INCLUSION OF 48 PACIFIC ISLANDERS WITHIN A COSMOPOLITAN REFERENCE PANEL IS SUFFICIENT FOR HIGH ACCURACY GENOTYPE IMPUTATION OF SAMOANS

by

Kevin Anderson

B.S., University of Pittsburgh, 2020

Submitted to the Graduate Faculty of the

Department of Human Genetics – Genome Bioinformatics

School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH

SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Kevin Anderson

It was defended on

April 22, 2022

and approved by

Jenna C. Carlson, Ph.D.
Assistant Professor, Departments of Biostatistics and Human Genetics
School of Public Health, University of Pittsburgh

Daniel E. Weeks, Ph.D.,
Professor, Departments of Human Genetics and Biostatistics
School of Public Health, University of Pittsburgh

Thesis Advisor: **Ryan L. Minster, Ph.D., M.S.I.S.**
Assistant Professor, Department of Human Genetics
School of Public Health, University of Pittsburgh

INCLUSION OF 48 PACIFIC ISLANDERS WITHIN A COSMOPOLITAN REFERENCE PANEL IS SUFFICIENT FOR HIGH ACCURACY GENOTYPE IMPUTATION OF SAMOANS

Kevin Anderson, M.S.

University of Pittsburgh, 2022

Imputation is a computational method for inferring genotypes based on previous knowledge of shared haplotype structure commonly used in genome-wide association studies. Genotype frequencies not only play an important role in imputation but also are highly variable around the world, meaning it is crucial to adjust for population bias in genetic studies. Common methods for imputation involve the use of publicly available haplotype panels from 1000 Genomes, TOPMed, or other consortia. However, these panels contain data mostly pulled from individuals of European ancestry. Population isolates such as Polynesians greatly benefit in genotype accuracy when using a population-specific haplotype reference panel. Here, I perform multiple imputations using the 1000 Genomes phase III reference panel and genome-wide data from 1285, 384, 96, 48, 24, and 1 Samoan on chromosomes 5 and 21 to determine how many fully sequenced individuals are needed to include in study-specific haplotype panels to achieve accurate imputation. I also investigated the accuracy of these multiple imputations on genotype frequencies of population-specific variants found in the *CREBRF* and *BTNL9* genes that are previously determined to be associated with higher BMI and lower HDL levels respectively. I demonstrate that the incorporation of 96 Samoans within the 1000 Genomes cosmopolitan panel produces accurate imputation quality of rare variants (minor allele frequency of 1%), and 24 Samoans for common variants (minor allele frequency greater than 5%). These results show that the creation of a study-specific reference panel utilizing a small subset of individuals from a population-isolate within a

cosmopolitan panel is a cost-effective strategy for accurate imputation. The ability to perform fine-mapping on rare population-specific variants will have broad public health implications such as better understanding of genetic disease etiology and function and improved genetic literacy when focusing on these population isolates.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank the Samoan participants of the study, local village authorities, and the many Samoan and other field workers over the years. I would also like to thank the Samoan government, particularly the Ministry of Health; the Ministry of Women, Community, and Social Development; the Office of the Prime Minister; and the Samoa Bureau of Statistics for their continued support of this work. This study was supported by the US National Institutes of Health grant R01HL093093 (Principal Investigator: Stephen McGarvey, Brown University) and grant R01HL133040 (Principal Investigator: Ryan L. Minster, University of Pittsburgh).

Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). Genome sequencing for the Soifua Manuia study, labeled as "NHLBI TOPMed: Genome-wide Association Study of Adiposity in Samoans" (phs000972.v4.p1) in the dbGaP, was performed at the Northwest Genomics Center (HHSN268201100037C) and the New York Genome Center (HHSN268201500016C). Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01-HL117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01-HL120393; U01-HL120393; contract HHSN268201800001I). I gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

1.0 Introduction

A key aim in genetic epidemiology studies is to identify locations, or loci, across the genome where differences in genetic variation are associated with different physical effects, or phenotypes, between individuals. For over 15 years, genome-wide association studies (GWAS) have been crucial in identifying common and rare genetic variants associated with diseases, such as Alzheimer disease (Kamboh, 2004), or complex traits, such as height and eye color (Uffelmann et al., 2021). An important step in conducting a GWAS experiment performing imputation on your gathered genotype data. Imputation is the process of inferring missing genotypes based on previous knowledge of haplotype structure from GWAS datasets (Marchini & Howie, 2010; Naj, 2019). This step solves a problem in using genotyping arrays for large-scale genome-wide studies wherein there are not enough DNA probes on the microarray to physically genotype all the areas of interest across the entire genome. Therefore, there are substantial gaps in data between genotyped loci that are potentially useful for association studies. Imputation is a computationally intensive process that uses statistical inference to predict the unobserved genotypes using known haplotypes, or groups of alleles inherited together, in a population together with the pattern of observed genotypes.

Today, GWAS analyses are common-practice and are significantly easier to perform due to decreased costs in sequencing and the development of easy-to-use computational pipelines for association analysis. Haplotype reference panels are publicly available through institutions such as the National Heart, Lung, and Blood Institution (NHLBI) Trans-Omics for Precision Medicine (TOPMed) Program (Taliun et al., 2021) and the 1000 Genomes Project (Auton et al., 2015) . However, most GWAS studies to date have been conducted on individuals of European ancestry

(Kowalski et al., 2019), and genetic studies of other ancestry groups around the world are limited. In total, 95.82% of all GWAS participants before January 2022 are of European ancestry (Mills & Rahal, 2020). This creates a problem where it is unhelpful to use large amounts of previously published GWAS data on non–European individuals due to differences in genetic variation across ethnic ancestry groups around the world. Without the support of known haplotypes, imputation on underrepresented populations becomes unreliable (Quick et al., 2020). This is especially true in the case of the Polynesian people of Samoa. Polynesians make up less than one percent of the world's population. Samoans are a subpopulation of the Polynesian ancestry group, where there are no reported haplotypes from Polynesian individuals in the 1000 Genomes database (Auton et al., 2015), and there are three individuals of calculated Polynesian ancestry reported in the TOP-Med freeze X dataset, excluding the Samoans that are included in those data.

To accurately and completely genotype individuals, you must fully sequence them. However, even today, fully sequencing enough individuals to gather complete genotype data for a powerful association study is still relatively expensive. Since 2017, cost of whole-genome sequencing (WGS) has dropped below $1,000 per sample (Karow, 2017). A solution to this problem is to first create a population-specific haplotype reference panel by sequencing a subset of the study group and incorporating that data into a larger, cosmopolitan reference panel (Ahmad et al., 2017). Then, use the combined panel in conjunction with genotypes from a genotyping array, which will have many markers with unmeasured genotypes, to fill in the genotypes at those unmeasured markers. Genotyping via DNA microarrays cost around $28–$90 per sample (Peng et al., 2017), therefore, genotyping 2,000 individuals with a genotyping array would save over $500,000 as opposed to sequencing them.. However, exactly how many individuals are necessary to incorporate into a cosmopolitan reference panel to achieve good imputation accuracy has not been determined . This

number, and the number of haplotypes is predicted to differ between different population groups (Mitt et al., 2017).

Previous work in obtaining microarray and WGS data has been conducted by members of the Obesity, Lifestyle, and Genetic Adaptations (OLaGA; “life” in Samoan) Study Group. The aim of OLaGA research is to assess the behavioral, environmental, and genetic determinants of adiposity and cardiometabolic risk in Samoans. . They have recently identified two variants, one missense variant in *CREBRF* associated with higher body mass index (BMI) and lower odds of type 2 diabetes (T2D) and one nonsense variant in *BTNL9* associated with lower high-density lipoprotein (HDL) levels. These variants are extremely rare in other populations but common in Samoans,and are hypothesized to have a reduced risk of type 2 diabetes.

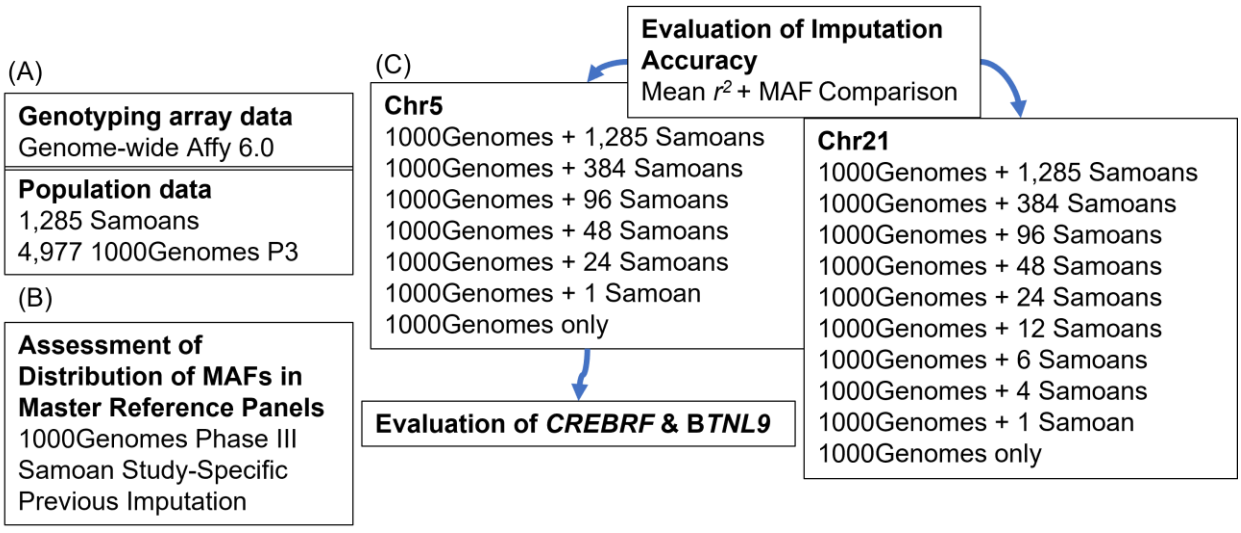

Figure 1. Study Overview

Table 1. Number of Polynesians in haplotype reference panels used in this study

Haplotype Reference Panel	Number of Individuals	Number of Polynesians
1000 Genomes	4,974	0
1000 Genomes + Study-Specific	6,259	1,285

In this study, I will compare the distribution of minor allele frequencies (MAFs) between the cosmopolitan reference panel from 1000 Genomes, the study-specific reference panel from Samoan sequencing data, and the previous imputation conducted by Minster et al. in 2016 (Figure 1B). I will then determine the minimum number of sequenced Samoans necessary to include with a cosmopolitan reference panel to achieve accurate imputation quality of Samoan genotypes by analyzing average r^2 values for variants on chromosomes 5 and 21. I will also examine the difference in allele frequencies between previously discovered variants in *CREBRF* (rs373863828) and *BTNL9* (rs200884524) on chromosome 5 when imputing with different numbers of Samoans (Figure 1C). The reference panels will be based on sequencing data from 1,285 Samoans and individuals from 1000 Genomes Phase 3 (Figure 1A).

The public health aim of this analysis is to provide better imputation results for future genetic studies on Samoans by incorporating Samoan-specific variants in an otherwise cosmopolitan haplotype reference panel Higher density of Samoan-specific alleles allows more straightforward fine mapping of rare variants such as the previously described variants in *CREBRF* and *BTNL9*. Increased genotyping accuracy will help understand disease etiology, risk, prevention, diagnosis and treatment (Molster et al., 2018). Also, by incorporating genomics techniques in a public heath-focused study on a population isolate, we can improve the genomic literacy of the public to further inform, educate, and empower people about health issues (McWalter & Gaviglio, 2015) thereby appropriately integrating genomic technologies into all aspects of healthcare (Bowen et al., 2012).

2.0 Methods

2.1 Participants, Phenotypes, and Genotypes

GWAS data statistics and WGS data are from a study conducted in 2010 by Hawley and colleagues (Hawley et al., 2014). This study aimed to assess the prevalence of adiposity and cardiometabolic genetic risk factors, specifically in Samoans. This study recruited 3,475 participants ranging from 24.5 to 65 years old (1,437 male; 2,038 female). Recruitment criteria involved participants from 33 villages located in the independent nation of Samoa who have four grandparents of Samoan origin, are non-pregnant, have no physical or cognitive impairment, and can complete an interview. Sample collection methods include fasted early morning blood samples, anthropometric measurements, blood pressure, and body composition. All participant recruitment, characterization, and genotyping described in this study were conducted by researchers who now comprise the Obesity, Lifestyle, and Genetic Adaptations (OLaGA; "life" in Samoan) Study Group.

Body mass index (BMI) was calculated using weight divided by the square height in meters. Serum samples were obtained after a 10 h fast in 10 ml vacutainers spray-coated with silica-containing polymer gel to separate the serum later stored in a −40 °C environment. LDL and HDL levels were determined using enzymatic in vitro tests on a Roche automated analyzer executing the glucose hexokinase method (H.U. Bergmeyer, K. Gawehn, 1974). Both verbal and signed consent was given through signed consent form, spoken and written in Samoan and English, which included uploading their genetic data to the Database of Genotypes and Phenotypes (dbGaP). The Brown University Institutional Review Board and the Health Research Committee of the Samoan Ministry of Health approved consent. Twenty-nine participants were later excluded based on the

inclusion criteria, and two based on incomplete data, totaling the study population to 3,475 participants. Of these participants, 91.1% provided a blood sample, and 84.6% provided a fasted serum sample. To better understand the genetic factors influencing BMI in Samoans, the OLaGA study group genotyped a discovery sample of 3,298 samples (3,194 participants, 34 duplicates, and 70 controls) using the Genome-Wide Human SNP Array 6.0 (Affymetrix, California, USA). Standard quality control measures were used, such as removal of probes with greater than 5% missingness, sex validation, relatedness, population substructure, discordance, and controlling for batch effects (Gogarten et al., 2012; Laurie et al., 2010). After quality control, complete phenotype and genotype data are available for 3,072 participants. Genotype data were phased using Eagle followed by imputation with the 1000 Genomes Phase 3 reference panel.

Ninety-six individuals underwent targeted sequencing on a 1.5 Mb region around the missense variant rs12513643 on chromosome 5q35.1. This site was highly associated with BMI ($p = 5.3 \times 10^{-14}$) and was replicated ($p = 1.2 \times 10^{-9}$) in 2,103 participants from previous studies (Minster et al., 2016). Sequencing data was imputed using SHAPEIT and IMPUTE2 against the December 2013 1000 Genomes Project Phase 1 Integrated variant set release haplotype reference panel. Initial imputation yielded poor accuracy results when imputing on a cosmopolitan haplotype panel, therefore a Samoan-based reference panel is necessary for accurate imputation. A subset of the discovery cohort (n = 1,285) underwent WGS by the National Heart, Lung, and Blood Institute (NHLBI) TOPMed Consortium (Minster et al., 2016). Sequencing data from these previous analyses were used to create a study-specific imputation panel containing 1,285 sequenced Samoans and 4,974 individuals from 1000 Genomes primarily of European ancestry, referred to as the master reference panel.

2.2 Analysis

All analysis was conducted on the University of Pittsburgh School of Public Health Department of Human Genetics' high-performance computational cluster, dubbed the GATTACA cluster. This cluster hosts the resources necessary to handle large amounts of imputation data. Phenotypic data from the three cohorts were imported into R. Each cohort contained subject identification numbers, age, sex, and BMI fields. The distribution of sex for each cohort was compared against age to visualize distribution. The BMI and HDL levels distribution was stratified by sex and regressed against age.

2.2.1 Previous Genotyping: Scaffold and Imputation

The OLaGA Study had previously created an imputation scaffold containing the genotype data from the Affymetrix 6.0 array used in their study and imputed participants of a targeted WGS study against the 1000 Genomes Project Phase I integrated variant set release haplotype reference panel. Pre-phasing of sequencing samples was performed using *SHAPEIT* and imputed using *IMPUTE2*. For each sequence variant contained in the scaffold, I extracted the MAFs and imputation r^2 values from the imputation variant call format (VCFs) files using *bcftools query* and visualized them using R. I calculated MAFs using *bcftools query* to pull the AC and AN from each variant and divide AC over AN to obtain the allele frequencies at each variant. Allele frequencies that were greater than 0.5 were flipped by subtracting their frequency by 1. The total number of variants with each MAF was calculated using Unix *sort* and *uniq* functions. I then plotted the MAFs in the scaffold and compared them against a subset of the master reference panel containing only the Samoan study participants.

Previous imputations on the targeted WGS participants of the OLaGA study were conducted by Minster et al., referred to as the discovery set. Imputation was performed against the December 2013 1000 Genomes Project Phase I Integrated variants set release haplotype panel. I pulled the chromosome number, base pair position, MAF, and r^2 values from each VCF using *bcftools query*. I calculated the number of variants at each MAF using the Unix tools *sort* and *uniq* and imported them into R. I plotted the MAFs for the discovery set separately as counts. Then, I compared MAFs in the discovery set to the MAFs found in the Samoan-only reference panel, stratified by an r^2 threshold of 0.8.

2.2.2 Creation of Study-Specific Imputation Panels

The master reference panel contains phased haplotype data from 1000 Genomes, and the 1,285 sequenced Samoans from the OlaGA study. To better understand the distribution of the allele frequencies, I partitioned the master reference panel using *bcftools* to obtain two datasets, one of just Samoan participants and one of 1000 Genomes participants. I used sample ID lists containing Samoan and 1000 Genomes ID numbers as the *--samples-file* parameter in *bcftools query* to do the partitioning. I calculated MAFs for each variant in both datasets using *bcftools* and Unix commands. I used *bcftools query* to pull the AC (total alternate, or minor, allele count) and AN (total alleles called in genotype) from each variant and divide AC over AN to obtain the allele frequency at each variant locus. Allele frequencies that were greater than 0.5 were flipped by subtracting their frequency by 1.

2.2.3 Imputing a Subset of Samoans Within a Cosmopolitan Reference Panel

I converted the phased haplotype data from the master reference panel, which includes the 1000 Genomes and the 1,285 sequenced Samoan participants, to vcf format and compressed using *bcftools* to prepare the data for imputation. The first step of imputing the haplotype data against the master reference panel I performed using Minimac3 (Das et al., 2016), creating a customized Minimac3 vcf file (m3vcf). Parameters for Minimac3 included parallel processing on five computer processing units (CPUs) with 20 GB of random-access memory (RAM). Lastly, I performed imputation using Minimac4 on the m3vcf against the Affy 6.0 array scaffold and using the *–allTypedSites* parameter with a chunk length of 10 Mb and an overlap of 3 Mb on 10 CPUs.

I conducted serial imputations, each based on the 1000 Genomes as a cosmopolitan core, with successively higher numbers of Samoan individuals included in the haplotype reference panel. Increasing the number of Samoan individuals included in the master reference panel will change the variants and allele frequencies found in that panel and improve how well imputation performs. I imputed using 0, 1, 24, 48, 96, 384, and 1,285 individuals on chromosome 5, and 0, 1, 4, 6, 12, 24, 48, 96, 384, and 1,285 individuals on chromosome 21. The individuals included were not random; instead, the Samoan participants were ranked by "informativeness" of their haplotype information. To visualize the effects of imputation accuracy for each increase of Samoan individuals to the reference panel, each log file was filtered to MAF and r^2 fields where each variant must have a MAF more significant than 0. Then, I calculated the average r^2 values for each MAF, stratified by the number of Samoans included in that imputation, and analyzed imputation accuracy at low MAF.

2.2.4 Allele Frequency Analysis of *CREBRF* & *BTNL9*

I identify differences in allele frequencies in the previously discovered variants in the *CREBRF* (5q35.1) and *BTNL9* (5q35.3) gene in addition to analyzing imputation accuracy (Minster et al., 2016). A set of 0, 24, 48, 96, 384, and 1,285 Samoans are included in a cosmopolitan haplotype panel and imputed independently on chromosome 5. Genotype frequencies were calculated from the imputed genotypes from the *vcf* file generated by Minimac4 using *Perl*. Genotype frequencies are compared from the study-specific panels and the imputed study-specific panels to analyze how accurate imputation is performed at these variants with varying numbers of Samoans.

3.0 Results

3.1 Phenotype Data

All GWAS statistics and imputation analyses were conducted on the University of Pittsburgh Graduate School of Public Health Department of Human Genetics' high-performance computational cluster, GATTACA.

Sex and age distribution was calculated from the GWAS data obtained from the OLaGA study. This study contained a sample size of 3,092 participants (1,247 male and 1,845 female) with an average age of 45.34 years and 44.69 years and average BMI levels of 31.31 and 34.90, respectively, with a correlation coefficient of 0.101 between male and female participants between age and BMI. HDL data were extracted from 1,211 male and 1,039 female participants with average levels of 44.14 mg/dL for male and 42.43 mg/dL for female participants, and a correlation coefficient of −0.248 between age and HDL.

3.2 Samoan and 1000 Genomes Haplotype Panel Comparisons

The master reference panel of the phased haplotype data from 4,974 1000 Genomes and 1,285 sequenced Samoan study participants contained 53,775,719 and 10,870,873 variants located on chromosomes 5 and 21 respectively. To compare the MAFs between the two panels I split them into two separate datasets. MAFs from each set were visualized as counts, with distribution highly skewed to the right because high MAF variants are much less common. The correlation of MAFs

between the 1000 Genomes and Samoan reference panels was visualized using R *tidyverse* and *ggplot* packages (Figure 2). Pearson's product-moment correlation of MAFs between these two panels was 0.791 on chromosome 5 (95% confidence interval 0.785, 0.797) and 0.782 on chromosome 21 (95% confidence interval 0.775, 0.788). Because there were more participants in the 1000 Genomes panel vs. the Samoan panel (4,975 and 2,953 respectively), there was a lot higher frequency of MAFs in the 1000 Genomes panel compared to the Samoan panel; however, the distribution of those frequencies was noted to be similar. With the majority of MAFs between 0.0002 and 0.00025 for both panels.

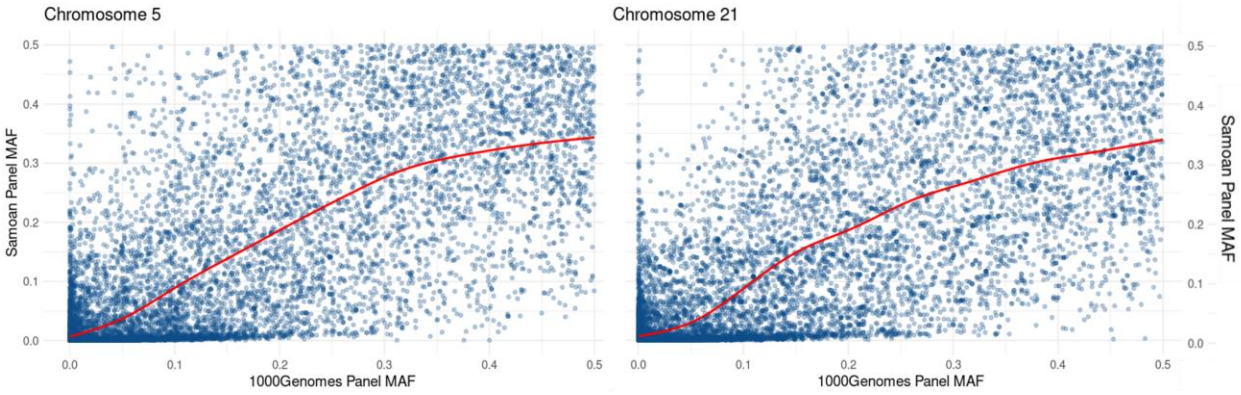

Figure 2. Comparison of panel MAFs

The 1000 Genomes panel (*x* axis) and the Samoan study-specific panel (*y* axis) on chromosome 5 (left) and 21 (right).

3.3 Scaffold and Previous Imputation Analysis

The Affymetrix (Affy) 6.0 imputation scaffold contained the genotype information for participants of the OLaGA study. There are 55,764 and 12,387 variants on chromosomes 5 and 21 on this scaffold. MAFs for the affy scaffold were plotted as counts, and the distribution of MAFs was

similarly skewed like the reference panel. The correlation of MAFs between the Affy scaffold (n = 1,834) and Samoan reference panels (n = 1,285) was visualized (Figure 3). These two datasets, were checked to observe how similar the allele frequencies were between the genotyped and sequenced Samoan participants. Pearson's product-moment correlation of MAFs between these two panels was 0.9978 on chromosome 5 (95% confidence interval 0.9976, 0.9979) and 0.9966 on chromosome 21 (95% confidence interval 0.9964, 0.9968). This very high correlation can be explained by both panels containing just Samoan participants of known Samoan ancestry; therefore, we expect the MAFs for each variant to be relatively the same.

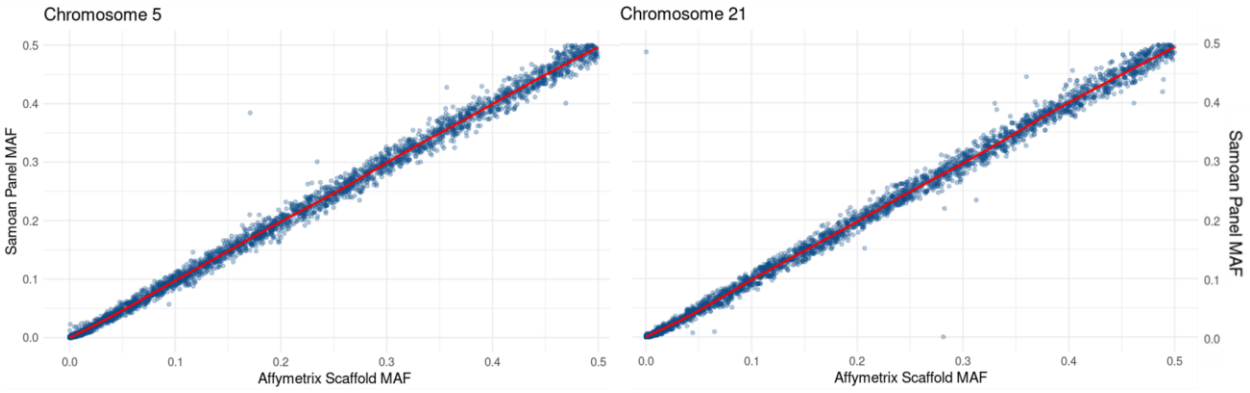

Figure 3. Comparison of scaffold and panel MAFs

The Affymetrix scaffold (x axis) and the Samoan specific panel (y axis) on chromosome 5 (left) and 21 (right).

Previous imputations on the genotyped Samoan individuals who were separated from the WGS subset of the OLaGA study were conducted by Minster et al., referred to as the discovery set. In the discovery imputation set, there are 3,392,329 and 727,199 variants on chromosomes 5 and 21 respectively from 5,623 participants; 3,119 genotyped Samoans and 2,504 samples from 1000 Genomes. MAFs were then compared from the existing imputation discovery set against the master reference panel containing only the 1,285 Samoan participant data, excluding

1000 Genomes data (Figure 4). The MAFs between the previous imputation set and the reference panel containing only Samoans were expected to be highly correlated, which is why I excluded 1000 Genomes data. The variants from the discovery set were joined with the variants from the Samoan panel by base pair position and plotted, stratifying by r^2 values above or below 0.8. I found 605,490 variants with an r^2 value below 0.8 and 955,688 variants above 0.8 r^2. The discovery set was highly correlated with the Samoan-only panel, with a coefficient of 0.9987 and 0.9988 on chromosomes 5 and 21 respectively.

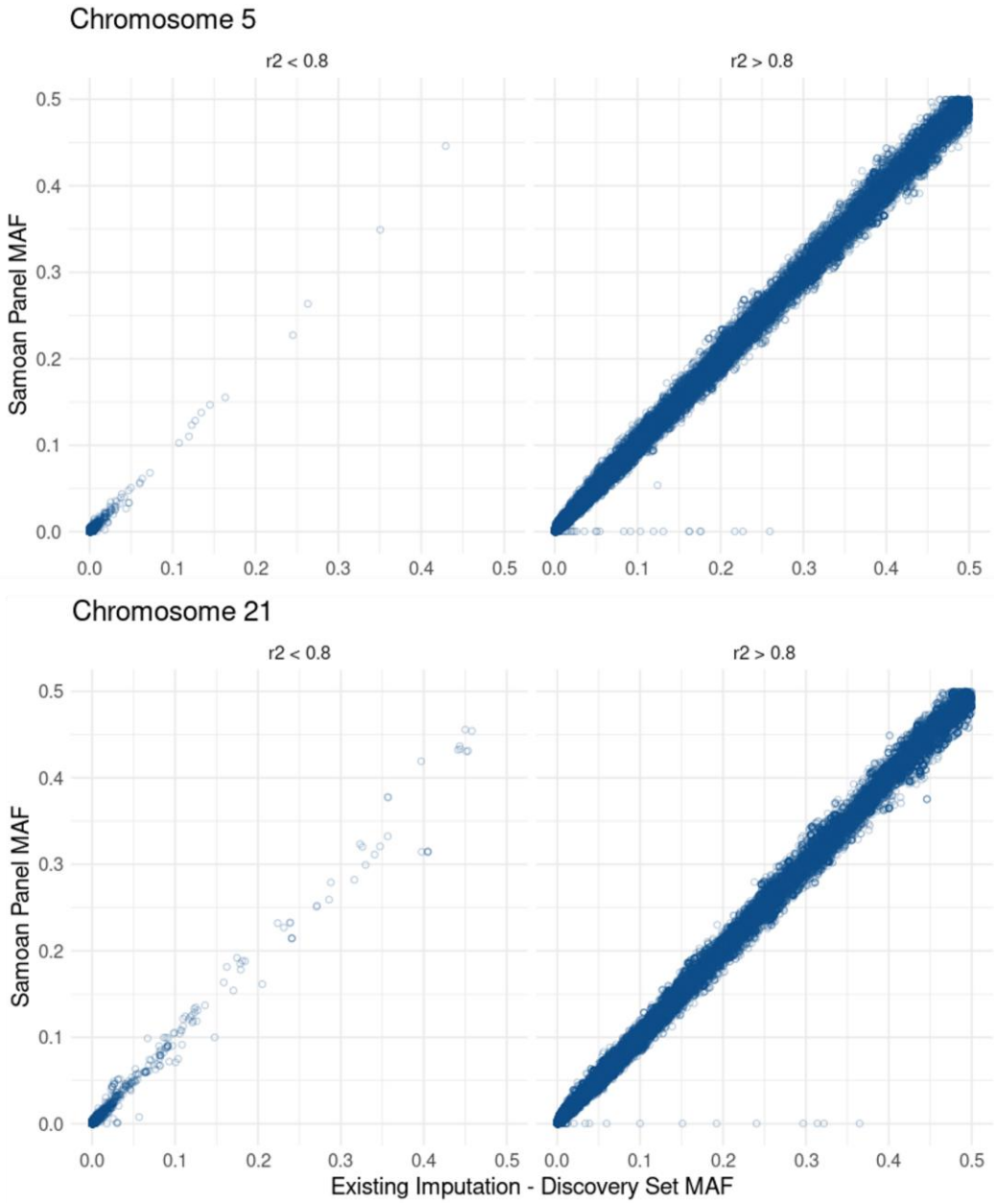

Figure 4. Comparison of panel and imputation MAFs

The existing imputation discovery set (*x* axis) and the Samoan specific panel (*y* axis) on chromosome 5 (top) and 21 (bottom) stratified by variants with r^2 values below 0.8 (left) and above 0.8 (right).

3.4 Serial Imputations

Decreasing the number of Samoan individuals included in the master reference panel will change the variants and allele frequencies found in that panel, changing how well imputation performs. Imputation was repeated using a subset of 0, 1, 24, 96, 48, 384, and 1,285 Samoan individuals on chromosome 5 (Figure 5) and 0, 1, 4, 6, 12, 24, 48, 96, 384, and 1,285 Samoans on chromosome 21 (Figure 6). Chromosome 21 was imputed with a wider array of individuals to further narrow down the number of individuals to include in a cosmopolitan haplotype panel to achieve accurate imputation. This chromosome was chosen for extra imputations due to the time saved of running imputation on a much smaller set of variants. The selection of samples used was not performed randomly; instead, the Samoan participants were ranked by "informativeness" of their haplotype information. Imputation analysis used data obtained from Minimac4 *.log* files, which contain a summary of imputation statistics such as SNP coordinates, reference, and alternative alleles, MAF, average call rate, and r^2 statistic. The r^2 statistic is crucial because it explains how well that variant is imputed. Minimac3 defines this value as the estimated correlation between imputed genotypes and true, unobserved genotypes, calculated by observed dosage variance over the expected dosage variance, given observed allele frequency, and assuming Hardy–Weinberg equilibrium (Das et al., 2016). As more Samoans are added to the reference panel with 1000 Genomes, imputation accuracy increases at every MAF, especially at frequencies below 1%.

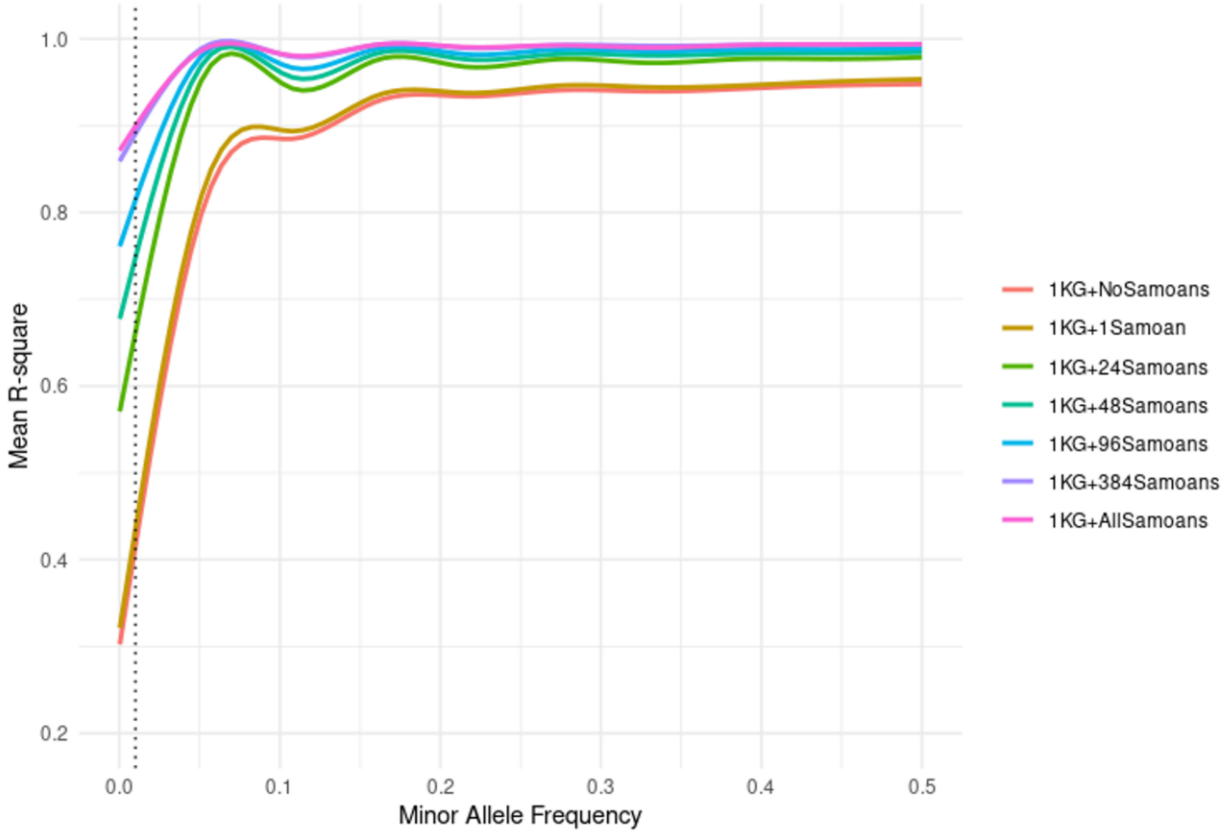

Figure 5. Imputation accuracy of chromosome 5 across MAFs

Imputations used 0, 1, 24, 48, 96, 384, and 1,285 Samoans within the 1000 Genomes haplotype panel. 1% MAF illustrated with dotted line.

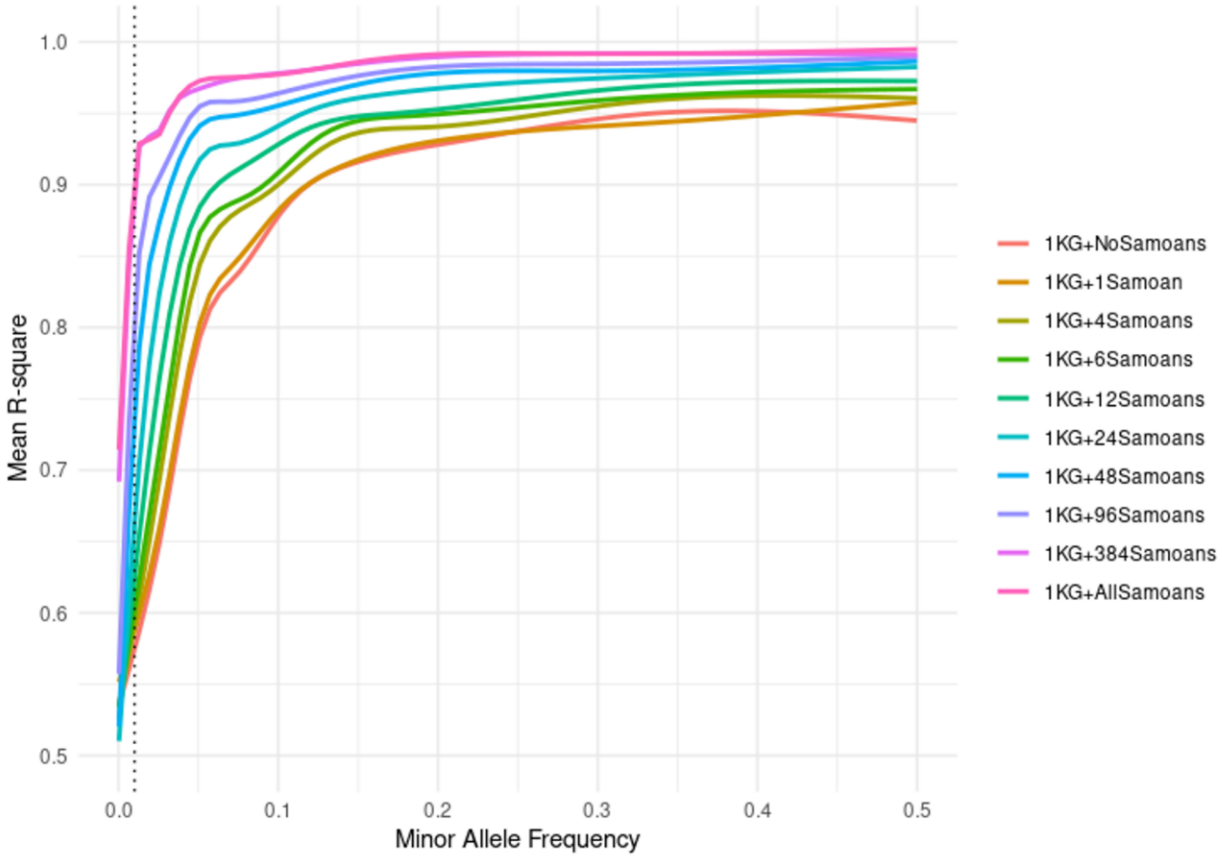

Figure 6. Imputation accuracy of chromosome 21 across MAFs

Imputations using 0, 1, 4, 6, 12, 24, 48, 96, 384, and 1,285 Samoans within the 1000 Genomes haplotype panel. 1% MAF illustrated with dotted line.

For imputation of rare variants (MAF =1%), the inclusion of 48 Samoans within the 1000 Genomes panel achieved accurate imputation with an r^2 value of 0.816 across all variants in chromosome 5. For chromosome 21, the inclusion of 384 Samoans within the 1000 Genomes panel achieved similar imputation accuracy at a MAF = 1% with $r^2 = 0.814$. However, when increasing the MAF threshold to 1.3%, you will get accurate imputation with 96 Samoans with an $r^2 = 0.798$ on chromosome 21. Complete information involving imputation accuracy across a variety of MAFs is contained in appendix tables 1 and 2.

Table 2. Imputation accuracy of chromosome 5 at MAF = 1%

Samoans Added	Mean r^2
0	0.432
1	0.423
24	0.677
48	0.816
96	0.831
384	0.938
1,285	0.912

Table 3. Imputation accuracy of chromosome 21 at MAF = 1%

Samoans Added	Mean r^2
0	0.358
1	0.293
4	0.395
6	0.287
12	0.675
24	0.760
48	0.747
96	0.704
384	0.813
1,285	0.886

3.5 Genotype Effects of *CREBRF* and *BTNL9* Variants with Different Number of Imputed Samples

Chromosome 5 is a chromosome of interest due to previously identified variants in *CREBRF* and *BTNL9* that are largely specific to the people of Samoa (Minster et al., 2016). To analyze the effects of imputation on a varying number of individuals added to a cosmopolitan reference panel, I looked at the genotypes from these variants in both the imputed and unimputed

panels. Table 4 demonstrates that imputation successfully predicted the genotype frequencies of *CREBRF* from the panel with 24 Samoans added. These imputed genotypes equate to a MAF of ~0.28 (r^2 = 0.99894), which is similar to what Minster et al. discovered in 2016 where the study determined a 0.276 MAF at this variant. Unfortunately, this variant was not seen in the haplotype of the one Samoan individual included in the 100Genomes reference panel. However, imputation on the *BTNL9* variant contained results for the reference panel containing one Samoan. For this variant, there is high imputation accuracy across every subset of Samoans added to the haplotype panel (Appendix B). Genotype frequencies also stay consistent across every panel (Table 4) leading to suggest that only one individual may be necessary to impute common variants that are population specific. Therefore, with these results and the results from Table 2, imputation performs accurately with only 48 individuals added to the 1000 Genomes reference panel for rare variants (MAF ~ 1%), and 24 individuals added for accurate imputation of common variants (MAF > 5%) depending on chromosome.

Table 4. Genotype frequencies of the rs373863828 in *CREBRF* and rs200884524 in *BTNL9*

Samoans Added		1	24	48	96	384	1,285
CREBRF (genotyped)	GG	100.0%	99.5%	99.0%	98.3%	94.0%	84.2%
	GA	0.0%	0.39%	0.08%	1.4%	5.0%	13.1%
	AA	0.0%	0.0008%	0.002%	0.3%	1.0%	2.7%
BTNL9 (genotyped)	CC	99.9%	99.8%	99.0%	98.5%	94.8%	86.5%
	CT	0.1%	0.2%	0.8%	1.2%	4.5%	11.7%
	TT	0.0%	0.0%	0.2%	0.3%	0.7%	1.8%
CREBRF (imputed)	GG	100.0%	52.1%	52.2%	52.1%	52.1%	52.2%
	GA	0.0%	39.5%	39.4%	39.5%	39.5%	39.4%
	AA	0.0%	8.4%	8.4%	8.4%	8.4%	8.4%
BTNL9 (imputed)	CC	63.2%	61.4%	61.8%	62.5%	60.9%	60.9%
	CT	31.6%	32.8%	32.5%	32.2%	33.3%	33.4%
	TT	5.2%	5.8%	5.7%	5.3%	5.8%	5.7%

4.0 Discussion

In this study, I combined increasing numbers of individuals of Samoan ancestry with the 1000 Genomes imputation panel to impute GWAS data and observe imputation accuracy. GWASs have provided excellent coverage for European ancestry populations for over ten years. Using population-specific reference panels will lead to better imputation. This experiment shows that the inclusion of 48 individuals in a population-specific reference panel will lead to higher imputation accuracy than using strictly cosmopolitan reference panels.

Phenotypes from the Obesity, Lifestyle and Genetic Adaptations (OLaGA; "life" in Samoan) study in 2010 were visualized and used to create summary statistics to understand the GWAS data distribution better. I found that the study contained a higher number of women than men, who had a higher average BMI. Sex would not affect imputation accuracy, as the chromosomes analyzed are autosomal, however phenotypic effects of rare variants potentially have a different effect size based on sex due to gene by environment interactions.

WGS on 1,285 participants provided sequencing data to create a study-specific haplotype reference panel specific to the Samoan haplotypes. This information was combined with the 1000 Genomes Phase III haplotype reference panel to create a custom imputation panel, containing both Samoan-specific haplotypes with common haplotypes found in most European samples, which was beneficial for the accurate imputation of genotyped Samoans. Chromosome-wide MAF distribution did not appear to differ between these two panels when analyzed as separate datasets. However, more individuals were included in the 1000 Genomes panel compared to the Samoan-specific panel. The Affymetrix (Affy) 6.0 Array was used for genotyping Samoans from the OLaGA study and the creation of the imputation scaffold. Scaffolds are phased genotype data that

do not include many variants and have substantial gaps of genetic information along the genome. These serve as the 'target' genotypes compared against the reference panel that will fill in these gaps during imputation.

When conducting the imputation accuracy comparisons between certain thresholds of Samoans added to a cosmopolitan panel, there are a few reasons I included certain thresholds. Imputation panels containing one and all Samoan participants were chosen as the extremes. Next, 384 of the top participants were selected as a rough quartile of the total number of individuals. Next, 96 individuals were chosen because that is the number of samples that comprise one well plate for genotyping. Then, a panel of 48 and 24 Samoans was chosen to narrow down imputation accuracy because I hypothesized the minimum number of Samoans would be around this number. And lastly, an imputation panel containing only one Samoan individual is compared. Chromosome 21 saw extra imputations with 12, 6, and 4 Samoan individuals added. This is because there are considerably fewer variants on chromosome 21 compared to chromosome 5, and it is possible to observe the behavior of imputation accuracy at more precise measurements while taking advantage of faster processing due to considerably fewer variants on this chromosome. Based on the results, there is a large increase in imputation accuracy between the single Samoan dataset and 96 datasets, which is expected. The 96 and 384 panels achieved similar results, with the 384 Samoan panel showing a minor increase in imputation accuracy. When looking at a rare MAF threshold (≈ 1%), imputation with 48 Samoans on chromosome 5 produced a fairly accurate outcome at $r^2 \approx 0.816$. However, imputation with 48 Samoans on chromosome 21 produced a less accurate imputation at $r^2 \approx 0.747$. This is possibly the result of chromosome 21 having 874,736 variants with a MAF greater than 0, while chromosome 5 has 3,865,258 variants with a MAF greater than 0. Imputation accuracy also correlates with MAF (Appendix A) where common variants are much more likely

to be accurately imputed even against a small haplotype panel. Therefore, imputation performs accurately with only 48 individuals added to the 1000 Genomes reference panel for rare variants (MAF ≈ 1%), and 24 individuals added for accurate imputation of common variants (MAF > 5%) depending on chromosome.

With the inclusion of ancestry group-specific haplotypes from these individuals, researchers can appreciate sequencing fewer individuals for genotyping studies on other ancestral populations. Haplotypes found in the 1000 Genomes panel will carry the bulk of the imputation burden and have the rarer alleles genotyped more accurately by referencing the haplotypes from the Samoan data within the combined panel. This results in a lowered study cost by maintaining high statistical power through large sample sizes while also performing WGS on a small number of individuals. Applications for this method include standard GWAS experiments, gene-set enrichment analysis, and expression quantitative trait loci (eQTL) mapping via transcriptome imputation.

There are some limitations to this study. This analysis was originally set to contain information from the TOPMed consortium. However, I was unable to perform imputation with the TOPMed panel included with 1000 Genomes and the Samoan data due to too many samples within that dataset. Therefore, the imputation analysis was conducted on the panel of the Samoan participants with the 1000 Genomes panel. This most likely will impact the imputation accuracy; however, there are still many different haplotypes within this reference panel to obtain meaningful results. The TOPMed panel contains a few individuals of Polynesian ancestry, which would have influenced imputation as this panel is not strictly cosmopolitan. A solution to this would have been removing these identified individuals in the TOPMed reference panel, but this was not necessary as it was not used for imputation. Only chromosomes 5 and 21 were imputed due to time restraints. However, I expect similar results to imputation genome-wide, and that data is available for

genome-wide analysis in the future. Lastly, it would be helpful to have the sequencing data for all imputed data to confirm allele frequencies for rs200884524 and rs373863828. Instead, I have the imputation statistics data from the imputation software to base accuracy on.

This study has broad public health implications despite being a computationally focused analysis. This panel will provide better imputation results for future GWAS studies on Samoan ancestry because it includes a higher density of Samoan-specific alleles allowing more straightforward fine mapping of rare variants. Specifically, a recently discovered variant in *CREBRF* is possibly associated with higher BMI and lower risk of type 2 diabetes. Obesity is a result of both highly polygenic and environmental factors such as food availability and exercise and is highly prevalent in Samoa. However, cases of type 2 diabetes remain lower than expected. rs373863828 within *CREBRF* and rs200884524 within *BTNL9* are hypothesized to be protective of type 2 diabetes despite increased risk from higher average BMI. These variants are found in high frequency within individuals of Samoan ancestry but not within the wider Polynesian group, suggesting genetic drift due to founder and bottleneck effects. Increased genotyping accuracy will help understand disease etiology, risk, prevention, diagnosis and treatment (Molster et al., 2018). Also, by incorporating genomics techniques in a public heath-focused study on a population isolate, we can improve the genomic literacy of the public to further inform, educate, and empower people about health issues (McWalter & Gaviglio, 2015) thereby appropriately integrating genomic technologies into all aspects of healthcare (Bowen et al., 2012).

In future studies, the creation of a study-specific haplotype reference panel for imputing genotypes is a cost-effective strategy for large-scale genome-wide analyses. In this study, I demonstrated that the inclusion of 48 Pacific Islanders within a cosmopolitan imputation panel provides enough haplotype information for adequate imputation accuracy in a Samoan study group.

However, this discovery may not contain a definitive number for all underrepresented populations, as the size of linkage disequilibrium (LD) blocks varies all around the world. This study could serve as an estimate for the number of individuals to sequence when conducting studies on population isolates.

Appendix A Summary of MAFs and r^2 values

Table 5. Summary of imputation accuracy across chromosome 5

Samoans Added	MAF	SNPs	
		$r^2 > 0.3$	$r^2 > 0.8$
0	0%–1%	9.4%	5.3%
	1%–5%	71.7%	37.5%
	5%–50%	99.4%	88.9%
1	0%–1%	10.6%	5.4%
	1%–5%	74.7%	40.5%
	5%–50%	99.6%	90.1%
24	0%–1%	13.0%	6.7%
	1%–5%	95.9%	74.1%
	5%–50%	99.9%	97.8%
48	0%–1%	14.8%	8.3%
	1%–5%	98.5%	84.2%
	5%–50%	99.9%	99.1%
96	0%–1%	17.9%	10.6%
	1%–5%	99.4%	90.7%
	5%–50%	99.9%	99.6%
384	0%–1%	28.3%	19.7%
	1%–5%	99.9%	97.5%
	5%–50%	99.9%	99.9%
1,285	0%–1%	33.5%	24.6%
	1%–5%	99.9%	97.1%
	5%–50%	99.9%	99.9%

Table 6. Summary of imputation accuracy across chromosome 21

Samoans Added	SNPs MAF	$r^2 > 0.3$	$r^2 > 0.8$
0	0%–1%	9.0%	4.0%
	1%–5%	66.4%	31.9%
	5%–50%	98.5%	83.5%
1	0%–1%	9.1%	4.1%
	1%–5%	69.4%	35.7%
	5%–50%	98.8%	84.5%
4	0%–1%	9.6%	4.5%
	1%–5%	76.9%	45.3%
	5%–50%	99.3%	88.4%
6	0%–1%	9.8%	19.7%
	1%–5%	80.9%	50.2%
	5%–50%	99.4%	90.0%
12	0%–1%	10.3%	4.9%
	1%–5%	87.6%	58.9%
	5%–50%	99.7%	92.9%
24	0%–1%	11.1%	5.4%
	1%–5%	92.9%	69.3%
	5%–50%	99.8%	95.8%
48	0%–1%	12.8%	6.6%
	1%–5%	96.8%	80.3%
	5%–50%	99.9%	97.6%
96	0%–1%	15.4%	8.4%
	1%–5%	98.5%	86.8%
	5%–50%	99.9%	98.5%
384	0%–1%	24.2%	15.9%
	1%–5%	99.5%	94.4%
	5%–50%	99.9%	99.2%
1,285	0%–1%	28.9%	21.2%
	1%–5%	98.8%	94.7%
	5%–50%	99.9%	98.9%

Appendix B Minor Allele Frequencies of *CREBRF* and *BTNL9*

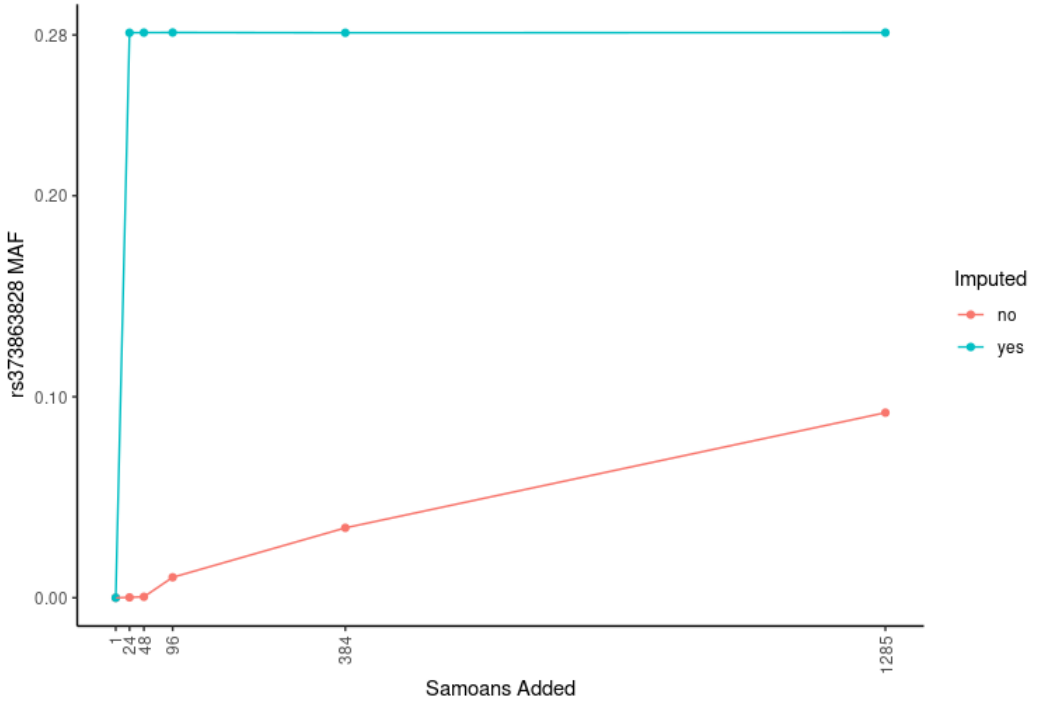

Figure 7. MAF of rs373863828 in *CREBRF* in the 1000 Genomes panel + Samoans.

R^2 imputation accuracies for 1 Samoan (0), 24 Samoans (0.99894), 48 Samoans (0.99905), 96 Samoans (0.99911), 384 Samoans (0.99951), and 1285 Samoans (0.99870).

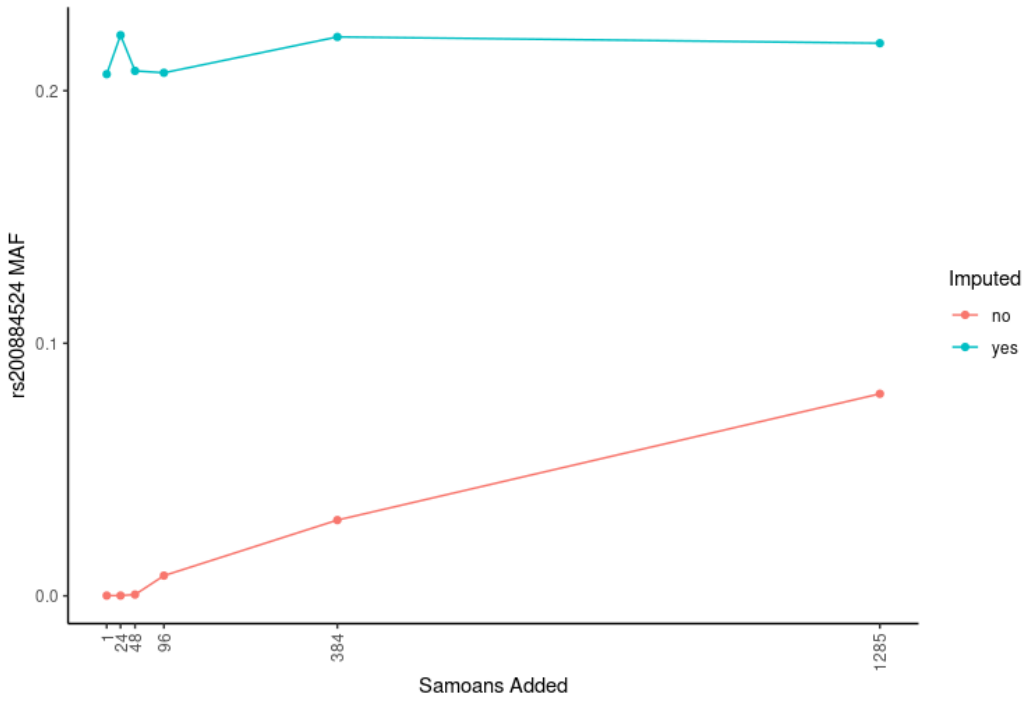

Figure 8. MAF of rs200884524 in *BTNL9* on the 1000 Genomes panel + Samoans.

R^2 imputation accuracies for 1 Samoan (0.92108), 24 Samoans (0.91276), 48 Samoans (0.86967), 96 Samoans (0.91062), 384 Samoans (0.94567), and 1285 Samoans (0.93106).

Appendix C Distribution of Variant Counts

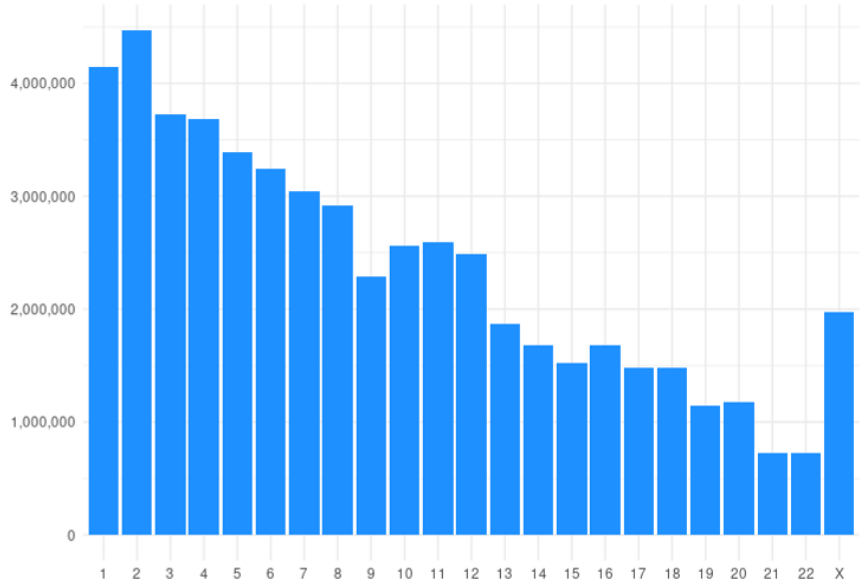

Figure 9. Counts of variants per chromosome in the 1000 Genome + Samoa master reference panel

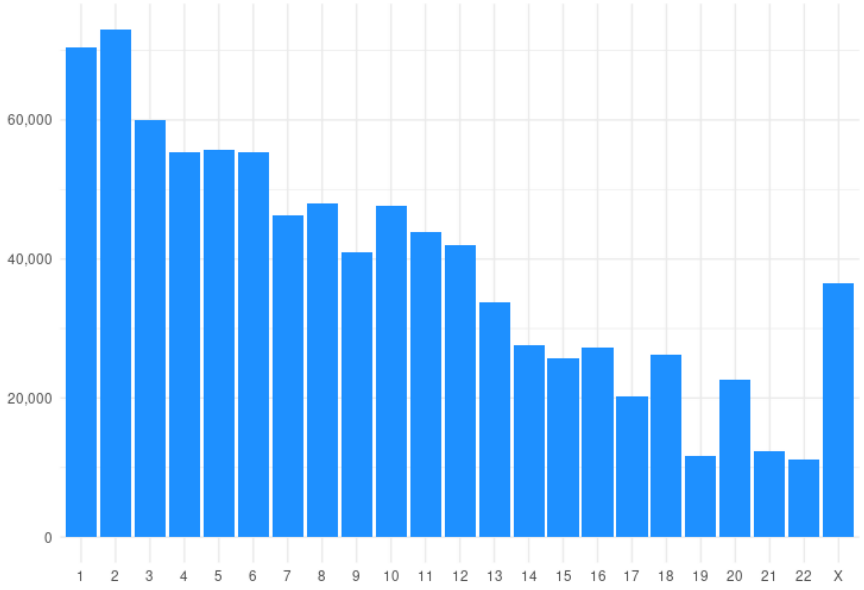

Figure 10. Counts of variants per chromosome on the Affymetrix 6.0 scaffold

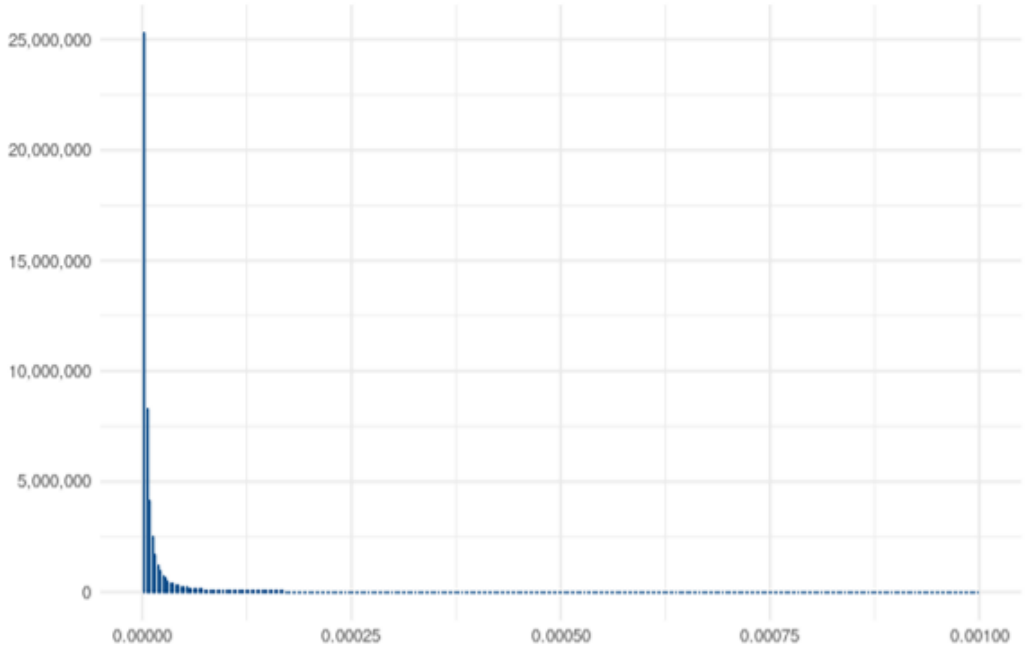

Figure 11. Distribution of MAFs (0%–0.1%) of the 1000 Genomes + Samoan master reference panel on chromosome 5

Appendix D Code

4.1 R Code

4.1.1 App.R

Below is the code used for the UI and server functions of the dashboard.

```
#-------------------------------------------------------------------------------
# Kevin Anderson
# kja34@pitt.edu
#
# HUGEN M.S GB Capstone - Descriptive Statistics and Visualization
#
# Main server file for the descriptive statistics dashboard. Contains UI and server functions.
#-------------------------------------------------------------------------------

#-------------------------------------------------------------------------------
# THIS IS THE LIBRARY SECTION - SHHHHHH!
#-------------------------------------------------------------------------------
library(shiny)
library(ggplot2)
library(tidyverse)
library(DT)
library(shinythemes)
library(psych)
library(shinyalert)
library(scales)

#-------------------------------------------------------------------------------
# Load helper functions
#-------------------------------------------------------------------------------
source("getStats.R")
source("getMAF.R")

#-------------------------------------------------------------------------------
# UI
#-------------------------------------------------------------------------------
ui <- fluidPage(theme = shinytheme("flatly"),
   sidebarLayout(
      sidebarPanel(width = 2,
            conditionalPanel(condition = "input.myTabs == 1",
                  radioButtons("cohortSel", "Choose cohort group",
                        c("1990" = "cohort1990",
                          "2002" = "cohort2002",
```

```
                                "2010" = "cohort2010",
                                "All" = "all")),
                        sliderInput("binSlider", "Set bins for histograms",
                                min = 10, max = 100, value = 50)
                ), # end of conditionalPanel

                conditionalPanel(condition = "input.myTabs == 2",
                        selectInput("chrSel", "Select chromosome",
                                choices = c(1:22, 'X'))
                ), # end of conditionalPanel

                conditionalPanel(condition = "input.myTabs == 3",
                        selectInput("chrSel_Scaffold", "Select chromosome",
                                choices = c(1:22))
                ), # end of conditionalPanel

                conditionalPanel(condition = "input.myTabs == 4",
                        selectInput("chrSel_Imputation", "Select chromosome",
                                choices = c(1:22, 'X'))
                ), # end of conditionalPanel

        ), # end of sidebarPanel
        mainPanel(
          tagList(tags$head(tags$script(type="text/javascript", src = "code.js")),
          navbarPage("Sāmoan Genotype Data", id = "myTabs", position = "static-top",
                tabPanel("Phenotypes", value = 1,
                        fluidRow(h3("Sex and Age Distribution"),
                          column(3, dataTableOutput("sexTable")),
                          column(9, plotOutput("ageHist"),
                                actionButton("ageStats", "Get age statistics",
                                        style="color: #fff; background-color: #104e8b; border-color: #2e6da4"))
                          ), # end of fluidRow

                        fluidRow(h3("BMI Distribution"),
                          column(6, plotOutput("bmiHist"),
                                actionButton("bmiStats", "Get BMI statistics",
                                        style="color: #fff; background-color: #104e8b; border-color: #2e6da4")),
                          column(6, plotOutput("bmiScatter"))
                        ), # end of fluidRow

                        fluidRow(h3("HDL Distribution"),
                          column(6, plotOutput("hdlHist"),
                              actionButton("hdlStats", "Get HDL statistics",
                                      style="color: #fff; background-color: #104e8b; border-color: #2e6da4")),
                          column(6, plotOutput("hdlScat"))
                        ), # end of fluidRow
                ), # end of tabPanel

                # Tried doing per sample tests using `bcftools query` however there are too many samples to generate an
                # output that is able to be read
                # for X in *.bcf; do bcftools query -f '%CHROM %POS %AN %AC{0}\n' $X | awk '{printf "%s %s %f\n",$1,$2,$4/$3}' >
/home/kja34/capstoneStats/$X.txt; done
                tabPanel("Reference Panels", value = 2,
                        fluidRow(h3("Samoa + TOPMed + 1000 Genomes Panels"),
                          column(6, plotOutput("mafHist"),
                                selectInput("limSet", "Select right limit",
                                        choices = c(.00001, .0001, .001, .01, .1, .5),
                                        selected = .001),
```

```
            actionButton("mafStats", "Get variant statistics",
                     style="color: #fff; background-color: #104e8b; border-color: #2e6da4")),
          column(6, plotOutput("mafScat")),
        ), # end of fluidRow

        fluidRow(
            column(6, h3("Samoan Panel"),
                plotOutput("mafHist_samoa"),
                selectInput("limSetSamoan", "Select right limit",
                        choices = c(.00001, .0001, .001, .01, .05, .1, .5),
                        selected = .05),
                actionButton("mafStatsSamoan", "Get variant statistics",
                         style="color: #fff; background-color: #104e8b; border-color: #2e6da4")),
            column(6, h3("1000 Genomes Panel"),
                plotOutput("mafHist_1000g"),
                    selectInput("limSet1000g", "Select right limit",
                            choices = c(.00001, .0001, .001, .01, .05, .1, .5),
                            selected = .05),
                    actionButton("mafStats1000g", "Get variant statistics",
                             style="color: #fff; background-color: #104e8b; border-color: #2e6da4"))
            ), # end of fluidRow

        fluidRow(
          column(6, h3("COMING SOON - Individual Samoan statistics")),
        ), # end of fluidRow
    ), # end of tabPanel

    tabPanel("Scaffold", value = 3,
        fluidRow(h3("Affymetrix Scaffold"),
            column(6, plotOutput("affyHist"),
                actionButton("mafStatsAffy", "Get variant statistics",
                         style="color: #fff; background-color: #104e8b; border-color: #2e6da4"),
                selectInput("limSetAffy", "Select right limit",
                        choices = c(.00001, .0001, .001, .01, .05, .1, .5),
                        selected = .05)),
            column(6, plotOutput("affyScat")),
        ),
    ), # end of tabPanel

    tabPanel("Existing Imputation", value = 4,
        fluidRow(h3("Existing Imputation - Replication"),
            column(6, plotOutput("replicationHist"),
                actionButton("mafStatsReplication", "Get variant statistics",
                         style="color: #fff; background-color: #104e8b; border-color: #2e6da4"),
                selectInput("limSelReplication", "Select right limit",
                        choices = c(.00001, .0001, .001, .01, .05, .1, .5),
                        selected = .05)),
            column(6, plotOutput("imputationScatReplication")),
        ), # end of fluidRow

        fluidRow(h3("Existing Imputation - Discovery"),
            column(6, plotOutput("discoveryHist"),
                actionButton("mafStatsDiscovery", "Get variant statistics",
                         style="color: #fff; background-color: #104e8b; border-color: #2e6da4"),
                selectInput("limSelDiscovery", "Select right limit",
                        choices = c(.00001, .0001, .001, .01, .05, .1, .5),
                        selected = .05)),
            column(6, plotOutput("imputationScatDiscovery")),
```

```
                    ), # end of fluidRow
                ) # end of tabPanel
            ) # end of navbarPage
            ) # end of tagList
        ) # end of mainPanel
    ) # end of sidebarLayout
) # end of fluidPage

#--------------------------------------------------------------------------------
# Server
#--------------------------------------------------------------------------------
server <- function(input, output, session) {

    # Load GWAS data
    load("/home/rminster/MSGB_Imputation_Projects/1990_BMI-phenotype.RData")
    load("/home/rminster/MSGB_Imputation_Projects/2002_BMI-phenotype.RData")
    load("/home/rminster/MSGB_Imputation_Projects/2010_BMI-phenotype.RData")
    load("/home/rminster/MSGB_Imputation_Projects/Samoan_Discovery_Phenotype_v3_2020-01-13.RData")
    replication_phenotypes <- read.table("/home/rminster/MSGB_Imputation_Projects/replication-phenotypes.txt", header = T)

    #--------------------------------------------------------------------------------
    # Phenotype Statistics
    #--------------------------------------------------------------------------------

    # Sex
    # This observeEvent listens for the selection of the radio button cohortSel
    # Sends the appropriate table to the output variable in the UI which displays on app
    observeEvent(input$cohortSel,{

        # reads in sex column from each dataset
        sex1990 <- annotDat_1990@data[["sex"]]
        sex2002 <- annotDat_2002@data[["sex"]]
        sex2020 <- annotDat_2020@data[["sex"]]

        # if and if else statements determining which graphs to display based on cohortSel
        if(input$cohortSel == "cohort1990"){
            sex1990df <- as.data.frame(table(sex1990)) %>%
                add_column(Prop = c(0.568, 0.468)) %>%
                rename(Sex = sex1990) %>%
                mutate(Sex = as.numeric(Sex))

            sex1990df$Sex[sex1990df$Sex == 1] <- "Female"
            sex1990df$Sex[sex1990df$Sex == 2] <- "Male"

            output$sexTable <- renderDataTable({
                datatable(sex1990df,
                          options = list(dom = 't'),
                          rownames = FALSE)
            })
        } else if(input$cohortSel == "cohort2002"){
            sex2002df <- as.data.frame(table(sex2002)) %>%
                add_column(Prop = c(0.545, 0.455)) %>%
                rename(Sex = sex2002) %>%
                mutate(Sex = as.numeric(Sex))

            sex2002df$Sex[sex2002df$Sex == 1] <- "Female"
            sex2002df$Sex[sex2002df$Sex == 2] <- "Male"
```

```
    output$sexTable <- renderDataTable({
      datatable(sex2002df,
                options = list(dom = 't'),
                rownames = FALSE)
    })
  } else if(input$cohortSel == "cohort2010"){
    sex2020df <- as.data.frame(table(sex2020)) %>%
      add_column(Prop = c(0.596, 0.403)) %>%
      rename(Sex = sex2020) %>%
      mutate(Sex = as.numeric(Sex))

    sex2020df$Sex[sex2020df$Sex == 1] <- "Female"
    sex2020df$Sex[sex2020df$Sex == 2] <- "Male"

    output$sexTable <- renderDataTable({
      datatable(sex2020df,
                options = list(dom = 't'),
                rownames = FALSE)
    })
  } else if(input$cohortSel == "all"){
    Sex <- c("Female", "Male")
    Freq <- c(2841, 2099) #total = 4940
    Prop <- c(0.575, 0.425)
    sexAllDf <- data.frame(Sex, Freq, Prop)

    output$sexTable <- renderDataTable({
      datatable(sexAllDf,
                options = list(dom = 't'),
                rownames = FALSE)
    })
  }
})

# Age
# Listens for when the user clicks the button to get statistics for age
# Displays a pop-up box that shows the embedded HTML code
observeEvent(input$ageStats, {
  getAgeStats()
})

# Listens for the radio button cohortSel selection
# These if and if else statements read in the appropriate columsn from each cohort's dataframe
# Manipulates the dataframe for more readable parameters for creating the plots
# Plots are interactable by allowing the user to adjust the number of bins via binSlider input
observeEvent(input$cohortSel,{

  if(input$cohortSel == "cohort1990"){
    age1990 <- data.frame(annotDat_1990@data[["age"]], annotDat_1990@data[["sex"]]) %>%
      rename("Age" = 1, "Sex" = 2)
    age1990$Sex[age1990$Sex == 0] <- "Female"
    age1990$Sex[age1990$Sex == 1] <- "Male"

    output$ageHist <- renderPlot({
      ggplot(age1990, aes(Age, fill = Sex)) +
        geom_histogram(bins = input$binSlider, color = 'black') + # this is where the slider adjusts bins
        scale_fill_manual(values=c("dodgerblue4", "#ff782a")) +
        theme_minimal()
      })
```

```
  } else if(input$cohortSel == "cohort2002"){
    age2002 <- data.frame(annotDat_2002@data[["age"]], annotDat_2002@data[["sex"]]) %>%
      rename("Age" = 1, "Sex" = 2)
    age2002$Sex[age2002$Sex == 0] <- "Female"
    age2002$Sex[age2002$Sex == 1] <- "Male"

    output$ageHist <- renderPlot({
      ggplot(age2002, aes(Age, fill = Sex)) +
        geom_histogram(bins = input$binSlider, color = 'black') +
        theme_minimal() +
        scale_fill_manual(values=c("dodgerblue4", "#ff782a"))
    })
  } else if(input$cohortSel == "cohort2010"){
    age2020 <- data.frame(annotDat_2020@data[["age"]], annotDat_2020@data[["sex"]]) %>%
      rename("Age" = 1, "Sex" = 2)
    age2020$Sex[age2020$Sex == 0] <- "Female"
    age2020$Sex[age2020$Sex == 1] <- "Male"

    output$ageHist <- renderPlot({
      ggplot(age2020, aes(Age, fill = Sex)) +
        geom_histogram(bins = input$binSlider, color = 'black') +
        theme_minimal() +
        scale_fill_manual(values=c("dodgerblue4", "#ff782a"))
    })
  } else if(input$cohortSel == "all"){
    age1990 <- data.frame(annotDat_1990@data[["age"]], annotDat_1990@data[["sex"]]) %>%
      rename("Age" = 1, "Sex" = 2)
    age2002 <- data.frame(annotDat_2002@data[["age"]], annotDat_2002@data[["sex"]]) %>%
      rename("Age" = 1, "Sex" = 2)
    age2020 <- data.frame(annotDat_2020@data[["age"]], annotDat_2020@data[["sex"]]) %>%
      rename("Age" = 1, "Sex" = 2)
    allAge <- rbind(age1990, age2002, age2020) # combines all of the cohorts together

    allAge$Sex[allAge$Sex == 0] <- "Female"
    allAge$Sex[allAge$Sex == 1] <- "Male"

    output$ageHist <- renderPlot({
      ggplot(allAge, aes(Age, fill = Sex)) +
        geom_histogram(bins = input$binSlider, color = 'black') +
        theme_minimal() +
        scale_fill_manual(values=c("dodgerblue4", "#ff782a"))
    })
  }
})

# BMI
# Displays window with statistics on BMI when the user clicks the button
observeEvent(input$bmiStats, {
  getBmiStats()
})

# Listens for the selection of cohorts via radio buttons
# Pulls appropriate columns from each annotated dataframe and manipulates data for plots
# Also creates scatterplot that is dependent on the cohort selection
observeEvent(input$cohortSel,{

  if(input$cohortSel == "cohort1990"){
```

```
        bmi1990 <- data.frame(annotDat_1990@data[["age"]], annotDat_1990@data[["sex"]], annotDat_1990@data[['BMIT1']])
%>%
          rename("Age" = 1, "Sex" = 2, "BMI" = 3)

        bmi1990$Sex[bmi1990$Sex == 0] <- "Female"
        bmi1990$Sex[bmi1990$Sex == 1] <- "Male"

        output$bmiHist <- renderPlot({
          ggplot(bmi1990, aes(BMI, fill = Sex)) +
            geom_histogram(bins = input$binSlider, color = 'black') +
            theme_minimal() +
            scale_fill_manual(values=c("dodgerblue4", "#ff782a"))
        })

        output$bmiScatter <- renderPlot({
          ggplot(bmi1990, aes(Age, BMI, color = Sex)) +
            geom_point() +
            geom_smooth(method=lm) + # adds regression lines
            theme_minimal() +
            scale_color_manual(values=c("dodgerblue4", "#ff782a")) +
            ggtitle("r^2 = 0.173") # calculated from cor.test()
        })
      } else if(input$cohortSel == "cohort2002"){
        bmi2002 <- data.frame(annotDat_2002@data[["age"]], annotDat_2002@data[["sex"]], annotDat_2002@data[['BMIT1']])
%>%
          rename("Age" = 1, "Sex" = 2, "BMI" = 3)

        bmi2002$Sex[bmi2002$Sex == 0] <- "Female"
        bmi2002$Sex[bmi2002$Sex == 1] <- "Male"

        output$bmiHist <- renderPlot({
          ggplot(bmi2002, aes(BMI, fill = Sex)) +
            geom_histogram(bins = input$binSlider, color = 'black') +
            theme_minimal() +
            scale_fill_manual(values=c("dodgerblue4", "#ff782a"))
        })

        output$bmiScatter <- renderPlot({
          ggplot(bmi2002, aes(Age, BMI, color = Sex)) +
            geom_point() +
            geom_smooth(method=lm) +
            theme_minimal() +
            scale_color_manual(values=c("dodgerblue4", "#ff782a")) +
            ggtitle("r^2 = 0.111")

        })
      } else if(input$cohortSel == "cohort2010"){
        bmi2020 <- data.frame(annotDat_2020@data[["age"]], annotDat_2020@data[["sex"]], annotDat_2020@data[['BMI']])
%>%
          rename("Age" = 1, "Sex" = 2, "BMI" = 3)

        bmi2020$Sex[bmi2020$Sex == 0] <- "Female"
        bmi2020$Sex[bmi2020$Sex == 1] <- "Male"

        output$bmiHist <- renderPlot({
          ggplot(bmi2020, aes(BMI, fill = Sex)) +
            geom_histogram(bins = input$binSlider, color = 'black') +
            theme_minimal() +
```

```
            scale_fill_manual(values=c("dodgerblue4", "#ff782a"))
        })

        output$bmiScatter <- renderPlot({
          ggplot(bmi2020, aes(Age, BMI, color = Sex)) +
            geom_point() +
            geom_smooth(method=lm) +
            theme_minimal() +
            scale_color_manual(values=c("dodgerblue4", "#ff782a")) +
            ggtitle("r^2 = 0.101")
        })
      } else if(input$cohortSel == "all"){
        bmi1990 <- data.frame(annotDat_1990@data[["age"]], annotDat_1990@data[["sex"]], annotDat_1990@data[['BMIT1']])
%>%
          rename("Age" = 1, "Sex" = 2, "BMI" = 3)
        bmi2002 <- data.frame(annotDat_2002@data[["age"]], annotDat_2002@data[["sex"]], annotDat_2002@data[['BMIT1']])
%>%
          rename("Age" = 1, "Sex" = 2, "BMI" = 3)
        bmi2020 <- data.frame(annotDat_2020@data[["age"]], annotDat_2020@data[["sex"]], annotDat_2020@data[['BMI']])
%>%
          rename("Age" = 1, "Sex" = 2, "BMI" = 3)
        allBMI <- rbind(bmi1990, bmi2002, bmi2020)

        allBMI$Sex[allBMI$Sex == 0] <- "Female"
        allBMI$Sex[allBMI$Sex == 1] <- "Male"

        output$bmiHist <- renderPlot({
          ggplot(allBMI, aes(BMI, fill = Sex)) +
            geom_histogram(bins = input$binSlider, color = 'black') +
            theme_minimal() +
            scale_fill_manual(values=c("dodgerblue4", "#ff782a"))
        })

        output$bmiScatter <- renderPlot({
          ggplot(allBMI, aes(Age, BMI, color = Sex)) +
            geom_point() +
            geom_smooth(method=lm) +
            theme_minimal() +
            scale_color_manual(values=c("dodgerblue4", "#ff782a")) +
            ggtitle("r^2 = 0.124")
        })
      }
    })

    # HDL
    hdlDf <- data.frame(replication_phenotypes$AGET3, replication_phenotypes$BMIT3, replication_phenotypes$HDLCT3,
replication_phenotypes$SEX) %>%
      rename("Age" = 1, "BMI" = 2, "HDL" = 3, "Sex" = 4)

    hdlDf$Sex[hdlDf$Sex == 1] <- "Female"
    hdlDf$Sex[hdlDf$Sex == 2] <- "Male"

    observeEvent(input$cohortSel,{

      output$hdlHist <- renderPlot({
        ggplot(hdlDf, aes(HDL, fill = Sex)) +
          geom_histogram(bins = input$binSlider, color = 'black') +
          theme_minimal() +
```

```
        scale_fill_manual(values=c("dodgerblue4", "#ff782a")) +
        scale_x_continuous(name = "HDL (mg/dL)")
    })

    cor.test(hdlDf$HDL, hdlDf$Age)

    output$hdlScat <- renderPlot({

      ggplot(hdlDf, aes(Age, HDL, color = Sex)) +
        geom_point() +
        geom_smooth(method = lm) +
        theme_minimal() +
        scale_color_manual(values=c("dodgerblue4", "#ff782a")) +
        scale_y_continuous(name = "HDL (mg/dL)") +
        ggtitle("r^2 = -0.248")
    })
})

observeEvent(input$hdlStats, {
    #describeBy(hdlDf$HDL, hdlDf$Sex)
    getHdlStats()
})

#-------------------------------------------------------------------------------------
# Reference panel statistics
#-------------------------------------------------------------------------------------

# MAF (minor allele frequency)
# These functions output graphs and statistics from the reference panel bcfs
# Outputted graphs are located on the second mainPanel - users have to click on the correct nav page

# This observe listens for when the user clicks the button to get statistics for MAF and displays a pop-up
observeEvent(input$mafStats, {
  getMafStats()
})

# Samoans and 1000g doesn't exclude variants with 0 MAF !!!!
observeEvent(input$mafStatsSamoan, {
    getMafStatsSamoan()
})
observeEvent(input$mafStats1000g, {
    getMafStats1000g()
})

# This listens for when the user changes which chromosome they are looking at from the selectizeInput chrSel
# Outputs bar graphs for the distribution of MAFs depending on the chromosome selected
# Also handles user input for adjusting the x-axis limit for each graph
#
### For Samoa + TOPMed + 1000Genomes Panels ###

observeEvent(input$chrSel,{
    output$mafHist <- renderPlot({
      getMAFAll(input$chrSel, as.numeric(input$limSet))
    })
})

### For Samoan panel only ###
observeEvent(input$chrSel,{
```

```
  output$mafHist_samoa <- renderPlot({
    getMAFSamoan(input$chrSel, as.numeric(input$limSetSamoan))
  })
})

### For 1000Genomes panel only ###
observeEvent(input$chrSel,{
    output$mafHist_1000g <- renderPlot({
      getMAF1000g(input$chrSel, as.numeric(input$limSet1000g))
    })
  })

### Samoan vs. 1000g ###
observeEvent(input$chrSel,{
  output$mafScat <- renderPlot({
    getMAFScat(input$chrSel)
  })
})

#-------------------------------------------------------------------------------
# Scaffolds
#-------------------------------------------------------------------------------

### Affymetrix ###
observeEvent(input$mafStatsAffy, {
  getAffyStats()
})

observeEvent(input$chrSel_Scaffold,{
  output$affyHist <- renderPlot({
    getAffyHist(input$chrSel_Scaffold, as.numeric(input$limSelAffy))
  })
})

observeEvent(input$chrSel_Scaffold,{
  output$affyScat <- renderPlot({
    getAffyScat(input$chrSel_Scaffold)
  })
})

#-------------------------------------------------------------------------------
# Existing Imputation Sets
#-------------------------------------------------------------------------------

observeEvent(input$mafStatsReplication, {
  getExistingReplicationStats()
})

observeEvent(input$mafStatsDiscovery, {
  getExistingDiscoveryStats()
})


observeEvent(input$chrSel_Imputation,{
  output$discoveryHist <- renderPlot({
    getMAFDiscovery(input$chrSel_Imputation, as.numeric(input$limSelDiscovery))
  })
})
```

```
  observeEvent(input$chrSel_Imputation,{
    output$replicationHist <- renderPlot({
      getMAFReplication(input$chrSel_Imputation, as.numeric(input$limSelReplication))
    })
  })

  observeEvent(input$chrSel_Imputation,{
    output$imputationScatDiscovery <- renderPlot({
      getImputationScat_Discovery(input$chrSel_Imputation)
    })
  })

  observeEvent(input$chrSel_Imputation,{
    output$imputationScatReplication <- renderPlot({
      getImputationScat_Replication(input$chrSel_Imputation)
    })
  })

} # end of Server

# Run the application
shinyApp(ui = ui, server = server)
```

4.1.2 RShiny Functions

Below is the code used mostly for reading in data requested by the user and generating visualizations for output.

```
#-------------------------------------------------------------------------------
# Kevin Anderson
# kja34@pitt.edu
#
# HUGEN M.S GB Capstone - Descriptive Statistics and Visualization
#
#-------------------------------------------------------------------------------

getMAFAll <- function(chr, lim) {
 df <- assign(paste0("chr", chr, "_hist"), read.table(paste0("~/capstoneStats/data/reference_panels/chr", chr, "_hist.txt")) %>%
          mutate(MAF = if_else(V2 > 0.5, 1-V2, V2)) %>%
          filter(V2 > 0) %>%
          filter(V2 < 1))

 ggplot(df, aes(MAF, V1)) +
  geom_col(color = 'dodgerblue4', fill = 'dodgerblue4') +
  scale_x_continuous(name = 'MAF', limits = c(0, lim)) + # user can adjust right limit
  scale_y_continuous(name = "Count", labels = scales::comma) + # changes axis labels from scientific notation to comma
  theme_minimal() +
  ggtitle(label = paste0("Chromosome ", chr))
}
```

```
getMAFSamoan <- function(chr, lim) {
  df <- assign(paste0("freeze.9b.chr", chr, ".phased_samoan"),
read.table(paste0("~/capstoneStats/data/reference_panels/freeze9b_chr", chr, "_samoans.txt_count.txt")) %>%
              mutate(MAF = if_else(V2 > 0.5, 1-V2, V2)) %>%
              filter(V2 > 0) %>%
              filter(V2 < 1))

  ggplot(df, aes(MAF, V1)) +
    geom_col(color = 'dodgerblue4', fill = 'dodgerblue4') +
    scale_x_continuous(name = 'MAF', limits = c(0, lim)) + # user can adjust right limit
    scale_y_continuous(name = "Count", labels = scales::comma) + # changes axis labels from scientific notation to comma
    theme_minimal() +
     ggtitle(label = paste0("Chromosome ", chr))
}

getMAF1000g <- function(chr, lim) {
  df <- assign(paste0("freeze.9b.chr", chr, ".phased_1000g"),
read.table(paste0("~/capstoneStats/data/reference_panels/freeze9b_chr", chr, "_1000g.txt_count.txt")) %>%
    mutate(MAF = if_else(V2 > 0.5, 1-V2, V2)) %>%
    filter(V2 > 0) %>%
    filter(V2 < 1))

  ggplot(df, aes(MAF, V1)) +
    geom_col(color = 'dodgerblue4', fill = 'dodgerblue4') +
    scale_x_continuous(name = 'MAF', limits = c(0, lim)) + # user can adjust right limit
    scale_y_continuous(name = "Count", labels = scales::comma) + # changes axis labels from scientific notation to comma
    theme_minimal() +
    ggtitle(label = paste0("Chromosome ", chr))
}

getMAFScat <- function(chr) {
  df_samoa <- assign(paste0("freeze.9b.chr", chr, ".phased_samoan"),
read.table(paste0("~/capstoneStats/data/reference_panels/freeze9b_chr", chr, "_samoans.txt_filtered.txt")) %>%
                mutate(MAF = if_else(V3 > 0.5, 1-V3, V3)) %>%
                filter(V3 > 0) %>%
                filter(V3 < 1))
  df_1000g <- assign(paste0("freeze.9b.chr", chr, ".phased_1000g"),
read.table(paste0("~/capstoneStats/data/reference_panels/freeze9b_chr", chr, "_1000g.txt_filtered.txt")) %>%
                mutate(MAF = if_else(V3 > 0.5, 1-V3, V3)) %>%
                filter(V3 > 0) %>%
                filter(V3 < 1))

  # doing a full join on the base pair position to match the MAFs between the two panels
  # there's going to be a lot of NAs
  df <- df_1000g %>%
    full_join(df_samoa, by = "V2")
  df2 <- df[sample(nrow(df), 100000), ] # taking only a small random sample of the data so it loads faster

  ggplot(df2, aes(MAF.x, MAF.y)) +
    geom_point(alpha = 0.3, col = 'dodgerblue4') +
    geom_smooth(se = FALSE, color = "red") +
    theme_minimal() +
    scale_x_continuous(name = "1000Genomes Panel MAF") +
    scale_y_continuous(name = "Samoan Panel MAF") +
    theme(axis.text=element_text(size=12)) +
    ggtitle(label = paste0("Chromosome ", chr))
}
```

```
getAffyHist <- function(chr, lim) {
  df <- assign(paste0("affy_chr", chr), read.table(paste0("~/capstoneStats/data/scaffold/affy_chr", chr, ".txt")) %>%
              mutate(MAF = if_else(V3 > 0.5, 1-V3, V3)) %>%
              filter(V3 < 1))

  ggplot(df, aes(MAF, V1)) +
    geom_col(color = 'dodgerblue4', fill = 'dodgerblue4') +
    scale_x_continuous(name = 'MAF', limits = c(0, lim)) + # user can adjust right limit
    scale_y_continuous(name = "Count", labels = scales::comma) + # changes axis labels from scientific notation to comma
    theme_minimal() +
    ggtitle(label = paste0("Chromosome ", chr))
}

getAffyScat <- function(chr) {
  df_samoa <- assign(paste0("freeze.9b.chr", chr, ".phased_samoan"),
read.table(paste0("~/capstoneStats/data/reference_panels/freeze9b_chr", chr, "_samoans.txt_filtered.txt")) %>%
                  mutate(MAF = if_else(V3 > 0.5, 1-V3, V3)) %>%
                  filter(V3 > 0) %>%
                  filter(V3 < 1))
  df_affy <- assign(paste0("affy_chr", chr), read.table(paste0("~/capstoneStats/data/scaffold/affy_chr", chr, ".txt")) %>%
                  mutate(MAF = if_else(V3 > 0.5, 1-V3, V3)) %>%
                  filter(V3 > 0) %>%
                  filter(V3 < 1))

  df <- df_samoa %>%
    full_join(df_affy, by = "V2")
  df2 <- df[sample(nrow(df), 100000), ] # taking only a small random sample of the data so it loads faster

  ggplot(df2, aes(MAF.x, MAF.y)) +
    geom_point(alpha = 0.3, color = 'dodgerblue4') +
    theme_minimal() +
    geom_smooth(se = FALSE, color = "red") +
    scale_x_continuous(name = "Affymetrix Scaffold MAF") +
    scale_y_continuous(name = "Samoan Panel MAF") +
    theme(text = element_text(size = 12)) +
    ggtitle(label = paste0("Chromosome ", chr))

}

getMAFDiscovery <- function(chr, lim) {
  df <- assign(paste0("freeze9b", chr, "_existing_discovery"),
read.table(paste0("~/capstoneStats/data/existing_imputation/freeze9b_chr", chr,
"_existing_imputation_discovery.txt_counts.txt")) %>%
              mutate(MAF = if_else(V2 > 0.5, 1-V2, V2)) %>%
              filter(V2 > 0) %>%
              filter(V2 < 1))

  ggplot(df, aes(MAF, V1)) +
    geom_col(color = 'dodgerblue4', fill = 'dodgerblue4') +
    scale_x_continuous(name = 'MAF', limits = c(0, lim)) + # user can adjust right limit
    scale_y_continuous(name = "Count", labels = scales::comma) + # changes axis labels from scientific notation to comma
    theme_minimal() +
    ggtitle(label = paste0("Chromosome ", chr))

}

getMAFReplication <- function(chr, lim) {
```

```
  df <- assign(paste0("freeze9b", chr, "_existing_replication"),
read.table(paste0("~/capstoneStats/data/existing_imputation/freeze9b_chr", chr,
"_existing_imputation_replication.txt_counts.txt")) %>%
              mutate(MAF = if_else(V2 > 0.5, 1-V2, V2)) %>%
              filter(V2 > 0) %>%
              filter(V2 < 1))


  ggplot(df, aes(MAF, V1)) +
    geom_col(color = 'dodgerblue4', fill = 'dodgerblue4') +
    scale_x_continuous(name = 'MAF', limits = c(0, lim)) + # user can adjust right limit
    scale_y_continuous(name = "Count", labels = scales::comma) + # changes axis labels from scientific notation to comma
    theme_minimal() +
    ggtitle(label = paste0("Chromosome ", chr))

}

getImputationScat_Replication <- function(chr) {
  df <- assign(paste0("freeze9b", chr, "_existing_replication_r2"),
read.table(paste0("~/capstoneStats/data/existing_imputation/freeze9b_chr", chr, "_existing_imputation_replication.txt")) %>%
              mutate(MAF = if_else(V3 > 0.5, 1-V3, V3)) %>%
              filter(V3 > 0) %>%
              filter(V3 < 1))

  df_samoa <- assign(paste0("freeze.9b.chr", chr, ".phased_samoan_r2"),
read.table(paste0("~/capstoneStats/data/reference_panels/freeze9b_chr", chr, "_samoans.txt_filtered.txt")) %>%
                mutate(MAF = if_else(V3 > 0.5, 1-V3, V3)) %>%
                filter(V3 > 0) %>%
                filter(V3 < 1))

  df2 <- df %>%
    full_join(df_samoa, by = "V2") %>%
    filter(V4 > 0) %>% # samoan panel doesn't have r^2 so they just appear as NA - removing them here
    mutate(R2 = as.factor(if_else(V4 < .8, "r2 < 0.8", "r2 > 0.8")))

  df2 <- df2[sample(nrow(df2), 100000), ] # taking only a small random sample of the data so it loads faster

# data:  df2$MAF.x and df2$MAF.y
# t = 4040.3, df = 59158, p-value < 2.2e-16
# alternative hypothesis: true correlation is not equal to 0
# 95 percent confidence interval:
#  0.9981636 0.9982218
# sample estimates:
#      cor
# 0.9981929

  ggplot(df2, aes(MAF.x, MAF.y)) +
    geom_point(pch = 21, alpha = 0.3, color = "dodgerblue4") +
    theme_minimal() +
    theme(text = element_text(size = 12)) +
    scale_x_continuous(name = "Existing Imputation - Replication MAF") +
    scale_y_continuous(name = "Samoan Panel MAF") +
    facet_grid(cols = vars(R2)) +
    ggtitle(label = paste0("Chromosome ", chr))

}

getImputationScat_Discovery <- function(chr) {
```

```
 df <- assign(paste0("freeze9b", chr, "_existing_discovery_r2"),
read.table(paste0("~/capstoneStats/data/existing_imputation/freeze9b_chr", chr, "_existing_imputation_discovery.txt")) %>%
            mutate(MAF = if_else(V3 > 0.5, 1-V3, V3)) %>%
            filter(V3 > 0) %>%
            filter(V3 < 1))

 df_samoa <- assign(paste0("freeze.9b.chr", chr, ".phased_samoan_r2"),
read.table(paste0("~/capstoneStats/data/reference_panels/freeze9b_chr", chr, "_samoans.txt_filtered.txt")) %>%
                mutate(MAF = if_else(V3 > 0.5, 1-V3, V3)) %>%
                filter(V3 > 0) %>%
                filter(V3 < 1))

 df2 <- df %>%
  full_join(df_samoa, by = "V2") %>%
  filter(V4 > 0) %>% # samoan panel doesn't have r^2 so they just appear as NA - removing them here
  mutate(R2 = as.factor(if_else(V4 < .8, "r2 < 0.8", "r2 > 0.8")))

 df2 <- df2[sample(nrow(df2), 100000), ] # taking only a small random sample of the data so it loads faster

 ggplot(df2, aes(MAF.x, MAF.y)) +
  geom_point(pch = 21, alpha = 0.3, color = "dodgerblue4") +
  theme_minimal() +
  theme(text = element_text(size = 14)) +
  scale_x_continuous(name = "Existing Imputation - Discovery Set MAF") +
  scale_y_continuous(name = "Samoan Panel MAF") +
  facet_grid(cols = vars(R2)) +
  ggtitle(label = paste0("Chromosome ", chr))

}
#-------------------------------------------------------------------------------
# Kevin Anderson
# kja34@pitt.edu
#
# HUGEN M.S GB Capstone - Descriptive Statistics and Visualization
#
# This script contains the functions that load all of the pop-up boxes whenever the user
# clicks one of the buttons to get the statistics for the respective graphs. I made this
# script because the main app.R file was getting too cluttered with all of the long HTML
# code contained below
#-------------------------------------------------------------------------------

getAgeStats <- function() {

 showModal(modalDialog(
  title = "Point Statistics for Age",
  HTML("<h4>1990 Cohort</h4>
            <p><strong>Female</strong></p>
            <ul>
              <li>n = 454</li>
              <li>Mean = 38.75 years</li>
              <li>Standard deviation = 9.34 years</li>
              <li>Range = 51 years</li>
            </ul>
            <p><strong>Male</strong></p>
            <ul>
              <li>n = 400</li>
              <li>Mean = 39.18 years</li>
              <li>Standard deviation = 9.35 years</li>
```

```
            <li>Range = 59 years</li>
           </ul>
           <br>
           <h4>2002 Cohort</h4>
           <p><strong>Female</strong></p>
           <ul>
            <li>n = 542</li>
            <li>Mean = 43.28 years</li>
            <li>Standard deviation = 16.50 years</li>
            <li>Range = 65.42 years</li>
           </ul>
           <p><strong>Male</strong></p>
           <ul>
            <li>n = 452</li>
            <li>Mean = 41.68 years</li>
            <li>Standard deviation = 16.44 years</li>
            <li>Range = 71.07 years</li>
           </ul>
           <br>
           <h4>2010 Cohort</h4>
           <p><strong>Female</strong></p>
           <ul>
            <li>n = 1845</li>
            <li>Mean = 44.69 years</li>
            <li>Standard deviation = 11.10 years</li>
            <li>Range = 42.33 years</li>
           </ul>
           <p><strong>Male</strong></p>
           <ul>
            <li>n = 1247</li>
            <li>Mean = 45.34 years</li>
            <li>Standard deviation = 11.41 years</li>
            <li>Range = 41.69 years</li>
           </ul>
           <br>
           <h4>All Cohorts</h4>
           <p><strong>Female</strong></p>
           <ul>
            <li>n = 2841</li>
            <li>Mean = 43.48 years</li>
            <li>Standard deviation = 12.26 years</li>
            <li>Range = 65.42 years</li>
           </ul>
           <p><strong>Male</strong></p>
           <ul>
            <li>n = 2099</li>
            <li>Mean = 43.38 years</li>
            <li>Standard deviation = 12.58 years</li>
            <li>Range = 71.07 years</li>
           </ul>"),
   easyClose = TRUE,
   footer = NULL
  ))
}

getBmiStats <- function() {

  showModal(modalDialog(
```

```
title = "Point Statistics for BMI",
HTML("<h4>1990 Cohort</h4>
          <p><strong>Female</strong></p>
          <ul>
            <li>n = 454</li>
            <li>Mean = 32.99</li>
            <li>Standard deviation = 6.45</li>
            <li>Range = 51.11</li>
          </ul>
          <p><strong>Male</strong></p>
          <ul>
            <li>n = 400</li>
            <li>Mean = 30.84</li>
            <li>Standard deviation = 5.67</li>
            <li>Range = 32.46</li>
          </ul>
          <br>
          <h4>2002 Cohort</h4>
          <p><strong>Female</strong></p>
          <ul>
            <li>n = 542</li>
            <li>Mean = 35.13</li>
            <li>Standard deviation = 8.14</li>
            <li>Range = 50.79</li>
           </ul>
           <p><strong>Male</strong></p>
           <ul>
            <li>n = 452</li>
            <li>Mean = 31.06</li>
            <li>Standard deviation = 6.80</li>
            <li>Range = 57.07</li>
           </ul>
           <br>
          <h4>2010 Cohort</h4>
          <p><strong>Female</strong></p>
          <ul>
            <li>n = 1845</li>
            <li>Mean = 34.90</li>
            <li>Standard deviation = 6.82</li>
            <li>Range = 50.30</li>
           </ul>
           <p><strong>Male</strong></p>
           <ul>
            <li>n = 1247</li>
            <li>Mean = 31.31</li>
            <li>Standard deviation = 5.94</li>
            <li>Range = 43.60</li>
           </ul>
           <br>
          <h4>All Cohorts</h4>
          <p><strong>Female</strong></p>
          <ul>
            <li>n = 2841</li>
            <li>Mean = 34.64</li>
            <li>Standard deviation = 7.07</li>
            <li>Range = 52.28</li>
           </ul>
           <p><strong>Male</strong></p>
```

```
            <ul>
              <li>n = 2099</li>
              <li>Mean = 31.16</li>
              <li>Standard deviation = 6.09</li>
              <li>Range = 57.80</li>
            </ul>"),
   easyClose = TRUE,
   footer = NULL
  ))
}

getHdlStats <- function() {

  showModal(modalDialog(
    title = "Point Statistics for HDL",
    HTML("<h4>All Samples</h4>
          <p><strong>Female</strong></p>
          <p>n = 1,039</p>
          <p>Mean = 42.43 mg/dL</p>
          <p>Standard deviation = 11.59</p>
          <p>Range = 80.14</p>
          <p><strong>Male</strong></p>
          <p>n = 1211</p>
          <p>Mean = 44.14 mg/dL</p>
          <p>Standard deviation = 10.88</p>
          <p>Range = 70.63</p>"),
    easyClose = TRUE,
    footer = NULL)
  )
}

getMafStats <- function() {

  showModal(modalDialog(
    title = "Point Statistics for MAF",
    HTML("
          <h4> Chromosome 1</h4>
          <p> Variant count: 66,411,703 </p>
          <br>
          <h4>Chromosome 2</h4>
          <p> Variant count: 72,336,188</p>
          <br>
          <h4>Chromosome 3</h4>
          <p> Variant count: 59,435,128 </p>
          <br>
          <h4>Chromosome 4</h4>
          <p> Variant count: 57,576,181 </p>
          <br>
          <h4>Chromosome 5</h4>
          <p> Variant count: 53,775,719 </p>
          <br>
          <h4>Chromosome 6</h4>
          <p> Variant count: 50,096,509 </p>
          <br>
          <h4>Chromosome 7</h4>
          <p> Variant count: 47,714,169 </p>
          <br>
          <h4>Chromosome 8</h4>
```

```
        <p> Variant count: 45,946,976 </p>
        <br>
        <h4>Chromosome 9</h4>
        <p> Variant count: 37,127,710 </p>
        <br>
        <h4>Chromosome 10</h4>
        <p> Variant count: 39,760,241 </p>
        <br>
        <h4>Chromosome 11</h4>
        <p> Variant count: 40,700,337</p>
        <br>
        <h4>Chromosome 12</h4>
        <p> Variant count: 39,243,751 </p>
        <br>
        <h4>Chromosome 13</h4>
        <p> Variant count: 29,242,598 </p>
        <br>
        <h4> Chromosome 14</h4>
        <p> Variant count: 26,341,588 </p>
        <br>
        <h4>Chromosome 15</h4>
        <p> Variant count: 24,305,349 </p>
        <br>
        <h4>Chromosome 16</h4>
        <p> Variant count: 27,058,669 </p>
        <br>
        <h4>Chromosome 17</h4>
        <p> Variant count: 23,324,911 </p>
        <br>
        <h4>Chromosome 18</h4>
        <p> Variant count: 22,889,891 </p>
        <br>
        <h4>Chromosome 19</h4>
        <p> Variant count: 17,608,125 </p>
        <br>
        <h4>Chromosome 20</h4>
        <p> Variant count: 18,290,624 </p>
        <br>
        <h4>Chromosome 21</h4>
        <p> Variant count: 10,870,873 </p>
        <br>
        <h4>Chromosome 22</h4>
        <p> Variant count: 11,337,104 </p>
        <br>
        <h4>X Chromosome</h4>
        <p> Variant count: 30,531,994 </p>
        <br> "),
   easyClose = TRUE,
   footer = NULL))
}

getMafStatsSamoan <- function() {

  showModal(modalDialog(
   title = "Point Statistics for MAF",
   HTML("
       <h4> Chromosome 1</h4>
        <p> Variant count: 66,411,703 </p>
```


<h4>Chromosome 2</h4>
 Variant count: 72,336,188

<h4>Chromosome 3</h4>
 Variant count: 59,435,128

<h4>Chromosome 4</h4>
 Variant count: 57,576,181

<h4>Chromosome 5</h4>
 Variant count: 53,775,719

<h4>Chromosome 6</h4>
 Variant count: 50,096,509

<h4>Chromosome 7</h4>
 Variant count: 47,714,169

<h4>Chromosome 8</h4>
 Variant count: 45,946,976

<h4>Chromosome 9</h4>
 Variant count: 37,127,710

<h4>Chromosome 10</h4>
 Variant count: 39,760,241

<h4>Chromosome 11</h4>
 Variant count: 40,700,337

<h4>Chromosome 12</h4>
 Variant count: 39,243,751

<h4>Chromosome 13</h4>
 Variant count: 29,242,598

<h4> Chromosome 14</h4>
 Variant count: 26,341,588

<h4>Chromosome 15</h4>
 Variant count: 24,305,349

<h4>Chromosome 16</h4>
 Variant count: 27,058,669

<h4>Chromosome 17</h4>
 Variant count: 23,324,911

<h4>Chromosome 18</h4>
 Variant count: 22,889,891

<h4>Chromosome 19</h4>
 Variant count: 17,608,125

<h4>Chromosome 20</h4>
 Variant count: 18,290,624


```
        <h4>Chromosome 21</h4>
        <p> Variant count: 10,870,873 </p>
        <br>
        <h4>Chromosome 22</h4>
        <p> Variant count: 11,337,104 </p>
        <br>
        <h4>X Chromosome</h4>
        <p> Variant count: 30,531,994 </p>
        <br>
      "),
  easyClose = TRUE,
  footer = NULL))
}


getMafStats1000g <- function() {

 showModal(modalDialog(
  title = "Point Statistics for MAF",
  HTML("
      <h4> Chromosome 1</h4>
        <p> Variant count: 66,411,703 </p>
        <br>
        <h4>Chromosome 2</h4>
        <p> Variant count: 72,336,188</p>
        <br>
        <h4>Chromosome 3</h4>
        <p> Variant count: 59,435,128 </p>
        <br>
        <h4>Chromosome 4</h4>
        <p> Variant count: 57,576,181 </p>
        <br>
        <h4>Chromosome 5</h4>
        <p> Variant count: 53,775,719 </p>
        <br>
        <h4>Chromosome 6</h4>
        <p> Variant count: 50,096,509 </p>
        <br>
        <h4>Chromosome 7</h4>
        <p> Variant count: 47,714,169 </p>
        <br>
        <h4>Chromosome 8</h4>
        <p> Variant count: 45,946,976 </p>
        <br>
        <h4>Chromosome 9</h4>
        <p> Variant count: 37,127,710 </p>
        <br>
        <h4>Chromosome 10</h4>
        <p> Variant count: 39,760,241 </p>
        <br>
        <h4>Chromosome 11</h4>
        <p> Variant count: 40,700,337</p>
        <br>
        <h4>Chromosome 12</h4>
        <p> Variant count: 39,243,751 </p>
        <br>
        <h4>Chromosome 13</h4>
        <p> Variant count: 29,242,598 </p>
```

```
        <br>
        <h4> Chromosome 14</h4>
        <p> Variant count: 26,341,588 </p>
        <br>
        <h4>Chromosome 15</h4>
        <p> Variant count: 24,305,349 </p>
        <br>
        <h4>Chromosome 16</h4>
        <p> Variant count: 27,058,669 </p>
        <br>
        <h4>Chromosome 17</h4>
        <p> Variant count: 23,324,911 </p>
        <br>
        <h4>Chromosome 18</h4>
        <p> Variant count: 22,889,891 </p>
        <br>
        <h4>Chromosome 19</h4>
        <p> Variant count: 17,608,125 </p>
        <br>
        <h4>Chromosome 20</h4>
        <p> Variant count: 18,290,624 </p>
        <br>
        <h4>Chromosome 21</h4>
        <p> Variant count: 10,870,873 </p>
        <br>
        <h4>Chromosome 22</h4>
        <p> Variant count: 11,337,104 </p>
        <br>
        <h4>X Chromosome</h4>
        <p> Variant count: 30,531,994 </p>
        <br>
      "),
  easyClose = TRUE,
  footer = NULL))
}

getAffyStats <- function() {

 showModal(modalDialog(
  title = "Point Statistics for Affymetrix Scaffold",
  HTML("
        <h4> Chromosome 1</h4>
        <p> Variant count: 70,380 </p>
        <br>
        <h4>Chromosome 2</h4>
        <p> Variant count: 73,062</p>
        <br>
        <h4>Chromosome 3</h4>
        <p> Variant count: 59,983 </p>
        <br>
        <h4>Chromosome 4</h4>
        <p> Variant count: 55,334 </p>
        <br>
        <h4>Chromosome 5</h4>
        <p> Variant count: 55,764 </p>
        <br>
        <h4>Chromosome 6</h4>
        <p> Variant count: 55,420 </p>
```

```
        <br>
        <h4>Chromosome 7</h4>
        <p> Variant count: 46,330 </p>
        <br>
        <h4>Chromosome 8</h4>
        <p> Variant count: 48,032 </p>
        <br>
        <h4>Chromosome 9</h4>
        <p> Variant count: 40,961 </p>
        <br>
        <h4>Chromosome 10</h4>
        <p> Variant count: 47,565 </p>
        <br>
        <h4>Chromosome 11</h4>
        <p> Variant count: 43,926</p>
        <br>
        <h4>Chromosome 12</h4>
        <p> Variant count: 42,010 </p>
        <br>
        <h4>Chromosome 13</h4>
        <p> Variant count: 33,739 </p>
        <br>
        <h4> Chromosome 14</h4>
        <p> Variant count: 27,656 </p>
        <br>
        <h4>Chromosome 15</h4>
        <p> Variant count: 25,774 </p>
        <br>
        <h4>Chromosome 16</h4>
        <p> Variant count: 27,218 </p>
        <br>
        <h4>Chromosome 17</h4>
        <p> Variant count: 20,308 </p>
        <br>
        <h4>Chromosome 18</h4>
        <p> Variant count: 26,211 </p>
        <br>
        <h4>Chromosome 19</h4>
        <p> Variant count: 11,693 </p>
        <br>
        <h4>Chromosome 20</h4>
        <p> Variant count: 22,554</p>
        <br>
        <h4>Chromosome 21</h4>
        <p> Variant count: 12,387 </p>
        <br>
        <h4>Chromosome 22</h4>
        <p> Variant count: 11,236 </p>
        <br>
        <h4>X Chromosome</h4>
        <p> Variant count: 36,465 </p>
        <br>

    "),
    easyClose = TRUE,
    footer = NULL))
}
```

```
getExistingReplicationStats <- function() {
  showModal(modalDialog(
    title = "Point Statistics for Existing Imputation - Replication Set",
    HTML("
          <h4> Chromosome 1</h4>
          <p> Variant count: 4,148,826 </p>
          <br>
          <h4>Chromosome 2</h4>
          <p> Variant count: 4,473,212</p>
          <br>
          <h4>Chromosome 3</h4>
          <p> Variant count: 3,723,954 </p>
          <br>
          <h4>Chromosome 4</h4>
          <p> Variant count: 3,682,018 </p>
          <br>
          <h4>Chromosome 5</h4>
          <p> Variant count: 3,394,288 </p>
          <br>
          <h4>Chromosome 6</h4>
          <p> Variant count: 3,245,542 </p>
          <br>
          <h4>Chromosome 7</h4>
          <p> Variant count: 3,037,910 </p>
          <br>
          <h4>Chromosome 8</h4>
          <p> Variant count: 2,911,943 </p>
          <br>
          <h4>Chromosome 9</h4>
          <p> Variant count: 2,293,025 </p>
          <br>
          <h4>Chromosome 10</h4>
          <p> Variant count: 2,557,466 </p>
          <br>
          <h4>Chromosome 11</h4>
          <p> Variant count: 2,591,727</p>
          <br>
          <h4>Chromosome 12</h4>
          <p> Variant count: 2,484,806 </p>
          <br>
          <h4>Chromosome 13</h4>
          <p> Variant count: 1,866,294 </p>
          <br>
          <h4> Chromosome 14</h4>
          <p> Variant count: 1,675,545 </p>
          <br>
          <h4>Chromosome 15</h4>
          <p> Variant count: 1,523,290 </p>
          <br>
          <h4>Chromosome 16</h4>
          <p> Variant count: 1,681,823 </p>
          <br>
          <h4>Chromosome 17</h4>
          <p> Variant count: 1,480,477 </p>
          <br>
          <h4>Chromosome 18</h4>
          <p> Variant count: 1,480,195</p>
          <br>
```

```
        <h4>Chromosome 19</h4>
        <p> Variant count: 1,147,567 </p>
        <br>
        <h4>Chromosome 20</h4>
        <p> Variant count: 1,176,819</p>
        <br>
        <h4>Chromosome 21</h4>
        <p> Variant count: 727,554 </p>
        <br>
        <h4>Chromosome 22</h4>
        <p> Variant count: 730,154 </p>
        <br>
        <h4>X Chromosome</h4>
        <p> Variant count: 1,969,365 </p>
        <br>
    "),
    easyClose = TRUE,
    footer = NULL))
}

getExistingDiscoveryStats <- function() {
  showModal(modalDialog(
    title = "Point Statistics for Existing Imputation - Discovery Set",
    HTML("
          <h4> Chromosome 1</h4>
        <p> Variant count: 4,145,679 </p>
        <br>
        <h4>Chromosome 2</h4>
        <p> Variant count: 4,468,841</p>
        <br>
        <h4>Chromosome 3</h4>
        <p> Variant count: 3,721,495 </p>
        <br>
        <h4>Chromosome 4</h4>
        <p> Variant count: 3,680,847 </p>
        <br>
        <h4>Chromosome 5</h4>
        <p> Variant count: 3,392,329 </p>
        <br>
        <h4>Chromosome 6</h4>
        <p> Variant count: 3,243,845 </p>
        <br>
        <h4>Chromosome 7</h4>
        <p> Variant count: 3,035,924 </p>
        <br>
        <h4>Chromosome 8</h4>
        <p> Variant count: 2,910,933 </p>
        <br>
        <h4>Chromosome 9</h4>
        <p> Variant count: 2,291,616 </p>
        <br>
        <h4>Chromosome 10</h4>
        <p> Variant count: 2,556,099</p>
        <br>
        <h4>Chromosome 11</h4>
        <p> Variant count: 2,589,050</p>
        <br>
        <h4>Chromosome 12</h4>
```

```
        <p> Variant count: 2,483,126 </p>
        <br>
        <h4>Chromosome 13</h4>
        <p> Variant count: 1,863,628</p>
        <br>
        <h4> Chromosome 14</h4>
        <p> Variant count: 1,674,384 </p>
        <br>
        <h4>Chromosome 15</h4>
        <p> Variant count: 1,521,610 </p>
        <br>
        <h4>Chromosome 16</h4>
        <p> Variant count: 1,680,287 </p>
        <br>
        <h4>Chromosome 17</h4>
        <p> Variant count: 1,477,226 </p>
        <br>
        <h4>Chromosome 18</h4>
        <p> Variant count: 1,479,501</p>
        <br>
        <h4>Chromosome 19</h4>
        <p> Variant count: 1,145,593 </p>
        <br>
        <h4>Chromosome 20</h4>
        <p> Variant count: 1,176,266</p>
        <br>
        <h4>Chromosome 21</h4>
        <p> Variant count: 727,199 </p>
        <br>
        <h4>Chromosome 22</h4>
        <p> Variant count: 729,593 </p>
        <br>
        <h4>X Chromosome</h4>
        <p> Variant count: 1,966,486 </p>
        <br>
    "),
    easyClose = TRUE,
    footer = NULL))
}
```

4.1.3 R Visualizations

This code was used to generate some of the visualizations used for the figures. Specifically imputation results for chromosomes 5 and 21 in addition to the visualizations for the *CREBRF* and *BTNL9* variants.

```
library(ggplot2)
library(reshape2)
library(tidyverse)
```

```
## Rsq viz ##
setwd("~/capstoneStats/data/imputation/chr21")

## Read in .info files from minimac4 ##
chr21_imputation_stats_no_samoan <- read.table("freeze9b.chr21.notopmed_no_samoan.info", header = T) %>%
  select(`MAF`, `Rsq`, `EmpRsq`) %>%
  na_if("-") %>%
  filter(`MAF` > 0) %>%
  mutate(rsquared = as.factor(if_else(Rsq < .8, "< 0.8", "> 0.8")))
chr21_imputation_stats_1_samoan <- read.table("freeze9b.chr21.notopmed_1_samoan.info", header = T) %>%
  select(`MAF`, `Rsq`, `EmpRsq`) %>%
  na_if("-") %>%
  filter(`MAF` > 0) %>%
  mutate(rsquared = as.factor(if_else(Rsq < .8, "< 0.8", "> 0.8")))
chr21_imputation_stats_4_samoan <- read.table("freeze9b.chr21.notopmed_4_samoan.info", header = T) %>%
  select(`MAF`, `Rsq`, `EmpRsq`) %>%
  na_if("-") %>%
  filter(`MAF` > 0) %>%
  mutate(rsquared = as.factor(if_else(Rsq < .8, "< 0.8", "> 0.8")))
chr21_imputation_stats_6_samoan <- read.table("freeze9b.chr21.notopmed_6_samoan.info", header = T) %>%
  select(`MAF`, `Rsq`, `EmpRsq`) %>%
  na_if("-") %>%
  filter(`MAF` > 0) %>%
  mutate(rsquared = as.factor(if_else(Rsq < .8, "< 0.8", "> 0.8")))
chr21_imputation_stats_12_samoan <- read.table("freeze9b.chr21.notopmed_12_samoan.info", header = T) %>%
  select(`MAF`, `Rsq`, `EmpRsq`) %>%
  na_if("-") %>%
  filter(`MAF` > 0) %>%
  mutate(rsquared = as.factor(if_else(Rsq < .8, "< 0.8", "> 0.8")))
chr21_imputation_stats_24_samoan <- read.table("freeze9b.chr21.notopmed_24_samoan.info", header = T) %>%
  select(`MAF`, `Rsq`, `EmpRsq`) %>%
  na_if("-") %>%
  filter(`MAF` > 0) %>%
  mutate(rsquared = as.factor(if_else(Rsq < .8, "< 0.8", "> 0.8")))
chr21_imputation_stats_48_samoan <- read.table("freeze9b.chr21.notopmed_48_samoan.info", header = T) %>%
  select(`MAF`, `Rsq`, `EmpRsq`) %>%
  na_if("-") %>%
  filter(`MAF` > 0) %>%
  mutate(rsquared = as.factor(if_else(Rsq < .8, "< 0.8", "> 0.8")))
chr21_imputation_stats_96_samoan <- read.table("freeze9b.chr21.notopmed_96_samoan.info", header = T) %>%
  select(`MAF`, `Rsq`, `EmpRsq`) %>%
  na_if("-") %>%
  filter(`MAF` > 0) %>%
  mutate(rsquared = as.factor(if_else(Rsq < .8, "< 0.8", "> 0.8")))
chr21_imputation_stats_384_samoan <- read.table("freeze9b.chr21.notopmed_384_samoan.info", header = T) %>%
  select(`MAF`, `Rsq`, `EmpRsq`) %>%
  na_if("-") %>%
  filter(`MAF` > 0) %>%
  mutate(rsquared = as.factor(if_else(Rsq < .8, "< 0.8", "> 0.8")))
chr21_imputation_stats_all_samoan <- read.table("freeze9b.chr21.notopmed_all_samoan.info", header = T) %>%
  select(`MAF`, `Rsq`, `EmpRsq`) %>%
  na_if("-") %>%
  filter(`MAF` > 0) %>%
  mutate(rsquared = as.factor(if_else(Rsq < .8, "< 0.8", "> 0.8")))

cutoff <- chr21_imputation_stats_no_samoan %>%
  filter(MAF > 0.0) %>%
  filter(MAF < 0.01) %>%
```

```
  mutate(strat = as.factor(if_else(Rsq < .3, "< 0.8", "> 0.8" )))


## group data by mean rsq at each MAF ##
tomerge_no <- chr21_imputation_stats_no_samoan %>%
  group_by(MAF) %>%
  summarize(`1KG+NoSamoans`=mean(Rsq))
tomerge_1 <- chr21_imputation_stats_1_samoan %>%
  group_by(MAF) %>%
  summarize(`1KG+1Samoan`=mean(Rsq))
tomerge_4 <- chr21_imputation_stats_4_samoan %>%
  group_by(MAF) %>%
  summarize(`1KG+4Samoans`=mean(Rsq))
tomerge_6 <- chr21_imputation_stats_6_samoan %>%
  group_by(MAF) %>%
 summarize(`1KG+6Samoans`=mean(Rsq))
tomerge_12 <- chr21_imputation_stats_12_samoan %>%
  group_by(MAF) %>%
  summarize(`1KG+12Samoans`=mean(Rsq))
tomerge_24 <- chr21_imputation_stats_24_samoan %>%
  group_by(MAF) %>%
  summarize(`1KG+24Samoans`=mean(Rsq))
tomerge_48 <- chr21_imputation_stats_48_samoan %>%
  group_by(MAF) %>%
  summarize(`1KG+48Samoans`=mean(Rsq))
tomerge_96 <- chr21_imputation_stats_96_samoan %>%
  group_by(MAF) %>%
  summarize(`1KG+96Samoans`=mean(Rsq))
tomerge_384 <- chr21_imputation_stats_384_samoan %>%
  group_by(MAF) %>%
  summarize(`1KG+384Samoans`=mean(Rsq))
tomerge_all <- chr21_imputation_stats_all_samoan %>%
  group_by(MAF) %>%
  summarize(`1KG+AllSamoans`=mean(Rsq))

## merge everything together and reshape df for plotting ##
merged1 <- merge(tomerge_no, tomerge_1, by="MAF")
merged2 <- merge(merged1, tomerge_4, by="MAF")
merged3 <- merge(merged2, tomerge_6, by="MAF")
merged4 <- merge(merged3, tomerge_12, by="MAF")
merged5 <- merge(merged4, tomerge_24, by="MAF")
merged6 <- merge(merged5, tomerge_48, by="MAF")
merged7 <- merge(merged6, tomerge_96, by="MAF")
merged8 <- merge(merged7, tomerge_384, by="MAF")
merged9 <- merge(merged8, tomerge_all, by="MAF")
p <- melt(merged9, id.vars = "MAF")

## visualize distribution of MAF and r-squared ##
ggplot(p, aes(MAF, value, fill=variable)) +
  geom_smooth(se=F, aes(color=variable)) +
  geom_vline(xintercept = 0.01, linetype="dotted") +
  #geom_vline(xintercept = 0.013, linetype="dotted") +
  scale_color_manual(values = c("1KG+NoSamoans" = "grey", "1KG+1Samoan" = "grey", "1KG+4Samoans" = "grey",
                                "1KG+6Samoans" = "grey", "1KG+12Samoans" = "grey", "1KG+24Samoans" = "grey",
                                "1KG+48Samoans" = "grey", "1KG+96Samoans" = "grey",
                                "1KG+384Samoans" = "#53B400", "1KG+AllSamoans" = "#619CFF")) +
  theme_minimal() +
  scale_x_continuous(name = "Minor Allele Frequency") +
```

```
  scale_y_continuous(name = "Mean R-square", limits = c(.5, 1)) +
  ggtitle("Chromosome 21")



## Variant stats viz - this is for all chromosomes ##
samoan_variants <- data.frame(Chromosome = c(1:22, "X"),
                              Variants = c(4148826, 4473212, 3723954, 3682018, 3394288,
                                           3245542, 3037910, 2911943, 2293025, 2557466,
                                           2591727, 2484806, 1866294, 1675545, 1523290,
                                           1681823, 1480477, 1480195, 1147567, 1176819,
                                           727554, 730154, 1969365))
ggplot(samoan_variants, aes(Chromosome, Variants)) +
  geom_col(fill = "dodgerblue") +
  scale_x_discrete(limits = c(1:22, "X")) +
  scale_y_continuous(labels = scales::comma) +
  theme_minimal()

affy_scaffold_variants <- data.frame(Chromosome = c(1:22, "X"),+
                              Variants = c(70380, 73062, 59983, 55334, 55764,
                                           55420, 46330, 48032, 40961, 47565,
                                           43926, 42010, 33739, 27656, 25774,
                                           27218, 20308, 26211, 11693, 22554,
                                           12387, 11236, 36465))
ggplot(affy_scaffold_variants, aes(Chromosome, Variants)) +
  geom_col(fill = "dodgerblue") +
  scale_x_discrete(limits = c(1:22, "X")) +
  scale_y_continuous(labels = scales::comma) +
  theme_minimal()

library(ggplot2)
library(reshape2)
library(tidyverse)

## Rsq viz ##
setwd("~/capstoneStats/data/imputation/chr5")

## Read in .info files from minimac4 ##
chr5_imputation_stats_no_samoan <- read.table("freeze9b.chr5.notopmed_no_samoan.info", header = T) %>%
  select(`MAF`, `Rsq`, `EmpRsq`) %>%
  na_if("-") %>%
  filter(`MAF` > 0) %>%
  mutate(rsquared = as.factor(if_else(Rsq < .8, "< 0.8", "> 0.8")))
chr5_imputation_stats_1_samoan <- read.table("freeze9b.chr5.notopmed_1_samoan.info", header = T) %>%
  select(`MAF`, `Rsq`, `EmpRsq`) %>%
  na_if("-") %>%
  filter(`MAF` > 0) %>%
  mutate(rsquared = as.factor(if_else(Rsq < .8, "< 0.8", "> 0.8")))
chr5_imputation_stats_24_samoan <- read.table("freeze9b.chr5.notopmed_24_samoan.info", header = T) %>%
  select(`MAF`, `Rsq`, `EmpRsq`) %>%
  na_if("-") %>%
  filter(`MAF` > 0) %>%
  mutate(rsquared = as.factor(if_else(Rsq < .8, "< 0.8", "> 0.8")))
chr5_imputation_stats_48_samoan <- read.table("freeze9b.chr5.notopmed_48_samoan.info", header = T) %>%
  select(`MAF`, `Rsq`, `EmpRsq`) %>%
  na_if("-") %>%
  filter(`MAF` > 0) %>%
  mutate(rsquared = as.factor(if_else(Rsq < .8, "< 0.8", "> 0.8")))
```

```
chr5_imputation_stats_96_samoan <- read.table("freeze9b.chr5.notopmed_96_samoan.info", header = T) %>%
 select(`MAF`, `Rsq`, `EmpRsq`) %>%
 na_if("-") %>%
 filter(`MAF` > 0) %>%
 mutate(rsquared = as.factor(if_else(Rsq < .8, "< 0.8", "> 0.8")))
chr5_imputation_stats_384_samoan <- read.table("freeze9b.chr5.notopmed_384_samoan.info", header = T) %>%
 select(`MAF`, `Rsq`, `EmpRsq`) %>%
 na_if("-") %>%
 filter(`MAF` > 0) %>%
 mutate(rsquared = as.factor(if_else(Rsq < .8, "< 0.8", "> 0.8")))
chr5_imputation_stats_all_samoan <- read.table("freeze9b.chr5.notopmed_all_samoan.info", header = T) %>%
 select(`MAF`, `Rsq`, `EmpRsq`) %>%
 na_if("-") %>%
 filter(`MAF` > 0) %>%
 mutate(rsquared = as.factor(if_else(Rsq < .8, "< 0.8", "> 0.8")))

### for creating data tables with num of variants at rsq threshold at specific MAFs ###
cutoff <- chr5_imputation_stats_24_samoan %>%
 filter(MAF > 0.0) %>%
 filter(MAF < 0.01) %>%
 mutate(strat = as.factor(if_else(Rsq < .3, "< 0.8", "> 0.8" )))

### finding mean rsq at each MAF ###
tomerge_no <- chr5_imputation_stats_no_samoan %>%
 group_by(MAF) %>%
 summarize(`1KG+NoSamoans`=mean(Rsq))
tomerge_1 <- chr5_imputation_stats_1_samoan %>%
 group_by(MAF) %>%
 summarize(`1KG+1Samoan`=mean(Rsq))
tomerge_24 <- chr5_imputation_stats_24_samoan %>%
 group_by(MAF) %>%
 summarize(`1KG+24Samoans`=mean(Rsq))
tomerge_48 <- chr5_imputation_stats_48_samoan %>%
 group_by(MAF) %>%
 summarize(`1KG+48Samoans`=mean(Rsq))
tomerge_96 <- chr5_imputation_stats_96_samoan %>%
 group_by(MAF) %>%
 summarize(`1KG+96Samoans`=mean(Rsq))
tomerge_384 <- chr5_imputation_stats_384_samoan %>%
 group_by(MAF) %>%
 summarize(`1KG+384Samoans`=mean(Rsq))
tomerge_all <- chr5_imputation_stats_all_samoan %>%
 group_by(MAF) %>%
 summarize(`1KG+AllSamoans`=mean(Rsq))

# tomerge_no <- chr5_imputation_stats_no_samoan %>%
#  group_by(MAF) %>%
#  summarize(`1KG+NoSamoans`=median(Rsq))
# tomerge_1 <- chr5_imputation_stats_1_samoan %>%
#  group_by(MAF) %>%
#  summarize(`1KG+1Samoan`=median(Rsq))
# tomerge_24 <- chr5_imputation_stats_24_samoan %>%
#  group_by(MAF) %>%
#  summarize(`1KG+24Samoans`=median(Rsq))
# tomerge_48 <- chr5_imputation_stats_48_samoan %>%
#  group_by(MAF) %>%
#  summarize(`1KG+48Samoans`=median(Rsq))
# tomerge_96 <- chr5_imputation_stats_96_samoan %>%
```

```
#   group_by(MAF) %>%
#   summarize(`1KG+96Samoans`=median(Rsq))
# tomerge_384 <- chr5_imputation_stats_384_samoan %>%
#   group_by(MAF) %>%
#   summarize(`1KG+384Samoans`=median(Rsq))
# tomerge_all <- chr5_imputation_stats_all_samoan %>%
#   group_by(MAF) %>%
#   summarize(`1KG+AllSamoans`=median(Rsq))

### merge all of the mean rsqs for each num samoan threshold and melt for plotting ###
merged1 <- merge(tomerge_no, tomerge_1, by="MAF")
merged2 <- merge(merged1, tomerge_24, by="MAF")
merged3 <- merge(merged2, tomerge_48, by="MAF")
merged4 <- merge(merged3, tomerge_96, by="MAF")
merged5 <- merge(merged4, tomerge_384, by="MAF")
merged6 <- merge(merged5, tomerge_all, by="MAF")

p <- melt(merged6, id.vars = "MAF")
#P <- P %>% filter(MAF > 0.7)

## visualize distribution of MAF and r-squared ##
ggplot(p, aes(MAF, value, fill=variable)) +
  geom_smooth(se=F, aes(color=variable)) +
#  geom_point() +
  geom_vline(xintercept = 0.01, linetype="dotted") +
  theme_minimal() +
#  coord_cartesian(ylim = c(.2, 1)) +
  scale_x_continuous(name = "Minor Allele Frequency") +
  scale_y_continuous(name = "Mean R-square", limits = c(.2, 1)) +
  ggtitle("Chromosome 5")

## CREBRF ##
crebrf_af <- data.frame(Samoans_Added = c(1, 24, 48, 96, 384, 1285, 1, 24, 48, 96, 384, 1285),
                        AF = c(0, .0002, .0005,.0102,.0348,.0921, 0, 0.28108, 0.28115,.28120,.28106,.28114),
                        Imputed = c("no", "no", "no", "no", "no", "no", "yes", "yes", "yes", "yes", "yes", "yes"),
                        rsq = c(NA, NA, NA, NA, NA, NA, 0, 0.99894, 0.99905 ,0.99911, 0.99951, 0.99870))
ggplot(crebrf_af, aes(Samoans_Added, AF, fill=Imputed)) +
  geom_point(aes(color=Imputed)) +
  geom_line(aes(color=Imputed)) +
  theme_classic() +
  scale_x_continuous(name = "Samoans Added", breaks = c(1,24,48,96,384,1285)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  scale_y_continuous(name = "rs373863828 MAF", breaks = c(0,.1,.2,0.28))

## BTNL9 ##
btnl9_af <- data.frame(Samoans_Added = c(1, 24, 48, 96, 384, 1285, 1, 24, 48, 96, 384, 1285),
                        AF = c(0.0001, 0.0001, 0.0005, 0.008, 0.03, 0.08, 0.20655, 0.22198, 0.20785, 0.20710, 0.22130, 0.21881),
                        Imputed = c("no", "no", "no", "no", "no", "no", "yes", "yes", "yes", "yes", "yes", "yes"),
                        rsq = c(NA, NA, NA, NA, NA, NA, 0, 0.92108, 0.91276, 0.86967, 0.91062, 94567))

ggplot(btnl9_af, aes(Samoans_Added, AF, fill=Imputed)) +
  geom_point(aes(color=Imputed)) +
  geom_line(aes(color=Imputed)) +
  theme_classic() +
  scale_x_continuous(name = "Samoans Added", breaks = c(1,24,48,96,384,1285)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  scale_y_continuous(name = "rs200884524 MAF", breaks = c(0,.1,.2,0.28))
```

4.2 Unix Code

4.2.1 Get_scaffold.sh

This code calculates the MAF of each variant from the Affymetrix 6.0 bcf files (repeated for all chromosomes)

```
for x in {1..22}
do
   bcftools view /home/mok36/imputation_code/freeze9b/discovery/7a_flip_nonseq_phased_data/noseq-flip-discovery-9b-
   chr${x}-phased.vcf.gz -G | bcftools query -f '%CHROM %POS %AN %AC{0}\n' | awk '{printf "%s %s %f\n",$1,$2,$4/$3}' >
   affy_chr${x}.txt
done
```

4.2.2 Get_samoans.sh

This code calculates the MAF of each variant from just the 1,285 Samons within the master reference panel then creates another file containing the count of each variant for every MAF.

```
for x in {1..22}
do
   bcftools view --force-samples --samples-file ~/capstoneStats/data/id_lists/ordered_samoans.txt
   /home/shared_data/samoa/wgs/freeze.9b/phased/freeze.9b.chr22.pass_only.phased.bcf | bcftools query -f '%CHROM
   %POS %AN %AC{0}\n' | awk '{printf "%s %s %f\n",$1,$2,$4/$3}' >
   /home/kja34/capstoneStats/data/reference_panels/freeze9b_chr${x}_samoans.txt
done

for x in *_samoans.txt
do
   awk '{print $3}' ${x} | sort | uniq -c > ${x}_count.txt
done
```

4.2.3 Get_1000g.sh

Does the same as get_samoans.sh but for the 1000genomes people

```
for x in {1..22}
do
```

```
    bcftools view --force-samples --samples-file ~/capstoneStats/data/id_lists/1000g_ids.txt /home/shared_data/sa-
    moa/wgs/freeze.9b/phased/freeze.9b.chr22.pass_only.phased.bcf | bcftools query -f '%CHROM %POS %AN %AC{0}\n' | awk
    '{printf "%s %s %f\n",$1,$2,$4/$3}' > /home/kja34/capstoneStats/data/reference_panels/freeze9b_chr${x}_1000g.txt
done

for x in *_1000g.txt
do
    awk '{print $3}' ${x} | sort | uniq -c > ${x}_count.txt
done
```

4.2.4 Existing Imputation

Creating the data to import into R for the existing imputations visualizations

```
for x in {1..22}
do
    bcftools query -f '%CHROM %POS %MAF %R2\n'
    /home/mok36/imputation_code/freeze9b/replication/${x}_replication_final_merge/replication-9b-hg38-final-merge-
    chr22.dose.vcf.gz | awk '{printf "%s %s %f %f\n",$1,$2,$3,$4}' >
    /home/kja34/capstoneStats/freeze9b_chrx{x}_existing_imputation_replication.txt
    bcftools query -f '%CHROM %POS %MAF %R2\n'
    /home/mok36/imputation_code/freeze9b/discovery/${x}_final_merge/discovery-9b-hg38-final-merge-chr${x}.dose.vcf.gz |
    awk '{printf "%s %s %f %f\n",$1,$2,$3,$4}' > /home/kja34/capstoneStats/freeze9b_chr{x}_existing_imputation_discovery.txt
done
```

4.2.5 Imputation

The code for running the actual imputations including splitting the panels up into certain amounts of Samoans. This is specifically for chromosome 21.

```
for x in 1 4 6 12 24 48 96 384 1285
do
    bcftools view -Oz --force-samples --samples-file /home/kja34/capstoneStats/data/id_lists/${x}_samoan_without_topmed.txt
    /home/shared_data/samoa/wgs/freeze.9b/phased/freeze.9b.chr21.pass_only.phased.bcf \
    -o /home/kja34/capstoneStats/data/imputation/chr21/freeze.9b.chr21.notopmed_${x}_samoan.vcf.gz

    Minimac3-omp --refHaps
    /home/kja34/capstoneStats/data/imputation/chr21/freeze.9b.chr21.notopmed_${x}_samoan.vcf.gz \
    --processReference \
    --myChromosome chr21 \
    --cpus 5 \
    --prefix /home/kja34/capstoneStats/data/imputation/chr21/freeze.9b.chr21.notopmed_${x}_samoan

    gunzip freeze9b.chr21.notopmed_${x}_samoan.m3vcf.gz
    sed -i 's/chr21/21/g' freeze9b.chr21.notopmed_${x}_samoan.m3vcf

    minimac4 --refHaps /home/kja34/capstoneStats/data/imputation/chr21/freeze9b.chr21.notopmed_${x}_samoan.m3vcf \
```

```
   --haps /home/mok36/imputation_code/freeze9b/discovery/7a_flip_nonseq_phased_data/noseq-flip-discovery-9b-chr21-
   phased.vcf.gz \
   --prefix /home/kja34/capstoneStats/data/imputation/chr21/freeze9b.chr21.notopmed_${x}_samoan \
   --format GT,DS \
   --allTypedSites \
   --ChunkLengthMb 20.00 \
   --ChunkOverlapMb 3.00 \
   --cpus 10
done
```

4.2.6 Genotype counts for *CREBRF* and *BTNL9*

This code gathers the vcf information from each variant from both before and after imputation, then calculates the genotype frequency.

```
for x in 1 24 48 96 384 1285
do
   echo >> crebrf
   echo ${x} samoan >> crebrf
   zgrep 173108771 freeze.9b.chr5.notopmed_${x}_samoan.vcf.gz >> crebrf
   done

   for x in 1 24 48 96 384 1285
   echo >> crebrf
   echo ${x} imputed samoans >> crebrf
   zgrep 173108771 freeze9b.chr5.notopmed_${x}_samoan.dose.vcf.gz >> crebrf
done


for x in 1 24 48 96 384 1285
do
   echo > btnl9
   echo ${x} samoan >> btnl9
   zgrep 181050285 freeze.9b.chr5.notopmed_${x}_samoan.vcf.gz >> btnl9

   echo >> btnl9
   echo ${x} imputed samoans >> btnl9
   zgrep 181050285 freeze9b.chr5.notopmed_${x}_samoan.dose.vcf.gz >> btnl9
done


cat /home/kja34/capstoneStats/data/imputation/chr5/crebrf \ |
perl -ane '
/^#/ and next;
%c = ();
foreach (@F[9..$#F]) { /^([^:]+)/ and $c{$1}++ }
print "$F[0]\t$F[1]";
foreach $gt (sort keys %c) { print "\t$gt:$c{$gt}" }
print "\n"
' > /home/kja34/capstoneStats/data/imputation/chr5/crebrf_af.txt

cat /home/kja34/capstoneStats/data/imputation/chr5/btnl9 \ |
perl -ane '
```

```
/^#/ and next;
%c = ();
foreach (@F[9..$#F]) { /^([^:]+)/ and $c{$1}++ }
print "$F[0]\t$F[1]";
foreach $gt (sort keys %c) { print "\t$gt:$c{$gt}" }
print "\n"
' > /home/kja34/capstoneStats/data/imputation/chr5/btnl9_af.txt
```

Bibliography

Ahmad, M., Sinha, A., Ghosh, S., Kumar, V., Davila, S., Yajnik, C. S., & Chandak, G. R. (2017). Inclusion of Population-specific Reference Panel from India to the 1000 Genomes Phase 3 Panel Improves Imputation Accuracy. *Scientific Reports*, *7*(1), 6733. https://doi.org/10.1038/s41598-017-06905-6

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., … National Eye Institute, N. I. H. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Bowen, M. S., Kolor, K., Dotson, W. D., Ned, R. M., & Khoury, M. J. (2012). Public health action in genomics is now needed beyond newborn screening. *Public Health Genomics*, *15*(6), 327–334. https://doi.org/10.1159/000341889

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P.-R., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., … Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*(10), 1284–1287. https://doi.org/10.1038/ng.3656

Gogarten, S. M., Bhangale, T., Conomos, M. P., Laurie, C. A., McHugh, C. P., Painter, I., Zheng, X., Crosslin, D. R., Levine, D., Lumley, T., Nelson, S. C., Rice, K., Shen, J., Swarnkar, R., Weir, B. S., & Laurie, C. C. (2012). GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics (Oxford, England)*, *28*(24), 3329–3331. https://doi.org/10.1093/bioinformatics/bts610

H.U. Bergmeyer, K. Gawehn, M. G. (1974). *Methods of Enzymatic Analysis, Vol. 1* (3rd ed.). Academic Press.

Hawley, N. L., Minster, R. L., Weeks, D. E., Viali, S., Reupena, M. S., Sun, G., Cheng, H., Deka, R., & Mcgarvey, S. T. (2014). Prevalence of adiposity and associated cardiometabolic risk factors in the Samoan genome-wide association study. *American Journal of Human Biology : The Official Journal of the Human Biology Council*, *26*(4), 491–501. https://doi.org/10.1002/ajhb.22553

Kamboh, M. I. (2004). Molecular genetics of late-onset Alzheimer's disease. *Annals of Human Genetics*, *68*(Pt 4), 381–404. https://doi.org/10.1046/j.1529-8817.2004.00110.x

Karow, J. (2017). Dante Labs Offers EUR 850 Whole Genome Sequencing and Interpretation for the First Time in the World. *PR Newswire*.

Kowalski, M. H., Qian, H., Hou, Z., Rosen, J. D., Tapia, A. L., Shan, Y., Jain, D., Argos, M., Arnett, D. K., Avery, C., Barnes, K. C., Becker, L. C., Bien, S. A., Bis, J. C., Blangero, J., Boerwinkle, E., Bowden, D. W., Buyske, S., Cai, J., … Li, Y. (2019). Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genetics*, *15*(12), e1008500. https://doi.org/10.1371/journal.pgen.1008500

Laurie, C. C., Doheny, K. F., Mirel, D. B., Pugh, E. W., Bierut, L. J., Bhangale, T., Boehm, F., Caporaso, N. E., Cornelis, M. C., Edenberg, H. J., Gabriel, S. B., Harris, E. L., Hu, F. B., Jacobs, K. B., Kraft, P., Landi, M. T., Lumley, T., Manolio, T. A., McHugh, C., … Weir, B. S. (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic Epidemiology*, *34*(6), 591–602. https://doi.org/10.1002/gepi.20516

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, *11*(7), 499–511. https://doi.org/10.1038/nrg2796

McWalter, K., & Gaviglio, A. (2015). Introduction to the Special Issue: Public Health Genetics and Genomics. In *Journal of genetic counseling* (Vol. 24, Issue 3, pp. 375–380). https://doi.org/10.1007/s10897-015-9825-9

Mills, M. C., & Rahal, C. (2020). The GWAS Diversity Monitor tracks diversity by disease in real time. In *Nature genetics* (Vol. 52, Issue 3, pp. 242–243). https://doi.org/10.1038/s41588-020-0580-y

Minster, R. L., Hawley, N. L., Su, C.-T., Sun, G., Kershaw, E. E., Cheng, H., Buhule, O. D., Lin, J., Reupena, M. S., Viali, S., Tuitele, J., Naseri, T., Urban, Z., Deka, R., Weeks, D. E., & McGarvey, S. T. (2016). A thrifty variant in CREBRF strongly influences body mass index in Samoans. *Nature Genetics*, *48*(9), 1049–1054. https://doi.org/10.1038/ng.3620

Mitt, M., Kals, M., Pärn, K., Gabriel, S. B., Lander, E. S., Palotie, A., Ripatti, S., Morris, A. P., Metspalu, A., Esko, T., Mägi, R., & Palta, P. (2017). Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *European Journal of Human Genetics*, *25*(7), 869–876. https://doi.org/10.1038/ejhg.2017.51

Molster, C. M., Bowman, F. L., Bilkey, G. A., Cho, A. S., Burns, B. L., Nowak, K. J., & Dawkins, H. J. S. (2018). The Evolution of Public Health Genomics: Exploring Its Past, Present, and Future. *Frontiers in Public Health*, *6*, 247. https://doi.org/10.3389/fpubh.2018.00247

Naj, A. C. (2019). Genotype Imputation in Genome-Wide Association Studies. *Current Protocols in Human Genetics*, *102*(1), e84. https://doi.org/10.1002/cphg.84

Peng, Z., Fan, W., Wang, L., Paudel, D., Leventini, D., Tillman, B. L., & Wang, J. (2017). Target enrichment sequencing in cultivated peanut (Arachis hypogaea L.) using probes designed from transcript sequences. *Molecular Genetics and Genomics : MGG*, *292*(5),

955–965. https://doi.org/10.1007/s00438-017-1327-z

Quick, C., Anugu, P., Musani, S., Weiss, S. T., Burchard, E. G., White, M. J., Keys, K. L., Cucca, F., Sidore, C., Boehnke, M., & Fuchsberger, C. (2020). Sequencing and imputation in GWAS: Cost-effective strategies to increase power and genomic coverage across diverse populations. *Genetic Epidemiology*, *44*(6), 537–549. https://doi.org/10.1002/gepi.22326

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., Pitsillides, A. N., LeFaive, J., Lee, S., Tian, X., Browning, B. L., Das, S., Emde, A.-K., Clarke, W. E., Loesch, D. P., … Consortium, N. T.-O. for P. M. (TOPMed). (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, *590*(7845), 290–299. https://doi.org/10.1038/s41586-021-03205-y

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), 59. https://doi.org/10.1038/s43586-021-00056-9