likelihood function below (Figure 8).

`ADMIXTURE` offers fast convergence compared to other methods discussed earlier.

```
# create .pop file
awk '{ if (NR!=1) {print $1 " " $6} }' $ref_dir/CEU-YRI.psam >
$outputFolder/temp.pop

sort $outputFolder/temp.pop -o $outputFolder/temp.pop

sort $outputFolder/allWindowNoDup_flipRefPruned.fam -o
$outputFolder/allWindowNoDup_flipRefPrunedSorted.fam

join -a 1 -a 2 -o0,2.2 -e ' -'
$outputFolder/allWindowNoDup_flipRefPrunedSorted.fam
$outputFolder/temp.pop > $outputFolder/temp2.pop

awk '{ print $2 }' $outputFolder/temp2.pop >
$outputFolder/allWindowNoDup_flipRefPruned.pop


# /storage/software/admixture_linux-1.3.0/admixture
$outputFolder/plink13/allWindowNoDup_flipRef4 --supervised

/storage/software/admixture_linux-1.3.0/admixture -B -s 314161 -j2 --supervised
$outputFolder/allWindowNoDup_flipRefPruned.bed 2
```

Additionally PCA was performed using `plink` using the `-pca` flag. Eigenvalues and eigenvectors were

extracted and fed into `R` in order to visualize principle components.

```
plink -bfile ../plink8/allWindow -pca --make-bed -out ../
```

Results

Cleaning and Exploratory Data Analysis - Results

As noted above, prior to further analysis, sex was checked in the genotype data. Four individuals needed to be removed due to sex discrepancies between inferred sex (from genotype data) and reported sex. This can be more clearly seen in Figure 9 below.

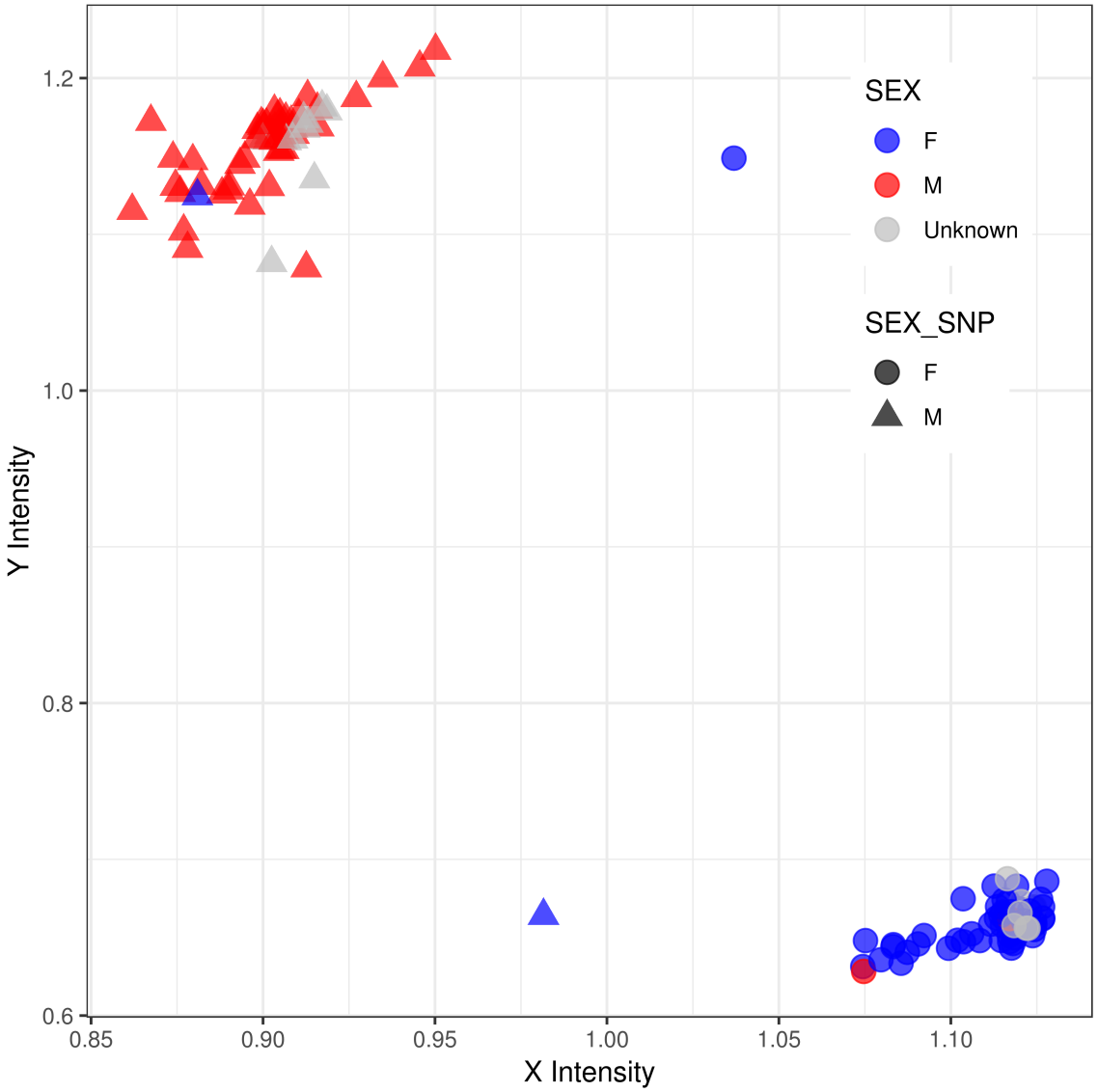

Figure 9: Sex Check

There are 2 distinct clusters of sexes (red for male, blue for female). There are four points that appear to be ambiguous and/or misclassified. A female grouped with the male cluster in the upper left and a male grouped with the females in the bottom right. Additionally, there are two points that show high leverage as they are distant from the other clouds of points.

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 10 and 11) below shows distribution of SNPs within the TRB locus expanded window as well as corresponding 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 $p = 0.05$.

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

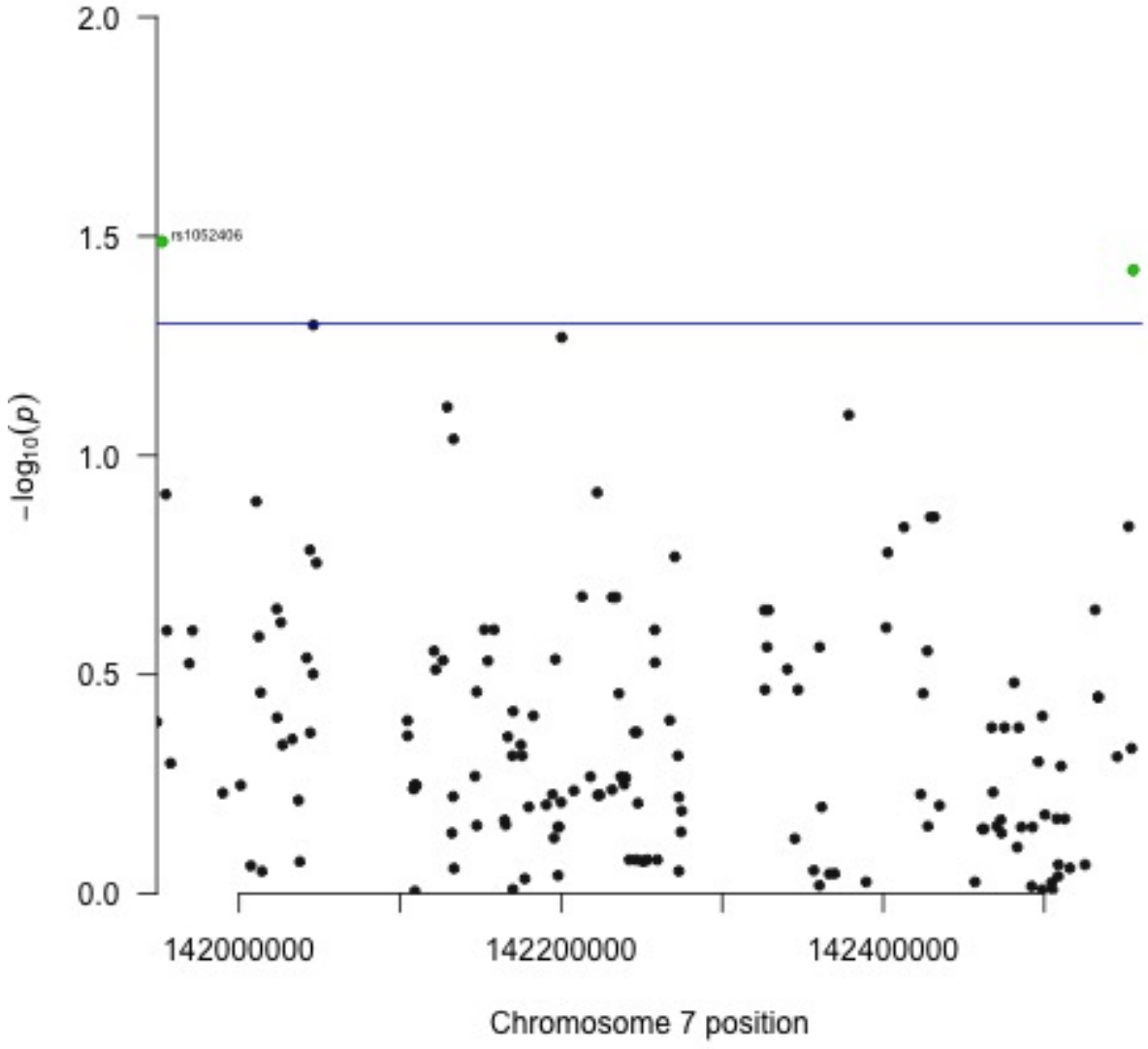

Figure 10: Manhattan TRB

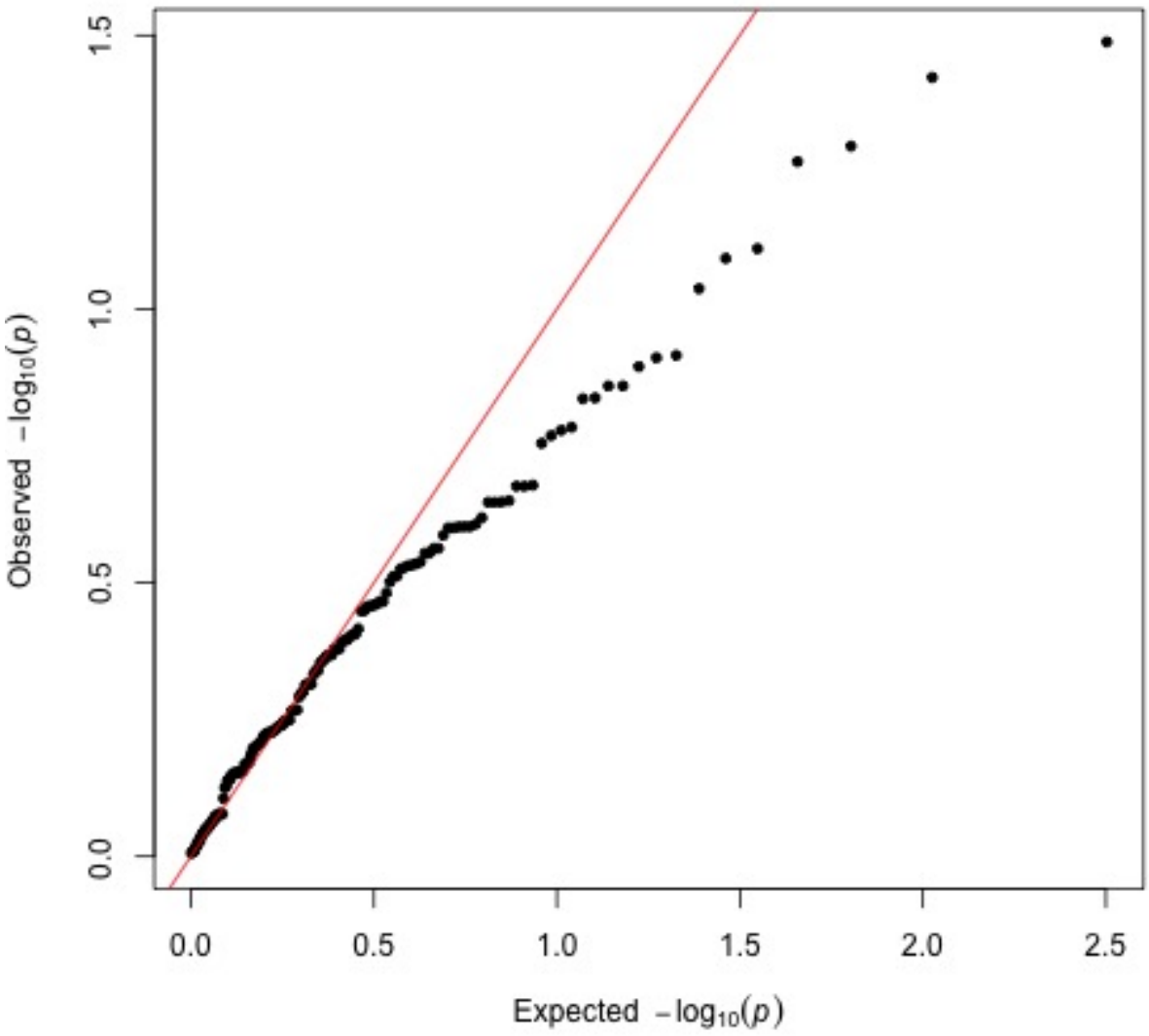

Figure 11: Q-Q Plot TRB

	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

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

1. rs1052406
2. rs1009848

As noted in the EDA above, Figure 12 shows these 2 significant SNPs with their corresponding metrics from `plink` in Figure 12.

After adding in age and sex, we see rs1009848's significance drops as shown in Figure 13

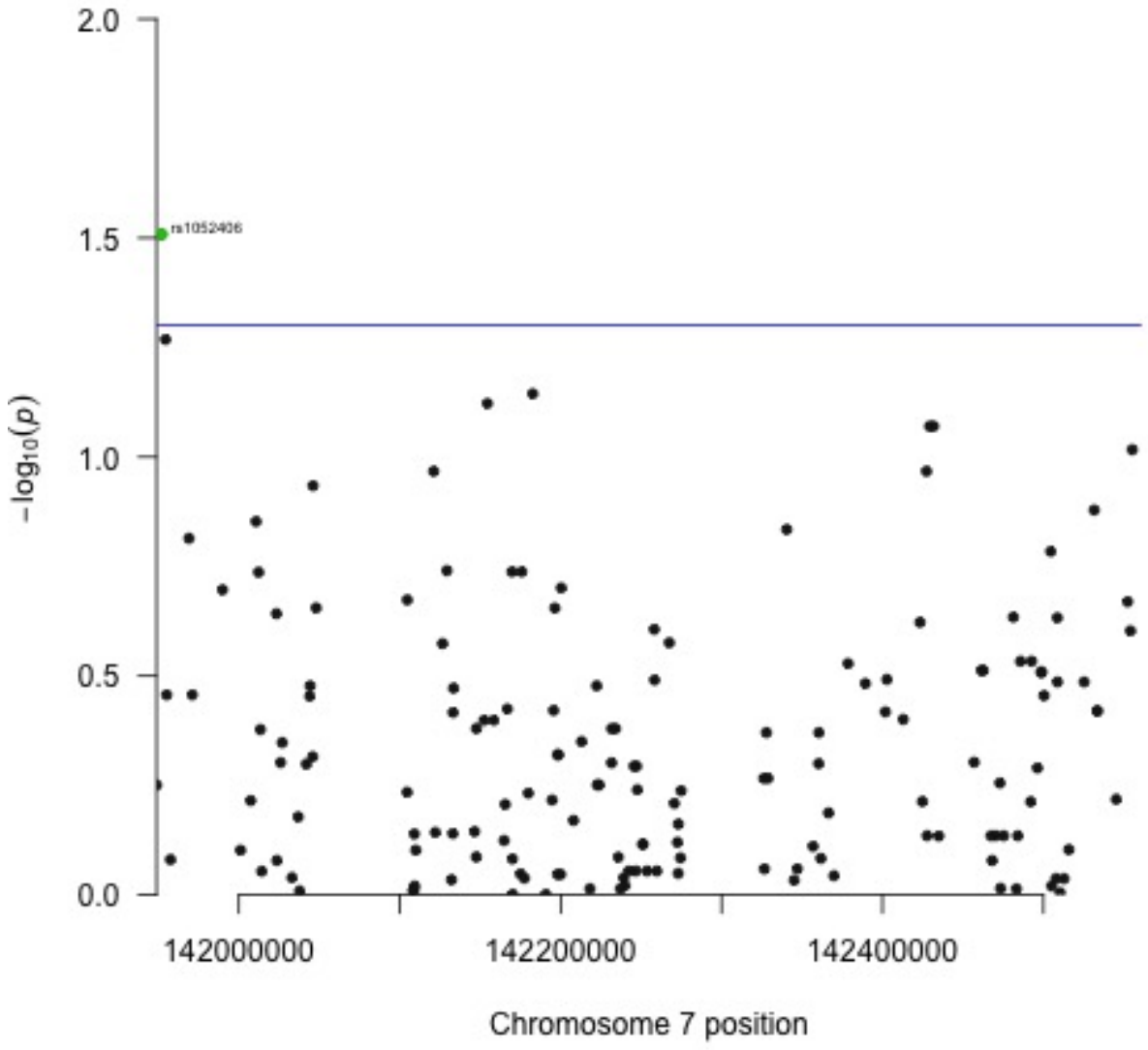

Figure 13: Manhattan with Additional Covariates

This may mean that in this data set, only rs1052406 shows significance when combined with other covariates.

In terms of productive clonaltiy, the Lorenz-curve in Figure 14, shows a diagonal line with a slope of 1, indicating high diversity. Lines further away from this line have less diversity. Each line is a different sample. One sees a modest spread in TCR diversity in the cohort of 15 patients.

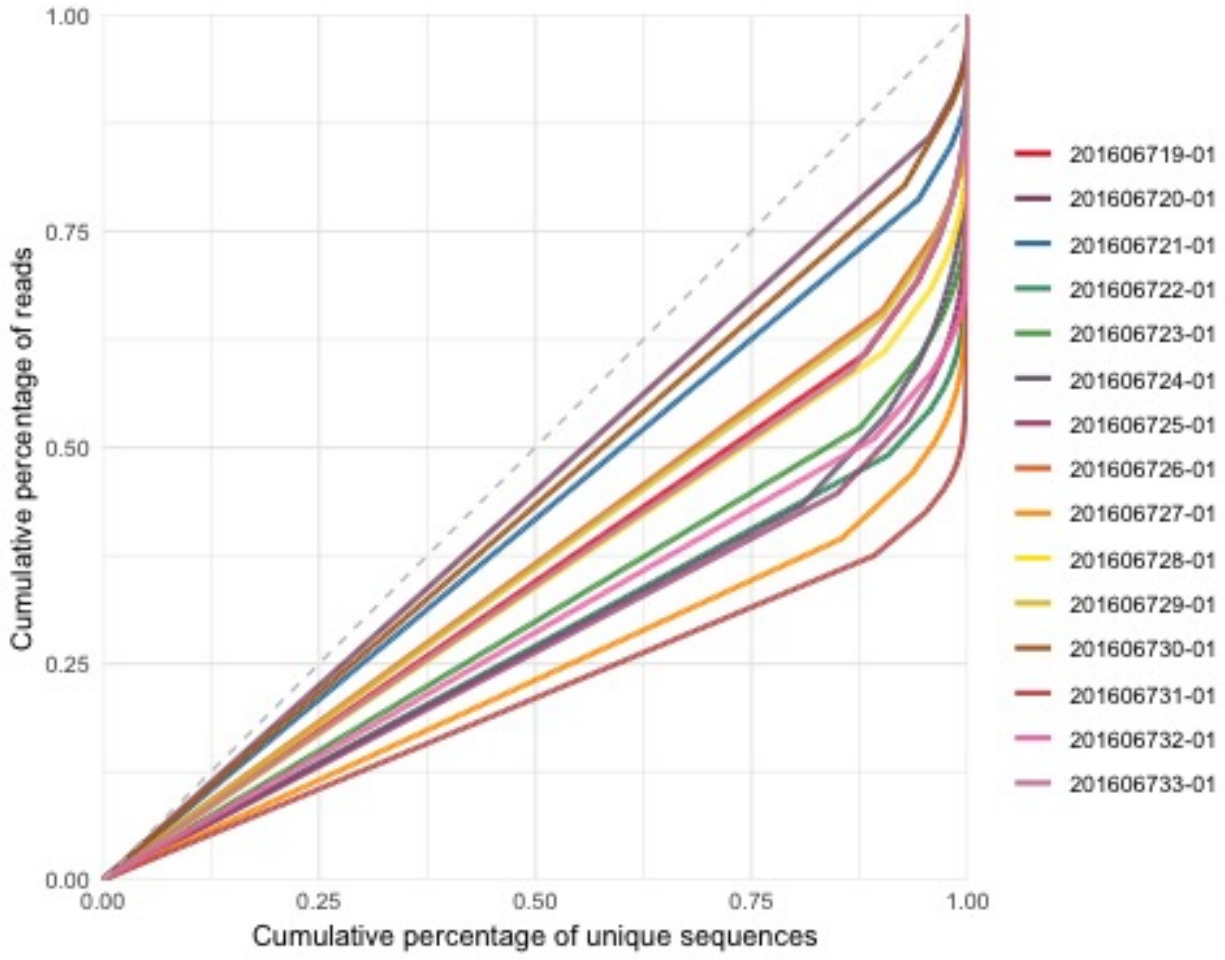

Figure 14: Lorenz Curve

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

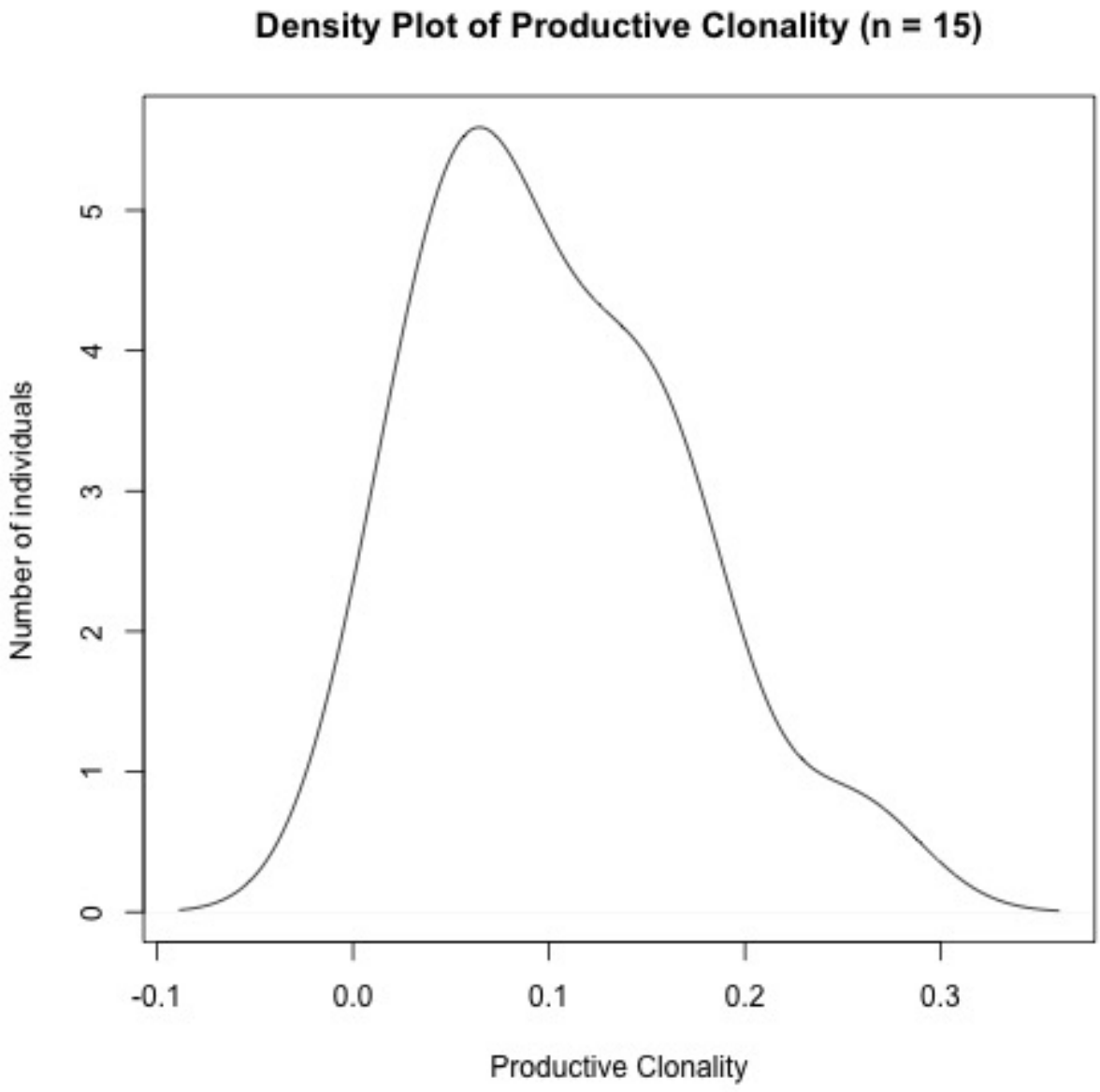

Figure 15: TCR Density

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 16.

```
# 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
```

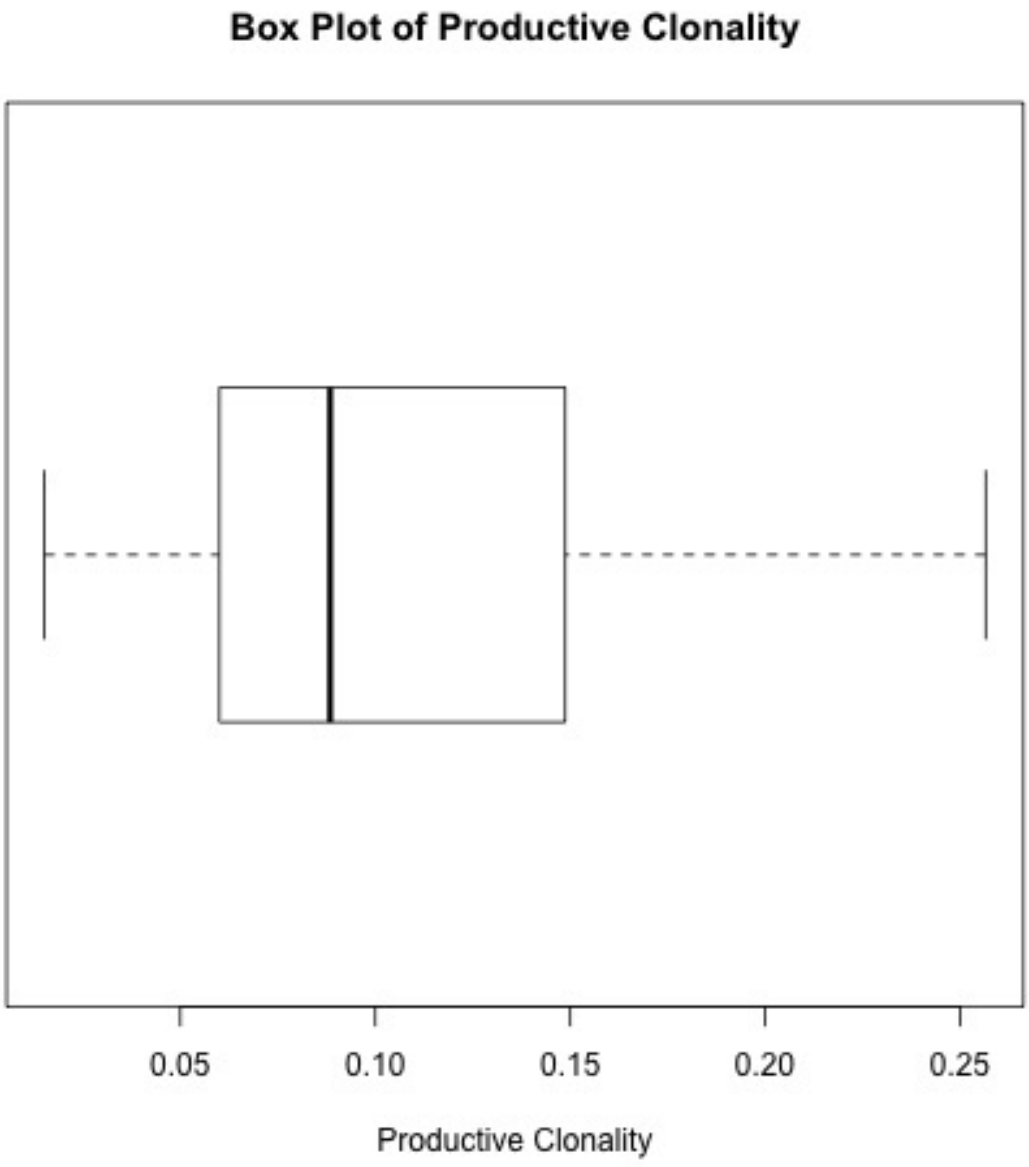

Figure 16: Box Plot of Productive Clonality

After extracting genetic ancestry and principle components (discussed below), a pairs plot was also produced showing correlations between various predictors as noted in the reference to the `obs` data frame above. Red denotes genotype values of 2, green genotype values of 1, and blue genotype values of 0 associated with SNP rs1052506. It seems that SNP rs1052506 shows the strongest correlation with productive clonality. See Figure 17.

Focusing on the TRB locus amongst these 15 samples resulted in 156 SNPs. ($p = 156$ and $n = 15$)

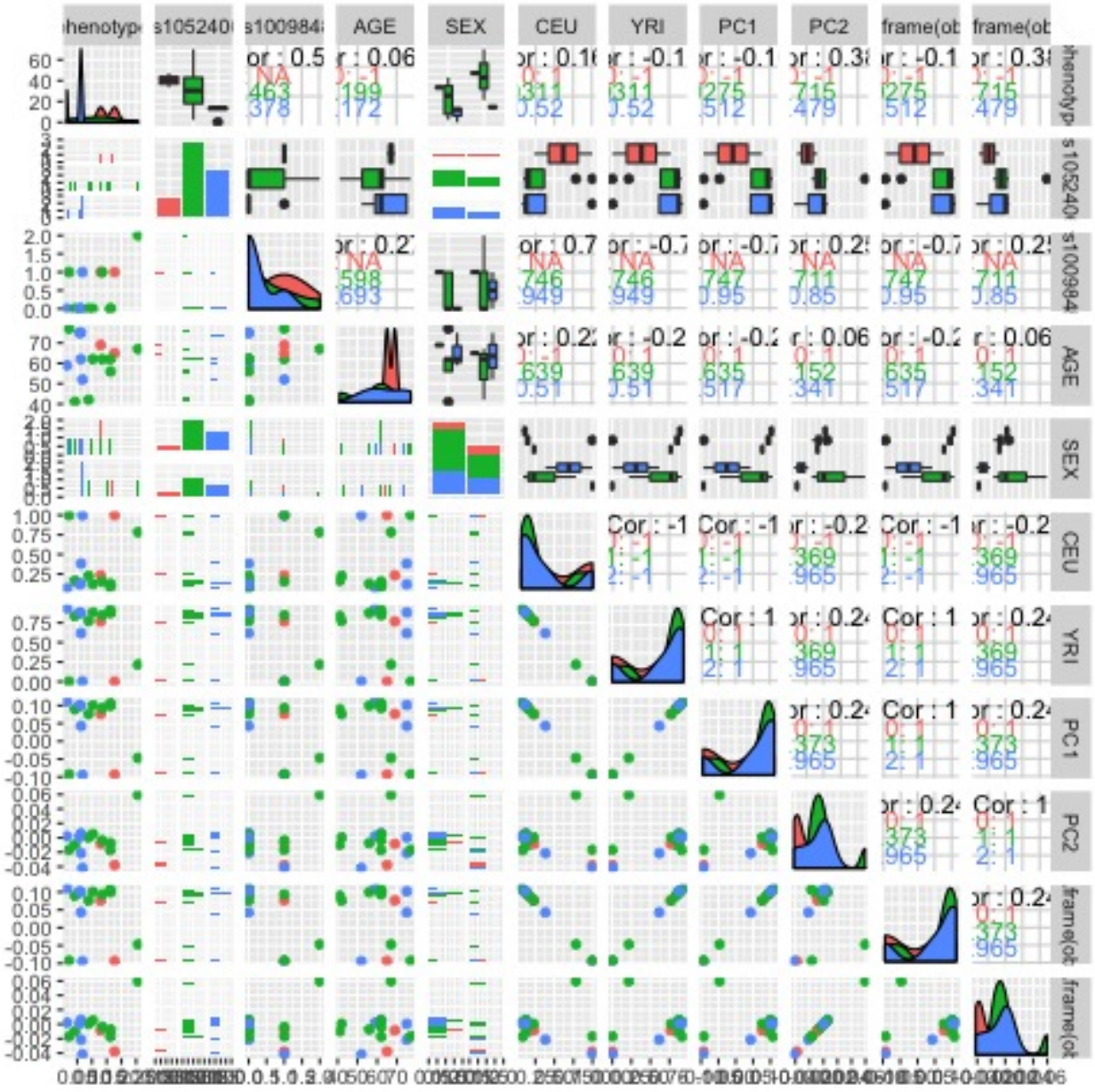

Figure 17: Pairs Plot All Variables

Modeling - Results

Initial Modeling

As mentioned in previous reports, dimension reduction methods of ridge and lasso regression were performed on the reduced sample size. The results showed that lasso regression yielded SNPs that were consistent with findings from earlier Manhattan plots. See Figure 18 and the table below.

SNP	Coefficient
rs1052406	-0.002969940
rs1009848	0.002077136

LOOCV showed that a *lambda* value of 0.03575439 should be used in this lasso regression. See Figure 19.

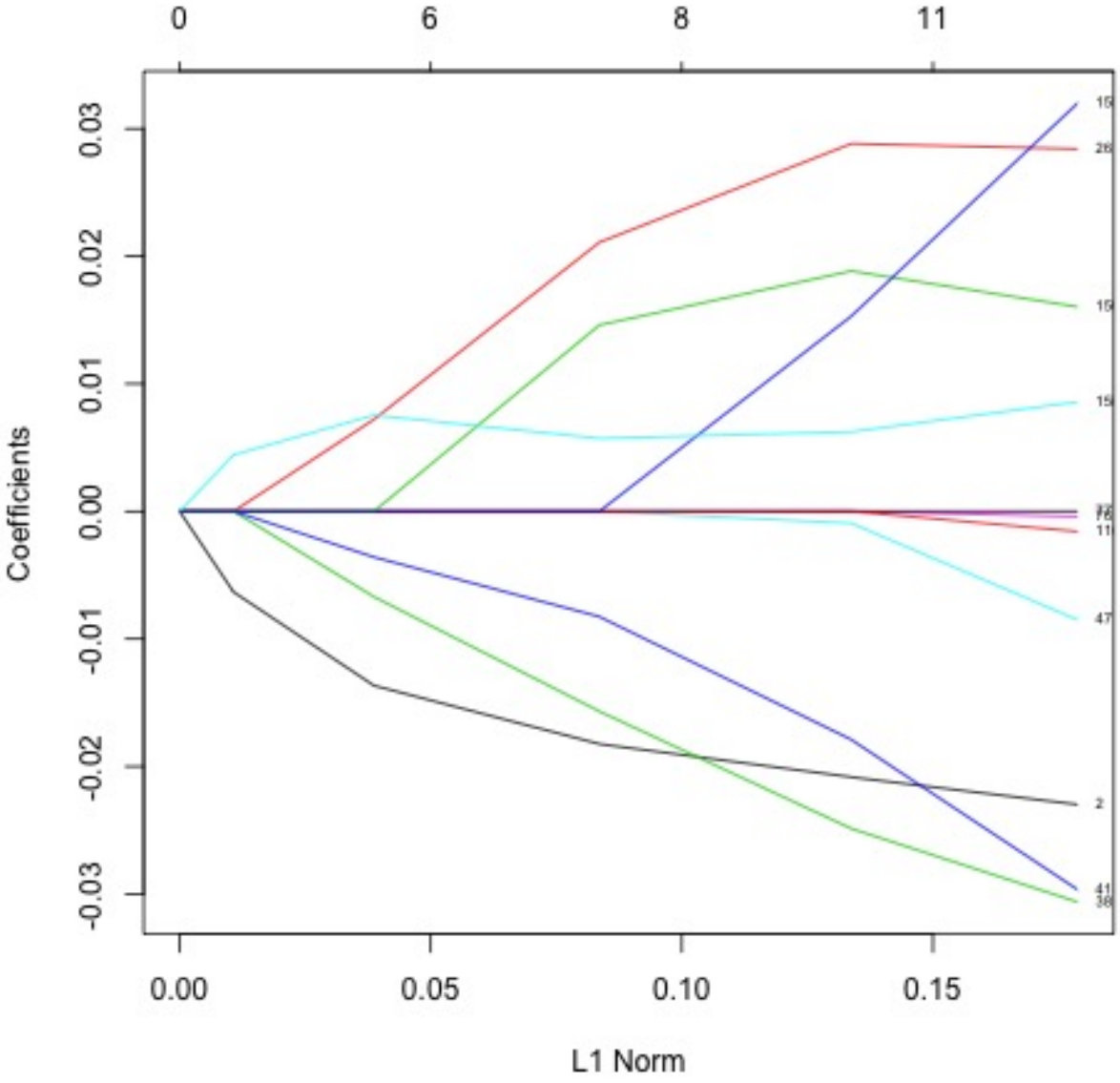

Figure 18: Lasso LOOCV

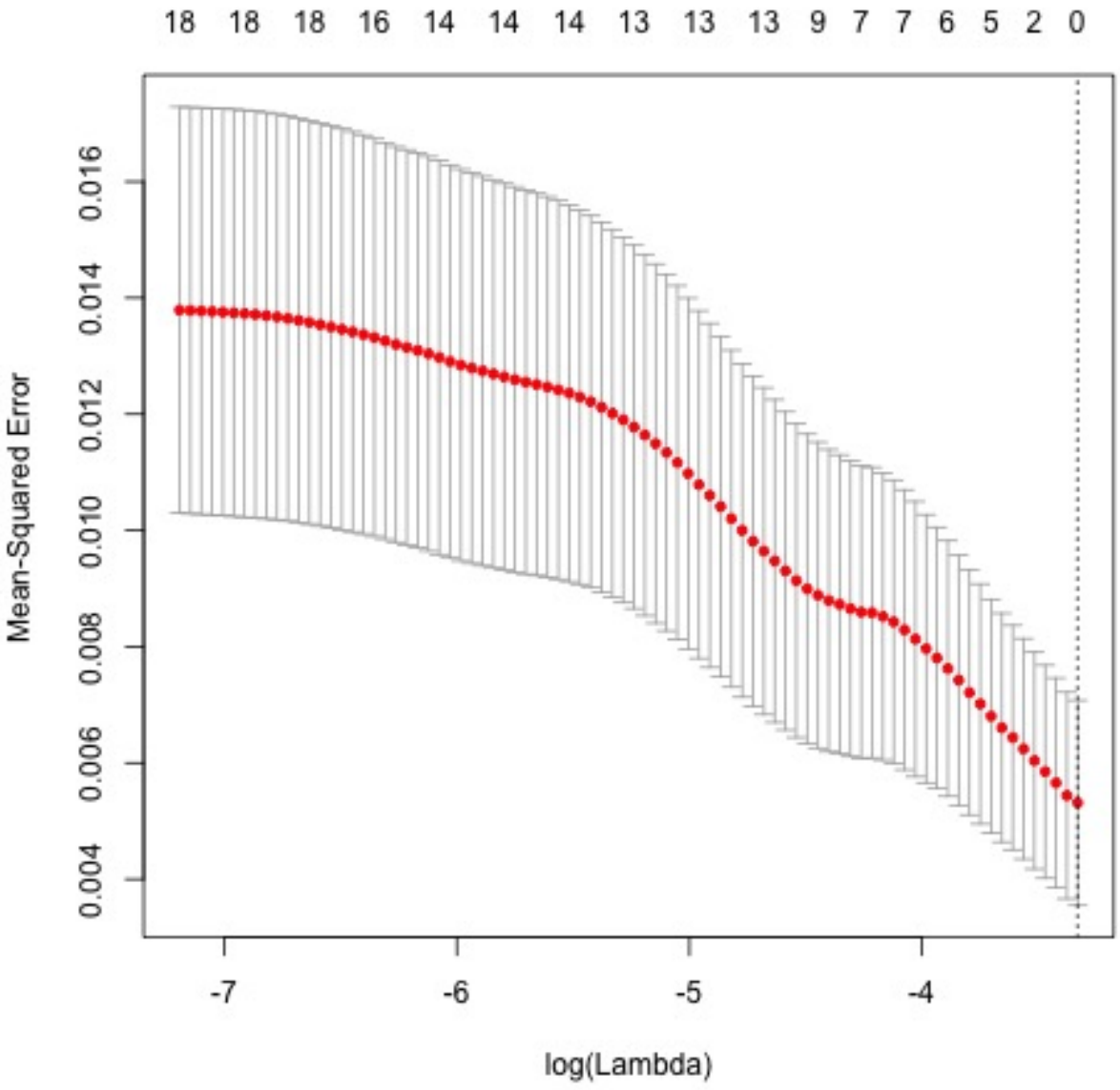

Figure 19: Lasso LOOCV

Subsequent modeling attempts were considered using lasso regression given the consistency presented with earlier results and the fact that lasso greatly reduced the number of SNP-predictors from 156 to 2, thereby simplifying the model.

Addition of Age and Sex Covariates

Adding in additional covariates of age and sex resulted in a Manhattan plot where only SNP rs1052406 persists as being significant at the $p < 0.05$ threshold. This is shown below in Figure 20.

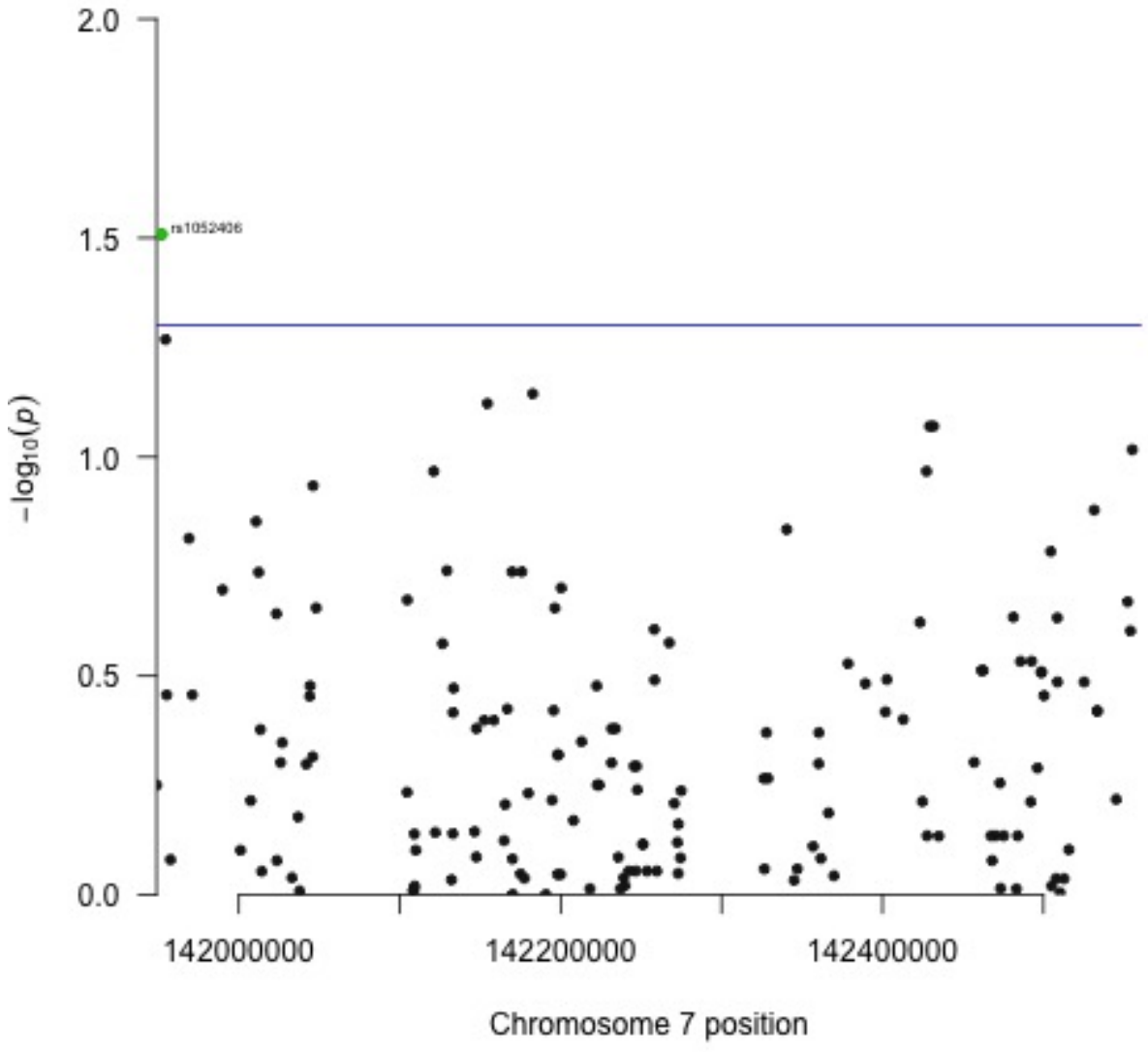

Figure 20: Manhattan with Age and Sex

ADMIXTURE and plink - Predicting Genetic Ancestry/Population Stratification

Recall that all 129 samples were analzyed in order to extract genetic ancestries. `ADMIXTURE`'s supervised learning approach produced proportions of the CEU and YRI ancestries. This is shown below in Figure 21.

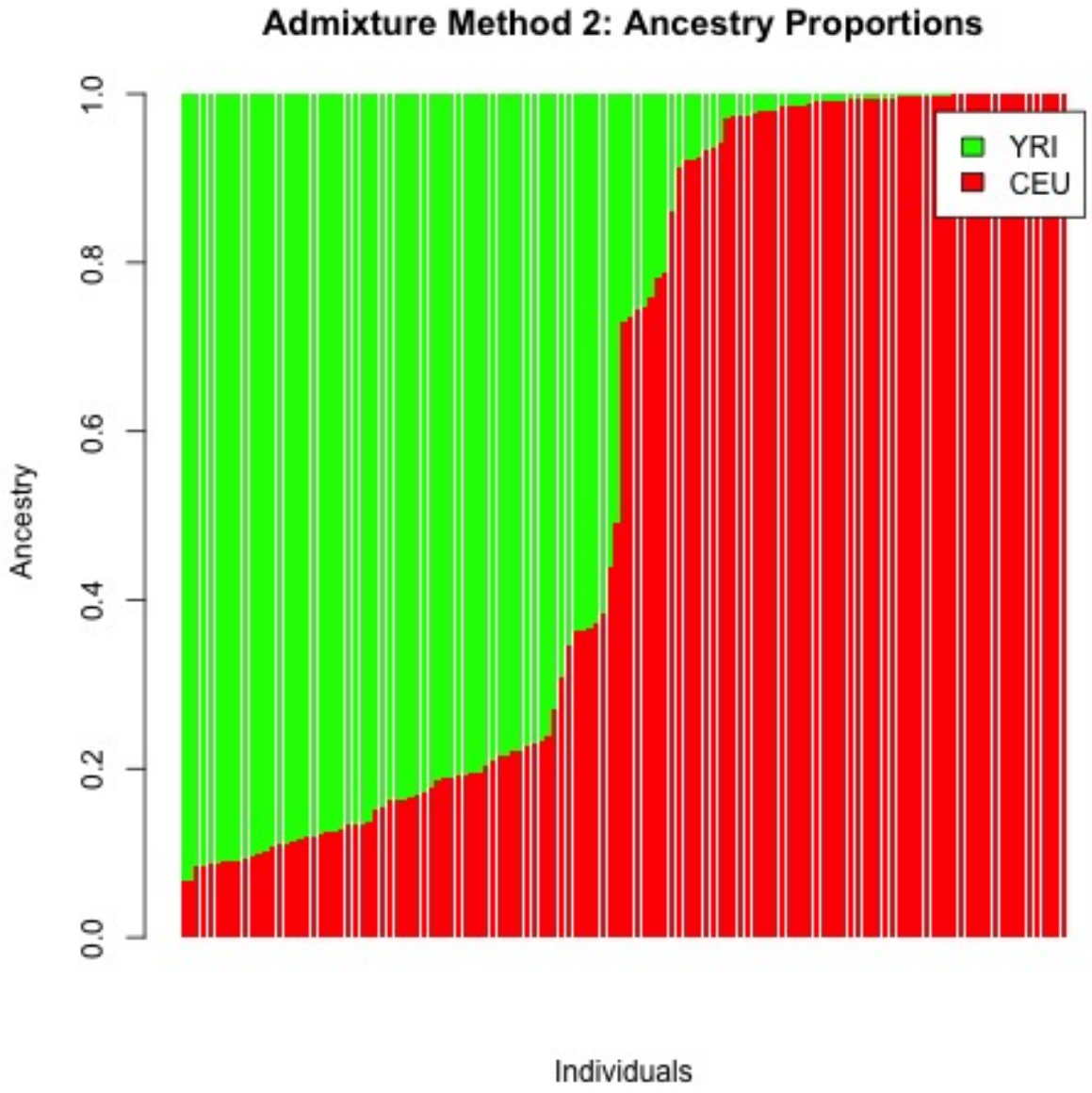

Figure 21: ADMIXTURE Supervised Learning

In comparing with self-reported races, one assumes reference-popuation-proportions greater than 50% should align with the self-reported race. Based on this, there were 2 samples (out of the 129 samples) with conflicts between self-reported races and proportions of ancestry observed. These were samples IDs MIPS062 and MIPS119 and may warrant further investigation. Despite this, `ADMIXTURE` showed a 98.45% accuracy between self-reported races and predicted ancestries.

In regards to `plink`'s method of projecting principle components, it appears that only PC1 and PC2 are likely to explain a majority of the variance given Figure 22 and Figure 23 presented below.

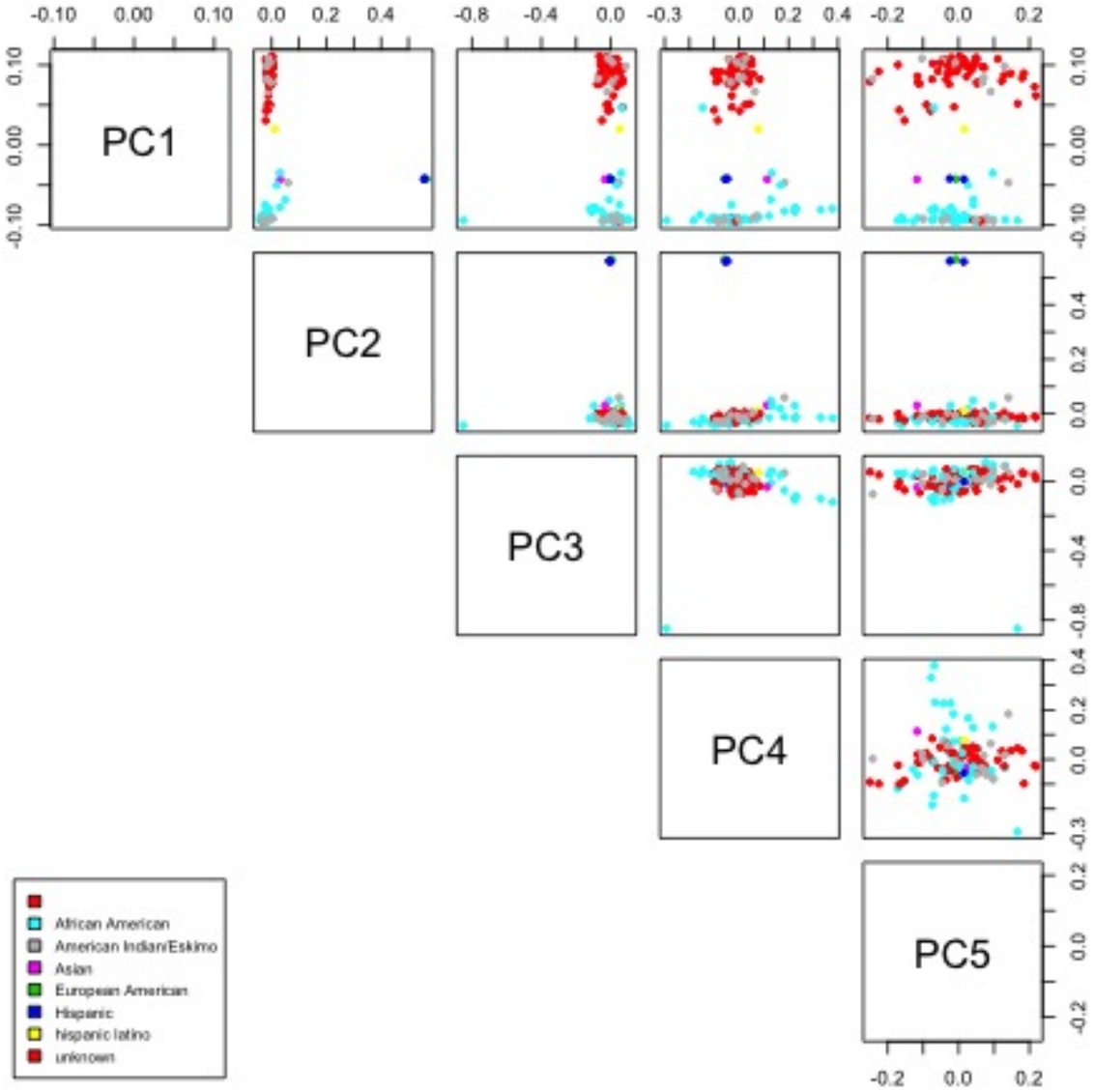

Figure 22: PCA Pairs Plot

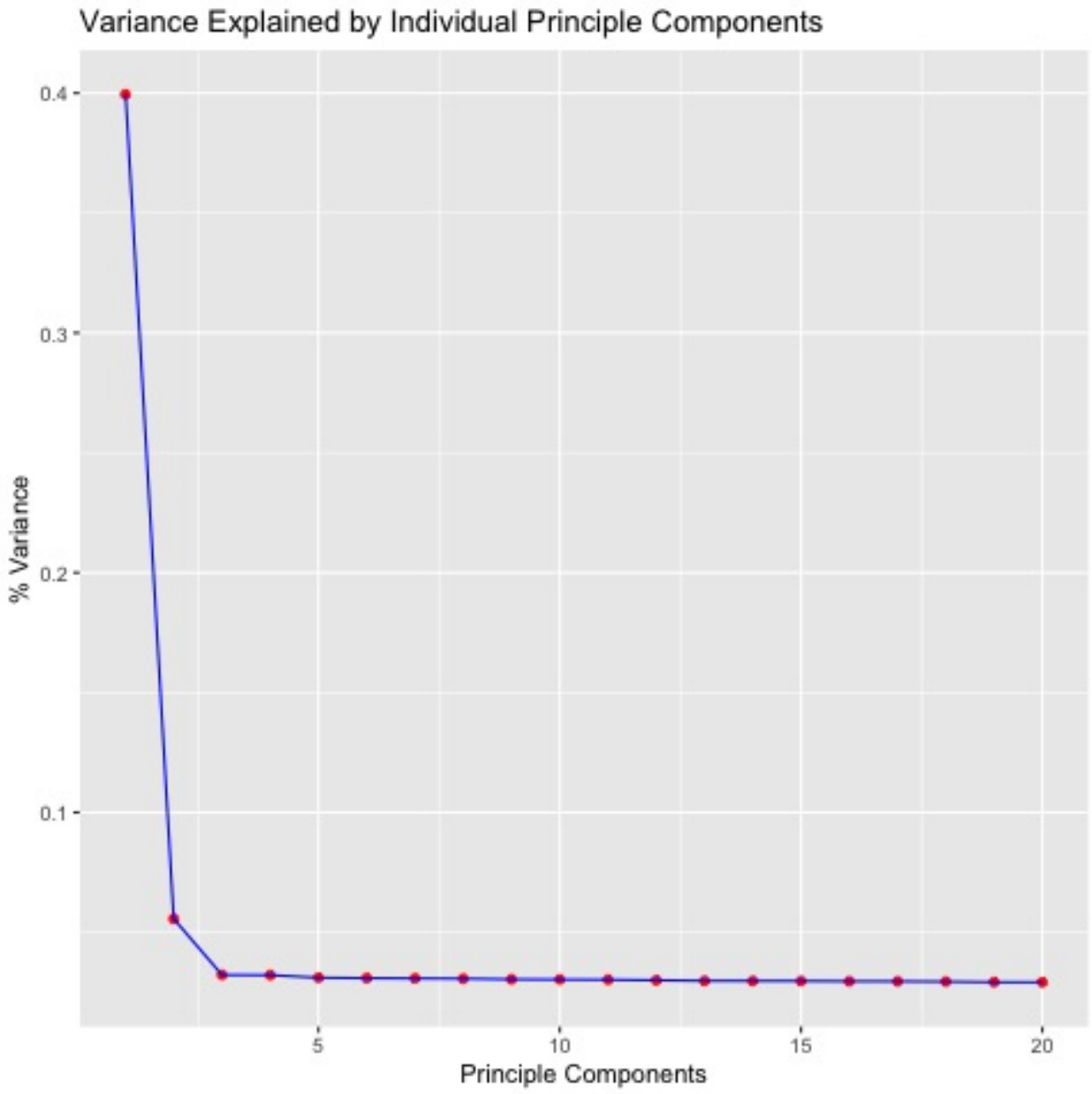

Figure 23: PCA Variance

Final Modeling and Comparisons

Two different implementations of lasso were compared. The lasso-admixture method produced a model which resulted in a single coefficient (SNP rs1052406). Surprisingly, the additional covariates of age, sex and genetic ancestry were not included in the lasso-admixture model. See Figure 24.

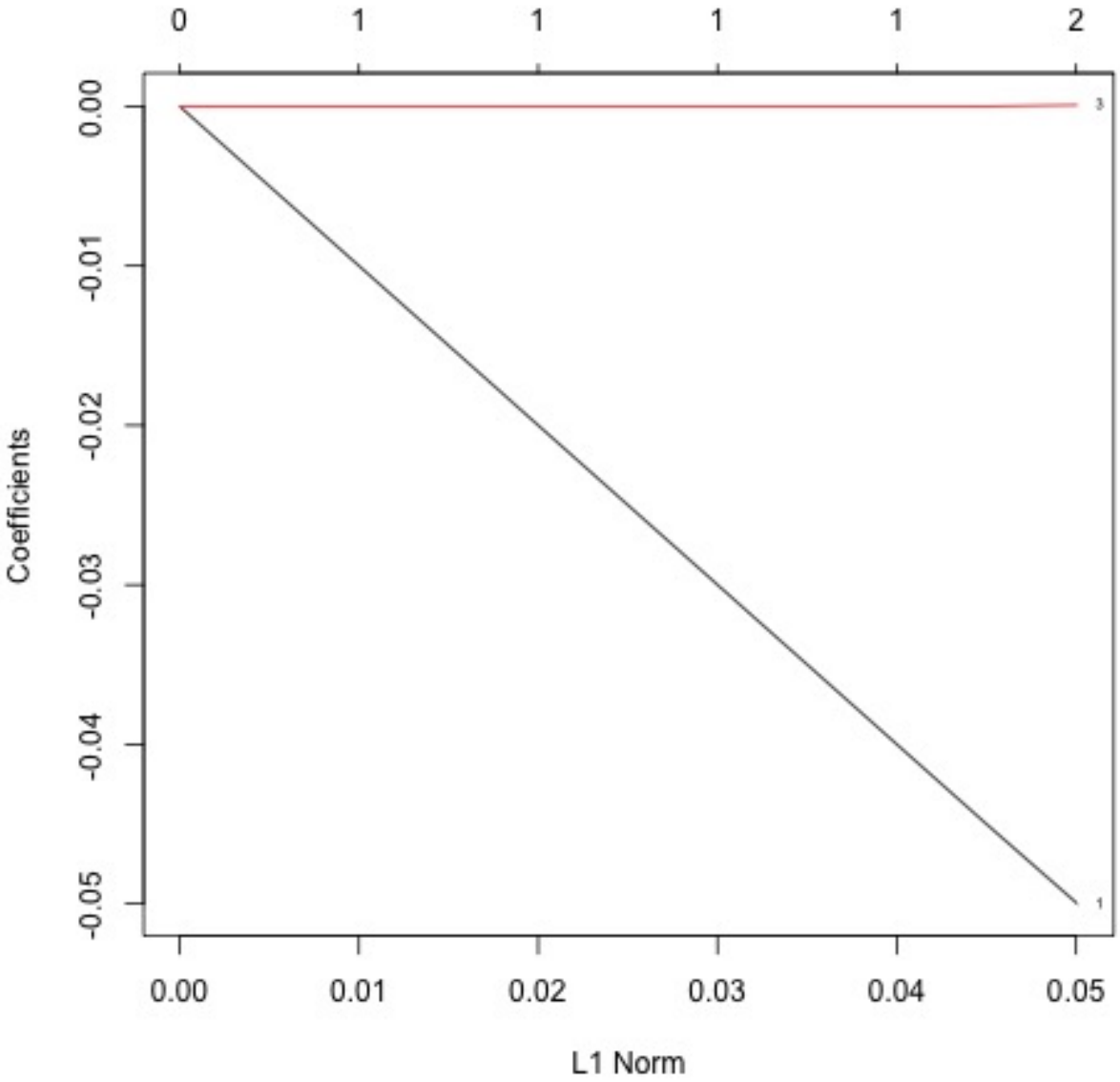

Figure 24: Lasso-Admixture Coefficients

The coefficient for SNP rs1052406 is shown below.

Parameter	Coefficient
rs1052406	-0.04389696

LOOCV showed that a λ value of 0.01396051 should be used. See Figure 25.

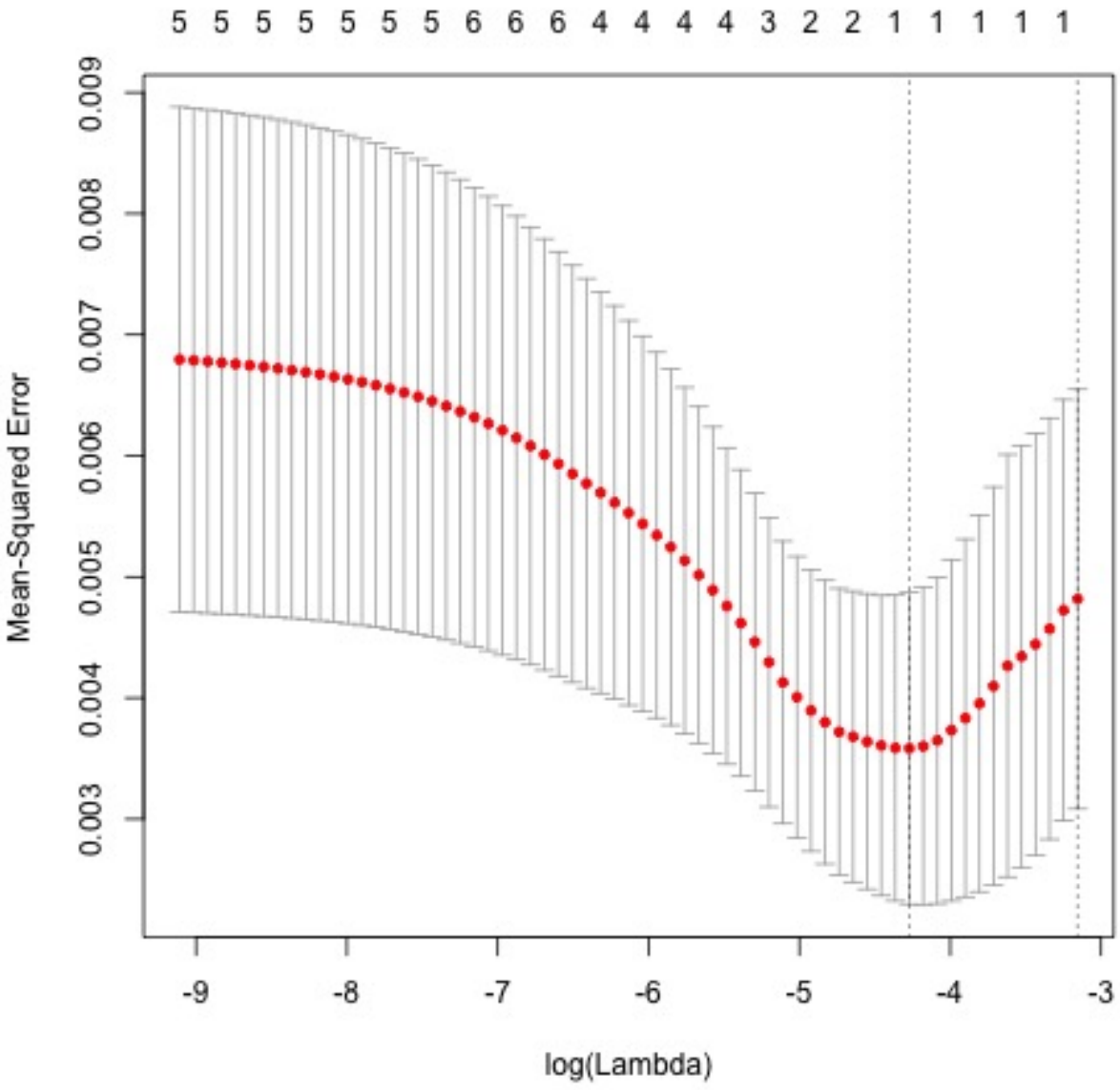

Figure 25: Lasso-Admixture LOOCV

In the lasso-pca method, coefficients were produced for SNP rs1052406 and PC2. Those coefficients are presented below in Figure 26.

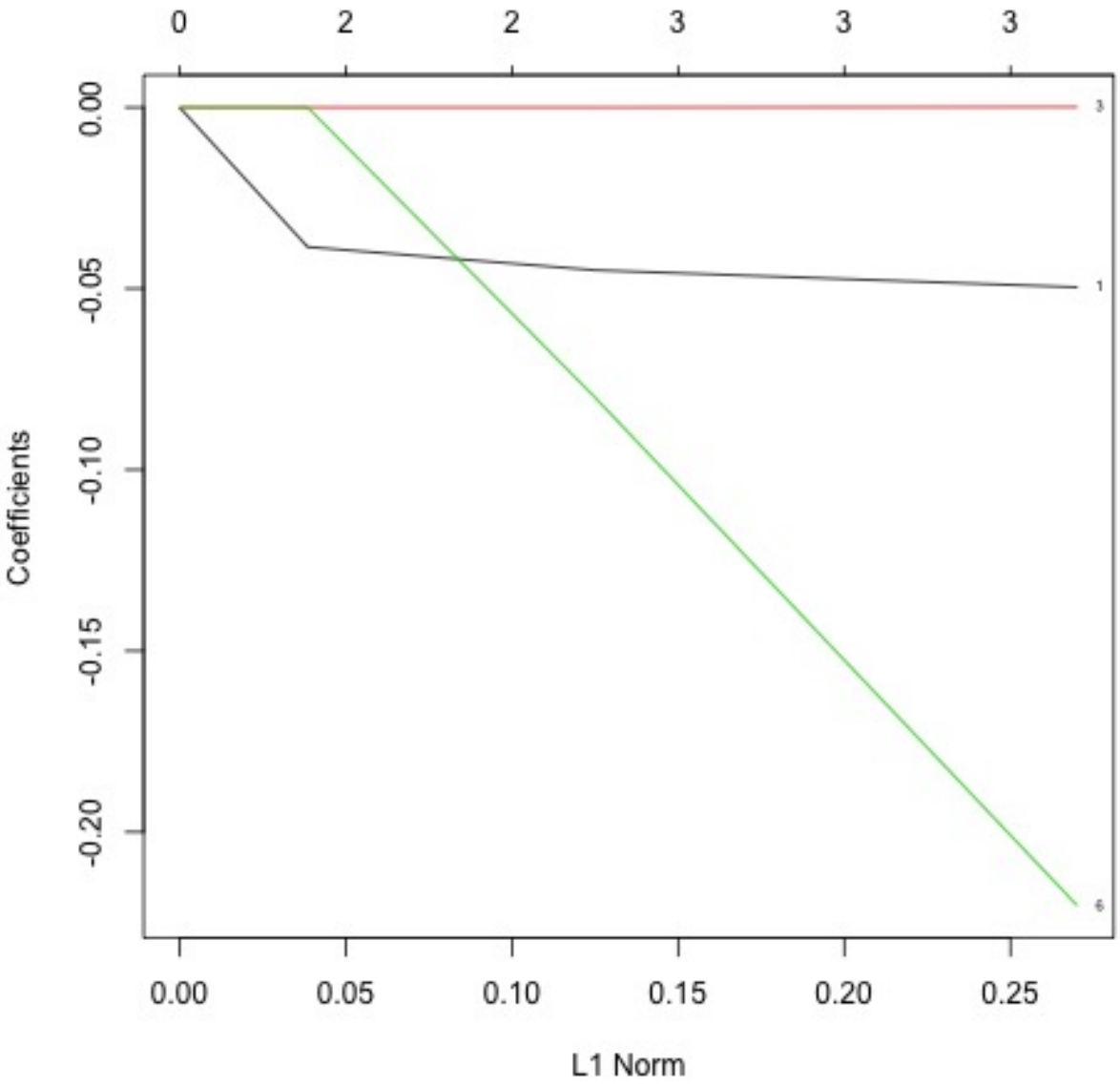

Figure 26: Lasso-PCA Coefficients

Parameter	Coefficient
rs1052406	-0.04379660
PC2	-0.06568802

LOOCV showed that a λ value of 0.01396051 should be used in this alternative approach as shown in Figure 27.

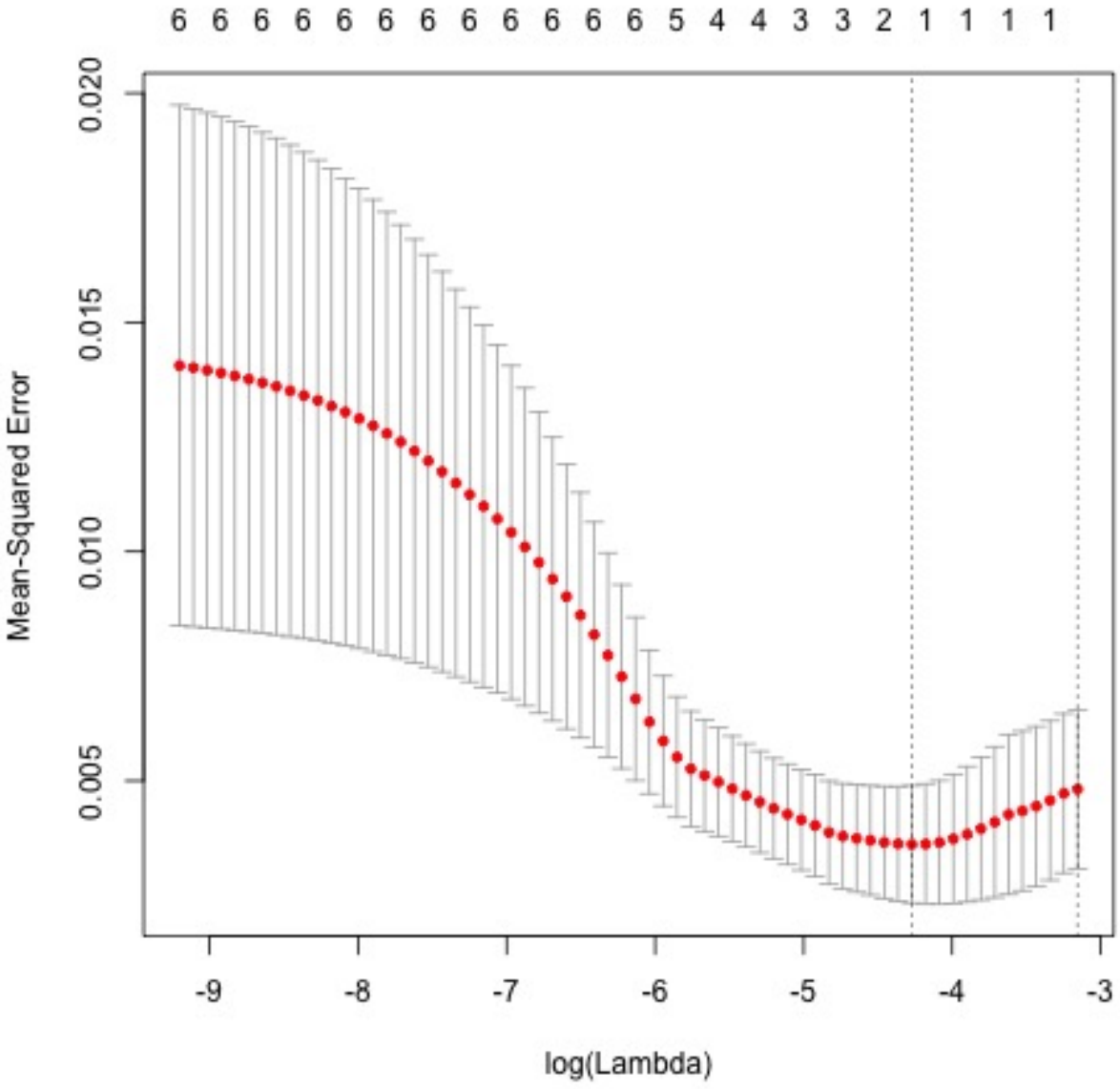

Figure 27: Lasso-PCA LOOCV

When comparing the MSEs (calcuated from the training data, due to small sample size) and AICs from the 2 approaches, the lasso-admixture method produced less error, albeit additional covariates do not seem to play a role in TCR diversity from this data set.

Method	MSE	AIC
Lasso-Admixture	0.002557937	2.283352
Lasso-PCA	0.00251772	4.975214

Discussion

Conclusion

The primary objective of this project was to explore whether there were any significant SNPs that could be associated with productive clonality from a limited sample data set. Overall, it appears that SNP rs1052506 may have effects on observed productive clonality. While SNP rs10009848 was shown to be possibly signficant, addition of more covariates and further regression analysis showed that only SNP rs1052506 was significant. Current associations in the literature bolster SNP rs1052506's relevance to productive clonality and T-Cell receptors. For instance, when looking at HaploReg (HaploReg annotated SNPs in same haplotyptes), although rs1052406 does not directly overlap with transcripts involved in TCRs, there does appear to be strongly correlated SNPs which are associated with the TRBV2 gene. The TRBV2 gene is a gene directly involved in the expression of the CDR3 region of T-Cell receptors.

SNP	Overlapping Genes	Function	Associated SNPs	Associated Genes
rs1052406	PRSS58	Serine Protease, Pancreatic Diseases	rs10231842, rs2960774, rs3020843	surround TRBV2 Genes

Additionally, the GTex Portal (showing tissue-specific gene expression), shows that varying rs1052406 genotypes may be associated with varying TRB4-1 expression levels in whole blood. See Figure 28.

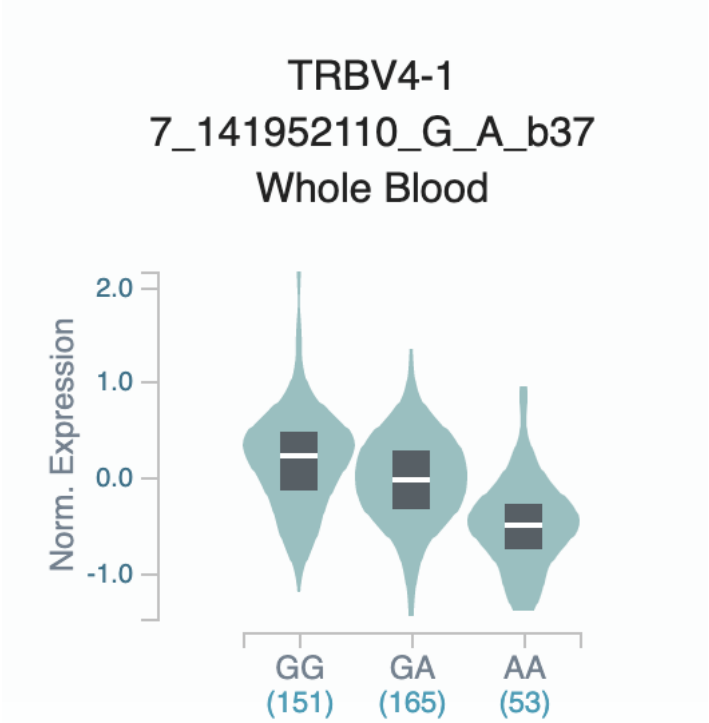

Figure 28: Violin TRBV4-1 for rs1052406

Further refinement of the data resulted in predicting proportions of CEU and YRI ancestries using `ADMIXTURE`. For exploratory purposes, PCA was also done using `plink` to account for similar population stratification. Lasso regressions were constructed for each method of population stratification taking the other covariantes of SNPs, age and sex into account. Interestingly the lasso-admixture implementation produced only 1 signficant predictor (SNP rs1052506) which is consistent with earlier results. In terms of AIC, the lasso-admixture showed a lower AIC. However, it should be noted the author is not confident in either of these models for productive clonality predictions. More so, from this study, there are likely SNPs within and around the TRB locus which contribute and/or show associative effects with productive clonality.

Limitations

The small sample size of this study should be acknowledged. Only 15 samples with productive clonality metrics were used from a larger cohort of 129 samples. This may be the main limiting factor in developing a predictive model for productive clonaltiy. If continuing development, additional samples should be sequenced to increase power and perhaps build stronger models. Future work remains to be decided, but the results thus far indicating some association with SNPs in and around the TRB locus and productive clonality are encouraging.

The Data Book

File Type	Variable	Units	Description
obs	phenotype	N/A	Same as plinkPedJoin.phenotype below
obs	rs1052506	N/A	Genotype values (0, 1, 2) for rs1052506
obs	rs10009848	N/A	Genotype values (0, 1, 2) for rs10009848

File Type	Variable	Units	Description
obs	AGE	years	Same as obs.AGE below
obs	SEX	N/A	Gender of patient, coded as dummy variable (1 for Male, 0 for female)
obs	CEU	N/A	Proportion of European ancestry
obs	YRI	N/A	Proportion of African ancestry
obs	PC1	N/A	PC1 eigenvector extracted from `plink`
obs	PC2	N/A	PC2 eigenvector extracted from `plink`
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)

R Packages

```
# immunoSeq data analysis tools
library(LymphoSeq)
# interact with biolync server
library(ssh)
# Read in binary genotype files
library(KRIS)
# Visualization of GWAS data
library(qqman)
# build regression models
```

```
library(glmnet)
# clean data
library(tidyverse)
```

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 strong support and instruction.
- This was truly an immersive and stimulating learning experience, incorporating various methods from data science and computational biology.

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. doi: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.

Bush, William S., and Jason H. Moore. 2012. "Chapter 11: Genome-Wide Association Studies." *PLoS Computational Biology* 8 (12). doi:10.1371/journal.pcbi.1002822.

Cole, Brian Sebasatian, and Endre Bakken Stovner. n.d. "No Title." https://github.com/biocore-ntnu/snpflip.

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.

DH, Alexander, Novembre J, and Lange K. 2009. "Fast model-based estimation of ancestry in unrelated individuals." *Genome Research* 19: 1655–64. doi:10.1101/gr.094052.109.vidual.

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

Kluiver, Hilde de, Florence Vorspan, Emmanuel Curis, Eske M. Derks, Cynthia Marie-Claire, Sven Stringer, and Andries T. Marees. 2018. "A tutorial on conducting genome-wide association studies: Quality control and statistical analysis." *International Journal of Methods in Psychiatric Research* 27 (2): e1608. doi:10.1002/mpr.1608.

M., Kate. 2016. "What are single nucleotide polymorphisms?" https://socratic.org/questions/what-are-single-nucleotide-polymorphisms.

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

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A.R. Ferreira, David Bender, Julian Maller, et al. 2007. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage

Analyses." *The American Journal of Human Genetics* 81 (3): 559–75. doi:10.1086/519795.

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. doi:10.1186/1753-6561-3-s7-s41.

Zurich, ETH. 2019. "Systems Immunology – Laboratory for Systems and Synthetic Immunology | ETH Zurich." Accessed March 6. https://www.bsse.ethz.ch/lsi/research/systems-immunology.html.