Imputation Efficacy Across Global Human Populations

Jordan L. Cahoon[1,2,3], Xinyue Rui[1,*], Echo Tang[2,*], Christopher Simons[2], Jalen Langie[1], Minhui Chen[1], Ying-Chu Lo[1], Charleston W. K. Chiang[1,2,4,†]

[1]Center for Genetic Epidemiology, Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA
[2]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA
[3]Department of Computer Science, University of Southern California, Los Angeles, CA 90089, USA
[4]Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA 90033, USA

* These authors contributed equally
† Correspondence: charleston.chiang@med.usc.edu

Abstract
Genotype imputation is now fundamental for genome-wide association studies but lacks fairness due to the underrepresentation of populations with non-European ancestries. The state-of-the-art imputation reference panel released by the Trans-Omics for Precision Medicine (TOPMed) initiative contains a substantial number of admixed African-ancestry and Hispanic/Latino samples to impute these populations with nearly the same efficacy as European-ancestry cohorts. However, imputation for populations primarily residing outside of North America may still fall short in performance due to persisting underrepresentation. To illustrate this point, we curated genome-wide array data from 23 publications published between 2008 to 2021. In total, we imputed over 43k individuals across 123 populations around the world. We identified a number of populations where imputation accuracy paled in comparison to that of European-ancestry populations. For instance, the mean imputation r-squared (Rsq) for 1-5% alleles in Saudi Arabians (N=1061), Vietnamese (N=1264), Thai (N=2435), and Papua New Guineans (N=776) were 0.79, 0.78, 0.76, and 0.62, respectively. In contrast, the mean Rsq ranged from 0.90 to 0.93 for comparable European populations matched in sample size and SNP content. Outside of Africa and Latin America, Rsq appeared to decrease as genetic distances to European reference increased, as predicted. Further analysis using sequencing data as ground truth suggested that imputation software may inflate imputation quality for non-European populations, implying that these quality estimates may be lower than initially estimated. To improve imputation quality, we assessed a strategy using meta-imputation to combine results from TOPMed with smaller population-specific reference panels, using 1496 whole genome sequenced individuals from Taiwan Biobank as an example reference. While we found that meta-imputation in this design did not improve Rsq genome-wide, Southeast Asian populations such as Filipino and Vietnamese experience a 0.16 and 0.11 increase in average imputation Rsq, respectively, for alleles extremely rare in Europeans (< 0.1% frequency) but more common (>1%) in East Asians. Taken together, our analysis suggests that meta-imputation may complement a large reference panel such as that of

TOPMed for underrepresented cohorts. Nevertheless, reference panels must ultimately strive to increase diversity and size to promote equity within genetics research.

Introduction

The genome-wide association study (GWAS) is a powerful tool that enables researchers to gain insight into disease mechanisms, discover potential drug targets, and forecast an individual's risk for health concerns in preventive medicine. Within the past decade, genotype imputation has become an essential component in GWAS. By predicting unobserved genotypes based on a reference panel of externally sequenced individuals, imputation increases marker density, improves the statistical power of GWAS, and enables large-scale meta-analysis across cohorts[1]. Importantly, the size, data quality, and haplotype diversity of imputation reference panels determine the accuracy of genotype imputation[2]. Target populations that are more genetically related to the reference populations will have better imputation quality compared to those that are more distantly related. Hence, better representation in imputation reference panels will improve the number of diverse populations that can benefit from genotype imputation and GWAS.

Historically, GWAS cohorts have overrepresented European-ancestry individuals[3–5]. Similarly, an abundance of European-ancestry cohorts being whole genome sequenced also meant that large-scale imputation reference panels were Euro-centric (e.g. ref [6]), which disproportionately performed worse for non-European cohorts. In 2020, the state-of-the-art imputation reference panel released by the Trans-Omics for Precision Medicine (TOPMed)[7] improved the diversity of reference haplotypes by including a substantial number of admixed African and Hispanic/Latino individuals. As a result, these populations are now much better imputed, with nearly the same efficacy as European-ancestry cohorts, though the performance still falls short for populations of other continental ancestries such as the East Asians and Polynesians[8,9].

As we push the frontier of genomic medicine to the global scale, it is necessary to be inclusive across diverse populations[10,11]. State-of-the-art imputation reference panels are still largely based on whole genome sequencing data from populations in North America and Europe. We anticipate the TOPMed Reference Panel may fall short in performance due to persisting underrepresentation of diverse global populations, particularly for ethnic minorities residing outside of North America and Europe. To illustrate this point, we curated and evaluated the imputation accuracy across 123 populations worldwide. Because cohort recruitments and GWAS are often conducted in (potentially arbitrarily defined) discrete populations[12] (though see ref. [13]) and for the ease of interpreting imputation quality metrics, our study necessitated defining and assigning descriptive labels to populations across the globe. Though we recognize that human variations exist on a continuum across space, we focused on identifying broad patterns across geographical space, rather than implicating specific populations; when we did represent results from a single or a few populations, we did so as illustrative examples.

By using the quality metric Imputation r-squared (Rsq), we first identified a number of populations where the imputation accuracy was substantially lower than that of populations residing in the European continent. We found that imputation accuracy remains relatively low for populations from Asia (East, Southeast, and South), the Middle East, Oceania, and the Pacific, particularly for intermediate frequency alleles (1-5%). The imputation accuracy may be even lower than estimated by Rsq when comparing available genotypes from sequencing data to calculate imputation quality. For selected populations that are poorly imputed with TOPMed references, we also assessed approaches to improve imputation quality by meta-imputing TOPMed-imputed genotypes with those from a smaller but more population-specific reference panel[14]. We found that meta-imputation only marginally improved the imputation accuracy with currently available resources, but it showed promise with imputing alleles that are enriched in populations underrepresented in the TOPMed Reference Panel. Our results thus suggest an urgent need to increase the size and diversity of cohorts prioritized for whole genome sequencing.

Results

Estimating Imputation Quality with the TOPMed Reference Panel

To identify the current insufficiencies of the TOPMed panel for imputing diverse global populations, we curated genome-wide array data from 23 studies published between 2008-2021 (**Methods**). In total, we imputed over 43k individuals across 123 populations, spanning multiple array platforms, three orders of magnitudes in sample sizes, and ten major geographical regions around the globe (**Supplemental Figures 1-2**, **Supplemental Tables 1-2**). We assessed the imputation quality of each population by comparing the distribution of Rsq, stratified by bins of minor allele frequencies, to that of a European cohort (the UK10K sequencing data)[15] subsampled to match in sample size and array SNP content (**Methods; Figure 1**). We generally observed that the imputation of common variants is more accurate than that of rare variants (**Supplemental Figure 3**). As expected, arrays with higher marker density also tend to perform better in imputation compared to those with lower marker density (**Supplemental Figure 3**).

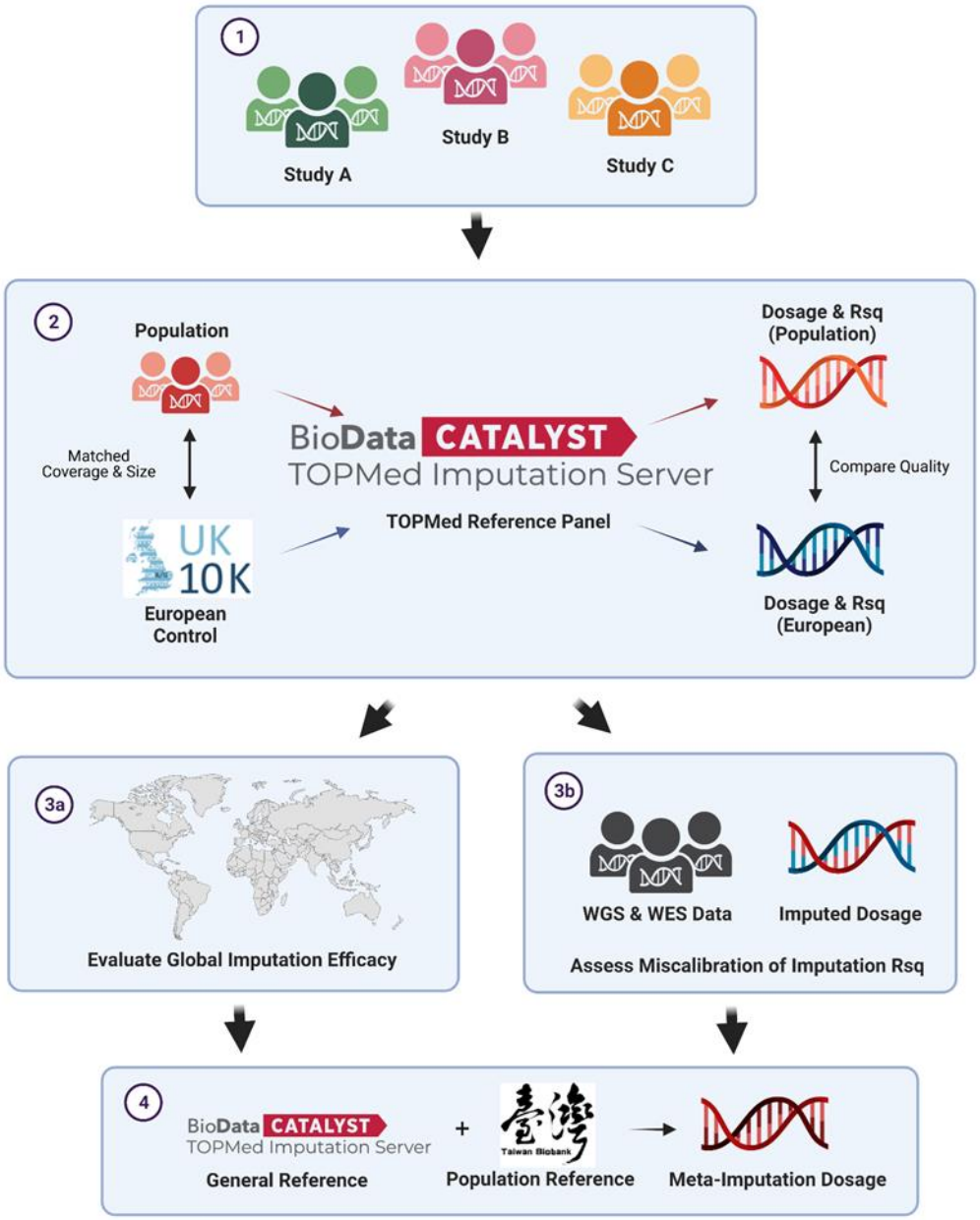

Figure 1: Study Design Overview. (1) We collected and curated genome-wide microarray data from 39 publications published between 2008-2021. In total we imputed 43k individuals and assessed the imputation efficacy in 122 populations across the globe. (2) For each target population we created a matched European control population with the same SNP content and sample size from the UK10K dataset. Both populations were imputed using the TOPMed Imputation Server. (3a) Imputation quality of each target population was evaluated by comparing the mean imputation Rsq in strata of allele frequencies to that of the matched European controls. (3b) For selected populations with available sequencing data, we assessed imputation quality using Pearson's r2 comparing the imputed dosage with the genotype from sequenced data. (4) We assessed meta-imputation to improve imputation quality by combining imputed dosages from TOPMed and Taiwan Biobank (TWB).

Most importantly, the imputation of the non-European populations tends to be worse than that of the matched European populations, particularly for alleles with less than 5% frequency (**Supplemental Figure 4, Supplemental Table 3**). To identify broad patterns across the globe, we visualized the ratio of mean Rsq between the target population and the matched European control population on a world map. From these comparisons, populations from Asia (East, Southeast, South), the Middle East, Siberia, Oceania, and the Pacific have the overall lowest imputation quality compared to matched controls from Europe, presumably due to underrepresentation of haplotypes from these regions within the TOPMed panel (**Figure 2A**). For instance, the mean Rsq for 1-5% alleles in Saudi Arabians (N=1061)[16], Vietnamese (N=1264)[17], Thai (N=2435)[17], and Papua New Guineans (N=776)[17] were 0.79, 0.78, 0.76, and 0.62, respectively. In contrast, the mean Rsq ranged from 0.90 to 0.93 for matched European populations (**Figure 2A**). On the other hand, African-ancestry populations as well as some American and Pacific populations imputed relatively well compared to European populations, likely due to their representation, directly or through shared ancestry from recent admixture, in TOPMed (**Figure 2A**). Expectedly, the imputation accuracy for a target population is inversely correlated with the genetic differentiation, as measured by the fixation index (FST), to the matched European populations (**Figure 2B**). The exceptions to this relationship are those populations from Africa or the Americas, which reflects the improved representation of African-ancestry individuals and Latinos in the TOPMed panel (**Figure 2B**). We observed qualitatively similar but less extreme patterns for the imputation efficacy of alleles 5% or greater in frequency (**Supplemental Figure 5**). Taken together, the discrepancies in imputation efficacy across the frequency spectrum suggest that some populations across the globe, notably those from Asia, Oceania, and the Pacific, will not reap the full benefits of imputation of a large reference like the TOPMed panel, and suffer a relative loss in statistical power in GWAS even if large-scale recruitment of study cohorts is feasible.

We developed an interactive website that stores and displays summaries of all of our imputation results for users to access: https://jlcah5.github.io/iaw/map.html

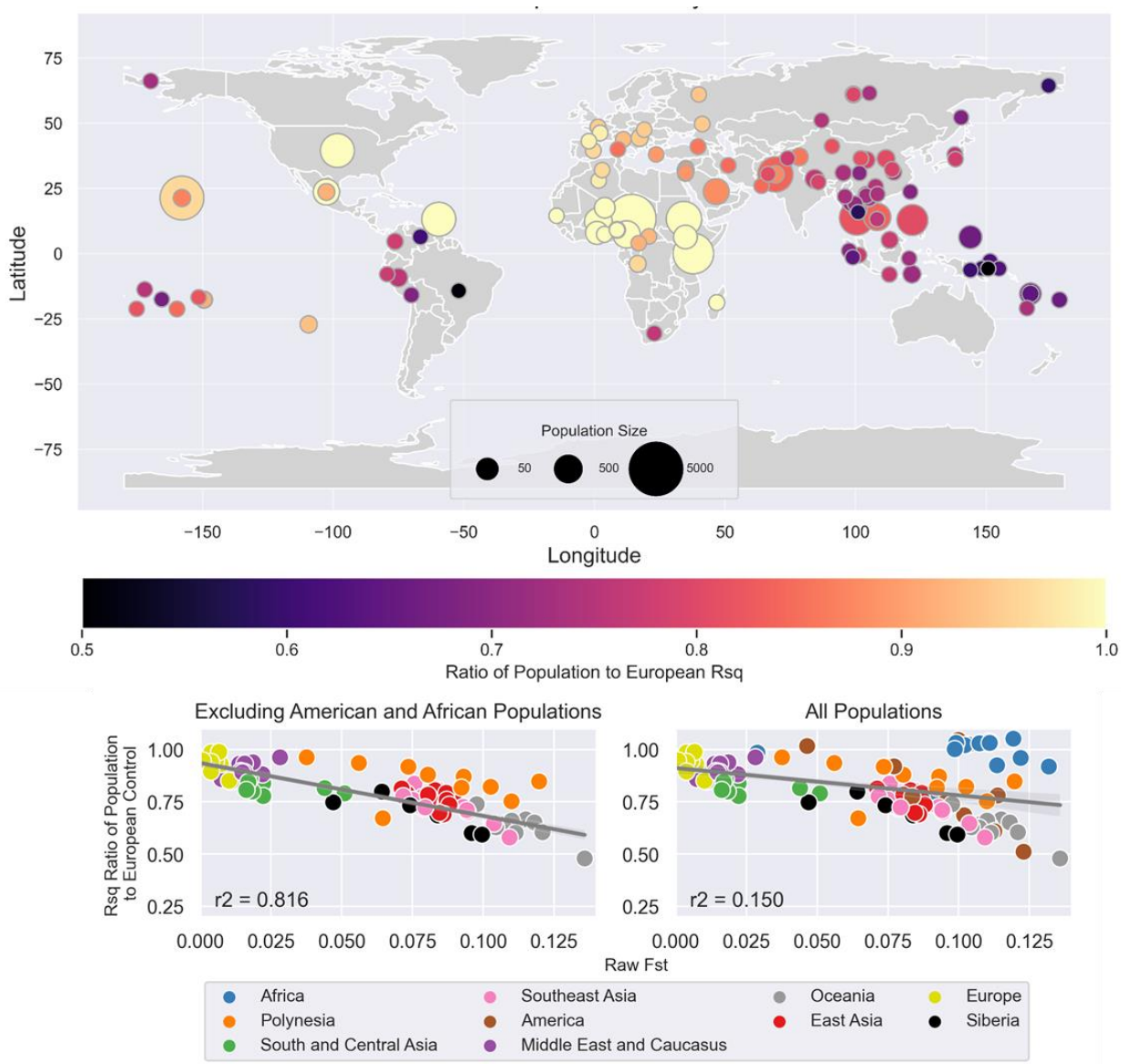

Figure 2: Disparity of TOPMed imputation efficacy across the globe. (A) Ratio of the mean imputation Rsq of 1-5% alleles for every target population to its corresponding matched european control are shown on the world map. Values closer to 1.0 indicate no relative imputation quality loss compared to a European cohort matched by sample size and SNP content. Size of the circle scales with the sample size of the dataset that was imputed, and color of the circle scales with imputation quality. (B) Relationship between the ratio of mean Rsq and Fst between each target population and its European control. Each population is colored according to its assigned broad geographical region. The left plot excluded African and American populations. The right plot contains all populations analyzed, where notably African-ancestry populations show strong genetic differentiation from Europeans by Fst, but are imputed with high efficacy using TOPMed.

Inferred Gaps in Imputation Quality across Populations May be Overly Optimistic

For many of the populations assessed, we relied on Rsq estimated by the imputation software as a metric of imputation quality. However, imputation Rsq is estimated from imputed frequencies without information from observable frequencies, and as a result may misrepresent the true imputation accuracies.[18]. A better estimate of imputation

accuracy, particularly in relation to the potential loss of statistical power in GWAS, is based on Pearson's correlation coefficient (r^2), computed between the imputed dosage and the observed genotype. For selected populations where we have access to publicly available high-coverage genome sequencing data (**Supplemental Table 4**), we used Pearson's r^2 to measure imputation quality.

In general, we found a qualitatively similar pattern using Pearson's r^2: common variants are imputed better than rare variants, and European-ancestry populations are imputed better than other non-European populations (**Supplemental Figure 6**). However, when comparing the two measures of imputation accuracy (imputation Rsq vs. Pearson's r^2), we observed that imputation Rsq is inflated compared to Pearson's r^2. While the inflation is evident across the frequency spectrum, it is more inflated among the rarer alleles (**Figure 3, Supplemental Table 5**). Importantly, imputation Rsq appears to be biased upwards in non-European populations than European populations compared to Pearson's r^2 (**Figure 3**). For instance, mean imputation Rsq for 10-50% frequency alleles genome-wide were similar between African-ancestry populations and their matched European counterpart (**Figure 2**), suggesting similar imputation accuracies. But for an allele between 10-50% frequency in Europeans with an estimated imputation Rsq of 0.5, the expected Pearson's r^2 is a relatively well-calibrated 0.42. In contrast, for an allele in the same frequency range and estimated Rsq in African Americans, it would have an expected Pearson's r^2 of 0.23, a 55% relative decrease in power when testing the association of this imputed variant with phenotype. We also observed excess inflation across the frequency spectrum for other populations from the Human Genome Diversity Project that we assessed (**Supplemental Figures 7-15**). Therefore, while we have shown reduced efficacy when imputing diverse populations using the current TOPMed imputation resources, our results above may suggest that the true imputation efficacy may be further lowered for non-European populations.

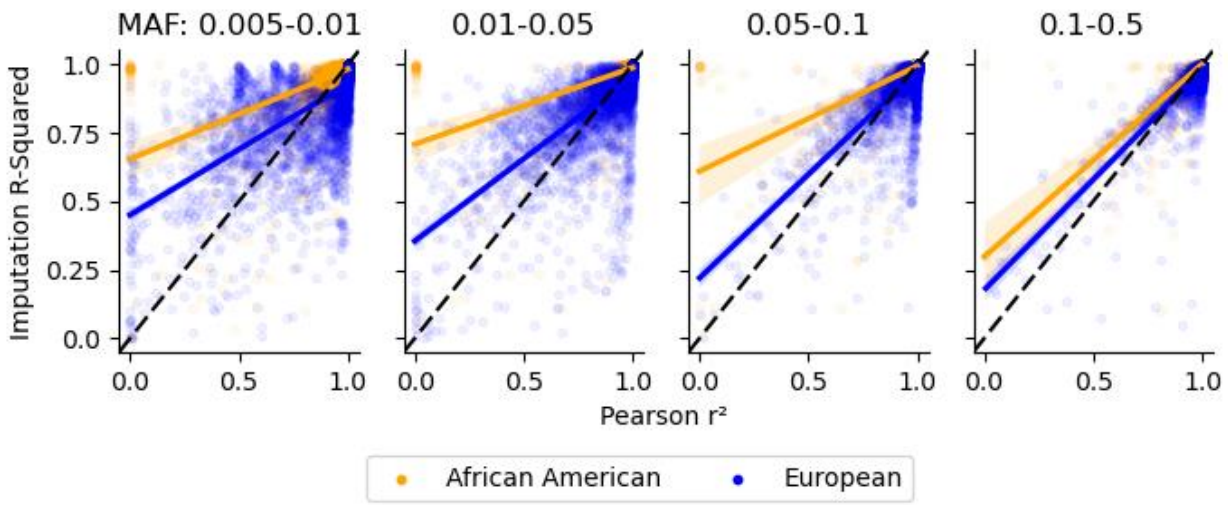

Figure 3: Imputation Rsq is an overly optimistic metric of imputation efficacy. The relationship between Pearson's r2 and Minimac4-Produced Imputation Rsq. A random set of 5,000 variants in each strata of minor allele frequency were sampled from 23andMe's African American Sequencing Project data (N=2269) and Europeans from the Human Genome Diversity Project (N=155). The dash line denotes unity, while the solid lines are fitted linear models. Comparing the Imputation Rsq to Pearson's r^2 for the same set of variants, Imputation Rsq appears to overestimate imputation quality on average. Imputations of African American samples experienced a higher inflation of Imputation Rsq compared to that of Europeans across all minor allele frequency bins.

Meta-Imputation Increases Imputation Quality for Population-Enriched Variants

We suggest that populations in key regions around the world are imputed with comparably lower quality using the TOPMed reference panel due to underrepresentation within the imputation reference. One approach to improve imputation efficacy and circumvent the issue of underrepresentation in the TOPMed panel is meta-imputation. Meta-imputation is a tool that combines the imputed dosages after imputation across multiple panels using weights based on empirically estimated imputation performance at genotyped sites[14]. Specifically, one could combine the imputation result from the TOPMed reference panel with a smaller, regional or population-specific imputation panel to improve the imputation efficacy, while still leveraging the large size of the TOPMed panel for better imputation of common variants.

We test the efficacy of meta-imputation using chromosome 11 from the Filipinos (N = 1,799)[19], Pakistani (N=2493)[20], Taiwanese (N=10,000)[21], and Vietnamese (N = 1,264)[17], chosen for their relatively large sample size and poor imputation using TOPMed alone (**Figure 2**). We meta-imputed the TOPMed-imputed results with that of a reference panel constructed from whole genome sequenced individuals from the Taiwan BioBank (TWB; N = 1,496)[21,22]. In total, 1,027,303 variants were found in both the TOPMed and TWB panels for chromosome 11 and were available for comparison. In general, imputation quality with the TWB panel was substantially poorer than that with the TOPMed panel, even for Taiwanese individuals (**Figure 4**). This is likely due to the sheer sample size of

the TOPMed imputation panel. As a result, meta-imputation between TOPMed and TWB panels overall rarely improved imputation efficacy, as quantified by imputation Rsq, and only marginally improved Rsq for common alleles, if at all (**Figure 4**). For instance, for 1-5% alleles, meta-imputation achieved a mean Rsq of 0.75, 0.64, 0.75, and 0.74, for Vietnamese, Filipino, Pakistani, and Taiwanese, respectively. This is comparable to the mean Rsq from imputing with the TOPMed panel alone (0.79, 0.68, 0.76, and 0.73 for the same populations, respectively).

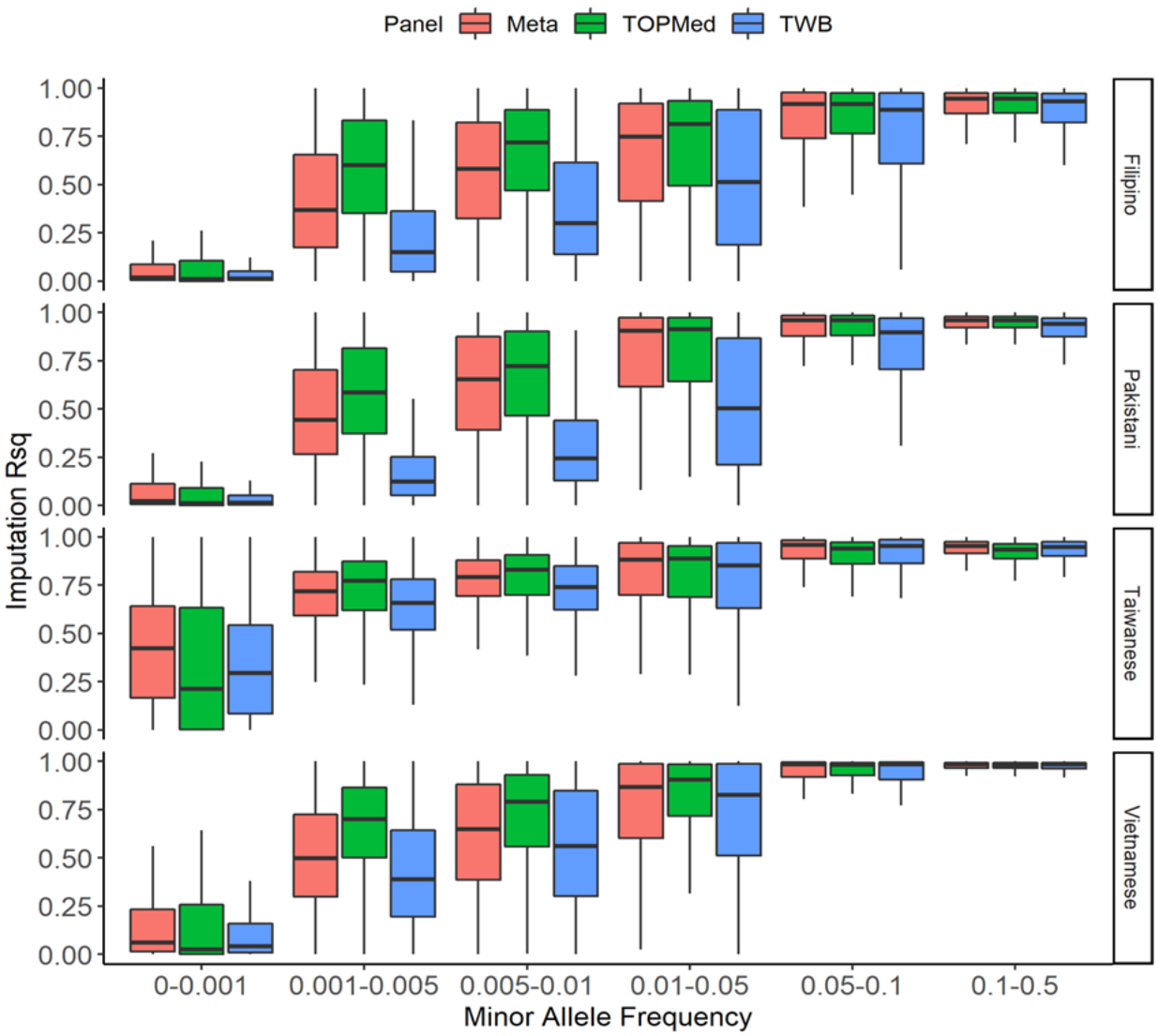

Figure 4: Meta-imputation does not improve imputation efficacy at overlapping intermediate frequency variants. Four populations, Filipino (N=1799), Pakistani (N=2493), Taiwanese (N=10000), and Vietnamese (N=1264) were independently phased and imputed with both Taiwan BioBank (TWB) and TOPMed. Imputation results at overlapping variants on chr11 were meta-imputed, and the distribution of Imputation Rsq across different strata of allele frequencies are shown for imputation with TOPMed (blue), TWB (red), and meta-imputation (green). Meta-imputation only marginally improves the imputation efficacy as measured by Rsq for common variants, if at all, even in cases when the target (e.g. Taiwanese) is from the same population as one of the references (TWB).

While meta-imputing TWB and TOPMed Reference Panels together achieved little improvements in imputation accuracy compared to imputing with the TOPMed reference panel alone, alleles enriched in East Asians experienced a boost in imputation quality (**Supplemental Figures 16-19**). For example, restricting to ~140,000 alleles on chr11 that are found in less than 0.1% frequency in Genome Aggregation Database (gnomAD)[23] non-Finnish Europeans but are found in 1% or greater frequency in gnomAD East Asians, we observed that Vietnamese, Filipino, and Taiwanese populations experienced an average of 16%, 11%, and 2.8% increase in imputation Rsq (**Figure 5**). In contrast,

Pakistani individuals did not experience an improvement in the imputation of East Asian-enriched variants, as these variants are likely rarer in South Asia, or exist on different haplotypic backgrounds that are underrepresented in the limited samples found in TWB. In addition to the improved imputation accuracy of population-enriched variants, meta-imputation also increased the variant coverage available for analysis. When meta-imputing with the TOPMed (14,564,684 SNPs on chr11) and TWB (1,957,999 SNPs on chr11) reference panels, there was an addition of 930,696 variants from TWB that would become available for analysis on chromosome 11, mostly due to extremely rare variants in TWB not found in TOPMed (**Supplemental Figure 20**). Meta-imputing TOPMed and TWB together led to a 6.4% increase in SNP coverage compared to imputing with the TOPMed panel alone.

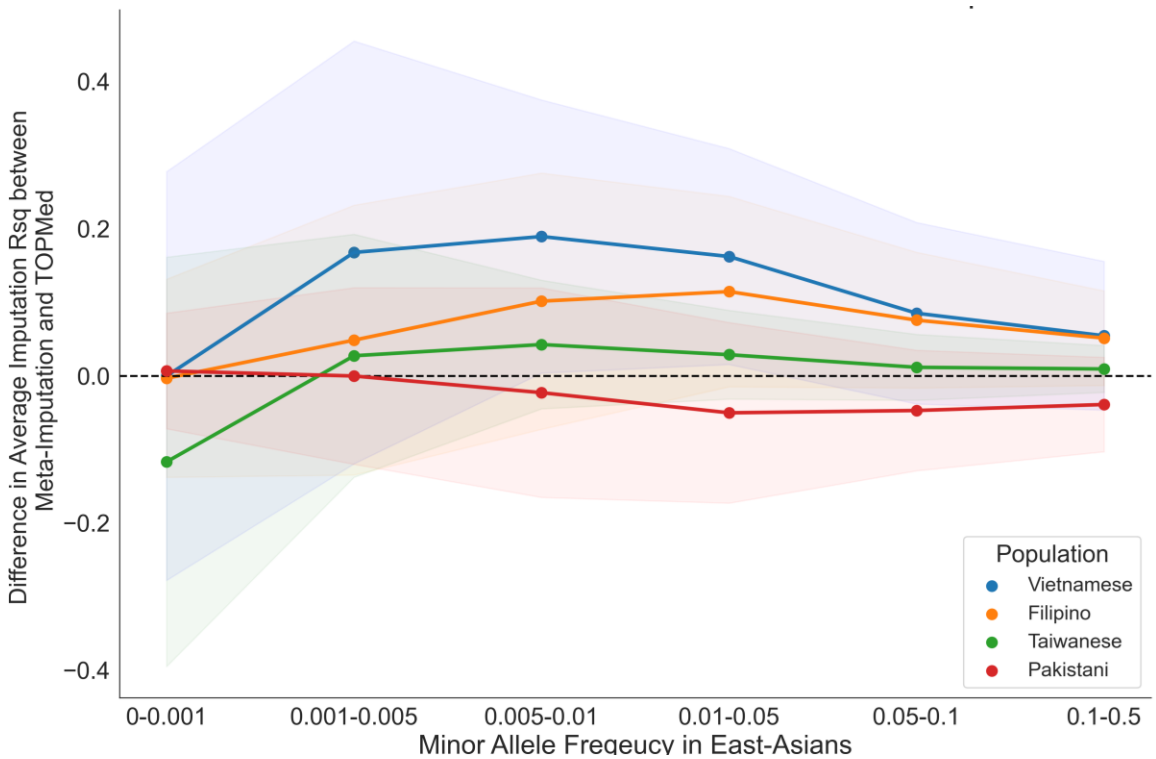

Figure 5: Meta-imputation improves the imputation efficacy at population-enriched variants. Restricted to variants with minor allele frequency less than 0.1% in gnomAD non-Finnish Europeans, meta-imputation combining TOPMed with TWB improves the imputation quality for variants that are enriched in East Asians, relative to just imputing with TOPMed alone. This improvement is particularly obvious for low (0.5% to 1%) to intermediate (1% to 5%) frequency alleles. No improvement was observed when Pakistani was used as the target population, presumably due to substantial differences in haplotypic patterns between Pakistani and TWB. Lines reflect the mean differences between meta-imputation and TOPMed-imputation alone. Shaded bands reflect the standard deviation across variants within the frequency bin.

Discussion

Genotype imputation is a critical tool that allows us to conduct effective, population-scale association studies. Our work reveals key discrepancies driven by the lack of global representation in the state-of-the-art imputation panel generated by the TOPMed consortium. Across 123 global populations, accounting for >43,000 individuals, we concluded that imputation quality is generally lower for diverse populations across the globe, especially those residing outside of the Americas and Europe, compared to matched European controls. For instance, for 1-5% alleles, populations from East Asia, the Middle East, Oceania, Polynesia, Siberia, Central and South Asia, and Southeast Asia are imputed with average reductions of 6.5-42% in imputation Rsq compared to controls (**Supplemental Table 3**). In addition, we found that Pearson's r^2, a metric of imputation accuracy that can be calculated when sequencing data is available, is on average lower than the imputation Rsq for variants across the entire frequency spectrum. Because Pearson's r^2 reflects the potential loss in statistical power in GWAS, our global assessment using Rsq may underestimate the degree to which disparities in imputation efficacy exist. The differences in imputation efficacy appears to be driven by the under-representation of non-European and non-American populations in TOPMed. By leveraging both TOPMed and a regional panel using individuals from TWB, we assessed meta-imputation as a method to improve imputation efficacy. Overall, we found meta-imputation to be a promising method that may increase the imputation quality of population-enriched variants and overall variant coverage.

While our analysis covered many global regions, there were a number of areas that we could not adequately gather data comprehensively. For example, our sub-Saharan African populations are dominated by Western and Eastern Africans. While our results showed that African populations are imputed with similar quality to the matched European populations (based on Rsq), this is likely reflecting the predominant West African ancestries that are found in the African American populations in the TOPMed reference panel[24]. The same level of imputation efficacy may not generalize to the Southern African regions. In fact, imputation Rsq for South Africans is 0.28 lower than that of their European controls for 5-50% alleles **(Supplemental Table 1, Supplemental Figures 4-5)**. More African datasets outside of West Africa are emerging[25,26] and future assessments are needed to better characterize the generalizability of current reference panels to the rest of the African continent.

Moreover, many of the African cohorts (*e.g.* Gambian, Ghanan, Kenyan, etc.), along with the cohorts from Southeast Asia (*e.g.* Thai, Vietnamese), were recruited as part of large epidemiological studies; study participants were sampled from major metropolitan areas, sometimes even residing outside of the country of origin. Hence, these larger cohorts may not fully represent the haplotypic diversity of all local populations in their respective

regions. Importantly, individuals with higher components of indigenous ancestries tend to be imputed with lower quality. For example, Native Hawaiian individuals with elevated levels of Polynesian ancestries[27] were imputed with ~0.1 lower Rsq in 1-5% alleles compared to randomly sampled admixed Native Hawaiians (**Figure 2A, Supplementary Table 1, Supplemental Figure 4**). Our results of imputation efficacy may thus still not be generalizable to indigenous or ethnic minority populations from the regions of the globe that we sampled. On the other hand, population genetic studies would focus on indigenous or ethnic minority populations, but they tend to be much smaller in sample size, thus preventing us from comprehensively evaluating the efficacy of imputation for rarer alleles. Nevertheless, we have observed a general decrease in imputation efficacy for intermediate frequency (1-5%) and common (> 5%) alleles. The disparity in imputation efficacy will likely widen for rarer alleles.

To evaluate meta-imputation as a means to improve imputation quality, we focused on a few Asian populations with sample sizes in the thousands as examples. One major challenge with the current meta-imputation framework is the drastic size differences between TOPMed and population-specific panels such as TWB. Meta-imputation of the 1000 Genomes Project and the Haplotype Reference Consortium (HRC) imputation panels were previously shown to have nearly the same imputation efficacy as combining both imputation panels and better imputation efficacy than imputing with either panel alone[14]. In our assessment, the TOPMed reference panel contains approximately 65-fold more individuals and nearly 7-fold the number of sites as TWB. On chr11, only ~7% and ~52% of variants from TOPMed and TWB, respectively, were found commonly on both panels and available for meta-imputation. As a result, meta-imputation may not consistently improve imputation quality when the TOPMed reference panel is involved, particularly for rare variants as these are less likely to be shared among imputation panels. Despite this, we observed a boost in imputation Rsq for variants enriched in East Asians for a number of regional populations (**Figure 5, Supplemental Figures 16-19**). Further, the additional reference panels would increase SNPs available for association testing by ~6% compared to imputation with TOPMed alone for chr11. As more whole genome sequencing data become available, particularly from diverse regions around the world, we expect the benefit of meta-imputation to become more apparent for non-European populations.

As global sequencing efforts progress, we can expect the release of more regional imputation panels. In our analysis, we focused on TWB due to its potential to improve the imputation quality of East Asian populations. Moreover, because many Taiwanese individuals are of Southern Chinese origin, and aboriginal Taiwanese may be related to their ancestors that expanded across the Pacifics[9,28], there may be some benefit to using TWB for the meta-imputation of Southeast Asian individuals. At the time of writing, only

small pilot sequencing efforts from other parts of the world, such as those from the GenomeAsia100K Consortium[29] were available for imputation through the Michigan Imputation Server[30]. As sequencing efforts across the world continue to release new data, we also expect meta-imputation to improve the imputation efficacy of variants enriched among individuals residing in regions such as Southeast and South Asia.

It is important that we maximize the number of individuals and populations who can benefit from large-scale association studies, and understanding the representativeness and performance of the current state-of-the-art imputation reference panel is an important first step. While we point to meta-imputation as a potential method to increase the quality of some imputed variants, we must continue to sequence diverse individuals and strive toward more representation in our scientific resources.

Online Methods

Dataset Collection

We gathered genotyping array datasets from 23 studies with publication dates ranging from 2008-2022. These represented 14 different genotyping array platforms and 43,507 individuals. These datasets were either publicly available, accessed through online databases such as dbGaP, or requested from the original author of the publication (**Supplemental Table 1**). Before data processing, all genotype and variant information was converted to PLINK Binary file format, and genomic coordinates were converted to GRCh38 using liftOver.

Quality Control & Population Consolidation

For each dataset, individuals with greater than 5% missingness and variants greater than 2% missingness were removed from the analysis. When genetic ancestry profiles are available from the original publication, we manually grouped populations with similar ancestry profiles into larger populations and then removed populations with less than 20 individuals from downstream analyses. Across all of our studies, we assembled 123 populations with 20 or more individuals for analysis (**Supplemental Tables 1-2**). Once the population to be imputed is defined, we then filtered variants that failed the Hardy-Weinberg equilibrium ($p < 10^6$) per population. The number of variants remaining at each step of pre-imputation filtering is provided in **Supplemental Table 6**. Next, the microarray data were converted to variant call format (VCF) files by chromosome for submission to the imputation server. Only autosomal markers were imputed and analyzed. If > 10% of strand flips were identified during the imputation server quality control that prevented imputation from proceeding, then we correct the strand flips by comparing the two alleles for each variant against those recorded in the 1000 Genomes sequencing data. For each population, we estimated latitude and longitude using the Geocoding API feature in the Google Maps Platform and location descriptors in the original sources; latitude and longitude used to generate **Figure 2** and **Supplemental Figures 1-2, 5** are provided in **Supplemental Table 1**.

Matched European-Ancestry Population for Comparisons

For each target population we curated and processed for imputation, a European-ancestry population matched in sample size and SNP content was created using the UK10K dataset[31] (European Genome-phenome Archive accession numbers EGAD00001000740 and EGAD00001000741), which contained whole genome sequencing data from the ALSPAC and TwinsUK cohorts. We extracted a random subset of individuals from the UK10K dataset matched in sample size. All SNPs in the target population passing quality control for imputation and also found in the UK10K dataset were also extracted from UK10K for imputation. Note that not all SNPs may be found in the UK10K dataset, thus our comparison population for imputation quality based on the

European-ancestry population necessarily has equal or fewer SNPs used for input of imputation. However, this is expected to lower the quality of imputation for the European-ancestry population and reduce any disparity in imputation quality observed, if any.

TOPMed Imputation

Unphased populations were submitted to the TOPMed imputation server for further server-based quality control and imputation with the mixed population setting. The Rsq filtering was turned off and Eagle v2.4 (ref [32]) phasing was selected. The defaults for the remaining settings were selected. For each population, including the European controls, the imputation R-squared (Rsq) metric was extracted for each imputed variant from info files produced by the imputation software, Minimac 4, and used as one of the metrics of imputation efficacy. Distributions of Rsq were examined in strata of minor allele frequencies based on the estimated frequency from imputation.

Pearson's Correlation r^2

Whole genome sequencing datasets from the Human Genome Diversity Project[33] (HGDP) and The African American Sequencing Project from 23andMe (AASP, dbGAP accession number phs001798.v2.p2) cohort were gathered. For HGDP, we processed the previously released genome-wide array data[34] as described above for imputation. For the AASP cohort, we extracted the SNPs found on Omni 2.5M array to mimic an array dataset. For overlapping individuals and SNPs between the imputed data and sequencing data, we computed Pearson's correlation r^2 between the imputed dosages and the genotypes from sequencing data as a metric of imputation efficacy.

Fixation Index (Fst) Calculation

Each target population was merged with the matched European-ancestry control from UK10K, and the intersection of variants was kept. Variants were pruned to achieve approximate linkage equilibrium using PLINK[35] (v1.9, using the --indep-pairwise command with a window size of 50kb, step size 5, and r-squared threshold 0.1). Fst was calculated using the --fst command. Raw Fst score was used to estimate population differentiation.

Meta-Imputation

To assess meta-imputation, we imputed selected populations with Taiwan BioBank [21,22] (TWB) as an imputation reference in-house. Whole genome sequenced individuals from TWB were phased without a reference using Eagle (version 2.4v). Phased genotypes were then converted into the reference file format (.m3vcf files) with Minimac 3[23].

To perform meta-imputation, selected target populations were phased independently without a reference using Eagle 2.4v first (rather than with the TOPMed imputation

server), and imputed with TWB imputation reference panel in-house using Minimac4[23], or on the TOPMed Imputation Server. The imputation results for TWB and TOPmed were meta-imputed with MetaMinimac2[14].

To evaluate the improvement of imputation Rsq at East Asian-enriched alleles, we stratified each meta-imputed variant according to their estimated frequencies from the Genome Aggregation Database (gnomAD) non-Finnish-Europeans and East Asians (release 3.1.1; https://gnomad. broadinstitute.org/)

Acknowledgements

The authors would like to thank Xin Sheng and Amellia Marvit for comments and feedback of this manuscript. This work was supported by funding from the National Institutes of Health (R35GM142783 to C.W.K.C.). J.L.C. and X.R. were supported by the Viterbi Merit Fellowship and the Provost Fellowship, respectively, through the University of Southern California. In addition, this study benefited from a number of authors who kindly shared their data for analysis, including Georgi Hudjasov, Phillip Endicott, Meshari Alwazae, Fowzan Alkuraya, Dang Liu, Mark Stoneking, and David Reich; the original publications are cited within the text. This study also accessed several datasets through online repositories such as dbGaP. For these studies, we acknowledge the use of:

(1) Data generated by MalariaGEN. A full list of the investigators who contributed to the generation of the data is available from www.MalariaGEN.net. Funding for this project was provided by Wellcome Trust (WT077383/Z/05/Z) and the Bill & Melinda Gates Foundation through the Foundation of the National Institutes of Health (566) as part of the Grand Challenges in Global Health Initiative.

(2) Data generated by the Pacific Islands Rheumatic Heart Disease Genetics Network. A full list of the investigators who contributed to the generation of the data is available from www.rhdgenetics.net. Funding for the project was provided by the British Heart Foundation (PG/14/26/30509; www.bhf.org.uk), the Medical Research Council UK (Fellowship G1100449; www.mrc.ac.uk), and the British Medical Association (Josephine Lansdell Grant; www.bma.org.uk) and cite the relevant primary Network publication (details of which can be found at www.rhdgenetics.net).

(3) Data generated by the Genetic Risk to Stroke in Smokers and Non-Smokers in Multiple Ethnic Groups study, funding for which was provided through the NIH Genes, Environment and Health Initiative [GEI] (U01HG004436). The Genetic Risk to Stroke in Smokers and Non-Smokers in Multiple Ethnic Groups study is one of the genome-wide association studies funded as part of the Gene Environment Association Studies (GENEVA) under GEI. Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by the GENEVA Coordinating Center (U01 HG004446). Assistance with data cleaning was provided by the National Center for Biotechnology Information. Funding support for genotyping, which was performed at the Johns Hopkins University Center for Inherited Disease Research, was provided by the NIH GEI (U01HG004438) and the NIH contract "High throughput genotyping for studying the genetic contributions to human disease"(HHSN268200782096C). Support for the fieldwork in Pakistan was funded from the NINDS (R21NS064908) and support for genotyping of common controls was from the Wellcome Trust.

(4) Data generated by the Diabetes in Mexico Study SIGMA Exome Sequencing Project, which was supported by the Carlos Slim Foundation. The data from the Diabetes in Mexico Study SIGMA Exome Sequencing Project reported here were supplied by the Broad Institute and National Institute of Genomic Medicine (INMEGEN). This manuscript was not prepared in collaboration with Investigators of the Diabetes in Mexico Study SIGMA Exome Sequencing Project study and does not necessarily reflect the opinions or views of the Diabetes in Mexico Study SIGMA Exome Sequencing Project study, or the Carlos Slim Foundation.

(5) Data generated by the Stanford University Global Reference Panel data. Funding support for the Population Architecture Using Genomics and Epidemiology (PAGE) Global Reference Panel was provided through the National Human Genome Research Institute (U01HG007417). Assistance with phenotype harmonization, SNP selection, data cleaning, meta-analyses, data management and dissemination, and general study coordination, was provided by the PAGE Coordinating Center (U01HG007419, U01HG004801 and its NHGRI ARRA supplement). The datasets used for the analyses described in this manuscript were obtained from dbGaP.

(6) Data generated by the African American Sequencing Project, which was supported by the National Human Genome Research Institute of the National Institutes of Health (grants R44HG009460 and R43HG009090). We thank the research participants and employees of 23andMe for contributing to this study.

Finally, computation for this work was supported by the University of Southern California's Center for Advanced Research Computing (CARC; https://carc.usc.edu).

References

1. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics* **11**, 499–511 (2010).
2. Bai, W.-Y. *et al.* Genotype imputation and reference panel: a systematic evaluation on haplotype size and diversity. *Briefings in Bioinformatics* **21**, 1806–1817 (2020).
3. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).
4. Atutornu, J., Milne, R., Costa, A., Patch, C. & Middleton, A. Towards equitable and trustworthy genomics research. *eBioMedicine* **76**, 103879 (2022).
5. Need, A. C. & Goldstein, D. B. Next generation disparities in human genomics: concerns and remedies. *Trends Genet* **25**, 489–494 (2009).
6. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279–1283 (2016).
7. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
8. Kowalski, M. H. *et al.* Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genet* **15**, e1008500 (2019).
9. Chiang, C. W. K. The Opportunities and Challenges of Integrating Population Histories Into Genetic Studies for Diverse Populations: A Motivating Example From Native Hawaiians. *Front Genet* **12**, 643883 (2021).
10. Martin, A. R. *et al.* Increasing diversity in genomics requires investment in equitable partnerships and capacity building. *Nat Genet* **54**, 740–745 (2022).
11. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 1080 (2019).
12. Peterson, R. E. *et al.* Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. *Cell* **179**, 589–603 (2019).
13. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).
14. Yu, K. *et al.* Meta-imputation: An efficient method to combine genotype data after imputation with multiple reference panels. *The American Journal of Human Genetics* **109**, 1007–1015 (2022).
15. Moayyeri, A., Hammond, C. J., Valdes, A. M. & Spector, T. D. Cohort Profile: TwinsUK and healthy ageing twin study. *Int J Epidemiol* **42**, 76–85 (2013).
16. Mineta, K., Goto, K., Gojobori, T. & Alkuraya, F. S. Population structure of indigenous inhabitants of Arabia. *PLoS Genet* **17**, e1009210 (2021).

17. Malaria Genomic Epidemiology Network. Insights into malaria susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania. *Nat Commun* **10**, 5732 (2019).
18. Sun, Q. *et al.* MagicalRsq: Machine-learning-based genotype imputation quality calibration. *The American Journal of Human Genetics* **109**, 1986–1997 (2022).
19. Adair, L. S. *et al.* Cohort profile: the Cebu longitudinal health and nutrition survey. *Int J Epidemiol* **40**, 619–625 (2011).
20. Traylor, M. *et al.* Genetic risk factors for ischaemic stroke and its subtypes (the METASTROKE collaboration): a meta-analysis of genome-wide association studies. *Lancet Neurol* **11**, 951–962 (2012).
21. Feng, Y.-C. A. *et al.* Taiwan Biobank: A rich biomedical research database of the Taiwanese population. *Cell Genomics* **2**, 100197 (2022).
22. Wei, C.-Y. *et al.* Genetic profiles of 103,106 individuals in the Taiwan Biobank provide insights into the health and history of Han Chinese. *NPJ Genom Med* **6**, 10 (2021).
23. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
24. Micheletti, S. J. *et al.* Genetic Consequences of the Transatlantic Slave Trade in the Americas. *The American Journal of Human Genetics* **107**, 265–277 (2020).
25. Fortes-Lima, C. A. *et al.* The genetic legacy of the expansion of Bantu-speaking peoples in Africa. 2023.04.03.535432 Preprint at https://doi.org/10.1101/2023.04.03.535432 (2023).
26. Kim, M. S. *et al.* Testing the generalizability of ancestry-specific polygenic risk scores to predict prostate cancer in sub-Saharan Africa. *Genome Biology* **23**, 194 (2022).
27. Sun, H. *et al.* The impact of global and local Polynesian genetic ancestry on complex traits in Native Hawaiians. *PLoS Genet* **17**, e1009273 (2021).
28. Chen, C.-H. *et al.* Population structure of Han Chinese in the modern Taiwanese population based on 10,000 participants in the Taiwan Biobank project. *Hum Mol Genet* **25**, 5321–5331 (2016).
29. Wall, J. D. *et al.* The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* **576**, 106–111 (2019).
30. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284–1287 (2016).
31. Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
32. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* **48**, 1443–1448 (2016).
33. Bergström, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, eaay5012 (2020).

34. Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104 (2008).
35. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaSci* **4**, 7 (2015).