Identification of Signficant SNPs associated with T-Cell Receptor Diversity

John Lin, Case Western Reserve University

March 07, 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 an overview of a cohort of patients' genotype and immunoSeq (TCR-related) data and initial approaches taken to visualize and analyze the data. Please note this work is in conjunction with Case Western Reserve Univeristy and is operating under the CC-BY-SA 4.0 License, currently.

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 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.

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.

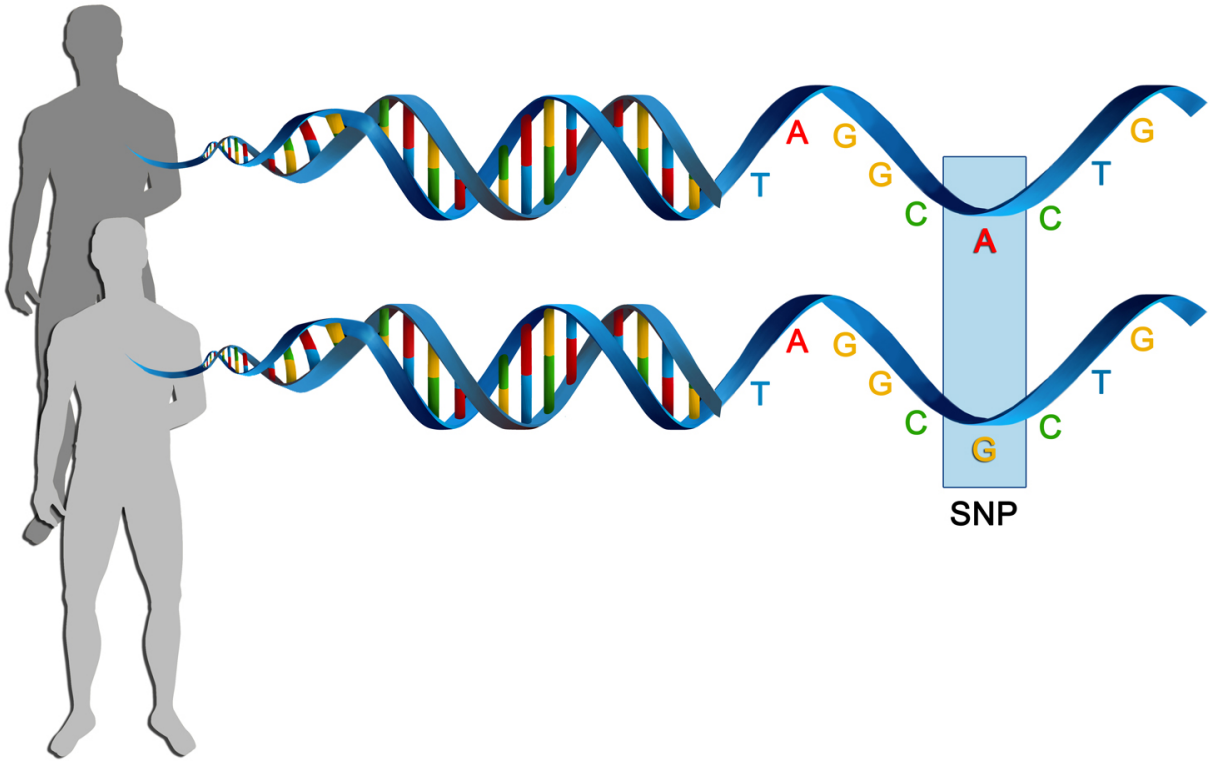


Figure 1: SNP Diagram from (M. 2016)

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) (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.

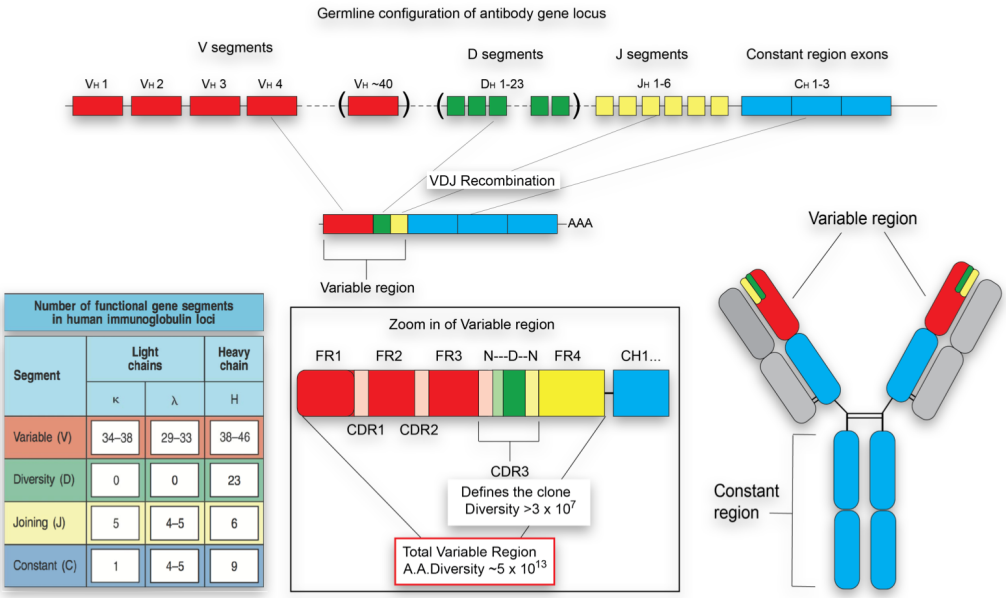


Figure 2: TCR Somatic Recombination from (Zurich n.d.)

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.

Data Science Methods

A high level pipeline of the tools used:

Genotype Data

- Illumina MEGA Chip Data (on biolync.case.edu) –> GenomeStudio –> PLINK WGA Toolkit <–> R

combined with...

Somatic (immunoSeq, TCR) Data

- immunoSeq Data –> R <–> PLINK WGA Toolkit

The above shows the use of PLINK (Purcell et al. 2007) to do much of the data processing as it is a toolkit for whole genome analysis. R will also be used for data processing. This is not a comprehensive representation of tools used and is anticipated to be updated in future reports.

Linear Regression under the Additive Model

In this model, independent tests of simple linear regression are performed for every SNP. In this additive model, 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)	Coding (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).

```
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")
```

It should be noted, the above is for demonstration purposes only. The linear regressions and p-values will be automated by PLINK. Additionally, there will be adjustments made to p-values to correct for multiple testing.

Principal Component Analysis

In addition to the above, there are statistical modeling aspects involved in cleaning the data, which are typical GWAS studies. Most striking, is the use of clustering to identify sub-populations. Systematic differences in allele frequences can be due to underlying populations and will confound the results. This is known as population stratification and will be applied the next phase of this analysis.

Clustering of Nucleotide and/or Amino Acid Sequences (if time allows)

The nature of this project requires significant background knowledge in genetics and familiarity with the tools conventionally used. If time allows after appropriate identification of SNPs, another possible analysis is clustering of nucleotide and/or amino acid sequences given similarities in productive clonality metrics.

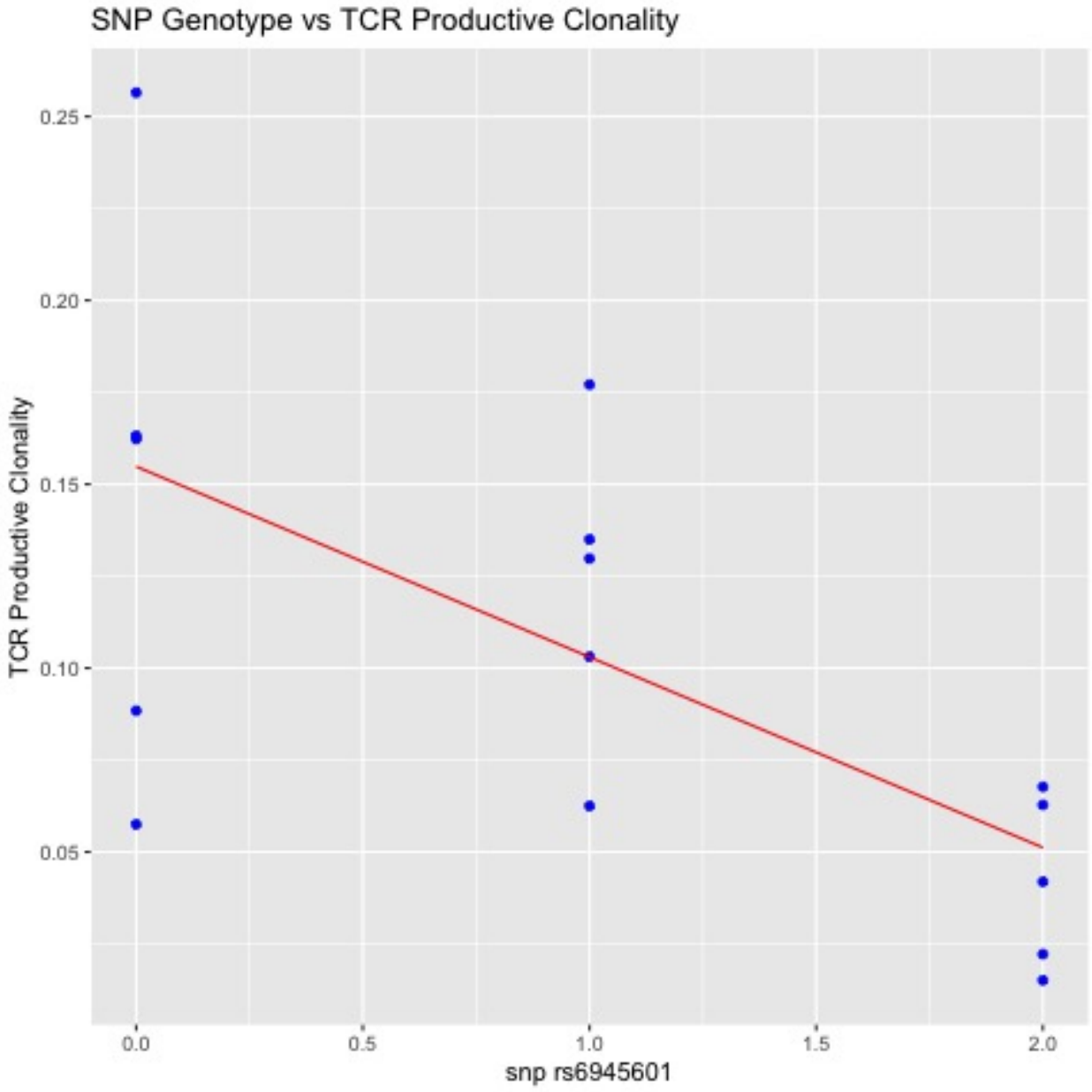


Figure 3: Linear Regression Example

Exploratory Data Analysis and Initial Modeling

The Data Set, Data Book and Packages to be used

Data Set

The raw version of the data comes in a variety of different forms. PLINK is optimized to work with BED files (binary, genotype data) and outputs FAM (showing phenotype data associated with individual identifiers) and BIM (SNP mapped data) as well. The number of observations is 15 (n = 15) with at most 82360 SNPs (after just filtering for SNPS in TRB gene). The features in this model are the different genotypes for different SNPs and the response of interest is productive clonality. Additionally, PLINK outputs a file (.assoc.linear) to easily create Manhattan plots (discussed later). The different file types and fields of interest are elaborated below.

Data Book

File Type	Variable	Units	Description
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)

Packages

Packages of interest are as follows:

```
# Genome browser interaction
library(rtracklayer)
```

```
# immunoSeq data analysis tools
library(LymphoSeq)
```

- Other:

```
# Read in binary genotype files
library(KRIS)
# Heat map generation
library(pheatmap)
# Visualization of GWAS data
library(qqman)
```

Overview of Methods and Initial Results

Quality Control

A significant portion of this project entails quality control and cleaning of the genotype data prior to analysis. The quality control steps below are typical in a GWAS study as noted by Kluiver et al. (2018). The primary aim is to remove confounding variables as well as possible factors relating to sample contamination. Some initial modeling was also performed to gain familiarity with the tool sets used.

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
8. Filter SNPs by appropriate window (50,000 bp +/- CDR3)
9. Filter phenotype data for productive sequences

Initial Modeling

10. Combine phenotype data with genotype data
11. Multiple Simple Linear Regression (Additive Model)
12. Visualization of immunoSeq data

Steps 1, 2, and 6 were completed with initial processing from Tyler Kinzy, a research associate in Dr. Jessica Cooke Bailey's research lab at Case Western Reserve University. Tyler checked for sex differences and found 4 individuals who needed to be removed due to sex discrepancies between inferred sex (from genotype data) and reported sex. This can be more clearly seen in the Figure 4 below.

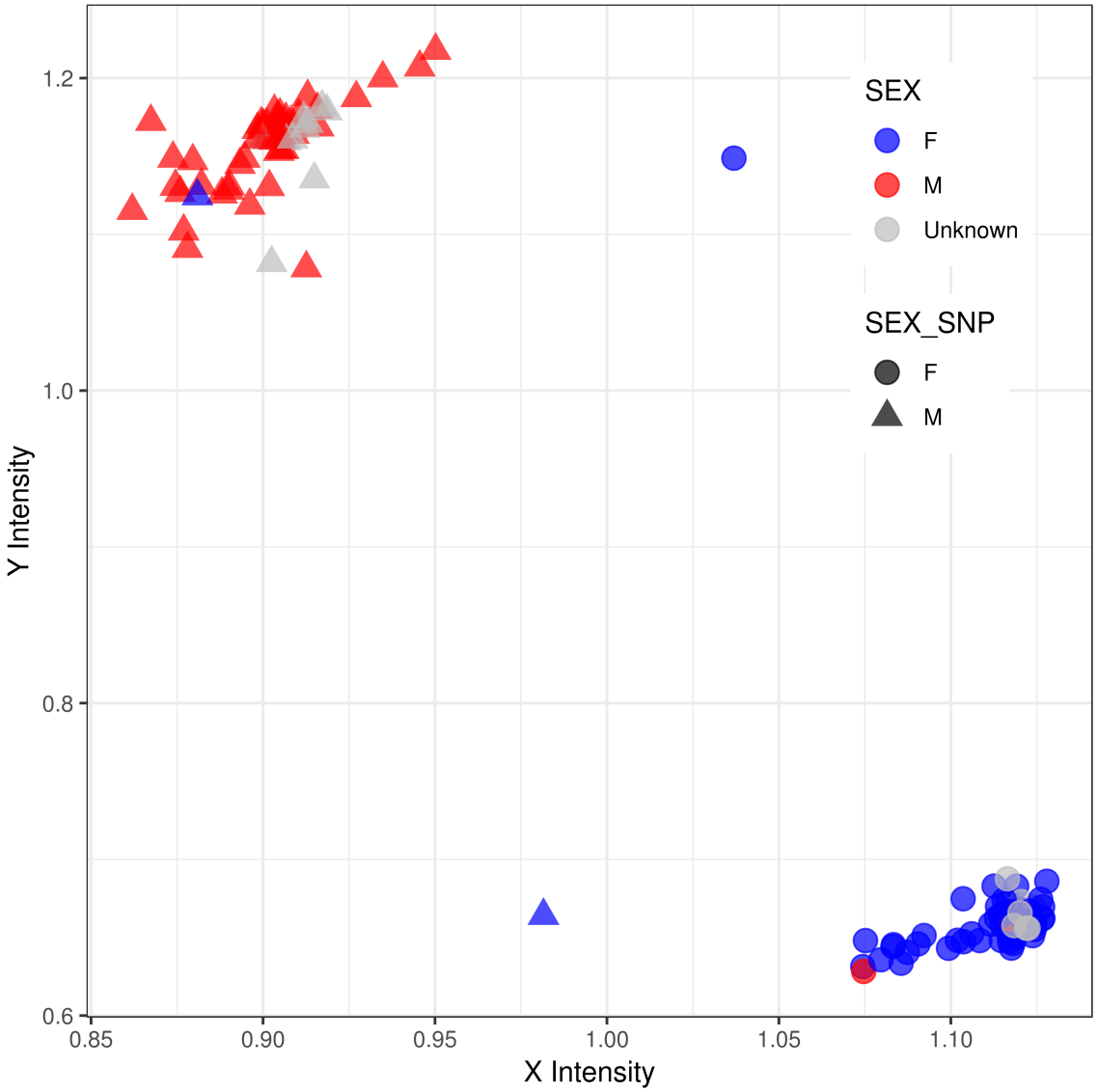


Figure 4: Sex Check

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.

Steps 3, 4, 5, and 7 remain to be done and can be easily accommodated with PLINK in subsequent arguments. See PLINK guide here.

Steps 9 and 10 were accomplished via R. Their outputs were referenced in a PLINK command which also happened to perform steps 8 and 11 simultaneously.

```
require(tidyverse)
require(RCurl)

# Create phenotype file tcrEmrPheno.txt to input into plink
# Read in all Phenotypes
# TODO: Need to store on the biolync server with RCurl
tcrEmrPheno <- read.delim(
  "/Users/linjo/Box\ Sync/MIPs/Immunoseq_EHR_2019-02-15_for_BOX.txt",
  sep = "\t", strip.white = TRUE)
# Create FIID and IID columns
tcrEmrPheno$FID <- paste(substring(tcrEmrPheno$MIPs.ID, 0, 4),
                         substring(tcrEmrPheno$MIPs.ID, 10, 13), sep = "")
tcrEmrPheno$IID <- paste(substring(tcrEmrPheno$MIPs.ID, 0, 4),
                         substring(tcrEmrPheno$MIPs.ID, 10, 13), sep = "")
tcrEmrPheno <- tcrEmrPheno[, c(47:48, 1:46)]
# Replace unwanted characters
tcrEmrPheno <- data.frame(lapply(tcrEmrPheno, function(x) {
  gsub(" ", "_", x)
}))
tcrEmrPheno <- data.frame(lapply(tcrEmrPheno, function(x) {
  gsub(",", "", x)
}))
tcrEmrPheno <- data.frame(lapply(tcrEmrPheno, function(x) {
  gsub("%", "", x)
}))
# Rearrange so that FIID and IID are first
tcrEmrPheno <- select(tcrEmrPheno, FID, IID, everything())
# Output for plink
# TODO: Need to store on the biolync server with RCurl
write_delim(tcrEmrPheno,
            path = "/Users/linjo/Box\ Sync/MIPs/tcrEmrPheno.txt",
            delim = "\t", col_names = TRUE, quote_escape = FALSE)
```

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

Then, in R again, one creates a Manhattan plot (see Figure 5) from the .assoc.linear output from PLINK.

```
# Read in linear regression results
# TODO: use Rcurl to read this in from server
# TODO: why is 10.3600 showing up as a value, it should be 0.2565
plinkLinear <- read.table("/Users/linjo/Desktop/tcr-project-desktop/MIPS_Updated.2019-02-21/jxl2059/plin
# Remove where snps with P values of NA
plinkLinear <- na.omit(plinkLinear, col = "P")
# Generate and save Manhattan plot
```

```r
# TODO: use Rcurl to read this in from server
jpeg('manhattan1.jpg')
manhattan(plinkLinear, ylim = c(0, 40), annotatePval = 0.05)
dev.off()
# Generate and save qq plot
# TODO: use Rcurl to read this in from server
jpeg('qqplot1.jpg')
qq(plinkLinear$P)
dev.off()
```

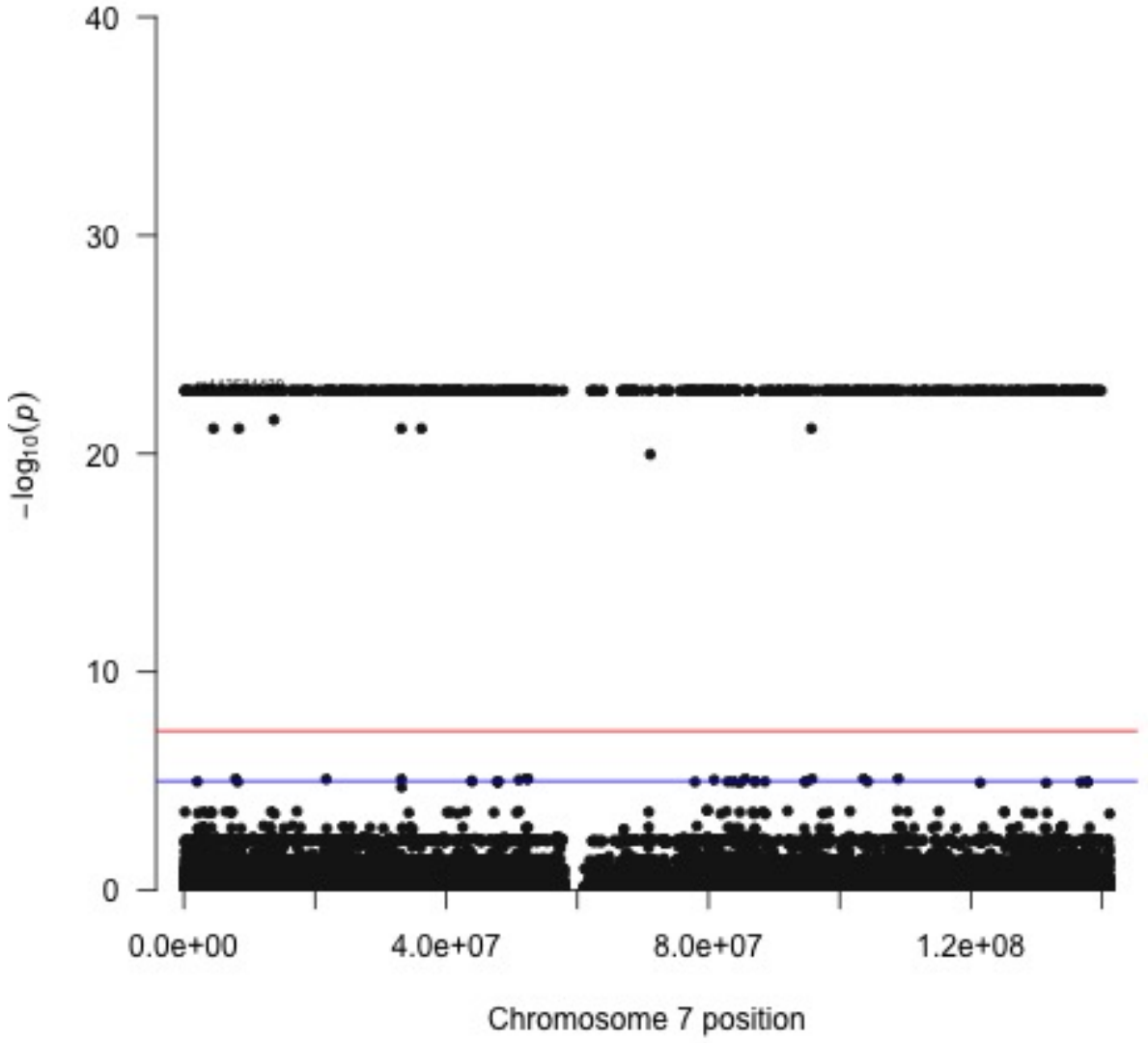
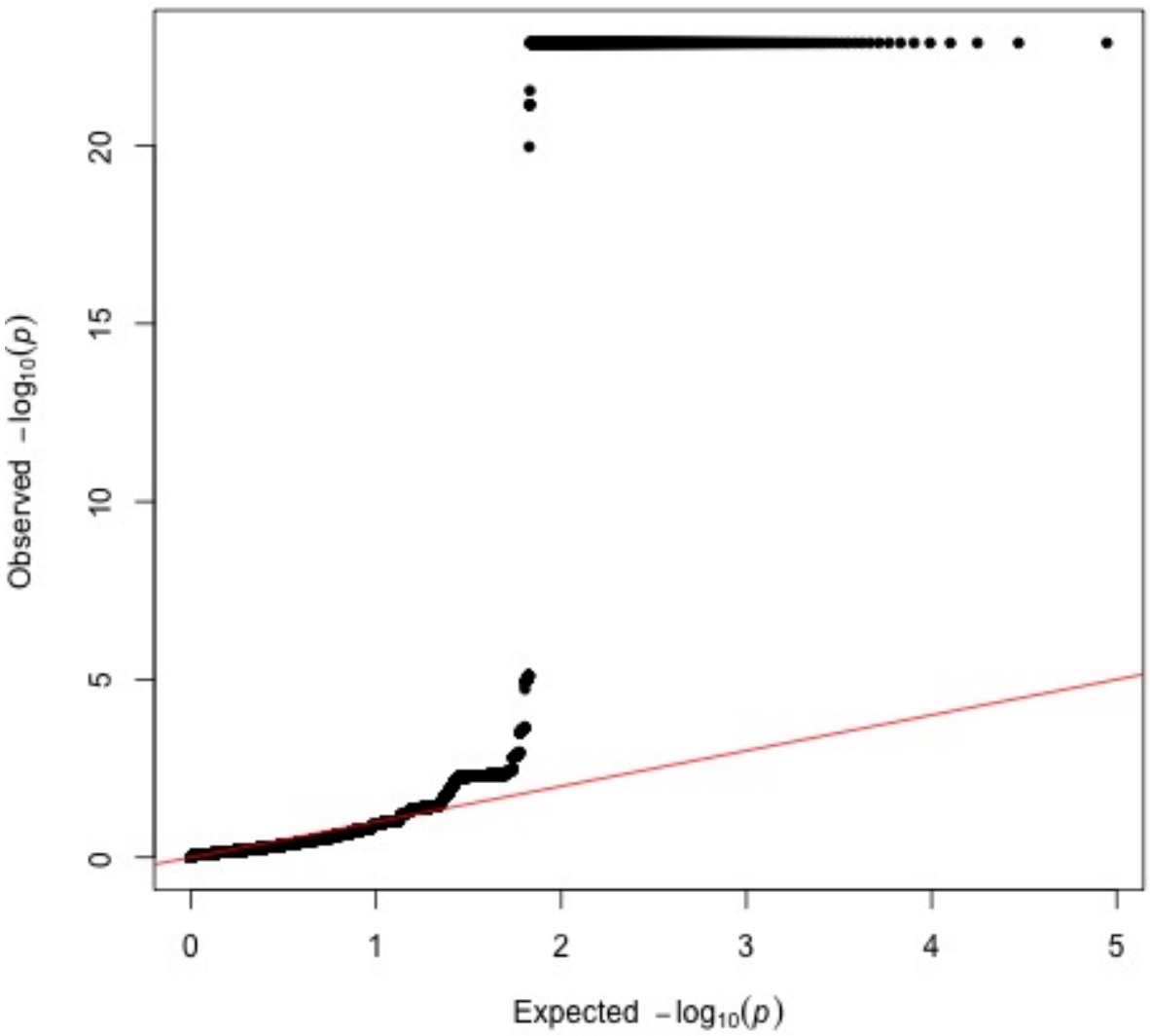


Figure 5: Manhattan Plot

When viewing this plot, the x-axis refers to chromosomal position. Chr 7 is of interest because TRB is located on that chromosome between 141.2 and 142 Mb. Every point is a different SNP and the y-axis represents $-log(p)$, so higher values are more significant. Initial inspection shows a plethora of SNPS that could possibly be significant in the current data set. However, after looking at the Q-Q plot (Figure 6), there is a strong deviation around $E(-log(p)) = 2$. This warrants further investigation and may stem from absence of an allele in those cases. MAF filtering may correct this.



Step 12 produced initial visualizations of the immunoSeq (TCR productive clonality) data, using the LymphoSeq package mentioned earlier. First, out of frame and early stop nucleotide sequences were filtered from the data to produce productive sequences.

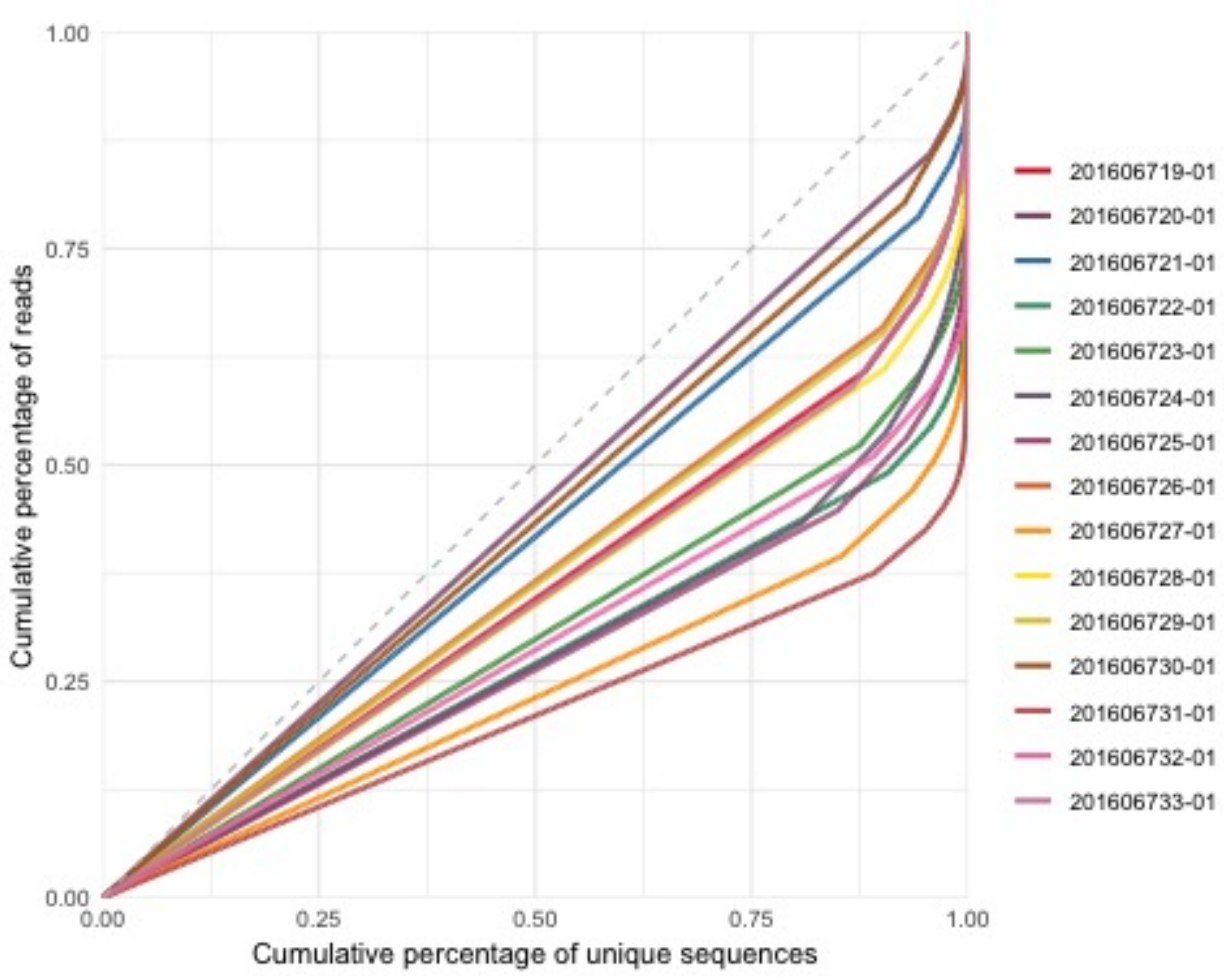
```
library(LymphoSeq)

# TODO: Need to store on the biolync server with RCurl
# Use LymphoSeq to read in all samples
TCR.list <- readImmunoSeq(path =
            "/Users/linjo/Box Sync/MIPS/ImmunoSeq_exported_samples")

# Filter for productive sequences
productive.TCR.nt <- productiveSeq(file.list = TCR.list,
                                   aggregate = "nucleotide", prevalence = FALSE)

# Generate and save visualization plots
# TODO: use Rcurl to read this in from server
jpeg('lorenz1.jpg')
lorenzCurve(samples = names(productive.TCR.nt), list = productive.TCR.nt)
dev.off()
jpeg('topSeq1.jpg')
topSeqsPlot(list = productive.TCR.nt, top = 10)
dev.off()
```

The Lorenz-curve in Figure 7, 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.


1.00
0.75
0.50
0.25
0.00
Cumulative percentage of reads
0.00
0.25
0.50
0.75
1.00
Cumulative percentage of unique sequences
201606719-01
201606720-01
201606721-01
201606722-01
201606723-01
201606724-01
201606725-01
201606726-01
201606727-01
201606728-01
201606729-01
201606730-01
201606731-01
201606732-01
201606733-01

Similarly, Figure 8, shows frequences of the top 10 clones in different colors for all 15 patients. The purple color denotes all other clone types.

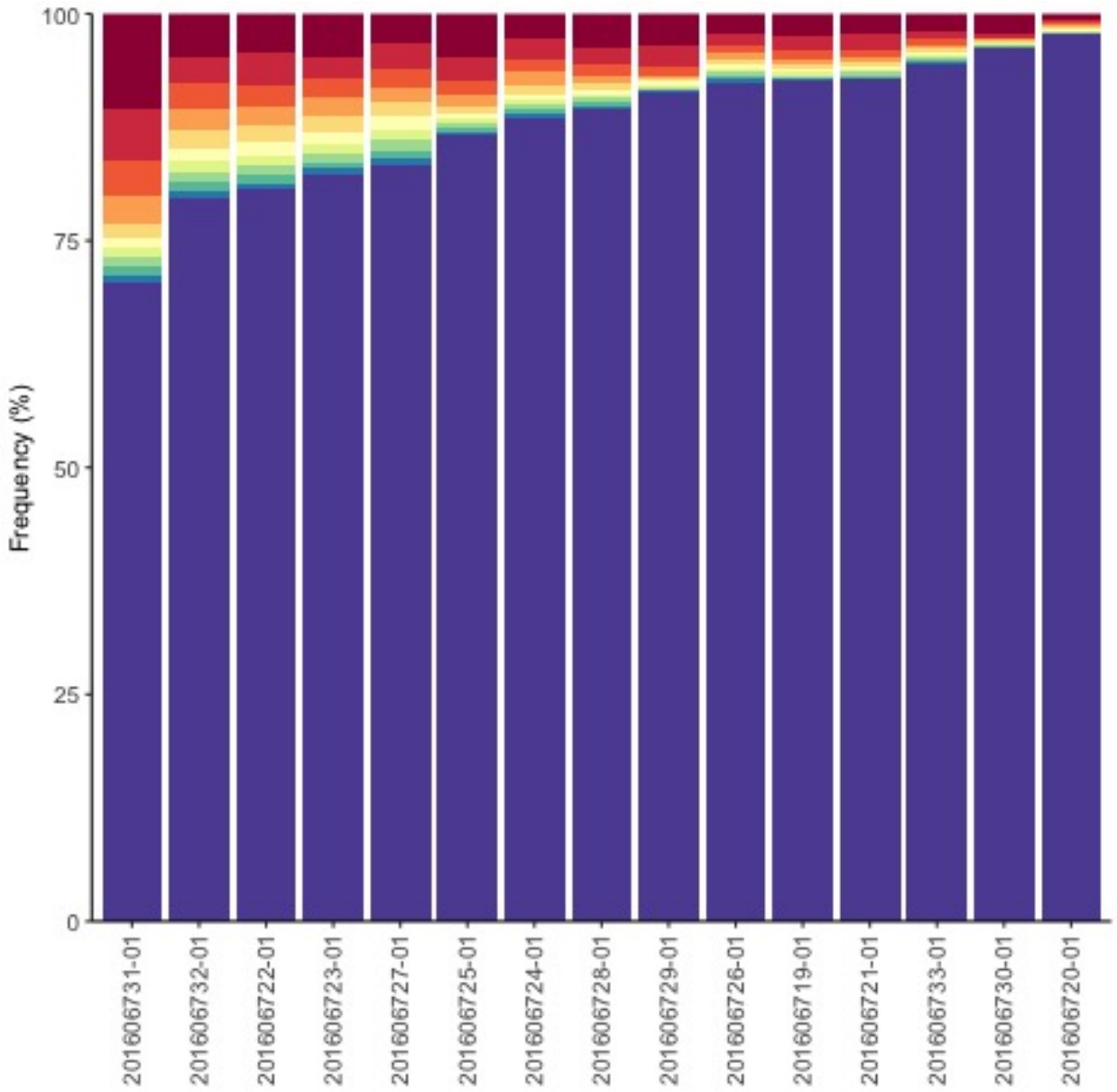


Figure 6: Lorenz Curve

Discussion

Conclusion thus far

We acknowledge the small sample size of this study, but also emphasize the exploratory rather than predictive aims. From the above, preliminary analysis of the genotype and immunoSeq data was accomplished. The Manhattan and QQ plots show a need for further refinement of the data, as noted in the steps for Quality Control. Much of work done thus far has entailed familiarizing with previous work, studying the genetic and biological prerequisite knowledge and becoming comfortable with the tools used. Therefore, there can be no significant SNP predictors in relation to TCR diversity identified at this time. It should be noted the number of SNPs may be reduced with further filtering.

Next Steps and Improvements

There are a number of additional tasks. Mainly, continuation of quality control to eliminate any noise as noted above. Additionally, the genetic window chosen (the TCRB gene) may be too wide and limiting this window may lead to different findings. Adjustments to the model used (the additive SNP model) could be made as well as adjusting p-values to account for multiple testing. Finally, the data and code base are spread across a variety of different platforms including servers, biolync.case.edu and hpc4, and a Box folder (for the immunoSeq data). Consolidating this in order to facilitate reproducible work is another endeavor.

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

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.

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. https://doi.org/10.1002/mpr.1608.

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

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. https://doi.org/10.1086/519795.

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