A harmonized public resource of deeply sequenced diverse human genomes

Zan Koenig[1,2,*], Mary T. Yohannes[1,2,*], Lethukuthula L. Nkambule[1,2], Julia K. Goodrich[2,3], Heesu Ally Kim[1], Xuefang Zhao[3,4,5], Michael W. Wilson[3], Grace Tiao[3], Stephanie P. Hao[3,4,5], Nareh Sahakian[6], Katherine R. Chao[3], gnomAD Project Consortium, Heidi L. Rehm[3,4], Benjamin M. Neale[1,2,3], Michael E. Talkowski[1,3,4,5], Mark J. Daly[1,2,3,7], Harrison Brand[3,4,5], Konrad J. Karczewski[1,2,3,4], Elizabeth G. Atkinson[1,2,8], Alicia R. Martin[1,2,3,4,†]

[1] Stanley Center for Psychiatric Research, The Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[2] Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA
[3] Program in Medical and Population Genetics, The Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[4] Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA
[5] Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA
[6] Broad Genomics, The Broad Institute of MIT and Harvard, 320 Charles Street, Cambridge, MA, 02141, USA
[7] Institute for Molecular Medicine Finland, Helsinki, Finland
[8] Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, 77030, USA
*These authors contributed equally
[†] Corresponding author: armartin@broadinstitute.org

Abstract

Underrepresented populations are often excluded from genomic studies due in part to a lack of resources supporting their analyses. The 1000 Genomes Project (1kGP) and Human Genome Diversity Project (HGDP), which have recently been sequenced to high coverage, are valuable genomic resources because of the global diversity they capture and their open data sharing policies. Here, we harmonized a high quality set of 4,096 whole genomes from HGDP and 1kGP with data from gnomAD and identified over 159 million high-quality SNVs, indels, and SVs. We performed a detailed ancestry analysis of this cohort, characterizing population structure and patterns of admixture across populations, analyzing site frequency spectra, and measuring variant counts at global and subcontinental levels. We also demonstrate substantial added value from this dataset compared to the prior versions of the component resources, typically combined via liftover and variant intersection; for example, we catalog millions of new genetic variants, mostly rare, compared to previous releases. In addition to unrestricted individual-level public release, we provide detailed tutorials for conducting many of the most common quality control steps and analyses with these data in a scalable cloud-computing environment and publicly release this new phased joint callset for use as a haplotype resource in phasing and imputation pipelines. This jointly called reference panel will serve as a key resource to support research of diverse ancestry populations.

Introduction

The 1000 Genomes Project (1kGP) and Human Genome Diversity Project (HGDP) have been among the most valuable genomic resources because of the breadth of global diversity they capture and their open sharing policies with consent to release unrestricted individual-level data (Bergström et al. 2020; Rosenberg et al. 2002; Li et al. 2008; 1000 Genomes Project Consortium et al. 2015, 2012). Consequently, genetic data from these resources have been routinely generated using the latest genomics technologies and serve as a ubiquitous resource of globally diverse populations for a wide range of disease, evolutionary, and technical studies. These projects are complementary; the 1000 Genomes Project is larger and has consisted of whole genome sequencing (WGS) data for many years; as such, it has been the default population genetic reference dataset, consisting of 3,202 genomes including related individuals that were recently sequenced to high coverage (Ebert et al. 2020; Byrska-Bishop et al. 2022). The 1000 Genomes Project has also been the most widely used haplotype resource, serving as a reference panel for phasing and imputation of genotype data for many genome-wide association studies (GWAS) (Lam et al. 2020; Howie et al. 2012). HGDP was founded three decades ago by population geneticists to study human genetic variation and evolution, and was designed to span a greater breadth of diversity, though with fewer individuals from each component population (Cavalli-Sforza et al. 1991; Cavalli-Sforza 2005). Originally assayed using only GWAS array data, the 948 individuals have recently undergone deep WGS and fill some major geographic gaps not represented in the 1000 Genomes Project, for example in the Middle East, sub-Saharan Africa, parts of the Americas, and Oceania (Bergström et al. 2020).

The 1kGP and HGDP datasets have been invaluable separately, but far larger genomic data aggregation efforts, such as gnomAD (Karczewski et al. 2020) and TOPMed (Taliun et al. 2021), have clearly demonstrated the utility of harmonizing such datasets through the broad uptake of their publicly released summaries of large numbers of high-quality whole genomes. For example, the gnomAD browser of allele frequencies has vastly improved clinical interpretation of rare disease patients worldwide (Karczewski et al. 2017). Additionally, the TOPMed Imputation Server facilitates statistical genetic analyses of complex traits by improving phasing and imputation accuracy compared to existing resources (Taliun et al. 2021). Yet, without individual-level data access from these larger resources due to more restrictive permissions, the 1kGP and HGDP genomes remain the most uniquely valuable resources for many of the most common genetic analyses. These include genetic simulations, ancestry analysis including local ancestry inference (Maples et al. 2013), genotype refinement of low-coverage genomes (Rubinacci et al. 2021), granular allele frequency comparisons at the subcontinental level, investigations of individual-level sequencing quality metrics, and many more.

Previously, researchers wishing to combine HGDP and 1kGP into a merged dataset were left with suboptimal solutions. Specifically, the sequenced datasets had been called separately, requiring intersection of previously called sites rather than a harmonized joint-callset. Additionally, they were on different reference builds, requiring lifting over of a large dataset prior to merging, which introduces errors and inconsistencies. Here, we have created a best-in-class publicly released harmonized and jointly called resource of HGDP+1kGP on GRCh38 that will facilitate analyses of diverse cohorts. This globally-representative haplotype resource better captures the breadth of genetic variation across diverse geographical regions than previous component studies. Specifically, we aggregated these genomes into gnomAD and then jointly processed these 4,096 high-coverage whole genomes; jointly called variants consisting of single nucleotide variants (SNVs), insertions/deletions (indels), and structural variants (SVs); conducted harmonized sample and variant QC; and separately released these individual-level genomes to facilitate a wide breadth of analyses. We quantify the number of variants identified in this new callset compared to existing releases and identify more variants as a result of joint variant calling; construct a resource of haplotypes for use as a phasing and imputation panel; examine the ancestry composition of this diverse set of populations; and publicly release these data without restriction alongside detailed tutorials illustrating how to conduct many of the most common genomic analyses.

Results

A harmonized resource of high-quality, high coverage diverse whole genomes

Here, we have developed a high-quality resource of diverse human genomes for full individual-level public release along with a guide for conducting the most common genetic analyses. To this end, we first harmonized project meta-data and jointly called variants from 4,150 whole genomes recently sequenced to high coverage from the 1kGP and HGDP into gnomAD (**Table S1**) (Bergström et al. 2020; Byrska-Bishop et al. 2022), the latter of which are new to gnomAD. **Figure 1A** shows the locations and sample sizes of populations included in this harmonized resource. After sample, variant, and genotype QC (Chen et al. 2022), including ancestry outlier removal (**Table S2, Methods**), we identified 159,795,273 high-quality variants across 4,096 individuals, 3,378 of whom are inferred to be unrelated (**Methods, Table S3**). We computed the mean coverage within each population and project (**Figure S1-2**) as well as the mean number of SNVs per individual within each population to better understand data quality and population genetic variation (**Table S4**). While coverage was more variable among samples in HGDP (μ=34, σ=6, range=23-75X) than in 1kGP (μ=32, σ=3, range=26-66X), consistent with older samples and more variable data generation strategies (Bergström et al. 2020), all genomes had sufficient coverage to perform population genetic analysis. Consistent with human population history and as seen before (1000 Genomes Project Consortium et al. 2015), African populations had the most genetic variation with 6.1M SNVs per individual, while out-of-Africa populations had an average of 5.3M SNVs (**Table S4, Figure 1B**). The San had the most genetic variants as well as singletons per genome on average overall (**Table S4**). This is consistent with previous studies, which showed that the San had the highest genetic variation of populations studied to date, likely explained by their history traditionally as hunter-gatherers who experienced no major bottlenecks out of Africa or within Africa (Schlebusch et al. 2017).

We generated a jointly genotyped structural variants (SVs) callset by detecting new SVs in the HGDP genomes (**Figure S3**) using the same ensemble SV detection tool, GATK-SV (Collins et al. 2020), as used to detect existing SVs called in the high-coverage 1kGP genomes (Byrska-Bishop et al. 2022). We then combined SVs to form a non-redundant SV set uniformly genotyped across all samples (**Figure S4**). In total, we identified 196,173 SV loci across all 4,150 HGDP and 1kGP samples. We detected a median of 8,123 SVs in each genome consisting primarily of deletions, duplications, and insertions (**Figure 1**). This is comparable sensitivity to previous studies of high-coverage whole genomes (~9,679 SVs / genome in (Byrska-Bishop et al. 2022) and ~7,439 SVs / genome in (Collins et al. 2020)) and higher than in low-coverage whole genomes, as expected (~3,431 SVs / genome in (Sudmant et al. 2015)). The frequencies of SVs were also consistent with Hardy-Weinberg Equilibrium as expected (**Figure S4**), and distributions matched expectations from previous cohorts with the vast majority of SVs being rare (84.2% SVs are <1% allele frequency among population). Additionally, SV size is inversely correlated with frequency (Sudmant et al. 2015; Byrska-Bishop et al. 2022; Collins et al. 2020), with notable exceptions of peaks consistent with known mobile elements, including ALU, LINE1, and SVA (**Figure 1C**). Consistent with shorter genetic variation, we observed a higher frequency of SVs in African populations. The quality of our variant call sets have been evaluated using both the short-read and long-read WGS data generated by the 1kGP and the human genome structural variation consortium (HGSVC, (Ebert et al. 2021;

Byrska-Bishop et al. 2022)). High precision was observed in the SV call set–among the 34 overlapping samples, 91.9% of the SVs were overlapped by either a short-read or long-read variant in the matched genome; the highest precision (97.6%) was observed for deletions followed by insertions (91.4%) and duplications (89.3%) (**Table S6**). We observed some differences in number SVs across samples from HGDP and 1kGP due to technical data generation differences, such as PCR status (**Figure S6**).

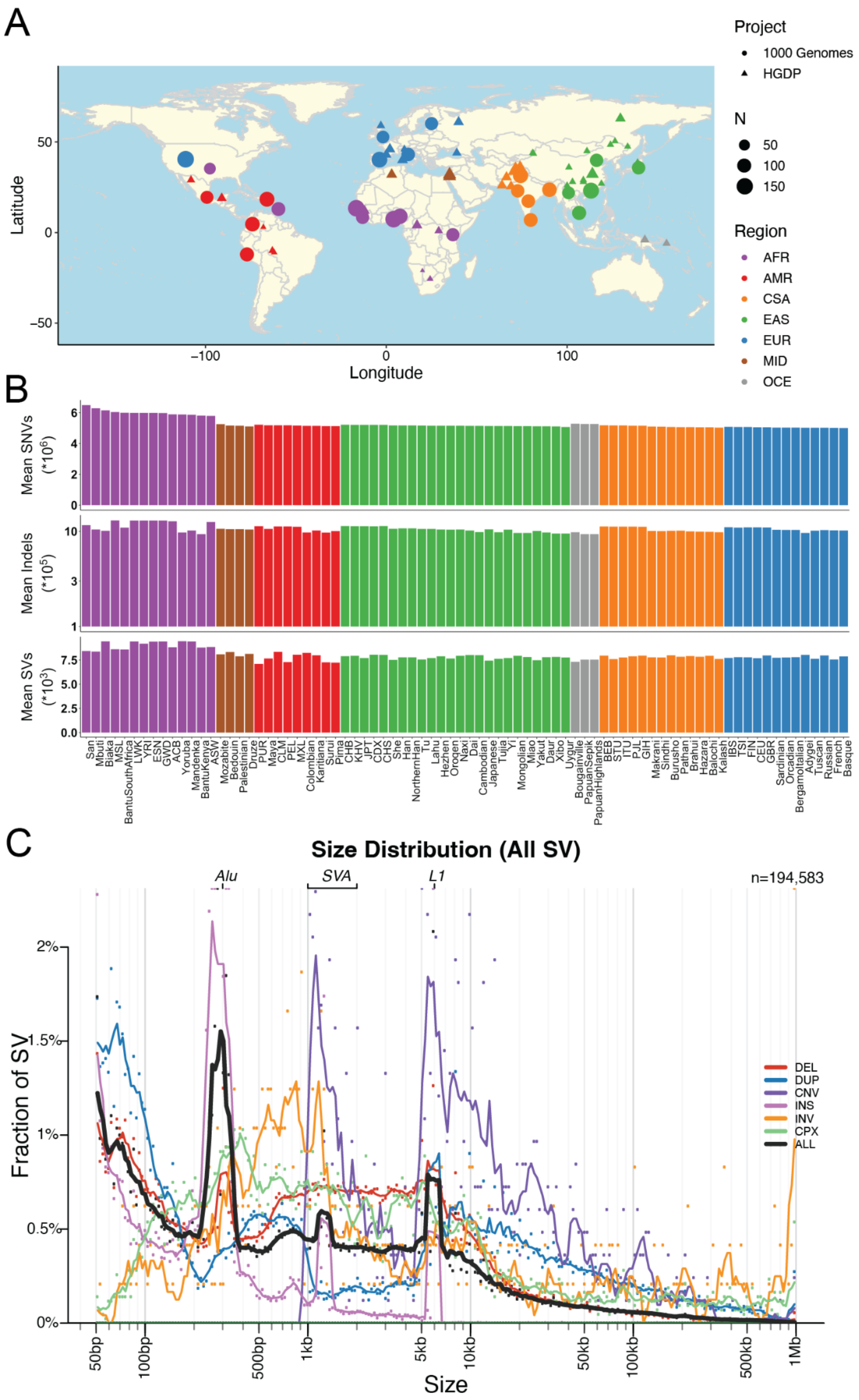


Figure 1 | Geographical locations and genetic variants across populations.

A) Global map indicating approximate geographical locations where samples were collected.

Coordinates were included for each population originating from the Geography of Genetic

Variants browser as well as meta-data from the HGDP (Bergström et al. 2020; Marcus and Novembre 2017). B) Mean number of SNVs (top), indels (middle), and SVs (bottom) per individual within each population. Colors are consistent with geographical/genetic regions in A-B), as follows: AFR=African, AMR=admixed American, CSA=Central/South Asian, EAS=East Asian, EUR=European, MID=Middle Eastern, OCE=Oceanian. C) Sizes of SVs decay in frequency with increasing size overall with notable exceptions of mobile elements, including Alu, SVA, and LINE1. Abbreviations are deletion (DEL), duplication (DUP), copy number variant (CNV), insertion (INS), inversion (INV), or complex rearrangement (CPX).

We examined global population genetic variation using principal component analysis (PCA) of the harmonized HGDP and 1kGP resource (**Figure 2**). As expected, we find PC1 differentiates AFR and non-AFR populations, PC2 differentiates EUR and EAS populations, and PC3-4 differentiate AMR and CSA populations. Subcontinental structure is also apparent in later PCs and within geographical/genetic regions (**Table S1, Figure S9-16**). These results are recapitulated with the likelihood model implemented in ADMIXTURE, where K=2 identifies similar structure in PC1, K=3 identifies similar structure in PC2, and so on (**Figure S7**). The best fit value of K=6 shown in **Figure 2** was chosen based on 5-fold cross-validation error (**Figure S8**).

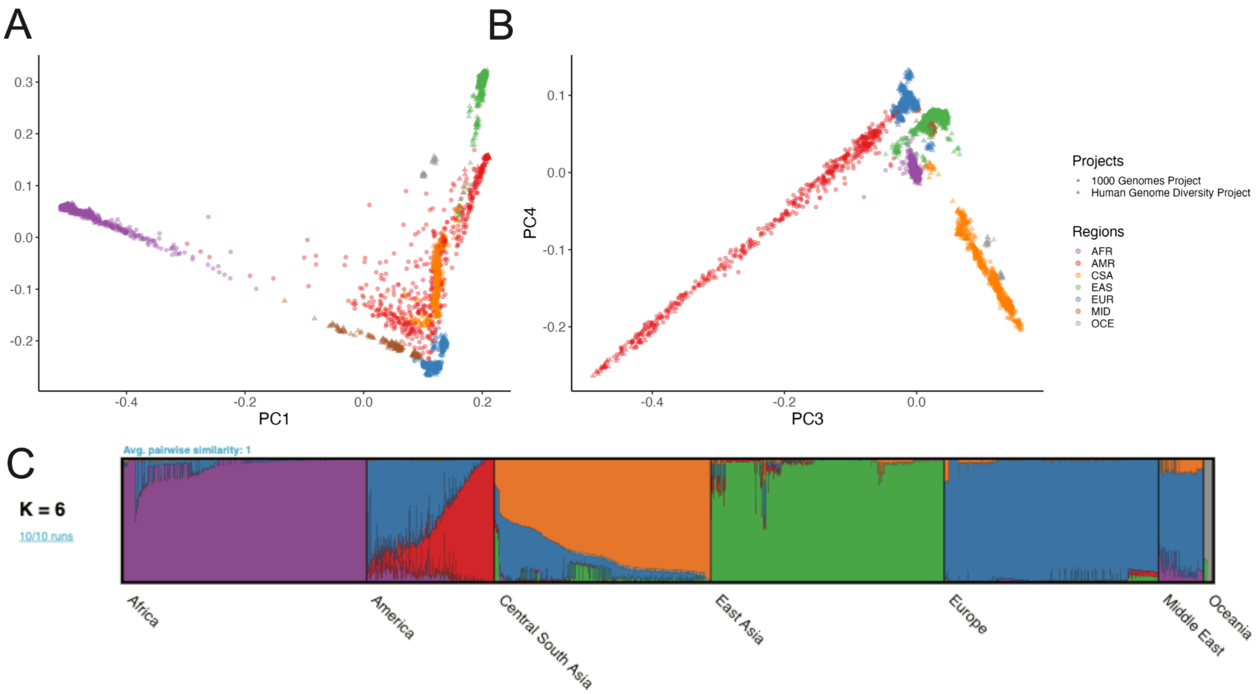


Figure 2 | Global ancestry analysis of genetic structure in the HGDP and 1kGP resource. Regional abbreviations are as in Figure 1. A-B) Principal components analysis (PCA) plots for A) PC1 versus PC2 and B) PC3 versus PC4 showing global ancestry structure across HGDP+1kGP. Subsequent PCs separated structure within geographical/genetic regions (**Figure S9-S16**). C) ADMIXTURE analysis at the best fit value of K=6.

Population genetic variation within and between subcontinental populations

We investigated the ancestry composition of populations within harmonized meta-data labels (AFR, AMR, CSA, EAS, EUR, MID, and OCE; **Table S1**) using principal component analysis (PCA) and ADMIXTURE analysis. Subcontinental PCA highlights finer scale structure within geographical/genetic regions (**Figure S9-S16**). For example, within the AFR, the first several PCs differentiate populations from South and Central African hunter-gatherer groups from others, then differentiate populations from East and West Africa (**Figure S10**). For AFR and AMR populations, individuals cluster similarly to the global PCA, reflecting some global admixture present in these populations (**Figure S10, S14**). The MID and OCE populations are only made up of samples from the HGDP dataset, as 1kGP did not contain samples from these regions (**Figure S15-16**).

We measured population genetic differentiation using common variants with Wright's fixation index, F_{ST} (**Figure 3A**). When populations are clustered according to pairwise F_{ST} between groups, they largely cluster by geographical/genetic region labels with a few exceptions (**Figure S18**). For example, AMR populations are interspersed with other populations, consistent with having variable ancestry proportions that span multiple continents. Additionally, MID populations are interspersed among the EUR populations (Bedouin and Palestinian cluster together while Mozabite and Druze cluster by themselves). A CSA population, Kalash, clusters among the EUR, and an EAS population, Uygur, clusters among the CSA. The OCE populations cluster together. There are no interspersed populations within the AFR cluster and no populations from AFR are interspersed among the other regions (**Table S7**).

We also compared F_{ST} versus geographical distance. We computed great circle distances using the haversine formula and pairwise geographic distances using five waypoints that reflect human migration patterns, recapitulating previous work (Ramachandran et al. 2005). The linear relationship between F_{ST} and geographical distance differs by project; specifically, HGDP has a steeper slope relating distance to F_{ST} (**Figure 3**), likely reflecting the anthropological design intended to capture more divergent populations compared to the samples in 1kGP that reflect some of the largest populations. We compared Pearson's correlation and Mantel tests to assess the change in the linear relationship between F_{ST} and geographical distance when incorporating waypoints. The Pearson's correlation coefficient and Mantel statistic are both higher when waypoints are incorporated, with the highest values being when both pairs of populations are from HGDP with a correlation coefficient of 0.76 (p-value < 2.2e-16) and Mantel statistic of 0.55, p-value: 0.01 (**Table S8**).

F_{ST} measurements require group comparisons and are only based on common variants, which typically arose early in human history. We also compared rare variant sharing via pairwise doubleton counts (f_2 analyses, **Figure 3A**). On average, pairs of individuals within a population share 51.62 doubletons, although this varies considerably as a function of demography. For example, due to the elevated number of variants in individuals of African descent (**Figure 1**), pairs of individuals within AFR populations share on average 76.38 doubletons, whereas pairs of individuals within out-of-Africa populations share 43.44 doubletons. The individual pairs that shared the most doubletons were largely from the San population, with the top 15 sharing between 14,717 and 24,374 doubletons. Very few doubletons are shared among pairs of individuals across populations within a geographical/genetic region (u=6.84, σ=19.1) with the highest doubleton count being 4,068 between individuals from BantuSouthAfrica and San, both AFR populations. Even fewer doubletons are shared among pairs of individuals across populations from different geographical/genetic regions (u=0.8, σ=1.78) with the highest doubleton count being 638 between a pair of individuals from CDX and BEB which are EAS and CSA populations, respectively. f_2 clustering tends to follow project meta-data labels by geographical/genetic region, with a few exceptions.

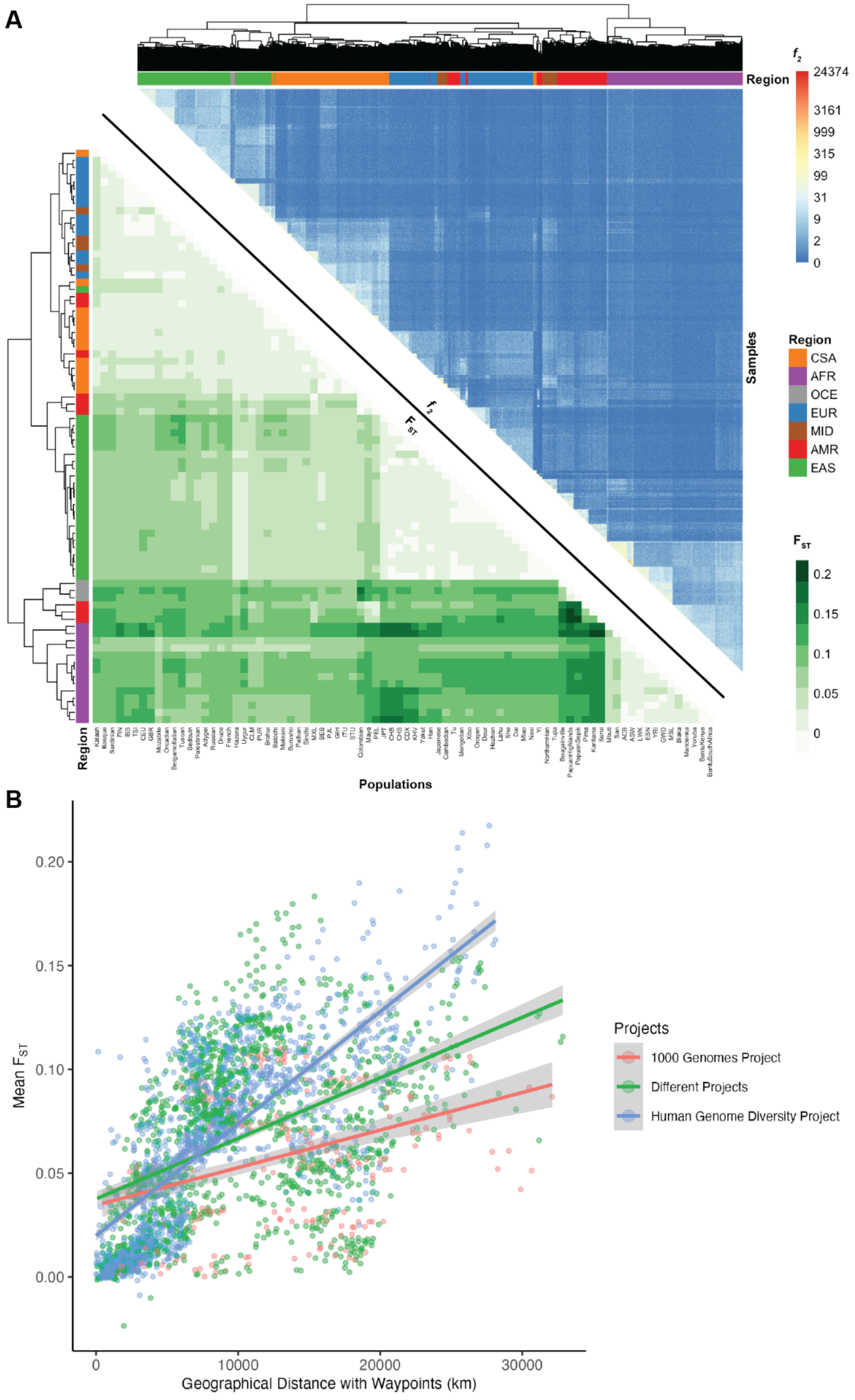


Figure 3 | Relationships between genetic differentiation measured from common variants (F_{ST}), rare variants (f_2), and geography.

A) Lower triangle: F_{ST} heatmap illustrating genetic divergence between pairs of populations.

Upper triangle: Heatmap of f_2 comparisons of doubleton counts between pairs of individuals.

Column and row colors at the leaves of the dendrogram show colors corresponding to meta-data geographical/genetic region and top right color bar indicates the number of doubletons shared across pairs of individuals, with more doubletons shared among individuals within the same versus across populations and geographical/genetic regions. B) Genetic divergence measured by F_{ST} versus geographical distance with five waypoints calculated using haversine formula.

A catalog of known versus novel genomic variation compared to existing datasets

To demonstrate the added benefit of jointly calling these two datasets, we have compiled metrics that compare our harmonized dataset with each individual dataset comprising it (Bergström et al. 2020; Byrska-Bishop et al. 2022), the previous phase 3 1kGP dataset sequenced to lower coverage (1000 Genomes Project Consortium et al. 2015), and the widely used gnomAD dataset (Chen et al. 2022). This jointly called HGDP+1kGP dataset contains 159,795,273 SNVs and indels that pass QC, whereas phase 3 1kGP has 73,257,633, high-coverage WGS of 1kGP (referred to here as NYGC 1kGP based on where they were sequenced) has 119,895,186, high-coverage WGS of HGDP (referred to here as Bergstrom HGDP based on the publication) has 75,310,370, and gnomAD has 644,267,978 high-quality SNVs and indels (Chen et al. 2022). Because gnomAD now contains both HGDP and 1kGP, we built a synthetic subset of gnomAD that removes allele counts contributed by HGDP and 1kGP. When comparing the HGDP+1kGP dataset to this synthetic version of gnomAD that excludes HGDP+1kGP, we show that variants unique to gnomAD are disproportionately rare (**Figure 4**). In contrast, compared to the comprising datasets of HGDP only, the NYGC 1kGP only, and phase 3 1kGP, the HGDP+1kGP dataset uniquely contributes a sizable fraction and number of variants spanning the full allele frequency spectrum, including both rare and common variants

(**Figure 4**). However, rare variants are particularly enriched; in all of the comparison datasets aside from gnomAD, the HGDP+1kGP dataset contains the largest proportion of rare variants. Few variants in the phase3 1kGP dataset were not in the HGDP+1kGP dataset or NYGC 1kGP because samples are entirely overlapping, as reported previously (Byrska-Bishop et al. 2022).

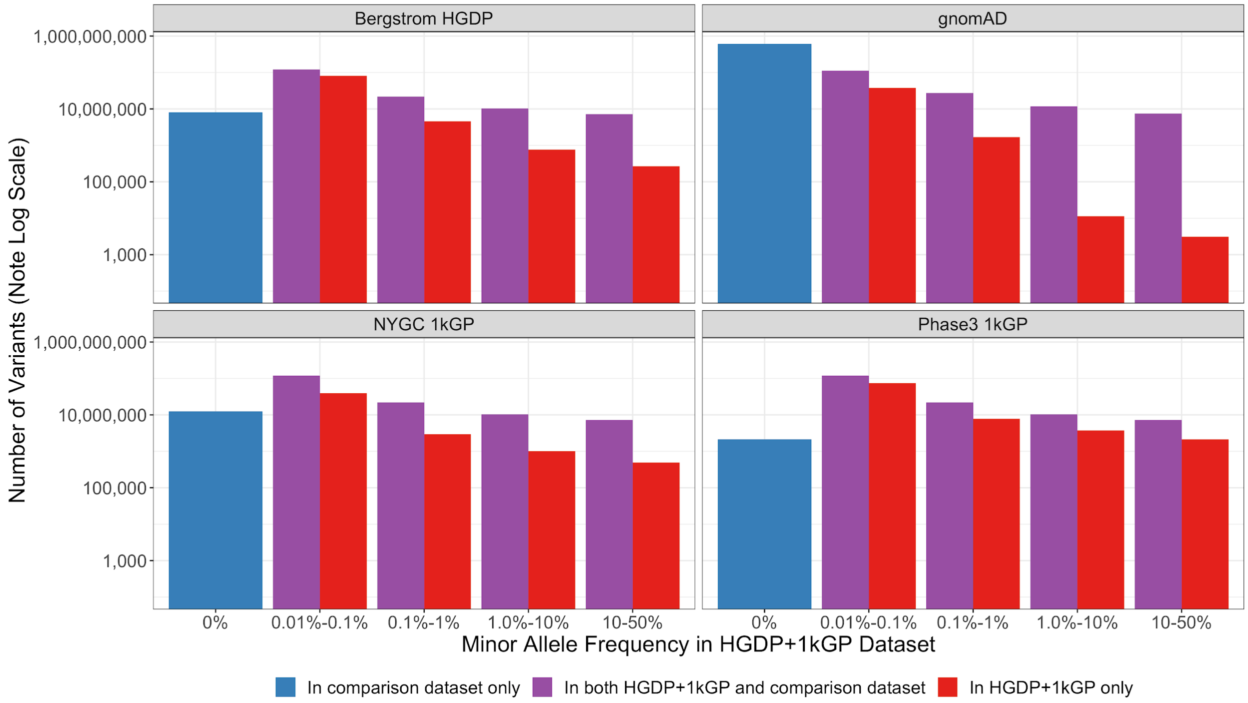


Figure 4 | Number of variants identified in this dataset compared to commonly used existing datasets as a function of allele frequency.

The number of variants on a log scale is plotted by minor allele frequency bin within the harmonized HGDP+1kGP dataset. We show variants found in the harmonized HGDP+1kGP dataset only (red), variants shared between the harmonized dataset and each comparison dataset (purple), and variants that are only found in each comparison dataset (blue).

Phased haplotypes improve imputation accuracy and flexibility compared to existing public resources

We next developed the HGDP+1kGP dataset as a haplotype resource by phasing variants together using SHAPEIT5 (Hofmeister et al. 2023), including information about trios (**Methods**). To evaluate imputation accuracy with common genetic data generation strategies, we used 93 downsampled whole genomes sequenced as part of the NeuroGAP Project consisting of East and South African participants to either 1) sites on relevant and commonly used GWAS arrays (Illumina GSA, MEGA, and H3Africa), or 2) lower depths of coverage (0.5X, 1X, 2X, and 4X), as previously (Martin et al. 2021). We imputed GWAS arrays using IMPUTE5 (Rubinacci et al. 2020) and low-coverage genomes using GLIMPSE (Rubinacci et al. 2021). We compared average imputation accuracy as a function of allele frequency estimated from gnomAD AFR frequency given our small sample size and with several haplotype reference panels. For arrays, we compared imputation accuracy using NYGC 1kGP, HGDP+1kGP, and TOPMed imputation panels (Kowalski et al. 2019). Because low-coverage genomes require individual-level haplotypes, we were unable to compare the TOPMed panel. As expected, HGDP+1kGP improves accuracy compared to 1kGP but not compared to the much larger TOPMed data (**Figure 5**). For low-coverage genomes, we find much higher imputation accuracies with HGDP+1kGP. At rarer variants, the imputation accuracy differences due to reference panel used are almost as high as those due to higher depths of sequencing; for example, rare variants sequenced to 2X depth and imputed with HGDP+1kGP are imputed almost as accurately as rare variants sequenced to 4X depth and imputed with 1kGP, thus highlighting the utility of the larger sample size and diversity in this resource (**Figure 5**).

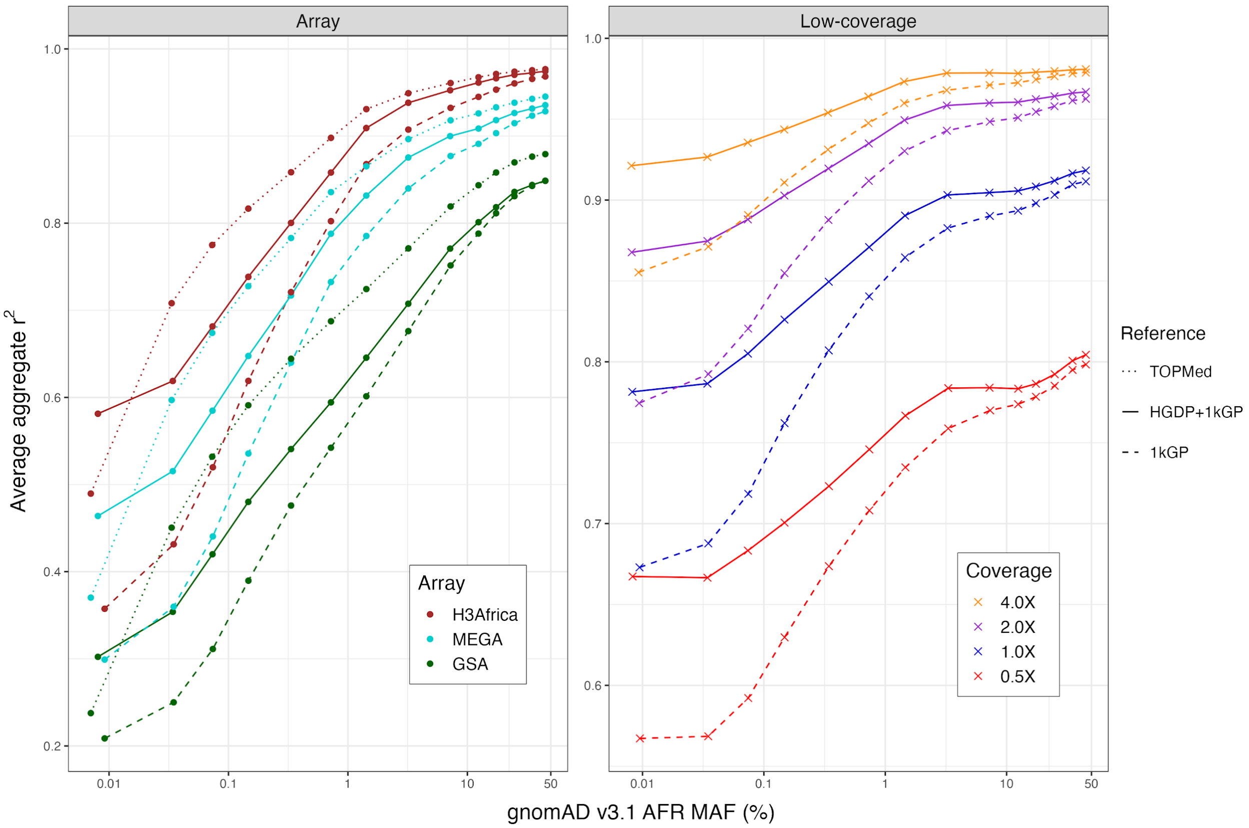
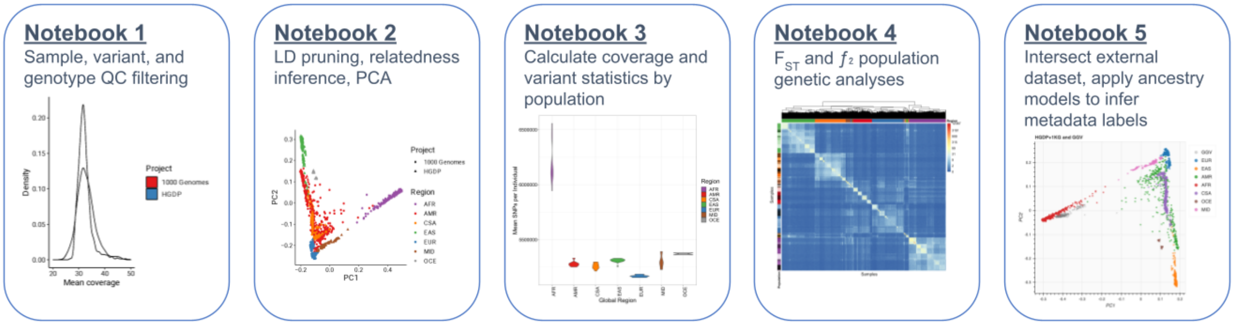


Figure 5 | SNP imputation performance in array and low-coverage datasets for three different reference panels using a validation set of 93 AFR individuals sequenced at 30X. SNP array and low-coverage sequencing data imputation performance for three reference panels as a function of minor allele frequency (MAF) for AFR in gnomAD v3.1 data. Aggregate r^2, which is the correlation between the imputed dosages and validation genotype calls, was computed in MAF bins and averaged across chromosomes 1-22. The validation set is 93 AFR individuals sequenced at 30X coverage (Martin et al. 2021).

Facilitating broad uptake of HGDP+1kGP as a public resource via development of detailed tutorials

In an effort to increase accessibility of this dataset, we have made publicly available tutorials of our analyses implemented primarily in Hail (https://hail.is/). Hail is an open source, Python-based, scalable tool for genomics that enables large-scale genetic analyses on the cloud. Tutorials can be accessed through Github via iPython notebooks (https://github.com/atgu/hgdp_tgp/tree/master/tutorials), and all underlying datasets are publicly available in requester-pays Google Cloud Platform buckets.

These tutorials cover various aspects of quality control (QC) and analysis, including sample and variant QC; visualizing distributions of QC statistics by metadata labels across diverse populations; filtering variants using LD, allele frequency, and missingness information; inferring relatedness; running PCA to infer ancestry; computing descriptive statistics including variant counts and coverage metrics; conducting population genetic analyses; and intersecting external datasets with HGDP+1kGP as a reference panel to apply ancestry models and infer metadata labels (**Figure 6**). For example, we intersected the publicly available Gambian Genome Variation (GGV) Project sequenced to low coverage with the HGDP+1kGP resource, trained a random forest on HGDP+1kGP geographical/genetic region meta-data labels, then applied this model to the GGV data to determine ancestry labels, which were all inferred to be AFR (**Figure S10**). When intersecting external datasets to apply ancestry labels, an important consideration is how many variants must overlap and how much missingness is tolerated to project external samples into the same PCA space as the reference panel and assign metadata labels given PCA shrinkage (Dey and Lee 2019). We find that <5% missingness is typically required to accurately assign ancestry labels (**Figure S20** and **Table S9**). In addition to all these analyses, we anticipate that there will be additional uses of this resource not documented in these tutorials, such as for phasing and imputation. To facilitate these uses, we have phased the HGDP+1kGP dataset and released these phased haplotypes that others can use to support phasing and imputation in their own datasets. We have also developed computational pipelines implemented in GWASpy that use these phased reference haplotypes, and tested these tools by applying phasing and imputation to diverse samples genotyped as part of other ongoing work.



Notebook 1
Sample, variant, and genotype QC filtering

Notebook 2
LD pruning, relatedness inference, PCA

Notebook 3
Calculate coverage and variant statistics by population

Notebook 4
F_{ST} and f_2 population genetic analyses

Notebook 5
Intersect external dataset, apply ancestry models to infer metadata labels

Figure 6 | Overview of tutorials that use cloud computing to conduct common genetic data analyses.

We have developed five iPython notebooks with tutorials for conducting many of the most common genetic analyses, including QC of sequencing data, relatedness inference and PCA, calculating statistics by population, analyzing genetic divergence, and applying ancestry analysis to a new dataset using HGDP+1kGP as a reference panel.

Discussion

The 1000 Genomes Project and Human Genome Diversity Project were landmark efforts to increase the unrestricted public availability of genomic data from a geographically and ancestrally diverse set of individuals. These resources have been widely used across research efforts for decades, including as reference panels for ancestry inference, phasing, imputation, genotype refinement, and investigations into population history and demography. However, these datasets have historically been discrete, leading to suboptimal intersections when a combined analysis of all samples is required.

The harmonized variant processing, quality control, and improved coverage of variants across the allele frequency spectrum in this jointly called resource will facilitate the improved study of diverse populations. Due to our rapid release of the data pre-publication, the callset formally released here has already been used as a resource of global diversity in the Genome Aggregation Database (gnomAD) (Chen et al. 2022), the Pan-UK Biobank Project (Karczewski et al.), the Global Biobank Meta-analysis Initiative (GBMI) (Zhou et al. 2022), and the Covid-19 Host Genetics Initiative (The COVID-19 Host Genetics Initiative and Ganna 2021). A primary use of this data is as a global reference for principal components analysis (PCA)--SNV loadings are freely shared so that user cohorts can be aligned to the same PC space as this optimized reference panel. In GBMI, harmonizing ancestry analysis with this resource served as a quality control measure to ensure that ancestral groupings are being applied consistently and that control for population stratification is being performed adequately (Zhou et al. 2022). Building on this approach and given the critical need for greater diversity in genomic studies, sequencing centers can use this resource in variant calling production pipelines to build dashboards that continuously monitor the diversity of samples being sequenced in real time.

This callset is also phased for use as a haplotype resource, potentially providing higher phasing and imputation accuracy particularly for underrepresented populations. While resources such as the Haplotype Reference Consortium (HRC) and TOPMed Imputation Panel are already useful (McCarthy et al. 2016; Kowalski et al. 2019), they either provide individual-level data but lack diversity (HRC) or are very large with significant diversity but do not share individual-level data (TOPMed). This limits the application of new methods, such as those needed to support low-coverage sequencing, which is receiving growing interest as it is comparable in cost to many genotype arrays and is especially beneficial to underrepresented populations (Martin et al. 2021). Combinations of high-coverage exome and low-coverage genome sequencing are also of growing interest and could be uniquely supported by this resource. This resource will also be critical for developing computational and analytical tools for genotype refinement, imputation, conducting data QC particularly across varying depths of coverage, and evaluating technical biases. For example, we observed fewer SVs in the HGDP genomes than 1kGP genomes among similar ancestry groups, which was primarily explained by PCR+ and PCR-free sequencing libraries.

This resource also provides a more complete and granular capture of the full spectrum of variation across the world that would be missed by intersecting the component datasets. Because a variant's frequency is one of its most informative features of its deleteriousness, the globally diverse allele frequencies that we have released on the gnomAD browser (Karczewski et al. 2017) provides additional scientific benefits by facilitating clinical variant interpretation across diverse populations. This GRCh38 release of this resource along with detailed tutorials for many of the most common genomic data analyses will also reduce barriers acknowledged by clinical labs which have not yet migrated to the latest genome build, citing that they do not feel the benefits outweigh the time and monetary costs and/or lack sufficient personnel to do so (Lansdon et al. 2021).

While this resource is more globally representative than many existing public datasets, certain geographic areas and ancestries are still underrepresented; for example, most genomic resources are enriched for participants in high-income countries (Fatumo et al. 2022) and there is particularly sparse coverage in central and southern Africa where genetic diversity is among the highest in the world. Some efforts that are already significantly underway, such as the H3Africa Initiative, will be critical for increasing representation from some of these ancestries. HGDP, designed over two decades ago alongside the Human Genome Project, was one of the earliest studies of its kind and therefore faced some ethical controversies, some of which remain relevant today (Resnik 1999; Greely 1999). For example, challenging issues of individual versus collective consent particularly among Amerindigenous communities parallel those currently being navigated by the *All of Us* Research Program; while criticisms have been raised, consensus has not been reached. HGDP responded to similar criticisms at the time, and developed the Model Ethical Protocol whose principles still guide all major genetic research projects to date (Weiss et al. 1997). The risk and beneficence of ongoing massive-scale efforts such as *All of Us*, whose mission is “to accelerate health research and medical breakthroughs, enable individualized prevention, treatment, and care for all of us” must be wrestled with to minimize risks and ensure adequate representativeness such that all can benefit from genomics research and ultimately precision medicine.

As genetically diverse datasets continue to grow to massive scales, it will be invaluable for researchers to be equipped with tools and resources that facilitate scalable,efficient, and equitable analysis. In the service of this goal, we concurrently release a series of detailed tutorials designed to be easily accessible in iPython notebooks demonstrating many common genomic analytic techniques as implemented in the cloud-native Hail software framework, which allows for flexible, computationally efficient, and parallelized analysis of big data. These tutorials lower the barrier for adoption of this resource and provide a code bank for researchers to conduct a variety of analyses, including conducting quality control of whole genome sequencing data, calculating variant and sample statistics within groups, analyzing population genetic variation, and applying ancestry labels from a reference panel to their own data. Overall, resources like this are essential for empowering genetic studies in diverse populations.

Methods

Genetic datasets

Human Genome Diversity Project (HGDP)

HGDP genomes sequenced and described previously (Bergström et al. 2020) were downloaded from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGDP/. Because the publicly available gVCFs were not the output of GATK HaplotypeCaller and were incompatible with joint calling, we reprocessed these genomes and conducted joint variant calling as part of gnomAD v3 (Chen et al. 2022). Most HGDP genomes were PCR-free (N=760), but some included PCR prior to sequencing (N=161). They were also sequenced at different times, for example as part of the Simons Genome Diversity Project (SGDP, N=120) or later at the Sanger Institute (N=801). More details are available from the source studies (Bergström et al. 2020; Mallick et al. 2016).

1000 Genomes Project (1kGP)

1kGP genomes have been sequenced as part of multiple efforts, first to mid-coverage as phase 3 of the 1kGP (1000 Genomes Project Consortium et al. 2015) and more recently to high-coverage (≥30X) at the New York Genome Center (NYGC) (Byrska-Bishop et al. 2022). We used the phase 3 1kGP genomes only for comparison to previous releases. The high-coverage 1kGP genomes sequenced at the NYGC were downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20201028_3202_raw_GT_with_annot/, which were harmonized with HGDP genomes to generate the HGDP+1kGP call set.

Human Genome Structural Variation Consortium (HGSVC)

The HGSVC generated high-coverage long-read WGS data and genomic variant calls from 34 samples in the 1kGP project (Ebert et al. 2021). We have evaluated precision of the SV callset by comparing against the long-read SV calls using these 34 genomes. The long-read SV calls were collected from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v1.0/integrated_callset/.

Genome Aggregation Database (gnomAD)

We compared the HGDP+1kGP resource to gnomAD v3.1.2, which includes both HGDP and 1kGP high-coverage whole genomes, to quantify the extent of novel variation across the allele frequency spectrum contributed by these genomes. To generate allele counts and numbers in gnomAD that would be consistent with a fully non-overlapping set of genomes, we subtracted allele counts and allele numbers in the gnomAD variant call set that were contributed specifically by the 1kGP and HGDP genomes, effectively creating a synthetic version of gnomAD without these genomes.

Gambian Genome Variation Project (GGVP)

As part of tutorials that demonstrate how we can intersect an external dataset with HGDP+1kGP and assign metadata labels, we intersected the HGDP+1kGP genomes with 394 Gambian Genome Variation Project genomes which are publicly available through the IGSR (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/gambian_genome_variation_project/data), as described previously (Network and Malaria Genomic Epidemiology Network 2019). Briefly, we first downloaded GGVP CRAM files. We then used GATK HaplotypeCaller to run variant calling in GVCF mode on the 394 Gambian genomes BAM files and generated per-sample gVCFs. The single-sample gVCFs were then combined into a multi-sample Hail

Sparse MatrixTable (MT) using Hail's run_combiner() function. The GGV Sparse MT was then combined using Hail's vcf_combiner, with the HGDP+1kGP Sparse MT to create a unique Sparse MT. Note that the Hail Sparse MatrixTable has since been replaced by the Hail VariantDataset.

Dataset Comparisons

All of the comparison datasets used GRCh38 as their reference genome aside from phase 3 1kGP, which was on hg19 prior to liftover. The comparison datasets consisted of phase 3 of 1kGP (1000 Genomes Project Consortium et al. 2015), gnomAD v3.1.2 (Chen et al. 2022), high coverage HGDP whole genome sequences (Bergström et al. 2020), and the New York Genome Center (NYGC) 1000 Genomes Project (Byrska-Bishop et al. 2022). All of these datasets were sequenced to high coverage (30X+) aside from the phase 3 1000 Genomes Project, which was sequenced to 4-8X coverage. The NYGC dataset includes all of the original 2,504 samples from phase3 1kGP as well as an additional 698 related samples.

Sample and variant QC

Quality control of samples was conducted according to procedures used in gnomAD, which include hard filtering with BAM-level metrics, sex inference, and ancestry inference described in greater depth previously (Chen et al. 2022), but was then modified to relax some gnomAD sample QC filters new to v3 in especially diverse or unique genomes. Specifically, the filters starting with 'fail_' indicate whether samples are outliers in number of variants after regressing out principal components, which can indicate a sample issue. However, we identified whole continental groups and populations that were removed due solely to SNV and indel residual filters, especially those that were most genetically unique (i.e., San, Mbuti, Biaka, Bougainville, and Papuan). Additional individuals from the LWK, Bantu Kenya, and Bantu South Africa populations were also removed solely on the basis of the fail_n_snp_residual filter, so we removed the gnomAD 'fail_' filters that quantify variant count residuals after regressing out PCs.

The raw dataset includes 189,381,961 variants (SNVs and indels) and 4,150 samples. We further filtered samples and variants according to gnomAD filters. Specifically, we excluded samples that failed gnomAD's sample QC hard filters, kept variants which were flagged as passing in the gnomAD QC pipeline, and applied genotype QC filters using a function imported from gnomAD, as described previously (Chen et al. 2022). This lowered the number of variants to 159,795,273 and removed 31 samples. Next, we conducted global and subcontinental Principal Component Analysis (PCA) within and among metadata geographical/genetic region labels (AFR, AMR, CSA, EAS, EUR, MID, and OCE) and identified 23 ancestry outliers who deviated substantially in PC space from others with the same metadata label along the first 10 PCs (these were identified visually when one to a few individuals defined the entire PC). After removing those individuals, we were left with 159,795,273 SNVs and indels in 4,096 individuals.

We calculated per-sample QC metrics such as the number of SNVs and call rate using the sample_qc() method in hail. Because singletons are especially sensitive to variation in sample size per population which is substantial across HGDP and 1kGP, we compared singleton counts by randomly downsampling to 6 unrelated samples, the minimum number of individuals per population, then removed monomorphic variants. We computed coverage data using the bam metrics field from gnomAD. We then calculated the mean of these metrics per individual within a population using Hail's hl.agg.stats() method (https://hail.is).

Relatedness

We computed relatedness using the PC-Relate algorithm (Conomos et al. 2016) implemented in Hail. Specifically, we considered SNVs with a minimum minor allele frequency of 0.05, 20 PCs, and allowed kinship coefficients up to 0.05 using the min_individual_maf=0.05, min_kinship=0.05, statistics='kin', k=20 arguments.

PCA and ADMIXTURE

We computed 20 PCs across global populations as well as within each continental ancestry group according to the “Genetic.region” project metadata label harmonized across HGDP and 1kGP as shown in **Table S1**. We first filtered to samples and variants that passed QC. We required that SNVs have MAF > 0.05 and missingness < 0.1%. We then performed LD pruning within a 500kb window, restricting to variants with $r^2 < 0.1$, leaving 255,666 variants for analysis. Finally, we computed relatedness as described above and restricted to a maximally independent set of unrelated individuals.

Using this filtered dataset, we ran PCA both globally and within metadata labels (AFR, AMR, CSA, EAS, EUR, MID, and OCE) in unrelated individuals using Hail’s hwe_normalized_pca() function, then projected related individuals into that PC space using a pc_project() function used in gnomAD and implemented in Hail.

The filtered dataset was also used to run ADMIXTURE (Alexander et al. 2009) across populations and geographical regions for values of K=2 through K=10 using ‘admixture {bed_file} {1-10}’. We conducted 10 runs for each value of K and performed a 5-fold cross-validation error for the first run of each K by adding ‘--cv=5’ to the command. Pong (Behr et al. 2016) was used to visualize ADMIXTURE results. We selected K=6 as the best fit value of

K based on a reduction in the rate of change of our 5-fold cross-validation as seen in **Figure S8**. The best fit value of K exhibits a low cross-validation error compared to other K values.

F_{ST} versus geographical distance

For each population pair that had an F_{ST} value, we calculated geographical distance using the haversine method (geosphere package in R) with the Earth's radius of 6371 km. This method of calculation did not account for human migration patterns so we additionally recalculated the pairwise geographical distances by incorporating five waypoints: Istanbul, Cairo, Phnom Penh, Anadyr and Prince Rupert, and set predetermined paths that go through certain waypoints depending on the geographical/genetic region to which the population pairs belong[22]. For example, to calculate the geographical distance between AMR and AFR populations, the path would go through Prince Rupert, Anadyr, and Cairo. In this example, the total distance between the pair would be the sum of the distances between the starting population and the first waypoint, pairs of waypoints in order (i.e. first to second, then second to third), and the third waypoint and the destination population. The distance between points was calculated using the haversine. We compared correlations between genetic divergence and geographical distance with and without waypoints using Pearson's correlation and Mantel tests.

Structural variants

Initial SV discovery and pruning

We applied GATK-SV (Collins et al. 2020) to integrate and genotype SVs from the HGDP and 1kGP samples. Briefly, the HGDP samples were split into batches, each consisting of ~190 samples, based on their initial cohort, PCR status, sex, and sequencing depth of the libraries (**Figure S3**). Raw Initial SVs were detected per sample by Manta (Chen et al. 2016), Wham (Kronenberg et al. 2015), cnMOPs (Klambauer et al. 2012), and GATK-gCNV (Babadi et al.

2022) and then were clustered across each batch and filtered through an initial random forest machine learning model to remove potential false positive SVs. We then jointly genotyped SVs across all batches using a non-redundant union of SVs. Partially overlapping SVs were either re-clustered into a unique SV or resolved into complex events. We observed mosaicism resulting from gain or loss of X and Y chromosomes for several samples (**Table S5**), likely due to a cell line artifact from passaging. While mosaic loss of the Y chromosome is the most common form of clonal mosaicism (Thompson et al. 2019), the non-canonical sex chromosome ploidies observed are not unique to these samples and have been previously observed in other datasets (Collins et al. 2020; Byrska-Bishop et al. 2022).

SV refinement and annotation

A series of refinements have been applied to improve the precision of SV calls while maintaining high sensitivity. First, two machine learning models have been developed and applied to prune false positive SVs. A lightGBM model(Byrska-Bishop et al. 2022) has been trained on the 9 1kGP samples that have been deep sequenced with long-read WGS data by the HGSVC (Chaisson et al. 2019; Ebert et al. 2021), and applied to all SVs except for large bi-allelic CNVs (>5Kb). Meanwhile, a minGQ model(Collins et al. 2020) has been trained using the inheritance information among trio families to filter bi-allelic CNVs that are 5Kb and above. Genomes that failed the machine learning models were assigned a null genotype, and the proportion of null genotypes among all samples were calculated as an “no call rate” (NCR) score. SV sites that have a 10% or higher NCR were labeled as low quality variants and removed from further analyses. Then, we examined the distribution of SVs per genome to identify potential outlier samples that carry significantly more SVs than average, and also compared the frequency of SVs across each batch to identify SVs that showed significant bias (i.e. batch effects). The resulting SV callset were annotated with their frequency by their ancestry.

Phased haplotypes and imputation accuracy evaluation

To create a haplotype reference panel that can be used for phasing or imputation, we first constructed a pedigree file with familial relationships between first degree relatives in the quality controlled harmonized dataset to improve phasing accuracy. We ran additional relatedness checks using the PC-Relate algorithm implemented in Hail. The PC-Relate results were filtered to sample pairs with a kinship statistic between 0.248 and 0.252. We then cross-checked the filtered PC-Relate results with the publicly available NYGC 1kGP pedigree file (Byrska-Bishop et al. 2022) and found that all parent-child relationships estimated by PC-Relate are reported in the 1kGP pedigree file. Of 602 previously reported trios, 9 samples failed QC (6 due to gnomAD sample QC, 3 due to ancestry outliers). In total, we therefore included 599 families, 6 of which were duos and the remaining 593 of which were trios. To investigate if there are any possible duplicate samples/monozygotic twins within or across projects, we filtered the PC-Relate results to sample pairs with kin statistic > 0.35 and found 5 pairs of samples, of which 3 have been reported before (Mountain and Ramakrishnan 2005). To verify if the 5 sample-pairs are indeed possible duplicates and/or monozygotic twins, we ran Identity-By-Descent (ref) as implemented in Hail and found that each sample-pair shared almost all alleles (IBS2). One sample from each of the 5 pairs was filtered out from the dataset.

As recommended by the SHAPEIT5 documentation, we applied additional QC filters to the dataset before phasing the haplotypes, keeping only variants with: (1) HWE>=1E-30; (2) F_MISSING<=0.1; and (3) ExcHet>=0.5 && ExcHet<=1.5. Common variants (MAF >= 0.1%) were phased in large chunks of length 20cM using the phase_common program in SHAPEIT5. The common variants chunks were then ligated together to create a haplotype scaffold containing partially phased haplotypes for each autosome (chr1-22). Using the partially phased scaffolds as input, rare variants (MAF < 0.1%) were then phased in small chunks of length 4cM using the phase_rare program in SHAPEIT5. Lastly, the fully phased chunks were then concatenated into chromosomes and indexed using bcftools. To improve the quality of phasing, pedigree information was used when phasing both common and rare variants.

To evaluate imputation accuracy, we used a filtering and downsampling strategy to simulate GWAS arrays and 93 whole genomes sequenced at various depths from previously sequenced high-coverage whole genomes from the Neuropsychiatric Genetics of African Populations-Psychosis (NeuroGAP-Psychosis) study, as previously (Martin et al. 2021). Briefly, these genomes were from participants in Ethiopia, Kenya, South Africa, and Uganda. Ethical and safety considerations are being taken across multiple levels, as described in greater detail previously (Stevenson et al. 2019). The GWAS arrays we evaluated included the widely used Illumina Global Screening Array (GSA) designed to increase scalability and improve imputation accuracy in European populations, the Multi-ethnic Genotyping Array (MEGA) designed to improve performance across globally diverse populations, and H3Africa array specialized for higher genetic diversity and smaller haplotype blocks in African genomes. We previously downsampled reads randomly to an average of 0.5X, 1X, 2X, and 4X using the GATK DownsampleSam module, which retains a random subset of reads and their mate pairs deterministically. More details on the downsampling strategy are in (Martin et al. 2021).

Code

https://github.com/atgu/hgdp_tgp

Data availability

All data are freely available and described more completely here:

https://gnomad.broadinstitute.org/news/2020-10-gnomad-v3-1-new-content-methods-annotation s-and-data-availability/#the-gnomad-hgdp-and-1000-genomes-callset. Phased haplotypes are available in BCFs here:

gs://gcp-public-data--gnomad/resources/hgdp_1kg/phased_haplotypes_v2/.

Author contributions

Z.K., M.T.Y., L.L.N., H.A.K., A.R.M. performed analysis. J.K.G., M.W.W., K.R.C., G.T., and K.J.K. processed these data with gnomAD. X.Z., S.P.H., M.E.T., and H.B. called structural variants. M.J.D., E.G.A., and A.R.M. conceptualized the project. A.R.M. provided funding. Z.K., M.T.K., K.J.K., E.G.A., and A.R.M. wrote the paper with input from all co-authors.

Acknowledgements

This work was supported by funding from the National Institutes of Health (K99/R00MH117229 to A.R.M., R01DE031261 to H.B. and A.R.M., R01MH115957 to M.E.T.). This project was also supported by funding from BroadIgnite and the Broad Next Generation Fund.

References

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56–65.

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. 2015. A global reference for human genetic variation. *Nature* **526**: 68–74.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**: 1655–1664.

Babadi M, Fu JM, Lee SK, Smirnov AN, Gauthier LD, Walker M, Benjamin DI, Karczewski KJ, Wong I, Collins RL, et al. 2022. GATK-gCNV: A Rare Copy Number Variant Discovery

Algorithm and Its Application to Exome Sequencing in the UK Biobank. *bioRxiv* 2022.08.25.504851. https://www.biorxiv.org/content/10.1101/2022.08.25.504851 (Accessed January 19, 2023).

Behr AA, Liu KZ, Liu-Fang G, Nakka P, Ramachandran S. 2016. pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics* **32**: 2817–2823.

Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, Chen Y, Felkel S, Hallast P, Kamm J, et al. 2020. Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**. https://science.sciencemag.org/content/367/6484/eaay5012/tab-pdf.

Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, et al. 2022. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**: 3426–3440.e19.

Cavalli-Sforza LL. 2005. The Human Genome Diversity Project: past, present and future. *Nat Rev Genet* **6**: 333–340.

Cavalli-Sforza LL, Wilson AC, Cantor CR, Cook-Deegan RM, King MC. 1991. Call for a worldwide survey of human genetic diversity: a vanishing opportunity for the Human Genome Project. *Genomics* **11**: 490–491.

Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, et al. 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* **10**: 1784.

Chen S, Francioli LC, Goodrich JK, Collins RL, Wang Q, Alföldi J, Watts NA, Vittal C, Gauthier LD, Poterba T, et al. 2022. A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. *bioRxiv* 2022.03.20.485034. https://www.biorxiv.org/content/biorxiv/early/2022/03/21/2022.03.20.485034 (Accessed August 15, 2022).

Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT. 2016. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**: 1220–1222.

Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, Khera AV, Lowther C, Gauthier LD, Wang H, et al. 2020. A structural variation reference for medical and population genetics. *Nature* **581**: 444–451.

Conomos MP, Reiner AP, Weir BS, Thornton TA. 2016. Model-free Estimation of Recent Genetic Relatedness. *Am J Hum Genet* **98**: 127–148.

Dey R, Lee S. 2019. Asymptotic properties of principal component analysis and shrinkage-bias adjustment under the generalized spiked population model. *J Multivar Anal* **173**: 145–164.

Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B. 2020. De novo assembly of 64 haplotype-resolved human genomes of diverse ancestry and integrated analysis of structural variation. *bioRxiv*. https://www.biorxiv.org/content/10.1101/2020.12.16.423102v1.abstract.

Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Serra Mari R, et al. 2021. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**. http://dx.doi.org/10.1126/science.abf7117.

Fatumo S, Chikowore T, Choudhury A, Ayub M, Martin AR, Kuchenbaecker K. 2022. A roadmap to increase diversity in genomic studies. *Nat Med* **28**: 243–250.

Greely HT. 1999. The overlooked ethics of the Human Genome Diversity Project. *Politics Life Sci* **18**: 297–299.

Hofmeister RJ, Ribeiro DM, Rubinacci S, Delaneau O. 2023. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. *Nat Genet* **55**: 1243–1249.

Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* **44**: 955–959.

Karczewski K, Atkinson E, Kanai M, Baya N, Turley P, Callier S, Sarma G, Walters R, Palmer D, Solomonson M, et al. Pan-UK Biobank. https://pan.ukbb.broadinstitute.org/ (Accessed June 22, 2020).

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, et al. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**: 434–443.

Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, Hamamsy T, Lek M, Samocha KE, Cummings BB, et al. 2017. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res* **45**: D840–D845.

Klambauer G, Schwarzbauer K, Mayr A, Clevert D-A, Mitterecker A, Bodenhofer U, Hochreiter S. 2012. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res* **40**: e69.

Kowalski MH, Qian H, Hou Z, Rosen JD, Tapia AL, Shan Y, Jain D, Argos M, Arnett DK, Avery C, et al. 2019. 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.

Kronenberg ZN, Osborne EJ, Cone KR, Kennedy BJ, Domyan ET, Shapiro MD, Elde NC, Yandell M. 2015. Wham: Identifying Structural Variants of Biological Consequence. *PLoS Comput Biol* **11**: e1004572.

Lam M, Awasthi S, Watson HJ, Goldstein J, Panagiotaropoulou G, Trubetskoy V, Karlsson R, Frei O, Fan C-C, De Witte W, et al. 2020. RICOPILI: Rapid Imputation for COnsortias PIpeLIne. *Bioinformatics* **36**: 930–933.

Lansdon LA, Cadieux-Dion M, Yoo B, Miller N, Cohen ASA, Zellmer L, Zhang L, Farrow EG, Thiffault I, Repnikova EA, et al. 2021. Factors affecting migration to GRCh38 in laboratories performing clinical next-generation sequencing. *J Mol Diagn*. http://dx.doi.org/10.1016/j.jmoldx.2021.02.003.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**: 1100–1104.

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, et al. 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**: 201–206.

Maples BK, Gravel S, Kenny EE, Bustamante CD. 2013. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am J Hum Genet* **93**: 278–288.

Marcus JH, Novembre J. 2017. Visualizing the geography of genetic variants. *Bioinformatics* **33**: 594–595.

Martin AR, Atkinson EG, Chapman SB, Stevenson A, Stroud RE, Abebe T, Akena D, Alemayehu M, Ashaba FK, Atwoli L, et al. 2021. Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations. *Am J Hum Genet* **108**: 656–668.

McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, Kang HM, Fuchsberger C, Danecek P, Sharp K, et al. 2016. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**: 1279–1283.

Mountain JL, Ramakrishnan U. 2005. Impact of human population history on distributions of individual-level genetic distance. *Hum Genomics* **2**: 4–19.

Network MGE, Malaria Genomic Epidemiology Network. 2019. Insights into malaria susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania. *Nature Communications* **10**. http://dx.doi.org/10.1038/s41467-019-13480-z.

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci U S A* **102**: 15942–15947.

Resnik DB. 1999. The Human Genome Diversity Project: ethical problems and solutions. *Politics Life Sci* **18**: 15–23.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. 2002. Genetic structure of human populations. *Science* **298**: 2381–2385.

Rubinacci S, Delaneau O, Marchini J. 2020. Genotype imputation using the Positional Burrows Wheeler Transform. *PLoS Genet* **16**: e1009049.

Rubinacci S, Ribeiro DM, Hofmeister RJ, Delaneau O. 2021. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat Genet* **53**: 120–126.

Schlebusch CM, Malmström H, Günther T, Sjödin P, Coutinho A, Edlund H, Munters AR, Vicente M, Steyn M, Soodyall H, et al. 2017. Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science* **358**: 652–655.

Stevenson A, Akena D, Stroud RE, Atwoli L, Campbell MM, Chibnik LB, Kwobah E, Kariuki SM, Martin AR, de Menil V, et al. 2019. Neuropsychiatric Genetics of African Populations-Psychosis (NeuroGAP-Psychosis): a case-control study protocol and GWAS in Ethiopia, Kenya, South Africa and Uganda. *BMJ Open* **9**: e025469.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH-Y, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**: 75–81.

Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, Taliun SAG, Corvelo A, Gogarten SM, Kang HM, et al. 2021. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**: 290–299.

The COVID-19 Host Genetics Initiative, Ganna A. 2021. Mapping the human genetic architecture of COVID-19 by worldwide meta-analysis. *bioRxiv*. http://medrxiv.org/lookup/doi/10.1101/2021.03.10.21252820.

Thompson DJ, Genovese G, Halvardson J, Ulirsch JC, Wright DJ, Terao C, Davidsson OB, Day FR, Sulem P, Jiang Y, et al. 2019. Genetic predisposition to mosaic Y chromosome loss in blood. *Nature* **575**: 652–657.

Weiss KM, Cavalli-Sforza LL, Dunston GM, Feldman M, Greely HT, Kidd KK, King M, Moore JA, Szathmary E, Twinn CM, et al. 1997. Proposed model ethical protocol for collecting DNA samples. *Houst Law Rev* **33**: 1431–1474.

Zhou W, Kanai M, Wu K-HH, Rasheed H, Tsuo K, Hirbo JB, Wang Y, Bhattacharya A, Zhao H, Namba S, et al. 2022. Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. *Cell Genomics* **2**. http://www.cell.com/article/S2666979X22001410/abstract (Accessed October 12, 2022).