Identity-by-descent captures Shared Environmental Factors at Biobank Scale

Franco Marsico[1, 2, †], Silvia Buonaiuto[1], Ernestine K Amos−Abanyie[1], Lokesh K Chinthala[3], Akram Mohammed[3], Regeneron Genetics Center[4], Robert J Rooney[1,5], Robert W Williams[1,6], Robert L Davis[3], Terri H Finkel[7], Chester W Brown[1, 8], Pjotr Prins[1], and Vincenza Colonna[1, 2, 5, †]

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

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

[3]Center for Biomedical Informatics, UTHSC, USA

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

[5]Dept of Pediatrics, UTHSC, USA

[6]Center for Integrative and Translational Genomics, UTHSC, USA

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

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

[†]Corresponding authors: vcolonna@uthsc.edu, fmarsic1@uthsc.edu

Abstract

The apple does not fall far from the tree is an old idiom that encapsulates a key concept: being related extends beyond merely sharing genetic material. It often implies sharing a common environment, including culture, language, dietary habits, and geographical location. In this study, we show that the analysis of genetic relatedness can serve as an indicator of health conditions by capturing the combined influences of

genetic inheritance and shared environmental factors. We mapped the genomic data and electronic health records from 13k individuals in the Integrative Genomics Biorepository cohort, to neighborhood-level geographic data and integrated census-based environmental metrics. We used an identity-by-descent (IBD) based hierarchical community detection algorithm to identify four main communities closely aligning with continental ancestry, and seventeen subcommunities. We found uneven exposure to elevated environmental stressors across subcommunities, and we were able to identify subcommuinities at high risk of respiratory and dermatological conditions, demonstrating the potential of our framework and its possible application to public health interventions. Interestingly, for conditions such as congenital disorders, the most important differences were detected between subcommunities within the same community, indicating the relevance of considering a fine-grained population structure. These findings show how genetic relatedness, even at distant levels, can reflect shared environments and social determinants of health, providing a framework for understanding health disparities.

1 Introduction

Identity-by-Descent (IBD) refers to DNA segments shared by two individuals that are identical due to inheritance from a common ancestor [1]. As the genealogical distance between individuals increases, the total amount of IBD shared decreases. This reduction occurs due to recombination events, which progressively fragment inherited genomic segments into shorter and fewer pieces across generations. High total IBD values are linked to close relationships (first and second degree) which are the foundation of many family systems [2]. Usually, these relationships reflect not only biological ties, but also shared culture, social norms, and environments. Smaller IBD values indicate more distant genetic connections, but these can still influence broader kinship structures. Even distant relatives may share cultural traits passed down through generations, shaped by slow-changing social norms, as observed in pioneering works by Cavalli-Sforza and Feldman [3, 4].

To study population structure, the traditional practice of categorizing individuals into discrete genetic groups faced significant epistemological challenges that extend beyond technical limitations. Common approaches impose artificial boundaries on what research has shown is a fundamentally continuous spectrum of human genetic variation [5, 6]. These boundaries often arise from methodological frameworks, such as imposing predefined reference populations [5, 7, 8, 9]. Despite the mentioned limitations, comparative analysis between groups remains relevant for understanding biological mechanisms, providing crucial insights into evolutionary processes and health conditions when properly analyzed [8, 10]. Alternative approaches, such as IBD-based categories, have been proposed as reference-free frameworks to study population structure and grouping individuals [11, 12, 13, 14, 15, 16, 17, 18, 19].

In this study, we used data from the Biorepository of Integrative Genomics (BIG) cohort [20]. A key feature of the BIG is its representation of high genetic diversity concentrated within a relatively confined geographic area. This characteristic enables us to investigate the complex interplay between genealogical relationships and environmental determinants using a standardized data source. The BIG resource provides comprehensive electronic health records, neighborhood-level environmental data linked through ZIP codes, and genomic information from 13,143 participants, with particular emphasis on pediatric populations. We show how IBD-based clustering bypasses reference-dependent limitations by directly grouping individuals into communities by pairwise genealogical relationships. This clustering also reflects shared environmental exposures. Furthermore, our analysis of health conditions reveals that subcommunities defined by IBD exhibit distinct patterns of disease risk that can inform targeted public health interventions.

2 Results

2.1 IBD network reveals population structure

To date, the BIG initiative has enrolled more than 42,000 participants with EHR and collected more than 15,000 biosamples at five sites (Supplementary Figure S1). The population is predominantly pediatric (87% under 18 years old), with ages from infancy to 90 years (mean: 8.4, median: 6.2) [20]. BIG is notable for its diverse ancestry and focus on children with diseases, unlike typical pediatric cohorts of healthy mother-child pairs [20]. Within the BIG cohort, we identified and phased 6.8 million high-confidence variable sites through exome sequencing and genotype-by-sequencing data from 13,143 individuals.

To study the genetic relationship between BIG individuals and also qualitatively describe the structure of our cohort by comparison with known labeled populations, we combined BIG data with 4,091 genomic sequences from the joint HGDP-1000GP resource [21]. In the combined database, pairwise IBD was calculated [22]. The segments obtained were further processed to avoid overlaps and, for each pair of individuals, the total IBD was calculated (Figure 1A). The observed exponential decay suggests that distant relative pairs far outnumber close kinships, a trend consistently documented in other cohorts [18, 14].

A network was constructed based on pairwise IBD. The 17,234 nodes represent individuals, with edge lengths inversely proportional to shared IBD. The representation layout was constructed using the Fruchterman-Reingold algorithm [23] (Figure 1B). Visual inspection indicates that four main dense groups ($M1$ to $M4$) can be easily distinguished. They are surrounded by smaller clusters in sparse regions.

Hierarchical community detection was performed using the Leiden algorithm to further characterize the cohort [24]. The first iteration identified ten communities, of which four contained more than 1,000 individuals: $C1$, $C2$, $C3$, and $C4$ (Figure 1C). Subsequent iterations of the Leiden algorithm within each community revealed 72 subcommunities in total. $C2$ exhibited the highest number of subcommunities, followed by $C4$, $C1$, and $C3$. Interestingly,

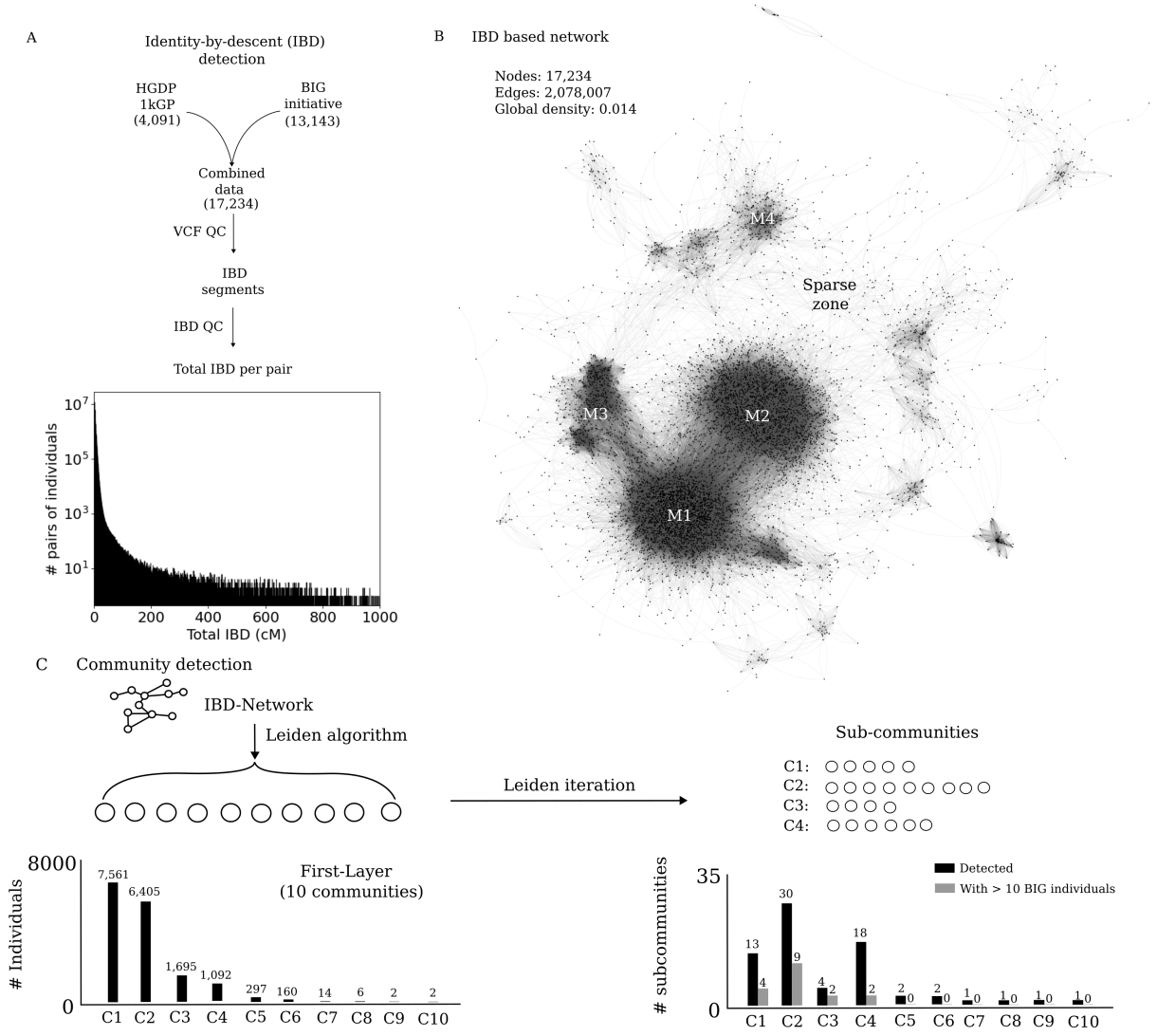

Figure 1: **Distribution of identity-by-descent (IBD) values and corresponding network analysis.** (A) Histogram showing the exponential decay of total IBD (cM). To simplify, the graph focuses on total IBD values below 1000 cM, highlighting lower values to investigate the structure of the population [14]. (B) IBD-based network with 17,234 nodes, revealing four dense communities (M1 to M4) and sparse zones. Fruchterman-Reingold layout was used to do the plot. (C) Bar plot of community sizes after the first Leiden iteration, with C1 to C4 containing over 1,000 individuals. (D) Number of subcommunities distribution, showing no direct correlation with parent community sizes.

the number of subcommunities did not correlate with the size of the parent community, indicating the potential impact of other factors, such as genetic diversity.

Subcommunities containing more than 10 individuals from the BIG cohort were selected for further analysis, resulting in 17 subcommunities that collectively represent 99.7% of the BIG samples, while the remaining 0.3% corresponds to individuals dispersed across the other 55 groups, placed closer to the reference populations.

2.2 BIG communities align with continental level ancestry

We first analyzed the top-level communities, which include individuals from the HGDP-1kGP and BIG datasets. The proportion of individuals from each HGDP-1kGP population was quantified within each community. As shown in Figure 2A, $C1$ is predominantly composed of individuals of European origin, including the CEPH, Iberian, Finnish, British, and Toscani populations. $C2$ is primarily associated with African populations, notably Yoruba, Esan, the African Caribbean, and the African Ancestry SW (African Americans from the USA). $C3$ predominantly represents American populations, including individuals from Puerto Rico, Colombia, Peru, and Mexico. $C4$ mainly consists of Asian populations, with a strong representation of Han Chinese individuals.

Global ancestry deconvolution for BIG individuals within each community was performed using RFMix [9]. A reference panel for continental-level ancestry was used (Supplementary Figure S2). Figure 2B illustrates the composition of the continental ancestry, including admixed individuals (two-way and multiway). Consistent with the patterns observed in Figure 2A, $C1$ shows enrichment in EUR (91%) ancestry, $C2$ in AFR (49%) and EUR-AFR (39%) admixed ancestry, $C3$ in EUR-AMR (51%) and multiway (43%) ancestry, and $C4$ in EAS (81%) ancestry. A chi-square test confirmed a highly significant association between community membership and inferred global ancestry (p-value < 0.0001; Cramér's V $= 0.81$).

After characterizing the top-level communities, we analyzed the derived subcommunities. We constructed a network where the nodes represent subcommunities and the edges are in-

versely proportional to the total IBD shared between the individuals belonging to them (Figure 2C). In particular, $C2$ produced the highest number and greatest diversity of subcommunities, as evidenced by the graph (Figure 2C) and the pairwise Fst analysis (Supplementary Figure S3). In contrast, the $C1$ subcommunities are more densely interconnected, reflecting closer genetic relationships between individuals. $C3$ subcommunities exhibit links to both $C1$ and $C2$, suggesting shared genetic characteristics between these groups.

2.3 Demographic patterns across subcommunities

To further characterize the cohort, we calculated the mean share of IBD between BIG individuals in subcommunities and HGDP-1kGP populations. As expected, the subcommunities derived from $C1$ were predominantly related to European populations, $C2$ to African, $C3$ to American, and $C4$ to Asian (Figure 3A). Within $C3$, distinct patterns emerged: the subcommunity $C3_2$ showed a stronger relationship with the Puerto Rican population, characterized by its three-way ancestry pattern [11, 25], while $C3_1$ was more closely associated with Peruvian and other Native American populations. In $C2$, similar trends were observed in all subcommunities, with the exception of $C2_6$, where the Surui showed relevant connections. For $C4$, $C4_0$ exhibited clearer relationships with reference Asian populations, while $C4_15$ showed sparse connections. In $C1$, subcommunities showed a uniform pattern, with higher connections to the British and CEPH HGDP-1kGP populations.

Figure 3B shows that the subcommunities derived from $C1$ were predominantly self-reported as white (91.3% of individuals), while those from $C2$ were primarily self-reported as black or African American (97.1%). In $C4$, the majority self-reported as Asian (62.4%) and Unknown (23.7%). In $C3$, different patterns were observed in subcommunities: $C3_1$ reported predominantly as OthersUnknown (73.1%) and white (8.5%); while $C3_2$ reported OtherUnknown (33.2%), white (21.1%) and Black or African American (21.2%). Consistently with previous community-level observations, $C1$ subcommunities reported mainly EUR inferred ancestry, $C2$ AFR and EUR-AFR, $C3$ EUR-AMR and Multiway, and $C4$ EAS (Figure 3C).

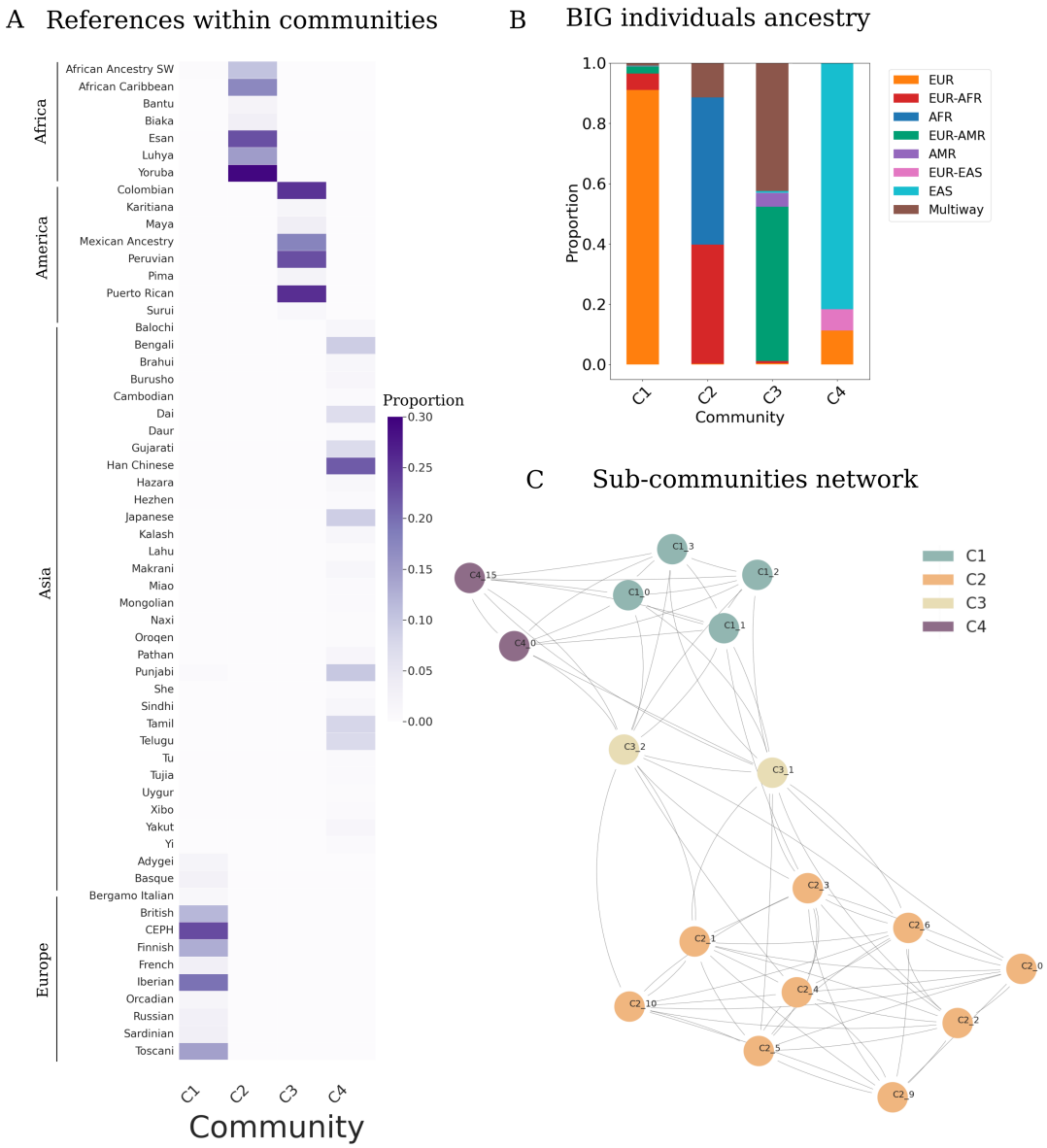

Figure 2: **Genetic community characterization based on ancestry.** (A) Heatmap of reference population proportions within the four main communities (C1: European, C2: African, C3: American and multiway, C4: Asian). (B) Proportions of inferred continental and admixed ancestries per community. (C) Subcommunity network showing IBD-based connections, highlighting diversity and interconnectivity (Fruchterman-Reingold layout).

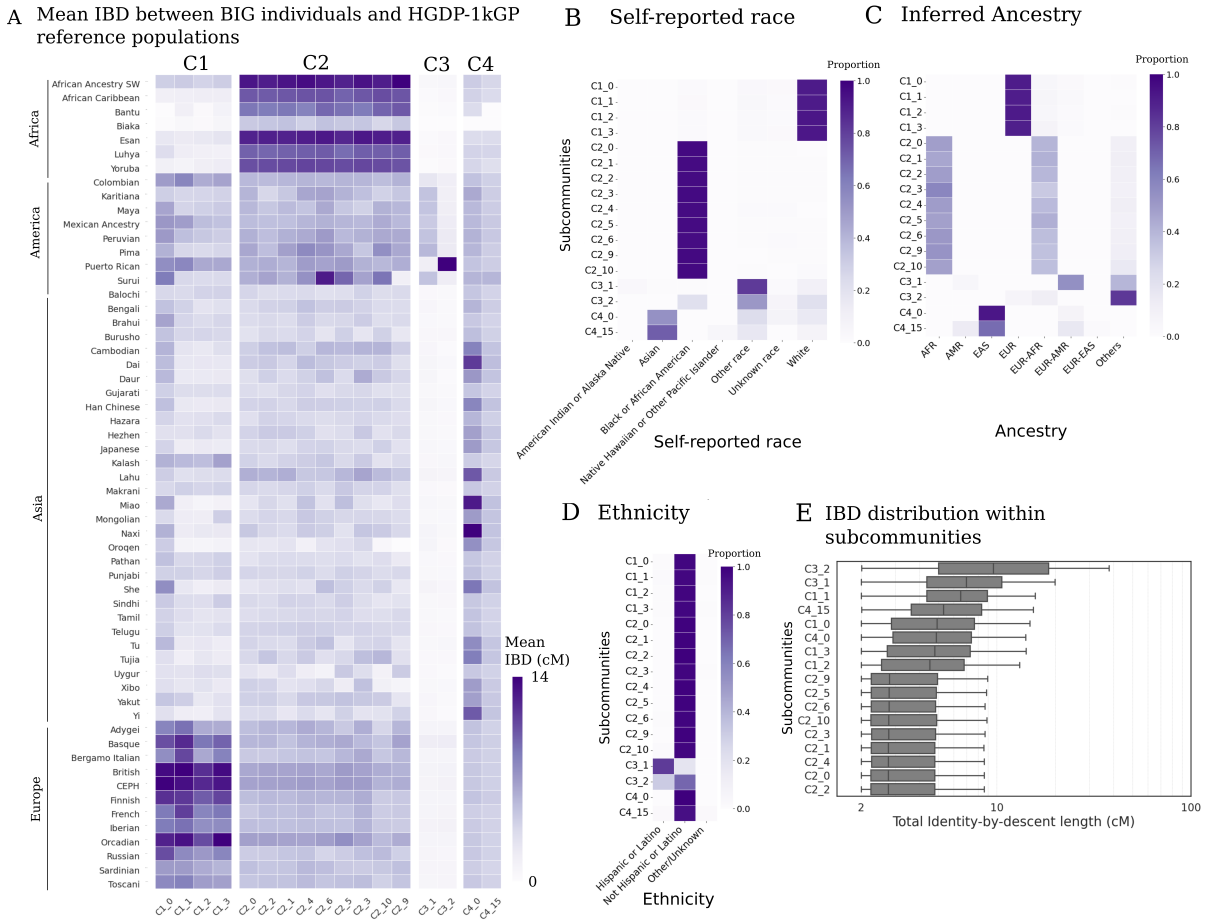

Figure 3: **Genetic and demographic patterns across sub-communities.** (A) Heatmap showing mean IBD values between subcommunities and reference populations, highlighting distinct genetic relationships across C1 to C4. (B) Distribution of self-reported race, (C) inferred ancestry and (D) ethnicity proportions within subcommunities. (E) Box plot showing total IBD values per subcommunity.

In terms of ethnicity (Figure 3D), almost all subcommunities were self-reported mainly as Not Hispanic or Latino. Notably, within $C3$, clear differences were observed: Individuals in $C3_1$ were more likely to self-identify as Hispanic or Latino (72.3%), while those in $C3_2$ predominantly self-reported as Not Hispanic or Latino (69.6%).

The distribution of IBD among the BIG individuals within each group (Figure 3E) revealed that $C3$ subcommunities exhibited the highest mean IBD values, followed by $C1$ and $C4$. In contrast, the $C2$ subcommunities had the lowest mean IBD values, reflecting less dense genetic relationships, consistent with the greater genetic diversity detected in this group.

2.4 IBD reflects shared environments

To elucidate the spatial distribution of genetic relationships, we examined whether subcommunities exhibit distinct geographical patterns and how these relate to IBD. We geolocated BIG participants' ZIP codes at the neighborhood level using publicly available spatial polygons [26]. We detected a strong association between genealogical and geographical distances (exponential fit $R^2 = 0.969$, p-value < 0.0001, Figure 4A, Supplementary Figure S4). Close relatives (first and second degree) lived, on average, between 4 and 6 km from each other, while 49% of the pairs shared the same location (considering the neighborhood level as the minimum granularity). The more distant kin, of the third and fourth degree, maintained on average a higher separation, but less than 20 Km, indicating substantial local clustering with 34.3 % sharing location (for distance references, see Supplementary Figure S1). In contrast, relationships beyond the seventh or eighth degree exhibited a marked increase in spatial separation. In particular, for eight-degree or more distant relatives, the proportion of pairs who shared the same neighborhood falls to 4.45%.

Figure 4B shows the median geographic distances between all individuals at the community level. For $C1$ the distances are the highest, indicating a more sparse population (88.6 Km). In contrast, for $C2$ (18.9 Km) and $C3$ (22.1 Km) the distances were the lowest, indicating a more concentrated population. The SDoH similarity index reveals that, although sparse, $C1$

subcommunities show similar environments (mean of 96%). More heterogeneity is observed for $C2$ (mean of 83%), also showing similarities with subcommunities $C3$.

We examined in more detail key SDoH factors due to their direct impact on clinical outcomes. As shown in Figure 4D, subcommunities within group $C2$ exhibit elevated $PM_{2.5}$ EJ index values, indicative of higher pollution, coupled with increased urban residency and smoking prevalence. Moreover, the Social Vulnerability Index reveals that a substantial fraction of $C2$ resides in high-risk areas, a trend also seen, though less pronounced, in $C3$. In line with Figure 4C, $C1$ shows a uniform SDoH profile.

2.5 Health condition patterns in subcommunities

To assess the use of health care, both to analyze inequalities and also to account for under-registration in specific groups, we calculated the rate of visits per year in each subcommunity (Figure 5A). We found similar trends in all subcommunities with an overall median of 7.7 visits per year, ranging from 6.9 in $C2_5$ to 9.1 in $C3_1$. Interestingly, the median of the total number of visits per individual considering the entire cohort was 38 times, ranging from 35 to 53 in the subcommunities, indicating the relevant monitoring of health conditions in all cases (Supplementary Figure S5).

Figure 5B provides a global comparison of the prevalence of selected phenotypes within each of the four main communities ($C1$, $C2$, $C3$, $C4$). Some conditions are enriched in the $C2$ community, such as respiratory (71.4%), infections (56.8%) and dermatological (50.7%). In contrast, others present a more homogeneous distribution, for example, neurological, genetic, and congenital.

Figure 5C refines this analysis by illustrating the proportion of conditions at the sub-community level. The results reveal distinct patterns at the community level, showing high proportions of individuals with respiratory (71.4%), infectious (56.2%), and dermatological (50.9%) conditions in $C2$ subcommunities. This finding is consistent with previous results (Figure 4D), where these subcommunities demonstrated high exposure to environmental risk

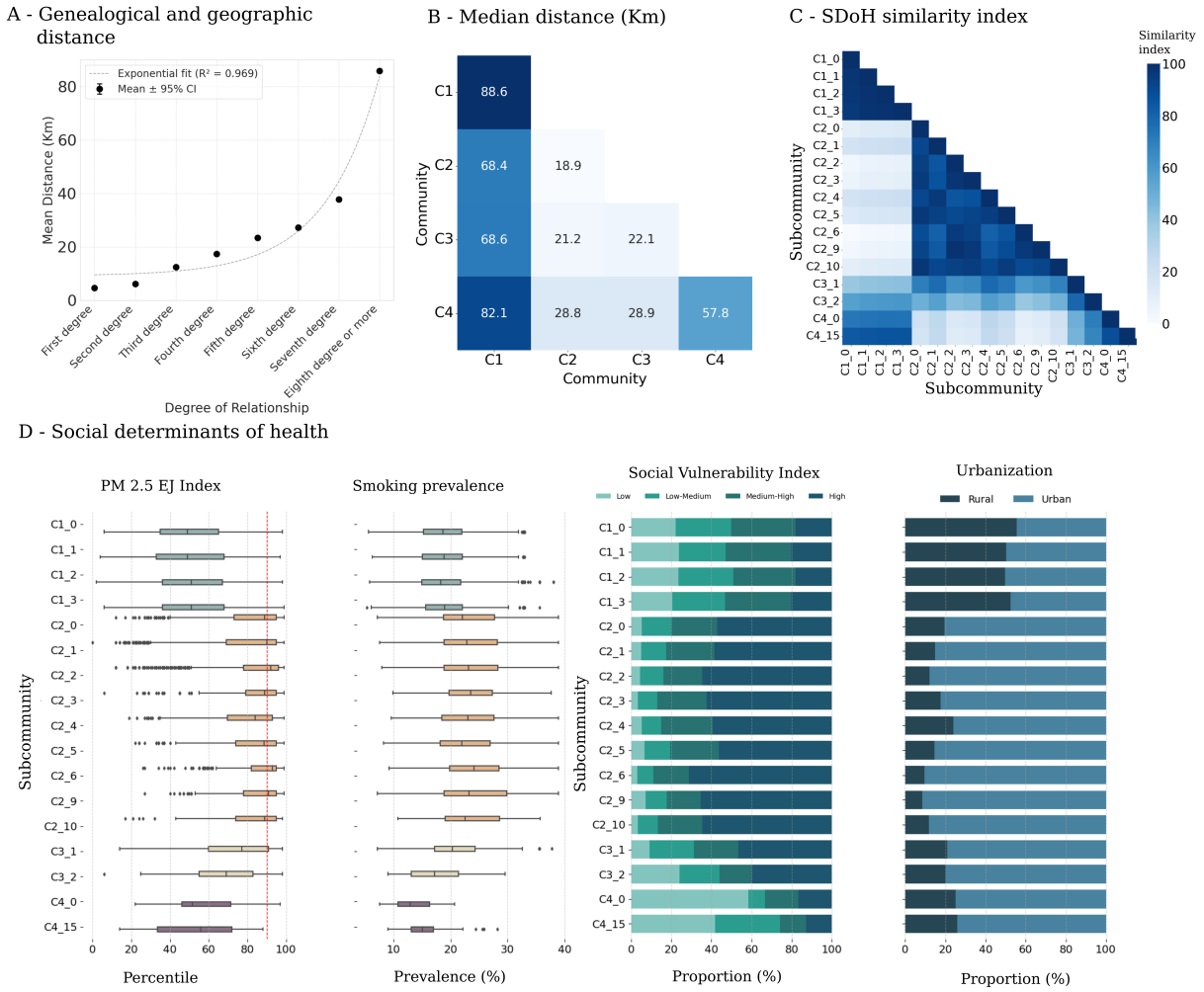

Figure 4: **Geographical analysis and environmental factors** (A) Genealogical (degree of relationship) and geographical (Km) distances. Exponential fit is shown. (B) Median geographic distances in Kilometers, highlighting variability in spatial distributions within and across communities. (C) Social determinants of health similarity index at neighborhood level between communities (the diagonal of the matrix shows the maximum (100%) identity. (D) Social determinants of health distributions for each subcommunity. PM 2.5 EJ index, Environmental Justice metric derived from the ambient concentration of fine particulate matter. Red line indicates percentile 90th, which implies high risk. Smoking prevalence, social vulnerability index and urbanization.

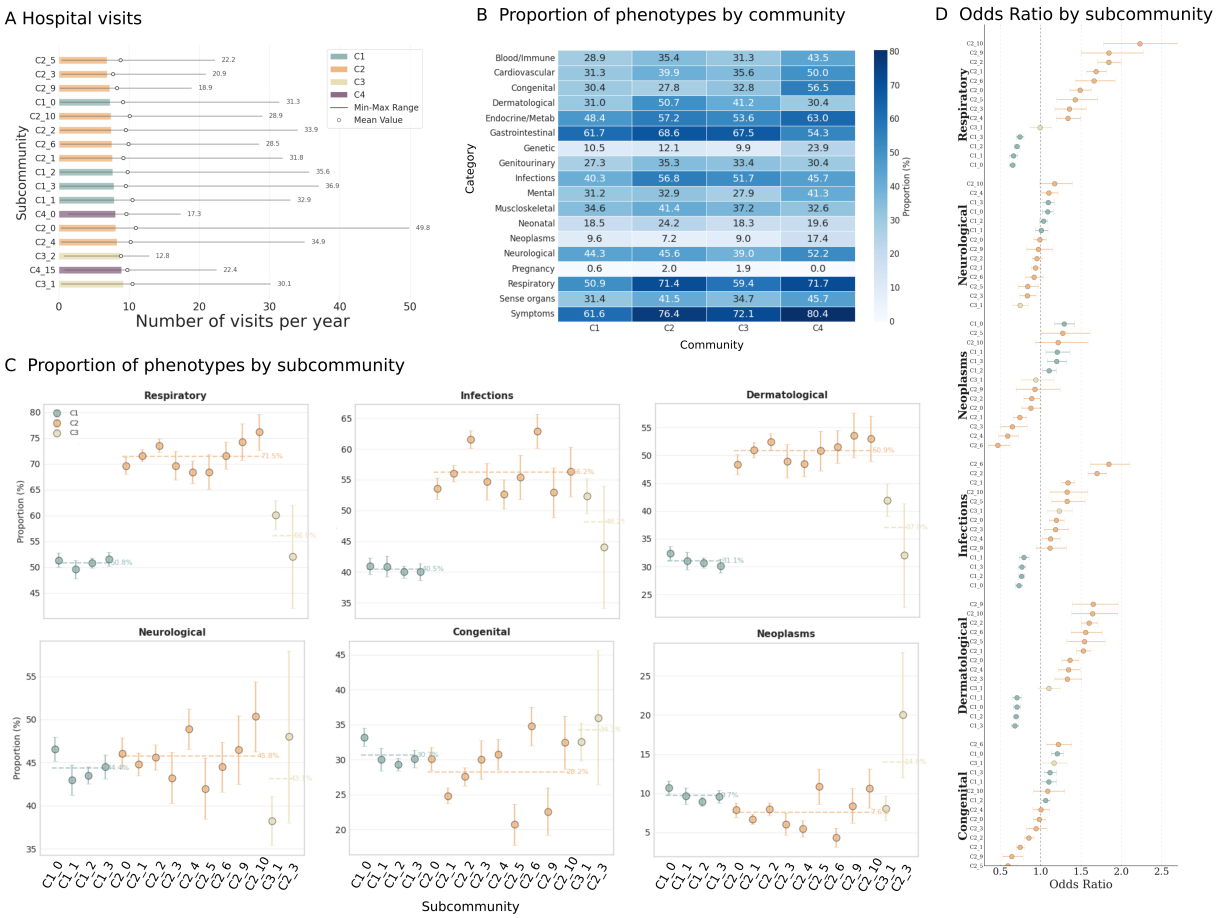

Figure 5: **Health conditions across subcommunities**(A) Bar plot illustrating the median number of hospital visits per year in each subcommunity. Mean and maximum are also indicated (B) Heatmap showing the proportion of phenotypes at the main community level (C1, C2, C3, C4), with each cell reflecting the percentage of individuals displaying a specific category (e.g., Respiratory, Neurological, Infections, etc.). (C) Proportion of phenotypes at the subcommunity level and CI95, colored by community C1, C2, C3. Subcommunities from C4 were excluded due to the high uncertainty associated with fewer individuals (complete results are shown in Supplementary Figure S6) (D) A Odds ratio (OR) of each subcommunity relative to the overall proportion (OR = 1 as the reference line). Points indicate the point estimates of the OR, and the horizontal whiskers represent CI95.

factors (for example, pollution).

Interestingly, in congenital disorders (that appeared generally similar across communities (Figure 5B)), subcommunity analysis reveals differences. For instance, subcommunities C2_5 and C2_9 exhibit lower prevalence of congenital conditions (Figure 5C).

Finally, Figure5D shows the odds ratios (OR) for each subcommunity, comparing their prevalence of the phenotype with the general cohort (OR = 1 as a reference). In particular, some C2 subcommunities are prominent with an OR of 2.22 [CI95 1.78-2.79] ($C2_2$) for respiratory conditions, 1.83 [CI95 1.61-2.01] ($C2_6$) for infections, and 1.65 [CI95 1.39-1.95] ($C2_9$) for dermatological conditions. Consistently, for congenital conditions, C2_5 and C2_9 exhibit low risk (OR 0.59 [CI95 0.49-0.71] and 0.64 [CI95 0.53-0.78] respectively).

2.6 IBD informativeness in pediatric-related health conditions

Here we compare the informativeness of IBD-based classification, ancestry, and SDoH for health conditions that are particularly relevant in pediatric populations. To assess the predictive utility of IBD-based community grouping, we fitted logistic regression models for two outcomes (Dermatitis and Asthma) using McFadden's pseudo R^2 as a measure of model fit. The rationale for selecting these pathologies is based on the trend observed in the respiratory and dermatological categories, which contain these health conditions. Three different sets of predictors were used: (i) Subcommunity, (ii) Ancestry, and (iii) SDoH.

The subcommunity variable showed the highest predictive utility for both Dermatitis (pseudo R^2=0.05), and Asthma (pseudo R^2=0.04), outperforming models based on Ancestry and SDoH (Table 1, Supplementary Tables S2, S3, S4, S5, S6, S7). These values align with those typically reported in pediatric epidemiological studies [27, 28, 29]. In particular, polygenic risk models for asthma and eczema often fall within the 0.03 to 0.1 range [27, 28]. Moreover, this study confirm the previous observation of C2 subcommunities with a higher prevalence of dermatitis and asthma (p-value < 0.001, Supplementary Tables S3, S6).

Health Condition	Subcommunity	Ancestry	SDoH
Dermatitis	0.05	0.03	0.04
Asthma	0.04	0.02	0.02

Table 1: Predictive utility of IBD-based community classification for pediatric health conditions. McFadden's pseudo R^2 values from logistic regression models using four predictor sets for three clinical outcomes.

These results also suggest that subcommunities may be capturing relevant information beyond measured environmental factors and inferred ancestry.

3 Discussion

In this work, we have shown that IBD can be used to group individuals without relying on reference populations. We also demonstrated that, at a broader level, this grouping aligns with continental ancestry. We then studied how detected subcommunities can inform about shared environments, leading to distinct patterns of social determinants of health. In addition, these subcommunities presented different prevalences of health conditions. Together, these results demonstrate the utility of considering both close and distant relationships between individuals for targeted public health interventions.

Recent discussions focused on the risks associated with the use of labels in human groups, especially those conflated with historically discriminatory categories such as race and ethnicity [30, 31, 32, 33, 34]. Typically, genetic studies categorize individuals based on continental-level labels, yet such classifications often overlook actual genetic relatedness. Our previous work demonstrated instances where individuals from different self-reported racial categories were closely related (i. e, second degree) [20]. Similarly, ancestry-based classifications can separate genetically related individuals into distinct groups, particularly in cases involving admixed ancestry [25]. In contrast, IBD-based hierarchical community detection allows grouping individuals by their direct genetic relationships without relying on externally imposed labels

based on arbitrarily defined reference populations.

Other studies have explored IBD-based grouping, also detecting its correlation with geographical patterns in the United Kingdom (UK) [18]. However, a particularity of the BIG cohort is its highly geographically concentrated genetic diversity, mainly in Memphis City [20]. To put in context, the UK is roughly 748 times larger than Memphis. We show that the correlation of IBD and geography is also relevant at the city and suburban level. In this case, we found neighborhood-level patterns, indicating that fine-grain characterization of the population is still relevant. This work focuses on the entire cohort characterization; most of the analyzes were based on the 17 main subcommunities that represent the predominant demographic structure. We show how subcommunity-level patterns emerge, providing information for risk assessment and public health interventions, identifying groups that can be at risk for genetic and environmental factors. Furthermore, the network-based approach offers additional clinical utility by identifying and characterizing outlier individuals within minor subcommunities, allowing targeted investigation of health conditions potentially linked to their distinct genetic background [14, 35].

In addition, population stratification, admixture, and shared environmental factors represent significant sources of confounders in genome-wide association studies (GWAS) and admixture mapping, potentially leading to false positives and negatives [36, 37, 38, 39]. Addressing these confounders through methods such as principal component adjustment, mixed-model association analyzes, and inclusion of local ancestry covariates has become critical to accurately identify true genotype-phenotype associations [39, 36]. Usually, in GWAS studies, only close relationships or ancestry are taken into account [36], and population structure adjustments are introduced by performing principal component analysis and incorporating the top principal components (PCs) as covariates. However, it has been observed that the introduction of PCs can lead to spurious associations in GWAS studies when including admixed populations [40]. IBD-based grouping strategy inherently integrates distant genetic relatedness, ancestry, and shared environmental factors. Using IBD grouping to adjust population structure in association studies can be a promising line of future research.

Despite its advantages, the IBD-based grouping approach has limitations. Its effectiveness may depend on the population density, degree of local genetic structure, and availability of high-quality genomic and geographic data [41]. Future work should evaluate the integration of IBD-based categorization with existing correction strategies in GWAS [42] and gene-based association studies [43], particularly in cohorts with varying levels of endogamy, admixture, and environmental heterogeneity [38]. Another aspect is the definition of relatedness, which lacks a universally accepted formulation. As discussed by Speed and Balding [44], different measures, based on pedigree, allele, and segment sharing, reflect different interpretations of what it means to be related. In this study, we define relatedness as the total length of genome-shared IBD, a direct, quantitative measure of recent co-ancestry. Although often categorized for interpretability, this metric captures a continuous spectrum of genetic similarity.

Finally, as we showed, a deeper understanding of our communities can be useful in studying health disparities and providing a framework for public health interventions.

4 Methods

4.1 Ethics

Our research followed the ethical guidelines established in the Declaration of Helsinki for human subject medical research. We conducted this study according to approved ethical standards with authorization from the University of Tennessee Health Science Center's Institutional Review Board (IRB number: 23-09204-NHSR). All participants provided written informed consent, with parents or legal guardians giving consent for children involved in the study. To protect privacy, we removed all identifying information from the data before conducting our analysis.

4.2 Sample Collection

Our study collected data from four key locations across Tennessee: (i) Le Bonheur Children's Hospital (Memphis) serves predominantly African American pediatric populations in a health-disparate area. Recruitment began in October 2015 across all hospital settings. Patient distribution follows a distance gradient (Supplementary Figure S5). Genomic DNA comes from leftover clinical blood linked to de-identified EHR data, with collection-consent discrepancies due to sample availability. (ii) Regional One Health (Memphis) joined in May 2022, focusing on adult genomic research from similar populations, with identical DNA collection procedures complementing LBCH's pediatric focus. (iii) East Tennessee State University (Johnson City) added in May 2023 represents Appalachian populations through dedicated adult blood draws, supporting BIG's commitment to rural communities. (iv) Family Resilience Initiative (Memphis), launched in January 2019, studies adverse childhood experiences through mother-child dyads, collecting biological samples quarterly over 18 months for DNA isolation, cortisol measurements, and clinical assessments to link biological and environmental data.

4.3 Clinical Data

BIG participants' clinical information is extracted from Electronic Health Records (EHR) as flat files and transferred to UTHSC using secure protocols. This information encompasses demographics, hospital visits, diagnoses, procedures, medication details (both prescribed and administered), laboratory results, and vital measurements. We convert these elements into a limited data set (LDS) and align them with the OMOP (Observational Medical Outcomes Partnership) common data model. To facilitate analysis, we map ICD9/10 diagnosis codes to their corresponding PheCodes. Each PheCode was also associated to a broader category (Respiratory, Dermatological, Neurological, among others, see Supplementary Figure S6. Disease conditions were defined using specific PheCodes: asthma was identified using Phecode RE_475; dermatitis using PheCodes beginning with 690, which includes various subtypes of skin inflammation conditions.

4.4 DNA Sequencing, quality control and variant analysis

All samples were sequenced on an Illumina NovaSeq 6000 platform. The 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, 95.2% achieved an average sequencing depth of at least 20X, and 99.3% of the samples had more than 90% of their bases covered at 20X or greater, 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%. All samples were sequenced on an Illumina NovaSeq 6000 system on S4 flow cells sequencer using 2×75 paired-end sequencing.

Sequence reads were aligned by the Burrows-Wheeler Aligner (BWA) MEM [45] 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 [46]. DeepVariant v0.10.0 with a custom exome model was used for variant calling [47], and the GLnexus v1.2.6 tool was used for joint variant calling [48]. The variants were annotated using a Variant Effect Predictor (VEP 110) [49]. Phasing was performed using ShapeIT v5 [50]. Our dataset comprised 6,886,631 variable sites after quality control, combining both exome capture and targeted sequencing data.

4.5 Ancestry Inference

To characterize genetic ancestry within the BIG cohort, we executed global ancestry inference employing RFMix [9]. Reference samples were extracted from the 1000 Genomes Project and Human Genome Diversity Project (HGDP), leveraging the recently developed joint call [21]. The merged dataset incorporated BIG participants and reference samples with autosomal variants.

To select reference samples, we implemented quality control protocols previously validated

in other investigations [25]. To eliminate reference samples with significant admixture, we performed unsupervised cluster analysis utilizing ADMIXTURE [51]. We established 4 groups (k = 4), and reference samples with a major group proportion >0.99 were retained for analysis. We opted for k = 4 to delineate continental level ancestry across four principal groups: African, American, European and Asian. This framework aligns with demographic patterns observed in major Americas cities, with the Asian group inclusion [25]. This last group was added considering self-reported race and ethnicity classifications.

Four-way ancestry deconvolution was conducted with the random forest classifier terminal nodes set to 5 (-n 5), expected addition generations average set to 12 (-G 12), and ten iterations of the expectation maximization algorithm (EM) (-e 10). Reference superpopulations at the continental level included African (AFR), American (AMR), European (EUR), and Asian (ASN). ASN incorporated East Asian (EAS) and Central South Asian (CSA) populations. CSA ancestry proved negligible (99% of the BIG cohort with values near 0, few cases below 0.075). As low global ancestry proportions are associated with inaccurate estimates, we excluded CSA from further analysis [52]. EAS was retained, showing a significant signal in cases consistent with self-reported Asian individuals. 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).

Discrete ancestry categories (AMR, AFR, EUR, EAS, EUR-AMR, EUR-AFR, and Multi-way) were established using these criteria: (i) individuals with >85% single ancestry were assigned to corresponding groups; (ii) individuals with >15% from two ancestries totaling >85% were classified as two-way admixed; (iii) individuals with >15% contribution from three or more ancestries were designated Multiway. The 85% threshold was selected because ancestry proportion is a continuous variable, requiring threshold definitions to be categorized, and ancestral contributions >10-15% are considered accurate and significant, while lower

proportions correlate with shorter ancestral segments and increased error rates [52, 53].

4.6 Identity-by-descent

To identify identity-by-descent (IBD) segments, we used hap-ibd in the phased dataset comprising 13,143 genomes from the BIG cohort and 4,091 genomes from the HGDP-1kGP, focusing exclusively on autosomal loci [22]. Hap-ibd was executed with a minimum seed parameter of 2 cM to focus on the detection of IBD segments reflecting recent common ancestry, given that segments of this length or greater provide information for fine-scale relationship inference [54].

Post-processing protocols established by Browning et al. [55, 56] were applied to refine detected segments. Specifically, the merge-ibd-segments tool was employed with default parameters, consolidating fragmented segments to improve continuity. Additionally, gaps containing at most one discordant homozygote and segments shorter than 0.6 cM were removed, significantly enhancing detection accuracy and efficiency in large-scale genomic datasets.

For quality control, segments overlapping centromeres, telomeres, the human leukocyte antigen region, and low complexity regions were excluded. To identify genome regions with abnormally high shared genetic material, we calculated total identity by descent at each marker position and removed those that deviated significantly from the genome-wide average. For downstream analysis, IBD segment lengths were summed across all chromosomes between each pair of individuals, generating the total shared IBD. Finally, only for illustrative purposes, we classified the degree of relationships based on total IBD by establishing non-overlapping thresholds based on previous studies [57, 54, 56].

4.7 Network and community detection

A genetic network was constructed wherein each individual was represented as a node, and edges were defined between individuals based on the total genome-wide IBD segment sharing.

To build this network, pairwise IBD data were processed and filtered, excluding pairs with total sharing longer than 1,000 cM to mitigate confounding introduced by close familial relationships [17]. For visualization and exploratory analyses, the network was represented using Fruchterman-Reingold layout.

We evaluated the Louvain and Leiden algorithms for community detection [24]. Both were implemented using a multilevel approach. This means that the algorithms were run iteratively, initially identifying broad-scale genetic structures and subsequently resolving increasingly finer-scale subcommunities up to three levels. Comparative analysis revealed an Adjusted Rand Index (ARI) of 0.975, indicating high similarity between community assignments from both methods. However, we selected the Leiden algorithm due to its improved resolution, robustness, and ability to avoid disconnected communities, a known limitation of Louvain [24].

The hierarchical Leiden algorithm was applied with modularity as the objective function, a resolution parameter of 0.3, and a beta value of 2. This method effectively detected hierarchical subcommunities within the network [17]. This means that, for example, a community detected at a broader level can be passed through the Leiden algorithm, leading to subcommunities.

Finally, genetic differentiation between communities was independently assessed using the fixation index (F_{ST}), calculated with PLINK. This analysis quantitatively characterized genetic distances between communities, offering independent validation and additional genetic context for the hierarchical community structures identified by the Leiden algorithm.

4.8 Geographical analysis

To analyze the geographical patterns of genetic relationships, we geolocated BIG cohort participants using ZIP code at neighborhood level data. Distances between neighborhoods were calculated using the sf R package [26], and therefore, we obtained a matrix of distances between individuals (with zero as the distance between those who shared the same neighborhood). A network based on the geographical distance was constructed, representing the subcommunities in the nodes and the inverse of the median distance for the edges. Geographical data was

integrated with the 24th USA census (2020)-based metrics. It provided several demographic measures and also social determinants of health.

4.9 About self-reported race

Patients self-identify their race when they enter the hospital. Hospital staff then choose the appropriate race category from a dropdown menu that follows HL7 standards for race and ethnicity (https://hl7-definition.caristix.com/v2/HL7v2.5/Tables/0005). The system allows selection of multiple race codes for individuals who identify with more than one racial background. However, due to inconsistencies in how records were collected in the past, some self-reported racial classifications were inaccurate or unsuitable. We therefore enhanced the data to create a more dependable classification system. Supplementary Table S1 outlines our specific refinement criteria in detail.

4.10 Social determinants of health

In our analysis, each subcommunity is characterized by aggregating Social Determinants of Health (SDoH) variables by neighborhood level. Therefore, we were able to assign SDoH to each BIG participant based on zip code. We analyzed SDoH variables from multiple sources. U.S. Census (2020) provided population density, poverty rates and Gini index. CDC contributed the Social Vulnerability Index (2020), obesity prevalence and smoking rates (2021). Geographic factors included USDA urban/rural designations (2019) and Treasury Department/CDFI metro classifications (2020). We incorporated unemployment rates (Treasury Department/CDFI, 2016-2020), and EPA's $PM_{2.5}$ particulate matter percentiles (2024).

Continuous variables (e.g., $PM_{2.5}$, $CSMOKING$) are averaged between individuals, while categorical variables ($A9_SVI$) are converted to ordinal scores by appropriate mapping and are similarly averaged. Each community is represented by a profile vector, and the Euclidean

distance between communities i and j is calculated as

$$d_{ij} = \sqrt{\sum_{k=1}^{C} (\bar{x}_{i,k} - \bar{x}_{j,k})^2}$$

where $\bar{x}_{i,k}$ is the mean value of the variable SDoH k for the community i and $\bar{x}_{j,k}$ the mean value of same variable for the community j. The distance is normalized into a similarity metric by

$$S_{ij} = 100 \times \left(1 - \frac{d_{ij}}{\max\{d_{ij}\}}\right)$$

ensuring that identical community profiles yield a similarity of 100%.

4.11 Logistic regression

To compare the utility of various sets of predictors for pediatric outcomes, we performed logistic regression analyses on two outcomes, Dermatitis and Asthma, using the (`smf`) package in Python. For each binary outcome (coded as 1 = Yes, 0 = No), we estimated three distinct models using the `logit` function: (1) a community model using IBD-based subcommunity classification as a categorical predictor ; (2) an ancestry model employing genetic ancestry as a categorical variable; and (3) a social determinants of health (SDoH) model. For the selection of SDoH variables, we previously performed a collinearity analysis to identify and retain only the most informative variables. This process allowed us to minimize redundancy among predictors and optimize the stability of the model incorporating continuous measures of the social vulnerability index, ambient $PM_{2.5}$ concentration, obesity prevalence, and smoking rates. In the subcommunities and ancestry models, only those categories with more than 100 individuals were incorporated (Supplementary Figure S7), to reduce bias due to small sample sizes and ensure stable parameter estimates. All models were fitted using maximum likelihood estimation with the default optimization algorithm (Newton-Raphson). Model performance was quantified using McFadden's pseudo R^2 (accessed via the `prsquared` attribute), which reflects the proportional improvement in log-likelihood relative to a null (intercept-only) model.

This systematic approach allowed us to assess the relative predictive contribution of genetic, environmental, and socioeconomic factors, both independently and in combination.

5 Data availability

This study draws its data from the Biorepository and Integrative Genomics (BIG), which is not publicly available due to privacy concerns and ethical considerations. We limit access to protect participants' confidentiality and ensure compliance with institutional and regulatory requirements. Researchers interested in accessing this data must first obtain approval from both the University of Tennessee Health Science Center (UTHSC) Institutional Review Board (IRB) and the BIG Research Oversight Committee. You can submit your request through the BIG portal at https://uthsc.edu/cbmi/big/. If you need assistance, please email biglist@uthsc.edu. We only grant access for legitimate research purposes, and all approved researchers must comply with our data use agreements. We typically process requests within 4 weeks of submission. Once approved, you can access the data throughout your research project's duration. We encourage responsible data sharing and invite interested researchers to contact either the authors or BIG directly for additional information.

6 Code availability

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

7 Competing interests

The Regeneron Genetic Center is a subsidiary of Regeneron Pharmaceuticals, Inc. All other authors declare no competing interests.

References

[1] Pier Francesco Palamara, Todd Lencz, Ariel Darvasi, and Itsik Pe'er. Length distributions of identity by descent reveal fine-scale demographic history. *The American journal of human genetics*, 91(5):809–822, 2012.

[2] Kenji Itao and Kunihiko Kaneko. Evolution of family systems and resultant socio-economic structures. *Humanities and Social Sciences Communications*, 8(1):1–11, 2021.

[3] Luigi Luca Cavalli-Sforza and Marcus W Feldman. *Cultural transmission and evolution: A quantitative approach*. Number 16. Princeton University Press, 1981.

[4] Luigi L Cavalli-Sforza. Cultural evolution. *American zoologist*, 26(3):845–856, 1986.

[5] Guido Barbujani and Vincenza Colonna. Human genome diversity: frequently asked questions. *Trends in Genetics*, 26(7):285–295, 2010.

[6] Theodosius Dobzhansky. *Mankind evolving*. JSTOR, 1962.

[7] Noah A Rosenberg, Jonathan K Pritchard, James L Weber, Howard M Cann, Kenneth K Kidd, Lev A Zhivotovsky, and Marcus W Feldman. Genetic structure of human populations. *science*, 298(5602):2381–2385, 2002.

[8] Guido Barbujani, Arianna Magagni, Eric Minch, and L Luca Cavalli-Sforza. An apportionment of human dna diversity. *Proceedings of the National Academy of Sciences*, 94(9):4516–4519, 1997.

[9] 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.

[10] María J Palma-Martínez, Yuridia S Posadas-García, Amara Shaukat, Brenda E López-Ángeles, and Mashaal Sohail. Evolution, genetic diversity, and health. *Nature Medicine*, pages 1–11, 2025.

[11] 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.

[12] Sharon R Browning and Brian L Browning. Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, 12(10):703–714, 2011.

[13] Peter Ralph and Graham Coop. The geography of recent genetic ancestry across europe. *PLoS biology*, 11(5):e1001555, 2013.

[14] Christa Caggiano, Arya Boudaie, Ruhollah Shemirani, Joel Mefford, Ella Petter, Alec Chiu, Defne Ercelen, Rosemary He, Daniel Tward, Kimberly C Paul, et al. Disease risk and healthcare utilization among ancestrally diverse groups in the los angeles region. *Nature Medicine*, 29(7):1845–1856, 2023.

[15] Edmund Gilbert, Ashwini Shanmugam, and Gianpiero L Cavalleri. Revealing the recent demographic history of europe via haplotype sharing in the uk biobank. *Proceedings of the National Academy of Sciences*, 119(25):e2119281119, 2022.

[16] Ardalan Naseri, Kecong Tang, Xin Geng, Junjie Shi, Jing Zhang, Pramesh Shakya, Xiaoming Liu, Shaojie Zhang, and Degui Zhi. Personalized genealogical history of uk individuals inferred from biobank-scale ibd segments. *BMC biology*, 19:1–12, 2021.

[17] Chengzhen L Dai, Mohammad M Vazifeh, Chen-Hsiang Yeang, Remi Tachet, R Spencer Wells, Miguel G Vilar, Mark J Daly, Carlo Ratti, and Alicia R Martin. Population histories of the united states revealed through fine-scale migration and haplotype analysis. *The American Journal of Human Genetics*, 106(3):371–388, 2020.

[18] Juba Nait Saada, Georgios Kalantzis, Derek Shyr, Fergus Cooper, Martin Robinson, Alexander Gusev, and Pier Francesco Palamara. Identity-by-descent detection across 487,409 british samples reveals fine scale population structure and ultra-rare variant associations. *Nature communications*, 11(1):6130, 2020.

[19] Gillian M Belbin, Sinead Cullina, Stephane Wenric, Emily R Soper, Benjamin S Glicksberg, Denis Torre, Arden Moscati, Genevieve L Wojcik, Ruhollah Shemirani, Noam D Beckmann, et al. Toward a fine-scale population health monitoring system. *Cell*, 184(8):2068–2083, 2021.

[20] Silvia Buonaiuto, Franco Marsico, Akram Mohammed, Lokesh K Chinthala, Ernestine K Amos-Abanyie, Regeneron Genetics Center, Pjotr Prins, Kyobeni Mozhui, Robert J Rooney, Robert W Williams, et al. The biorepository and integrative genomics resource for inclusive genomics: insights from a diverse pediatric and admixed cohort. *Nature Communications, in press*, 2025.

[21] 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.

[22] 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.

[23] Petr Gajdoš, Tomáš Ježowicz, Vojtěch Uher, and Pavel Dohnálek. A parallel fruchterman–reingold algorithm optimized for fast visualization of large graphs and swarms of data. *Swarm and evolutionary computation*, 26:56–63, 2016.

[24] Siti Haryanti Hairol Anuar, Zuraida Abal Abas, Norhazwani Mohd Yunos, Nurul Hafizah Mohd Zaki, Nurul Akmal Hashim, Mohd Fariddudin Mokhtar, Siti Azirah Asmai, Zaheera Zainal Abidin, and Ahmad Fadzli Nizam. Comparison between louvain and leiden algorithm for network structure: a review. In *Journal of Physics: Conference Series*, volume 2129, page 012028. IOP Publishing, 2021.

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

[26] E Pebesma and R Bivand. sf: simple features for r. r package version 0.6-3. *URL: cran. r-project. org/package= sf*, 2, 2018.

[27] Bahram Namjou, Michael Lape, Edyta Malolepsza, Stanley B DeVore, Matthew T Weirauch, Ozan Dikilitas, Gail P Jarvik, Krzysztof Kiryluk, Iftikhar J Kullo, Cong Liu, et al. Multiancestral polygenic risk score for pediatric asthma. *Journal of Allergy and Clinical Immunology*, 150(5):1086–1096, 2022.

[28] Kara E Hennelly, Angela M Ellison, Mark I Neuman, and Jeffrey A Kline. Clinical variables that increase the probability of pulmonary embolism diagnosis in symptomatic children. *Research and Practice in Thrombosis and Haemostasis*, 4(1):124–130, 2020.

[29] Christopher H Arehart, Michelle Daya, Monica Campbell, Meher Preethi Boorgula, Nicholas Rafaels, Sameer Chavan, Gloria David, Jon Hanifin, Mark K Slifka, Richard L Gallo, et al. Polygenic prediction of atopic dermatitis improves with atopic training and filaggrin factors. *Journal of Allergy and Clinical Immunology*, 149(1):145–155, 2022.

[30] Max Kozlov. 'all of us' genetics chart stirs unease over controversial depiction of race. *Nature*, 2024.

[31] 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.

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

[33] Jedidiah Carlson, Brenna M Henn, Dana R Al-Hindi, and Sohini Ramachandran. Counter the weaponization of genetics research by extremists. *Nature*, 610(7932):444–447, 2022.

[34] Anna CF Lewis, Santiago J Molina, Paul S Appelbaum, Bege Dauda, Anna Di Rienzo, Agustin Fuentes, Stephanie M Fullerton, Nanibaa'A Garrison, Nayanika Ghosh, Eve-

lynn M Hammonds, et al. Getting genetic ancestry right for science and society. *Science*, 376(6590):250–252, 2022.

[35] Megan C Lancaster, Hung-Hsin Chen, M Benjamin Shoemaker, Matthew R Fleming, Teresa L Strickland, James T Baker, Grahame F Evans, Hannah G Polikowsky, David C Samuels, Chad D Huff, et al. Detection of distant relatedness in biobanks to identify undiagnosed cases of mendelian disease as applied to long qt syndrome. *Nature communications*, 15(1):7507, 2024.

[36] Taotao Tan and Elizabeth G Atkinson. Strategies for the genomic analysis of admixed populations. *Annual Review of Biomedical Data Science*, 6, 2023.

[37] Bernie Devlin and Kathryn Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, 1999.

[38] Amke Caliebe, Fasil Tekola-Ayele, Burcu F Darst, Xuexia Wang, Yeunjoo E Song, Jiang Gui, Ronnie A Sebro, David J Balding, Mohamad Saad, Marie-Pierre Dubé, et al. Including diverse and admixed populations in genetic epidemiology research. *Genetic epidemiology*, 46(7):347–371, 2022.

[39] Jinghua Liu, Juan Pablo Lewinger, Frank D Gilliland, W James Gauderman, and David V Conti. Confounding and heterogeneity in genetic association studies with admixed populations. *American journal of epidemiology*, 177(4):351–360, 2013.

[40] Kelsey E Grinde, Brian L Browning, Alexander P Reiner, Timothy A Thornton, and Sharon R Browning. Adjusting for principal components can induce collider bias in genome-wide association studies. *PLoS genetics*, 20(12):e1011242, 2024.

[41] Kecong Tang, Ardalan Naseri, Yuan Wei, Shaojie Zhang, and Degui Zhi. Open-source benchmarking of ibd segment detection methods for biobank-scale cohorts. *GigaScience*, 11:giac111, 2022.

[42] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7):821–824, 2012.

[43] Takiy-Eddine Berrandou, David Balding, and Doug Speed. Ldak-gbat: Fast and powerful gene-based association testing using summary statistics. *The American Journal of Human Genetics*, 110(1):23–29, 2023.

[44] Doug Speed and David J Balding. Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics*, 16(1):33–44, 2015.

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

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

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

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

[53] 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.

[54] Chad D Huff, David J Witherspoon, Tatum S Simonson, Jinchuan Xing, W Scott Watkins, Yuhua Zhang, Therese M Tuohy, Deborah W Neklason, Randall W Burt, Stephen L Guthery, et al. Maximum-likelihood estimation of recent shared ancestry (ersa). *Genome research*, 21(5):768–774, 2011.

[55] 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.

[56] Sharon R Browning and Brian L Browning. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *The American Journal of Human Genetics*, 97(3):404–418, 2015.

[57] Alexander Gusev, Pier Francesco Palamara, Gregory Aponte, Zhong Zhuang, Ariel Darvasi, Peter Gregersen, and Itsik Pe'er. The architecture of long-range haplotypes shared within and across populations. *Molecular biology and evolution*, 29(2):473–486, 2012.

Regeneron Genetics Center

Aris Baras[4], Goncalo Abecasis[4], Adolfo Ferrando[4], Giovanni Coppola[4], Andrew Deubler[4], Aris Economides[4], Luca A Lotta[4], John D Overton[4], Jeffrey G Reid[4], Alan Shuldiner[4], Katherine Siminovitch[4], Jason Portnoy[4], Marcus B Jones[4], Lyndon Mitnaul[4], Alison Fenney[4], Jonathan Marchini[4], Manuel Allen Revez Ferreira[4], Maya Ghoussaini[4], Mona Nafde[4], William Salerno[4], John D Overton[4], Christina Beechert[4], Erin Fuller[4], Laura M Cremona[4], Eugene Kalyuskin[4], Hang Du[4], Caitlin Forsythe[4], Zhenhua Gu[4], Kristy Guevara[4], Michael Lattari[4], Alexander Lopez[4], Kia Manoochehri[4], Prathyusha Challa[4], Manasi Pradhan[4], Raymond Reynoso[4], Ricardo Schiavo[4], Maria Sotiropoulos Padilla[4], Chenggu Wang[4], Sarah E Wolf[4], Hang Du[4], Kristy Guevara[4], Amelia Averitt[4], Nilanjana Banerjee[4], Dadong Li[4], Sameer Malhotra[4], Justin Mower[4], Mudasar Sarwar[4], Deepika Sharma[4], Sean Yu[4], Aaron Zhang[4], Muhammad Aqeel[4], Jeffrey G Reid[4], Mona Nafde[4], Manan Goyal[4], George Mitra[4], Sanjay Sreeram[4], Rouel Lanche[4], Vrushali Mahajan[4], Sai Lakshmi Vasireddy[4], Gisu Eom[4], Krishna Pawan Punuru[4], Sujit Gokhale[4], Benjamin Sultan[4], Pooja Mule[4], Eliot Austin[4], Xiaodong Bai[4], Lance Zhang[4], Sean O'Keeffe[4], Razvan Panea[4], Evan Edelstein[4], Ayesha Rasool[4], William Salerno[4], Evan K Maxwell[4], Boris Boutkov[4], Alexander Gorovits[4], Ju Guan[4], Lukas Habegger[4], Alicia Hawes[4], Olga Krasheninina[4], Samantha Zarate[4], Adam J Mansfield[4], Lukas Habegger[4], Goncalo Abecasis[4], Manuel Allen Revez Ferreira[4], Joshua Backman[4], Kathy Burch[4], Adrian Campos[4], Liron Ganel[4], Sheila Gaynor[4], Benjamin Geraghty[4], Arkopravo Ghosh[4], Salvador Romero Martinez[4], Christopher Gillies[4], Lauren Gurski[4], Joseph Herman[4], Eric Jorgenson[4], Tyler Joseph[4], Michael Kessler[4], Jack Kosmicki[4], Adam Locke[4], Priyanka Nakka[4], Jonathan Marchini[4], Karl Landheer[4], Olivier Delaneau[4], Maya Ghoussaini[4], Anthony Marcketta[4], Joelle Mbatchou[4], Arden Moscati[4], Aditeya Pandey[4], Anita Pandit[4], Jonathan Ross[4], Carlo Sidore[4], Eli Stahl[4], Timothy Thornton[4], Sailaja Vedantam[4], Rujin Wang[4], Kuan-Han Wu[4], Bin Ye[4], Blair Zhang[4], Andrey Ziyatdinov[4], Yuxin Zou[4], Jingning Zhang[4], Kyoko Watanabe[4], Mira Tang[4], Frank Wendt[4], Suganthi Balasubramanian[4], Suying Bao[4], Kathie Sun[4], Chuanyi Zhang[4], Adolfo Ferrando[4], Giovanni Coppola[4], Luca A Lotta[4], Alan Shuldiner[4],

Katherine Siminovitch[4], Brian Hobbs[4], Jon Silver[4], William Palmer[4], Rita Guerreiro[4], Amit Joshi[4], Antoine Baldassari[4], Cristen Willer[4], Sarah Graham[4], Ernst Mayerhofer[4], Erola Pairo Castineira[4], Mary Haas[4], Niek Verweij[4], George Hindy[4], Jonas Bovijn[4], Tanima De[4], Parsa Akbari[4], Luanluan Sun[4], Olukayode Sosina[4], Arthur Gilly[4], Peter Dornbos[4], Juan Rodriguez-Flores[4], Moeen Riaz[4], Manav Kapoor[4], Gannie Tzoneva[4], Momodou W Jallow[4], Anna Alkelai[4], Ariane Ayer[4], Veera Rajagopal[4], Sahar Gelfman[4], Vijay Kumar[4], Jacqueline Otto[4], Neelroop Parikshak[4], Aysegul Guvenek[4], Jose Bras[4], Silvia Alvarez[4], Jessie Brown[4], Jing He[4], Hossein Khiabanian[4], Joana Revez[4], Kimberly Skead[4], Valentina Zavala[4], Jae Soon Sul[4], Lei Chen[4], Sam Choi[4], Amy Damask[4], Nan Lin[4], Charles Paulding[4], Marcus B Jones[4], Esteban Chen[4], Michelle G LeBlanc[4], Jason Mighty[4], Jennifer Rico-Varela[4], Nirupama Nishtala[4], Nadia Rana[4], Jaimee Hernandez[4], Alison Fenney[4], Randi Schwartz[4], Jody Hankins[4], Anna Han[4], Samuel Hart[4], Ann Perez-Beals[4], Gina Solari[4], Johannie Rivera-Picart[4], Michelle Pagan[4], Sunilbe Siceron[4].

[4]Regeneron Genetics Center, Tarrytown, NY, USA.

Supplementary Figures

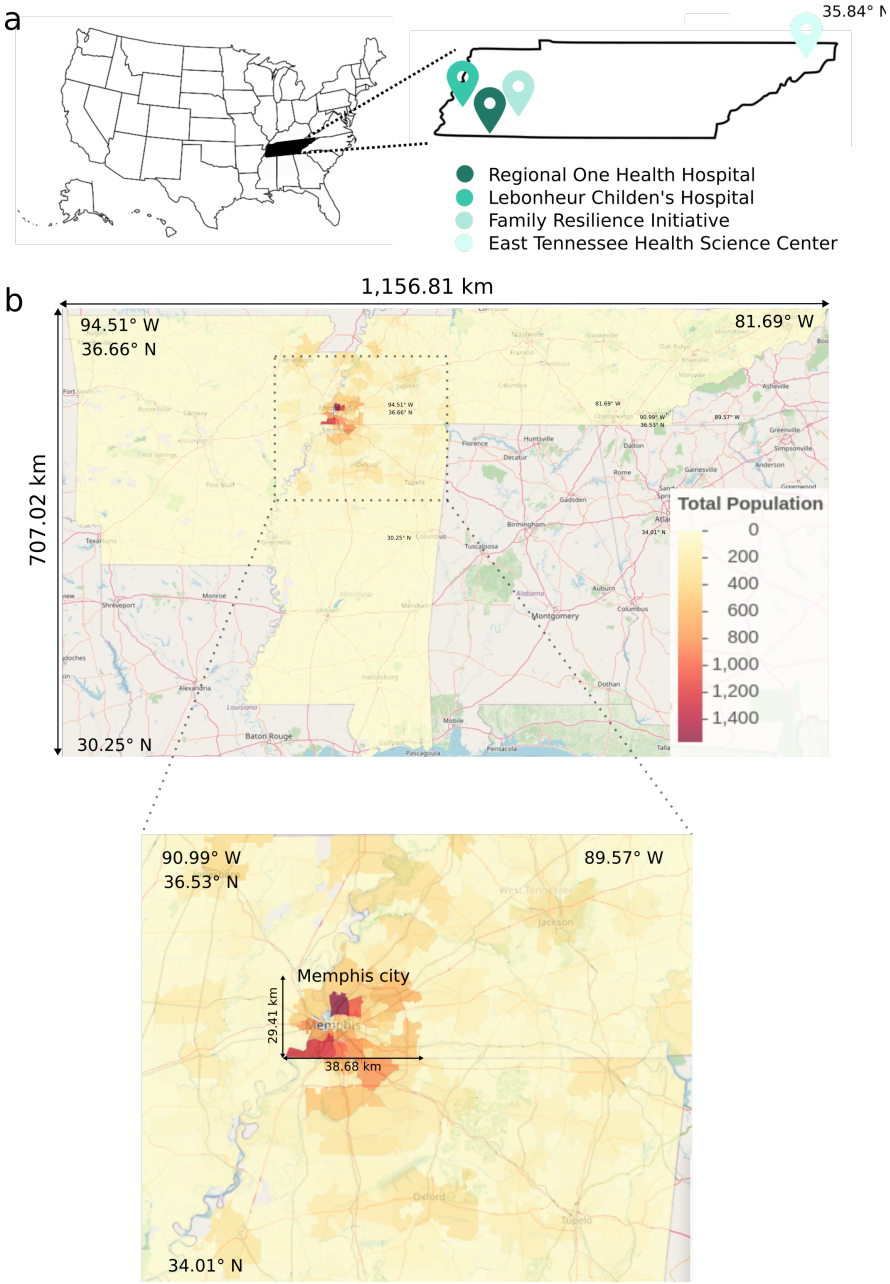

Figure S1: **Demography of enrolled participants** (a) The four sample sites (b) Number of enrolled participants by ZIP code. The region surrounding Memphis City is zoomed in, and distances in Km are shown. Maps were produced with the `leaflet` package (v. 2.2.1) using GeoJSON data for state ZIP-code boundaries publicly available.

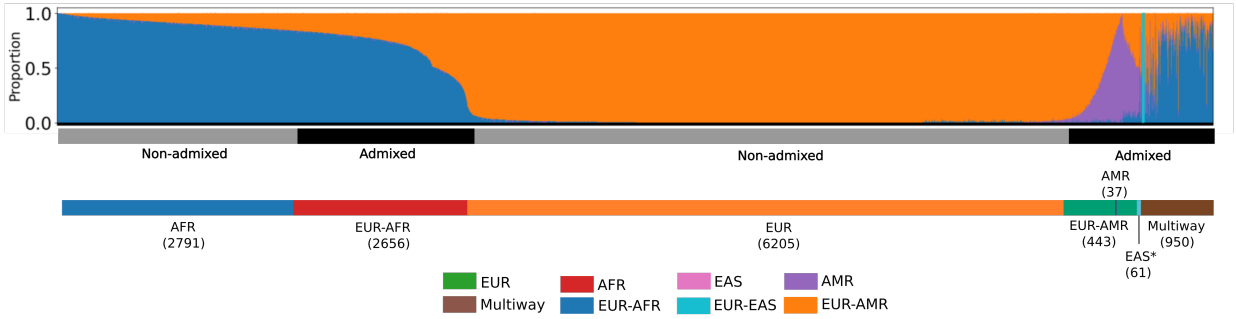

Figure S2: **Global ancestry deconvolution of sequenced individuals**. It is based on RFMix [9] and using reference populations in the 1000 Genomes and Human Genome Diversity Project (HGDP) data sets. Each vertical bar represents one individual, colors are proportional to inferred ancestry. For further 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) as described in the text.

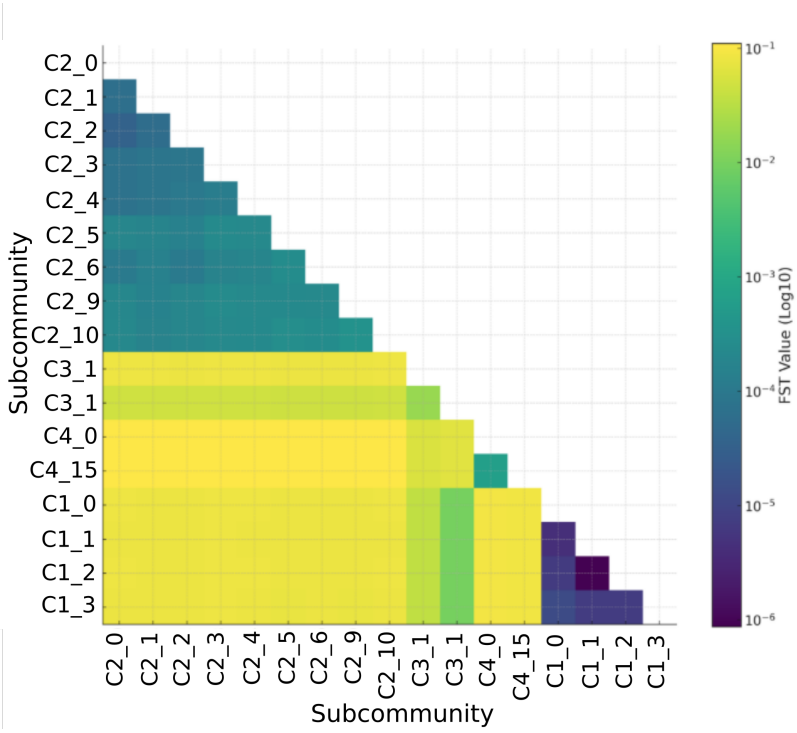

Figure S3: **Pairwise Fst values between subcommunities**. This analysis was performed using PLINK. It can be seen that subcommunities within C1 have the lowest fst values when analyzed each other. Contrastingly, C2 show higher values, indicative of more genetic diversity in the group.

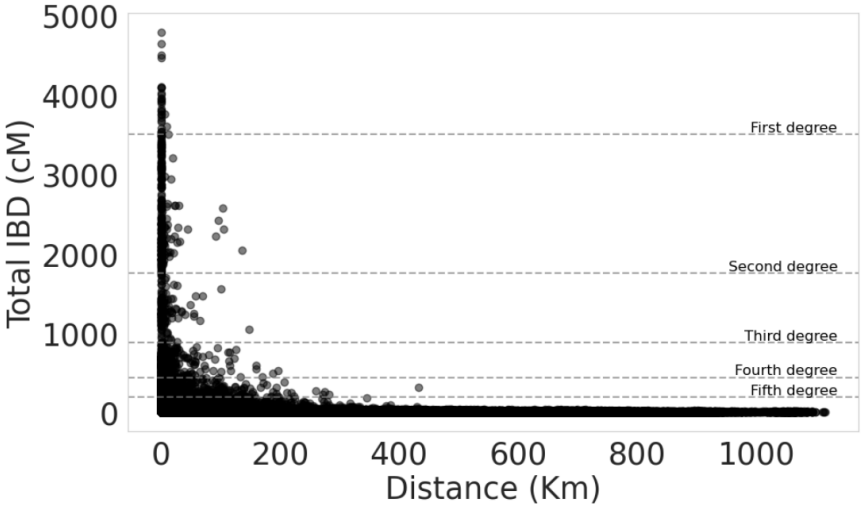

Figure S4: **Total IBD (cM) and Distance (Km)**. Different cutoffs for degrees of relationships are shown. The plot elucidates how higher total IBD values are related with closer distanteces. Geographical distances have been computed with sf R package [26].

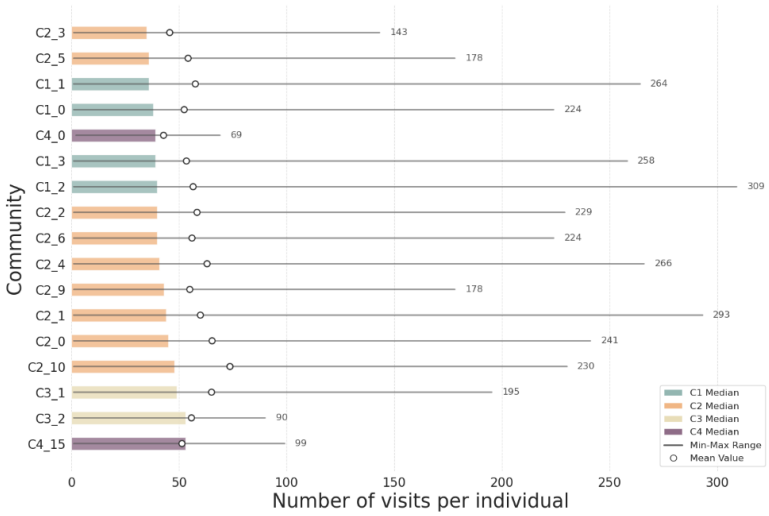

Figure S5: **Number of visits per individual across subcommunities**. Median (bars), mean and maximum are shown. Colors indicate the community level value.

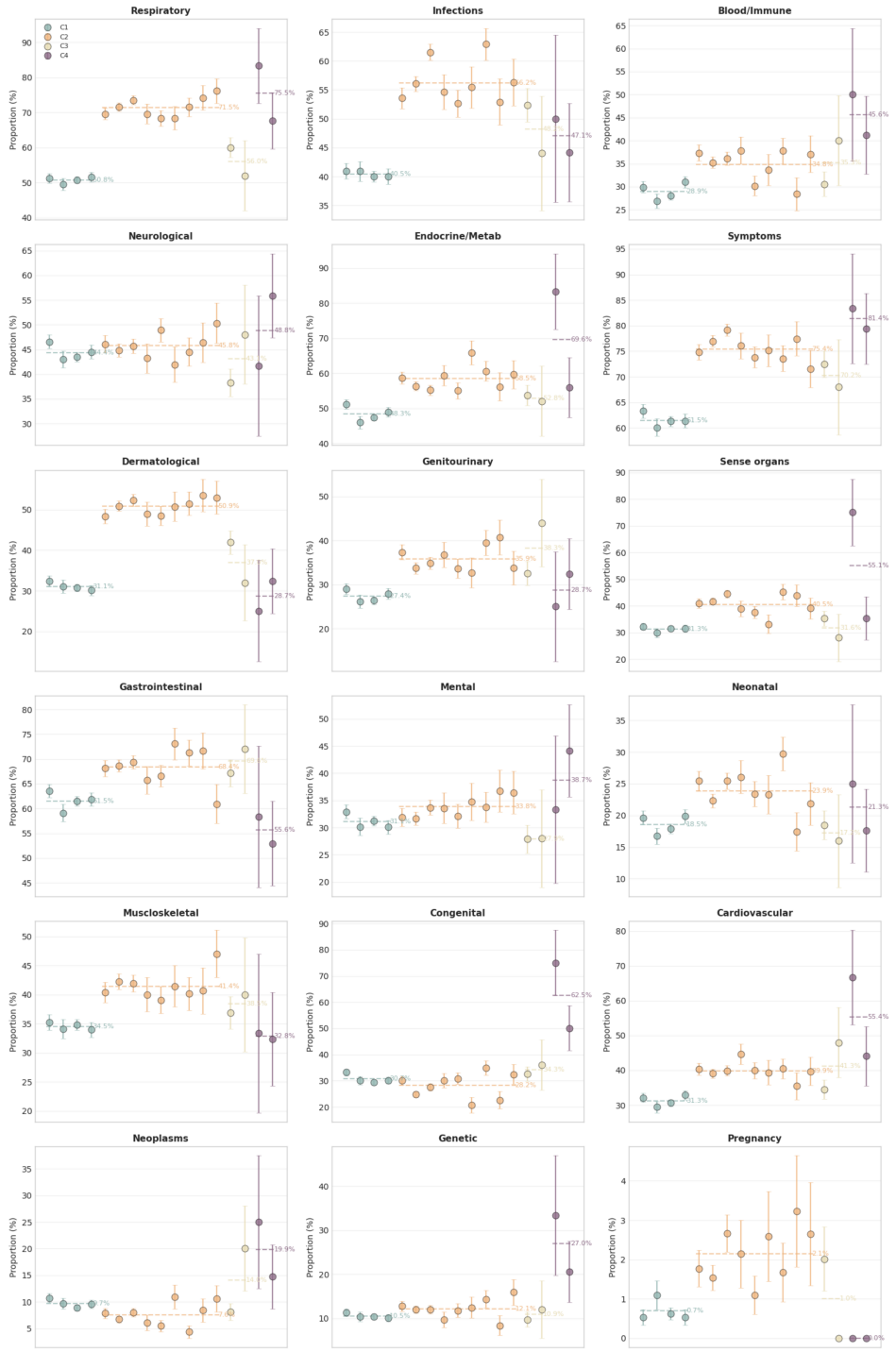

Figure S6: **Proportion of phenotypes at the subcommunity level**. CI95 for all proportions are shown and colored by community C1, C2, C3 and C4. The proportion is defined as the number of individuals in each community that reported, in at least one visit, a health condition in the corresponding category over the total. Error bars indicate 95% confidence intervals for proportions calculated using the Wilson score method.

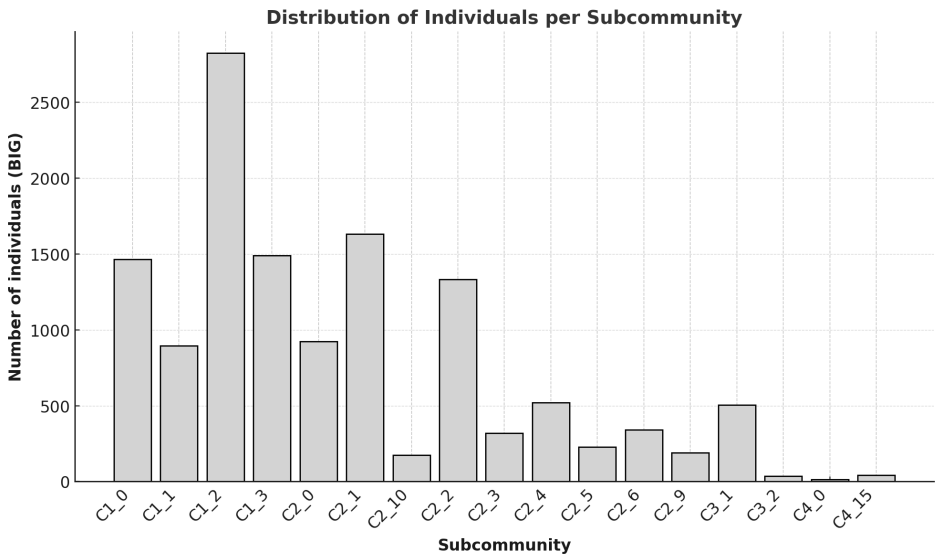

Figure S7: **Number of BIG individuals in each subcommunity**. We show the number of BIG individuals in each subcommunity , from the total of 13,143 incorporated.

Supplementary Tables

Table S1: **Standardization of self-reported race categories in electronic health records** We harmonized diverse race classifications to ensure analytical consistency and precision, following HL7 v2.5 standard terminologies while eliminating potentially problematic or outdated entries. This standardization process maintains demographic insights while supporting rigorous statistical analysis across our cohort [32].

Self-Reported Race	**Number of individuals in self-reported race**	**Standardized HL7 v2.5 Race Category**	**Number of individuals in Standardized HL7 v2.5 Race Category**
Black or African American	4905	Black or African American	5279
African American	392		
White	5672	White	6115
White or Caucasian	89		
Caucasian	354		
Amer. Indian/Alaska Native	8	American Indian or Alaska Native	35
Native American	27		
Asian	68	Asian	69
Asian or Pacific Islander	1		
Native Hawaiian/Other Pacific Islander	4	Native Hawaiian or Other Pacific Islander	6
Pacific Islander	2		
Other	14	Other race	922
Other/Unknown	410		
Amharic (East African)	1		
East Indian	2		
Hispanic	268		
Middle Eastern	3		
Multiple	224	Unknown race	45
Unavailable	37		
Declined	4		
Patient Declined to Answer	3		

Table S2: **Details of the logistic regression analysis considering subcommunity classification for asthma** Examining the relationship between IBD-based subcommunity classification and asthma. The model estimates log-odds of asthma diagnosis (dependent variable) based on community classification (independent variable), with C1_0 subcommunity serving as the reference level.

Term	Estimate	Std. Error	p-value
Intercept	-2.216	0.088	< 0.001
C1_1	-0.193	0.150	0.199
C1_2	-0.121	0.110	0.271
C1_3	-0.051	0.125	0.685
C2_0	0.549	0.126	< 0.001
C2_1	0.581	0.110	< 0.001
C2_2	0.724	0.113	< 0.001
C2_3	0.474	0.179	0.008
C2_4	0.623	0.146	< 0.001
C2_5	0.628	0.196	0.001
C2_6	0.529	0.172	0.002
C2_9	0.613	0.207	0.003
C3_1	-0.524	0.205	0.011

Table S3: **Details of the logistic regression analysis considering genetic ancestry for asthma** Examining the relationship between genetic ancestry and asthma. The model estimates log-odds of asthma diagnosis (dependent variable) based on genetic ancestry (independent variable), with AFR ancestry serving as the reference level.

Term	Estimate	Std. Error	p-value
Intercept	-1.610	0.051	< 0.001
EUR	-0.729	0.068	< 0.001
EUR-AFR	0.042	0.073	0.563
EUR-AMR	-0.831	0.183	< 0.001
EUR-EAS	-0.875	1.042	0.401
Others/AMR/EAS	-0.258	0.108	0.017

Table S4: **Details of the logistic regression analysis considering SDoH for asthma** Examining the relationship between social determinants of health (SDoH) and asthma. The model estimates log-odds of asthma diagnosis (dependent variable) based on continuous measures of Social Vulnerability Index (SVI), $PM_{2.5}$ concentration, obesity prevalence, and smoking rates.

Term	Estimate	Std. Error	p-value
Intercept	-3.052	0.266	< 0.001
SVI	0.015	0.005	0.006
PM25	0.010	0.002	< 0.001
URBAN	0.008	0.011	0.469
CSMOKING	0.005	0.010	0.640

Table S5: **Details of the logistic regression analysis considering subcommunity classification for dermatitis** Examining the relationship between IBD-based subcommunity classification and dermatitis. The model estimates log-odds of dermatitis diagnosis (dependent variable) based on community classification (independent variable), with C1_0 subcommunity serving as the reference level.

Term	Estimate	Std. Error	p-value
Intercept	-3.385	0.147	< 0.001
C1_1	-0.088	0.244	0.720
C1_2	-0.101	0.184	0.583
C1_3	-0.085	0.211	0.688
C2_0	1.080	0.186	< 0.001
C2_1	1.113	0.170	< 0.001
C2_2	1.237	0.172	< 0.001
C2_3	0.994	0.249	< 0.001
C2_4	1.173	0.207	< 0.001
C2_5	0.857	0.292	0.003
C2_6	1.169	0.233	< 0.001
C2_9	0.618	0.332	0.063
C3_1	-0.261	0.317	0.411

Table S6: **Details of the logistic regression analysis considering genetic ancestry for dermatitis** Examining the relationship between genetic ancestry and dermatitis. The model estimates log-odds of dermatitis diagnosis (dependent variable) based on genetic ancestry (independent variable), with AFR ancestry serving as the reference level.

Term	Estimate	Std. Error	p-value
Intercept	-2.296	0.066	< 0.001
EUR	-1.194	0.100	< 0.001
EUR-AFR	0.007	0.094	0.944
EUR-AMR	-1.360	0.312	< 0.001
EUR-EAS	-0.189	1.043	0.856
Others/AMR/EAS	-0.242	0.141	0.087

Table S7: **Details of the logistic regression analysis considering SDoH for dermatitis** Analysis of the relationship between Social Determinants of Health (SDoH) and Dermatitis. The model estimates log-odds of dermatitis diagnosis based on SVI, $PM_{2.5}$ concentration, obesity, and smoking rates.

Term	Estimate	Std. Error	p-value
Intercept	-4.555	0.375	< 0.001
SVI	-0.003	0.007	0.686
PM25	0.019	0.003	< 0.001
OBESITY	0.012	0.016	0.470
CSMOKING	0.005	0.013	0.708