INCLUSION OF CUSTOM MARKERS DOES NOT IMPROVE GENOTYPE IMPUTATION ACCURACY IN A POPULATION ISOLATE

by

Devin Alexander Dikec

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
This thesis was presented

by

Devin Alexander Dikec

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
The field of genetics has been expanding as researchers can now study the entire genome to observe associations between diseases and specific allele markers. Due to the expensive nature of genotyping, an important set of markers can instead be genotyped so the remainder of the genome can be inferred from patterns computationally. This process of genotype imputation significantly reduces costs, but references patterns from mainly European populations when making computational predictions. The island nation of Samoa shows significant genetic differences as a population isolate with bottleneck effects, so little data shows how well these selected markers predict the Samoan genome. I aim to compare a standardized scaffold of allele markers using an Illumina chip against an enhanced scaffold with an additional 5,000 custom Samoan markers added for associations with traits of interest. I will be imputing the missing genotypes from each scaffold using a master reference panel of Samoan individuals with additional 1000 Genomes individuals sequenced by TopMed. The imputations are run for chromosome 21 as a test set for its smaller size, followed by chromosome 5 for its connection to Samoan specific traits of higher BMI and HDL levels. The generated results provided minor allele frequency (MAF) and two measures of imputation accuracy, $r^2$ and empirical $r^2$, to observe from each imputation using the unique scaffolds. These outputs were compared by various graphical representations. While the output values were not identical, their comparisons showed no discernable difference by scaffold. Most outputs
by scaffold were similar but a unique result of interest is $r^2$ comparison on chromosome 5. The custom Samoan markers showed unusual clumps in $r^2$ value outputs by scaffold. This change could show an improvement from the inclusion of the new markers but is the only significant result of difference from all comparisons. Therefore, the main conclusion seems to be that these included markers do not significantly alter imputation, but if available may be helpful for future accuracy in Samoan data.
# Table of Contents

Preface ................................................................................................................................. x
Acknowledgements .............................................................................................................. xi
1.0 Introduction ..................................................................................................................... 1
2.0 Methods .......................................................................................................................... 5
   2.1 Sample ....................................................................................................................... 5
   2.2 Scaffold Design ......................................................................................................... 6
   2.3 Reference Panel ....................................................................................................... 7
   2.4 Imputation ............................................................................................................... 7
   2.5 Imputation Statistics .............................................................................................. 8
   2.6 Comparison of Imputations ..................................................................................... 9
3.0 Results ............................................................................................................................ 11
   3.1 Scaffold Custom Content Differences .................................................................. 11
   3.2 Minor Allele Frequency Graph Comparison ....................................................... 12
   3.3 $r^2$ Graph Comparison ......................................................................................... 14
   3.4 Empirical $r^2$ Graph Comparison ......................................................................... 16
   3.5 Minor Allele Frequency by $r^2$ Comparison ....................................................... 18
4.0 Discussion ....................................................................................................................... 20
   4.1 Explanation of Scaffold Comparison .................................................................... 20
   4.2 Limitations and Strengths ...................................................................................... 20
   4.3 Previous Studies ...................................................................................................... 21
   4.4 Importance of Inclusion ....................................................................................... 21
4.5 Public Health Relevance .................................................................................................................. 22

4.6 Future Directions ............................................................................................................................. 22

Appendix A Code Snippets .................................................................................................................. 24

Bibliography .......................................................................................................................................... 29
List of Tables

Table 1. Scaffold count differences

........................................................................................................................................... 11
List of Figures

Figure 1. Visual example of imputation (Das 2018) .......................................................... 2
Figure 2. Imputation code .................................................................................................... 8
Figure 3. Example of imputation output file data table ....................................................... 9
Figure 4. Minor allele frequency comparison for chromosome 21 ..................................... 13
Figure 5. Minor allele frequency comparison for chromosome 5 ...................................... 14
Figure 6. \( r^2 \) comparison for chromosome 21 ................................................................. 15
Figure 7. \( r^2 \) comparison for chromosome 5 .................................................................. 16
Figure 8. Empirical \( r^2 \) comparisons for chromosome 21 .............................................. 17
Figure 9. Empirical \( r^2 \) comparison on chromosome 5 .................................................... 17
Figure 10. MAF by \( r^2 \) by scaffold comparison on chromosome 21 ............................... 18
Figure 11. MAF by \( r^2 \) comparison on chromosome 5 .................................................... 19
Preface

I’d like to thank my family and friends for their continuous support during the challenges I’ve faced throughout this process. I also want to thank the department of Human Genetics, my program advisor Ryan Minster, and my cohort for working together through these new experiences.
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 HHSN26820180002I). 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 HHSN268201800011). I gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.
1.0 Introduction

The human genome is a diverse underexplored set of information that science has yet to fully understand. The development of genome wide association studies (GWAS) has grown, allowing researchers to learn more about the genome by using associations with genetic markers between cases and controls or across a distribution for many phenotypes (Reay 2021, Shu 2018). The goal of these studies is to identify markers or locations that might ultimately inform prevention or treatment (Shu 2018). An obstacle to this understanding is the acquisition of genotypes, whether by genotyping arrays or by whole-genome sequencing, is currently a very expensive process. For instance, whole genome sequencing for the human genome can cost $1350 per sample, and even for smaller arrays one chip to cover 48 samples still costs around $800. The more customization required in the sequences, the higher the costs will rise. Additionally, most GWASs are centered on European populations. According to the GWAS diversity monitor, 96% of all GWAS participants are from European backgrounds, and so data for many populations in the world is lacking (GWAS Diversity 2022).

One solution to this issue is imputation. Imputation is the computational estimation of unknown genotypes using a smaller set of important sequenced allele sites and observed genetic patterns (Marchini 2010). Using a known reference set of standard allele patterns, researchers can produce estimated genotypes of multiple genomes. Rather than spending hundreds of dollars for each individual sequencing several thousand samples, the researchers can sequence a subset, array genotype the rest and estimate the missing genotypes at a highly reduced budget. (Quick 2010, Pistis 2014).
Two main data sets are necessary for imputation: a genotype scaffold and reference panel. The scaffold is a set of genotyped variants for one set of individuals with missing information at many variant sites, and the reference panel is a set of individuals with a larger number of genotyped variants, often measured through sequencing, used to fill in missing scaffold data through haplotype patterns. A visual example of imputation is provided in Figure 1 below, where the bottom block of genotypes represents a scaffold of important alleles that were sequenced, and the reference haplotypes at the top are the reference panel filling in the bottom data based on the common patterns.

![Visual example of imputation](image)

Figure 1. Visual example of imputation (Das 2018)

Even though imputation has its benefits, this process still estimates genotypes, and the accuracy of the results relies on several factors, including the reference panel used and the scaffold provided (Huang 2014). If the scaffold of allele markers was not designed with important
population patterns in mind, these starting scaffolds may be missing important alleles, and this missing data could alter the accuracy of the imputation results by predicting the wrong patterns (Mitt 2017).

The main population of interest for this study is from the Pacific Islands nation of Samoa. Samoans are interesting because they are an isolated population understudied in human genetics, and their historic isolation likely led to unique genetic patterns through bottleneck effects (Tsai 2004, Friedlaender 2008). Better imputation of genotypes for Samoan individuals could provide more accurate data for GWAS of phenotypes.

The scaffold for imputation in this study is genotyped markers from the Infinium Global Screening Array-24 v3.0 BeadChip (Illumina, San Diego, CA, USA). The “global” content of the Global Screening Array (GSA) is constructed from variants observed in the twenty-six population samples in Phase 3.1 of the 1000 Genomes Consortium. None of those population samples is of Polynesian peoples, and so up to 5,000 “Samoan-specific” markers were included in the scaffold of about 660,000 current markers. These additional markers were selected because they are more specific to patterns in Samoan individuals, so the goal is to test how their addition will change predictions from the original sequencing chip. More details about the scaffold are provided in Section 2.2.

This leads me to the main question I would like to answer. When using GSA data of Samoan participants as a scaffold for imputation, does including the additional 5,000 Samoan-specific markers improve imputation accuracy? The main goal being to observe whether there is evidence that custom markers used to impute genotypes leads to significant improvements in imputation for underrepresented populations.
Smaller non-European populations are typically ignored in research and unconsidered in designs of new diagnosis algorithms and disease interventions (GWAS Diversity 2022). If the inclusion of custom markers significantly improves imputation, then future researchers genotyping with chips can better understand how different populations may need additional markers for better accuracy. Better imputation can lead to a greater understanding of marginalized genomes and allow for better specific treatments and prevention measures. Public health and medicine are expanding into an era of precision interventions for communities and individuals (Gameiro 2018). A better understanding of how to improve imputation can inform whether these ideas are ready to treat all individuals or just the represented majority. The more we incorporate marginalized and underrepresented populations into research, the more health outcomes will improve around the globe.
2.0 Methods

2.1 Sample

The participants for this study come from the union of two participant samples recruited by the OLaGA Study Group (Obesity, Lifestyle, and Genetic Adaptations; "őlaga means “life” in Samoan): a longitudinal study from the early 1990s of 550 Samoans ages 29–88 (McGarvey 1990, Chin-Hong 1996, Galanis 1999) and a family study from 2002–03 of 908 adults ages 18–88 (Dai 2007, Åberg 2008, Dai 2008, Åberg 2009a, Åberg 2009b). The average age of the combined data set for both male and female participants is 34 years old at time of sample collection. All participants were genotyped using the Infinium Global Screening Array-24 v3.0 BeadChip (Illumina, San Diego, CA, USA). The participants in both studies resided either in the U.S. territory of American Samoa or the Independent State of Samoa. The research team also captured body mass index (BMI), high-density lipoprotein (HDL), and low-density lipoprotein (LDL) levels. BMI is each person’s weight in kilograms divided by their height in meters squared. HDL and LDL are measured through blood tests to look at overall levels of cholesterol. All participants provided written and spoken consent in both English and Samoan languages.
2.2 Scaffold Design

A scaffold is the set of allele markers sequenced that the imputation software will use to predict the missing alleles in between it. The genotyping array that was used to measure the genotypes for the scaffold was the GSA.

The OLaGA Study Group added 5,000 specific markers to the standard genotyping chip because of their connection to regions of interest in Samoans. These markers are considered the custom content to be compared for how the change the accuracy of imputation when included or removed from the scaffold. All data underwent quality control measures to ensure accuracy before continuing evaluations (Moors 2021). The main objective of this study, again, is to compare imputation quality when using the scaffold without this custom content to that when using the scaffold with this custom content. Due to resource constraints, I confined my work to two chromosomes: 21 and 5. Chromosome 21 was initially selected because it is the smallest chromosome and thus the most amenable to repeated imputation. I also worked with chromosome 5 because there are two markers on it associated with variance in BMI and HDL levels in Samoans.

The starting scaffold here was the genotypes from the GSA with the custom Samoan-specific markers included. These data were phased and divided into scaffolds for each individual chromosome. Next, I removed the custom markers from these scaffolds to create a second set of scaffolds for each chromosome that lacks those markers. To accomplish this removal, I extracted lists of the custom markers on chromosomes 21 and 5 from a master file of custom markers into two comma-separated value (CSV) files, one for each of those chromosomes. I used these two files as exclusion lists with bcftools (Danecek 2021) to remove the custom content from complete chromosome 21 and 5 scaffolds.
2.3 Reference Panel

A haplotype reference panel is needed to match patterns in the scaffold data and impute the missing genotypes. The reference panel here is the union of 1,285 Samoans and 4,974 individuals from the 1000 Genomes Project all sequenced by the TOPMed Program (Taliun 2021). These data are large enough and diverse enough to be used properly in imputation. The reference panel data are in binary variant call format (BCF); however, Minimac4 (Abecasis 2019), the imputation software I will be using to perform imputation, requires the input reference panel to be in Minimac3 variant call format (M3VCF) format, which can be produced by Minimac3. I used bcftools to convert the BCF into a compressed variant call format (VCF) file, the Minimac3 to convert the compressed VCF into M3VCF. Reference panels were created for both chromosomes 21 and 5 separate from the whole genome to reduce memory use and overall runtime.

2.4 Imputation

Imputation was conducted using Minimac4 (Abecasis 2019). The M3VCF reference files and the compressed VCF scaffold files were the input for imputation. I conducted four imputation operations: two for chromosome 21 using the chromosome 21 reference panel and the two scaffolds for chromosome 21, one with and one without the custom content; and two for chromosome 5 using the chromosome 5 reference panel and the two scaffolds for chromosome 5, one with and one without the custom content. Example code that I used to perform imputation is shown in Figure 2. Minimac4 produces a compressed VCF file that contains the imputed genotypes and
imputation statistics for each marker and an information file that contains just the imputation statistics for each marker.

```
#!/bin/sh
#$ -N minimac4
#$ -1 h_vmen=18G
minimac4 --refHaps /home/riemer/samoa10006-ref-chr21-M3VCF.m3/vcf.gz \
    --haps /home/riemer/minimac4/imputation_code/freezefb/replication/7_repl_genotype_phasing/replication-9b-chr21-phased.vcf.gz \
    --prefix imputed_chr21 \
    --format GT,DS \
    --allTypesDiffs \
    --chunkLengthMb 20.00 --chunkOverlapMb 3.00 \
    --cpus 10
```

Figure 2. Imputation code

### 2.5 Imputation Statistics

The minor allele frequency, $r^2$, and empirical $r^2$ are the three main pieces of data extracted from these outputs. The minor allele frequency shows how common the alternate allele is predicted at each SNP site. The $r^2$ value shows the generated imputation quality from the output. It is a calculation of observed variance over expected variance if the alleles were perfectly imputed. Empirical $r^2$ is the square of the correlation between imputed dosage and true genotype values.

Figure 3 below highlights the first twelve lines of information generated from a Minimac4 imputation process. The three main columns we will be investigate are minor allele frequency labeled as MAF, $r^2$ labeled as Rsq, and empirical $r^2$ labeled as EmpRsq. Other information presented with this data is the exact chromosome location of each SNP, and which alleles serve as reference and alternate. From this output data, graphs are generated between the two alternate scaffold data sets. Four output files were generated with similar data split by chromosome and scaffold type respectively.
Figure 3. Example of imputation output file data table

<table>
<thead>
<tr>
<th>SNP</th>
<th>REF.0.</th>
<th>ALT.1.</th>
<th>ALT_Frq</th>
<th>MAF</th>
<th>AvgCall</th>
<th>Rsq</th>
<th>Genotyped</th>
<th>LooRsq</th>
<th>EmpR</th>
<th>EmpRsq</th>
</tr>
</thead>
<tbody>
<tr>
<td>21:5032067:G:A</td>
<td>G</td>
<td>A</td>
<td>0.00008</td>
<td>0.00008</td>
<td>0.99992</td>
<td>0.00054</td>
<td>Imputed</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>21:5032115:G:A</td>
<td>G</td>
<td>A</td>
<td>0.00030</td>
<td>0.00030</td>
<td>0.99970</td>
<td>0.00037</td>
<td>Imputed</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>21:5032117:G:T</td>
<td>G</td>
<td>T</td>
<td>0.00041</td>
<td>0.00041</td>
<td>0.99959</td>
<td>0.00047</td>
<td>Imputed</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>21:5032377:G:T</td>
<td>G</td>
<td>T</td>
<td>0.00026</td>
<td>0.00026</td>
<td>0.99974</td>
<td>0.00343</td>
<td>Imputed</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>21:5032486:G:A</td>
<td>G</td>
<td>A</td>
<td>0.00033</td>
<td>0.00033</td>
<td>0.99967</td>
<td>0.00044</td>
<td>Imputed</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>21:5032585:G:A</td>
<td>G</td>
<td>A</td>
<td>0.00038</td>
<td>0.00038</td>
<td>0.99962</td>
<td>0.00023</td>
<td>Imputed</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>21:5032596:A/G</td>
<td>A</td>
<td>G</td>
<td>0.00020</td>
<td>0.00020</td>
<td>0.99980</td>
<td>0.00045</td>
<td>Imputed</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>21:5032613:A/C</td>
<td>A</td>
<td>C</td>
<td>0.00016</td>
<td>0.00016</td>
<td>0.99984</td>
<td>0.00053</td>
<td>Imputed</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>21:5032643:A/C</td>
<td>A</td>
<td>C</td>
<td>0.00027</td>
<td>0.00027</td>
<td>0.99973</td>
<td>0.00031</td>
<td>Imputed</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>21:5032723:T/C</td>
<td>T</td>
<td>C</td>
<td>0.00078</td>
<td>0.00078</td>
<td>0.99926</td>
<td>0.11204</td>
<td>Imputed</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>21:5032932:G:C</td>
<td>G</td>
<td>C</td>
<td>0.00036</td>
<td>0.00036</td>
<td>0.99964</td>
<td>0.00114</td>
<td>Imputed</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>21:5032982:C:T</td>
<td>C</td>
<td>T</td>
<td>0.00031</td>
<td>0.00031</td>
<td>0.99969</td>
<td>0.00135</td>
<td>Imputed</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

2.6 Comparison of Imputations

The imputation results were processed and compared using the coding language R (R Core Team 2021) via RStudio. Eight main comparison graphs were made through the ggplot2 package (Wickham 2016). For chromosome 21, a scatter plot compares the minor allele frequency (MAF) of each SNP as estimated in the imputation with the imputation without the custom content. A diagonal line with a slope of 1 and intercept of 0 serves as a standard for what values we expect if there is no difference. For chromosome 21, a similar visualization was generated of the $r^2$ values from the two imputations. Again, the points on the plot will represent the paired output values, and deviation from the expected trendline is a noticeable difference from scaffolds. The final scatter-plot comparison for chromosome 21 used the empirical $r^2$ values from the two imputations. Missing values in columns required filtering out before observations from the trendline could be observed.
Similarly, a scatter plot of MAFs, a scatter plot of \(r^2\) values, and scatter plot of empirical \(r^2\) values were created comparing the imputations of genotypes on chromosome 5. The larger size of the chromosome will generate more points on the scatterplot, but no changes were made to the input code.

Finally, I generated a comparison of the \(r^2\) values by minor allele frequency for each of the two imputations for chromosome 21 and chromosome 5. Because the density of points made visualizing the trend difficult, I originally used geom_smooth() to graph smooth trendlines for each imputation (with and without custom content) to see if there is any observable difference in quality across the MAF range on either chromosome. However, a large clustering of points caused the smooth plot to overshoot the maximum \(r^2\) value, so geom_hex() was used as a replacement comparison.

Chromosome 5 is observed in more detail due to the nature of the overarching study. Since Samoan data takes particular interest in high-density lipoprotein and BMI differences, two main genes of interest in Samoan populations include \textit{CREBRF} and \textit{BTNL9}. By comparing highlighted SNPs within the regions of these two genes, we could look for distinct prediction differences to estimate the Samoan-content’s predicted effects.
3.0 Results

3.1 Scaffold Custom Content Differences

Table 1 below highlights the differences between the separated and combined scaffold for how many markers were present before imputation. 42 custom Samoan markers were in the original chromosome 21 combined scaffold out of the 7698 total markers. After their removal, the total markers dropped to 7656 showing successful removal of the custom markers. The original chromosome 5 scaffold had 34,099 alleles and the custom showed 213 Samoan markers. After removal, the total size drops to 33,886, this output again confirmed proper creation of the new scaffold. The significantly smaller size of chromosome 21 made it an optimal candidate for early imputation practice. Then, chromosome 5 showed more custom content markers which is a positive sign for imputing the genes of interest. Even though original custom content lengths were provided of 56 and 277, not all the custom content could be incorporated into the master scaffold, so the length removed is different than expected. These changes in length still demonstrate successful creation of the modified scaffolds.

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Scaffold</th>
<th>Marker Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>Combined</td>
<td>7698</td>
</tr>
<tr>
<td>21</td>
<td>Custom Markers</td>
<td>42</td>
</tr>
<tr>
<td>21</td>
<td>Separated</td>
<td>7656</td>
</tr>
<tr>
<td>5</td>
<td>Combined</td>
<td>34099</td>
</tr>
<tr>
<td>5</td>
<td>Custom Markers</td>
<td>213</td>
</tr>
<tr>
<td>5</td>
<td>Separated</td>
<td>33886</td>
</tr>
</tbody>
</table>
3.2 Minor Allele Frequency Graph Comparison

For all scatterplots generated in ggplot, the $x$ axis shows the custom imputation output values while the $y$ axis shows the imputation outputs for when custom markers are removed. Points in the top left of the graphs represent SNPs where inclusion of custom markers results in lower prediction of values. Points in the bottom right show where inclusion of custom markers results in higher imputation values. Points matching the diagonal red line show that the imputation outputs were equal regardless of custom marker inclusion.

The generated minor allele frequencies between each scaffold on chromosome 21 (Figure 4) show similar outputs. Most points matching with the red line show that generated minor allele frequency was the same from each imputation even with changes in the custom markers from the starting haplotypes. No one point looks to be significantly different, but clusters of small deviation could merit future further investigation. 726,942 SNPs are shown on this chromosome 21 graph for each of the imputed regions.
The minor allele frequency comparison within chromosome 5 (Figure 5) appears more highly correlated from chromosome 21. The points appear to nearly match the expected trendline in red, and little output can be seen that show a preference to either scaffold. 3,392,113 SNPs are graphed for this comparison and very few points deviate from the standard trendline.
The $r^2$ comparisons between the two scaffolds show more significant deviation from the minor allele frequency data for both chromosomes. In figure 6, chromosome 21 SNPs appear in both the top left and bottom right corners. This output shows that certain SNPs have significant differences in $r^2$ when custom markers are compared. The expected trend is more points should appear below the standard red line since the custom markers should provide a greater accuracy and understanding for imputation. However, the trend still appears even in disparity.
The $r^2$ graph for chromosome 5 (Figure 7) highlights some of the most interesting results. Matching the MAF graphs, 3,392,113 specific regions are displayed compared to the 726,942 from chromosome 21. A large spread of points is shown below the trendline demonstrating higher predictability from the custom marker scaffold. Yet, the left side of the graph shows a large grouping of SNPs where the $r^2$ is around 0 for the custom scaffold, and much higher when the custom markers are removed. This high variability and clustering along the axis are unexpected results, and a cause for further investigation.
3.4 Empirical $r^2$ Graph Comparison

The final direct comparison to make for chromosome 21 is looking at the empirical $r^2$ values (Figure 8). This dataset had only about 1% of the original data included in chromosome 21 due to SNPs with no output values leading to only 7,656 SNPs. This graph has a few points spread out on each side of the red trendline, but many do line up and show the empirical $r^2$ values match regardless of custom content inclusion.
Figure 8. Empirical $r^2$ comparisons for chromosome 21

Figure 9 shows empirical $r^2$ trends on chromosome 5, and they match trends from chromosome 21. The size of the data is again about 1% the $r^2$ comparison due to missing data leaving only 33,886 SNPs to be graphed. The points lining up with the trendline show a prominent level of matching or similar values by SNP across the dataset. The most spread from the expected values appear in the top right corner when both empirical $r^2$ values are above 0.5.

Figure 9. Empirical $r^2$ comparison on chromosome 5
3.5 Minor Allele Frequency by $r^2$ Comparison

One of the final comparisons completed between the scaffolds was to observe how each MAF value correlated to $r^2$ values by SNP on chromosome 21. The side-by-side comparison of scaffolds shows the similarity between imputation outputs. A large cluster of SNPs shows values with low MAF and high $r^2$. Significantly low MAF and $r^2$ were filtered out for comparison of results, but no observable difference was recorded so the same process was checked for chromosome 5.

![Figure 10. MAF by $r^2$ by scaffold comparison on chromosome 21](image)

The chromosome 5 output showed almost identical results to chromosome 21 with no observable difference present between scaffolds. Similar trends were present with large clusters of data having low MAF and high $r^2$. Due to the similarity of scaffolds, extra tests were run to check for identical data between scaffold columns and chromosomes. It was confirmed that the
imputations all produced unique results, but the additional markers did not create a significant difference observable by graph.

Figure 11. MAF by $r^2$ comparison on chromosome 5
4.0 Discussion

4.1 Explanation of Scaffold Comparison

The comparisons shown in minor allele frequency, empirical $r^2$, and full scaffold comparison graphs (Figures 4, 5, 8, 9, 10 and 11) show almost no change from scaffold modifications. The $r^2$ graphs (Figures 6 and 7) show that these modifications alter results, but the greater accuracy and precision with the custom markers is still difficult to confirm. The most interesting observation is the $r^2$ values for chromosome 5 in figure 7. The custom content shows many $r^2$ values of 0 that without this included content ranges from high to low. This could be showing that the original scaffold is predicting incorrect $r^2$ values that are corrected by additional markers. The scaffold set with custom content included is expected to be better and so these results do not fully align with the original expectations. The scaffold is shown to have been correctly modified and imputation runs properly so this change in custom markers may not produce meaningful results.

4.2 Limitations and Strengths

A limitation to consider for this imputation is the relative size of the study. Due to the computational resources required and time available the study was limited to only two human chromosomes. These chromosomes were chosen for a specific purpose, but it still does limit results missing large chunks of the genome. A major strength of this study is the reliability of the test set. Since we can test on individuals who have been fully genotyped, we know exactly what to expect
from imputation results. This ability to double check our accuracy is a great method for confirmation of a correct imputation. Imputation can be hard to evaluate from its basis being an estimation, but the full set of data is a massive advantage.

4.3 Previous Studies

Previous imputation modifications show changes in accuracy but modified a different aspect of the process. Ahmad et al. 2017, and Yoo et al. 2019, demonstrate how the inclusion of other underrepresented groups adds to the accuracy of the study. Ahmad showed this research with Indian populations, and Yoo with rare northeast Asian population variants. A main difference between these studies and my own was where the new data was applied. Many previous studies apply their changes to their larger reference panels since this is the data being used to generate imputed genomes. My data has more custom Samoan markers within the haplotypes being imputed by the reference panel. Past research has shown that with more background information, computational predictions can operate in a more accurate fashion. This study’s results did not negate this idea with either improved or at least similar imputation results.

4.4 Importance of Inclusion

The global population has significant levels of variation across the genome brought about by factors such as ancestry and environment. The less diversity researchers account for with these imputation studies, the more likely predictions will be incorrect. Due to the cost of genotyping and
the new basis of these computational techniques, many regions of the world are ignored in the design of these reference panels and scaffold chips. The more we can include other underrepresented populations in future studies, the better these techniques can become and the treatments they support can thrive.

4.5 Public Health Relevance

Computational methods are not normally considered in a public health context, but their relevance can still be observed. Many complex diseases and rare variants are shifting to computational analysis to make stronger predictions for preventative actions. For the island nation of Samoa, a better understanding of its high prevalence of obesity along with trends for type 2 diabetes can provide more detail into these health conditions. The better understanding provided by this study can allow for more direct care by medical professionals to address issues on an individual level.

4.6 Future Directions

There are many ways this study could be expanded into the future. An easy start is to expand into testing reference panels with higher inclusivity of Samoan custom data. The analysis is currently in progress, but trends may follow those of this study and previous imputation studies. If more resources are available or sequencing becomes cheaper, sequencing more of the Samoan population can provide more unique markers to use as backbones for imputation. Since these
results can be shown to improve accuracy, the future of this research could also move to other smaller and underrepresented geographical regions to continue gathering data that can be applied for better research and treatment of the public.
Appendix A Code Snippets

Creation of chromosome 21 scaffold without custom content

```
#$ -S /bin/bash

bcftools view ~/../mok36/imputation_code/freeze9b/replication/7_replication_genotype_phasing/replication-9b-chr21-phased.vcf.gz -T ^sorted_custom2.csv -Oz -o scaffold_nocustom_21.vcf.gz
```

Creation of chromosome 5 scaffold without custom content

```
#$ -S /bin/bash

bcftools view ~/../mok36/imputation_code/freeze9b/replication/7_replication_genotype_phasing/replication-9b-chr5-phased.vcf.gz -T ^custom_content_5.csv -Oz -o scaffold_nocustom_chr5.vcf.gz
```

Imputation of chromosome 21 for scaffold with no custom content

```
#$ -S /bin/sh
#$ -N minimac4_nocustom_chr21
#$ -l h_vmem=18G
cd ~
minimac4 --refHaps ~/../rminster/MSGB_Imputation_Projects/samoa1000G-ref-chr21-M3VCF.m3vcf.gz \
--haps scaffold_nocustom_21.vcf.gz \
--prefix imputed_nocustom_chr21 \
--format GT,DS \
--allTypedSites \
--ChunkLengthMb 20.00 --ChunkOverlapMb 3.00 \
--cpus 10
#tabix -p vcf imputed_nocustom_chr21.vcf.gz
```

Imputation of chromosome 21 for scaffold with custom content

```
#$ -S /bin/sh
#$ -N minimac4 
#$ -l h_vmem=18G

minimac4 --refHaps /home/rminster/samoa1000G-ref-chr21-M3VCF.m3vcf.gz \
--haps /home/mok36/imputation_code/freeze9b/replication/7_replication_genotype_phasing/replication-9b-chr21-phased.vcf.gz \
--prefix imputed_chr21 \
--format GT,DS \
--allTypedSites \
--ChunkLengthMb 20.00 --ChunkOverlapMb 3.00 \
--cpus 10
#tabix -p vcf imputed_chr21.vcf.gz
```
Imputation for chromosome 5 for scaffold without custom content

```bash
#S -S /bin/sh
#S -N minimac4_nocustom_chr5
#S -I h_vmem=18G
cd ~
minimac4 --refHaps ~/../rminster/MSGB_Imputation_Projects/samoa1000G-ref-chr5-M3VCF.m3vcf.gz \  
--haps scaffold_nocustom_chr5.vcf.gz \  
--prefix imputed_nocustom_chr5 \  
--format GT,DS \  
--allTypedSites \  
--ChunkLengthMb 20.00 --ChunkOverlapMb 3.00 \  
--cpus 10
#tabix -p vcf imputed_nocustom_chr5.vcf.gz
```

Imputation for chromosome 5 for scaffold with custom content

```bash
#S -S /bin/sh
#S -N minimac4_chr5_full
#S -I h_vmem=18G
minimac4 --refHaps /home/rminster/MSGB_Imputation_Projects/samoa1000G-ref-chr5-M3VCF.m3vcf.gz \  
--haps /home/mok36/imputation_code/freeze9b/replication/7_replciation_genotype_phasing/replication-9b-chr5-phased.vcf.gz \  
--prefix imputed_chr5 \  
--format GT,DS \  
--allTypedSites \  
--ChunkLengthMb 20.00 --ChunkOverlapMb 3.00 \  
--cpus 10
#tabix -p vcf imputed_chr5.vcf.gz
```

Importing imputed data

```r
imputed_nocustom_5_output <- read.csv("~/imputed_nocustom_chr5.info", header=T, sep = "\t")
imputed_nocustom_21_output <- read.csv("~/imputed_nocustom_chr21.info", header=T, sep = "\t")
imputed_custom_5_output <- read.csv("~/imputed_chr5.info", header=T, sep = "\t")
imputed_custom_21_output <- read.csv("~/imputed_chr21.info", header=T, sep = "\t")
```

Creation of data frames for MAF and $r^2$ variables on each chromosome

```r
library(ggplot2)
R2_chr21 <- data.frame(imputed_custom_21_output$Rsq, imputed_nocustom_21_output$Rsq)
colnames(R2_chr21) <- c("custom_Rsq", "nocustom_Rsq")

R2_chr5 <- data.frame(imputed_custom_5_output$Rsq, imputed_nocustom_5_output$Rsq)
colnames(R2_chr5) <- c("custom_Rsq", "nocustom_Rsq")

MAF_chr21 <- data.frame(imputed_custom_21_output$MAF, imputed_nocustom_21_output$MAF)
colnames(MAF_chr21) <- c("custom_MAF", "nocustom_MAF")
```
Creation of empirical $r^2$ data frame and filtering process

```r
MAF_chr5 <- data.frame(imputed_custom_5_output$MAF, imputed_nocustom_5_output$MAF)
colnames(MAF_chr5) <- c("custom_MAF", "nocustom_MAF")

EmpR2_chr21 <- data.frame(imputed_custom_21_output$EmpRsq, imputed_nocustom_21_output$EmpRsq)
colnames(EmpR2_chr21) <- c("custom_EmpRsq", "nocustom_EmpRsq")

EmpR2_chr5 <- data.frame(imputed_custom_5_output$EmpRsq, imputed_nocustom_5_output$EmpRsq)
colnames(EmpR2_chr5) <- c("custom_EmpRsq", "nocustom_EmpRsq")

Filtered_EmpR2_chr21 <- EmpR2_chr21[!grepl("-", EmpR2_chr21$custom_EmpRsq),]
Filtered_EmpR2_chr21 <- Filtered_EmpR2_chr21[!grepl("-", Filtered_EmpR2_chr21$nocustom_EmpRsq),]

Filtered_EmpR2_chr5 <- EmpR2_chr5[!grepl("-", EmpR2_chr5$custom_EmpRsq),]
Filtered_EmpR2_chr5 <- Filtered_EmpR2_chr5[!grepl("-", Filtered_EmpR2_chr5$nocustom_EmpRsq),]
```

**Figure 4** Graph Code-MAF 21

```r
ggplot(data = MAF_chr21, aes(x=custom_MAF, y=nocustom_MAF)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color='#E41A1C') +
  ggtitle("Chromosome 21 MAF Comparison") +
  xlab("Custom MAF values") +
  ylab("No Custom MAF values") +
  theme_bw()
```

**Figure 5** Graph Code-MAF 5

```r
ggplot(data = MAF_chr5, aes(x=custom_MAF, y=nocustom_MAF)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color='#E41A1C') +
  ggtitle("Chromosome 5 MAF Comparison") +
  xlab("Custom MAF values") +
  ylab("No Custom MAF values") +
  theme_bw()
```

**Figure 6** Graph Code-R-squared 21

```r
ggplot(data = R2_chr21, aes(x=custom_Rsq, y=nocustom_Rsq)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color='#E41A1C') +
  ggtitle("Chromosome 21 R-squared Comparison") +
  xlab("Custom R-squared values") +
  ylab("No Custom R-squared values") +
  theme_bw()
```

**Figure 7** Graph Code-R-squared 5

```r
ggplot(data = R2_chr5, aes(x=custom_Rsq, y=nocustom_Rsq)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color='#E41A1C') +
  ggtitle("Chromosome 5 R-squared Comparison") +
  xlab("Custom R-squared values") +
  ylab("No Custom R-squared values") +
  theme_bw()
```
Figure 8 Graph Code: EmpR-squared 21
```r
ggplot(data = Filtered_EmpR2_chr21, aes(x=custom_EmpRsq, y=nocustom_EmpRsq)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color='#E41A1C') +
  ggtitle("Chromosome 21 EmpR-squared Comparison") +
  xlab("Custom EmpR-squared values") +
  ylab("No Custom EmpR-squared values") +
  theme_bw()
```

Figure 9 Graph Code: EmpR-squared 5
```r
ggplot(data = Filtered_EmpR2_chr5, aes(x=custom_EmpRsq, y=nocustom_EmpRsq)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color='#E41A1C') +
  ggtitle("Chromosome 5 EmpR-squared Comparison") +
  xlab("Custom EmpR-squared values") +
  ylab("No Custom EmpR-squared values") +
  theme_bw()
```

Preparing data frame for Figure 10
```r
chr21_custom <- data.frame(imputed_custom_21_output$SNP, imputed_custom_21_output$MAF,
                            imputed_custom_21_output$Rsq)
chr21_nocustom <- data.frame(imputed_nocustom_21_output$SNP, imputed_nocustom_21_output$MAF,
                            imputed_nocustom_21_output$Rsq)
colnames(chr21_custom) <- c("SNP", "MAF", "Rsq")
colnames(chr21_nocustom) <- c("SNP", "MAF", "Rsq")
chr21_rbind <- dplyr::bind_rows(list("custom scaffold"=chr21_custom, "no custom scaffold"=chr21_nocustom), .id = 'source')
compare_21 <- chr21_rbind %>%
  arrange(source, MAF, Rsq)
```

Figure 10 Graph Code: MAF by \( r^2 \) 21
```r
library(tidyverse)
library(tidylog)
library(tidyquant)

ggplot(data = compare_21 %>%
  filter(Rsq > 0.05) %>%
  filter(MAF > 0.005), aes(x = MAF, y = Rsq)) +
  geom_hex(bins = 10) +
  geom_ma(ma_fun = SMA,n = 5000, linetype = 1, color = "red") +
  ggtitle("Chromosome 21 MAF by R-squared by Scaffold") +
  xlab("MAF") +
  ylab("R-squared") +
  facet_grid(. ~ source) +
  scale_fill_viridis_c()
```

Preparing data frame for Figure 11
```r
chr5_custom <- data.frame(imputed_custom_5_output$SNP, imputed_custom_5_output$MAF,
                          imputed_custom_5_output$Rsq)
chr5_nocustom <- data.frame(imputed_nocustom_5_output$SNP, imputed_nocustom_5_output$MAF,
                          imputed_nocustom_5_output$Rsq)
colnames(chr5_custom) <- c("SNP", "MAF", "Rsq")
```

27
colnames(chr5_nocustom) <- c("SNP", "MAF", "Rsq")

chr5_rbind <- dplyr::bind_rows(list("custom scaffold"=chr21_custom, "no custom scaffold"=chr21_nocustom), .id = 'source')

compare_5 <- chr5_rbind %>%
  arrange(source, MAF, Rsq)

**Figure 11 Graph Code-MAF by $r^2$ 5**

```r
ggplot(data = compare_5 %>%
  filter(Rsq > 0.05) %>%
    filter(MAF > 0.005), aes(x = MAF, y = Rsq)) +
  geom_hex(bins = 10) +
  geom_ma(ma_fun = SMA, n = 5000, linetype = 1, color = "red") +
  ggtitle("Chromosome 5 MAF by R-squared by Scaffold") +
  xlab("MAF") +
  ylab("R-squared") +
  facet_grid(. ~ source) +
  scale_fill_viridis_c()
```

**Confirmation of Difference between scaffold $r^2$**

```r
all(chr21_custom$Rsq == chr21_nocustom$Rsq)
identical(chr21_custom$Rsq, chr21_nocustom$Rsq)
```
Bibliography


