The Biorepository and Integrative Genomics resource for inclusive genomics: insights from a diverse pediatric and admixed cohort

Silvia Buonaiuto[1*], Franco Marsico[1*], Akram Mohammed[2], Lokesh K Chinthala[2], Ernestine K Amos-Abanyie[1], Regeneron Genetics Center[3], Pjotr Prins[1], Kyobeni Mozhui[1,4], Robert J Rooney[5], Robert W Williams[4], Robert L Davis[2], Terri H Finkel[3,5], Chester W Brown[1,3], and Vincenza Colonna[1,5,6]

[*] *Equal Contribution*

[1] Dept of Genetics, Genomics and Informatics, UTHSC, USA

[2] Center for Biomedical Informatics, UTHSC, USA

[3] Regeneron Genetics Center, Tarrytown, NY, USA

[4] Department of Preventive Medicine, Division of Preventive Medicine, UTHSC, USA

[3] Dept of Pediatrics, Division of Genetics, UTHSC, USA

[5] Dept of Pediatrics, Division of Rheumatology, UTHSC, USA

[6] Institute of Genetics and Biophysics, National Research Council, Naples, 80111, Italy

Abstract

The Biorepository and Integrative Genomics (BIG) Initiative in Tennessee has developed a pioneering resource to address gaps in genomic research by linking genomic, phenotypic, and environmental data from a diverse Mid-South population, including underrepresented groups. We analyzed 13,152 genomes from BIG and found significant genetic diversity, with 50% of participants inferred to have non-European or several types of admixed ancestry. Ancestry within the BIG cohort is stratified, with distinct geographic and demographic patterns, as

African ancestry is more common in urban areas, while European ancestry is more common in suburban regions. We observe ancestry-specific rates of novel genetic variants, which are enriched for functional or clinical relevance. Disease prevalence analysis linked ancestry and environmental factors, showing higher odds ratios for asthma and obesity in minority groups, particularly in the urban area. Finally, we observe discrepancies between self-reported race and genetic ancestry, with related individuals self-identifying in differing racial categories. These findings underscore the limitations of race as a biomedical variable. BIG has proven to be an effective model for community-centered precision medicine. We integrated genomics education, and fostered great trust among the contributing communities. Future goals include cohort expansion, and enhanced genomic analysis, to ensure equitable healthcare outcomes.

1 Introduction

The history of human genetic research, from foundational concepts of genetic mapping to the development of genome-wide association studies, highlights how how a series of technical breakthroughs have progressively dismantled biases, and paved the way for more inclusive studies. Comprehensive human linkage studies began in the 1980s with the discovery that restriction length fragment polymorphisms could map Mendelian loci [1]. Several years before the human genome was sequenced, Risch and Merikangas proposed that genotyping large numbers of *’diallelic polymorphisms’* could uncover the genetic basis of common diseases [2], laying the foundation for genome-wide association studies in the early 2000s. However, limited genome data and informative markers led early genetic studies to focus on low-heterogeneity populations like Finnish and Icelandic cohorts, in which power to detect linkage were well matched to informative Single Nucleotide Polymorphisms (SNPs) [3, 4, 5, 6]. The initial SNP arrays had relative low coverage — barely adequate for the most genetically tractable European populations —, and much more limited utility for more diverse populations, especially highly heterogeneous African cohorts[7, 8].

Newer high density arrays has enabled the inclusion of non-European populations [9, 10, 11], but they have now been overshadowed by affordable whole exome and genome sequencing and by the welcome addition of deep sequencing of a much more representative swath of humanity [12, 13, 14, 15, 16]. However, even supposedly unbiased sequencing suffers from a strong structural bias — namely the reliance on the procrustean expedient of using a single “reference” genome. These final barriers to genetic equity were finally overcome last year with the publication of comprehensive human pangenome assemblies that do not reify a single haploid genome as “The Human” [17, 18]. This advancement will allow a minimally biased exploration of genome-phenome-environment relations of almost any human population.

Because of the slow progress in technology and data availability described above, along with other important contributing factors, to date most genetic data available for human research has predominantly originated from European populations, introducing a bias in medical research and healthcare that fails to accurately represent the genetic diversity of the global human population [19, 20, 21, 22, 23]. Genetic risk assessments based on European ancestry cohorts yield less accurate outcomes for non-European populations, as seen with *CYP2C19* gene variants, which affect drug metabolism and increase risks of misdiagnosis or delayed treatment [24, 25, 26]. While the importance of including ethnically diverse populations in studies of quantitative trait evolution is well known [27], the underrepresentation of diverse populations in genetic research exacerbates health inequities and limits understanding of disease genetics across ancestries, further deepening existing treatment disparities. This underrepresentation underscores the urgent need for more inclusive and diverse genetic studies to improve global health outcomes, leading to a surge of initiatives aimed at addressing these disparities (e.g., [16, 28, 29, 30]).

The Biorepository and Integrative Genomics (BIG) Initiative of Tennessee (US), is a multi-institute initiative that has developed a biorepository resource from a diverse Mid-South population in the US, including African Americans, and rural populations in Appalachia, which are disproportionately impacted by chronic diseases and the associated costs of healthcare [31, 32]. The BIG biospecimens and their genomic data are linked to de-identified electronic health records, with the purpose of creating a platform for genomics-based research that includes underrepresented populations and to support future personalized healthcare delivery platforms [33]. The initial focus of BIG on building a large and diverse cohort for genetically informed treatment and prevention of pediatric conditions, has now been expanded to a state-wide program that enrolls participants of any age with the goal of building genome-phenome-environment data for 100,000 Tennesseans.

Here we report on the analysis of 13,152 genomes from the BIG collection. We demonstrate that the BIG is a genetically diverse and ethnically rich study population, representing a unique and valuable resource for inclusive genomics. Our findings highlight ancestry-specific diversity and genetic burden, underscoring the critical need of inclusive sets of data. Finally, we show that self-reported race does not accurately reflect genetic ancestry and should be cautiously applied as a covariate in genetic analyses.

2 Results

A robust foundation for inclusive genomics studies

To date, the BIG initiative has consented over 42,000 participants with electronic health records and collected more than 15,000 biosamples from five collection sites **(Fig 1A)**. The BIG cohort is predominantly pediatric, with 87% of participants under 18 years old. At the time of sample collection, participant ages ranged from infancy to 90 years, with an average age of 8.4 years and a median age of 6.2 years (Fig. S1). BIG stands out as one of the largest cohorts focused on diverse ancestries, providing a substantial representation of different ethnic backgrounds [34, 35, 36, 37, 38, 39, 40] (**Table S1**). Notably, it is among the few cohorts specifically enriched for children with diseases, unlike most pediatric cohorts that typically recruit healthy mother-child pairs during pregnancy [41, 42, 43, 35, 36, 38, 40].

Since 2017, the BIG initiative has developed the Memphis Genomics Educational Network (*MEMGEN*) to engage the Memphis Shelby County Public School District (MSCPSD) community in genomics education. *MEMGEN* has reached students in seven public high schools (with plans to expand to 25), providing hands-on genomic experiences and ethical discussions that inspire STEM careers and academic growth in underserved communities. Community engagement is strengthened through advisory boards like the Le Bonheur Family Partners Council, supporting the BIG initiative since 2015, and the UTHSC Community Advisory Board, representing seventeen grassroots organizations. These boards ensure research and educational efforts align with community needs, fostering a community-centered approach to precision medicine and addressing health disparities.

Capturing broad diversity and several types of admixture

Within the BIG cohort, we identified and phased 6.8 million high-confidence variable sites, evenly distributed across the genome (**Fig S2**) through exome sequencing and genotype-by-sequencing data from 13,152 individuals. We used this genetic information to understand the ancestry composition of BIG by performing supervised ancestry deconvolution [44], with 1000 Genomes and HGDP as reference populations [12, 13]. While we observe a clear, uninterrupted continuum of ancestry, we subdivided the data set into seven ancestry groups to account for admixture and further characterize our cohort (**Fig 1B**). In practice, individuals were classified as not-admixed if more than 85% of their global ancestry corresponded to a single group. The choice of an 85% threshold reflects the understanding that genetic ancestry exists on a continuum, therefore defining discrete categories implies setting thresholds and making arbitrary decisions ([29] see Methods section). Furthermore, ancestral contributions over 10-15% are generally considered accurate and significant, while lower proportions are often linked to shorter ancestral segments and higher error rates [45].

According to this ancestry-based grouping, 50% of participants relate to individuals of non-European origin in the reference data sets. In particular, 20% of the BIG individuals are similar to Africans in the reference sets, and 30% present admixed origins, with two-way and multiple-admixture patterns **(Fig 1B)**. These figures, projected on all consented individuals, indicate that over 20k consented samples are likely of non-European or admixed origin, placing BIG among the largest pediatric cohorts with many admixed children (**Table S1**).

The distribution of inferred ancestry groups by zip code shows ancestry stratification, with prevalence of European ancestry in the suburbs and areas surrounding Memphis **(Fig 1C, S3)**. Stratification appears even more marked when visualized by single ancestry **(Fig 1D)**. A high dissimilarity index [46] between EUR and AFR (0.67) is observed, highlighting relevant geographic difference, while AFR and EUR-AFR (0.24) are the most evenly distributed pair, indicating much closer spatial overlap **(Fig S3C)**. This evidence indicates that BIG individuals with similar ancestry often share a similar environment, implying that geography could act as a confounding factor if not accounted for in association analyses.

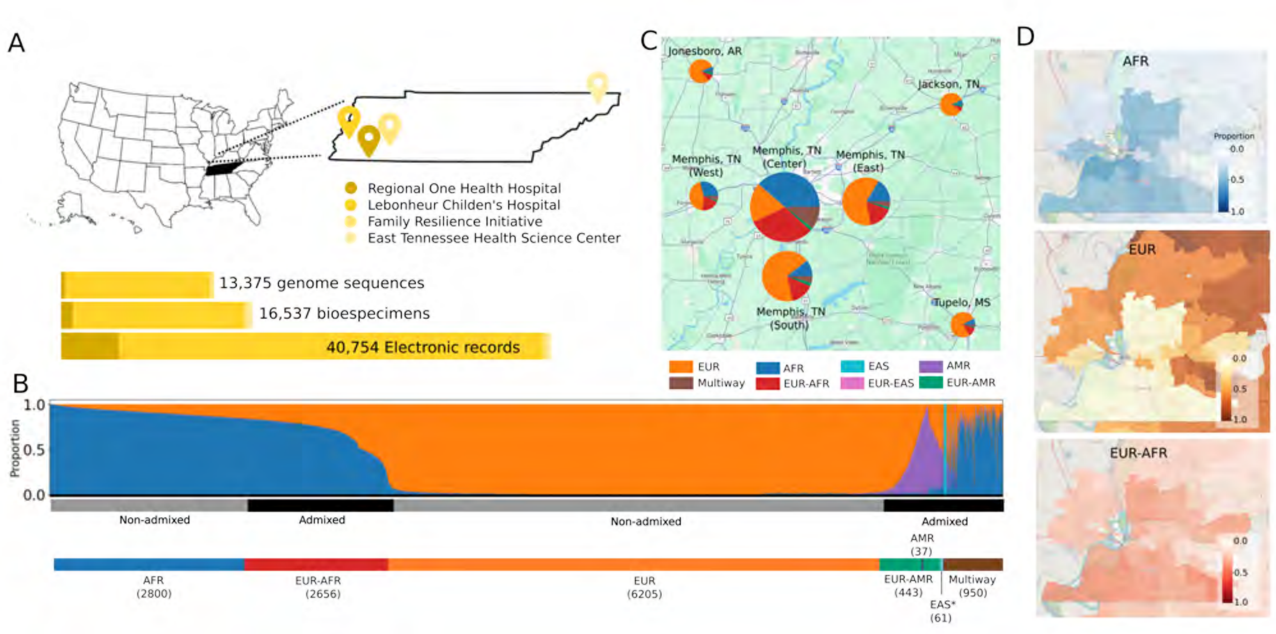

Figure 1: **Geographic distribution and global ancestry deconvolution of individuals from the BIG initiative - (A)** Overview of data collected across four sites in Tennessee, US. **(B)** Global ancestry deconvolution of 13,152 sequenced individuals, based on reference populations in the 1000 Genomes and HGDP data sets. Each vertical bar represents one individual, colors are proportional to inferred ancestry. For further analyses, individuals were grouped based on the ancestry proportions in seven categories (colored bar, number of individuals per category in parentheses), and classified as admixed or not (black and gray bar) **(C)** Proportion of individuals corresponding to each ancestry stratified by the zip code. **(D)** Prevalence of ancestries by zip code - EUR: European; AFR: African; EAS: East-Asian; AMR: Indigenous-American.

Integrating genetic, phenotypic, and environmental information

Electronic health records are an integral part of the BIG cohort, covering a range of Phecode categories [47], with gastrointestinal and respiratory medical conditions among the most represented **(Fig S4)**. We examined the prevalence of obesity, hypertension, diabetes and asthma, four health conditions commonly associated with minority groups and local environmental influences [48]. BIG children have a high incidence of diabetes and asthma (363 and 697 cases, respectively, **Fig 2A**), while adults have a more balanced incidence across these same four diseases **(Fig S5)**. Ancestry categories such as AFR and EUR-AFR, are major contributors across conditions, and we observed higher odds ratios for obesity and asthma in minority groups (all individuals self-identified as belonging to non-White racial groups) compared to 200 randomly selected conditions **(Fig 2B)**.

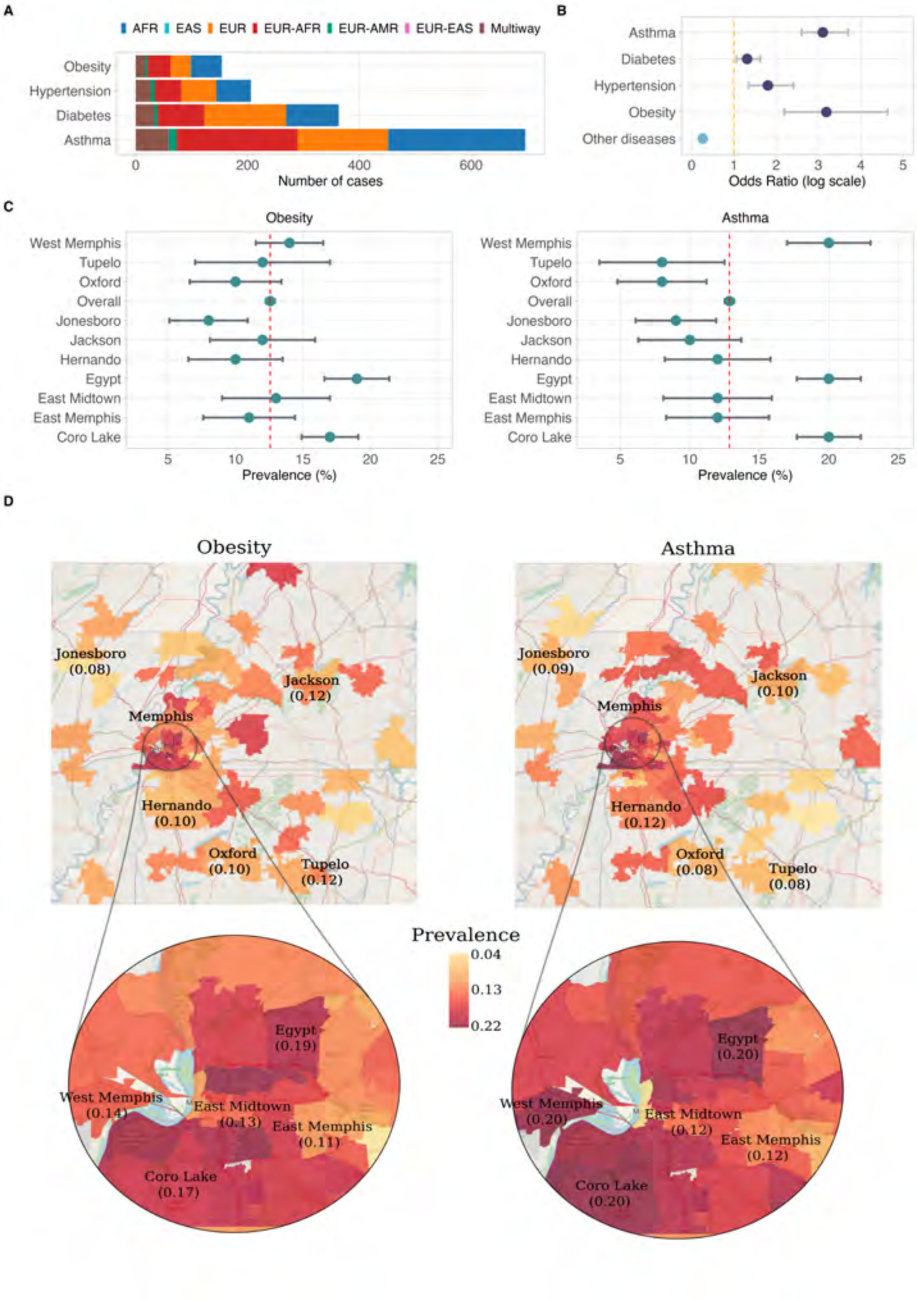

Figure 2: **Prevalence of diseases common in health disparities populations.** **(A)** Number of cases stratified by inferred ancestry categories. **(B)** Odds ratios for asthma, diabetes, hypertension, obesity compared to odds ratio of two hundred random diseases, observed among individuals self-identifying as belonging to non-White racial groups. **(C)** Prevalence of obesity and asthma by zone. This is defined as the proportion of cases in the total population. The 95% confidence intervals are calculated using the Wald method. **(D)** The map displays zones color-coded by prevalence levels in locations with more than 100 total individuals. The Memphis Metropolitan area, characterized by high population density, is zoomed in.

Analysis of disease prevalence by zip code suggests a notable environmental component for obesity and asthma. In particular, three suburban areas around Memphis exhibit above-average prevalence for both conditions, with asthma being 1.7 times more prevalent in these zones compared to the overall prevalence in BIG ($\approx$ 20% versus 12.8% CI95 [12.51-13.19] **Fig 2C**). While these analyses are only preliminary, the resulting observations underscore the value of the BIG dataset in linking genetic, phenotypic, and environmental information, enabling a multidimensional understanding of health disparities.

Ancestry-specific diversity and genetic burden

Our joint principal component analysis (PCA) of the BIG and 1000 Genomes datasets **(Fig 3A)** reveals significant genetic diversity in the BIG dataset, with mixed ancestry groups contributing to the spread and overlap between clusters corresponding to African, American, East Asian, and European individuals in the 1000 Genomes. In contrast, the 1000 Genomes dataset exhibits more distinct clustering with minimal overlap, reflecting more clearly defined ancestral groups. These results underscore the BIG dataset's value in capturing admixture and genetic diversity not represented in the 1000 Genomes, highlighting the importance of including diverse and admixed populations in genetic studies to better capture the full spectrum of human variation.

As expected, the average number of genetic differences from the reference human genome varies by ancestry [12]. Individuals with African or admixed African ancestry typically have, on average, $\sim$85k more variable sites compared to other ancestry groups **(Fig 3B)**. This observation underscores the risk of bias in using a single reference sequence and its associated genomic annotations. The genetic diversity represented within BIG would be more accurately modeled by a pangenomic approach [17].

Our dataset includes 771,717 novel single nucleotide variants (11.2% of the total), which are absent from major databases such as gnomAD, 1000 Genomes Project, or Human Genome Diversity Project [12, 13, 49]. Novel variants are mostly rare and private to ancestries, as expected (**Fig S6**). The rough number of novel variants per individual is higher within inferred admixed ancestries, Americans, and Asians (**Fig 3C**). Some novel variants have important functional consequences on the gene product (**Fig S6**, VEP classification [50]: 2.8% high impact, including

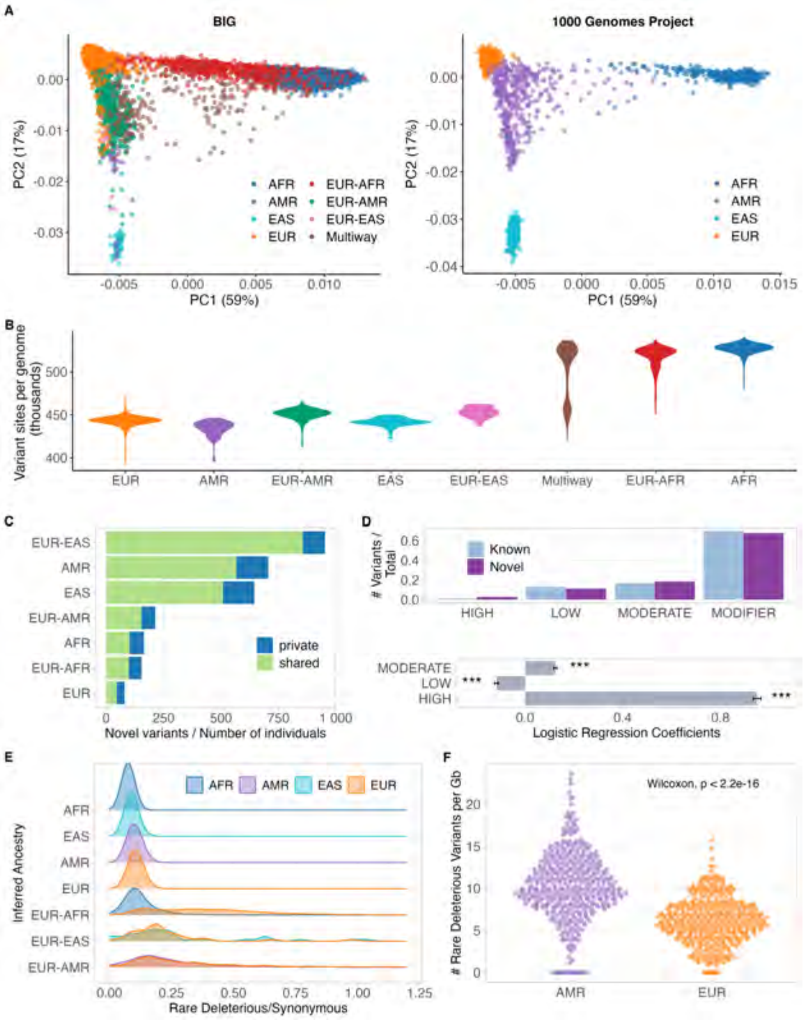

Figure 3: **Genetic variability and genetic burden in the BIG cohort - (A)** Joint principal component analysis of genetic data from individuals in the BIG and in the 1000 Genomes populations, represented separately for clarity. Colors represent inferred genetic ancestry. The first two principal components explain 76% of the variance captured by the first 20 PCs. **(B)** Number of variable sites per genome compared to the reference sequence as a function of inferred ancestry. **(C)** Estimate of the number of novel variants by individuals per ancestry with indication of variants private to the ancestry **(D)** Proportion of known and novel variants across different impact categories. **(E)** Rare deleterious-to-synonymous variant ratio across inferred ancestries. The peaks and spreads of these distributions highlight variation in the frequency of deleterious mutations across ancestries, reflecting potential differences in genetic diversity, mutation load, and evolutionary pressures. **(F)** Count of rare deleterious variants in EUR-AMR admixed individuals, which have the highest deleterious-to-synonymous ratio. Variant counts are assigned based on the inferred ancestry of the genomic regions where they are found. This means individuals are counted twice: once for their AMR ancestry regions and once for their EUR ancestry regions.

frameshift variants, stop/start gain/loss and splicing affecting variants; 19.7%: missense) and potential implications for disease association (11.0% predicted to be deleterious by SIFT [51]; 7.9% considered probably or possibly damaging by PolyPhen [52]). Notably, the rate of high impact annotation in novel variants is double compared to known (logistic regression coefficient β=0.95, p-value<0.001, **Table S3, Fig. 3D**)

Genetic burden by ancestry was evaluated as the distribution of rare deleterious (alternate allele frequency <1% in the total BIG samples, predicted to have high impact or missense with SIFT<0.05 and Polyphen>0.85) versus rare synonymous genetic variants across different ancestral groups. Among non-admixed groups, African individuals display the lowest deleterious/synonymous ratio, whereas European individuals exhibit the highest **(Fig 3C)**. Admixed populations show broader distributions in deleterious/synonymous ratios, with the European-American group demonstrating the highest ratios. In EUR-AMR group, the average number of rare deleterious variants per Gb is significantly higher in the AMR tracts compared to EUR ones **(Fig 3D, Fig S7)** as shown in other studies [53], likely due to demography and founder effect [54, 55].

Overall, the remarkable breadth of genetic diversity observed underscores BIG's value as a comprehensive resource for exploring genetic variation, enhancing disease association studies, and promoting equitable genomic research in underrepresented populations.

Discrepancies between self-reported race and inferred genetic ancestry

We compared counts of individuals in self-reported racial categories with those in inferred genetic ancestry categories, with some racial categories aggregated for simplicity (Table S2). The number of self-reported White individuals aligns closely with those inferred as Europeans, while participants identifying as Black or African American appear distributed between two genetic ancestry categories: Africans and admixed African-Europeans. For other racial groups, the patterns are more diverse and complex **(Fig 4A)**.

We eavluated the fraction of the genome shared identical by descent (IBD) among all possible pairs of individuals and compared with self-reported race. Predictably, IBD genome sharing was higher among individuals within the same self-reported race. However, we also detected IBD relationships greater than the 2nd degree (compatible with 1st cousin or uncle-nephew relationship) between individuals of different self-reported races **(Fig 4B)**. This observation suggests that genetically related individuals may self-identify differently with respect to socially constructed categories like race.

The relationship between self-reported race and inferred ancestry was further examined among pairs of individuals who identified as belonging to the same race. In some instances, the self-reported race of a pair differed from that of other pairs within the same ancestry category (**Fig 4C)**). For example, one pair of first-degree relatives (sharing approximately 50% of their genome) who both self-reported as White were found to have differing inferred ancestries: one individual was classified as having African ancestry, while the other showed a mixture of African and European ancestries (represented by the orange triangle in the AFR; EUR-AFR category in **Fig 4C)**). Similarly, among three pairs of individuals self-reporting as Black or African American, one member of each pair was inferred to have European ancestry (represented by the purple triangle in the EUR; EUR-AFR category in **Fig 4C)**). These findings highlight the limitations of using self-reported race as a category for analyzing genetic variation.

3 Discussion

The BIG cohort is a genetically diverse and ethnically inclusive pediatric resource, addressing the historic underrepresentation of non-European populations in genomics research. With 87% of participants under 18 and 50% of non-European ancestry—including 20% closely aligning with African reference populations and 30% exhibiting complex admixture patterns—it offers broad genetic variability and significant potential to represent human genomic diversity. Previous comparative studies have shown that admixed African populations from Tennessee rank among those with the highest proportion of African ancestry in the United States [56]. Notably, individuals from Memphis exhibit the greatest genetic diversity within their African ancestry component compared to thirteen other similar populations [57]. Although our study is not explicitly comparative, these findings position the African and admixed African individuals in the BIG cohort as being among the most genetically diverse populations globally.

This diversity facilitated the discovery of new genetic variants, many with clinical relevance. We have indications of ancestry-specific burden in admixed individuals. While this is an intrigu-

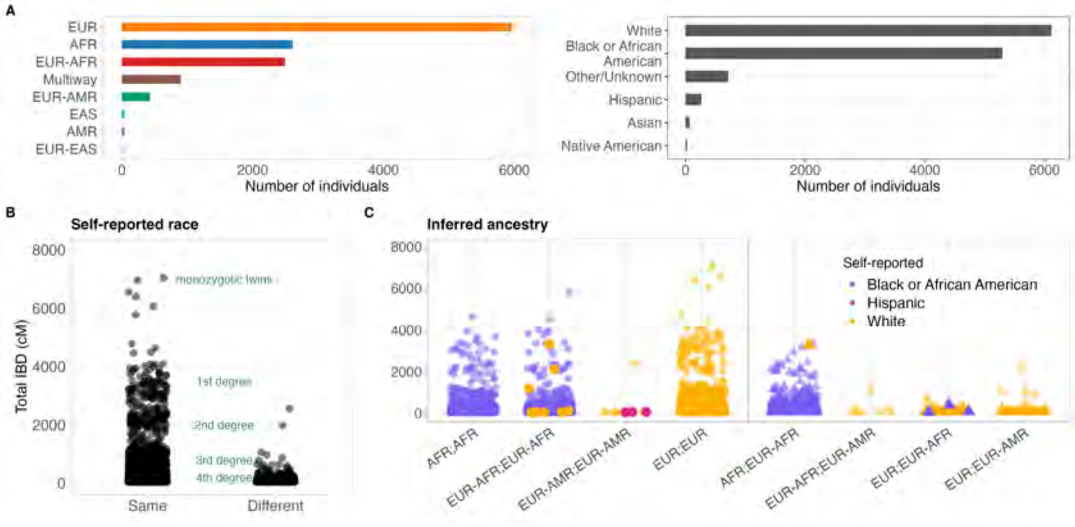

Figure 4: **Poor alignment between self-reported race and genetic ancestry.** **(A)** Counts of individuals per inferred ancestry (left) and self-reported race (right). **(B)** Genome segments shared Identical By Descent (IBD) in centimorgans (cM) between all individual pairs in BIG, categorized by whether individuals self-reported the same or different race. In some instances, individuals who self-report as belonging to different races are related at the third-degree level (e.g., first cousins) or even as close as second-degree relatives (e.g., half-siblings), as indicated by the IBD analysis. **(C)** IBD genome sharing and inferred ancestry among individuals self-reporting the same race (color-coded). In some cases the self-reported race of a pair deviates from the patterns observed in other pairs within the same ancestry category.

ing observation, it certainly deserves further investigation before any definitive conclusions can be reached. We believe that several factors, including sample size, stratification effects, and demography, must be carefully considered to achieve a more solid conclusion. This again underscores the importance of ensuring that relevant populations are well represented, as failing to do so risks leading to erroneous conclusions.

As a model for studying health disparities, the BIG cohort reveals higher odds ratios for obesity and asthma among minority groups, driven by genetic and environmental factors, as reflected in zip-code-specific disease patterns. We show that the BIG cohort has the potential to integrate genomic data, electronic health records, and environmental information to thoroughly analyze these and other common diseases. [58] With relevance to disease mapping, our study highlights how self-identified racial categories often fail to align with genetic ancestry, as seen in other studies [59]. The value of using race in biomedical research has been a longstanding topic of debate [60, 61]. Race is predominantly a socio-cultural construct, reflecting identity and social experiences rather than genetic heritage [62]. Nevertheless, race can serve as a useful framework for describing health disparities in societies where racial categories are deeply embedded in social structures [59], and there have been increasing calls for greater inclusion of underrepresented individuals in genetic and biomedical research to help clarify the relationship between race and ancestry [63, 64].

A peculiar feature of the BIG cohort is the inclusion of many admixed individuals, encompassing three distinct types of admixture. Admixed populations constitute a significant part of global genetic diversity and present unique statistical challenges in the analysis of genetic variation, leading to their frequent exclusion from genomics and medical research. Admixture can be used to map quantitative traits and to detect positive selection [65, 66], requiring smaller sample sizes compared to other mapping techniques [67]. Admixture mapping leverages local ancestry inference to associate traits with an unusually high proportion of ancestry from one of the parental populations around the disease-causing locus [68, 69, 70] and it has been successfully used - as an example - to map Alzheimer's disease [71].

All the findings from the BIG study hold significant implications primarily for the scientific community, however, and most importantly, BIG pioneers a model for inclusive genomic studies, emphasizing community engagement to align research efforts with the needs of the contributing communities. Clinically, the insights gained from BIG can inform precision medicine initiatives for historically underserved populations, particularly in regions of Tennessee, where African Americans and others face a disproportionate burden of chronic disease. Through *MEMGEN* local students and families engage with hands-on genomics education and ethical aspects of genetic research, which demystifies the science and inspires interest in STEM fields, promoting inclusivity by respecting cultural contexts and building trust.

A future key priority for the BIG initiative is to expand its participant base to include adults, allowing for a comprehensive study across all age groups and an even broader spectrum of genetic diversity. Continued community education is also a priority to sustain engagement and participation in the BIG initiative. Another important priority is to adopt a pangenomic approach in genetic data analysis to better represent the genetic diversity within the cohort. Moving toward an inclusive genome model that integrates multiple ancestries and population-specific variants will enhance the accuracy of variant identification and genetic association studies for individuals in the BIG cohort.

By embracing this pangenomic approach, the BIG initiative can establish a new benchmark for inclusive genomics, ensuring that research benefits all participants by reflecting their unique genetic backgrounds.

In conclusion, the BIG initiative can continue to lead in inclusive genomics, creating a resource that supports equitable health outcomes and advances the field toward a truly representative model of precision medicine.

4 Methods

Ethics

This study adhered to the ethical principles outlined in the Declaration of Helsinki for medical research involving human subjects. This study was conducted in accordance with ethical standards and is approved by the Institutional Review Board (IRB) of UTHSC (IRB number: 23-09204-NHSR). Written informed consent was obtained from all participants; for pediatric subjects, consent was provided by their legal guardians or next of kin. To ensure confidentiality, all data were de-identified prior to analysis.

Sample collection sites

Le Bonheur Children's Hospital (LBCH, Memphis, TN) - LBCH is the primary pediatric care center in Memphis, and serves a predominantly African American population in an area marked by significant health disparities. Recruitment at this site was launched in October 2015 and spans inpatient rooms, ICUs, outpatient clinics, and the emergency department. Information from genomic DNA extracted from leftover blood collected during routine care is linked to de-identified electronic health record data. Leftover samples are not always available for collection, although they can be collected on a subsequent visit. This explains the discrepancy between the number of consented participants and collected biosamples.

Regional One Health (ROH, Memphis, TN) - ROH is a leading healthcare provider in Memphis, providing comprehensive care to underserved and vulnerable communities in the same geographical area of LBCH. In May 2022, the BIG Initiative extended its reach to ROH, focusing on adult genomic research. Participants are recruited across hospital settings, with DNA collected from leftover blood during standard care and linked to de-identified EHR data. This expansion complements BIG's pediatric focus at LBCH by including a diverse adult population.

East Tennessee State University (ETSU, Johnson City, TN) - The BIG Initiative expanded to ETSU in May 2023 to include the Appalachian region, emphasizing adult participant recruitment. DNA samples are collected through dedicated blood draws and linked to de-identified EHR data. ETSU's inclusion aligns with BIG's commitment to engaging rural and underserved populations, complementing efforts at LBCH and ROH to create a robust, diverse genomic database for advancing precision medicine across the Mid-South and Appalachia.

Family Resilience Initiative (FRI, Memphis, TN) - Launched in January 2019, the Family Resilience Initiative (FRI) examines the impact of adverse childhood experiences (ACEs) and social determinants of health on long-term outcomes. The program enrolls mother-child dyads from the Memphis region, collecting sputum and/or blood samples at four visits spaced six months apart. Samples are processed through BIG's operational pipeline for DNA isolation, cortisol measurements, and clinical assessments. By linking biological and environmental data, FRI aims to understand ACEs' physiological and epigenetic effects, providing insights to guide tailored interventions and improve family health in vulnerable communities.

DNA sequencing

The 13,152 samples were processed with NEB/Kapa reagents, captured with the Twist Comprehensive Exome Capture design, enhanced by Regeneron-designed spikes targeting sequencing genotyping sites. Among the sequenced samples average coverage is 20X for 95.2%, with 99.3% above 90%, highlighting the overall quality of the data. The genotyping spike targets an additional ≈1.4M variants in the human genome. Genotyping call rate (percentage of SNP / indels targeted genotyping at which a call can be made) is 99.0%. CHIP targets mean coverage, crucial for detecting low-frequency variations, averages at 100X. All samples were sequenced on an Illumina NovaSeq 6000 system on S4 flow cells sequencer using 2×75 paired-end sequencing.

Variant identification

Sequence reads were aligned by the Burrows-Wheeler Aligner (BWA) MEM [72] to the GRCh38 assembly of the human reference genome in an alt-aware manner. Duplicates were marked using Picard, and mapped reads were sorted using sambamba [73]. DeepVariant v0.10.0 with a custom exome model was used for variant calling [74], and the GLnexus v1.2.6 tool was used for joint variant calling [75]. The variants were annotated using a Variant Effect Predictor (VEP 110) [50]. Phasing was performed using ShapeIT v5 [76].

Global and Local Ancestry inference

To characterize the genetic admixture within the BIG cohort, we performed a global and local ancestry inference (LAI) analysis using RFMix `v.2.0;https://github.com/slowkoni/rfmix` [44]. Reference samples included those of the 1000 Genomes Project and the Human Genome Diversity Project (HGDP), using the recently developed joint call [77]. The merged genotyping dataset, which combined BIG participants with reference samples, consisted of autosomal variants. To select the reference samples, we followed a quality control previously used in other studies [78]. To exclude reference samples with extensive admixture, we performed an unsupervised cluster analysis using ADMIXTURE [79]. We selected 4 groups (k = 4), and reference samples with a major group proportion > 0.99 were considered for the analysis. Four-way LAI was performed with the number of terminal nodes for the random forest classifier set to 5 (-n 5), the average number of generations since the expected addition set to 12 (-G 12), and ten rounds of the expectation maximization algorithm (EM) (-e 10). Reference superpopulations selected at the continental level were African (AFR), American (AMR), European (EUR), and Asian (EAS). Specifically, AFR is represented by YRI (101), LWK (30), MSL (16), Mbuti (10), GWD (48), ESN (64), Bantu South Africa (3), Bantu Kenya (10) and Biaka (21) groups. EUR contains Tuscan (6), Sardinian (12), Orcadian (13), IBS (117), GBR (103), French (24), Bergamo Italian (9), Basque (17) and CEU (114). AMR by Surui (6), Pima (10), PEL (10), Maya (16), Karitiana (7), and CLM (7). Finally, EAS is represented by CHS (106) and CHB (39). Local ancestry inference with RFMix2 was used to classify rare alleles (AF <0.01), both synonymous and deleterious, by ancestry. A custom script was developed to process phased VCFs with local ancestry calls, assigning each allele to an ancestral population and generating ancestry-specific haplotype counts. This approach enables the precise tracking of allelic ancestry in samples.

Discrete ancestry categories (AMR, AFR, EUR, EAS, EUR-AMR, EUR-AFR, and Multiway) were defined based on the following criteria: (i) individuals with more than 85% of a single ancestry were categorized into single-ancestry groups; (ii) individuals with at least 15% contribution from two ancestries, and a combined total of over 85%, were classified as two-way admixed; (iii) individuals with significant contributions (greater than 15%) from three or more ancestries were classified as Multiway. The number of individuals per ancestry group by ZIP code (based on ZCTA5 Code Tabulation Areas from the 2020 U.S. Census) was used to map the proportion of each ancestry within each location. The dissimilarity index [46] was calculated for ancestry categories with populations exceeding 500 individuals. To ensure reliable calculations, ZIP codes with fewer than 100 total individuals were excluded from the analysis.

About inferred population labels

In this study, we use self-reported race and ethnicity, which are socially constructed and categorical, alongside genetic ancestry proxies derived from methods like RFMix [44]. Although race and ethnicity are discrete categories that reflect social and historical contexts, genetic ancestry arises from continuous biological processes that capture paths through the ancestral recombination graph [80]. To facilitate our analysis, we categorize genetic ancestry into regional groupings such as AMR (ancestries from the Americas) or EUR (ancestries from Europe), but it is important to clarify that these labels are not fixed or essentialized categories [81]. This grouping is useful only because it helps us explore the demographic and environmental histories that shape the variation of complex genetic traits. This discretization is merely one arbitrary scale, and in several analyses, we examine finer ancestral variation within these groupings using dimensionality reduction techniques (PCA), unsupervised clustering (ADMIXTURE) and relatedness (e.g., IBD segment analyses). We emphasize that such proxy cannot be equated with historical racial categories that have been used to justify inequality [82]. In fact, a part of the results section is focused on showing the discrepancies between both categories.

About self-reported race

Race is self-reported by enrolled patients at the time of admission to the hospital. The admission staff select the race code from a drop-down list of possible race categories according to HL7 standards for race and ethnicity https://hl7-definition.caristix.com/v2/HL7v2.5/Tables/0005. It is possible to select multiple race codes from the drop-down list in case people associate themselves with multiple races.

Clinical Data

The clinical data associated with BIG participants are extracted from the EHR (Electronic Health Records) system in flat files and shared with UTHSC through a secure file transfer protocol. These data include demographics, visits, diagnoses, procedures, prescribed and administered medications, labs, and vital signs. These data elements are converted to a limited data set (LDS) and mapped to a common data model, the OMOP (Observational Medical Outcomes Partnership) CDM. To support the analysis, the ICD9/10 diagnosis codes are assigned to PheCodes.

Diversity and population structure analyses

Joint PCA, considering BIG and 1000GP cohorts, was performed in order to compare genetic diversity. We used the bigsnpr R package protocol for PCA analysis (https://privefl.github.io/bigsnpr) [83]. Briefly, this involved using King software [84] to estimate kinship coefficients and remove first and second-degree relatives (cutoff < 0.0884). LD clumping ($r < 0.2$) and exclusion of long-range LD regions were based on Mahalanobis distances. Outliers were identified with K-nearest-neighbor. The first 20 PCs were computed using truncated SVD. After excluding outliers, we projected related individuals in the PC space. Variants with MAF < 0.01 were excluded. For ADMIXTURE analyses, we performed unsupervised clustering with k = 3, 4, 5, and 6. We applied standard quality control filters, including LD pruning and removal of variants with MAF < 0.01. Logistic regression was performed in R.

Relatedness and Identical By Descent analysis

To analyze relatedness and infer family relationships, we used KING software to calculate kinship coefficients and determine the probability of sharing zero IBD (identity by descent) [84]. Quality control for kinship inference included removing variants with high missingness, filtering by MAF > 0.01, and performing LD pruning.

To identify IBD segments, we used hap-ibd in the phased data set comprising 13,152 genomes, focusing on autosomal loci [85]. Hap-ibd was executed with a minimum seed parameter of 2 cM to detect IBD segments of at least this length. The inferred IBD segments were post-processed using the protocol developed by Browning et al. [86], particularly the merge-ibd-segments tool, with default parameters. Gaps with at most one discordant homozygote and less than 0.6 cM were removed.

Code availability

The scripts used for QC, PCA, local and global ancestry deconvolution, and IBD analysis are available on https://github.com/SilviaBuonaiuto/BIG

5 Data availability

The BIG data presented here is potentially identifiable human data, and therefore its availability is somewhat restricted. However, we strongly support data availability in general. Data used for this study can be shared after University of Tennessee Health Science Center institutional IRB and BIG Research Oversight Committee review and approval https://uthsc.edu/cbmi/big/. Please contact the authors for further information.

6 Acknowledgments

We extend our gratitude to all the individuals and their families who generously contributed to the BIG initiative. We would like to thank Carol Hendrix and the consent teams in Memphis and in Johnson City for oversight of recruitment and sample collection; Kito Lord, from ROH;

James Adkins, and Jonathan Patrick Moorman from ETSU; Jason Yaun, Sandra Arnold from FRI; Marcella Vacca; Scott Strome; Jon McCullers; David Haines; Peter Buckley, G. Nicholas Verne, and Pamela Beckley from UTHSC; Trey Eubanks from Le Bonheur Children's Hospital; the BIG Community Advisory Board.

The authors gratefully acknowledge support from the Center for Integrative and Translational Genomics at UTHSC (SB, FM, RWW, PP, VC); NIH/NIGMS (R01GM123489 to PP); NSF (PPoSS Award 2118709 to PP); the NIH/NHLBI (RO1 HL170151 to THF); The Rady Children's Institute for Genomic Medicine (THF); the Children's Foundation of Memphis (THF); the Urban Child Institute; the Children's Foundation Research Institute, Children's Foundation of Memphis; the Assisi Foundation (CWB).

7 Author contribution

SB, FM, PP, KM, RWW, RLD, THF, CWB, VC; Data curation: SB, FM, AM, LKC; Formal Analysis: SB, FM, AM, EKA, VC; Funding acquisition: PP, RJR, RWW, RLD, THF, CWB, VC; Investigation: SB, FM, VC; Methodology: SB, FM, VC; Project administration: ; Resources: PP, RWW, CWB; Software: PP; Supervision: PP, RJR, RWW, RLD, THF, CWB, VC; Validation: ; Visualization: SB, FM, VC; Writing – original draft: SB, FM, EKA, RWW, RLD, THF, CWB, VC; Writing – review & editing: SB, FM, AM, LKC, PP, KM, EKA, RJR, RWW, RLD, THF, CWB, VC.

Extended Affiliation [3]:

3. Regeneron Genetics Center, Tarrytown, NY, USA.

RGC Management & Leadership Team Aris Baras, Goncalo Abecasis, Adolfo Ferrando, Giovanni Coppola, Andrew Deubler, Aris Economides, Luca A Lotta, John D Overton, Jeffrey G Reid, Alan Shuldiner, Katherine Siminovitch, Jason Portnoy, Marcus B Jones, Lyndon Mitnaul, Alison Fenney, Jonathan Marchini, Manuel Allen Revez Ferreira, Maya Ghoussaini, Mona Nafde, William Salerno.

Sequencing & Lab Operations John D Overton, Christina Beechert, Erin Fuller, Laura M Cremona, Eugene Kalyuskin, Hang Du, Caitlin Forsythe, Zhenhua Gu, Kristy Guevara, Michael Lattari, Alexander Lopez, Kia Manoochehri, Prathyusha Challa, Manasi Pradhan, Raymond Reynoso, Ricardo Schiavo, Maria Sotiropoulos Padilla, Chenggu Wang, Sarah E Wolf, Hang Du, Kristy Guevara.

Clinical Informatics Amelia Averitt, Nilanjana Banerjee, Dadong Li, Sameer Malhotra, Justin Mower, Mudasar Sarwar, Deepika Sharma, Sean Yu, Aaron Zhang, Muhammad Aqeel.

Genome Informatics & Data Engineering Jeffrey G Reid, Mona Nafde, Manan Goyal, George Mitra, Sanjay Sreeram, Rouel Lanche, Vrushali Mahajan, Sai Lakshmi Vasireddy, Gisu Eom, Krishna Pawan Punuru, Sujit Gokhale, Benjamin Sultan, Pooja Mule, Eliot Austin, Xiaodong Bai, Lance Zhang, Sean O'Keeffe, Razvan Panea, Evan Edelstein, Ayesha Rasool, William Salerno, Evan K Maxwell, Boris Boutkov, Alexander Gorovits, Ju Guan, Lukas Habegger, Alicia Hawes, Olga Krasheninina, Samantha Zarate, Adam J Mansfield, Lukas Habegger.

Analytical Genetics & Data Science Goncalo Abecasis, Manuel Allen Revez Ferreira, Joshua Backman, Kathy Burch, Adrian Campos, Liron Ganel, Sheila Gaynor, Benjamin Geraghty, Arkopravo Ghosh, Salvador Romero Martinez, Christopher Gillies, Lauren Gurski, Joseph Herman, Eric Jorgenson, Tyler Joseph, Michael Kessler, Jack Kosmicki, Adam Locke, Priyanka Nakka, Jonathan Marchini, Karl Landheer, Olivier Delaneau, Maya Ghoussaini, Anthony Marcketta, Joelle Mbatchou, Arden Moscati, Aditeya Pandey, Anita Pandit, Jonathan Ross, Carlo Sidore, Eli Stahl, Timothy Thornton, Sailaja Vedantam, Rujin Wang, Kuan-Han Wu, Bin Ye, Blair Zhang, Andrey Ziyatdinov, Yuxin Zou, Jingning Zhang, Kyoko Watanabe, Mira Tang, Frank Wendt, Suganthi Balasubramanian, Suying Bao, Kathie Sun, Chuanyi Zhang.

Therapeutic Area Genetics Adolfo Ferrando, Giovanni Coppola, Luca A Lotta, Alan Shuldiner, Katherine Siminovitch, Brian Hobbs, Jon Silver, William Palmer, Rita Guerreiro, Amit Joshi, Antoine Baldassari, Cristen Willer, Sarah Graham, Ernst Mayerhofer, Erola Pairo Castineira, Mary Haas, Niek Verweij, George Hindy, Jonas Bovijn, Tanima De, Parsa Akbari, Luanluan Sun, Olukayode Sosina, Arthur Gilly, Peter Dornbos, Juan Rodriguez-Flores, Moeen Riaz, Manav Kapoor, Gannie Tzoneva, Momodou W Jallow, Anna Alkelai, Ariane Ayer, Veera Rajagopal, Sahar Gelfman, Vijay Kumar, Jacqueline Otto, Neelroop Parikshak, Aysegul Guvenek, Jose Bras, Silvia Alvarez, Jessie Brown, Jing He, Hossein Khiabanian, Joana Revez, Kimberly Skead, Valentina Zavala, Jae Soon Sul, Lei Chen, Sam Choi, Amy Damask, Nan Lin, Charles Paulding.

Research Program Management and Strategic Initiatives Marcus B Jones, Esteban Chen, Michelle G LeBlanc, Jason Mighty, Jennifer Rico-Varela, Nirupama Nishtala, Nadia Rana, Jaimee Hernandez.

Senior Partnerships and Business Operations

Alison Fenney, Randi Schwartz, Jody Hankins, Anna Han, Samuel Hart.

Business Operations and Administrative Coordinators

Ann Perez-Beals, Gina Solari, Johannie Rivera-Picart, Michelle Pagan, Sunilbe Siceron.

8 Supplementary Figures

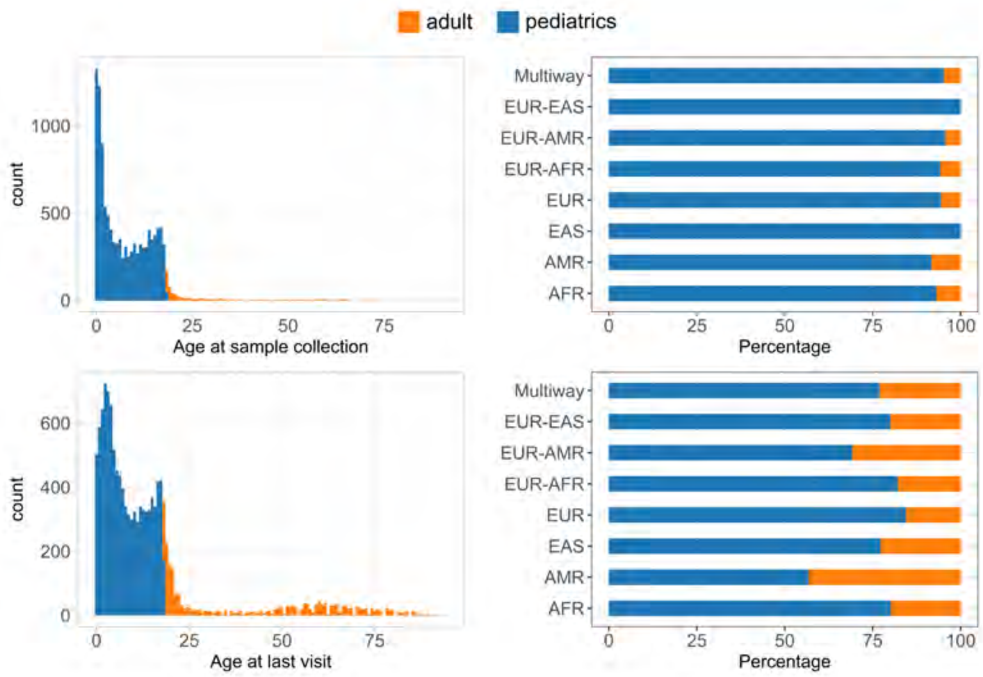

Figure S1: **Age distribution.** Distribution of age at sample collection (top) and at the last visit (bottom). The right panels: bar plots illustrating the percentage distribution of demographic categories, stratified by inferred ancestry.

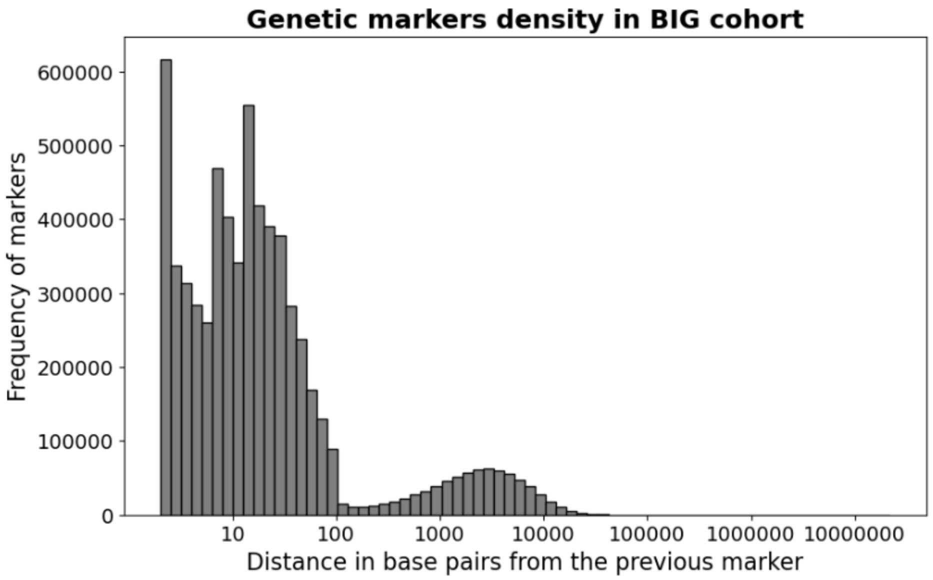

Figure S2: **Distance between consecutive variable sites.** Distribution of distances for all pair of variants. The distribution is multimodal, roughly corresponding to modal distance for markers in the coding (lower modes) and non-coding (higher mode) regions. Overall, the markers appear to be evenly distributed, indicating good coverage across the entire genome.

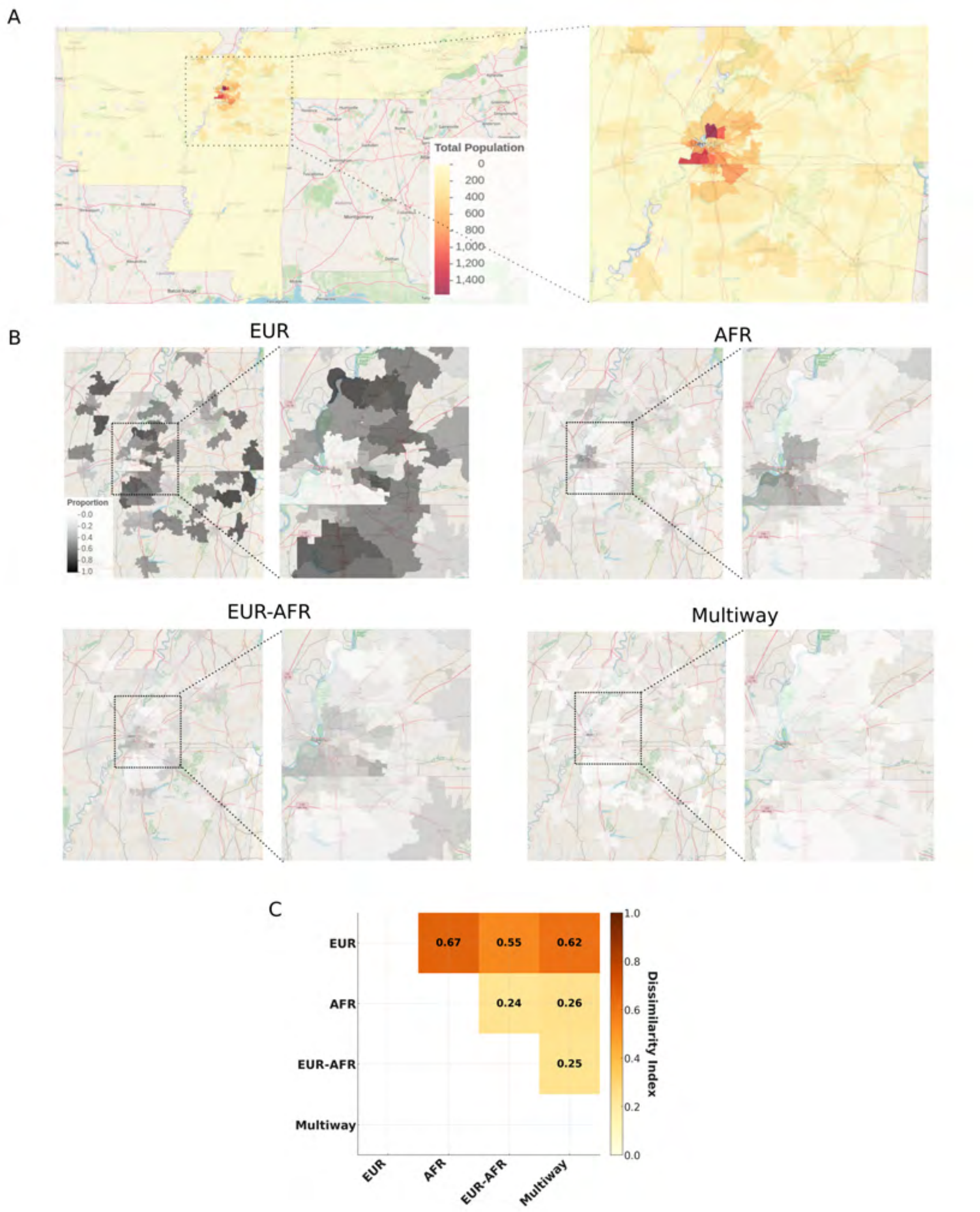

Figure S3: **Demography of enrolled participants.** **(A)** Number of enrolled participants by ZIP code. The region surrounding Memphis City is zoomed in. **(B)** Proportion of individuals by zip code for inferred ancestries with more than 500 individuals. **(C)** Pairwise Dissimilarity index between ancestries, considering the proportion in each zip code. The dissimilarity index measures the extent of segregation between two groups across geographic areas, indicating the proportion of one group that would need to relocate to achieve an even distribution relative to the other group, with values ranging from 0 (perfect integration) to 1 (complete segregation).

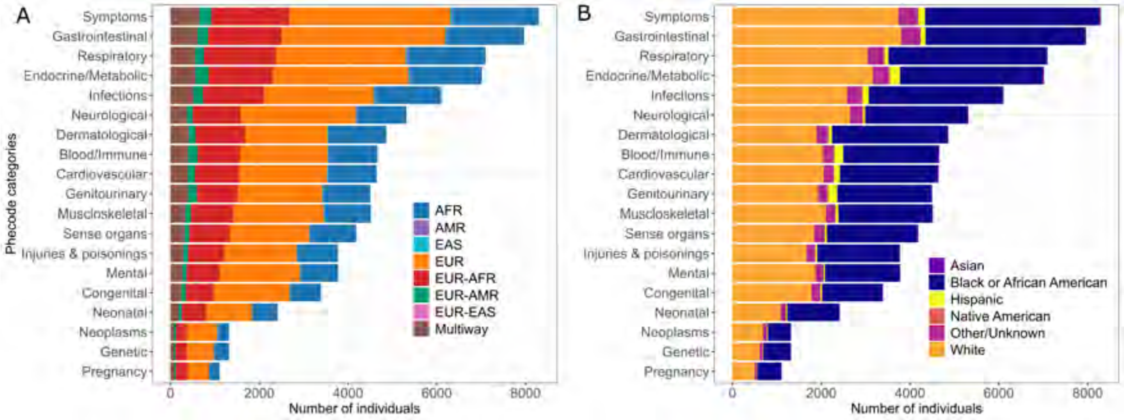

Figure S4: **Phenotypes prevalence in electronic health records in participants with sequence data** stratified by inferred ancestry **(A)** and self-reported race **(B)**. Phenotypes are grouped into Phecode categories. Distribution of ancestries in Phecode categories reflect the global distribution of ancestry

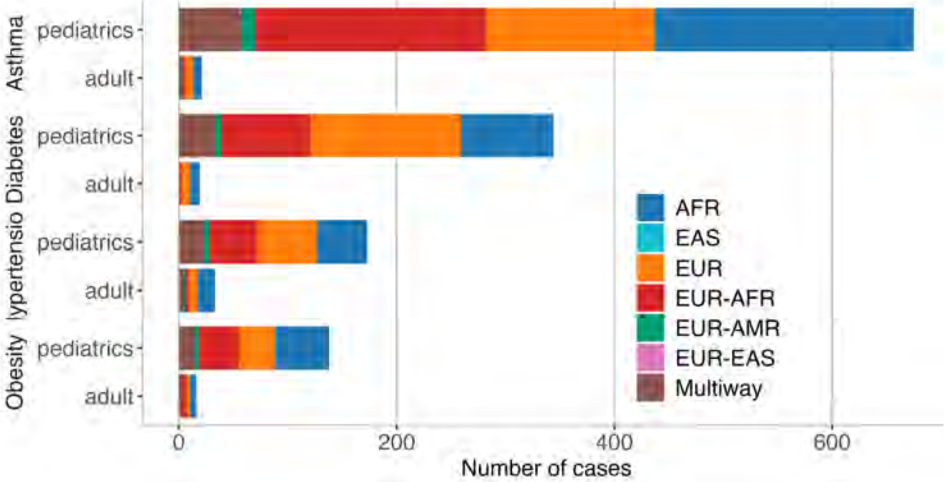

Figure S5: **Count of cases across Asthma, Diabetes, Hypertension, and Obesity**. Cases are categorized by pediatric and adult populations and color-coded by inferred ancestry groups: AFR (African), EAS (East Asian), EUR (European), EUR-AFR (European-African), EUR-AMR (European-American), EUR-EAS (European-East Asian), and Multiway.

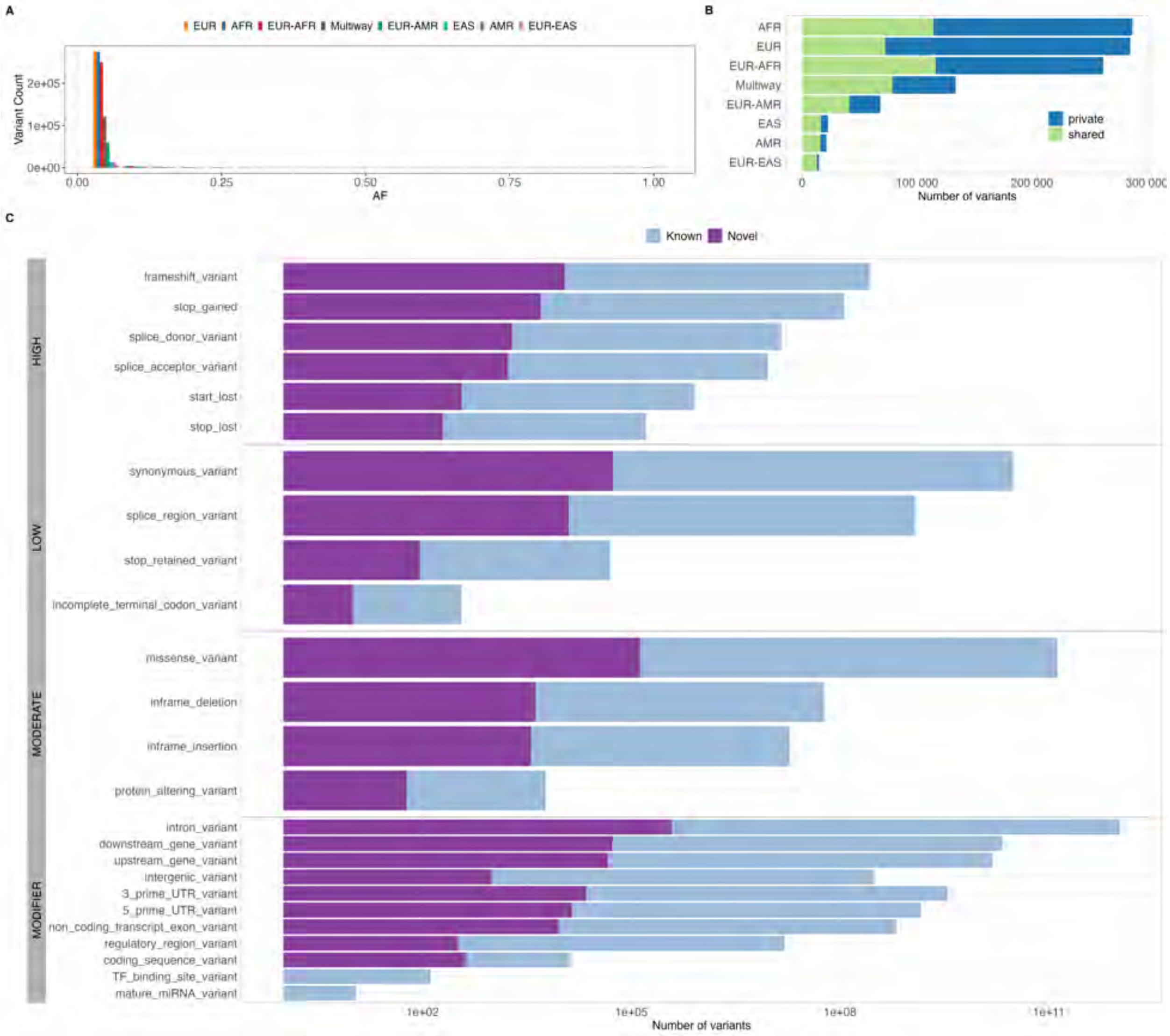

Figure S6: **Features of novel variants.** **(A)** Allele frequency spectrum showing the prevalence of rare variants. **(B)** Counts of variants by ancestry stratified by private and shared with another one or more ancestries. **(C)** Counts of variants (log scale) by annotated consequences for novel and known variants.

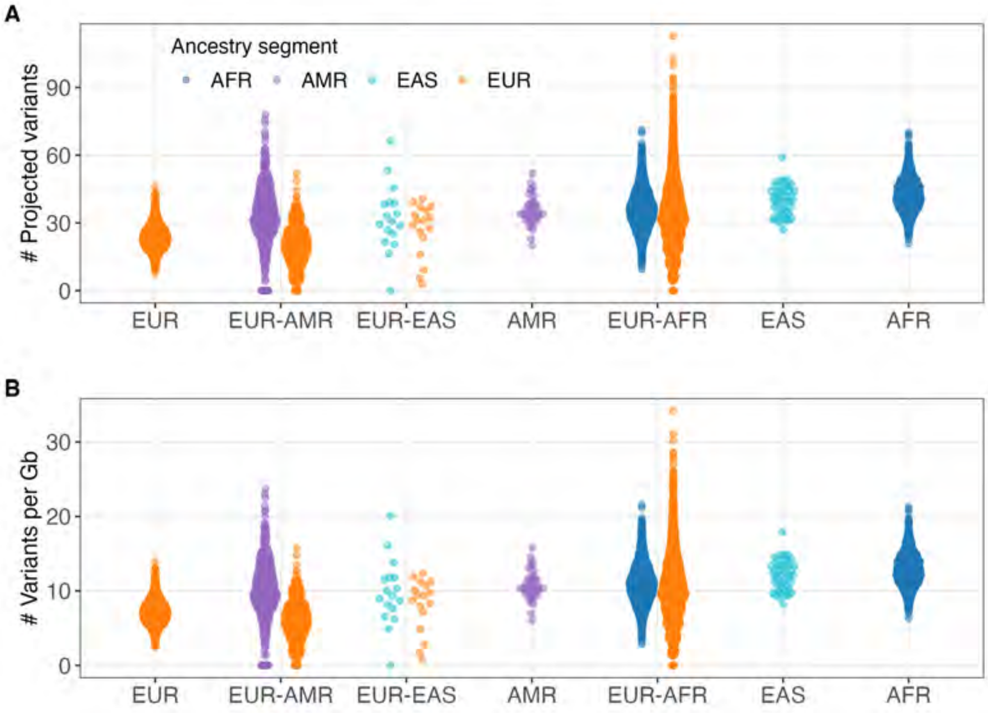

Figure S7: **Counts per individual of rare deleterious variant by ancestry.** Rare deleterious variants are defined as having alternate allele frequency <1% in the total BIG samples, and classified as high impact or missense with SIFT<0.05 and Polyphen>0.85. Variants counts take into account the inferred ancestry of the genomic tract in which they are located, therefore individuals in admixed groups individuals are represented twice. In panel **(A)** counts per ancestry tract are normalized by the proportion of ancestry and therefore the y-axis represent the projection as the ancestry tract was as long as the whole genome. In panel **(B)** we report counts per Gb.

9 Supplementary Tables

Table S1: **Examples of large pediatric cohorts.** Although the list is not exhaustive, it is intended to provide context for understanding BIG's position in terms of size, diversity, and data availability.

Reported Ancestry Representation	Cohort Name	Size	Start Date	Reported Ancestry	Study Group Type	EHR Availability	Genetic Data
Predominantly one ancestry	Avon Longitudinal Study of Parents and Children (ALSPAC) [41]	14,000 children	1991	Predominantly European descent, reflecting the population of the Avon area in the UK	Mother-Child	Has EHR	Has genetic data
	Copenhagen Prospective Studies on Asthma in Childhood (COPSAC) [87]	700 children	1998	Primarily Danish, reflecting the population of Denmark	Children Only	No EHR	Has genetic data
	The Norwegian Mother and Child Cohort Study (MoBa) [42]	114,500 children	1999	Predominantly Norwegian, reflecting the population of Norway	Mother-Child	Has EHR	Has genetic data
	Longitudinal Study of Australian Children (LSAC) [88]	10,000 children	2004	Predominantly Australian, with representation from various ethnic backgrounds	Children Only	No EHR	No genetic data
	All Our Families (AOF) Cohort [43]	3,000 families	2008	Primarily of European descent, reflecting the population of Calgary, Canada	Mother-Child	Has EHR	No genetic data
Diverse ancestries	Children of Philadelphia (CHOP)	100,000 children	2006	Diverse, reflecting the population of Philadelphia	Mother-Child	Yes	Has genetic data
	Childhood Cancer Survivor Study (CCSS) [34]	24,000 survivors	1994	Diverse, reflecting the population of North America	Children Only	Has EHR	Has genetic data
	The Boston Birth Cohort [35]	8,000 births	1998	Predominantly African American and Hispanic participants	Mother-Child	Has EHR	Has genetic data
	Generation R Study [36]	10,000 children	2002	Multi-ethnic urban population, including Dutch, Surinamese, Turkish, Moroccan, and others	Mother-Child	Has EHR	Has genetic data
	Pediatric Imaging, Neurocognition, and Genetics (PING) Study [37]	1,400 children	2009	Diverse, including African American, Asian, Hispanic, and non-Hispanic White participants	Children Only	No EHR	Has genetic data
	NICHD Fetal Growth Studies [38]	2,400 pregnancies	2009	Diverse, including African American, Asian, Hispanic, and non-Hispanic White participants	Mother-Child	Has EHR	Has genetic data
	Biorepository for Integrative Genomics (BIG) [33]	42,000	-	Diverse, including African American, Asian, Hispanic, and non-Hispanic White participants	Children Only	Has EHR	Has genetic data
	Healthy Brain Network (HBN) [39]	10,000 children	2015	Diverse, with efforts to include underrepresented populations	Children Only	No EHR	Has genetic data
	Environmental Influences on Child Health Outcomes (ECHO) [40]	50,000 children	2016	Diverse, with efforts to include underrepresented populations	Mother-Child	Has EHR	Has genetic data

Table S2: **Simplification of self-reported race entries in electronic health records.** The purpose of grouping is to simplify the analyses and eliminate the use of inaccurate or inappropriate terminology [89].

Original Category	Grouped Category
White or Caucasian Caucasian White	White
Black or African American African American	Black or African American
Asian Asian or Pacific Islander	Asian
Other/Unknown Other Declined Patient Declined to answer Unavailable Multiple	Other/Unknown

Table S3: **Prevalence of high impact variants among novel variants.** Estimates from the logistic regression.

Term	Estimate	Std. error	Statistic	p.value	conf.low
Intercept	-2.09	0.001	-1430	$< 2e - 16$	-2.10
HIGH	0.95	0.008	116	$< 2e - 16$	0.93
LOW	-0.118	0.004	-30.3	$< 2e - 16$	-0.125
MODERATE	0.12	0.003	38.8	$< 2e - 16$	0.118

References

[1] David Botstein, Raymond L White, Mark Skolnick, and Ronald W Davis. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American journal of human genetics*, 32(3):314, 1980.

[2] Neil Risch and Kathleen Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517, 1996.

[3] Leena Peltonen, Anu Jalanko, and Teppo Varilo. Molecular genetics the finnish disease heritage. *Human molecular genetics*, 8(10):1913–1923, 1999.

[4] Jeff Gulcher, Agnar Helgason, and Kári Stefánsson. Genetic homogeneity of icelanders. *Nature Genetics*, 26(4):395–395, 2000.

[5] Yali Xue, Massimo Mezzavilla, Marc Haber, Shane McCarthy, Yuan Chen, Vagheesh Narasimhan, Arthur Gilly, Qasim Ayub, Vincenza Colonna, Lorraine Southam, et al. Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated european populations. *Nature communications*, 8(1):15927, 2017.

[6] Kalliope Panoutsopoulou, Konstantinos Hatzikotoulas, Dionysia Kiara Xifara, Vincenza Colonna, Aliki-Eleni Farmaki, Graham RS Ritchie, Lorraine Southam, Arthur Gilly, Ioanna Tachmazidou, Segun Fatumo, et al. Genetic characterization of greek population isolates reveals strong genetic drift at missense and trait-associated variants. *Nature communications*, 5(1):5345, 2014.

[7] Dan L Nicolae, Xiaoquan Wen, Benjamin F Voight, and Nancy J Cox. Coverage and characteristics of the affymetrix genechip human mapping 100k snp set. *PLoS Genetics*, 2(5):e67, 2006.

[8] Lon R Cardon and Goncalo R Abecasis. Using haplotype blocks to map human complex trait loci. *TRENDS in Genetics*, 19(3):135–140, 2003.

[9] DM Altshuler, RA Gibbs, L Peltonen, DM Altshuler, RA Gibbs, L Peltonen, et al. Interna-

tional hapmap 3 consortium: Integrating common and rare genetic variation in diverse human populations. *Nature*, 467:52, 2010.

[10] Guanjie Chen, Daniel Shriner, Jie Zhou, Ayo Doumatey, Hanxia Huang, Norman P Gerry, Alan Herbert, Michael F Christman, Yuanxiu Chen, Georgia M Dunston, et al. Development of admixture mapping panels for african americans from commercial high-density snp arrays. *BMC genomics*, 11:1–12, 2010.

[11] Arti Tandon, Nick Patterson, and David Reich. Ancestry informative marker panels for african americans based on subsets of commercially available snp arrays. *Genetic epidemiology*, 35(1):80–83, 2011.

[12] Adam Auton, Gonçalo R Abecasis, David M Altshuler, Richard M Durbin, David R Bentley, Aravinda Chakravarti, Andrew G Clark, Peter Donnelly, Evan E Eichler, Paul Flicek, Stacey B Gabriel, Richard A Gibbs, Eric D Green, Matthew E Hurles, and Gil McVean. A global reference for human genetic variation. *Nature*, 526:68–74, 2015.

[13] Anders Bergström, Shane A McCarthy, Ruoyun Hui, Mohamed A Almarri, Qasim Ayub, Petr Danecek, Yuan Chen, Sabine Felkel, Pille Hallast, Jack Kamm, et al. Insights into human genetic variation and population history from 929 diverse genomes. *Science*, 367(6484):eaay5012, 2020.

[14] H3Africa Consortium et al. Enabling the genomic revolution in africa: H3africa is developing capacity for health-related genomics research in africa. *Science (New York, NY)*, 344(6190):1346, 2014.

[15] Swapan Mallick, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo, Mengyao Zhao, Niru Chennagiri, Susanne Nordenfelt, Arti Tandon, et al. The simons genome diversity project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206, 2016.

[16] Mayo Blegen Ashley L. 18 Wirkus Samantha J. 18 Wagner Victoria A. 18 Meyer Jeffrey G. 18 Cicek Mine S. 10 18 Biobank and All of Us Research Demonstration Project Teams Choi Seung Hoan 14 http://orcid. org/0000-0002-0322-8970 Wang Xin 14 http://orcid. org/0000

0001-6042-4487 Rosenthal Elisabeth A. 15. Genomic data in the all of us research program. *Nature*, 627(8003):340–346, 2024.

[17] Wen-Wei Liao, Mobin Asri, Jana Ebler, Daniel Doerr, Marina Haukness, Glenn Hickey, Shuangjia Lu, Julian K Lucas, Jean Monlong, Haley J Abel, et al. A draft human pangenome reference. *Nature*, 617(7960):312–324, 2023.

[18] Erik Garrison, Andrea Guarracino, Simon Heumos, Flavia Villani, Zhigui Bao, Lorenzo Tattini, Jörg Hagmann, Sebastian Vorbrugg, Santiago Marco-Sola, Christian Kubica, et al. Building pangenome graphs. *Nature Methods*, pages 1–5, 2024.

[19] Christina V. Van Hout, Ioanna Tachmazidou, Joshua D. Backman, J. Scott Hoffman, Di Liu, Ashutosh K. Pandey, Claudia Gonzaga-Jauregui, Shaista Khalid, Bingshan Ye, Niranjan Banerjee, Matthew R. Nelson, and Gonçalo R. Abecasis. Whole exome sequencing and characterization of coding variation in 49,960 individuals in the uk biobank. *Nature Communications*, 11(1):1–11, Jul 2020.

[20] Christina C. Kyriazis, Wei Wei, Jemma Danesh, Danish Saleheen, Roland D. Schmid, and Emanuele Di Angelantonio. Human genetic diversity and disease: from outside africa to within europe. *Communications Biology*, 6:353, 2023.

[21] Pardis C. Sabeti and David Reich. Genetic and archeological evidence for early human population structure. *Cell*, 179(6):1462–1474, 2019.

[22] Giorgio Sirugo, Scott M Williams, and Sarah A Tishkoff. The missing diversity in human genetic studies. *Cell*, 177(1):26–31, 2019.

[23] Segun Fatumo, Tinashe Chikowore, Ananyo Choudhury, Muhammad Ayub, Alicia R Martin, and Karoline Kuchenbaecker. A roadmap to increase diversity in genomic studies. *Nature medicine*, 28(2):243–250, 2022.

[24] Sonia Moreno-Grau, Manvi Vernekar, Arturo Lopez-Pineda, Daniel Mas-Montserrat, Míriam Barrabés, Consuelo D Quinto-Cortés, Babak Moatamed, Ming Ta Michael Lee, Zhenning Yu, Kensuke Numakura, et al. Polygenic risk score portability for common diseases across genetically diverse populations. *Human Genomics*, 18(1):93, 2024.

[25] Ola AlAzzeh and Youssef M Roman. The frequency of rs2231142 in abcg2 among native hawaiian and pacific islander subgroups: implications for personalized rosuvastatin dosing. *Pharmacogenomics*, 24(3):173–182, 2023.

[26] David Twesigomwe, Britt I Drögemöller, Galen EB Wright, Clement Adebamowo, Godfred Agongo, Palwendé R Boua, Mogomotsi Matshaba, Maria Paximadis, Michèle Ramsay, Gustave Simo, et al. Characterization of cyp2d6 pharmacogenetic variation in sub-saharan african populations. *Clinical Pharmacology & Therapeutics*, 113(3):643–659, 2023.

[27] Michael A McQuillan, Chao Zhang, Sarah A Tishkoff, and Alexander Platt. The importance of including ethnically diverse populations in studies of quantitative trait evolution. *Current opinion in genetics & development*, 62:30–35, 2020.

[28] Edra K Ha, Daniel Shriner, Shawneequa L Callier, Lorinda Riley, Adebowale A Adeyemo, Charles N Rotimi, and Amy R Bentley. Native hawaiian and pacific islander populations in genomic research. *NPJ Genomic Medicine*, 9(1):45, 2024.

[29] Mashaal Sohail, María J Palma-Martínez, Amanda Y Chong, Consuelo D Quinto-Cortés, Carmina Barberena-Jonas, Santiago G Medina-Muñoz, Aaron Ragsdale, Guadalupe Delgado-Sánchez, Luis Pablo Cruz-Hervert, Leticia Ferreyra-Reyes, et al. Mexican biobank advances population and medical genomics of diverse ancestries. *Nature*, 622(7984):775–783, 2023.

[30] Luisa Pereira, Leon Mutesa, Paulina Tindana, and Michèle Ramsay. African genetic diversity and adaptation inform a precision medicine agenda. *Nature Reviews Genetics*, 22(5):284–306, 2021.

[31] Richard Crespo, Matthew Christiansen, Kim Tieman, and Richard Wittberg. An emerging model for community health worker–based chronic care management for patients with high health care costs in rural appalachia. *Preventing chronic disease*, 17:E13, 2020.

[32] Kate Beatty, Olivia Egen, John Dreyzehner, and Randy Wykoff. Poverty and health in tennessee. *South Med J*, 113(1):1–7, 2020.

[33] Rony Jose, Robert Rooney, Naga Nagisetty, Robert Davis, and David Hains. Biorepository and integrative genomics initiative: Designing and implementing a preliminary platform for predictive, preventive and personalized medicine at a pediatric hospital in a historically disadvantaged community in the usa. *EPMA Journal*, 9:225–234, 2018.

[34] Leslie L Robison, Gregory T Armstrong, John D Boice, Eric J Chow, Stella M Davies, Sarah S Donaldson, Daniel M Green, Sue Hammond, Anna T Meadows, Ann C Mertens, et al. The childhood cancer survivor study: a national cancer institute–supported resource for outcome and intervention research. *Journal of clinical oncology*, 27(14):2308–2318, 2009.

[35] Colleen Pearson, Tami Bartell, Guoying Wang, Xiumei Hong, Serena A Rusk, LingLing Fu, Sandra Cerda, Blandine Bustamante-Helfrich, Wendy Kuohung, Christina Yarrington, et al. Boston birth cohort profile: rationale and study design. *Precision nutrition*, 1(2):e00011, 2022.

[36] Marjolein N Kooijman, Claudia J Kruithof, Cornelia M van Duijn, Liesbeth Duijts, Oscar H Franco, Marinus H van IJzendoorn, Johan C de Jongste, Caroline CW Klaver, Aad van der Lugt, Johan P Mackenbach, et al. The generation r study: design and cohort update 2017. *European journal of epidemiology*, 31:1243–1264, 2016.

[37] Terry L Jernigan, Timothy T Brown, Donald J Hagler Jr, Natacha Akshoomoff, Hauke Bartsch, Erik Newman, Wesley K Thompson, Cinnamon S Bloss, Sarah S Murray, Nicholas Schork, et al. The pediatric imaging, neurocognition, and genetics (ping) data repository. *Neuroimage*, 124:1149–1154, 2016.

[38] Germaine M Buck Louis, Jagteshwar Grewal, Paul S Albert, Anthony Sciscione, Deborah A Wing, William A Grobman, Roger B Newman, Ronald Wapner, Mary E D'Alton, Daniel Skupski, et al. Racial/ethnic standards for fetal growth: the nichd fetal growth studies. *American journal of obstetrics and gynecology*, 213(4):449–e1, 2015.

[39] Lindsay M Alexander, Jasmine Escalera, Lei Ai, Charissa Andreotti, Karina Febre, Alexander Mangone, Natan Vega-Potler, Nicolas Langer, Alexis Alexander, Meagan Kovacs, et al. An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific data*, 4(1):1–26, 2017.

[40] Christina H Park, Carol J Blaisdell, S Sonia Arteaga, Clay Mash, Susan Laessig, Manjit Hanspal, Erin Luetkemeier, Leslie C Thompson, and Matthew W Gillman. How the environmental influences on child health outcome (echo) cohort can spur discoveries in environmental epidemiology. *American Journal of Epidemiology*, page kwae073, 2024.

[41] Deborah A Lawlor, Melanie Lewcock, Louise Rena-Jones, Claire Rollings, Vikki Yip, Daniel Smith, Rebecca M Pearson, Laura Johnson, Louise AC Millard, Nashita Patel, et al. The second generation of the avon longitudinal study of parents and children (alspac-g2): a cohort profile. *Wellcome Open Research*, 4, 2019.

[42] Per Magnus, Lorentz M Irgens, Kjell Haug, Wenche Nystad, Rolv Skjærven, and Camilla Stoltenberg. Cohort profile: the norwegian mother and child cohort study (moba). *International journal of epidemiology*, 35(5):1146–1150, 2006.

[43] Suzanne C Tough, Sheila W McDonald, Beverly Anne Collisson, Susan A Graham, Heather Kehler, Dawn Kingston, and Karen Benzies. Cohort profile: the all our babies pregnancy cohort (aob). *International Journal of Epidemiology*, 46(5):1389–1390k, 2017.

[44] Brian K Maples, Simon Gravel, Eimear E Kenny, and Carlos D Bustamante. Rfmix: a discriminative modeling approach for rapid and robust local-ancestry inference. *The American Journal of Human Genetics*, 93(2):278–288, 2013.

[45] Simon Gravel. Population genetics models of local ancestry. *Genetics*, 191(2):607–619, 2012.

[46] Otis Dudley Duncan and Beverly Duncan. A methodological analysis of segregation indexes. *American sociological review*, 20(2):210–217, 1955.

[47] Lisa Bastarache. Using phecodes for research with the electronic health record: from phewas to phers. *Annual review of biomedical data science*, 4(1):1–19, 2021.

[48] Christine Cole Johnson, Aruna Chandran, Suzanne Havstad, Xiuhong Li, Cynthia T McEvoy, Dennis R Ownby, Augusto A Litonjua, Margaret R Karagas, Carlos A Camargo, James E Gern, et al. Us childhood asthma incidence rate patterns from the echo consortium to identify high-risk groups for primary prevention. *JAMA pediatrics*, 175(9):919–927, 2021.

[49] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, 2020.

[50] William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham RS Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. *Genome biology*, 17:1–14, 2016.

[51] Robert Vaser, Swarnaseetha Adusumalli, Sim Ngak Leng, Mile Sikic, and Pauline C Ng. Sift missense predictions for genomes. *Nature protocols*, 11(1):1–9, 2016.

[52] Ivan Adzhubei, Daniel M Jordan, and Shamil R Sunyaev. Predicting functional effect of human missense mutations using polyphen-2. *Current protocols in human genetics*, 76(1):7–20, 2013.

[53] Zachary A Szpiech, Angel CY Mak, Marquitta J White, Donglei Hu, Celeste Eng, Esteban G Burchard, and Ryan D Hernandez. Ancestry-dependent enrichment of deleterious homozygotes in runs of homozygosity. *The American Journal of Human Genetics*, 105(4):747–762, 2019.

[54] Marcos Araújo Castro e Silva, Tiago Ferraz, Caina M Couto-Silva, Renan B Lemes, Kelly Nunes, David Comas, and Tábita Hünemeier. Population histories and genomic diversity of south american natives. *Molecular biology and evolution*, 39(1):msab339, 2022.

[55] Sara D Niedbalski and Jeffrey C Long. Novel alleles gained during the beringian isolation period. *Scientific Reports*, 12(1):4289, 2022.

[56] Soheil Baharian, Maxime Barakatt, Christopher R Gignoux, Suyash Shringarpure, Jacob Errington, William J Blot, Carlos D Bustamante, Eimear E Kenny, Scott M Williams, Melinda C Aldrich, et al. The great migration and african-american genomic diversity. *PLoS genetics*, 12(5):e1006059, 2016.

[57] Sharon R Browning, Brian L Browning, Martha L Daviglus, Ramon A Durazo-Arvizu, Neil

Schneiderman, Robert C Kaplan, and Cathy C Laurie. Ancestry-specific recent effective population size in the americas. *PLoS genetics*, 14(5):e1007385, 2018.

[58] Kirsten Voorhies, Akram Mohammed, Lokesh Chinthala, Sek Won Kong, In-Hee Lee, Alvin T Kho, Michael McGeachie, Kenneth D Mandl, Benjamin Raby, Melanie Hayes, et al. Gsdmb/ormdl3 rare/common variants are associated with inhaled corticosteroid response among children with asthma. *Genes*, 15(4):420, 2024.

[59] Tesfaye B Mersha and Tilahun Abebe. Self-reported race/ethnicity in the age of genomic research: its potential impact on understanding health disparities. *Human genomics*, 9(1):1, 2015.

[60] Esteban González Burchard, Elad Ziv, Natasha Coyle, Scarlett Lin Gomez, Hua Tang, Andrew J Karter, Joanna L Mountain, Eliseo J Pérez-Stable, Dean Sheppard, and Neil Risch. The importance of race and ethnic background in biomedical research and clinical practice, 2003.

[61] Richard S Cooper. Race and genomics. *The New England journal of medicine*, 348(12):1166, 2003.

[62] Kellee White, Jourdyn A Lawrence, Nedelina Tchangalova, Shuo J Huang, and Jason L Cummings. Socially-assigned race and health: a scoping review with global implications for population health equity. *International journal for equity in health*, 19:1–14, 2020.

[63] Daphne O Martschenko, Hannah Wand, Jennifer L Young, and Genevieve L Wojcik. Including multiracial individuals is crucial for race, ethnicity and ancestry frameworks in genetics and genomics. *Nature genetics*, 55(6):895–900, 2023.

[64] Giorgio Sirugo, Sarah A Tishkoff, Scott M Williams, et al. The quagmire of race, genetic ancestry, and health disparities. *The Journal of clinical investigation*, 131(11), 2021.

[65] Iman Hamid, Katharine L Korunes, Sandra Beleza, and Amy Goldberg. Rapid adaptation to malaria facilitated by admixture in the human population of cabo verde. *Elife*, 10:e63177, 2021.

[66] Iman Hamid, Katharine L Korunes, Daniel R Schrider, and Amy Goldberg. Localizing post-admixture adaptive variants with object detection on ancestry-painted chromosomes. *Molecular Biology and Evolution*, 40(4):msad074, 2023.

[67] Meng Lin, Danny S Park, Noah A Zaitlen, Brenna M Henn, and Christopher R Gignoux. Admixed populations improve power for variant discovery and portability in genome-wide association studies. *Frontiers in genetics*, 12:673167, 2021.

[68] Nick Patterson, Neil Hattangadi, Barton Lane, Kirk E Lohmueller, David A Hafler, Jorge R Oksenberg, Stephen L Hauser, Michael W Smith, Stephen J O'Brien, David Altshuler, et al. Methods for high-density admixture mapping of disease genes. *The American Journal of Human Genetics*, 74(5):979–1000, 2004.

[69] Eva Suarez-Pajes, Ana Díaz-de Usera, Itahisa Marcelino-Rodríguez, Beatriz Guillen-Guio, and Carlos Flores. Genetic ancestry inference and its application for the genetic mapping of human diseases. *International journal of molecular sciences*, 22(13):6962, 2021.

[70] Michael W Smith, Nick Patterson, James A Lautenberger, Ann L Truelove, Gavin J McDonald, Alicja Waliszewska, Bailey D Kessing, Michael J Malasky, Charles Scafe, Ernest Le, et al. A high-density admixture map for disease gene discovery in african americans. *The American Journal of Human Genetics*, 74(5):1001–1013, 2004.

[71] Andréa RVR Horimoto, Diane Xue, Timothy A Thornton, and Elizabeth E Blue. Admixture mapping reveals the association between native american ancestry at 3q13. 11 and reduced risk of alzheimer's disease in caribbean hispanics. *Alzheimer's Research & Therapy*, 13:1–14, 2021.

[72] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, 25(14):1754–1760, 2009.

[73] Artem Tarasov, Albert J Vilella, Edwin Cuppen, Isaac J Nijman, and Pjotr Prins. Sambamba: fast processing of ngs alignment formats. *Bioinformatics*, 31(12):2032–2034, 2015.

[74] Ryan Poplin, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T Afshar, et al. A universal snp

and small-indel variant caller using deep neural networks. *Nature biotechnology*, 36(10):983–987, 2018.

[75] Michael F Lin, Ohad Rodeh, John Penn, Xiaodong Bai, Jeffrey G Reid, Olga Krasheninina, and William J Salerno. Glnexus: joint variant calling for large cohort sequencing. *BioRxiv*, page 343970, 2018.

[76] Olivier Delaneau, Bryan Howie, Anthony J Cox, Jean-François Zagury, and Jonathan Marchini. Haplotype estimation using sequencing reads. *The American Journal of Human Genetics*, 93(4):687–696, 2013.

[77] Zan Koenig, Mary T Yohannes, Lethukuthula L Nkambule, Xuefang Zhao, Julia K Goodrich, Heesu Ally Kim, Michael W Wilson, Grace Tiao, Stephanie P Hao, Nareh Sahakian, et al. A harmonized public resource of deeply sequenced diverse human genomes. *Genome Research*, 2024.

[78] Alex Mas-Sandoval, Sara Mathieson, and Matteo Fumagalli. The genomic footprint of social stratification in admixing american populations. *Elife*, 12:e84429, 2023.

[79] David H Alexander and Kenneth Lange. Enhancements to the admixture algorithm for individual ancestry estimation. *BMC bioinformatics*, 12:1–6, 2011.

[80] Anna CF Lewis, Santiago J Molina, Paul S Appelbaum, Bege Dauda, Anna Di Rienzo, Agustin Fuentes, Stephanie M Fullerton, Nanibaa'A Garrison, Nayanika Ghosh, Evelynn M Hammonds, et al. Getting genetic ancestry right for science and society. *Science*, 376(6590):250–252, 2022.

[81] Ethnicity Committee on the Use of Race, Engineering National Academies of Sciences, and Medicine. Using population descriptors in genetics and genomics research: A new framework for an evolving field. *NATIONAL ACADEMIES PRESS*, 2023.

[82] Helena Machado and Rafaela Granja. Emerging dna technologies and stigmatization. *Forensic genetics in the governance of crime*, pages 85–104, 2020.

[83] Florian Privé, Hugues Aschard, Andrey Ziyatdinov, and Michael G.B. Blum. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics*, 34(16):2781–2787, 2018.

[84] Ani Manichaikul, Josyf C Mychaleckyj, Stephen S Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, 2010.

[85] Ying Zhou, Sharon R Browning, and Brian L Browning. A fast and simple method for detecting identity-by-descent segments in large-scale data. *The American Journal of Human Genetics*, 106(4):426–437, 2020.

[86] Brian L Browning and Sharon R Browning. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, 194(2):459–471, 2013.

[87] Hans Bisgaard. The copenhagen prospective study on asthma in childhood (copsac): design, rationale, and baseline data from a longitudinal birth cohort study. *Annals of Allergy, Asthma & Immunology*, 93(4):381–389, 2004.

[88] Matthew Gray and Diana Smart. Growing up in australia: the longitudinal study of australian children: a valuable new data source for economists. *Australian Economic Review*, 42(3):367–376, 2009.

[89] Alice B Popejoy. Too many scientists still say caucasian. *Nature*, 596(7873):463–463, 2021.