It is made available under a CC-BY-NC-ND 4.0 International license .

1	Evaluation of imputation performance of multiple reference panels in a Pakistani population
2	
3	Jiayi Xu ^{1*} , Dongjing Liu ² , Arsalan Hassan ^{3,4} , Giulio Genovese ^{5,6,7} , Alanna C. Cote ² , Brian
4	Fennessy ² , Esther Cheng ² , Alexander W. Charney ² , James A. Knowles ⁸ , Muhammad Ayub ⁹ ,
5	Roseann E. Peterson ¹⁰ , Tim B. Bigdeli ¹⁰ , Laura M. Huckins ^{1*}
6	
7	1. Department of Psychiatry, Yale School of Medicine, New Haven, CT, USA
8	2. Icahn School of Medicine at Mount Sinai, New York, NY, USA
9	3. University of Peshawar, Peshawar, Khyber Pakhtunkhwa, Pakistan
10	4. Institute of Omics and Health Research, Lahore, Pakistan
11	5. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard,
12	Cambridge, MA, USA
13	6. Stanley Center, Broad Institute of MIT and Harvard, Cambridge, MA, USA
14	7. Department of Genetics, Harvard Medical School, Boston, MA, USA
15	8. The Human Genetics Institute of New Jersey, Rutgers University, Piscataway, NJ, USA
16	9. University College London, London, UK
17	10. Department of Psychiatry and Behavioral Sciences, Institute for Genomics in Health,
18	State University of New York Downstate Health Sciences University, Brooklyn, NY, USA
19	
20	*Corresponding authors
21	Correspondence should be addressed to Laura M. Huckins, PhD: laura.huckins@yale.edu;
22	Jiayi Xu, PhD: <u>jiayi.xu@yale.edu</u> .
23	
24	

It is made available under a CC-BY-NC-ND 4.0 International license .

25 Abstract

26	Genotype imputation is crucial for GWAS, but reference panels and existing benchmarking
27	studies prioritize European individuals. Consequently, it is unclear which publicly available
28	reference panel should be used for Pakistani individuals, and whether ancestry composition or
29	sample size of the panel matters more for imputation accuracy. Our study compared different
30	reference panels to impute genotype data in 1814 Pakistani individuals, finding the best
31	performance balancing accuracy and coverage with meta-imputation with TOPMed and the
32	expanded 1000 Genomes (ex1KG) reference. Imputation accuracy of ex1KG outperformed
33	TOPMed despite its 30-fold smaller sample size, supporting efforts to create future panels with
34	diverse populations.
35	
36	Keywords:
37	Genetics, Genome-Wide Association Studies, Imputation, Imputation Panels, South Asian
38	Ancestry, Pakistan.
39	
40	
41	
42	
43	
44	
45	
46	
47	
48	
49	
50	

It is made available under a CC-BY-NC-ND 4.0 International license .

51 Background

52 Genotype imputation allows easy, accurate, and significant increases in the number of 53 variants available for genome wide association studies (GWAS), facilitating data harmonization 54 and meta-analyses across cohorts and genotyping platforms, as well as statistical fine-mapping 55 (1).

56 Currently, >85% of GWAS comprise individuals of European ancestry (EUR) (2), and 57 available reference panels are predominantly European. Inclusion of non-EUR populations is 58 critical to increase equity in genetic research (3) and advance understanding of genetic 59 architecture of disease. Individuals of South Asian (SAS) ancestry, especially those in Pakistan, 60 are severely under-represented in genetic studies, with few published GWAS (4). While 61 imputation accuracy has been evaluated in an Indian population (5), no study has yet assessed 62 imputation accuracy in Pakistani individuals based on reference panels (i.e., TOPMed (6), high-63 coverage expanded 1000 Genomes (ex1KG) (7), low-coverage 1000 Genomes (1KG) (8), 64 GenomeAsia (9)). These panels include a small proportion of Pakistani samples, suggesting 65 limited utility to studies of Pakistani populations and as-yet-determined imputation accuracy. 66 Although TOPMed is by far the largest imputation panel (N = 97,256), only 0.1% are Pakistani 67 individuals (n=139) (6,10). The ex1KG and GenomeAsia-Pilot panels have 146 and 113 68 Pakistani samples with a total sample size of 3,202 and 1,739, respectively (7,9). 69 Here, we compare accuracy of imputation panels in a cohort of 1814 Pakistani individuals. We include comparison of true $R^2 (R_{True}^2$ through leveraging targeted sequencing 70 71 data) and estimated R^2 (R_{rst}^2). We assess 5 imputation panels, including TOPMed, ex1KG, 72 1000G GRCh38 (1KG38), 1000G GRCh37 SAS (1KG37-SAS), GenomeAsia-Pilot, as well as a 73 meta-imputation approach combining TOPMed and ex1KG (meta). In this study, we for the first 74 time examine which of the publicly available panels yields the best performance for Pakistani

- 75 populations, whether merging existing panels further improves imputation accuracy, and
- 76 whether ancestry or sample size is a more critical determinant of imputation accuracy.

It is made available under a CC-BY-NC-ND 4.0 International license .

77 Results and Discussion

We initially evaluated imputation accuracy among common variants (minor allele 78 79 frequency (MAF) \geq 1% in the Pakistani individuals). To evaluate the imputation accuracy 80 against true genotypes, 1748 Pakistani individuals in this study (96%) also have sequenced 81 genotype data via targeted sequencing. When comparing the imputed genotypes to targeted 82 sequencing data, broadly, there was no difference in imputation accuracy for common variants 83 across the six panels except for MAF bins of (0.01, 0.015] and (0.4, 0.5] (Figure 1a, average R_{True}^2 = 0.61-0.91 across MAF bins of common variants). Genome-wide, including variants for 84 85 which we do not have matched targeted sequencing data, ex1KG and meta-imputation had the highest imputation accuracy (R_{Est}^2) for common variants (mean $R_{Est}^2 = 0.75 - 0.94, 0.74 - 0.93$ 86 87 respectively), while GenomeAsia-Pilot and 1KG37-SAS panels had the lowest imputation 88 accuracy (R_{Est}^2 = 0.62-0.86, 0.64-0.86 respectively, Figures 2a). ex1KG also had the highest imputation accuracy measured via empirical $R^2 (R_{Emp}^2)$ for single nucleotide polymorphisms 89 90 (SNPs) that were both imputed and genotyped on the Illumina Infinium Global Screening Array 91 (GSA) (Fig. S9).

92 Next, we evaluated which imputation panel offered the best coverage for imputation of 93 common variants, defined as percentage of targeted sequencing SNPs that were also imputed 94 by a specific imputation panel. 90.9% - 99.4% of common variants called in our targeted 95 sequencing data were included in at least one imputation panel (Figure 1b), of which 61.3% (GenomeAsia-Pilot) to 83.8% (meta-imputation) were well-imputed ($R_{Est}^2 \ge 0.8$) (Figure 1e). In 96 97 particular, among the 1896 common variants through targeted sequencing, 1589 were well-98 imputed by meta-imputation, followed by ex1KG with 1578 SNPs. On a genome-wide scale, 99 imputation with 1KG38 resulted in the most well-imputed SNPs (8.02M), followed by meta-100 imputation (7.93M) and ex1KG (7.85M); for average R_{Est}^2 , ex1KG is the best, followed by meta-

It is made available under a CC-BY-NC-ND 4.0 International license .

imputation and then 1KG38. Imputation with GenomeAsia-Pilot led to the fewest imputed and
well-imputed SNPs (Figures 1b, 1e, 2b, 2c).

Overall, for common variants, balancing genomic coverage and imputation accuracy, meta-imputation had the best performance, followed by ex1KG which had the highest imputation accuracy but slightly fewer well-imputed variants (7.85M vs 7.93M). However, metaimputation is computationally intensive and expensive in terms of both time and data storage, taking ~70 hours (and 328G disk space) compared to ~2 hours (and 87G) for ex1KG for the 1814 Pakistani individuals in this study (Table S3).

109 We found that imputation quality was low for rarer variants (MAF < 1%) irrespective of 110 the imputation reference panel used. However, for lower frequency variants with MAF between 111 0.5% and 1%, average R_{True}^2 and R_{Est}^2 were acceptable (\geq 0.6) for imputed SNPs using meta-112 imputation and/or ex1KG (Figs. 1a, 2a). When comparing against sequenced data, meta-

113 imputation and ex1KG had the highest mean R_{True}^2 for SNPs with MAF 0.1% - 1% ($p_{meta vs ex1KG} \ge$

114 0.35, Figure 1a). On a genome-wide level, meta-imputation had the highest mean R_{Est}^2 for SNPs

115 with MAF 0.05% -1%. For the lowest MAF category (MAF \leq 0.05%), findings were inconsistent:

116 GenomeAsia-Pilot had the highest mean R_{True}^2 , while 1KG38 had the highest mean R_{Est}^2 , but

both had poor imputation quality on an absolute scale (0.38 for GenomeAsia-Pilot R_{True}^2 , 0.092

118 for 1KG38 R_{Est}^2). It is likely that the mean R_{True}^2 and R_{Est}^2 for meta-imputation was diluted at the

119 ultra-rare end due to its substantial number of poorly imputed rare variants ($R_{Est}^2 < 0.8$)

120 compared to other imputation panels (Fig. S10).

121 Imputation coverage was also low for rare variants. Although meta-imputation and 122 TOPMed imputation yielded 98.4%-99.4% of targeted sequencing rare variants (Figure 1c), this 123 number drastically dropped when including only well-imputed SNPs (2.6%-2.9% of targeted 124 sequencing variants, Figure 1f). Therefore, sequencing remains the best approach to retrieve 125 high-quality genotypes for rare variants. Nonetheless, we still obtained a large absolute number

It is made available under a CC-BY-NC-ND 4.0 International license .

of high-quality genotypes for rare variants through imputation (e.g., 4.8M genome-wide rare
variants via meta-imputation for Pakistani individuals). Across imputation panels, metaimputation had the highest count of both imputed and well-imputed rare variants (286.9M, 4.8M
respectively), followed by TOPMed (282.8M, 4.2M). 1KG37-SAS and GenomeAsia-Pilot had the
fewest well-imputed variants (1.43M and 0.7M; Figs. 2b, 2c).
Overall, for rare variants, meta-imputation yielded the highest imputation quality and

132 generated the highest count of well-imputed SNPs, followed by imputation with TOPMed.

133 However, meta-imputation required ~70 hours of computational time, compared to ~5 hours for

134 TOPMed for the 1814 Pakistani individuals (Table S3). Further, meta-imputation required

135 ~328G of disk space, compared to ~150G using TOPMed. We do not recommend the current

136 GenomeAsia-Pilot and 1KG37-SAS datasets for imputation of rare variants given the relatively

137 low counts of well-imputed rare variants obtained.

138 Next, we tested whether SNPs with greater MAF differences between Pakistani and EUR populations had significant differences in imputation accuracy (measured by R_{True}^2), since most 139 140 reference panels included in this study are primarily of European origin (with the exception of 141 GenomeAsia-Pilot). In fact, deviation from EUR MAF did not have a negative impact on SNP 142 imputation quality (Figs. S11-13), across both common and rare variants. Since all the 143 imputation panels included in this study have 0.1% - 6.5% sequenced individuals from Pakistan 144 (n ranges from 113 to 146 for Pakistani individuals, and n ranges from 489 to 724 for SAS 145 individuals in general; Table S1), this suggests that imputation guality is unlikely to impaired by 146 a large proportion of EUR ancestry in reference panels as long as there is some inclusion of 147 individuals from the targeted population (11).

For comparison of coverage across different SNP categories, we mainly focused on wellimputed SNP counts since these are the ones typically included in GWAS (Figs. 2d-2i, S14-S15). For common variants, meta-imputation imputed the most SNVs (7.4M) whereas 1KG38 imputed the most indels (780.8K) (Fig. 2d). Meta-imputation also produced the highest SNP

It is made available under a CC-BY-NC-ND 4.0 International license .

152 count in most of the SNP categories for common variants, including SNVs (7.4M, Fig. 2d), 153 missense variants, non-coding transcript exonic variants, variants in splice region, and start/stop 154 variants (Fig. 2g-2i), whereas 1KG38 produced the highest SNP count for indels (780.8K; Fig. 155 2d), intronic variants, and upstream and downstream gene variants (Fig. 2f). For rare variants, meta-imputation produced the highest SNP counts across all categories, followed by TOPMed. 156 157 Regardless of common or rare variants, imputation using GenomeAsia-Pilot produced the least 158 number of well-imputed SNPs across different SNP types, with no imputed indel or multi-allelic 159 variants. Notably, although far fewer high-guality rare variants are available through imputation 160 panels compared to sequencing (Fig. 1f), on a genome-wide scale, current imputation panels 161 still provide a substantial number of well-imputed rare variants across different SNP categories 162 (Fig. 2d-2i) for further investigation. For example, meta-imputation generated 4.75 M well-163 imputed rare variants, including 29933 well-imputed rare missense variants (compared to 24616 164 well-imputed common missense variants). Given high rate of consanguineous marriage 165 increases the probability of homozygous rare variants, genetic studies of rare variants in the 166 Pakistani population as well as other South Asian populations could provide valuable insights 167 into genetic etiology of various diseases for future therapeutic development (5). 168 Together, our results imply that meta-imputation is currently the best choice for imputation of 169 both common and rare SNPs in Pakistani individuals, taking into account both imputation 170 accuracy and variant coverage. Other panels may also be appropriate; for example, ex1KG has 171 high imputation accuracy for common variants, and lower computational demands, while 172 TOPMed has high coverage of well-imputed rare variants. To address the question of whether 173 sample size or ancestry matching matters more for imputation in the Pakistani population, our 174 findings suggest that ancestry matters more for genetic imputation of common variants, as 175 ex1KG outperformed TOPMed for the number of well-imputed common SNPs (7.85M vs 7.37M, 176 Figs. 1e, 2c) with a smaller total sample size (N_{ex1KG} = 3202 vs N_{TOPMed} = 97256) but a larger 177 Pakistani-specific sample size (N_{ex1KG}= 146 vs N_{TOPMed} = 139). For rare variants, while whole

It is made available under a CC-BY-NC-ND 4.0 International license .

178 genome sequencing remains the gold standard to obtain a large number of rare variants, the 179 absolute number of well-imputed SNPs genome-wide by imputation is still massive (>4M for 180 TOPMed and meta-imputation, Fig. 2c) and more cost-friendly compared to whole genome 181 sequencing (on a relative scale of \$100 vs \$1000 per sample). While sample size is important to 182 increase the likelihood of finding certain rare variants in the population, ancestry is also 183 important for Pakistani-specific rare variants given the high rate of consanguineous marriages 184 (12). As GWAS diversity initiatives gain momentum in the recent years, it is crucial to explore 185 the impact of reference panel ancestry and sample size on imputation. Our study highlighted the 186 significance of these factors in the context of genetic research diversification using the Pakistani 187 population as a case study. 188 Further, although the GenomeAsia-Pilot had lower performance amongst all the imputation 189 panels tested in this study, a future release of its whole reference panel of 100K (9), assuming 190 6.5% of Pakistani individuals, would provide the largest Pakistani-specific imputation panel 191 (estimated N = 6500 versus ex1KG which has 146 Pakistani individuals, Table S1). In addition, 192 with the option of meta-imputation, imputation power could be further improved by meta-193 imputing the future release of GenomeAsia with other large panels (e.g., TOPMed). Another 194 alternative approach to improve imputation accuracy while controlling the sequencing cost in the 195 future is to sequence a subset of population-specific individuals and meta-impute with other

196 large reference panels (13).

197

198 Conclusions

In this study, we evaluated imputation performance of various imputation panels in the Pakistani
population for the first time. Overall, we found meta-imputation to have the highest genomewide imputation accuracy for rare variants whereas ex1KG had the highest genome-wide
imputation accuracy for common variants. Balancing imputation quality with genomic coverage,
our study shows that meta-imputation of TOPMed and ex1KG is the best for imputation of both

It is made available under a CC-BY-NC-ND 4.0 International license .

204 rare and common variants in Pakistani individuals when the computational resources are not 205 limited. Otherwise, we recommend TOPMed for rare variant imputation and ex1KG for common 206 variant imputation. Further, this study suggests that SNPs with MAF deviated from EUR MAF 207 did not have reduced imputation quality in Pakistani individuals. Increased European sample 208 sizes in imputation reference panels seem unlikely to increase imputation accuracy in non-209 Europeans; we found the ex1KG panel, with fewer overall individuals but more Pakistani 210 individuals, outperformed the overall larger TOPMed panel for common variants. Taken together, 211 our study supports the importance of including more diverse populations into the current 212 repertoire of human genome reference panels for future genetic research. 213 214 Methods 215 Study Subjects 216 The Pakistani individuals included in this study were collected as pilot in preparation for the 217 Genetics of Schizophrenia in Pakistan (GEN-SCRIP) Study. The samples were collected by 218 Lahore Institute of Research and Development and University of Peshawar. The diagnoses 219 were made based on clinician's interview according to the Diagnostic and Statistical Manual of 220 Mental Disorders IV. The study was approved by the Institutional Review Board (IRB) 221 committee at University of Peshawar in Pakistan, the IRB of Lahore Institute of Research and 222 Development and IRB of University of Health Sciences Lahore. 223

224 Genotyping and Quality Control

The Infinium[™] Global Screening Array-24 v3.0 BeadChip (Illumina Inc., California, USA) was
used for genotyping of the Pakistani individuals. Genotyping was performed at the Genomic
Core Facility of Icahn School of Medicine at Mount Sinai. The MoChA pipeline was used on
Google Cloud to convert the Illumina genotype intensity files (.idat) to VCF files via Illumina
GenomeStudio (14,15) (Fig. S1, Supplemental methods). Genetic variants with GenTrain score

It is made available under a CC-BY-NC-ND 4.0 International license .

230 < 0.4 and cluster separation score < 0.3 were excluded to remove variants with poor genotype cluster separation (Figs. S2, S3). Quality control (QC) was then performed using PLINK (16). At 231 232 the SNP level, we removed SNPs with low call rate (<0.95), duplicated SNPs with a lower call 233 rate, and those with low minor allele frequency (MAF < 0.001) (Figs. S2, S3). Additionally, given 234 that Pakistan has a tradition of consanguineous marriages which could result in a high level of 235 autozygosity in the population, we also excluded variants that deviate from Hardy-Weinberg Equilibrium (HWE, $p < 1 \times 10^{-6}$) only in individuals with low autozygosity (Fig. S3) (4). Low 236 237 autozygosity is defined as an individual's consanguinity coefficient F_{ROH} less than 0.5% (4) (F_{ROH} 238 = $\sum L_{ROH} / L_{aenome}$, where L_{ROH} is the sum of the length of all runs of homozygosity (ROH) 239 detected in a subject, and L_{genome} is the total length of the human genome, estimated to be 240 2772.7 megabases (17)). ROH was called via the PLINK command ---homozyg on a pruned set 241 of common autosomal SNPs with pair-wise correlation <0.8 (--indep-pairwise 50 10 0.8, --maf 242 0.01). At the individual level, we removed samples with high missingness (>5%) and high 243 heterozygosity ($F_{het} < 0.23$ or > 0.33, Figs. S2, S4), individuals with a lower call rate for 244 duplicated sample pairs, related individuals (3rd degree or closer with a kinship coefficient > 245 0.088; those with the highest call rate were retained), those with ambiguous sex (0.2 < F246 estimate < 0.8), and any samples indicative of swap/mismatch issues (Figs. S2, S4). After the 247 thorough QC, a total of 520,234 variants were retained for 1814 Pakistani samples for phasing 248 (Fig. S2). In addition, the principal component analysis (PCA) was constructed, and all study 249 subjects were confirmed to have South Asian ancestry (Fig. S5).

250

251 Phasing and Imputation

Phasing was performed based on the 1000 Genome Project Phase 3 GRCh38 reference panel
(8) using SHAPEIT4 via the MoChA pipeline on Google Cloud (14,15) to generate phased vcf
files (Fig. S1, Supplemental methods). The MoChA pipeline has its default filtering steps to
exclude SNPs with more than 3% missingness and SNPs with excess heterozygosity (i.e.,

It is made available under a CC-BY-NC-ND 4.0 International license .

p<1×10⁻⁶ in a HWE test) before the phasing process. Next, we uploaded the sorted
chromosome-separated phased VCF files to the TOPMed and Michigan imputation servers (18)
for imputation via Minimac4 version 1.7.3. The imputation reference panels evaluated in this
study include TOPMed (GRCh38) (6), ex1KG (GRCh38) (7), 1KG38 (GRCh38) (8), 1KG37-SAS
(GRCh37) (8) and GenomeAsia-Pilot (GRCh37) (9). The sample size, in particular SAS and
Pakistani sample size, in each imputation panel as well as the mean read depth are included in
Table S1.

263 In addition, we performed meta-imputation using both TOPMed and ex1KG imputed 264 results through the MetaMinimac2 tool, which was designed to improve imputation performance 265 by statistically combined imputed results from different reference panels without the need to 266 physically merging the imputation panels together (19). Since meta-imputation can only run on 267 imputed results of the same genome build, we selected TOPMed (GRCh38) and ex1KG 268 (GRCh38) for meta-imputation, given that TOPMed is currently the largest imputation reference 269 panel whereas ex1KG has the largest sample size of Pakistani individuals (Table S1). An 270 overview of the study design for this paper is outlined in Fig. S1.

271

272 <u>Targeted Sequencing</u>

273 Besides genotyping on SNP arrays, targeted sequencing of 49234 exonic variants on 161 274 genes on 22 autosomal chromosomes was also performed in these Pakistani individuals, using 275 the Ion Torrent platform at Sema4, Inc. (Mount Sinai Genomics Inc., Connecticut, USA) with an 276 average sequencing depth of $224 \times (20)$. The Ion AmpliSeg technology was used to create the 277 sequencing library, in which the amplicons for the 161 genes were designed using Ion AmpliSeq 278 Designer version 6.13. Individual genotype called from the targeted sequencing was considered 279 as the true genotype to assess the accuracy of imputed genotype using different imputation 280 reference panels. The detailed QC of the targeted sequencing data is described elsewhere (20).

281

It is made available under a CC-BY-NC-ND 4.0 International license .

282 Imputation accuracy evaluation

<u>True R-square</u>: The true R^2 (R^2_{True}) is the squared correlation between the imputed genotype 283 dosage and the sequenced genotype. It is considered as the most robust measure to assess 284 285 the imputation accuracy given it is based on comparison against directly measured genotype 286 from targeted sequencing with rich read depth (mean: 224x), compared to the mean read depth 287 used for included imputation reference panels (7.4x to 38.2x, Table S1). Theoretically, $R_{True}^2 = 1$ 288 indicates that the imputed genotypes are the same as the sequenced genotypes, and therefore a R_{True}^2 value closer to 1 suggests a higher imputation accuracy. R_{True}^2 was calculated for each 289 290 individual SNP by the aggRSquare tool (19) and averaged across the default MAF bins in the 291 aggRSguare tool, including MAF cutoffs of 0.0005, 0.001, 0.002, 0.005, 0.010, 0.015, 0.020, 292 0.035, 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5. Given we only have targeted sequencing data for exonic variants of 161 genes, the R_{True}^2 can only be evaluated on a subset of genome-wide SNPs on 293 294 the 22 autosomal chromosomes (n=49234).

295

Estimated R-square: Without sequenced genotypes as a gold standard reference, R² may be 296 297 estimated (R_{Est}^2) based on their posterior allele probabilities for each imputed SNP on a 298 genome-wide scale (Table S2). Calculation of R_{Est}^2 is an embedded function in the Minimac tool 299 used by the online imputation server (11) as well as in the MetaMinimac2 tool for metaimputation (19). When Hardy Weinberg equilibrium holds (11), R_{Est}^2 is equivalent to the INFO 300 301 score produced by the IMPUTE software (21), another imputation quality score frequently used for SNP QC for GWAS. R_{Est}^2 may be less accurate than R_{True}^2 when there are insufficient 302 samples for imputation (11), and/or when the SNPs are rare (correlation with R_{True}^2 dropped to 303 304 0.72-0.77 for rare variants across imputation panels, Fig. S6). Well-imputed SNPs in this study 305 are defined as $R_{Est}^2 \ge 0.8$.

It is made available under a CC-BY-NC-ND 4.0 International license .

To test whether R_{Est}^2 is an appropriate proxy for R_{True}^2 in our study, we calculated imputation accuracy for SNPs with high R_{Est}^2 (≥ 0.8), a filter often applied in GWAS, and observed a corresponding improvement in R_{True}^2 across MAF, with most mean $R_{True}^2 \geq 0.8$ across different MAF bins (Figure 1d; correlation of R_{True}^2 and R_{Est}^2 =0.82-0.86, Fig. S6), confirming that R_{Est}^2 can be used as a proxy for R_{True}^2 when there is no sequencing data available to serve as reference (Fig. S7).

312

<u>Empirical R-Square</u>: Empirical R² (R_{Emp}^2) is the squared correlation between the imputed 313 314 genotype dosage and the directly measured genotype on the GSA array. Imputed SNP dosages 315 of these genotyped loci were calculated by masking the measured genotype as if they were unknown and then imputing these SNPs. Therefore, R_{Emp}^2 can be calculated only for SNPs 316 directly genotyped on the GSA array (i.e., nearly 500K SNPs) (Table S2). The R_{Emp}^2 calculation 317 is included in Minimac, but not in the MetaMinimac2 tool for meta-imputation (19). R_{Emp}^2 was not 318 319 strongly correlated with R_{Est}^2 (r = 0.43-0.54 across imputation panels, Figs. S8-S9) and therefore it is a less accurate measure for imputation accuracy, compared to R_{Est}^2 , which has a high 320 correlation with R_{True}^2 . Table S4 explains the difference between the three types of R². 321 322

323 SNP type

In this study, we defined rare variants as MAF < 0.01, and common if MAF \ge 0.01. The

325 distinction between single nucleotide variant (SNV) and insertion-deletions (indel) was based on

the number of nucleotides at one locus (i.e., if both reference and alternate alleles are single

327 nucleotides, then it was classified as a SNV; if any of the alleles had > 1 nucleotide, it was

- 328 classified as an indel). We used SnpEff to functionally annotate all the imputed variants (22).
- 329 For imputed data using the GenomeAsia-Pilot and 1KG37-SAS panels, the GRCh37.75 version

of the pre-built human database was used for annotation, whereas for imputed data using otherimputation panels, the GRCh38.99 version was used.

- 332
- 333 Statistical analyses

334 To compare imputation accuracy across different imputation panels at each MAF bin, we used 335 the Kruskal-Wallis rank sum test for significance testing, which is a non-parametric test of one-336 way ANOVA when ANOVA assumptions (i.e., homogeneity of variance and normality) are not 337 met. Further, to test for pairwise significance, we applied a pairwise t test without assuming 338 equal variance in the two groups, setting two-sided nominal significance at 0.05 for all statistical 339 analyses, while Bonferroni correction was applied to correct for multiple testing to reduce false 340 positive findings (e.g., Bonferroni-corrected significance set at $P_{Bonferroni} = 0.05/15 = 0.003$ for 341 pairwise comparison across 6 imputation panels). To evaluate whether SNPs with allele 342 frequency (AF) deviated from EUR AF would have lower imputation guality, we performed linear regressions between true R² and the absolute difference of AF using SAS AF based on targeted 343 344 sequencing and 1000G-based EUR AF. All analyses were conducted in the R statistical 345 software (version 4.2.0, Vienna, Austria). 346 347 348 349 350 351 352

It is made available under a CC-BY-NC-ND 4.0 International license .

354 **Declarations**

- 355 <u>Ethics approval and consent to participate</u>: Interviewers at the hospitals/enrollment centers from
- 356 Pakistan collected informed consent from the participants. The study was approved by the
- 357 Ethics committee at University of Peshawar, University of Health Sciences, Lahore and Lahore
- 358 Institute of Research and Development in Pakistan.
- 359
- 360 <u>Availability of data and materials</u>: Datasets used in this study are not publicly available because
- it contains private patient data in Pakistan, but they are available from the corresponding author
- 362 on request. The source code for the analyses is available and deposited in GitHub
- 363 (https://github.com/xuj18/).
- 364
- 365 <u>Competing interests</u>: The authors declare no competing interests.
- 366
- 367 <u>Funding</u>: J.X. and L.M.H. are both supported by the National Institute of Mental Health grant
- 368 R01MH118278. L.M.H. also acknowledges funding from NIMH (R01MH124839, RM1MH132648,
- 369 R01MH125938) and National Institute of Environmental Health Sciences (R01ES033630). D.L.,
- B.F., E.C., and A.W.C. were supported for this work by NIH R01MH109536. A.H., J.A.K., M.A.,
- and T.B.B. are supported by NIMH grants R01MH112904 and R01MH123775. T.B.B. and G.G.
- 372 were both supported by R01MH123451. G.G. was also supported for this work by NIH
- 373 R01MH104964. R.E.P., T.B.B., and L.M.H. are supported by NIMH grant R01MH125938. R.E.P.
- also received support from the Brain & Behavior Research Foundation NARSAD grant 28632
- 375 PS Fund. The funding body is not involved in the study design, data collection, data analysis,
- 376 result interpretation, and writing of the manuscript.
- 377
- 378 <u>Author contributions</u>: J.X. analyzed the data and drafted the manuscript for this study. J.X.,
- 379 R.E.P., T.B.B., and L.M.H. conceptualized and designed the study. J.X. and L.M.H. revised the

It is made available under a CC-BY-NC-ND 4.0 International license .

manuscript. A.H., J.A.K., and M.A. initiated the GEN-SCRIP study and collected data including blood samples from the study participants in Pakistan. D.L., B.F., E.C., and A.W.C. contributed to the production of both genotyping and targeted sequencing data for these Pakistani individuals. D.L. provided targeted sequencing data with quality control for the analysis in this manuscript. G.G. provided consultation on the use of MoChA pipeline. A.C.C. contributed to figure production in this manuscript. All authors read and approved the final manuscript for submission. Acknowledgement: This work was supported in part through the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai. Research reported in this paper was supported by the Office of Research Infrastructure of the National Institutes of Health under award number S10OD018522 and S10OD026880. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We would also like to thank Dr. Ketian Yu in the Department of Biostatistics at University of Michigan-Ann Arbor for her technical assistance with the MetaMinimac2 tool and the aggRSguare tool.

It is made available under a CC-BY-NC-ND 4.0 International license .

406 References

- 407 1. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev
 408 Genet. 2010 Jul;11(7):499–511.
- 2.Mills MC, Rahal C. A scientometric review of genome-wide association studies. Commun Biol.
 2019 Jan 7;2(1):1–11.
- Peterson RE, Kuchenbaecker K, Walters RK, Chen CY, Popejoy AB, Periyasamy S, et al.
 Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities,
 Methods, Pitfalls, and Recommendations. Cell. 2019 Oct 17;179(3):589–603.
- 414 4. Huang QQ, Sallah N, Dunca D, Trivedi B, Hunt KA, Hodgson S, et al. Transferability of
 415 genetic loci and polygenic scores for cardiometabolic traits in British Pakistani and
 416 Bangladeshi individuals. Nat Commun. 2022 Aug 9;13(1):4664.
- 417 5. Wall JD, Sathirapongsasuti JF, Gupta R, Rasheed A, Venkatesan R, Belsare S, et al. South
 418 Asian medical cohorts reveal strong founder effects and high rates of homozygosity. Nat
 419 Commun. 2023 Jun 8;14(1):3377.
- 420 6. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of
 421 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021
 422 Feb;590(7845):290–9.
- 7. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, et al. High-coverage
 whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios.
 Cell. 2022 Sep 1;185(18):3426-3440.e19.
- 426 8. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global
 427 reference for human genetic variation. Nature. 2015 Oct;526(7571):68–74.
- 428 9. Wall JD, Stawiski EW, Ratan A, Kim HL, Kim C, Gupta R, et al. The GenomeAsia 100K
 429 Project enables genetic discoveries across Asia. Nature. 2019 Dec;576(7785):106–11.
- 430 10. TOPMed Imputation Server [Internet]. [cited 2023 May 16]. Available from:
 431 https://imputation.biodatacatalyst.nhlbi.nih.gov/#!pages/about
- 432 11. Das S, Abecasis GR, Browning BL. Genotype Imputation from Large Reference Panels.
 433 Annual Review of Genomics and Human Genetics. 2018;19(1):73–96.
- 434 12. Anwar I, Taroni F. Genetic peopling of Pakistan: Influence of consanguinity on
 435 population structure and forensic evaluation of traces. Forensic Science International:
 436 Genetics Supplement Series. 2019 Dec 1:7(1):232–3.
- 437 13. Quick C, Anugu P, Musani S, Weiss ST, Burchard EG, White MJ, et al. Sequencing and
 438 imputation in GWAS: Cost-effective strategies to increase power and genomic coverage
 439 across diverse populations. Genetic Epidemiology. 2020;44(6):537–49.
- 440 14. Loh PR, Genovese G, Handsaker RE, Finucane HK, Reshef YA, Palamara PF, et al.
 441 Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. Nature.
 442 2018 Jul;559(7714):350–5.

- 443 15. Genovese G. The MOsaic CHromosomal Alterations (MoChA) WDL Pipeline [Internet].
 444 2023 [cited 2023 May 15]. Available from: https://github.com/freeseek/mochawdl
- 445 16. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A
 446 Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum
 447 Genet. 2007 Sep;81(3):559–75.
- 448 17. Christofidou P, Nelson CP, Nikpay M, Qu L, Li M, Loley C, et al. Runs of Homozygosity:
 449 Association with Coronary Artery Disease and Gene Expression in Monocytes and
 450 Macrophages. The American Journal of Human Genetics. 2015 Aug 6;97(2):228–37.
- 451 18. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation
 452 genotype imputation service and methods. Nat Genet. 2016 Oct;48(10):1284–7.
- Yu K, Das S, LeFaive J, Kwong A, Pleiness J, Forer L, et al. Meta-imputation: An
 efficient method to combine genotype data after imputation with multiple reference panels.
 The American Journal of Human Genetics. 2022 Jun 2;109(6):1007–15.
- 456 20. Liu D, Meyer D, Fennessy B, Feng C, Cheng E, Johnson JS, et al. Schizophrenia risk
 457 conferred by rare protein-truncating variants is conserved across diverse human populations.
 458 Nat Genet. 2023 Mar;55(3):369–76.
- 459 21. Howie BN, Donnelly P, Marchini J. A Flexible and Accurate Genotype Imputation Method
 460 for the Next Generation of Genome-Wide Association Studies. PLOS Genetics. 2009 Jun
 461 19;5(6):e1000529.
- 462 22. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for
 463 annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin).
 464 2012 Apr 1;6(2):80–92.
- 465
- 466
- 467
- 468
- 469
- 470
- 471
- 472
- 473
- 474
- 475

It is made available under a CC-BY-NC-ND 4.0 International license .

476 Figures





478 Figure 1. Imputation accuracy and coverage compared against targeted sequencing in 479 the Pakistani individuals. a) Imputation accuracy of different imputation panels across 480 different MAF bins, measured by , for variants that were also included in the targeted 481 sequencing (i.e., 49234 exonic variants on 161 genes). is the squared correlation 482 between imputed genotype dosage and sequenced genotype, with 1 being a perfect match. 483 Significance: * indicates a nominally significant difference in the mean across different imputation panels (Kruskal-Wallis rank sum test P < 0.05) at a specific MAF bin; ** denotes a 484 Bonferroni-corrected significant difference (P < 0.003, correcting for the number of pairwise 485 486 panels tested). The horizontal grey lines indicate of 0.6 (dashed line, indicating fairly high 487 imputation guality) and 0.8 (solid line, indicating high imputation guality). The vertical grey lines 488 indicate MAF > 1% (dashed line) and > 5% (solid line); b) The number of common variants 489 imputed by different panels compared to targeted sequencing; c) The number of rare variants 490 imputed by different panels compared to targeted sequencing; d) Imputation accuracy of 491 different imputation panels, measured by , for variants with an of 0.8 or greater. SNPs

It is made available under a CC-BY-NC-ND 4.0 International license .

with high R_{Est}^2 also have high R_{True}^2 values (average ≥ 0.65), suggesting the appropriateness of using R_{Est}^2 as a proxy for R_{True}^2 when there is no sequencing data available as the gold standard reference; e) The number of well-imputed common variants by different imputation panels compared against targeted sequencing; f) The number of well-imputed rare variants (defined as $R_{Est}^2 \ge 0.8$) by different imputation panels compared against targeted sequencing. Abbreviation: 1000G, 1000 Genomes; MAF, minor allele frequency; SAS, South Asian ancestry; SNP, single nucleotide polymorphism.

It is made available under a CC-BY-NC-ND 4.0 International license .



500 Figure 2. Genome-wide imputation accuracy and coverage in the Pakistani individuals. a)

499

Imputation accuracy of different imputation panels measured by , which is the squared
correlation between imputed genotype dosage and true genotype based on posterior allele
probabilities. A value closer to 1 indicates higher imputation accuracy. Significance: *
indicates a nominally significant difference in the mean across different imputation panels
(Kruskal-Wallis rank sum test P < 0.05) at a specific MAF bin; ** denotes a Bonferroni-corrected

It is made available under a CC-BY-NC-ND 4.0 International license .

506 significant difference (P < 0.003, correcting for the number of pairwise panels tested). The 507 horizontal grey lines indicate R_{Fst}^2 of 0.6 (dashed line, indicating fairly high imputation quality) 508 and 0.8 (solid line, indicating high imputation guality). The vertical grey lines indicate MAF > 1%509 (dashed line) and > 5% (solid line); b) The number of imputed SNPs genome-wide by 510 common/rare variants; c) The number of well-imputed SNPs ($R_{Est}^2 \ge 0.8$) genome-wide by 511 common/rare variants; d) The number of well-imputed SNPs genome-wide by common/rare and 512 SNV/indel status; e) The number of well-imputed SNPs genome-wide by common/rare and bi-513 allelic/multi-allelic status; f-i) The number of well-imputed SNPs genome-wide by common/rare 514 status and SNP types via functional annotation; f) includes 4 SNP types with the highest SNP 515 counts (i.e., intron, intergenic, downstream, upstream); g) includes 5 SNP types with following 516 highest SNP counts (i.e., non-coding transcript exon, 3' UTR, missense, synonymous, 5' UTR); 517 h) includes 6 SNP types with low SNP counts (i.e., splice region, 5' UTR premature start codon 518 gain, intragenic, splice donor, splice acceptor, stop gained); i) includes 5 SNP types with ultra-519 low SNP counts (i.e., start lost, stop lost, stop retained, initiator codon, start retained). 520 Abbreviation: 1000G, 1000 Genomes; indel, insertion-deletion; MAF, minor allele frequency; 521 SAS, South Asian ancestry; SNP, single nucleotide polymorphism; SNV, single nucleotide 522 variant; UTR, untranslated region.