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] Muagututi‘a 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
4. Naseri & Associates Public Health Consultancy Firm and Family Health Clinic, Apia, Samoa
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

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 $0.01 \leq MAF \leq 0.05$. 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 $MAF \geq 0.001$. 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

71 etiology comes from majority ethnic groups, even though disease presentation, progression,
72 and response to treatment differ by ancestry[11,14–21]. Specific to genetic studies, Pacific Islanders
73 more broadly are often grouped into an “other/mixed” ancestral category which still only
74 comprises <1% of GWAS participants[22]. However, because of their unique population history,
75 there are regions of the genome for which Samoans have different variant (or allele) frequencies
76 and haplotype structure (LD) from other majority populations, making the aggregation of them
77 with other Oceanic or Asian-ancestry individuals inappropriate. Moreover, advances in
78 knowledge on the genetic architecture of cardiometabolic traits from majority populations are
79 often assumed to apply to all populations globally, which is premature at best (but, more likely,
80 contributing to racial health disparity)[23,24].

81 To combat this, greater representation of Pacific Islanders is needed in gene-mapping studies of
82 cardiometabolic traits. However, this cannot happen without improved genotype imputation.
83 Evidence of this comes from studies based on cohorts of Samoans and American Samoans
84 recruited in 1990-1995[25–27], 2002-2003[28,29], and 2010[12]. We have performed GWAS to dissect
85 the genetic architecture of cardiometabolic disease traits in these cohorts. Through this, we
86 have identified variants of high frequency in Samoa not appearing (at least at high frequency) in
87 other non-Pacific-Islander populations associated with body mass index[30] and lipid traits[31],
88 which may be at high frequency due to these founding events and bottlenecks. Interestingly,
89 these variants were not present in initial GWAS using genotyping array data only. They were
90 discovered using imputation based on Samoan haplotypes. For both loci, the genetic
91 architecture of the GWAS signal looks markedly different after imputation, with the location of
92 the sentinel variants shifting dramatically post-imputation. The discovery of these associations
93 offers a case study into how genetic studies in Samoans and other Pacific-Islander populations
94 would benefit from improved imputation.

95 In this study, we describe the strategy we employed to create a Samoan-specific reference
96 panel containing phased haplotypes from 1000G participants and Samoan participants with
97 whole-genome sequencing (WGS) from the Samoan Study, part of the TOPMed Consortium.
98 Importantly, the Samoan Study, which consists of 1,285 participants of Samoan ancestry that
99 are part of the TOPMed Consortium, was not included in the TOPMed reference panel because
100 the needed approvals were not in place when it was being constructed. Samoan Study leaders
101 are working with the Health Research Committee of the Ministry of Health of Samoa to
102 determine to what extent whole-genome sequences from the Samoan Study will be part of
103 public-facing services like the TOPMed imputation server and variant browser.

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

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

114 Subjects and Methods

115 Participants for this study came from a population-based sample of 3,504 Samoan adults
116 studied in 2010[12]. Genotyping was performed for a subset of 3,182 Samoan adults using
117 Genome-Wide Human SNP 6.0 arrays (Affymetrix) and extensive quality control was conducted
118 via a pipeline developed by Laurie et al.[32]. Additional details for sample genotyping and

119 genotype quality control are described in Minster et al.[30] A subset (n=1,285) from this cohort
120 also had WGS data from TOPMed. These 1,285 individuals with WGS were optimally chosen
121 using INFOSTIP[33] to capture the cohort's genetic diversity
122 (http://www.cs.columbia.edu/~itsik/software/INFOSTIP.html). This study has been approved by
123 the Health Research Committee of the Samoa Ministry of Health and the institutional review
124 boards of Brown University and the University of Pittsburgh. All participants gave informed
125 consent.

126 *Construction of the reference panels*

127 For this paper, imputation was performed on chromosomes 5 and 21 for computational
128 efficiency and to examine imputation of a variant on chromosome 5 associated with atherogenic
129 lipid profiles that is enriched in Polynesians (rs200884524, [hg38] chr5:181050285 C>T in
130 *BTNL9*[31]). Various imputation protocols (nine in total, described in Table 1; Figure S1) were
131 implemented to assess imputation accuracy compared to extant panels, measure the impact of
132 additional population-specific samples added to the panel, and investigate how the phasing of
133 the reference panel impacts the imputation of population-specific variation.

134 Samoan-specific genotype reference panels for imputation were created using the WGS data
135 from TOPMed. The in-house reference panels were comprised of high-quality (i.e., passing all
136 QC filters and with a minimum depth of 10), bi-allelic markers from the freeze 9 call subset of
137 2,504 1000G samples and up to 1,285 individuals from the Samoan Study. Separate reference
138 panels were constructed with 4, 24, 48, 96, and 384 Samoan individuals (i.e., fractions/multiples
139 of a 96 well plate) and all available 1,285 Samoan individuals. For protocols #3-7, the first 4, 24,
140 48, 96, or 384 individuals based on the INFOSTIP ranking were used to construct the reference
141 panels to capture the most genetic diversity possible with that sample size (Figure S2).

142 For protocol #9, phasing was performed on the entire reference panel using Eagle v2.4.1[34]. For
143 protocols #3-8, the reference panel was not phased in-house; rather, phased haplotypes for the

144 2,504 1000G samples and 4, 24, 48, 96, 384, or 1,285 Samoan individuals were extracted from
145 the entire set of calls from freeze 9; prior to extraction, the entirety of freeze 9 was phased
146 altogether in a population-unaware manner by the TOPMed Informatics Research Center[3].
147 Importantly, protocol #8 only differs from protocol #9 in the way the reference panel was
148 phased.

149 *Imputation*

150 The reference panels were then used to impute genotype data for the remaining Samoan
151 participants in the 2010 cohort who were not part of any reference panel (*n* = 1,834). To achieve
152 this, the Affymetrix genotyped variant coordinates were converted to the hg38 genome build
153 using LiftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver). Then, data were aligned to the
154 reference panel using Genotype Harmonizer[35], using the mafAlign option to align to the minor
155 allele when LD alignment failed and the minor allele frequency (MAF) was ≤ 30% in both the
156 input and reference sets. The resulting variants were then phased with Eagle v2.4.1[34] and
157 genome-wide imputation was performed using Minimac4[36]. An example of the imputation code
158 is available in at https://github.com/jennaccarlson/imputation.

159 To benchmark the performance of the in-house panels, imputation for the same 1,834 non-
160 sequenced Samoans was performed using the 1000G Phase III reference panel, via the
161 Michigan Imputation Server[36] (Table 1, protocol #2) and the TOPMed reference panel via the
162 TOPMed Imputation Server[3] (Table 1, protocol #1).

163 To compare imputation results, imputation information metrics were extracted from info files.
164 The number of well-imputed ($r^2 \geq 0.80$), non-monomorphic variants was tabulated for each
165 imputation. Additionally, imputation sets were merged by position and alleles, accounting for
166 possible strand flips; within MAF bins determined by the imputed variants, mean and median r^2
167 were calculated for the intersection of variants. Empirical r^2 was also examined for the

168 genotyped variants. We also assessed the impact of the reference panel to impute
169 rs200884524, a population-specific variant of interest, by examining imputed MAF and r^2.

170 To assess if the imputation quality gains from using a reference panel that includes Samoan
171 haplotypes were due to allele frequency differences or haplotype size, we calculated Weir and
172 Cockerham's fixation index (F_{ST}) and LD score. F_{ST} was calculated using the 1000G+1285S
173 reference panel (protocol #9) with vcftools[37]; the population labels used were derived from the
174 super populations listed in 1000G (AFR, AMR, EAS, EUR, SAS)[1], with the 1,285 Samoans
175 listed as a separate sixth population. LD scores were calculated using the 1000G+1285S
176 reference panel (protocol #9) in GTCA[38] with the default block size (1 Mb with an overlap of 500
177 kb between blocks). For visualization, F_{ST} and LD scores were each stratified into low and high
178 at the median value within a given MAF bin.

179 We also investigated if the phasing of the reference panel impacted the imputation of
180 population-specific variation. We hypothesized that for Samoan-specific variation, the phasing of
181 genotypes when done alongside super majority populations with very different allele frequencies
182 might be biased toward the haplotypes not common in Samoans, thereby decreasing the
183 Samoan-specific allele frequency. To investigate this, we compared imputation quality (r^2) and
184 MAF for the 1000G + 1285S panel (protocol #9) and 1000G + 1285S multiethnic-phased panel
185 (protocol #8) for Samoan-specific variants, defined as those in the 99th percentile of F_{ST} and with
186 MAF > 0.05 in the 1,834 Samoans imputed with the 1000G + 1285S reference panel (protocol
187 #9).

188 Results

189 We compared the average imputation quality (r^2) and the number of well-imputed variants ($r^2 \geq$
190 0.8) across 9 sets of imputed genotypes on chromosomes 5 and 21 to assess the performance

191 of the various imputation reference panels which included TOPMed, 1000G Phase III, and
192 seven in-house panels constructed with 1000G plus varying numbers of Samoans (Table 1).

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

194 The 1000G + 1285S panel resulted in the largest number of well-imputed variants, yielding
195 243,534 more variants on chromosome 5 than the TOPMed panel, with 16,112 of these having
196 a MAF of at least 0.05 (Figure 1; Table S1). The yield in well-imputed variants increased as the
197 number of Samoans included in the reference panel increased across all MAF bins (an up-to-
198 42% increase in well-imputed variants compared to the TOPMed imputation). The largest gains
199 were seen for variants with MAF < 0.05, with up to a 118% increase in well-imputed variants
200 compared to the TOPMed imputation. Moreover, the 1000G + 1285S panel had over 250,000
201 well-imputed unique variants (not seen in TOPMed or 1000G imputations) on chromosome 5
202 and over 60,000 on chromosome 21 in addition to the variants shared across panels (Figure 2).

203 We also examined "lost" variants, those with imputation quality $r^2 < 0.80$ in the 1000G
204 imputation, to see how many could be "rescued" (had imputation quality $r^2 \geq 0.80$) across other
205 panels (Figure 3; Table S2). The proportion of lost variants rescued was highest for variants
206 with greater MAF across all panels, with nearly 100% of lost variants recovered in some cases
207 (MAF ≥ 0.05 for the 1000G + 1285S panel). The panels with at least 48 Samoans included
208 (1000G + 48S, 1000G + 96S, 1000G + 384S, 1000G + 1285S) rescued more variants overall
209 than TOPMed (15.0% to 35.5% for Samoan panels compared to 12.2% to 13.2% for TOPMed).
210 Over 95% of lost variants with MAF ≥ 0.05 were rescued for the reference panels with 384 or
211 more Samoans. Even for the 1000G + 4S panel, 29.4% to 37.1% of variants with MAF ≥ 0.05
212 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 r^2 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

238 than the median LD score of all variants within the same MAF bin (Table S5)) being imputed
239 better on average than low LD score ones across the MAF spectrum, consistent with
240 expectations (Figure 9).

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

252 *Population-specific variation and the effect of phasing*

253 For rs200884524 the imputation quality is highest (r^2 = 0.89-0.95) for the reference panels that
254 include Samoan haplotypes (Table 2). Additionally, the MAF among the imputed genotypes in
255 the 1,834 non-reference-panel Samoans from protocols #3-9 (0.207-0.222) is much closer to
256 what is expected – targeted genotyping in separate cohorts of similar ancestry yielded a MAF of
257 0.202-0.233[31]. The TOPMed imputation yielded a much more accurate imputed MAF than
258 1000G, which may suggest the presence of some Pacific haplotypes at this locus.

259 We also compared r^2 and MAF between imputed variants from the protocol #8 and #9 panels for
260 variants in the 99th percentile of F_{ST} that were common in Samoans ($MAF \geq 0.05$ in the protocol
261 #8 imputation). For both chromosome 5 and 21, r^2 and MAF were similar for these variants. The
262 estimated correlation between r^2 from the 1000G +1285S panel and that from the 1000G +

263 1285S in-house phasing panel for the 4,468 variants above the 99th percentile F_{ST} on
264 chromosome 5 was 0.999343 (Figure S9). For the 1,011 common variants above the 99th
265 percentile F_{ST} on chromosome 21, the estimated correlation in r^2 was 0.999989 (Figure S9).
266 However, the imputed MAF tended to be slightly lower among these variants in the imputation
267 from the 1000G + 1285S in-house phasing reference panel (Table S6).

268 Discussion

269 The 1000G+1285S panel outperformed all other panels tested, especially for low-frequency
270 variants. There were substantial gains in imputation quality (r^2) and the number of well-imputed
271 variants, primarily for low-frequency variants. Compared to the commonly recommended
272 TOPMed panel, the number of well-imputed variants available increased 4% to 118% with the
273 1000G + 1285S panel and average imputation quality increased 3% to 27%.

274 In looking at variants with low imputation quality in the 1000G imputation, a large proportion of
275 them were “rescued”, i.e., had high imputation quality, in panels containing 48 or more
276 Samoans. These panels rescued more variants, especially common variants, than the TOPMed
277 panel did. Improved imputation performance for low-frequency variants ($0.001 \leq MAF < 0.01$)
278 was achieved with as few as 48 Samoans included in the reference panel; for variants with MAF
279 ≥ 0.01, imputation quality was high with as few as 24 Samoans included in the reference panel.
280 Importantly, this sample size is not necessarily prescriptive for future studies planning to
281 perform WGS to create their own population-specific as these 48 individuals were not randomly
282 selected – they were chosen to maximize the genetic variation present in the original set of
283 samples, which were all of Samoan ethnicity (defined by participant self-reporting four Samoan
284 grandparents) and therefore comprised a quite homogeneous population with little to no
285 admixture.[8] These findings may reflect the fact that in this cohort, the 48 individuals yielded over
286 70% of the information content from INFOSTIP (Figure S2). In other populations, the number of

287 individuals needed to see improved imputation will depend on the genetic variation of the target
288 population. For populations with larger genetic heterogeneity, including those with lots of
289 admixture or different haplotype sizes, more individuals will likely be needed to achieve
290 improved imputation accuracy over extant panels.

291 The trends in imputation quality by F_{ST} and LD score suggest that these improvements in r^2 for
292 the 1000G + 1285S panel are driven by differences in both the LD patterns and haplotypes of
293 Samoans and allele frequency differences across populations. While imputation quality was
294 highest for high LD score variants, the gains in imputation quality for the 1000G + 1285S panel
295 over the TOPMed panel were the largest for low LD score variants across the MAF spectrum.
296 This suggests that the panel inclusive of Samoan haplotypes is enriched for the imputation of
297 population-specific variants in small LD blocks that would not be predicted well by other
298 reference panels, even though the T (minor) allele count of this variant is still low (22 alleles
299 present in Freeze 5 according to BRAVO[39]; https://bravo.sph.umich.edu/freeze5/hg38/).

300 As expected, imputation quality using the panels including Samoan haplotypes was higher for a
301 population-specific variant, rs200884524 in *BTNL9*. Moreover, the frequency of rs200884524
302 changed as the composition of the reference panel changed. This has implications for post-
303 imputation association testing for variants common in only one population as statistical power is
304 lower for variants with lower MAF at the same effect size. More broadly, the poorer imputation
305 quality makes accurate inference and association testing more difficult.

306 We hypothesized that the phasing of reference panel haplotypes would affect imputation quality,
307 specifically, that phasing taking place in large multiethnic consortia might bias the phasing of the
308 reference panel haplotypes more common in majority-represented populations. Here, we did not
309 see evidence of this. While it may impact specific variants, the phasing of the reference panel
310 haplotypes does not appear to impact imputation quality or imputed genotype frequencies on
311 average. There was a small difference in the number of well-imputed variants, but neither panel

312 consistently produced more well-imputed variants than the other. However, it is worth noting
313 that the in-house phasing of the panel containing 1000G and 1,285 Samoans phased the
314 reference panel haplotypes altogether, rather than in a population-aware manner. Further
315 investigation is needed to examine if stratifying by super population before phasing would yield
316 a difference in imputed genotypes than phasing altogether, especially when working with
317 population isolates like Samoans. This would be particularly important if haplotypes harboring
318 population-specific variants appeared at very different frequencies in the less-represented
319 population in the reference panel.

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

331

332 Acknowledgements

333 We would like to thank the Samoan participants of the study, local village authorities, and the
334 many Samoan and other field workers over the years. We acknowledge the Samoan Ministry of
335 Health and the Samoa Bureau of Statistics for their support of this research. We give particular

336 thanks to two research assistants, Melania Selu and Vaimoana Lupematasila, who contributed
337 to the 2010 recruitment and continue to assist us in our work in Samoa.

338 Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was supported by
339 the National Heart, Lung, and Blood Institute (NHLBI). Genome Sequencing for NHLBI
340 TOPMed: Samoan (phs000972.v5.p1) was performed at New York Genome Center Genomics
341 (HHSN268201500016C) and Northwest Genomics Center (HHSN268201100037C). Core
342 support including centralized genomic read mapping and genotype calling, along with variant
343 quality metrics and filtering were provided by the TOPMed Informatics Research Center
344 (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype
345 harmonization, data management, sample-identity QC, and general program coordination were
346 provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract
347 HHSN268201800001I). We gratefully acknowledge the studies and participants who provided
348 biological samples and data for TOPMed.

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

353

354 Author Contributions

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

357 MK – Methodology, Software, Writing – review & editing

358 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.

379 ## Data and code availability

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

384

385 ## References

386 1. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P.,
387 Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global
388 reference for human genetic variation. Nature *526*, 68–74. 10.1038/nature15393.
389 2. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang,
390 H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976
391 haplotypes for genotype imputation. Nature Genetics *48*, 1279–1283. 10.1038/ng.3643.
392 3. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G.,
393 Corvelo, A., Gogarten, S.M., Kang, H.M., et al. (2021). Sequencing of 53,831 diverse
394 genomes from the NHLBI TOPMed Program. Nature *590*, 290–299. 10.1038/s41586-021-
395 03205-y.
396 4. Kowalski, M.H., Qian, H., Hou, Z., Rosen, J.D., Tapia, A.L., Shan, Y., Jain, D., Argos, M.,
397 Arnett, D.K., Avery, C., et al. (2019). Use of >100,000 NHLBI Trans-Omics for Precision
398 Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and
399 detection of rare variant associations in admixed African and Hispanic/Latino populations.
400 PLoS Genet *15*, e1008500. 10.1371/journal.pgen.1008500.

401 5. Friedlaender, J.S., Friedlaender, F.R., Reed, F.A., Kidd, K.K., Kidd, J.R., Chambers, G.K.,
402 Lea, R.A., Loo, J.-H., Koki, G., Hodgson, J.A., et al. (2008). The Genetic Structure of Pacific
403 Islanders. PLOS Genetics *4*, e19. 10.1371/journal.pgen.0040019.
404 6. Skoglund, P., Posth, C., Sirak, K., Spriggs, M., Valentin, F., Bedford, S., Clark, G.R.,
405 Reepmeyer, C., Petchey, F., Fernandes, D., et al. (2016). Genomic insights into the
406 peopling of the Southwest Pacific. Nature *538*, 510–513. 10.1038/nature19844.
407 7. Tsai, H.-J., Sun, G., Smelser, D., Viali, S., Tufa, J., Jin, L., Weeks, D.E., McGarvey, S.T.,
408 and Deka, R. (2004). Distribution of genome-wide linkage disequilibrium based on
409 microsatellite loci in the Samoan population. Human Genomics *1*, 327. 10.1186/1479-7364-
410 1-5-327.
411 8. Harris, D.N., Kessler, M.D., Shetty, A.C., Weeks, D.E., Minster, R.L., Browning, S.,
412 Cochrane, E.E., Deka, R., Hawley, N.L., Reupena, M.S., et al. (2020). Evolutionary history
413 of modern Samoans. Proceedings of the National Academy of Sciences *117*, 201913157.
414 10.1073/pnas.1913157117.
415 9. World Health Organization. Regional Office for the Western Pacific (2021). Progress on the
416 prevention and control of noncommunicable diseases in the Western Pacific Region:
417 country capacity survey 2019 (WHO Regional Office for the Western Pacific).
418 10. Graham, G. (2015). Disparities in Cardiovascular Disease Risk in the United States. Current
419 Cardiology Reviews *11*, 238–245.
420 11. Anand, S., Bradshaw, C., and Prabhakaran, D. (2020). Prevention and management of
421 CVD in LMICs: why do ethnicity, culture, and context matter? BMC Med *18*, 7.
422 10.1186/s12916-019-1480-9.
423 12. Hawley, N.L., Minster, R.L., Weeks, D.E., Viali, S., Reupena, M.S., Sun, G., Cheng, H.,
424 Deka, R., and Mcgarvey, S.T. (2014). Prevalence of adiposity and associated
425 cardiometabolic risk factors in the samoan genome-wide association study. American
426 Journal of Human Biology *26*, 491–501. 10.1002/ajhb.22553.

427 13. McGarvey, S.T. (2001). Cardiovascular disease (CVD) risk factors in Samoa and American
428 Samoa, 1990-95. Pacific health dialog *8*, 157–162.
429 14. Klag, M.J., Whelton, P.K., Randall, B.L., Neaton, J.D., Brancati, F.L., and Stamler, J. (1997).
430 End-stage Renal Disease in African-American and White Men: 16-Year MRFIT Findings.
431 JAMA *277*, 1293–1298. 10.1001/jama.1997.03540400043029.
432 15. SIGMA Type 2 Diabetes Consortium, Williams, A.L., Jacobs, S.B., Moreno-Macías, H.,
433 Huerta-Chagoya, A., Churchhouse, C., Márquez-Luna, C., García-Ortíz, H., Gómez-
434 Vázquez, M.J., Burtt, N.P., et al. (2014). Sequence variants in SLC16A11 are a common
435 risk factor for type 2 diabetes in Mexico. Nature *506*, 97–101. 10.1038/nature12828.
436 16. Nair, A.K., and Baier, L.J. (2015). Complex Genetics of Type 2 Diabetes and Effect Size:
437 What have We Learned from Isolated Populations? Rev Diabet Stud *12*, 299–319.
438 10.1900/RDS.2015.12.299.
439 17. Staimez, L.R., Deepa, M., Ali, M.K., Mohan, V., Hanson, R.L., and Narayan, K.M.V. (2019).
440 Tale of two Indians: Heterogeneity in type 2 diabetes pathophysiology. Diabetes/Metabolism
441 Research and Reviews *35*, e3192. 10.1002/dmrr.3192.
442 18. Vartiainen, E., Dianjun, D., Marks, J.S., Korhonen, H., Guanyi, G., Ze-Yu, G., Koplan, J.P.,
443 Pietinen, P., Guang-Lin, W., Williamson, D., et al. (1991). Mortality, cardiovascular risk
444 factors, and diet in China, Finland, and the United States. Public Health Rep *106*, 41–46.
445 19. Gupta, A.K., Poulter, N.R., Dobson, J., Eldridge, S., Cappuccio, F.P., Caulfield, M., Collier,
446 D., Cruickshank, J.K., Sever, P.S., Feder, G., et al. (2010). Ethnic Differences in Blood
447 Pressure Response to First and Second-Line Antihypertensive Therapies in Patients
448 Randomized in the ASCOT Trial. American Journal of Hypertension *23*, 1023–1030.
449 10.1038/ajh.2010.105.
450 20. Ojji, D.B., Mayosi, B., Francis, V., Badri, M., Cornelius, V., Smythe, W., Kramer, N., Barasa,
451 F., Damasceno, A., Dzudie, A., et al. (2019). Comparison of Dual Therapies for Lowering

452 Blood Pressure in Black Africans. New England Journal of Medicine *380*, 2429–2439.
453 10.1056/NEJMoa1901113.
454 21. James, P.A., Oparil, S., Carter, B.L., Cushman, W.C., Dennison-Himmelfarb, C., Handler,
455 J., Lackland, D.T., LeFevre, M.L., MacKenzie, T.D., Ogedegbe, O., et al. (2014). 2014
456 Evidence-Based Guideline for the Management of High Blood Pressure in Adults: Report
457 From the Panel Members Appointed to the Eighth Joint National Committee (JNC 8). JAMA
458 *311*, 507–520. 10.1001/jama.2013.284427.
459 22. Mills, M.C., and Rahal, C. (2020). The GWAS Diversity Monitor tracks diversity by disease
460 in real time. Nat Genet *52*, 242–243. 10.1038/s41588-020-0580-y.
461 23. Precision medicine needs an equity agenda (2021). Nat Med *27*, 737–737. 10.1038/s41591-
462 021-01373-y.
463 24. Burke, W. (2021). Utility and Diversity: Challenges for Genomic Medicine. Annual Review of
464 Genomics and Human Genetics *22*, 1–24. 10.1146/annurev-genom-120220-082640.
465 25. McGarvey, S.T., Levinson, P.D., Bausser-Man, L., Galanis, D.J., and Hornick, C.A. (1993).
466 Population change in adult obesity and blood lipids in American Samoa from 1976–1978 to
467 1990. American Journal of Human Biology *5*, 17–30. 10.1002/ajhb.1310050106.
468 26. Chin-Hong, P.V., and McGarvey, S.T. (1996). Lifestyle incongruity and adult blood pressure
469 in Western Samoa. Psychosomatic medicine *58*, 131–137.
470 27. Galanis, D.J., McGarvey, S.T., Quested, C., Sio, B., and Afele-Fa'amuli, S.A. (1999).
471 Dietary intake of modernizing Samoans: implications for risk of cardiovascular disease.
472 Journal of the American Dietetic Association *99*, 184–190.
473 28. Dai, F., Keighley, E.D., Sun, G., Indugula, S.R., Roberts, S.T., Åberg, K., Smelser, D.,
474 Tuitele, J., Jin, L., Deka, R., et al. (2007). Genome-wide scan for adiposity-related
475 phenotypes in adults from American Samoa. International Journal of Obesity *31*, 1832–
476 1842. 10.1038/sj.ijo.0803675.

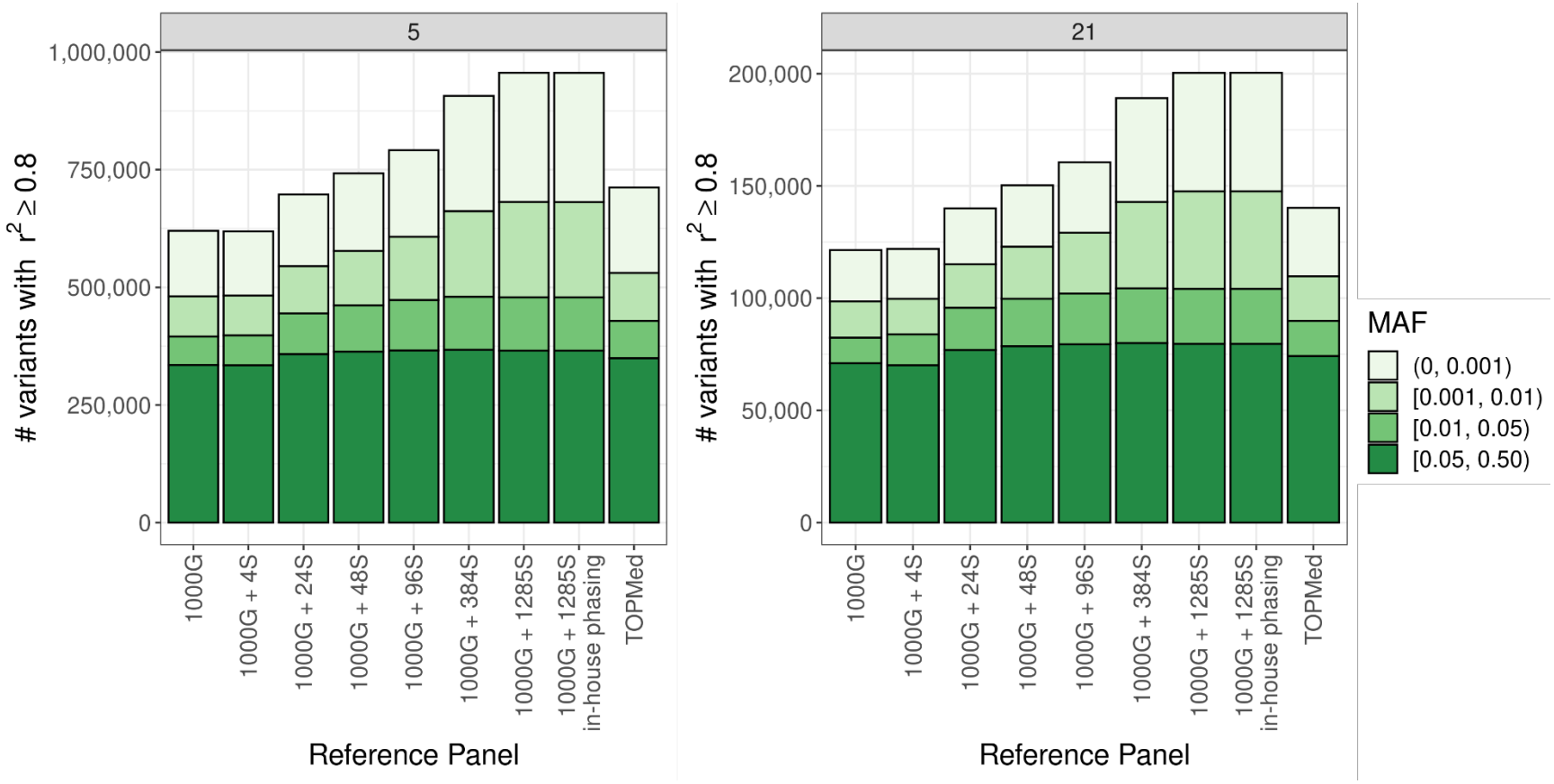
477 29. Dai, F., Sun, G., Åberg, K., Keighley, E.D., Indugula, S.R., Roberts, S.T., Smelser, D., Viali,
478 S., Jin, L., Deka, R., et al. (2008). A Whole Genome Linkage Scan Identifies Multiple
479 Chromosomal Regions Influencing Adiposity-Related Traits among Samoans. Annals of
480 Human Genetics *72*, 780–792. 10.1111/j.1469-1809.2008.00462.x.
481 30. Minster, R.L., Hawley, N.L., Su, C.-T., Sun, G., Kershaw, E.E., Cheng, H., Buhule, O.D., Lin,
482 J., Reupena, M.S., Viali, S., et al. (2016). A thrifty variant in CREBRF strongly influences
483 body mass index in Samoans. Nature Genetics *48*, 1049–1054. 10.1038/ng.3620.
484 31. Carlson, J.C., Krishnan, M., Rosenthal, S.L., Russell, E.M., Zhang, J.Z., Hawley, N.L.,
485 Moors, J., Cheng, H., Dalbeth, N., de Zoysa, J.R., et al. (2023). A stop-gain variant in
486 *BTNL9* is associated with atherogenic lipid profiles. HGG Adv *4*, 100155.
487 10.1016/j.xhgg.2022.100155.
488 32. Laurie, C.C., Doheny, K.F., Mirel, D.B., Pugh, E.W., Bierut, L.J., Bhangale, T., Boehm, F.,
489 Caporaso, N.E., Cornelis, M.C., Edenberg, H.J., et al. (2010). Quality control and quality
490 assurance in genotypic data for genome-wide association studies. Genetic Epidemiology
491 *34*, 591–602. 10.1002/gepi.20516.
492 33. Gusev, A., Shah, M.J., Kenny, E.E., Ramachandran, A., Lowe, J.K., Salit, J., Lee, C.C.,
493 Levandowsky, E.C., Weaver, T.N., Doan, Q.C., et al. (2012). Low-Pass Genome-Wide
494 Sequencing and Variant Inference Using Identity-by-Descent in an Isolated Human
495 Population. Genetics *190*, 679–689. 10.1534/genetics.111.134874.
496 34. Loh, P.-R., Danecek, P., Palamara, P.F., Fuchsberger, C., A Reshef, Y., K Finucane, H.,
497 Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R., et al. (2016). Reference-based
498 phasing using the Haplotype Reference Consortium panel. Nat Genet *48*, 1443–1448.
499 10.1038/ng.3679.
500 35. Deelen, P., Bonder, M.J., Van Der Velde, K.J., Westra, H.J., Winder, E., Hendriksen, D.,
501 Franke, L., and Swertz, M.A. (2014). Genotype harmonizer: Automatic strand alignment and

502 format conversion for genotype data integration. BMC Research Notes *7*, 901.
503 10.1186/1756-0500-7-901.
504 36. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew,
505 E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and
506 methods. Nature Genetics *48*, 1284–1287. 10.1038/ng.3656.
507 37. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker,
508 R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and
509 VCFtools. Bioinformatics *27*, 2156–2158. 10.1093/bioinformatics/btr330.
510 38. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: A Tool for Genome-
511 wide Complex Trait Analysis. Am J Hum Genet *88*, 76–82. 10.1016/j.ajhg.2010.11.011.
512 39. The NHLBI Trans-Omics for Precision Medicine (TOPMed) Whole Genome Sequencing
513 Program. (2018). BRAVO variant browser: University of Michigan and NHLBI.
514

515 Figure Titles and Legends

516 **Figure 1. Count of well-imputed variants by reference panel.** The number of variants with imputation quality $r^2 \geq 0.80$ by reference

517 panel is plotted across minor allele frequency (MAF) bins for chromosomes 5 and 21. (These counts are printed in Table S1.)

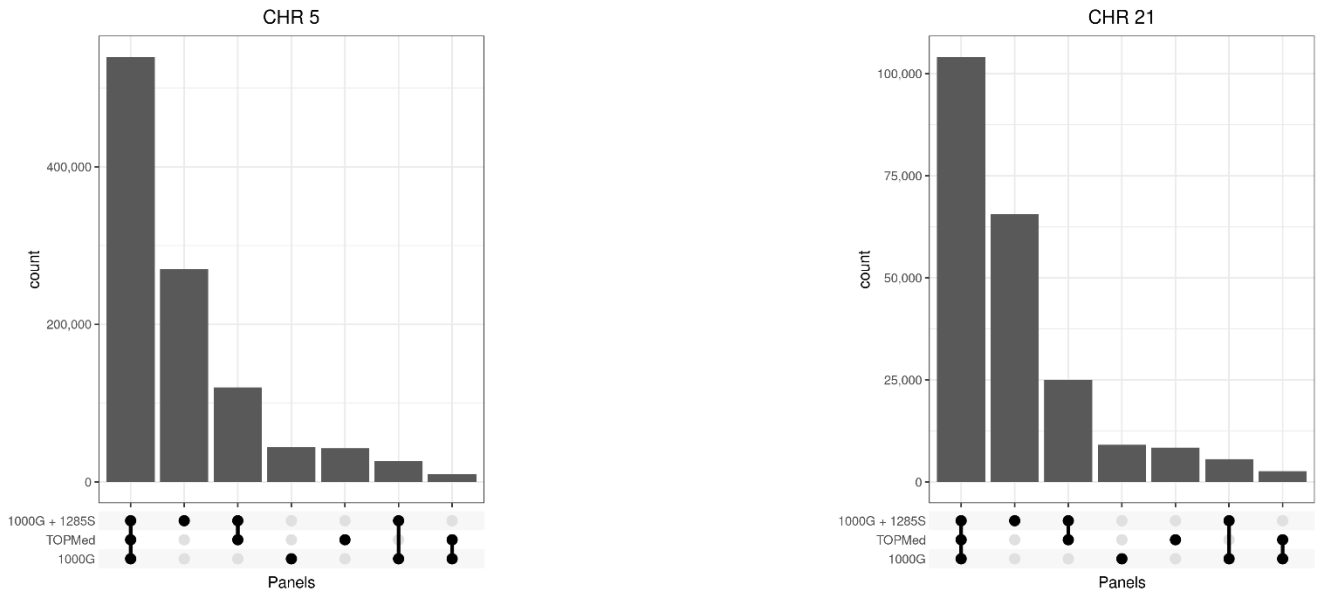


518

519

520 **Figure 2. Count of uniquely imputed variants by reference panel.** The number of well imputed variants ($r^2 \geq 0.80$) unique to the

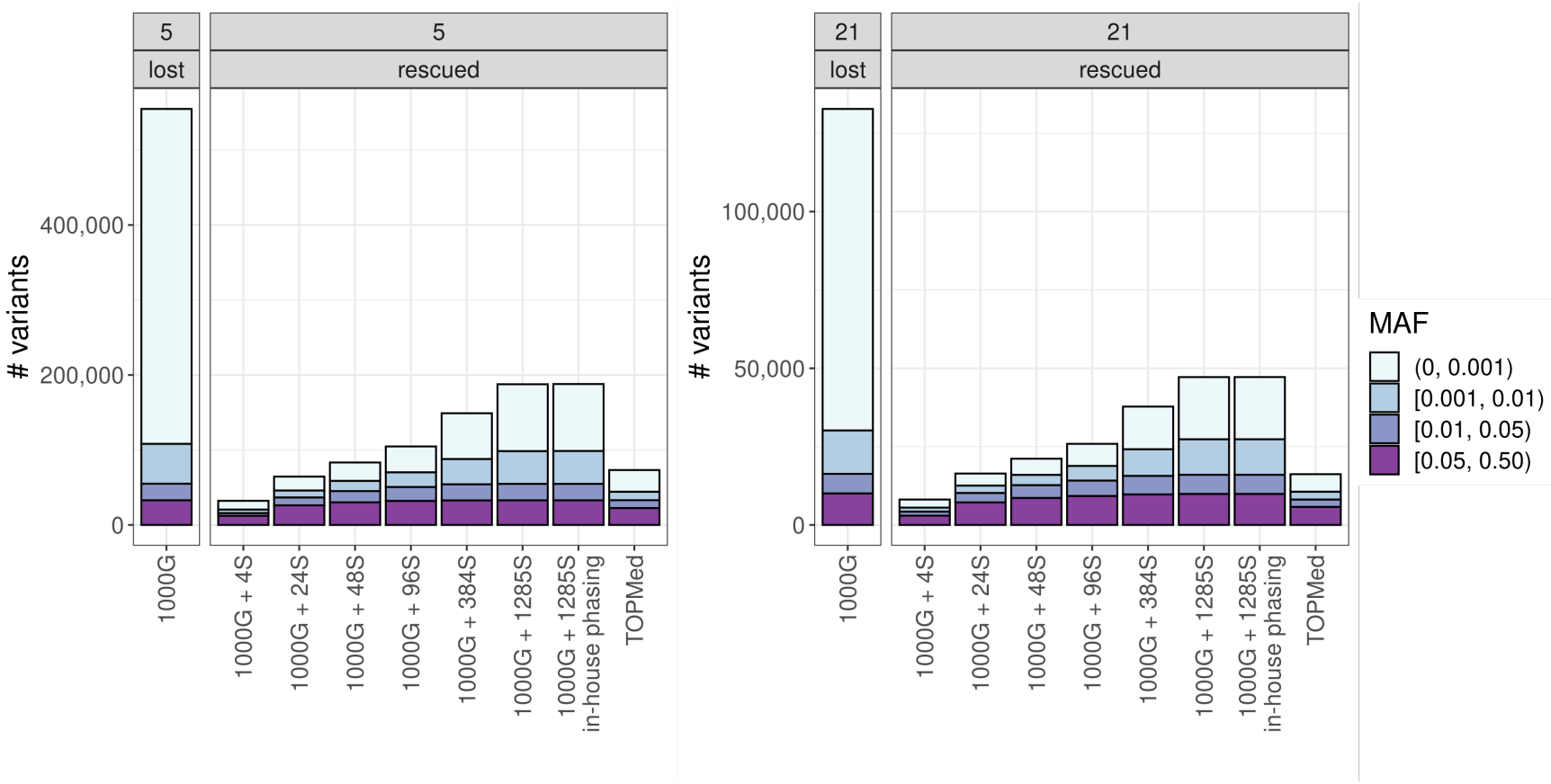
521 1000G + 1285S, TOPMed, and 1000G reference panel (and all possible combinations) is plotted for chromosomes 5 and 21.



522

523

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.

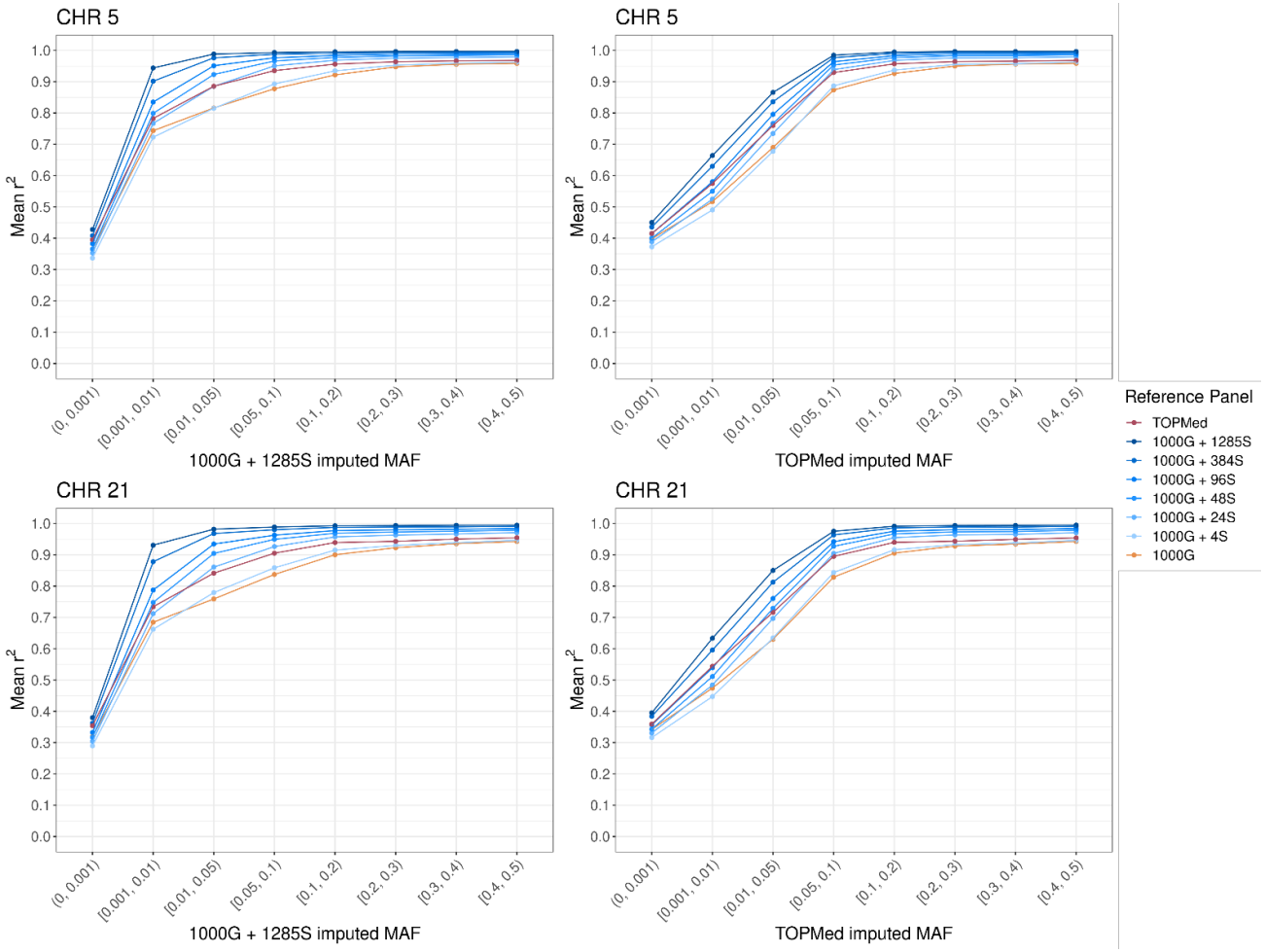


527 **Figure 4. Imputation quality versus minor allele frequency by reference panel.** Mean imputation quality (r^2) by minor allele

528 frequency (MAF) bin (left column: MAF based on 1000G + 1285S imputation; right column: MAF based on the TOPMed imputation)

529 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

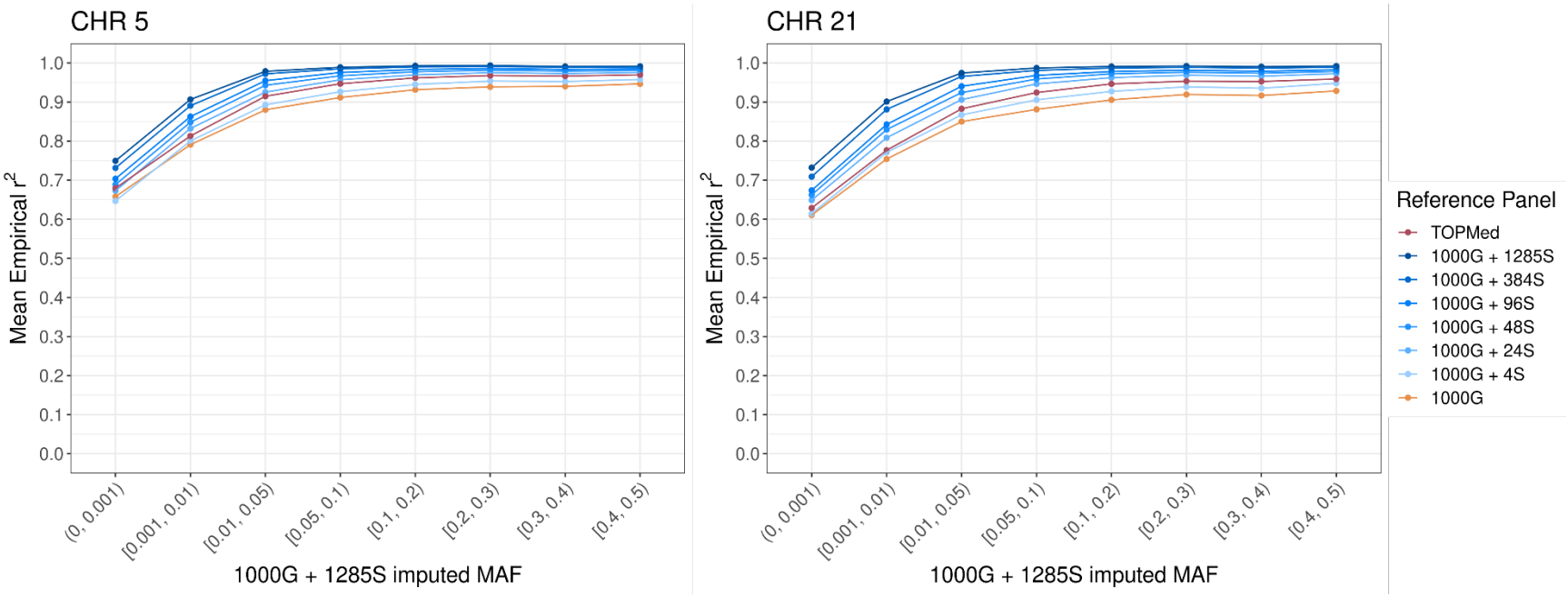
530 reference panel are in Table S3.)


CHR 5
CHR 5
CHR 21
CHR 21
Mean r2
1000G + 1285S imputed MAF
TOPMed imputed MAF
(0, 0.001)
[0.001, 0.01)
[0.01, 0.05)
[0.05, 0.1)
[0.1, 0.2)
[0.2, 0.3)
[0.3, 0.4)
[0.4, 0.5)
Reference Panel
TOPMed
1000G + 1285S
1000G + 384S
1000G + 96S
1000G + 48S
1000G + 24S
1000G + 4S
1000G

533 **Figure 5. Empirical r^2 versus minor allele frequency by reference panel for genotyped variants.** Mean empirical r^2 by minor

534 allele frequency (MAF) bin (based on 1000G + 1285S imputation) is plotted by reference panel on chromosomes 5 and 21. (Plots

535 with median empirical r^2 are given in Figure S4.)



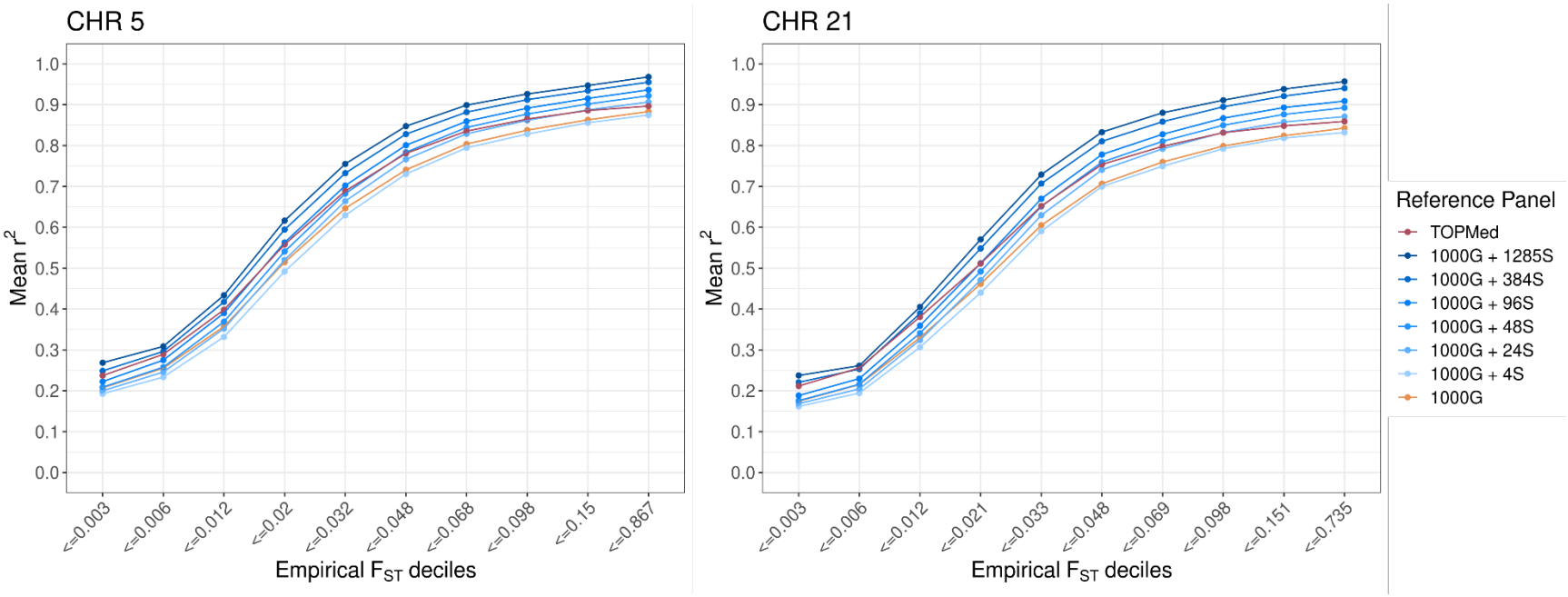
536

537

538 **Figure 6. Imputation quality versus fixation index by reference panel.** Mean imputation quality (r^2) by empirical Weir and

539 Cockerham Fixation Index (F_{ST}) deciles is plotted by reference panel on chromosomes 5 and 21. (Plots with median r^2 are given in

540 Figure S5.)



541

542

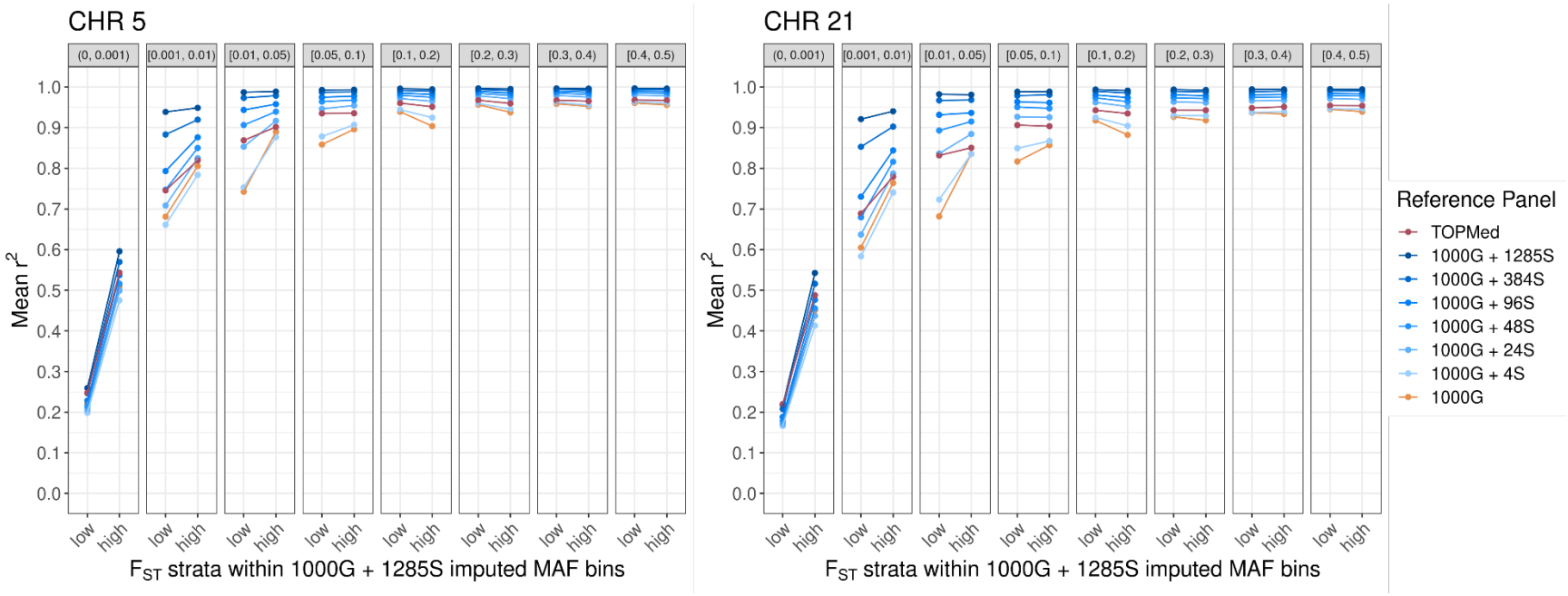
543 **Figure 7 Imputation quality versus fixation index by reference panel and MAF bin.** Mean imputation quality (r^2) by median-

544 stratified Weir and Cockerham Fixation Index (F_{ST}) across MAF (based on 1000G + 1285S imputation) is plotted by reference panel

545 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

546 S5). (Plots with median r^2 are given in Figure S6.)

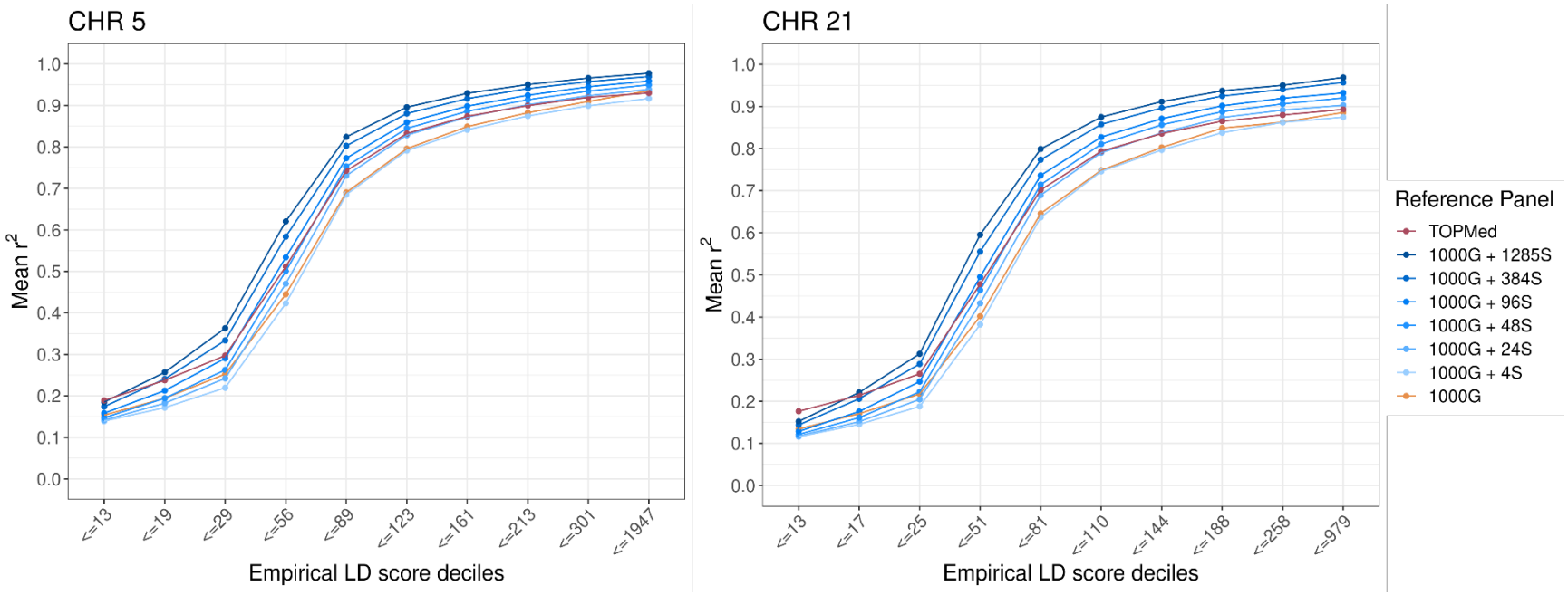
547



548

549 **Figure 8 Imputation quality versus LD score by reference panel.** Mean imputation quality (r^2) by empirical LD score deciles is

550 plotted by reference panel on chromosomes 5 and 21. (Plots with median r^2 are given in Figure S7.)



551

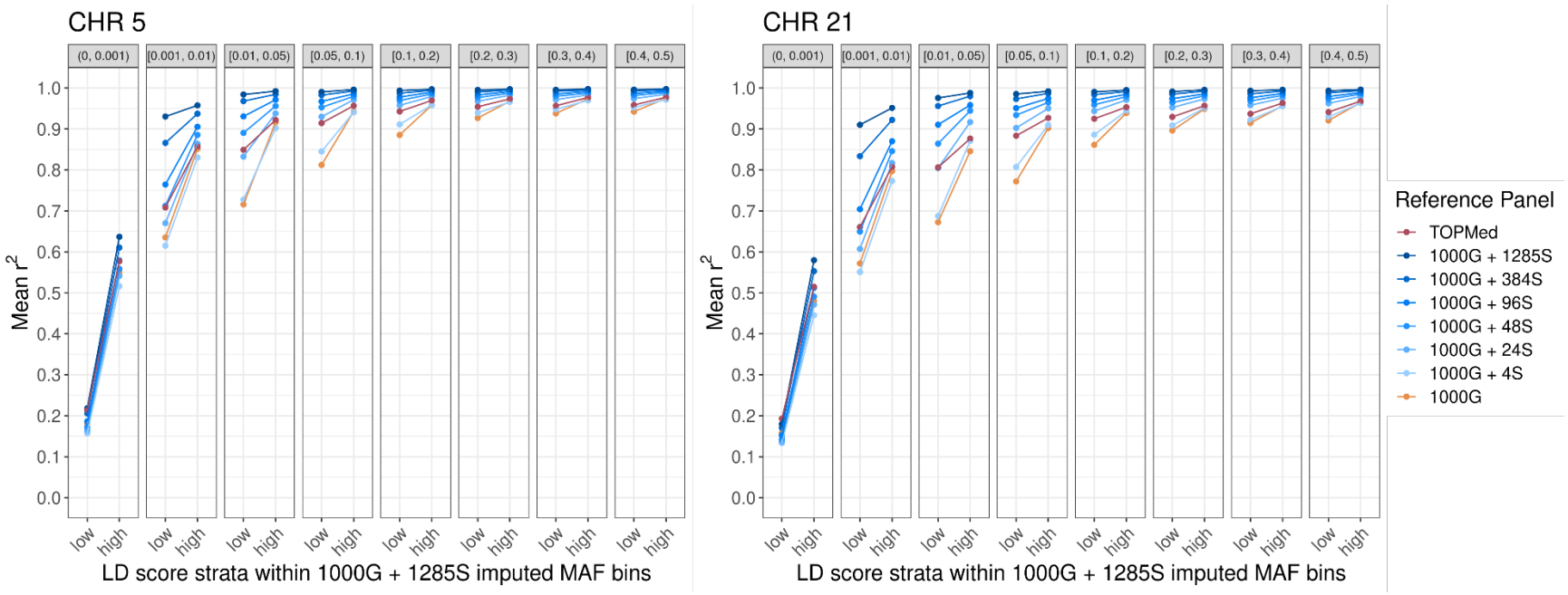
552

553 **Figure 9 Imputation quality versus LD score by reference panel and MAF bin.** Mean imputation quality (r^2) by median-stratified

554 LD score across MAF (based on 1000G + 1285S imputation) is plotted by reference panel on chromosomes 5 and 21. LD score was

555 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

556 given in Figure S8.)



557

558

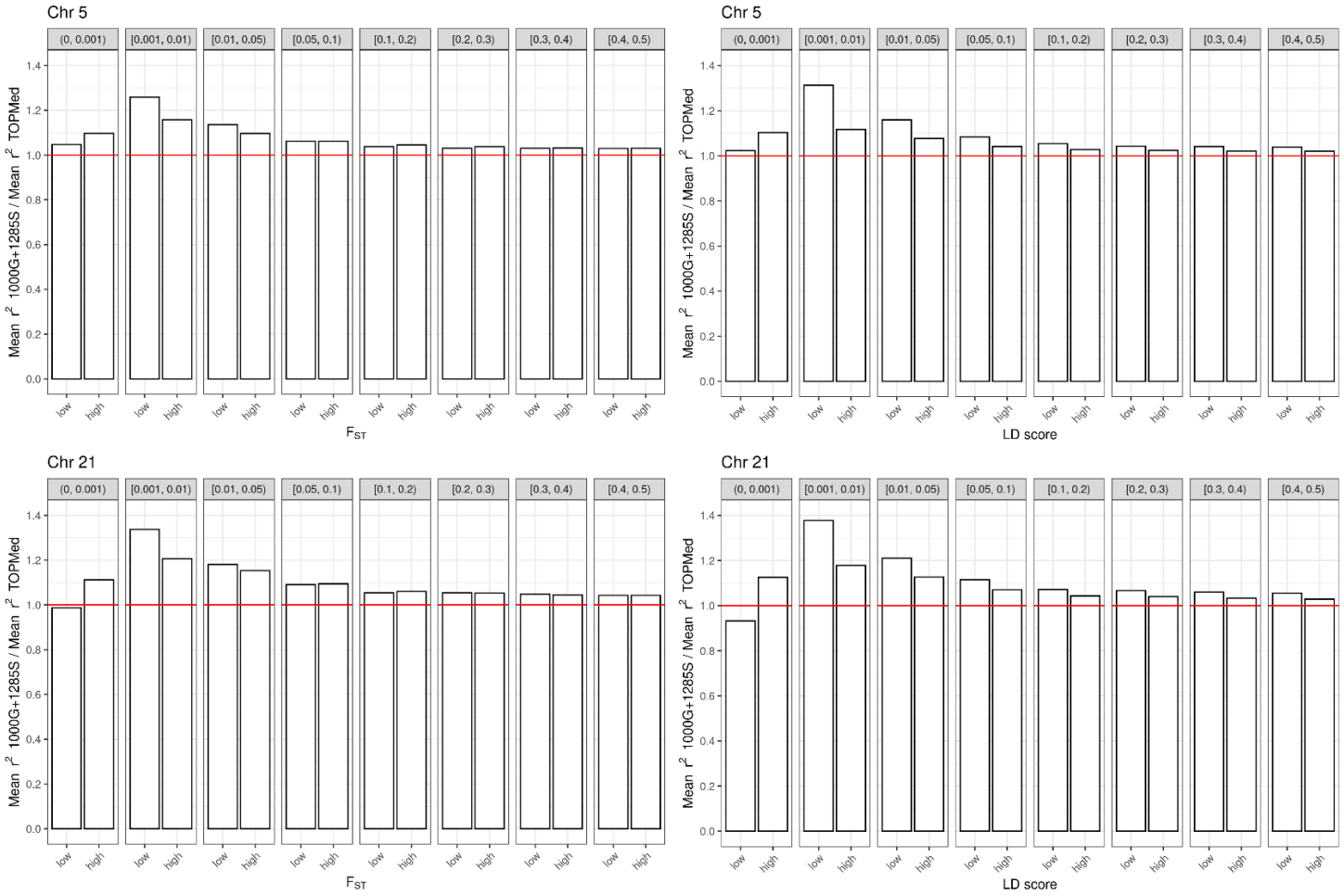
559 **Figure 10. Gains in imputation quality from TOPMed panel to 1000G + 1285S panel by MAF for F_{ST} and LD score.** Fold change

560 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

561 1000G + 1285S imputation) is plotted for chromosomes 5 and 21. Both F_{ST} and LD score were stratified into 'low' and 'high' about the

562 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

563 between TOPMed and 1000G + 1285S panels.



564

565 **Figure 11. Differences in gains in imputation quality from TOPMed panel to 1000G +**

566 **1285S panel by MAF for F_{ST} and LD score.** Both F_{ST} and LD score were stratified into 'low'

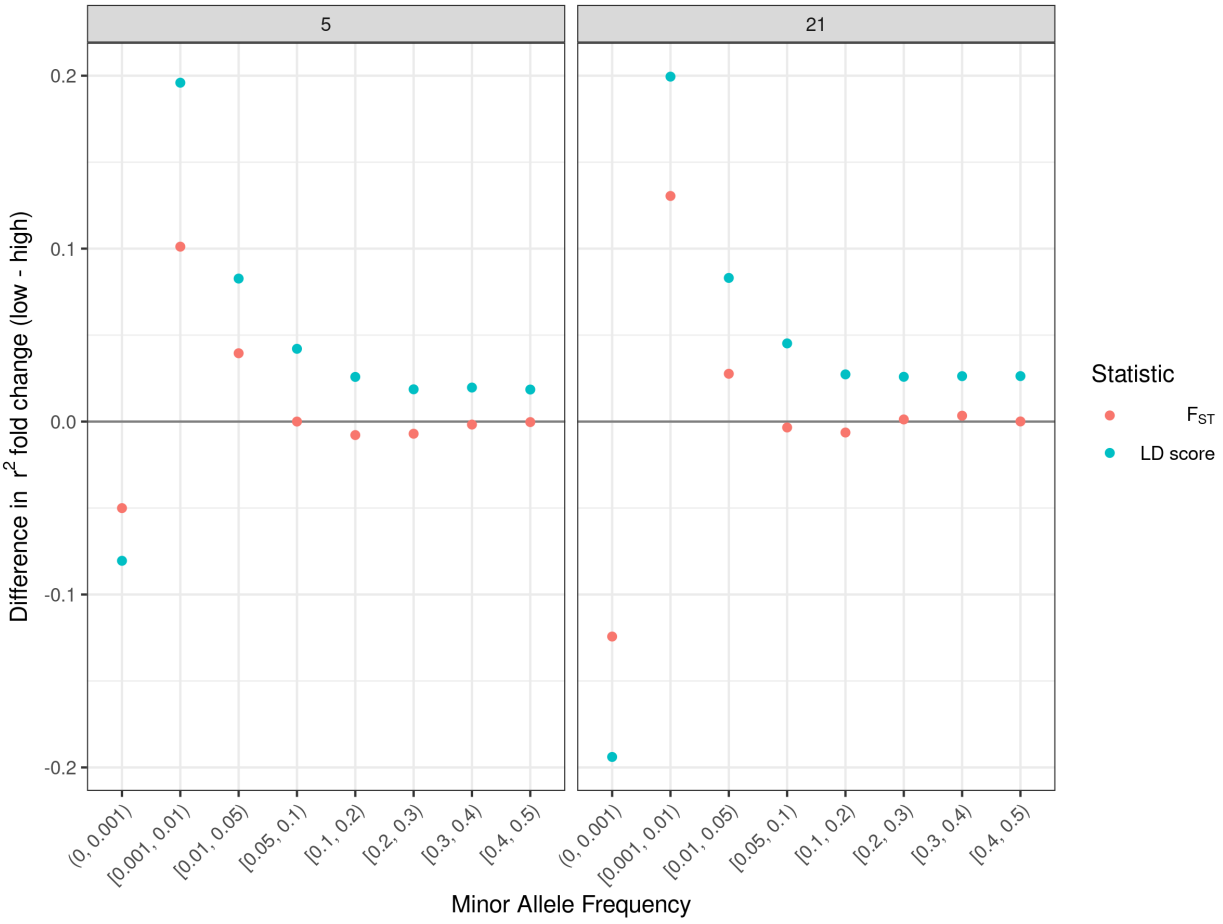
567 and 'high' about the median F_{ST}/LD score within each MAF bin (values given in Table S5). The

568 difference in fold change between low and high scoring variants (low – high) from (Figure 6) is

569 plotted across MAF separately for F_{ST} and LD score for chromosomes 5 and 21. Reference line

570 at Y=0 (gray) is given to indicate no difference between high and low F_{ST}/LD Score

571



572

573

574 # Tables

575 **Table 1. Names and descriptions of tested imputation protocols**

Protocol Number	Reference Panel Name	Server	Reference Panel Samples	Phasing of reference panel performed with…
1	TOPMed	TOPMed Imputation Server	TOPMed	Reference panel samples only
2	1000G	Michigan Imputation Server	1000G Phase III (GRCh38/hg38)	Reference panel samples only
3	1000G + 4S	In-house	1000G + 4 Samoans	All TOPMed freeze 9 samples
4	1000G + 24S	In-house	1000G + 24 Samoans	All TOPMed freeze 9 samples
5	1000G + 48S	In-house	1000G + 48 Samoans	All TOPMed freeze 9 samples
6	1000G + 96S	In-house	1000G + 96 Samoans	All TOPMed freeze 9 samples
7	1000G + 384S	In-house	1000G + 384 Samoans	All TOPMed freeze 9 samples
8	1000G + 1285S	In-house	1000G + 1285 Samoans	All TOPMed freeze 9 samples
9	1000G + 1285S in-house phasing	In-house	1000G + 1285 Samoans	Reference panel (1000G + 1285S) samples only

576

577 **Table 2. Imputation quality and minor allele frequency (MAF) for rs200884524**

578 **(chr5:181050285 C/T [hg38]) in *BTNL9* across reference panels.** Targeted genotyping in two

579 separate cohorts of 1,666 Samoan and American Samoan adults yielded a MAF of 0.202-

580 0.233[31].

Reference panel	r^2	MAF
1000G	0.17098	0.00076
1000G + 4S	0.90687	0.20424
1000G + 24S	0.91276	0.22198
1000G + 48S	0.86967	0.20785
1000G + 96S	0.91062	0.2071
1000G + 384S	0.94567	0.2213
1000G + 1285S	0.94360	0.21852
1000G + 1285S in-house phasing	0.94360	0.21852
TOPMed	0.79515	0.16620

581