Improving imputation quality in Samoans through the integration of population-specific sequences into existing reference panels

Jenna C. Carlson,1,2,* Mohanraj Krishnan,1 Shuwei Liu,1 Kevin J. Anderson,1 Jerry Z. Zhang,2 Toni-Ann J. Yapp,1 Elizabeth A. Chiyka,1 Devin A. Dikec,1 Hong Cheng,3 Take Naseri,4,5 Muagututia Sefuiva Reupena,6 Satupa’itea Viali,7,8 Ranjan Deka,3 Nicola L. Hawley,8 Stephen T. McGarvey,5,9 Daniel E. Weeks,1,2 Ryan L. Minster.1

1. Department of Human Genetics, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America
2. Department of Biostatistics, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America
3. Department of Environmental and Public Health Sciences, College of Medicine, University of Cincinnati, Cincinnati, Ohio, United States of America
5. International Health Institute, Department of Epidemiology, Brown University, Providence, Rhode Island, United States of America
6. Lutia i Puava ae Mapu i Fagalele, Apia, Samoa
7. School of Medicine, National University of Samoa, Apia, Samoa
8. Department of Chronic Disease Epidemiology, Yale School of Public Health, New Haven, Connecticut, United States of America
9. Department of Anthropology, Brown University, Providence, Rhode Island, United States of America

* Corresponding author: jnc35@pitt.edu

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Abstract

Genotype imputation is fundamental to association studies, and yet even gold standard panels like TOPMed are limited in the populations for which they yield good imputation. Specifically, Pacific Islanders are poorly represented in extant panels. To address this, we constructed an imputation reference panel using 1,285 Samoan individuals with whole-genome sequencing, combined with 1000 Genomes (1000G) samples, to create a reference panel that better represents Pacific Islander, specifically Samoan, genetic variation. We compared this panel to 1000G and TOPMed panels based on imputed variants using genotyping array data for 1,834 Samoan participants who were not part of the panels. The 1000G + 1285 Samoan panel yielded up to 2.25-2.76 times more well-imputed ($r^2 \geq 0.80$) variants than TOPMed and 1000G. There was improved imputation accuracy across the minor allele frequency (MAF) spectrum, although it was more pronounced for variants with $0.01 \leq \text{MAF} \leq 0.05$. Imputation accuracy ($r^2$) was greater for population-specific variants (high fixation index, $F_{ST}$) and those from larger haplotypes (high LD score). The gain in imputation accuracy over TOPMed was largest for small haplotypes (low LD score), reflecting the Samoan panel’s ability to capture population-specific variation not well tagged by other panels. We also augmented the 1000G reference panel with varying numbers of Samoan samples and found that panels with 48 or more Samoans included outperformed TOPMed for all variants with $\text{MAF} \geq 0.001$. This study identifies variants with improved imputation using population-specific reference panels and provides a framework for constructing other population-specific reference panels.
Introduction

Genotype imputation is fundamental to modern genome-wide association studies (GWAS), yielding a much denser association landscape in which fine-mapping of causal variants and genes can occur than what can be obtained using genotyping array data alone. However, the utility of genotype imputation depends upon the haplotypes that make up the reference panels. This is especially critical in populations that are underrepresented in genetic research, as imputation is much poorer when the reference panel does not contain good haplotype matches to that of the target population. Existing panels like 1000 Genomes (1000G)\(^1\), the Haplotype Reference Consortium (HRC)\(^2\), and the Trans-Omics for Precision Medicine (TOPMed)\(^3\) panels catalog human genome variation of numerous majority populations worldwide, especially those with Western European ancestries. The TOPMed panel contains the most ancestrally diverse set of samples and has addressed disparities in representation, especially for Hispanic- and African-admixed populations, greatly improving imputation accuracy in these populations\(^3,4\).

However, there are still populations for which these reference panels are inadequate. This is obvious in looking at the representation of Pacific Islanders, who are underrepresented in health research, and when included, are often aggregated along with those of Asian ancestry. In common, publicly available reference panels (1000G, HRC, and TOPMed), there are no samples explicitly from Oceania, and so the totality of genetic variation present in Pacific Islanders cannot be adequately represented in genetic association studies. This may further exacerbate the health disparities faced by Pacific Islanders, who are not only underrepresented in research but have a disproportionate burden of non-communicable diseases compared to majority populations.

For example, the Pacific Island nation of Samoa, which has a unique population history consisting of founder effects and bottlenecks\(^5-8\), also faces some of the highest rates of non-communicable disease, globally\(^9-13\). Despite this, most existing knowledge about disease
etiology comes from majority ethnic groups, even though disease presentation, progression, and response to treatment differ by ancestry\textsuperscript{11,14–21}. Specific to genetic studies, Pacific Islanders more broadly are often grouped into an “other/mixed” ancestral category which still only comprises <1\% of GWAS participants\textsuperscript{22}. However, because of their unique population history, there are regions of the genome for which Samoans have different variant (or allele) frequencies and haplotype structure (LD) from other majority populations, making the aggregation of them with other Oceanic or Asian-ancestry individuals inappropriate. Moreover, advances in knowledge on the genetic architecture of cardiometabolic traits from majority populations are often assumed to apply to all populations globally, which is premature at best (but, more likely, contributing to racial health disparity)\textsuperscript{23,24}.

To combat this, greater representation of Pacific Islanders is needed in gene-mapping studies of cardiometabolic traits. However, this cannot happen without improved genotype imputation. Evidence of this comes from studies based on cohorts of Samoans and American Samoans recruited in 1990-1995\textsuperscript{25–27}, 2002-2003\textsuperscript{28,29}, and 2010\textsuperscript{12}. We have performed GWAS to dissect the genetic architecture of cardiometabolic disease traits in these cohorts. Through this, we have identified variants of high frequency in Samoa not appearing (at least at high frequency) in other non-Pacific-Islander populations associated with body mass index\textsuperscript{30} and lipid traits\textsuperscript{31}, which may be at high frequency due to these founding events and bottlenecks. Interestingly, these variants were not present in initial GWAS using genotyping array data only. They were discovered using imputation based on Samoan haplotypes. For both loci, the genetic architecture of the GWAS signal looks markedly different after imputation, with the location of the sentinel variants shifting dramatically post-imputation. The discovery of these associations offers a case study into how genetic studies in Samoans and other Pacific-Islander populations would benefit from improved imputation.
In this study, we describe the strategy we employed to create a Samoan-specific reference panel containing phased haplotypes from 1000G participants and Samoan participants with whole-genome sequencing (WGS) from the Samoan Study, part of the TOPMed Consortium. Importantly, the Samoan Study, which consists of 1,285 participants of Samoan ancestry that are part of the TOPMed Consortium, was not included in the TOPMed reference panel because the needed approvals were not in place when it was being constructed. Samoan Study leaders are working with the Health Research Committee of the Ministry of Health of Samoa to determine to what extent whole-genome sequences from the Samoan Study will be part of public-facing services like the TOPMed imputation server and variant browser.

We investigated this panel derived from Samoan samples with WGS and compared it to several other reference panels, to determine the extent to which adding Samoan participants to the existing 1000G reference panel impacted imputation quality. Specifically, we sought to address these questions:

1) How does the imputation for a reference panel of 1,285 Samoans and the 1000G panel compare to other existing panels?
2) How many population-specific sequences are necessary to add to the 1000G panel to achieve good imputation accuracy?
3) Does the phasing of the reference panel impact imputed accuracy and genotype frequency?

Subjects and Methods

Participants for this study came from a population-based sample of 3,504 Samoan adults studied in 2010\textsuperscript{12}. Genotyping was performed for a subset of 3,182 Samoan adults using Genome-Wide Human SNP 6.0 arrays (Affymetrix) and extensive quality control was conducted via a pipeline developed by Laurie et al.\textsuperscript{32}. Additional details for sample genotyping and
genotype quality control are described in Minster et al.30 A subset (n=1,285) from this cohort
also had WGS data from TOPMed. These 1,285 individuals with WGS were optimally chosen
using INFOSTIP33 to capture the cohort’s genetic diversity
(http://www.cs.columbia.edu/~itsik/software/INFOSTIP.html). This study has been approved by
the Health Research Committee of the Samoa Ministry of Health and the institutional review
boards of Brown University and the University of Pittsburgh. All participants gave informed
consent.

Construction of the reference panels
For this paper, imputation was performed on chromosomes 5 and 21 for computational
efficiency and to examine imputation of a variant on chromosome 5 associated with atherogenic
lipid profiles that is enriched in Polynesians (rs200884524, [hg38] chr5:181050285 C>T in
BTNL931). Various imputation protocols (nine in total, described in Table 1; Figure S1) were
implemented to assess imputation accuracy compared to extant panels, measure the impact of
additional population-specific samples added to the panel, and investigate how the phasing of
the reference panel impacts the imputation of population-specific variation.

Samoan-specific genotype reference panels for imputation were created using the WGS data
from TOPMed. The in-house reference panels were comprised of high-quality (i.e., passing all
QC filters and with a minimum depth of 10), bi-allelic markers from the freeze 9 call subset of
2,504 1000G samples and up to 1,285 individuals from the Samoan Study. Separate reference
panels were constructed with 4, 24, 48, 96, and 384 Samoan individuals (i.e., fractions/multiples
of a 96 well plate) and all available 1,285 Samoan individuals. For protocols #3-7, the first 4, 24,
48, 96, or 384 individuals based on the INFOSTIP ranking were used to construct the reference
panels to capture the most genetic diversity possible with that sample size (Figure S2).

For protocol #9, phasing was performed on the entire reference panel using Eagle v2.4.134. For
protocols #3-8, the reference panel was not phased in-house; rather, phased haplotypes for the
2,504 1000G samples and 4, 24, 48, 96, 384, or 1,285 Samoan individuals were extracted from the entire set of calls from freeze 9; prior to extraction, the entirety of freeze 9 was phased altogether in a population-unaware manner by the TOPMed Informatics Research Center. Importantly, protocol #8 only differs from protocol #9 in the way the reference panel was phased.

**Imputation**

The reference panels were then used to impute genotype data for the remaining Samoan participants in the 2010 cohort who were not part of any reference panel (n = 1,834). To achieve this, the Affymetrix genotyped variant coordinates were converted to the hg38 genome build using LiftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver). Then, data were aligned to the reference panel using Genotype Harmonizer, using the mafAlign option to align to the minor allele when LD alignment failed and the minor allele frequency (MAF) was ≤ 30% in both the input and reference sets. The resulting variants were then phased with Eagle v2.4.1 and genome-wide imputation was performed using Minimac. An example of the imputation code is available in at https://github.com/jennaccarlson/imputation.

To benchmark the performance of the in-house panels, imputation for the same 1,834 non-sequenced Samoans was performed using the 1000G Phase III reference panel, via the Michigan Imputation Server (Table 1, protocol #2) and the TOPMed reference panel via the TOPMed Imputation Server (Table 1, protocol #1).

To compare imputation results, imputation information metrics were extracted from info files. The number of well-imputed (r² ≥ 0.80), non-monomorphic variants was tabulated for each imputation. Additionally, imputation sets were merged by position and alleles, accounting for possible strand flips; within MAF bins determined by the imputed variants, mean and median r² were calculated for the intersection of variants. Empirical r² was also examined for the
genotyped variants. We also assessed the impact of the reference panel to impute rs200884524, a population-specific variant of interest, by examining imputed MAF and \( r^2 \).

To assess if the imputation quality gains from using a reference panel that includes Samoan haplotypes were due to allele frequency differences or haplotype size, we calculated Weir and Cockerham’s fixation index (\( F_{ST} \)) and LD score. \( F_{ST} \) was calculated using the 1000G+1285S reference panel (protocol #9) with vcftools\(^3\); the population labels used were derived from the super populations listed in 1000G (AFR, AMR, EAS, EUR, SAS)\(^1\), with the 1,285 Samoans listed as a separate sixth population. LD scores were calculated using the 1000G+1285S reference panel (protocol #9) in GTCA\(^3\) with the default block size (1 Mb with an overlap of 500 kb between blocks). For visualization, \( F_{ST} \) and LD scores were each stratified into low and high at the median value within a given MAF bin.

We also investigated if the phasing of the reference panel impacted the imputation of population-specific variation. We hypothesized that for Samoan-specific variation, the phasing of genotypes when done alongside super majority populations with very different allele frequencies might be biased toward the haplotypes not common in Samoans, thereby decreasing the Samoan-specific allele frequency. To investigate this, we compared imputation quality (\( r^2 \)) and MAF for the 1000G + 1285S panel (protocol #9) and 1000G + 1285S multiethnic-phased panel (protocol #8) for Samoan-specific variants, defined as those in the 99th percentile of \( F_{ST} \) and with MAF > 0.05 in the 1,834 Samoans imputed with the 1000G + 1285S reference panel (protocol #9).

Results

We compared the average imputation quality (\( r^2 \)) and the number of well-imputed variants (\( r^2 \geq 0.8 \)) across 9 sets of imputed genotypes on chromosomes 5 and 21 to assess the performance
of the various imputation reference panels which included TOPMed, 1000G Phase III, and seven in-house panels constructed with 1000G plus varying numbers of Samoans (Table 1).

**The number of well-imputed \( r^2 \geq 0.80 \) variants depends on the reference panel**

The 1000G + 1285S panel resulted in the largest number of well-imputed variants, yielding 243,534 more variants on chromosome 5 than the TOPMed panel, with 16,112 of these having a MAF of at least 0.05 (Figure 1; Table S1). The yield in well-imputed variants increased as the number of Samoans included in the reference panel increased across all MAF bins (an up-to-42% increase in well-imputed variants compared to the TOPMed imputation). The largest gains were seen for variants with MAF < 0.05, with up to a 118% increase in well-imputed variants compared to the TOPMed imputation. Moreover, the 1000G + 1285S panel had over 250,000 well-imputed unique variants (not seen in TOPMed or 1000G imputations) on chromosome 5 and over 60,000 on chromosome 21 in addition to the variants shared across panels (Figure 2).

We also examined “lost” variants, those with imputation quality \( r^2 < 0.80 \) in the 1000G imputation, to see how many could be “rescued” (had imputation quality \( r^2 \geq 0.80 \)) across other panels (Figure 3; Table S2). The proportion of lost variants rescued was highest for variants with greater MAF across all panels, with nearly 100% of lost variants recovered in some cases (MAF ≥ 0.05 for the 1000G + 1285S panel). The panels with at least 48 Samoans included (1000G + 48S, 1000G + 96S, 1000G + 384S, 1000G + 1285S) rescued more variants overall than TOPMed (15.0% to 35.5% for Samoan panels compared to 12.2% to 13.2% for TOPMed). Over 95% of lost variants with MAF ≥ 0.05 were rescued for the reference panels with 384 or more Samoans. Even for the 1000G + 4S panel, 29.4% to 37.1% of variants with MAF ≥ 0.05 were rescued.
Panels with as few as 48 Samoans outperform TOPMed

Imputation quality increased as the number of Samoans represented in the panel increased. The panels with 48 or more Samoans included (1000G + 48S, 1000G + 96S, 1000G + 384S, 1000G + 1285S) outperformed the TOPMed panel for all variants with imputed MAF ≥ 0.001 (based on the 1000G + 1285S imputation) (Figure 4; Table S3-4). When looking at empirical r² for genotyped variants only, which is a more robust metric for evaluating imputation for lower frequency variants, the panels including Samoans outperform both the TOPMed and 1000G panels (Figure 5). Average imputation quality tended to be lower when examining variants based on TOPMed-imputed MAF instead of 1000G + 1285S-imputed MAF (Figure 4), consistent with expectations that Samoan allele frequencies would differ from those in TOPMed due to lower-quality matches of local haplotypes in the TOPMed reference panel. However, even when using TOPMed-imputed MAF, the panels containing at least 96 Samoans (1000G + 96S, 1000G + 384S, 1000G + 1285S) outperformed the TOPMed panel across the allele frequency spectrum.

Gains in imputation quality from Samoan representation in the reference panel are greatest for low LD score variants

Imputation quality increased for all panels as F_ST increased, as expected due to higher confidence in matching haplotypes with population-specific variation (Figure 6). We also stratified F_ST into high and low strata about the median value within each MAF bin to examine the relative difference in imputation quality by F_ST (Table S5). For variants with MAF < 0.05, those with high F_ST (i.e., F_ST greater than the median F_ST of all variants within the same MAF bin) tended to be imputed better across all reference panels. For variants with MAF ≥ 0.05, there did not appear to be a substantial difference in imputation quality by F_ST except for the 1000G panel (Figure 7). A similar trend was observed for LD scores, with imputation quality increasing as LD score increased (Figure 8) and high LD score variants (i.e., LD score greater
than the median LD score of all variants within the same MAF bin (Table S5) being imputed better on average than low LD score ones across the MAF spectrum, consistent with expectations (Figure 9).

To identify for which variants imputation with the 1000G + 1285S panel was better than the TOPMed panel, we compared the difference in mean $r^2$ across the two sets of imputed variants. The gains in average $r^2$ for the 1000G + 1285S panel compared to the TOPMed panel were more pronounced for lower frequency variants ($0.001 \leq \text{MAF} < 0.05$) with low $F_{ST}$ and low LD scores (Figure 10). However, for variants with MAF $\geq 0.05$, there was no discernible difference in average imputation quality across low and high $F_{ST}$ variants but a small difference comparing low and high LD score variants (Figure 10). Low LD score variants had a larger gain in imputation quality than low $F_{ST}$ variants for the 1000G + 1285S panel over the TOPMed panel (Figure 11). This suggests that the gains in imputation from the 1000G + 1285S panel over TOPMed were not only from allele frequency differences but also from haplotype structure and that the imputation was the most improved for lower-frequency variants on smaller haplotypes.

Population-specific variation and the effect of phasing

For rs200884524 the imputation quality is highest ($r^2 = 0.89-0.95$) for the reference panels that include Samoan haplotypes (Table 2). Additionally, the MAF among the imputed genotypes in the 1,834 non-reference-panel Samoans from protocols #3-9 (0.207-0.222) is much closer to what is expected – targeted genotyping in separate cohorts of similar ancestry yielded a MAF of 0.202-0.233\(^3\). The TOPMed imputation yielded a much more accurate imputed MAF than 1000G, which may suggest the presence of some Pacific haplotypes at this locus.

We also compared $r^2$ and MAF between imputed variants from the protocol #8 and #9 panels for variants in the 99th percentile of $F_{ST}$ that were common in Samoans (MAF $\geq 0.05$ in the protocol #8 imputation). For both chromosome 5 and 21, $r^2$ and MAF were similar for these variants. The estimated correlation between $r^2$ from the 1000G +1285S panel and that from the 1000G +
1285S in-house phasing panel for the 4,468 variants above the 99\textsuperscript{th} percentile $F_{ST}$ on chromosome 5 was 0.999343 (Figure S9). For the 1,011 common variants above the 99\textsuperscript{th} percentile $F_{ST}$ on chromosome 21, the estimated correlation in $r^2$ was 0.999989 (Figure S9). However, the imputed MAF tended to be slightly lower among these variants in the imputation from the 1000G + 1285S in-house phasing reference panel (Table S6).

**Discussion**

The 1000G+1285S panel outperformed all other panels tested, especially for low-frequency variants. There were substantial gains in imputation quality ($r^2$) and the number of well-imputed variants, primarily for low-frequency variants. Compared to the commonly recommended TOPMed panel, the number of well-imputed variants available increased 4\% to 118\% with the 1000G + 1285S panel and average imputation quality increased 3\% to 27\%.

In looking at variants with low imputation quality in the 1000G imputation, a large proportion of them were “rescued”, i.e., had high imputation quality, in panels containing 48 or more Samoans. These panels rescued more variants, especially common variants, than the TOPMed panel did. Improved imputation performance for low-frequency variants (0.001 ≤ MAF < 0.01) was achieved with as few as 48 Samoans included in the reference panel; for variants with MAF ≥ 0.01, imputation quality was high with as few as 24 Samoans included in the reference panel. Importantly, this sample size is not necessarily prescriptive for future studies planning to perform WGS to create their own population-specific as these 48 individuals were not randomly selected – they were chosen to maximize the genetic variation present in the original set of samples, which were all of Samoan ethnicity (defined by participant self-reporting four Samoan grandparents) and therefore comprised a quite homogeneous population with little to no admixture.\textsuperscript{8} These findings may reflect the fact that in this cohort, the 48 individuals yielded over 70\% of the information content from INFOSTIP (Figure S2). In other populations, the number of
individuals needed to see improved imputation will depend on the genetic variation of the target  

13

population. For populations with larger genetic heterogeneity, including those with lots of  

different haplotype sizes, more individuals will likely be needed to achieve  

improved imputation accuracy over extant panels.

The trends in imputation quality by FST and LD score suggest that these improvements in $r^2$ for  

the 1000G + 1285S panel are driven by differences in both the LD patterns and haplotypes of  

Samoans and allele frequency differences across populations. While imputation quality was  

highest for high LD score variants, the gains in imputation quality for the 1000G + 1285S panel  

over the TOPMed panel were the largest for low LD score variants across the MAF spectrum.  

This suggests that the panel inclusive of Samoan haplotypes is enriched for the imputation of  

population-specific variants in small LD blocks that would not be predicted well by other  

reference panels, even though the T (minor) allele count of this variant is still low (22 alleles  

present in Freeze 5 according to BRAVO39; https://bravo.sph.umich.edu/ 

freeze5/hg38/).

As expected, imputation quality using the panels including Samoan haplotypes was higher for a  

population-specific variant, rs200884524 in $BTNL9$. Moreover, the frequency of rs200884524  

changed as the composition of the reference panel changed. This has implications for post-  

imputation association testing for variants common in only one population as statistical power is  

lower for variants with lower MAF at the same effect size. More broadly, the poorer imputation  

quality makes accurate inference and association testing more difficult.

We hypothesized that the phasing of reference panel haplotypes would affect imputation quality,  

specifically, that phasing taking place in large multiethnic consortia might bias the phasing of the  

reference panel haplotypes more common in majority-represented populations. Here, we did not  

see evidence of this. While it may impact specific variants, the phasing of the reference panel  

haplotypes does not appear to impact imputation quality or imputed genotype frequencies on  

average. There was a small difference in the number of well-imputed variants, but neither panel
consistently produced more well-imputed variants than the other. However, it is worth noting that the in-house phasing of the panel containing 1000G and 1,285 Samoans phased the reference panel haplotypes altogether, rather than in a population-aware manner. Further investigation is needed to examine if stratifying by super population before phasing would yield a difference in imputed genotypes than phasing altogether, especially when working with population isolates like Samoans. This would be particularly important if haplotypes harboring population-specific variants appeared at very different frequencies in the less-represented population in the reference panel.

Taken together, this work highlights the utility of augmenting extant panels with population-specific sequences, resulting in more well-imputed variants for association testing, higher imputation quality overall, and better imputation of population-specific variation. In this work, these gains were the greatest for low-frequency population-specific variants in smaller LD blocks. These results suggest that augmenting extant reference panels with even a small number of population-specific haplotypes can substantially increase the number of well-imputed variants available for analysis compared to gold-standard extant panels, which may allow for better inclusion of even more diverse populations in genetic association studies. This improvement in the representativeness of genetic studies, by extension, will lay the foundation for more equitable access to personalized medicine innovation in this and other Pacific populations.

Acknowledgements

We would like to thank the Samoan participants of the study, local village authorities, and the many Samoan and other field workers over the years. We acknowledge the Samoan Ministry of Health and the Samoa Bureau of Statistics for their support of this research. We give particular
thanks to two research assistants, Melania Selu and Vaimoana Lupematasila, who contributed to the 2010 recruitment and continue to assist us in our work in Samoa.

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 NHLBI TOPMed: Samoan (phs000972.v5.p1) was performed at New York Genome Center Genomics (HHSN268201500016C) and Northwest Genomics Center (HHSN268201100037C). 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 (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

This work was funded by the National Institute of Health grants R01HL093093 (STM), R01HL133040 (RLM), R01AG09375 (STM), R01HL52611 (MI Kamboh), R01DK59642 (STM), and R01DK55406 (RD). Genotyping was performed in the Core Genotyping Laboratory at the University of Cincinnati, funded by National Institutes of Health grant P30ES006096 (S Pinney).

Author Contributions

JCC – Conceptualization, Data curation, Formal Analysis, Methodology, Visualization, Writing – original draft, Writing – review & editing

MK – Methodology, Software, Writing – review & editing

SL – Methodology, Software, Writing – review & editing
KA – Formal Analysis, Writing – review & editing

JZZ – Methodology, Software

TJY – Visualization, Writing – review & editing

EAC – Formal Analysis

DAD – Formal Analysis

HC – Data curation, Resources, Investigation

TN – Project administration

MSR – Data curation, Project administration

SV – Project administration, Writing – review & editing

RD – Data curation, Resources, Investigation, Funding acquisition, Writing – review & editing

NLH – Data curation, Funding acquisition, Project administration, Writing – review & editing

STM – Data curation, Funding acquisition, Project administration, Writing – review & editing

DEW – Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – review & editing

RLM – Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – review & editing

Declaration of Interests

The authors declare no competing interests.
Data and code availability

The discovery cohort data used for this study are available through dbGaP (accession numbers: phs000914.v1.p1 and phs000972.v5.p1). A template for the imputation code is available at https://github.com/jennaccarlson/imputation; additional code for creating the imputation reference panels and imputation quality analysis will be made available upon request.

References


5. Friedlaender, J.S., Friedlaender, F.R., Reed, F.A., Kidd, K.K., Kidd, J.R., Chambers, G.K.,
Islanders. PLOS Genetics 4, e19. 10.1371/journal.pgen.0040019.
6. Skoglund, P., Posth, C., Sirak, K., Spriggs, M., Valentin, F., Bedford, S., Clark, G.R.,
Reepmeyer, C., Petchey, F., Fernandes, D., et al. (2016). Genomic insights into the
and Deka, R. (2004). Distribution of genome-wide linkage disequilibrium based on
microsatellite loci in the Samoan population. Human Genomics 1, 327. 10.1186/1479-7364-
1-5-327.
8. Harris, D.N., Kessler, M.D., Shetty, A.C., Weeks, D.E., Minster, R.L., Browning, S.,
10.1073/pnas.1913157117.
prevention and control of noncommunicable diseases in the Western Pacific Region:
country capacity survey 2019 (WHO Regional Office for the Western Pacific).
Cardiology Reviews 11, 238–245.
cardiometabolic risk factors in the samoan genome-wide association study. American


format conversion for genotype data integration. BMC Research Notes 7, 901.


**Figure 1. Count of well-imputed variants by reference panel.** The number of variants with imputation quality \( r^2 \geq 0.80 \) by reference panel is plotted across minor allele frequency (MAF) bins for chromosomes 5 and 21. (These counts are printed in Table S1.)
Figure 2. Count of uniquely imputed variants by reference panel. The number of well imputed variants \( (r^2 \geq 0.80) \) unique to the 1000G + 1285S, TOPMed, and 1000G reference panel (and all possible combinations) is plotted for chromosomes 5 and 21.
Figure 3. Count of rescued variants by reference panel. Number of imputed variants that were lost ($r^2 < 0.80$) in the 1000G imputation but were rescued ($r^2 \geq 0.80$) in other panels by minor allele frequency (MAF) and reference panels across chromosome.
Figure 4. Imputation quality versus minor allele frequency by reference panel. Mean imputation quality ($r^2$) by minor allele frequency (MAF) bin (left column: MAF based on 1000G + 1285S imputation; right column: MAF based on the TOPMed imputation) is plotted by reference panel on chromosomes 5 and 21. (Plots with median $r^2$ are given in Figure S3. Tables with mean $r^2$ across reference panel are in Table S3.)
**Figure 5. Empirical $r^2$ versus minor allele frequency by reference panel for genotyped variants.** Mean empirical $r^2$ by minor allele frequency (MAF) bin (based on 1000G + 1285S imputation) is plotted by reference panel on chromosomes 5 and 21. (Plots with median empirical $r^2$ are given in Figure S4.)
Figure 6. Imputation quality versus fixation index by reference panel. Mean imputation quality ($r^2$) by empirical Weir and Cockerham Fixation Index ($F_{ST}$) deciles is plotted by reference panel on chromosomes 5 and 21. (Plots with median $r^2$ are given in Figure S5.)
Figure 7 Imputation quality versus fixation index by reference panel and MAF bin. Mean imputation quality ($r^2$) by median-stratified Weir and Cockerham Fixation Index ($F_{ST}$) across MAF (based on 1000G + 1285S imputation) is plotted by reference panel on chromosomes 5 and 21. $F_{ST}$ was stratified into ‘low’ and ‘high’ about the median $F_{ST}$ within each MAF bin (values given in Table S5). (Plots with median $r^2$ are given in Figure S6.)
Figure 8 Imputation quality versus LD score by reference panel. Mean imputation quality ($r^2$) by empirical LD score deciles is plotted by reference panel on chromosomes 5 and 21. (Plots with median $r^2$ are given in Figure S7.)
**Figure 9** Imputation quality versus LD score by reference panel and MAF bin. Mean imputation quality ($r^2$) by median-stratified LD score across MAF (based on 1000G + 1285S imputation) is plotted by reference panel on chromosomes 5 and 21. LD score was stratified into ‘low’ and ‘high’ about the median LD score within each MAF bin (values given in Table S5). (Plots with median $r^2$ are given in Figure S8.)

**Figure 10.** Gains in imputation quality from TOPMed panel to 1000G + 1285S panel by MAF for $F_{ST}$ and LD score. Fold change in mean imputation quality ($r^2$ 1000G+1285S / $r^2$ TOPMed) by median-stratified $F_{ST}$ (left) and LD scores (right) across MAF (based on...
1000G + 1285S imputation) is plotted for chromosomes 5 and 21. Both $F_{ST}$ and LD score were stratified into ‘low’ and ‘high’ about the median $F_{ST}$/LD score within each MAF bin (values given in Table S5). Reference line at $Y=1$ (red) is given to indicate no difference between TOPMed and 1000G + 1285S panels.
Figure 11. Differences in gains in imputation quality from TOPMed panel to 1000G + 1285S panel by MAF for $F_{ST}$ and LD score. Both $F_{ST}$ and LD score were stratified into ‘low’ and ‘high’ about the median $F_{ST}$/LD score within each MAF bin (values given in Table S5). The difference in fold change between low and high scoring variants (low – high) from (Figure 6) is plotted across MAF separately for $F_{ST}$ and LD score for chromosomes 5 and 21. Reference line at Y=0 (gray) is given to indicate no difference between high and low $F_{ST}$/LD Score.
### Table 1. Names and descriptions of tested imputation protocols

<table>
<thead>
<tr>
<th>Protocol Number</th>
<th>Reference Panel Name</th>
<th>Server Name</th>
<th>Reference Panel Samples</th>
<th>Phasing of reference panel performed with…</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TOPMed</td>
<td>TOPMed Imputation Server</td>
<td>TOPMed</td>
<td>Reference panel samples only</td>
</tr>
<tr>
<td>2</td>
<td>1000G</td>
<td>Michigan Imputation Server</td>
<td>1000G Phase III (GRCh38/hg38)</td>
<td>Reference panel samples only</td>
</tr>
<tr>
<td>3</td>
<td>1000G + 4S</td>
<td>In-house</td>
<td>1000G + 4 Samoans</td>
<td>All TOPMed freeze 9 samples</td>
</tr>
<tr>
<td>4</td>
<td>1000G + 24S</td>
<td>In-house</td>
<td>1000G + 24 Samoans</td>
<td>All TOPMed freeze 9 samples</td>
</tr>
<tr>
<td>5</td>
<td>1000G + 48S</td>
<td>In-house</td>
<td>1000G + 48 Samoans</td>
<td>All TOPMed freeze 9 samples</td>
</tr>
<tr>
<td>6</td>
<td>1000G + 96S</td>
<td>In-house</td>
<td>1000G + 96 Samoans</td>
<td>All TOPMed freeze 9 samples</td>
</tr>
<tr>
<td>7</td>
<td>1000G + 384S</td>
<td>In-house</td>
<td>1000G + 384 Samoans</td>
<td>All TOPMed freeze 9 samples</td>
</tr>
<tr>
<td>8</td>
<td>1000G + 1285S</td>
<td>In-house</td>
<td>1000G + 1285 Samoans</td>
<td>All TOPMed freeze 9 samples</td>
</tr>
<tr>
<td>9</td>
<td>1000G + 1285S</td>
<td>In-house</td>
<td>1000G + 1285 Samoans</td>
<td>Reference panel (1000G + 1285S) samples only</td>
</tr>
</tbody>
</table>
Table 2. Imputation quality and minor allele frequency (MAF) for rs200884524

(chr5:181050285 C/T [hg38]) in BTNL9 across reference panels. Targeted genotyping in two separate cohorts of 1,666 Samoan and American Samoan adults yielded a MAF of 0.202-0.233.

<table>
<thead>
<tr>
<th>Reference panel</th>
<th>$r^2$</th>
<th>MAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000G</td>
<td>0.17098</td>
<td>0.00076</td>
</tr>
<tr>
<td>1000G + 4S</td>
<td>0.90687</td>
<td>0.20424</td>
</tr>
<tr>
<td>1000G + 24S</td>
<td>0.91276</td>
<td>0.22198</td>
</tr>
<tr>
<td>1000G + 48S</td>
<td>0.86967</td>
<td>0.20785</td>
</tr>
<tr>
<td>1000G + 96S</td>
<td>0.91062</td>
<td>0.2071</td>
</tr>
<tr>
<td>1000G + 384S</td>
<td>0.94567</td>
<td>0.2213</td>
</tr>
<tr>
<td>1000G + 1285S</td>
<td>0.94360</td>
<td>0.21852</td>
</tr>
<tr>
<td>1000G + 1285S in-house phasing</td>
<td>0.94360</td>
<td>0.21852</td>
</tr>
<tr>
<td>TOPMed</td>
<td>0.79515</td>
<td>0.16620</td>
</tr>
</tbody>
</table>