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], Rima Zahr[4], Regeneron Genetics Center[5], Robert J Rooney[1,6], Robert W Williams[1,7], Robert L Davis[3], Terri H Finkel[8], Chester W Brown[1, 9], Marcella Vacca[2], Pjotr Prins[1], and Vincenza Colonna[1, 2, 6, †]

[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]Pediatric Nephrology and Hypertension, UTHSC, USA
[5]Regeneron Genetics Center, Tarrytown, NY, USA
[6]Dept of Pediatrics, UTHSC, USA
[7]Center for Integrative and Translational Genomics, UTHSC, USA
[8]Dept of Pediatrics, Division of Rheumatology, UTHSC, USA
[9]Dept of Pediatrics, Division of Genetics, UTHSC, USA
[†]Corresponding authors: vcolonna@uthsc.edu, fmarsic1@uthsc.edu

Abstract

The *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 to include shared environments and culture. Here, we show that identity-by-descent (IBD)–based

genetic similarity, independent of reference populations, provides a powerful framework for studying this interplay. Using genomic and electronic health record data from 13,143 individuals in the Biorepository for Integrative Genomics, we applied a hierarchical community detection algorithm to classify individuals by the proportion of the genome shared identical by descent. This approach identified four major communities concordant with continental ancestry and seventeen subcommunities capturing fine-scale demographic structure beyond conventional ancestry classifications.

Integrating neighborhood-level geographic data and census-derived environmental metrics revealed unequal exposure to environmental stressors across IBD-defined subcommunities. IBD-based grouping improved statistical modeling of disease risk and identified subpopulations with elevated rates of respiratory and dermatological conditions. We implemented an open-source dashboard that enables real-time clinical decision support and public health surveillance by correlating IBD-defined subcommunities with disease prevalence and environmental exposures—a unique capability that transforms biobank data into actionable insights for participants while maintaining privacy.

Together, these findings demonstrate that IBD-based clustering jointly captures genetic and environmental determinants of health, offering a scalable framework for precision health and population genetics.

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 faces significant epistemological challenges that extend beyond technical limitations. Common approaches impose artificial boundaries on what research has shown to be 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 categorization, 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] to build IBD-based communities that integrate genomic, clinical, and environmental information, providing a scalable framework to study health disparities and gene–environment interactions at biobank scale. The BIG resource provides comprehensive electronic health records (EHR), neighborhood-level environmental data linked through ZIP codes, and genomic information from 13,143 participants, with particular emphasis on pediatric populations. A key feature of the BIG is its representation of high genetic diversity concentrated within a relatively confined geographic area, enabling the investigation of genealogical-environmental interactions using standardized data. Here, IBD-based clustering is used as a reference-free framework for population structure analysis. We demonstrate that IBD communities correspond to both inferred and self-reported ancestry while capturing fine-scale genetic communities with distinct environmental exposures and health outcomes, reflecting geographic stratification. We show that community-specific health patterns can be traced to differential environmental exposures, providing insights into gene-environment interactions at the population level. Our findings establish IBD communities as a valuable tool for precision public health, enabling targeted

interventions based on community-specific genetic and environmental profiles.

2 Results

2.1 Capturing population structure using information from genomic segments shared identical by descent

The BIG initiative has enrolled over 42,000 participants with linked EHRs and collected more than 15,000 biosamples (**Supplementary Figure S1**). The cohort is predominantly pediatric (87% under 18 years), spanning ages from infancy to 90 years (mean = 8.4, median = 6.2) [20]. Unlike typical pediatric cohorts of healthy mother–child pairs, BIG focuses on children with diverse diseases and ancestry backgrounds.

To assess population structure, we performed identity-by-descent (IBD) analysis, which infers genetic relationships directly from shared genomic segments [21]. Total pairwise IBD was calculated as the sum of the length (cM) of all shared segments between individuals. The distribution of total IBD values followed an exponential decay (**Figure 1A**), indicating that distant relatives pairs far outnumber close kinships, consistent with previous studies [14, 18, 22].

We constructed an IBD network, where nodes represent individuals (n = 13,143) and edge lengths are inversely proportional to shared IBD. The network's global density (0.018) and average degree (241.15) indicate a well-connected structure. Four main dense clusters are visible, surrounded by smaller peripheral groups (**Figure 1B**). Using the Leiden algorithm [23], we applied two rounds of hierarchical community detection. The first iteration identified four major communities (C1–C4), two of which contained more than 5,000 individuals. Subsequent iterations within each community revealed 17 subcommunities (**Figure 1C**).

The community classification corresponds well with the population structure revealed by PCA of autosomal genetic markers, with clear separation between communities (**Figure 1D**).

This concordance between discrete community assignments and continuous genetic variation validates that communities represent genuine genealogical differences rather than arbitrary divisions, suggesting that the IBD-based communities capture the underlying genetic structure present in the data. Despite this overall trend, we also notice individuals positioned between the main groups in the PCA space, as well as some individuals from one community who fall within the region predominantly occupied by a different community. As an example, C1 individuals in C2 space would be assigned to C2 based on their position in the PCA-based classification; however, the IBD-based approach places them in C1 due to detected signals of relatedness (up to third degree), highlighting the additional power of relationship-based community detection over purely dimensional-reduction-based methods.

At a higher level, a network of subcommunities (**Figure 1E**) shows that the number of subgroups does not correlate with parent community size, suggesting that internal genetic diversity influences subdivision. Community C2 exhibits the highest heterogeneity and the greatest number of subcommunities, supported by pairwise Fst values (**Supplementary Figure S2**). In contrast, C1 subcommunities are more closely connected, indicating closer relationships among members, while C3 connects to both C1 and C2, probably reflecting shared ancestry patterns.

2.2 IBD communities reflect geographical ancestry with fine-scale substructure

Having determined IBD-based communities, we next asked whether they corresponded to global ancestry previously inferred from reference populations [20]. We found an enrichment of European ancestry (EUR, 91 %) in C1, African (AFR, 49 %) and admixed EUR–AFR (39 %) in C2, American-admixed (EUR–AMR, 51 %) and multiway (43 %) in C3, and East Asian (EAS, 81%) in C4 (**Figure 2A**). The association between community and ancestry was highly significant (χ^2 $p < 0.0001$; Cramér's $V = 0.81$).

We characterized the relationship between ancestry and network topology by computing

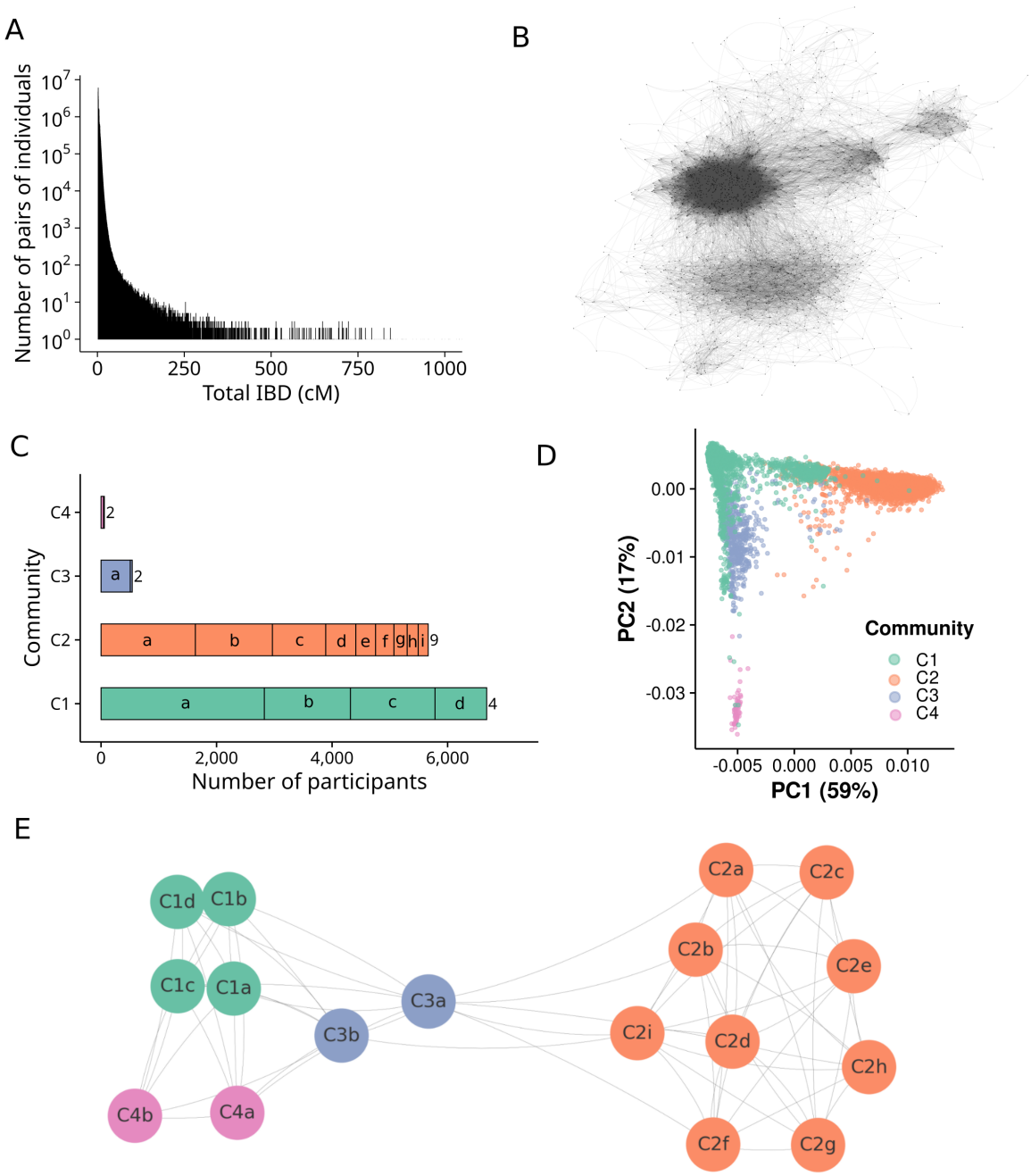

Figure 1: **Identity-by-descent (IBD) community detection on the BIG cohort. (A)** Exponential decay of total IBD (cM). **(B)** IBD-based network with 13,143 nodes, revealing four denser areas surrounded by sparse zones, suggesting a structured network. Fruchterman-Reingold layout was used to do the plot [24]. **(C)** Community sizes after the first Leiden iteration (C1 to C4), and size of the subcommunities (small case letters on bars) derived from the second Leiden iteration. **(D)** PCA plot showing samples distribution across the two first PCs. Colors represent the communities. The first two principal components explain 76% of the variance captured by the first 20 PCs. **(E)** Network of subcommunities highlighting diversity and inter-connectivity (Fruchterman-Reingold layout). Connections between subcommunities are based on total IBD.

each BIG individual's connectivity (degree link strength) and betweenness (path-based centrality). Stratification by ancestry revealed that Europeans in C1, Africans in C2, and Americans in C3 tend to act as hub–bridge nodes, mediating many connections and shortest paths (**Figure 2B**). In contrast, admixed individuals exhibit lower connectivity and betweenness, occupying more peripheral network positions. Multiway-admixed individuals consistently fall into low-degree, low-betweenness niches, while certain moderately connected nodes display elevated betweenness, acting as critical inter-community bridges. C4 communities were not analyzed due to their smaller size.

To further characterize communities and subcommunities, we measured IBD sharing between BIG subcommunities and the HGDP-1kGP reference panels. The IBD sharing patterns align with the ancestry correlations we previously inferred (**Figure 2C**). However, the fine-scale subcommunity structure suggests subtle demographic events that shaped genetic diversity within broader continental groups. For example, subcommunity C3b showed a stronger relationship with the Puerto Rican population, characterized by its three-way ancestry pattern [11, 25], while C3a was more closely associated with Peruvian and other Native American populations. This subdivision suggests multiple waves or routes of population movement and mixing within C3. Likewise, the connections within C2 (predominantly related to Africans) of C2g with Surui (an Indigenous American population) might indicate shared ancestry components that transcend simple continental divisions, possibly reflecting more recent demographic events.

After characterizing community structure through genetic measures, we next examined how communities aligned with participants' self-reported ancestry and ethnicity (**Figure 2D-E**). Similar to what was observed for IBD sharing with reference populations, while community-level patterns align predictably with self-reported race/ethnicity (C1: 91% white, C2: 97% Black/African American, C4: 62% Asian), subcommunity analysis reveals important nuances. Most notably, C3 subcommunities diverge markedly. In terms of self-reported race, C3a is predominantly Other race (73%), whereas C3b shows a three-way split across Other race (33%), White (21%), and Black/African American (21%). In terms of self-reported

ethnicity, individuals in C3a were more likely to self-identify as Hispanic or Latino (72.3%), while those in C3b predominantly self-report as Not Hispanic or Latino (69.6%).

Overall, our findings demonstrate that IBD-based communities capture population stratification consistent with both genetic ancestry and self-reported race/ethnicity. While major communities reflect continental ancestry patterns, subcommunity analysis reveals finer population substructure and demographic complexity, including admixture patterns and the nuanced relationships between genetic ancestry and self-identity. This hierarchical approach provides the resolution to understand both historical demographic processes and contemporary genetic diversity within the BIG cohort.

2.3 IBD sharing reflects shared environments in the BIG cohort

To elucidate the spatial distribution of genetic relationships, we examined whether subcommunities exhibit distinct geographical patterns and how these relate to IBD. Specifically, we geolocated BIG participants' ZIP codes at the neighborhood level using publicly available spatial polygons [26], and then analyzed whether genetically determined subcommunities cluster geographically. We detected a strong association between genealogical and geographical distances (exponential fit $R^2 = 0.969$, p-value < 0.0001, **Figure 3A**, **Supplementary Figure S3**). Nearly half (49%) of close relatives (first- and second-degree) lived in the same neighborhood, while the remaining pairs typically resided 4-6 km apart. Third- and fourth-degree relatives showed greater separation, though still within 20 km, with 34.3% sharing the same neighborhood. Beyond the seventh or eighth degree, geographic distance increased sharply, only 4.45% of eighth-degree or more distant relatives lived in the same neighborhood.

We used ZIP codes to link communities to social determinants of health (SDoH), i.e., non-medical factors influencing health outcomes. All SDoH values were obtained from publicly available datasets spanning 2016 to 2024. Complete results, including interactive visualizations and detailed methodology, are available on our dedicated website (https://github.com/MarsicoFL/BIG_Communities). We quantified SDoH similarity across subcommunities using

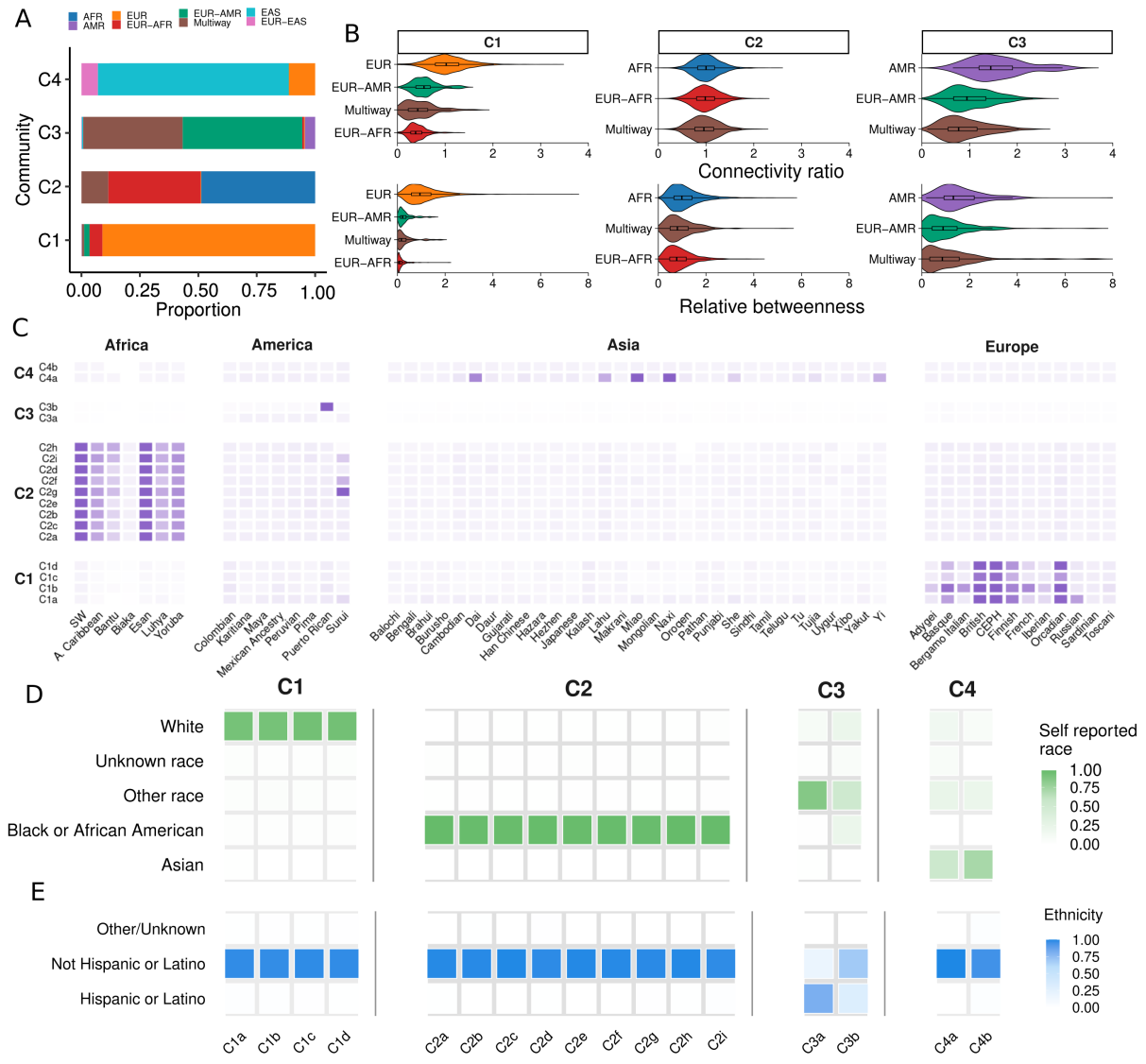

Figure 2: **Genetic and demographic patterns across sub-communities.** **(A)** Inferred continental and admixed ancestries per community from previus work. **(B)** Ancestry's role in the IBD-network. Connectivity ratio for each node was defined as its degree divided by the mean degree of its IBD community. Values greater than 1 indicate higher connectivity compared to the average. Betweenness measures the centrality of individuals in communities; individuals with betweenness above 1 are more central, while values below 1 indicate more peripheral individuals. **(C)** Heatmap showing mean IBD values between subcommunities and reference populations, highlighting distinct genetic relationships across C1 to C4. **(D)** Distribution of self-reported race, and **(E)** ethnicity proportions within subcommunities.

mean SDoH and found that subcommunities within the same community shared highly similar social environments, while communities differed markedly from each other (**Figure 3B**). This pattern held even for C1, whose subcommunities maintained 96% SDoH similarity despite being the most geographically dispersed (median distance 88.6 km between individuals vs. 18.9-22.1 km for C2 and C3, **Supplementary Figure S4B**). Conversely, C2 exhibited the greatest social heterogeneity among its subcommunities while occupying a smaller geographic area than C1 (**Supplementary Figure S4B**). SDoH similarity is strongly correlated with mean IBD sharing within and between subcommunities. To account for non-independent pairwise entries, we tested the association using a Mantel permutation. We find a monotonic association (Mantel Spearman $\rho = 0.747$, $p = 2.0 \times 10^{-4}$), demonstrating that communities with similar SDoH share greater genetic relatedness (**Figure 3C**). In addition, the nuanced variation in SDoH within C2 is reflected in overall lower IBD sharing within C2. The spatial separation of C2 subcommunities likely explains their SDoH variation. For example, subcommunities C2a, C2e, and C2h occupy distinct western, central, and eastern regions of Memphis respectively (**Figure 3D**, Supplementary Figure S4A), demonstrating how IBD-based clustering captures the interplay between genetic relatedness, geographic separation, and differential social environments.

Large differences between communities and more subtle differences between subcommunities also appear when examining in more detail key SDoH factors with direct impacts on clinical outcomes. As shown in **Figure 3E**, subcommunities within group C2 exhibit elevated $PM_{2.5}$ Environmental Justice (EJ) index values, indicative of higher pollution levels, 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 that is also seen, though less pronounced, in C3.

Overall, these findings suggest that in the BIG cohort IBD-based community membership strongly reflects social environmental exposure. IBD-based communities likely represent distinct socioeconomic or cultural groups with differential access to resources and environmental conditions, which may have important implications for understanding health disparities across

these populations.

2.4 Health condition patterns in subcommunities reflects differential environmental exposure

Having assessed the differential environmental exposures across communities and subcommunities, we next examined whether these differences translate into disparities in disease incidence. We mapped the 14.7k ICD-9/10 codes from the BIG electronic health records to phecode categories and found overall distinct disease prevalence patterns across communities (**Figure 4B**). Community-level analysis revealed differences in environmentally-influenced conditions [27, 28]. C2 showed a particularly high prevalence of respiratory diseases (71.4%), infections (56.8%), and dermatological conditions (50.7%). In contrast, conditions with stronger genetic components [29, 30], including neoplasms, neurological disorders, and congenital show more homogeneous distributions across communities. We evaluated disease prevalence across the aforementioned six categories at the subcommunity level. Although these categories are broad and warrant detailed case-by-case analysis, we observed a clear pattern: Communities showed variable prevalence for environmentally-driven diseases but consistent prevalence for genetically-driven diseases, suggesting that differential environmental exposures influence disease risk (**Figure 4D**). In line with what has been described so far, C2 has the largest variance across subcommunities in many cases, especially in congenital disorders that appeared generally similar across communities, with some cases, C2i and C2j, exhibiting a lower prevalence of congenital conditions.

Combining the population structure, environmental factors, and health conditions, we detected that fine particulate matter exposure was strongly associated with the respiratory disease burden across subcommunities: $PM_{2.5}$ explained most of the variation in respiratory prevalence (Spearman $\rho = 0.85$, $R^2 = 0.90$), whereas no meaningful association was observed for neurological disorders ($\rho = 0.10$, $R^2 = 0.02$; **Figure 4C**). C4 subcommunities were not included due to an insufficient number of individuals. These

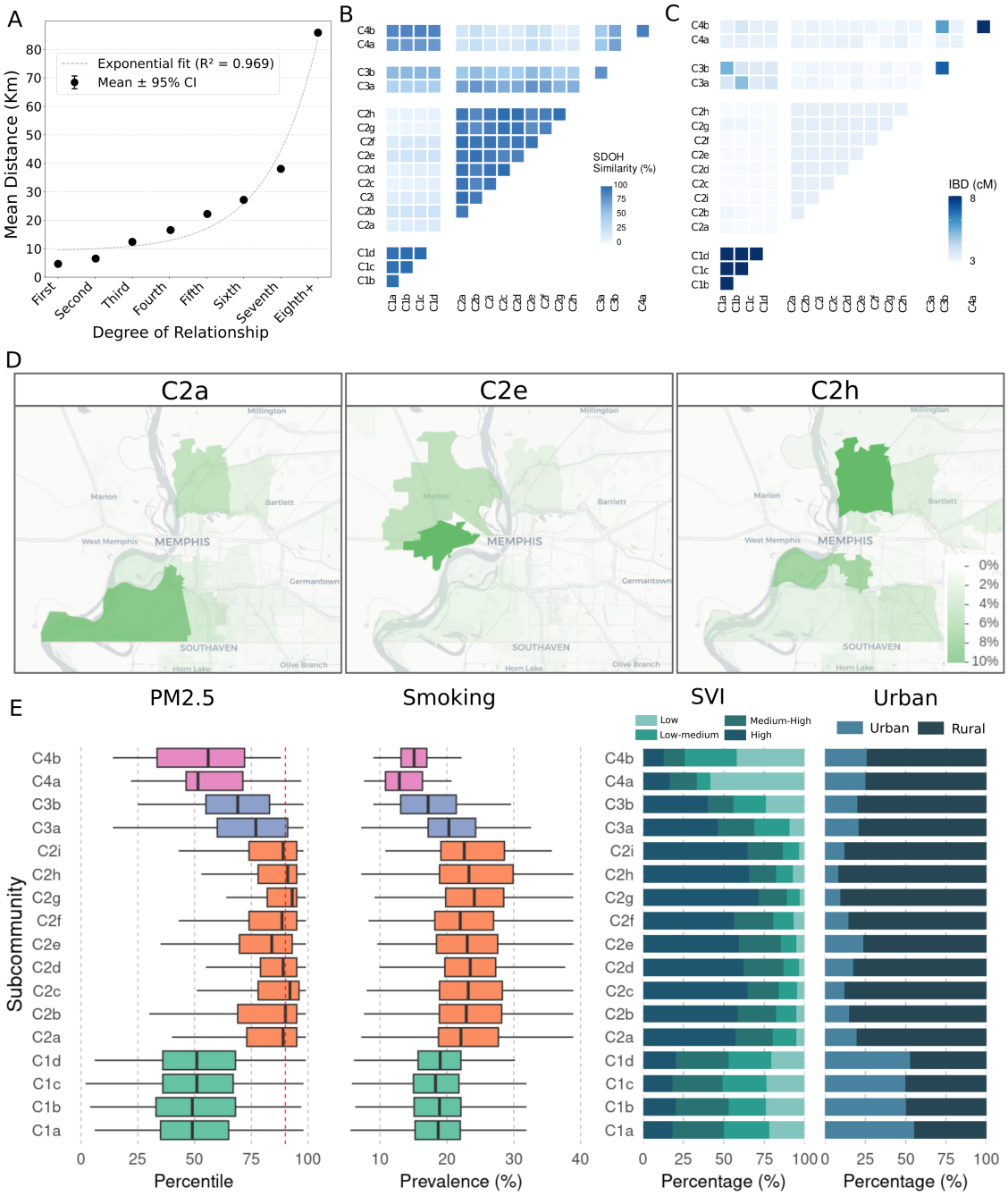

Figure 3: **Genealogy, environment and geography of BIG communities. (A)** Relationship between genealogical (degree of relationship) and geographical (Km) distances. Exponential fit is shown. **(B)** Social determinants of health similarity index at neighborhood level across communities. The diagonal of the matrix is not shown and corresponds to the maximum (100%) identity. **(C)** IBD sharing measured as median pairwise IBD (cM) between pairs of different subcommunities **(D)** Maps centered on the city of Memphis showing the geographical distribution of individuals from three subcommunities (C2a, C2e and C2h) from community C2. **(E)** Social determinants of health by subcommunities. PM 2.5 EJ index is an environmental justice metric derived from the ambient concentration of fine particulate matter, red line indicates the 90th percentile, above which indicats high risk. Smoking prevalence, social vulnerability index and urbanization are also reported.

exposure–disease relationships, along with additional diagnostic categories and environmental metrics, can be interactively explored using the web application accompanying this work (https://github.com/MarsicoFL/BIG_Communities).

Importantly, individuals across all subcommunities showed consistent healthcare access, with a median of 8.7 visits annually and 38 total visits per individual, suggesting that differences in disease patterns between communities are unlikely to be explained by differential access to care or under-diagnosis (**Figure 4A**).

2.5 IBD informativeness in pediatric-related health conditions

To evaluate the clinical utility of IBD-based classification, we compared its predictive ability alongside ancestry and SDoH for for health conditions that are particularly relevant in pediatric populations. We focused on dermatitis, for which we identified 779 participants with the condition, and asthma, with 1655. The rationale for selecting these pathologies is based on the trend observed in the respiratory and dermatological phecode categories, which contain these health conditions. We fitted logistic regression models for both conditions using McFadden's pseudo R^2 as a measure of model fit. 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 [31, 32, 33]. In particular, polygenic risk models for asthma and eczema often fall within the 0.03 to 0.1 range [31, 32]. Moreover, this study confirms the previous observation of C2 subcommunities with a higher prevalence of dermatitis and asthma (p-value < 0.001, Supplementary Tables S3, S6).

Having established the clinical relevance of IBD-derived structure for pediatric conditions, we next examined its role as an adjustment variable in genetic association models. Specifically, we assessed whether incorporating IBD information improves the precision and stability of

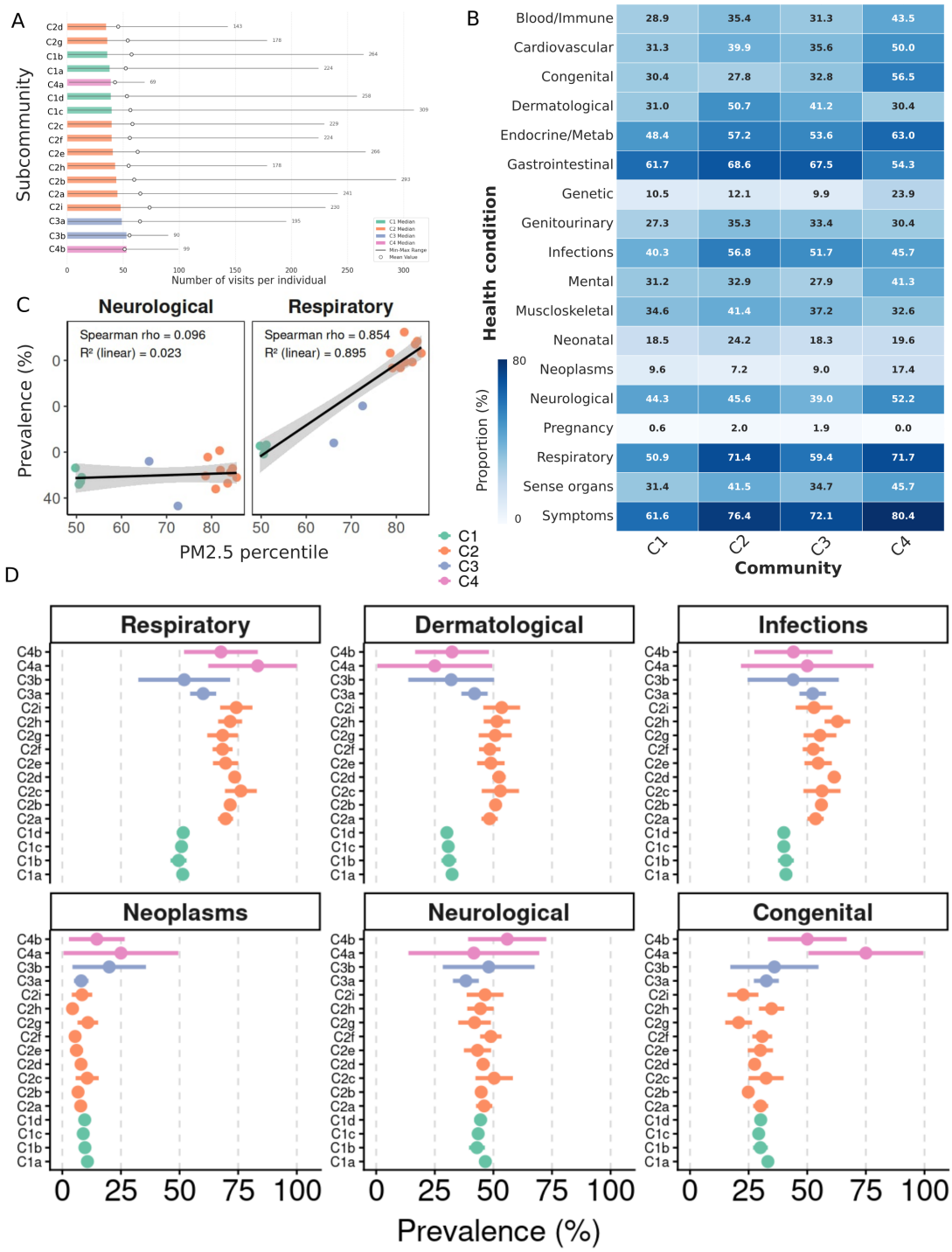

Figure 4: **Health conditions across subcommunities.** **(A)** Bar plot illustrating the median number of hospital visits per year in subcommunities. Mean and maximum are also indicated. **(B)** Heatmap showing the proportion of phenotypes coded by phecodes at the main community level (C1, C2, C3, C4), 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 95% confidence intervals, 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 S5).

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 three predictor sets for two clinical outcomes.

associations between genetic risk variants and kidney dysfunction.

To further examine the robustness of the association between APOL1 high-risk (HR) genotype and advanced kidney dysfunction, we evaluated logistic regression models defining CKD (Chronic Kidney disease) as eGFR (Estimated Glomerular Filtration Rate) <45 mL/min/1.73 m^2, sequentially adjusting for (i) age and sex, (ii) age, sex, and identity-by-descent (IBD), and (iii) age, sex, and principal components (PCs) of ancestry. Across these models, the estimated odds ratios (ORs) for *APOL1* HR genotypes progressively strengthened, suggesting improved model precision and attenuation of potential confounding by relatedness and population structure. In the minimally adjusted model (age and sex), the APOL1 HR genotype was associated with modestly increased odds of eGFR <45 (OR=1.89, 95% CI 1.64–2.20). Inclusion of IBD information to account for cryptic relatedness modestly improved model fit and yielded a slightly higher OR (1.98, 95% CI 1.71–2.30), consistent with more accurate separation of genetically independent observations. When further adjusted for ancestry PCs, the association strengthened further (OR=2.02, 95% CI 1.73–2.34), achieving greater statistical significance and narrower confidence intervals. This stepwise improvement highlights the importance of accounting for population structure in genetic association studies, particularly in ancestrally diverse pediatric cohorts. Overall, the progressive adjustment demonstrated that ancestry correction enhances the stability and interpretability of *APOL1*-mediated CKD associations. The final model incorporating age, sex, IBD, and PCs provided the best fit and most consistent estimates, reinforcing that the *APOL1* HR genotype confers significantly elevated risk of reduced kidney function (eGFR <45 mL/min/1.73 m^2) among African American children.

2.6 Public health applications of IBD-based stratification

Leveraging the data in the BIG biobank, we implemented an open-source dashboard (https://github.com/MarsicoFL/BIG_Communities) that is publicly available and can benefit healthcare stakeholders. The IBD-based classification at the subcommunity level enables at least two direct applications. First, in clinical settings, when a patient's zip code is shared with the clinician and the patient presents with symptoms such as skin irritation or respiratory issues, an EHR plugin automatically displays their subcommunity's elevated risks for specific conditions (e.g., dermatitis, asthma) and relevant environmental exposures, guiding clinicians toward targeted diagnostic tests rather than broad screening. Secondly, for public health surveillance, when investigating disease outbreaks or chronic condition clusters, health departments can map IBD networks to identify transmission pathways or shared exposures among genetically related families living across different neighborhoods—revealing epidemiological connections that ZIP code analysis alone would miss. Both applications rely on existing biobank genotypes without additional data collection.

Finally, this network-based approach accommodates dynamic biobank growth, as demonstrated by our analysis incorporating 4,091 individuals from the HGDP-1kGP joint release [34]. It showed that existing community assignments remain stable upon the addition of new participants, and new communities arise capturing new genetic diversity (**Supplementary Figure S7**). Our application can be adapted to other cohorts, enabling joint exploration of IBD-based communities, health conditions, and environmental factors. Similar approaches have demonstrated utility in other biobank initiatives, as described by Caggiano et al. [14] and Dai et al. [17].

3 Discussion

In this work, we have shown that IBD can be used to group individuals without relying on reference populations, and that in our cohort the grouping is informative about environment

and population health factors. Our findings demonstrate that IBD-based clustering captures population structure at multiple scales, from continental ancestry patterns to fine-scale genetic subcommunities that reflect shared environmental exposures and social determinants of health. Beyond broad inter-community differences, we observed distinct health condition prevalences among subcommunities, suggesting that genetic relatedness correlates with both biological and environmental factors influencing disease risk. Importantly, incorporating IBD information into genotype-phenotype models accounts for both genetic similarity and shared environmental effects, providing a more comprehensive framework for association studies. 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 [35, 36, 37, 38, 39]. 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 for grouping individuals by their direct genetic relationships without relying on externally imposed labels based on arbitrarily defined reference populations.

While other studies have identified correlations between genetics and geographical patterns at continental [40] and national levels, notably in the United Kingdom [18], the BIG cohort uniquely captures exceptional genetic diversity within a single urban area—Memphis City. 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, 41].

In addition, population stratification, admixture, and shared environmental factors represent significant sources of confounding in genome-wide association studies (GWAS) and admixture mapping, potentially leading to false positives and negatives [42, 43, 44, 45]. 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 [45, 42]. Usually, in GWAS studies, only close relationships or ancestry are taken into account [42], 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 [46]. The 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 applicability may depend on the population density, the degree of local genetic structure, and the availability of high-quality genomic and geographic data [47]. Future work should evaluate the integration of IBD-based categorization with existing correction strategies in GWAS [48] and gene-based association studies [49], particularly in cohorts with varying levels of endogamy, admixture, and environmental heterogeneity [44]. Another aspect is the definition of relatedness, which lacks a universally accepted formulation. As discussed by Speed and Balding [50], 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.

Ultimately, our findings highlight how a deeper understanding of genetically defined communities can enhance the study of health disparities, uncover hidden confounders in genomic analyses, and open new avenues for integrating gene–environment interactions into population-scale research.

4 Methods

4.1 Ethics

Our research followed the ethical guidelines established in the Declaration of Helsinki for human subjects' 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 the 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 **??**). 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 S5). 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 [51] 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 [52]. DeepVariant v0.10.0 with a custom exome model was used for variant calling [53], and the GLnexus v1.2.6 tool was used for joint variant calling [54]. The variants were annotated using a Variant Effect Predictor (VEP 110) [55]. Phasing was performed using ShapeIT v5 [56]. 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 in the BIG cohort, we inferred global ancestry with RFMix [9] using methodology described in [20]. Reference samples from 1000 Genomes and HGDP [34] were merged with BIG participants, and quality-controlled references [25] were filtered using ADMIXTURE (k = 4) [57] to retain samples with major ancestry proportion >0.99, defining African (AFR), American (AMR), European (EUR), and Asian (ASN) ancestries. Four-way deconvolution used RFMix with standard parameters (5 terminal nodes, 12 generations, 10 EM iterations); CSA ancestry was excluded due to negligible proportions. Discrete categories were defined as: >85% single ancestry for unadmixed groups; >15% from two ancestries summing to >85% for two-way admixed (EUR-AMR, EUR-AFR); and >15% from three or more ancestries for Multiway, with the 85% threshold reflecting that contributions >15% are accurate while lower values show higher error [58, 59].

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 [21]. 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 [60].

Post-processing protocols established by Browning et al. [61, 62] were applied to refine the detected segments (using Refined IBD [61]. 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 a 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 IBD sharing [60, 63, 62] using the pedsuite R package [64].

4.7 Network, community detection and topology

A genetic network was constructed in which each BIG participant was represented as a node, and edges were defined based on the total genome-wide IBD segment sharing. Edge weights were proportional to the total IBD shared between individuals. 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, as suggested in previous work [17]. For visualization and exploratory analyzes, the network was represented using the Fruchterman-Reingold layout.

We evaluated the Louvain and Leiden algorithms for community detection [23]. 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. 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 [23].

The hierarchical Leiden algorithm was applied, with modularity as the objective function, using a resolution parameter of 0.3 and a beta value of 2. These specific values were determined empirically to optimize the balance between identifying meaningful broad-scale community structures and maintaining balanced sizes of subcommunities. This approach aligns conceptually with previously published methods for hierarchical community detection based on IBD networks [17], but it explicitly specifies parameter values adapted to our dataset characteristics and study objectives.

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.

Network topology analysis was performed using NetworkX to characterize the structural properties of IBD-based genetic networks within each community partition. For each individual node, the connectivity ratio was computed as the normalized degree centrality, calculated by dividing the number of direct connections by the mean degree within the respective community. Betweenness was quantified using the standard algorithm that measures the

fraction of shortest paths between all node pairs that traverse through each individual node, subsequently normalized by the community-specific mean betweenness score to facilitate comparative analysis across communities with different topological characteristics. The normalization procedure ensures that values above 1.0 indicate above-average connectivity or centrality within the community context, while values below 1.0 represent below-average network importance. To be consistent and informative in our context, network topology analyses require more than 30 nodes; for this reason, C4 subcommunities were not included.

To contextualize the genetic structure of the BIG cohort and validate the robustness of the method to the inclusion of new participants, we performed a sensitivity analysis incorporating 4,091 genomic sequences from the joint HGDP-1000GP resource [34]. In this combined dataset (n=17,234 individuals), pairwise IBD was recalculated and processed using identical parameters and filtering thresholds as the BIG-only analysis. The same hierarchical Leiden algorithm was applied S7.

The first iteration identified ten communities, four of which (C1–C4) contained 99.7% of BIG individuals, corresponding to the previously applied BIG-only approach. Subsequent iterations revealed 72 subcommunities in total, with C2 exhibiting the highest number of subcommunities, followed by C4, C1, and C3. At the subcommunity level, classification remained concordant with the BIG-only analysis. Notably, many subcommunities were composed specifically of HGDP-1000GP individuals, demonstrating the model's sensitivity to detecting novel communities of individuals genetically distant from those currently identified in BIG. The number of subcommunities did not correlate directly with parent community size, suggesting the influence of genetic diversity rather than sample size alone.

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 S6), 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), as it provides an intuitive measure of goodness-of-fit specifically suited for logistic regression models by assessing the proportional improvement in log-likelihood relative to a null (intercept-only) model. This systematic approach allowed us to evaluate the relative predictive contribution of genetic, environmental, and socioeconomic factors, both independently and in combination.

4.12 Assessment of *APOL1* genotype association with kidney dysfunction

To evaluate the relationship between the APOL1 high-risk (HR) genotype and kidney dysfunction, logistic regression models were fitted using CKD status (defined as eGFR <45 mL/min/1.73 m^2) as the binary outcome. CKD status is registered in ICD-10 under the Genitourinary category. The primary predictor was APOL1 genotype categorized as HR (two risk alleles) versus low-risk (zero or one risk allele). Models were sequentially adjusted to assess the influence of demographic and genetic structure covariates: (i) age and sex; (ii) age, sex, and identity-by-descent proportion to account for cryptic relatedness; and (iii) age, sex, and the top principal components (PCs) of ancestry to correct for population stratification. Odds ratios (ORs) with 95% confidence intervals (CIs) were reported to evaluate model fit and the robustness of associations.

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, Pjotr Prins, Khyobeni Mozhui, Robert J Rooney, Robert W Williams, et al. Insights from the biorepository and integrative genomics pediatric resource. *Nature Communications*, 16(1):4750, 2025.

[21] 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.

[22] Ruhollah Shemirani, Gillian M Belbin, Christy L Avery, Eimear E Kenny, Christopher R Gignoux, and José Luis Ambite. Rapid detection of identity-by-descent tracts for mega-scale datasets. *Nature communications*, 12(1):3546, 2021.

[23] 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.

[24] 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.

[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] Huan Minh Tran, Feng-Jen Tsai, Yueh-Lun Lee, Jer-Hwa Chang, Li-Te Chang, Ta-Yuan Chang, Kian Fan Chung, Han-Pin Kuo, Kang-Yun Lee, Kai-Jen Chuang, et al. The impact of air pollution on respiratory diseases in an era of climate change: A review of the current evidence. *Science of the Total Environment*, 898:166340, 2023.

[28] Joseph NS Eisenberg, Manish A Desai, Karen Levy, Sarah J Bates, Song Liang, Kyra Naumoff, and James C Scott. Environmental determinants of infectious disease: a framework for tracking causal links and guiding public health research. *Environmental health perspectives*, 115(8):1216–1223, 2007.

[29] Yue Huang, Sui Yu, Zhanhe Wu, and Beisha Tang. Genetics of hereditary neurological disorders in children. *Translational pediatrics*, 3(2):108, 2014.

[30] Kyung-Shin Lee, Yoon-Jung Choi, Jinwoo Cho, Hyunji Lee, Heejin Lee, Soo Jin Park, Joong Shin Park, and Yun-Chul Hong. Environmental and genetic risk factors of congenital anomalies: an umbrella review of systematic reviews and meta-analyses. *Journal of Korean medical science*, 36(28), 2021.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

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

[38] 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.

[39] 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.

[40] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, et al. Genes mirror geography within europe. *Nature*, 456(7218):98–101, 2008.

[41] 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.

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

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

[44] 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.

[45] 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.

[46] 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.

[47] 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.

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

[49] 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.

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

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

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

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

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

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] Magnus Dehli Vigeland. *Pedigree Analysis in R*. Academic Press, 2021.

Regeneron Genetics Center

RGC Management and Leadership Team

Aris Baras, Goncalo Abecasis, Adolfo Ferrando, Giovanni Coppola, Andrew Deubler, 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, Cristen Willer, Lourdes Crane.

Sequencing and 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.

Genome Informatics and Data Engineering

Jeffrey G Reid[5], Mona Nafde[5], Manan Goyal[5], George Mitra[5], Sanjay Sreeram[5], Rouel Lanche[5], Vrushali Mahajan[5], Sai Lakshmi Vasireddy[5], Gisu Eom[5], Krishna Pawan Punuru[5], Sujit Gokhale[5], Shehroze Aamer[5], Pooja Mule[5], Mudasar Sarwar[5], Muhammad Aqeel[5], Xiaodong Bai[5], Lance Zhang[5], Sean O'Keeffe[5], Razvan Panea[5], Evan Edelstein[5], Ayesha Rasool[5], William Salerno[5], Evan K Maxwell[5], Boris Boutkov[5], Alexander Gorovits[5], Ju Guan[5], Alicia Hawes[5], Olga Krasheninina[5], Samantha Zarate[5], Adam J Mansfield[5], Lukas Habegger[5], Stephen Tahan[5].

Analytical Genetics and Data Science

Goncalo Abecasis[5], Manuel Allen Revez Ferreira[5], Joshua Backman[5], Kathy Burch[5], Adrian Campos[5], Liron Ganel[5], Sheila Gaynor[5], Benjamin Geraghty[5], Arkopravo Ghosh[5], Christopher Gillies[5], Lauren Gurski[5], Eric Jorgenson[5], Tyler Joseph[5], Michael Kessler[5], Jack Kosmicki[5], Adam Locke[5], Priyanka Nakka[5], Jonathan Marchini[5], Karl Landheer[5], Olivier Delaneau[5], Maya Ghoussaini[5], Anthony Marcketta[5], Joelle Mbatchou[5], Jonathan Ross[5], Carlo Sidore[5], Eli Stahl[5], Timothy Thornton[5], Rujin Wang[5], Kuan-Han Wu[5], Bin Ye[5], Blair Zhang[5], Andrey Ziyatdinov[5], Yuxin Zou[5], Jingning Zhang[5], Kyoko Watanabe[5], Mira Tang[5], Frank Wendt[5], Suganthi Balasubramanian[5], Suying Bao[5], Kathie Sun[5], Chuanyi Zhang[5], Sean Yu[5], Aaron Zhang[5], David Corrigan[5], Dhruv Shidhaye[5], Chen Wang[5], Keyrun Adhikari[5], Alexander Lachmann[5], Anna Alkelai[5], Mark Weiner[5].

Therapeutic Area Genetics

Adolfo Ferrando[5], Giovanni Coppola[5], Luca A. Lotta[5], Alan Shuldiner[5], Katherine Siminovitch[5], Brian Hobbs[5], Jon Silver[5], William Palmer[5], Rita Guerreiro[5], Amit Joshi[5], Antoine Baldassari[5], Cristen Willer[5], Sarah Graham[5], Ernst Mayerhofer[5], Erola Pairo Castineira[5], Mary Haas[5], Niek Verweij[5], George Hindy[5], Jonas Bovijn[5], Tanima De[5], Luanluan Sun[5], Olukayode Sosina[5], Arthur Gilly[5], Peter Dornbos[5], Moeen Riaz[5], Manav Kapoor[5], Gannie Tzoneva[5], Veera Rajagopal[5], Sahar Gelfman[5], Vijay Kumar[5], Jacqueline Otto[5], Jose Bras[5], Silvia Alvarez[5], Jessie Brown[5], Hossein Khiabanian[5], Joana Revez[5], Kimberly Skead[5], Jae Soon Sul[5], Lei Chen[5], Sam Choi[5], Amy Damask[5], Nan Lin[5], Charles Paulding[5], Sameer Malhotra[5], Joseph Herman[5], Jacob McPadden[5], David Blair[5].

Research Program Management and Strategic Initiatives

Marcus B Jones[5], Michelle G LeBlanc[5], Nadia Rana[5], Jennifer Rico-Varela[5], Jaimee Hernandez[5], Larizbeth Romero[5], Ashley Paynter[5].

Senior Partnerships and Business Operations

Randi Schwartz[5], Lourdes Crane[5], Alison Fenney[5], Jody Hankins[5], Anna Han[5], Samuel Hart[5], Ryan Smith[5].

Business Operations and Administrative Coordinators

Ann Perez-Beals[5], Gina Solari[5], Johannie Rivera-Picart[5], Michelle Pagan[5], Sunilbe Siceron[5].

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

Supplementary Figures

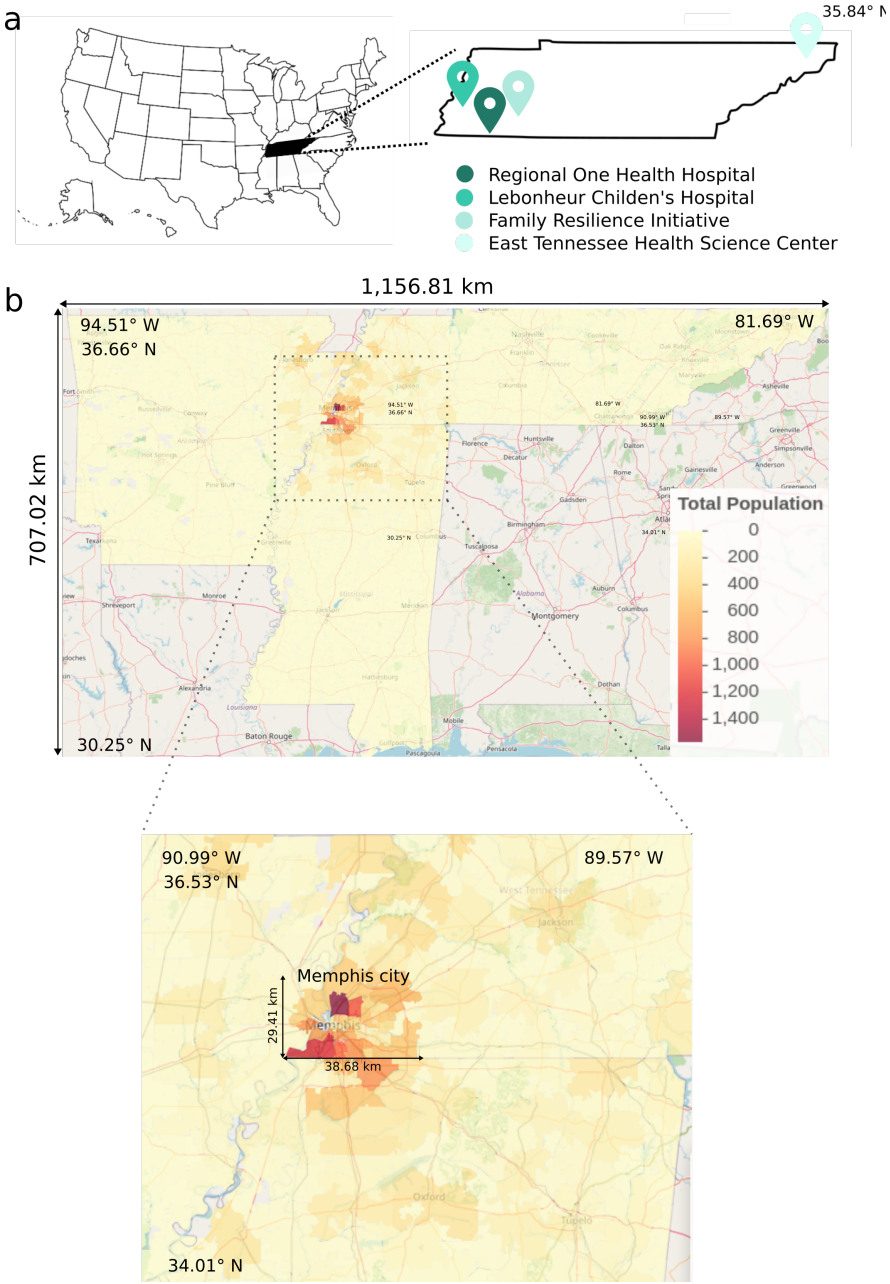

Figure S1: **Demography of enrolled participants** (a) The four sample sites (b) Number of enrolled participants by ZIP code. The region surrounding Memphis City is zoomed in, and distances in Km are shown. Maps were produced with the `leaflet` package (v. 2.2.1) using GeoJSON data for state ZIP-code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

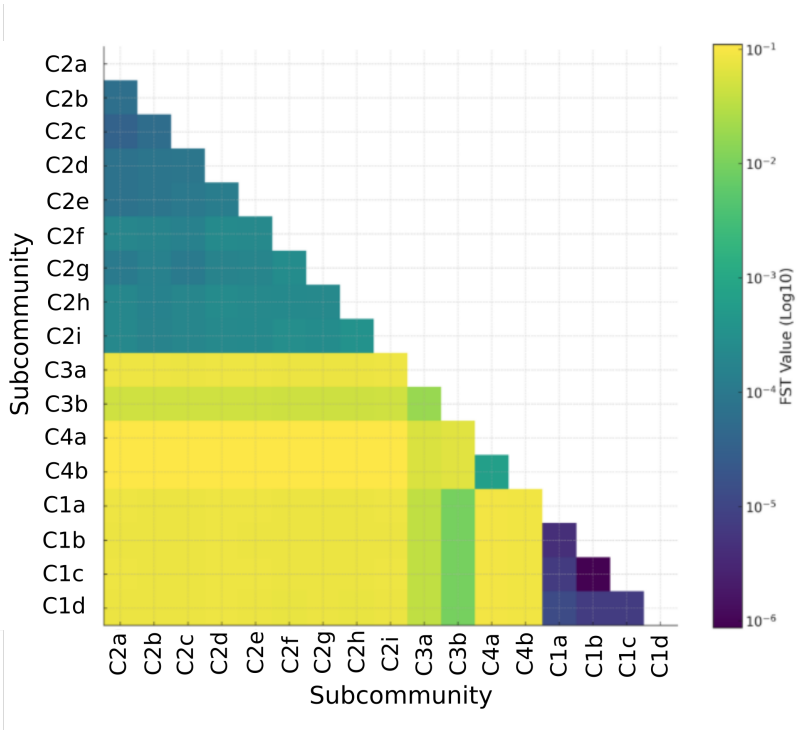

Figure S2: **Pairwise Fst values between subcommunities**. This analysis was performed using PLINK. It can be seen that subcommunities within C1 have the lowest fst values when analyzed each other. Contrastingly, C2 show higher values, indicative of more genetic diversity in the group.

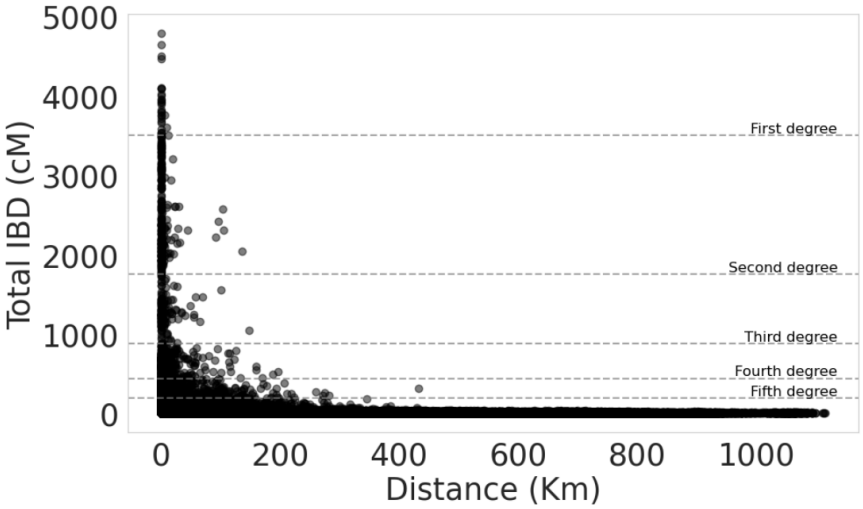

Figure S3: **Total IBD (cM) and Distance (Km)**. Different cutoffs for degrees of relationships are shown. The plot elucidates how higher total IBD values are related with closer distances. Geographical distances have been computed with sf R package [26].

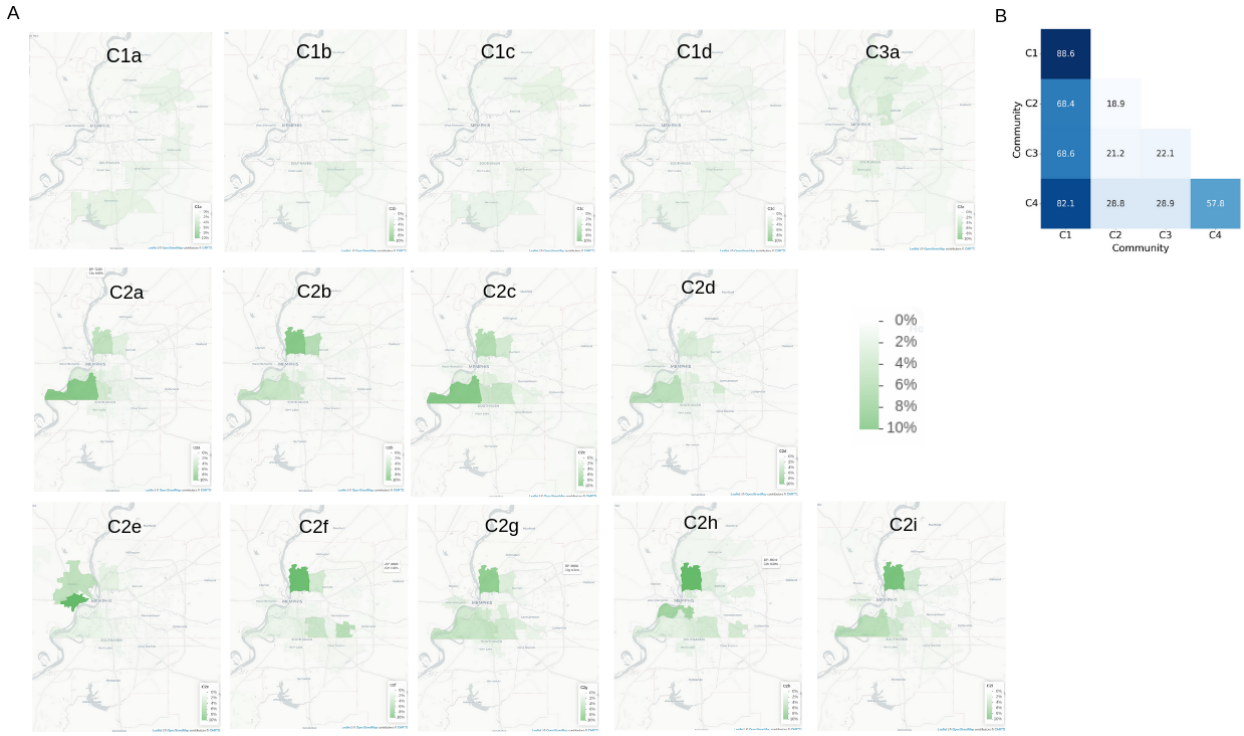

Figure S4: **Geographical distribution of subcommunities around Memphis area** A. Maps centered in Memphis city (34.96°-35.36°N: 88.70°-90.70° W) showing the proportional distribution of individuals from each subcommunity across zip codes. Maps were produced with the leaflet package (v. 2.2.1) using GeoJSON data for state ZIP-code boundaries publicly available. B. Median geographic distances in Kilometers, highlighting variability in spatial distributions within and across communities.

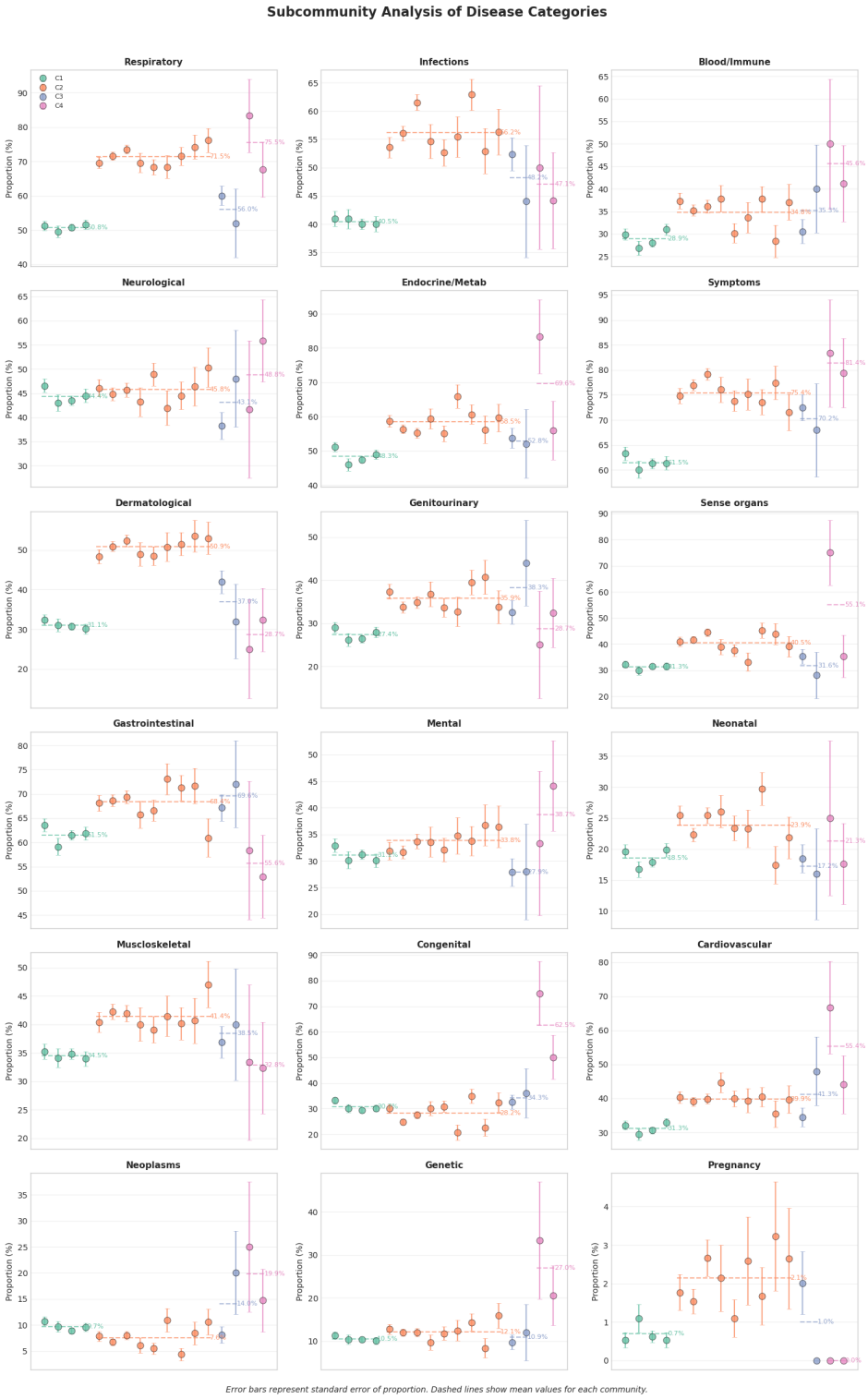

Figure S5: **Proportion of phenotypes at the subcommunity level**. CI95 for all proportions are shown and colored by community C1, C2, C3 and C4. The proportion is defined as the number of individuals in each community that reported, in at least one visit, a health condition in the corresponding category over the total. Error bars indicate 95% confidence intervals for proportions calculated using the Wilson score method.

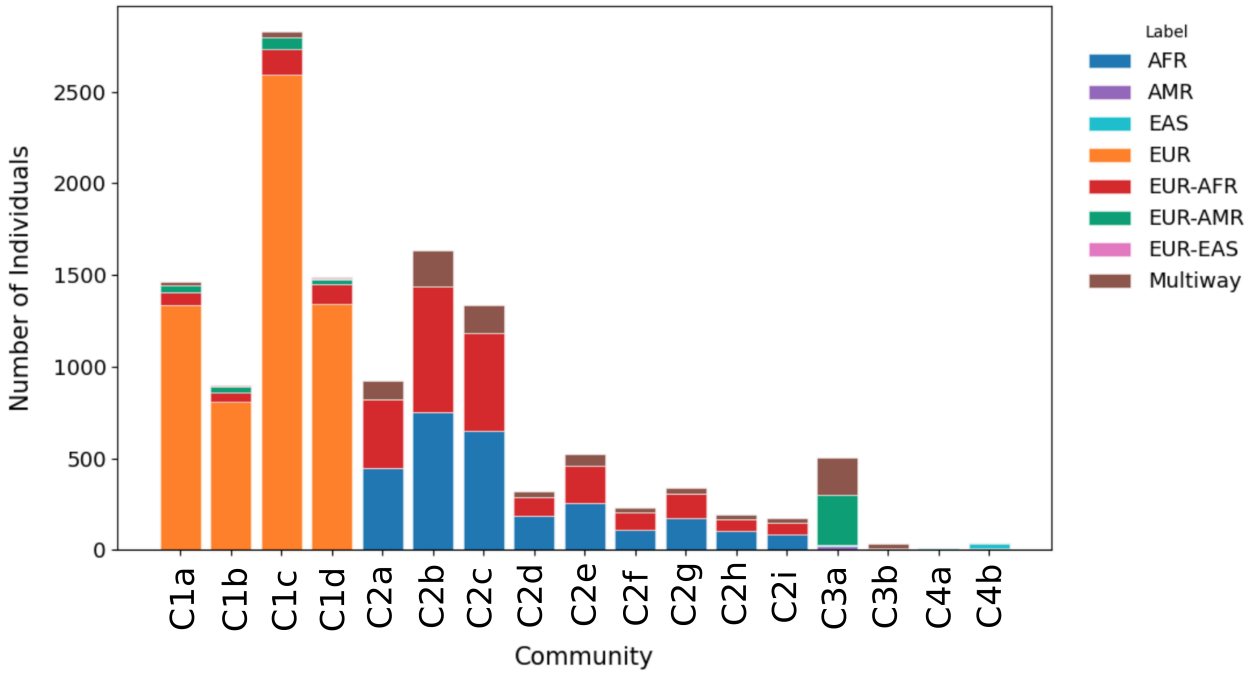

Figure S6: **Number of BIG individuals in each subcommunity**. We show the number of BIG individuals in each subcommunity , from the total of 13,143 incorporated. Inferred ancestries categories proportion by subcommunity are shown.

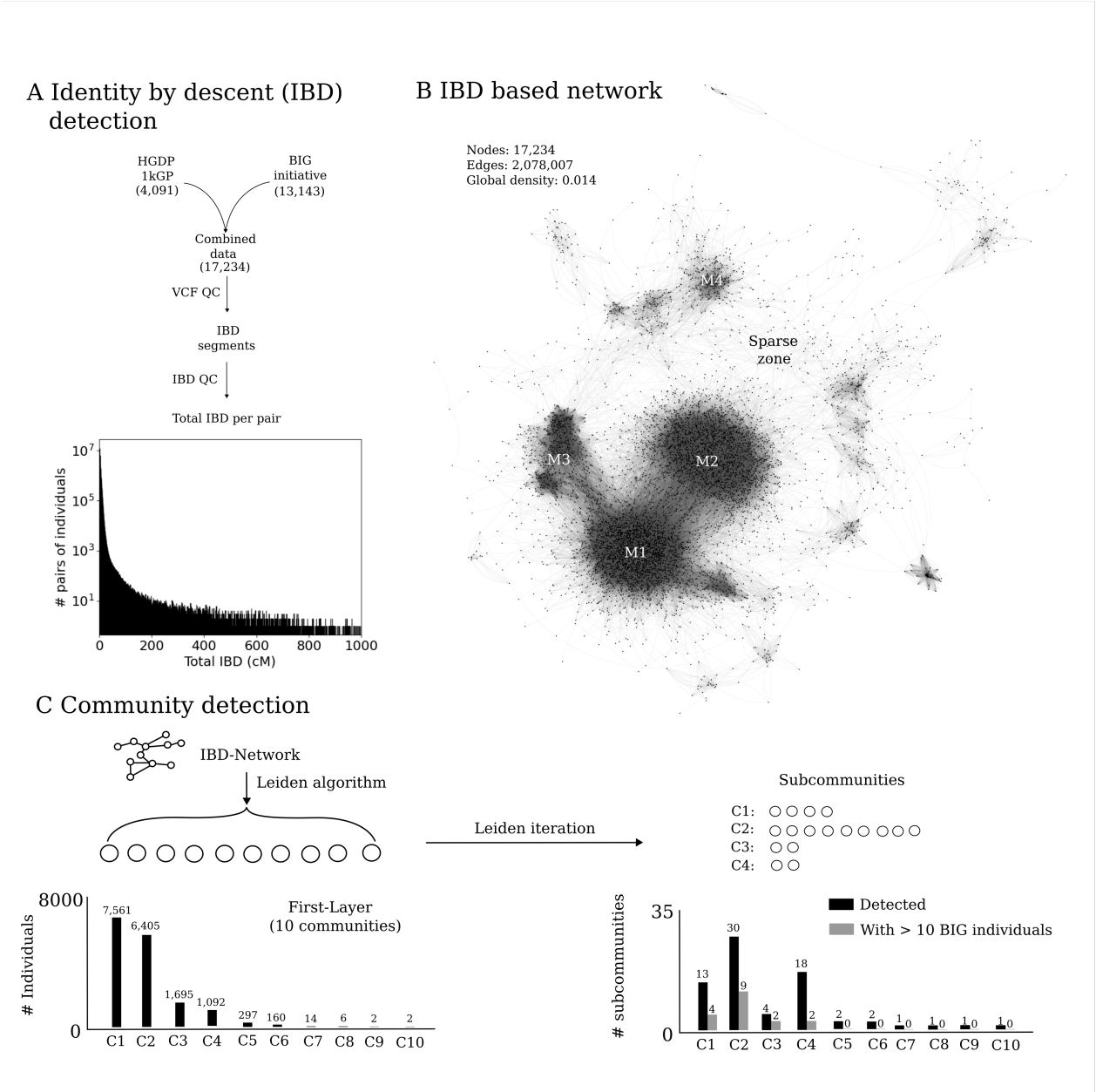

Figure S7: **Distribution of identity-by-descent (IBD) values and corresponding network analysis**. (A) Histogram showing the exponential decay of total IBD (cM). (B) IBD-based network with 17,234 nodes, revealing four dense communities (M1 to M4) and sparse zones. Fruchterman-Reingold layout was used 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.

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 [37].

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
C1a	-0.193	0.150	0.199
C1b	-0.121	0.110	0.271
C1c	-0.051	0.125	0.685
C2a	0.549	0.126	< 0.001
C2b	0.581	0.110	< 0.001
C2c	0.724	0.113	< 0.001
C2d	0.474	0.179	0.008
C2e	0.623	0.146	< 0.001
C2f	0.628	0.196	0.001
C2g	0.529	0.172	0.002
C2h	0.613	0.207	0.003
C3a	-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
C1a	-0.088	0.244	0.720
C1b	-0.101	0.184	0.583
C1c	-0.085	0.211	0.688
C2a	1.080	0.186	< 0.001
C2b	1.113	0.170	< 0.001
C2c	1.237	0.172	< 0.001
C2d	0.994	0.249	< 0.001
C2e	1.173	0.207	< 0.001
C2f	0.857	0.292	0.003
C2g	1.169	0.233	< 0.001
C2h	0.618	0.332	0.063
C3a	-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