Sequencing and health data resource of children of African ancestry

Leah C. Kottyan[1,2,3,4*], Scott Richards[3,4], Morgan E. Tracy[3,5], Lucinda P. Lawson[2], Beth Cobb[4,6], Steve Esslinger[3,5], Margaret Gerwe[3,5], James Morgan[3,5], Alka Chandel[7], Leksi Travitz[1,4], Yongbo Huang[1,4], Catherine Black[3,4], Agboade Sobowale[3,4,8], Tinuke Akintobi[8], Monica Mitchell [1,8,9], Andrew F. Beck[1,10,11,12,13], Ndidi Unaka[1,10,14], Michael Seid[1,13,15], Sonja Fairbanks[11], Michelle Adams[16], Tesfaye Mersha[1,17], Bahram Namjou[1,3,4], Michael W. Pauciulo[1,3,5], Jeffrey R. Strawn[18], Robert T. Ammerman[19], Daniel Santel[20], John Pestian[1,20,21], Tracy Glauser[1,22], Cynthia A. Prows[3], Lisa J. Martin[1,3], Louis Muglia[1,3], John B. Harley[23], Iouri Chepelev[23,24], Kenneth M. Kaufman[1,3,4,23*]

1. Department of Pediatrics. College of Medicine. University of Cincinnati. Cincinnati, Ohio.
2. Division of Allergy & Immunology. Cincinnati Children's Hospital Medical Center. Cincinnati, Ohio.
3. Division of Human Genetics. Cincinnati Children's Hospital Medical Center. Cincinnati, Ohio.
4. Center for Autoimmune Genomics and Etiology. Cincinnati Children's Hospital Medical Center. Cincinnati, Ohio.
5. Discover Together Biobank. Cincinnati Children's Hospital Medical Center. Cincinnati, Ohio.
6. Center for Stem Cell & Organoid Medicine (CuSTOM), Cincinnati Children's Hospital Medical Center. Cincinnati, Ohio.
7. Information Services for Research (IS4R). Cincinnati Children's Hospital Medical Center. Cincinnati, Ohio.
8. Office of Community Relations. Cincinnati Children's Hospital Medical Center. Cincinnati, Ohio.

9. Division of Behavioral Medicine and Clinical Psychology. Cincinnati Children's Hospital Medical Center. Cincinnati, Ohio.
10. Division of General & Community Pediatrics. Cincinnati Children's Hospital Medical Center. Cincinnati, Ohio.
11. Division of Hospital Medicine. Cincinnati Children's Hospital Medical Center. Cincinnati, Ohio.
12. Office of Population Health and Michael Fisher Child Health Equity Center. Cincinnati Children's Hospital Medical Center. Cincinnati, Ohio.
13. Anderson Center. Cincinnati Children's Hospital Medical Center. Cincinnati, Ohio.
14. Department of Pediatrics, Stanford University School of Medicine. Stanford, California.
15. Division of Pulmonary Medicine. Cincinnati Children's Hospital Medical Center. Cincinnati, Ohio.
16. Cincinnati Children's Research Foundation. Cincinnati Children's Hospital Medical Center. Cincinnati, Ohio.
17. Division of Asthma Research. Cincinnati Children's Hospital Medical Center. Cincinnati, Ohio.
18. Department of Psychiatry and Behavioral Neuroscience, University of Cincinnati School of Medicine. Cincinnati, Ohio.
19. Division of Behavioral Medicine and Clinical Psychology, Cincinnati Children's Hospital Medical Center, Department of Pediatrics, University of Cincinnati. Cincinnati, Ohio.
20. Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center
21. Computational Medicine Center, Cincinnati Children's Hospital Medical Center. Cincinnati, Ohio.
22. Division of Neurology. Cincinnati Children's Hospital Medical Center. Cincinnati, Ohio.
23. US Department of Veterans Affairs Medical Center, Cincinnati, Ohio. Cincinnati, Ohio.
24. Research Service, US Department of Veterans Affairs Medical Center, Cincinnati, Ohio

* Corresponding authors: Leah Kottyan, Leah.Kottyan@cchmc.org, 513-636-1316; Kenneth Kaufman, Kenneth.Kaufman@cchmc.org, 513-803-5385

Abstract:

Purpose

Individuals who self-report as Black or African American are historically underrepresented in genome-wide studies of disease risk, a disparity particularly evident in pediatric disease research. To address this gap, Cincinnati Children's Hospital Medical Center (CCHMC) established a biorepository and developed a comprehensive DNA sequencing resource including 15,684 individuals who self-identified as African American or Black and received care at CCHMC.

Methods

Participants were enrolled through the CCHMC Discover Together Biobank and sequenced. Admixture analyses confirmed the genetic ancestry of the cohort, which was then linked to electronic medical records.

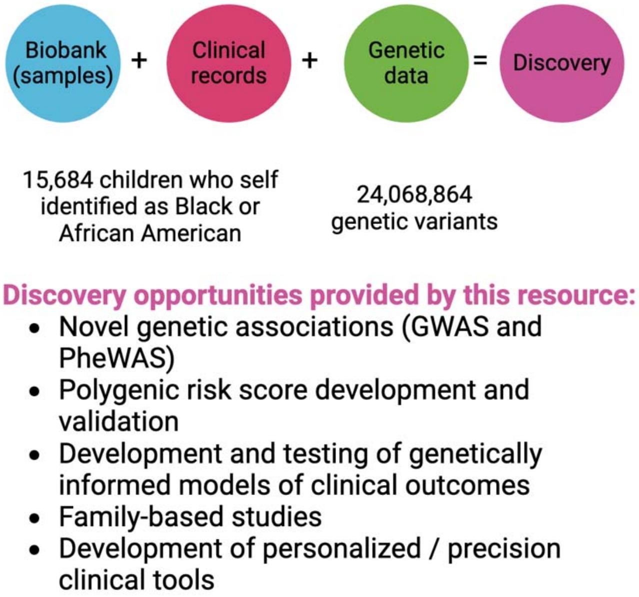
Results

High-quality genome-wide genotypes from common variants accompanied by medical record-sourced data are available through the Genomic Information Commons. This dataset performs well in genetic studies. Specifically, we replicated known associations in sickle cell disease (*HBB*, $p = 4.05 \times 10^{-1□□}$), anxiety (*PLAA3*, $p = 6.93 \times 10^{-□}$), and asthma (*PCDH15*, $p = 5.6 \times 10^{-1□}$), while also identifying novel loci associated with asthma severity.

Conclusion

We present the acquisition and quality of genetic and disease-associated data and present an analytical framework for using this resource. In partnership with a community advisory council, we have co-developed a valuable framework for data use and future research.

Graphical abstract



15,684 children who self identified as Black or African American

24,068,864 genetic variants

Discovery opportunities provided by this resource:

- Novel genetic associations (GWAS and PheWAS)
- Polygenic risk score development and validation
- Development and testing of genetically informed models of clinical outcomes
- Family-based studies
- Development of personalized / precision clinical tools

Introduction

Individuals of self-reported African ancestry have historically been underrepresented in genetic studies of diseases, reflecting broad disparities in genetic research and clinical medicine[1, 2]. This underrepresentation stems from systemic inequities in research funding, recruitment practices, and access to healthcare, which together limit the inclusion of diverse populations in study cohorts[1-4]. As a result, most genomic data have been derived from individuals of European ancestry, leading to significant gaps in understanding the genetic underpinnings of diseases affecting Black or African American adults and children[5, 6]. This lack of diversity undermines the ability to identify ancestry-specific genetic variants, which reduces the applicability of genomic findings, precision medicine interventions, and risk prediction models for certain populations[1, 7, 8]. Addressing ancestry-based disparities in research is critical for ensuring equitable advances in healthcare and bridging gaps in disease prevention and treatment outcomes for underrepresented groups[1].

The underrepresentation of individuals of African ancestry in genome-wide studies of disease risk has been a persistent issue, particularly in the context of pediatric diseases[2, 9]. This lack of representation not only limits our understanding of disease etiology across ancestries, but hampers the development of effective, equitable healthcare solutions[1, 3]. In 2024, 19.8% of the patients who received care at CCHMC and 15.6% of patients enrolled in the Discover Together Biobank with a DNA sample available identified as Black or African American. To address this critical gap, CCHMC has undertaken a significant initiative to create a comprehensive sequencing resource focused on broadly expanding opportunities to identify the genetic basis of health and disease in children of African ancestry.

The primary objective of this study was to acquire high-quality genetic data from participants in the Discover Together Biobank who self-identified as Black or African American and had accompanying electronic medical records. To achieve this goal, we employed a low-pass sequencing strategy[10-12], which yielded an average genomic coverage of around 3X (i.e. meaning that the average number of reads per base is ~3). The low pass approach enabled us to significantly broaden the scope of our study compared to what would have been feasible with standard 30X read depth whole genome sequencing.

To demonstrate the utility of this resource, our secondary objective was to conduct GWAS analyses for three diverse human disorders; sickle cell disease (OMIM #603903), anxiety disorder (OMIM #607834), and asthma (OMIM #600807). Sickle cell disease is a severe inherited blood disorder that predominantly affects Black and African American children in the United States, with an estimated prevalence of approximately 1 in 365 African American births in this population[13, 14]. It is almost always caused by at least one hemoglobin S allele in combination with a second pathogenic variant in the *HBB* gene, resulting in the production of abnormal hemoglobin S[14, 15]. When deoxygenated, hemoglobin S polymerizes, leading to the characteristic sickling of red blood cells, increased hemolysis, and vaso-occlusive events[16]. Clinically, sickle cell disease manifests with episodes of severe pain (vaso-occlusive crises), anemia, increased susceptibility to infections, and progressive organ damage, significantly impacting quality of life and life expectancy[13].

Anxiety disorders are among the most common mental health conditions in children, with prevalence estimates ranging from 7% to 10%, though rates may be higher due to underdiagnosis in some populations[17]. Black and African American children experience anxiety at similar or higher rates than their White peers, yet they are less likely to receive a diagnosis or access appropriate mental health care[18, 19]. While genetic studies have identified variants associated with anxiety risk, there has been limited research specifically examining genetic contributors in African American youth, leading to gaps in understanding potential ancestry-specific risk factors.

Asthma is one of the most common chronic diseases in children, with a disproportionate burden among Black and African American children, who experience higher prevalence, increased severity, and greater rates of hospitalization and mortality compared to other racial and ethnic groups[20, 21]. In the United States, Black children have nearly twice the prevalence of asthma as White children, with environmental, socioeconomic, and genetic factors contributing to this disparity[22, 23]. Genetic studies have identified multiple loci associated with asthma risk, including variants in genes such as *TSLP*[24]**,** *GSDMB*[25-27], and *IL1RL1*[28], though much of the research has been conducted in populations of European ancestry, limiting insights into genetic contributors in African American pediatric cohorts[29].

Herein, we present a CCHMC-based biobank resource, which encompasses deep genetic and phenotypic data from 15,684 individuals who self-identified as African American or Black and received care at CCHMC. We detail the acquisition and quality of the genetic and health data contained in this resource, and we also recommend an analytical framework for using these data alongside various analyses that demonstrate the power and potential of this unique resource. Importantly, this initiative has been developed in partnership with a community advisory council, ensuring that the framework for data use and future research is aligned with community needs and perspectives (see accompanying perspectives piece).

Methods

Collaboration with a community advisory committee

Meeting every other month over the course of two years, the committee facilitated discussions focused on ensuring that data and biological samples were kept safe while promoting transparency in research practices. A key focus was increasing the representation of historically underrepresented populations, particularly Black and African American children, in genomic studies to address longstanding disparities in research participation. The committee worked collaboratively to develop and review community engagement strategies, identify effective methods of communication, and strengthen relationships between community-engaged research and partners. The insights from this group directly influenced how we presented data from the cohort and shaped our decision to prioritize genetic studies of asthma and anxiety, conditions that disproportionately impact Black and African American children (**See co-submitted Perspective article submitted by members of this advisory committee**).

Discover Together Biobank

The CCHMC Discover Together Biobank is an actively-enrolling, multifaceted initiative that collects biospecimens with associated phenotypic and genotypic data while combining clinical research protocols with appropriate governance and accession processes. Discover Together Biobank serves as a key piece of infrastructure to aid the research of CCHMC investigators and currently contains samples and electronic medical record data from over 100,000 participants. Discover Together Biobank is constructed from a patient-based participant population that spans multiple institutional biobanking consent periods at CCHMC.

Participants for this study were consented to the CCHMC IRB approved study during a consent period as part of our “Better Outcomes for Children” cohort, which used broad consent to enroll

CCHMC patients to ascertain clinical residual blood samples for approved research use. The Better Outcomes for Children consent allows for use of DNA and associated medical record data for research, including genomics.

Deposition of Data in the Genetic Information Commons

Data from this sequencing resource is available through the Genomic Information Commons (GIC). The GIC is a research initiative allowing researchers to collaborate across multiple academic medical centers to share consented patient genomic data for large-scale research projects aimed at discovery of genetic mechanisms of health and disease and improving patient health outcomes (https://www.genomicinformationcommons.org/)[30]. The GIC facilitates collaborative research by enabling sharing of genomic data, phenotypic information, and biospecimen metadata across participating institutions.

DNA preparation and sequencing: See Supplemental Note (**Supplemental Methods**)

Assessment of relatedness between samples

Family relationships were determined using the KING software package (version 2.1.6)[31]. **Supplemental Dataset 3** provides relatedness information for this cohort.

Principal component analysis and ancestry assessment

Principal component analysis was performed using the software package Plink2 (version 2.0-161017). An LD pruned data set was generated with Plink2 (independent-pairwise 50kb window, 5 step, 0.8 r^2 threshold and minor allele frequency between 0.15 and 0.4). The first 10 principal components were approximated using Plink2. Admixture for each sample was calculated using the software program admixture version 1.3.0 assuming 4 populations[32]. 155 individuals had less than 2% African admixture and were removed from this cohort.

Genome-wide association studies

Study cohorts were generated by identifying cases in the biorepository samples with matching ICD 10 codes. Only one case per family was used for analysis. All families with an affected sample were dropped as a source for potential control samples. Cases were matched to controls using the R package PCAmatchR (5th release) based on the first three principal components[33]. Cases were matched to controls with a match distance less than 0.1 as determined by PCAmatchR. The number of controls per case was dependent on the total number of cases and available potential controls.

Quality control of genotype data was performed on each cohort. Variants with a case minor allele frequency less than 1% or a variant call rate below 85% were dropped. In addition, any variant with a control HWE p-value below 0.0001 was dropped. All QC measurements and association analyses were performed with the Plink1.90b6.26[34].

Interactive Locus Zoom results are available for the association studies presented in this manuscript:

Sickle cell disease vs controls:

https://my.locuszoom.org/gwas/378516/?token=788c8c35d21348b996cd25c638d4ebc3

Anxiety vs controls:

https://my.locuszoom.org/gwas/852976/?token=57f16f9e8123412192f061c71246f89c

Asthma vs controls:

https://my.locuszoom.org/gwas/407471/?token=ab55b1868cb141ac979a84b3e6859c10

Severe versus controlled asthma:

https://my.locuszoom.org/gwas/48795/?token=6823af24c2f94f89af20ce3a20b0b6ca

See **Supplemental Note** for a suggested protocol for case - control studies using this resource.

Partnerships to connect expert clinical phenotyping with high quality genetic data: acute asthma exacerbation

We partnered with clinical experts in the Asthma Learning Health System,[21] an asthma-focused learning network at Cincinnati Children's, to identify carefully phenotyped groups and perform clinically directed analyses. The Asthma Learning Health System aims to advance equitable asthma care, with the goal that all children with asthma can live their best lives. The network's clinical team identified children treated at Cincinnati Children's to manage a diagnosis of asthma, further identifying two key groups: 99 participants who required acute care in the emergency department for asthma and 393 participants who managed their condition without needing acute interventions during the study period. Leveraging genomic data from this sequencing resource, we conducted a genome-wide association study (GWAS) to investigate genetic factors contributing to asthma severity and healthcare utilization.

Results

A Community Advisory Committee played a critical role in guiding the way we developed this sequencing research. Including Community Advisory Committee perspectives was and remains critical as we strive for maximally ethical and inclusive conduct of pediatric genetic research, particularly in relation to biorepositories and data security (**see co-submitted Perspectives article authored by members of advisory committee**).

The final dataset contains high quality sequence data for 15,684 samples that correspond to consented patients with electronic medical record data at Cincinnati Children's. In this sequencing resource, 53.5% of the samples were from females (**Figure 1A**). The average genomic coverage was 2.66 (median: 2.60, interquartile range: 1.12) (**Figure 1B**). The mean age of enrollment was 9.68 (with a standard deviation of 7.33) years old, and the average current age of participants is 19.77 years. Samples had an average of 3,155,811 called variants with allele frequencies greater than 1%, with a range from 2,542,625 to 3,417,031 (**Figure 1C**). The average per variant call rate was 93.2%, ranging from 84.4%-96.7% (**Figure 1D**).

A total of 77.8% of the participants in this study self-identified as Black or African American in their corresponding electronic medical record data, with the remaining individuals identifying as more than one race and/or ethnicity or did not provide data (**Figure 2A**). We performed principal component and admixture analysis to identify genetic ancestry, using 1000 genomes data as a reference (**Figure 2 B and C**). As expected, most participants in this study had 75-85% African admixture. Strikingly, there was a large range of African admixture in this cohort, ranging from 2-98% (**Figure 2C** and **Supplemental Dataset 1**). The number of non-reference (alternative) alleles called within each sample was strongly correlated with the proportion of African ancestry (**Figure 2D**), which is consistent with the enhanced genetic diversity of the populations from Africa relative to other global populations [35, 36].

As expected from a pediatric biobank-based study, there was substantial relatedness between participants in the cohort (**Table 1**). A total of 1,050 siblings were identified in addition to 41 monozygotic twin pairs. Also, 98 parent-offspring pairs were identified, providing opportunities to study genetic heritability in the context of specific health outcomes; 134 2nd degree relatives were also identified (**Table 1**). A relatedness matrix was developed to support additional studies using this cohort (**Supplemental Dataset 2**). Altogether, these results highlight opportunities to study health outcomes within families and underscore the importance of accounting for relatives in case-control based studies within this resource.

Clinical characteristics of the cohort

Electronic medical records linked to the subjects in this study include 10,472 distinct ICD 10 codes, with the top 20 codes listed in (**Table 2**). The most common codes include those for infections, common symptoms of childhood illnesses, gastrointestinal complaints, attention-deficit hyperactivity disorder, obesity, and dermatitis. Other top ICD 10 codes were associated with routine medical care of children including encounters for immunizations. Current Procedure Terminology (CPT) codes are used to report medical, surgical, and diagnostic procedures and services, while Healthcare Common Procedure Coding System (HCPCS) codes report medical procedures, supplies, and services that are not included in CPT. In this cohort, 18,708 distinct CPT and HCPCS codes were reported. The top reported codes in this cohort were for medical visits (ranging from routine visits for established patients to complex emergent encounters), venipuncture (with associated blood tests), vaccination, urinalysis, tonsillectomy with adenoidectomy, and spirometry (**Table 3**). Of the 11,717 distinct medication codes used in this cohort, the most common were analgesics, intravenous saline, antiemetics, albuterol, antibiotics, and steroids (**Table 4**). Although participants are largely from Metropolitan Cincinnati, the area surrounding CCHMC, home addresses were reported across most of the United States (**Figure 3**). Because street addresses and zip codes are available from each medical encounter of a participant within the CCHMC's system, studies integrating genetic data with environmental data gleaned from geocoding are possible.

Exemplary case control genome wide association studies using this cohort.

Sickle-cell Disorder: In our cohort, 754 individuals had the D57 ICD-10 code (coding for sickle-cell disorders) and were identified as “cases”. Using the framework presented in **Supplemental Figure 3**, 4,484 controls were identified. Because this sequencing resource includes common variants, this analysis was intended to identify modifiers of sickle-cell disorder. The most significant genetic association identified in this analysis consisted of a group of variants in a genetic haplotype tagged by rs4426157 (p-value = $4.05x10^{-148}$; odds ratio = 4.38) (**Figure 4A** and **Supplemental Table 3**), which is located six kilobases downstream of HBB. Variants in this haplotype have also been identified in independent cohorts assessing hemolysis in sickle cell anemia (10^{-10} < p-values <10^{-29})[37] and thromboembolic events in sickle cell disease (10^{-6} < p-values <10^{-8})[38].

Anxiety disorders: Using a clinical phenotyping algorithm developed by CCHMC investigators, we identified 3,733 subjects with anxiety in this cohort. After identifying 7,341 controls, we performed an association study and found three variants with genome-wide association (**Figure 4B** and **Supplemental Dataset 3**): rs78667939 near *PLAAT3* (p-value = $6.93x10^{-9}$; odds ratio = 1.55), rs77200838 in an intergenic region (p-value = $2.75x10^{-8}$; odds ratio of risk allele = 1.58), and rs181943174 near *FAM98A* and *RASGRP3* (p-value = $2.12x10^{-8}$; odds ratio = 2.09).

Asthma: After identifying 4,303 cases with electronic medical record codes of ICD-10 J45 and 8,342 principal component-matched controls, we performed a genome wide association study. We identified one locus with a variant exceeding genome-wide significance: rs185118029 near *PCDH15* ($5.6x10^{-10}$; odds ratio = 1.71) (**Figure 4C** and **Supplemental Dataset 3**). Variants at the *PCDH15* locus have not previously been identified for asthma. Additionally, numerous risk loci from our analysis with p-values ranging from 10^{-4} to 10^{-7} included variants that were previously reported as increasing asthma risk in independent studies. These included replication of known asthma risk loci at *GSDMB*[26, 27], *IL1RL1*[39, 40], *TSLP*[41], *RAD50*[42], *D2HGDH*[39, 40, 42-50], *CDHR3*[42], and *PRDM11*[44] (**Supplementary Dataset 4**).

Asthma Learning Health System acute asthma GWAS results

The Asthma Learning Health System dataset analysis yielded suggestive associations at rs35317849 near *CUL2* (p-value = 3.98 x10^{-7}; odds ratio = 2.14) and rs73846614 between *VGLL3* and *CHMP2B* (p-value = 8.16x10^{-7}; odds ratio = 3.48), highlighting potential genetic markers that may influence asthma exacerbations and emergency care needs (**Figure 5A**).

To further understand the role of combined genetic individual risk for individuals in this cohort, we calculated polygenic risk scores using a tool that was previously developed and validated across multiancestral asthma cohorts[51]. While not significantly different, the participants who required more acute care trended towards having a higher burden of asthma risk variants (one sided p-value of 0.109) (**Figure 5B**). Moreover, because we were able to provide a single datapoint that quantifies the cumulative genetic risk for each individual in this study, the clinical investigators were able to use this score to adjust for genetic risk when assessing other environmental and clinical factors. Together, these results highlight opportunities for integrating genetic research into precision medicine approaches for asthma management.

Discussion

In this study, we established a comprehensive sequencing and phenotyping resource that addresses the historical underrepresentation of individuals of African ancestry in genetic studies, particularly in pediatric disease research. By leveraging the Discover Together Biobank at CCHMC, we produced a rich dataset comprising deep genetic and phenotypic data from 15,684 children who self-identified as African American or Black. The study was made possible using low-coverage (3.3X) whole genome sequencing and an innovative imputation method to accurately call genotypes. We demonstrate that the resulting data are of high quality using heterozygous concordance of samples from the 1000 genomes project as well as samples from this cohort for which we had corresponding 30X whole genome sequencing data. We present a framework for case-control assessment using this cohort and show opportunities for discovery by presenting exemplary analyses. Our GWAS identified new and replicated previously known significant genetic associations for sickle cell disease, asthma, anxiety, and severe asthma - demonstrating the power of this resource to advance our understanding of the genetic underpinnings of pediatric diseases.

Our collaboration with a Community Advisory Committee has been instrumental in ensuring that our research is conducted ethically and inclusively. The committee's insights guided our efforts to increase the representation of historically underrepresented populations in genomic studies and to develop community engagement strategies that promote transparency and trust. This partnership has also informed our decision to prioritize genetic studies of conditions that disproportionately impact Black and African American children, such as asthma and anxiety.

The genetic diversity observed in our cohort, of African ancestry ranging from 2% to 99%, highlights the potential to uncover ancestry-specific genetic variants that may contribute to disease risk. This range of admixture also creates an opportunity for admixture-based genetic studies that leverage the differences in local and global admixture to identify variants associated with phenotypes with ancestry-specific differences. These opportunities are particularly relevant for conditions such as sickle cell disease, anxiety, and asthma, each with morbidity which disproportionately affect Black and African American children. Gaining information from people of diverse ancestral backgrounds is also essential to fully capture the biological mechanisms underlying health and disease [52].

The inclusion of related individuals, such as multi-generational parent-child dyads and siblings, further enhances the utility of this resource for future family-based studies. This allows for the investigation of heritability and the identification of genetic factors that contribute to health outcomes within families. Additionally, the integration of electronic medical record data with genetic data enables the exploration of gene-environment interactions and the impact of socio-environmental factors on health outcomes.

Since this sequencing dataset was derived from Discover Together Biobank and made searchable via the Genomic Information Commons cohort builder, the dataset is well positioned to be easily accessible via a transparent, equitable, and tracked request system. The review process is built around the Genomic Information Commons sample request tool, with review procedures following the Discover Together Biobank protocols in place for institutional samples and data.

In conclusion, the unique sequencing resource we developed represents a significant step towards addressing the disparities in genomic research and advancing precision medicine for all. By providing high-quality genetic and phenotypic data, this resource has the potential to drive discoveries that improve health outcomes for children of African ancestry. Future research should continue to leverage this resource to explore the genetic basis of pediatric diseases and to develop targeted interventions that promote health equity.

Data Availability

The data described in this manuscript have been deposited in the GIC (https://www.genomicinformationcommons.org/).

Acknowledgements

The authors would like to acknowledge the CCHMC Genomics Sequencing Facility (RRID:SCR_022630). This study used samples, data, and/or services from Cincinnati Children's Discover Together Biobank [RRID: SCR_022632]. We thank the Discover Together Biobank for support of this study and the participants and their families, whose help and participation made this work possible. Schematic figures were made in BioRender.

Funding Statement

Cincinnati Children's funded this work through recruitment packages to LM and JBH as well as funding of an Academic Research Committee award to JBH, LCK, and MTW. This research was partially funded by National Institutes of Health (NIH) U01 HG011172 and U19 AI07235 to L.C.K and P30 AR070549 to LCK and M.T.W. and R01HG011411 to T.B.M.

Author Contributions

Conceptualization: LCK, IC, LM, LBH, KMK, ; Data curation: LCK, SR, MT, LPL, AC, BN, IC, KMK; Formal analysis: LCK, KMK, BN; Funding acquisition: LM, JBH; Visualization: LCK, KMK; Writing-original draft: LCK and KMK; Writing-review & editing: LCK, SR, MT, LPL, SE, MH, JM, AC, LT, YH, CF, AS, TA, MM, AB, MS, NU, SF, BC, MB, MA, TM, BN, MWP, JRS, RTA, DS, JP, TG, CP, LJM, IC, LM, JH, KMK.

Ethics Declaration

The CCHMC Institutional Review Board (IRB) reviewed and approved this study. As detailed in the methods section, samples were obtained from the Discover Together Biobank after approval from their sample use committee.

Conflict of Interest

The authors report no conflicts of interests.

Figure Legends

Figure 1. Genotype data of the cohort. A. Distribution of female and male participants. B. Genomic coverage with a line indicating the median. C. Distribution of total genetic variants called with a line indicating the median. D. Per participant call rate of variants included in the final dataset.

Figure 2. Ancestry of children in the cohort. A. Self-reported race as collected in the electronic medical record for each participant. B. Principal component analysis of each participant (black dots) in the context of reference individuals from the 1000 Genomes Project, whose ancestry is indicated by color (see legend). C. Distribution of African admixture across individuals in the cohort. D. The number of variants called per sample (dots) in relationship to the genomic proportion of African admixture. CCHMC: Cincinnati Children's Hospital Medical Center.

Figure 3. Geographic distribution of children in this cohort. Each dot represents a zip code with at least one participant in this study. Dots are sized and have transparency adjusted based on the number of participants who lived in the corresponding zip code. All zip codes were geocoded and plotted with Esri ArcGIS Pro 3.4. Map sources: Esri, TomTom, Garmin, FAO, NOAA, USGS, EPA, USFWS.

Figure 4. Exemplary case-control genome-wide associations studies using this cohort. Genome-wide association studies were performed using the framework presented in Supplemental Figure 3 for sickle cell disease (note break on y-axis) (A), Anxiety (B), and Asthma (C). For each study, the negative log p-value of a logistic regression analysis that

accounts for the first three principal components is shown. Genomic inflation for these analyses were 1.033, 1.048, and 1.042 for sickle cell disease, anxiety, and asthma, respectively.

Figure 5. Partnerships to connect expert clinical phenotyping with high quality genetic data: acute asthma exacerbation. A. Individuals who were identified by clinicians as having controlled asthma and those with acute asthma exacerbations (labeled as severe asthma) were compared in a genome-wide association study. The genomic inflation for this analysis was 1.021. B. A polygenic risk score for asthma was used to quantify the cumulative burden of asthma genetic risk loci for each individual in the study[51]. The difference in the distribution of polygenic risk was not statistically significant (one sided p-value of 0.108).

Supplemental Figure 1. Quality assessment of sequencing data. Heterozygous concordance refers to the rate at which a variant is correctly identified as heterozygous (having one copy of each allele) across sequencing studies. A. Heterozygous concordance (y-axis) of GM12878 for each sequencing run of this study was assessed using corresponding 30X sequencing and is shown in the context of genomic coverage (x-axis). Heterozygous concordance (B) and full concordance (C) is presented for each control sample from which there was corresponding 30X sequencing data (see Methods).

Supplemental Figure 2. Impact of variant-level quality control thresholds. Heterozygous concordance for low-pass sequencing of GM12878 is presented with no filters, with a 1% allele frequency filter, with a Hardy-Weinberg equilibrium threshold of 0.0001, and with both allele frequency and Hardy-Weinberg filters.

Supplemental Figure 3. General recommended analytical strategy for case control analyses using this cohort.

Tables

Table 1. Relatedness of cohort out of 15,684 individuals. A full list of individuals in our cohort and their relatedness to other participants in the cohort can be found in Supplemental Dataset 2.

Relation	Count
Monozygotic twins	41
Parent - Offspring	98
Siblings	1,050
2nd degree	134

Table 2. Top 20 International Classification of Diseases (ICD)-10 codes associated with cohort

Number	ICD-10 code	Description
2,916	J06.9	Acute upper respiratory infection, unspecified
2,446	R50.9	Fever, unspecified
2,137	R05	Cough
2,033	J45.909	Unspecified asthma, uncomplicated
1,963	R11.10	Vomiting, unspecified
1,915	K59.00	Constipation, unspecified
1,813	B34.9	Viral infection, unspecified
1,769	R10.9	Unspecified abdominal pain
1,751	Z23	Encounter for immunization
1,747	J02.9	Acute pharyngitis, unspecified
1,745	Z20.822	Contact with and (suspect exposure to covid-19)
1,548	R51	Headache
1,326	R09.81	Nasal congestion
1,305	Z32.02	Encounter for pregnancy test, result negative
1,257	B97.89	Other viral agents as the cause of diseases classified elsewhere
1,235	E66.9	Obesity, unspecified
1,220	Z53.21	Procedure and treatment not carried out due to patient leaving prior to being seen by health care provider
1,204	F90.9	Attention-deficit hyperactivity disorder, unspecified type
1,155	R19.7	Diarrhea, unspecified
1,086	L30.9	Dermatitis, unspecified

Table 3. Top 20 Current Procedural Terminology (CPT) and Healthcare Common Procedure Coding System (HCPCS) Codes associated with cohort

Number	CPT or HCPCS code	Description
14,078	99213	Established patient office visit, 20-29 minutes
13,444	36415	Routine venipuncture
11,569	85025	Comprehensive blood count (CBC) with automated differential count
11,216	99283	Emergency department visits that require moderate complexity medical decision making
10,231	99214	Office or outpatient visit for an established patient
10,058	90471	The administration of a single vaccine or combination vaccine by injection
9,946	99212	An evaluation and management (E/M) visit for an established patient
9,572	94760	A single noninvasive pulse oximetry measurement
8,790	99284	An emergency department (ED) visit for a patient with a high-severity problem
8,176	99211	An office or outpatient visit for an established patient that may not require a physician
7,991	84443	Laboratory test that measures thyroid-stimulating hormone (TSH) levels
7,831	85027	Complete blood count (CBC) that doesn't include a white blood cell differential
7,695	90472	Additional vaccine given after the first vaccine
7,534	80061	Lipid panel
7,418	99282	Emergency department visit for a new or established patient.
7,405	1445253	Closed Reduction, Splinting Right Thumb
7,354	81001	Urinalysis performed with a dipstick or tablet reagent and microscopy
7,182	506675	Hypertrophy of Tonsils and Adenoids
7,150	1435700	Spirometry Pre & Post Bronchodilators
7,123	1481540	Tonsillectomy and Adenoidectomy

Table 4. Top 20 Medications associated with cohort

Number	Medicine Name
8,692	Ibuprofen 100 mg/5mL oral suspension
8,678	Normal saline flush for medications
8,598	Acetaminophen 160 mg/5mL oral suspension
7,962	Intravenous fluids
7,094	Sodium chloride 0.9% intravenous / injectable solution
6,601	Influenza virus vaccine split intramuscular suspension
6,502	Sodium chloride 0.9% injectable solution
6,444	Acetaminophen 325 mg oral tabs
6,303	Ibuprofen 200 mg oral tabs
6,049	Ondansetron HCL 4 mg/2mL injectable solution
5,526	Hepatitis A vaccine 720 el u/0.5mL intramuscular suspension
5,303	Morphine sulfate (preservative-free) 1 mg/mL injectable solution
5,273	Sodium chloride 0.9 % intravenous bolus infusion
5,158	Albuterol sulfate HFA 108 (90 base) mcg/act inhaled Aerosol
4,935	Albuterol sulfate (2.5 mg/3mL) 0.083% inhaled nebulizer
4,892	Amoxicillin 400 mg/5mL oral suspension
4,891	Lactated ringers intravenous solution
4,694	Fentanyl citrate 0.05 mg/mL injectable solution
4,569	Dexamethasone sodium phosphate 4 mg/mL injectable solution
4,545	Acetaminophen 10 mg/mL intravenous solution

References

1. Bailey ZD, Krieger N, Agenor M, Graves J, Linos N, Bassett MT. Structural racism and health inequities in the USA: evidence and interventions. Lancet. 2017;389(10077):1453-63. doi: 10.1016/S0140-6736(17)30569-X. PubMed PMID: 28402827.

2. Popejoy AB, Ritter DI, Crooks K, Currey E, Fullerton SM, Hindorff LA, Koenig B, Ramos EM, Sorokin EP, Wand H, Wright MW, Zou J, Gignoux CR, Bonham VL, Plon SE, Bustamante CD, Clinical Genome Resource A, Diversity Working G. The clinical imperative for inclusivity: Race, ethnicity, and ancestry (REA) in genomics. Hum Mutat. 2018;39(11):1713-20. doi: 10.1002/humu.23644. PubMed PMID: 30311373; PMCID: PMC6188707.

3. Unaka N, Kahn RS, Spitznagel T, Henize AW, Carlson D, Michael J, Quinonez E, Anderson J, Beck AF, Cincinnati Children's Health Equity Network Study G. An Institutional Approach to Equity and Improvement in Child Health Outcomes. Pediatrics. 2024;154(2). doi: 10.1542/peds.2023-064994. PubMed PMID: 38953125; PMCID: PMC11464011.

4. Landry LG, Ali N, Williams DR, Rehm HL, Bonham VL. Lack Of Diversity In Genomic Databases Is A Barrier To Translating Precision Medicine Research Into Practice. Health Aff (Millwood). 2018;37(5):780-5. doi: 10.1377/hlthaff.2017.1595. PubMed PMID: 29733732.

5. Duncan L, Shen H, Gelaye B, Meijsen J, Ressler K, Feldman M, Peterson R, Domingue B. Analysis of polygenic risk score usage and performance in diverse human populations. Nat Commun. 2019;10(1):3328. Epub 20190725. doi: 10.1038/s41467-019-11112-0. PubMed PMID: 31346163; PMCID: PMC6658471.

6. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019;51(4):584-91. Epub 20190329. doi: 10.1038/s41588-019-0379-x. PubMed PMID: 30926966; PMCID: PMC6563838.

7. Marquez-Luna C, Loh PR, South Asian Type 2 Diabetes C, Consortium STD, Price AL. Multiethnic polygenic risk scores improve risk prediction in diverse populations. Genet Epidemiol. 2017;41(8):811-23. Epub 2017/11/08. doi: 10.1002/gepi.22083. PubMed PMID: 29110330; PMCID: PMC5726434.

8. Grinde KE, Qi Q, Thornton TA, Liu S, Shadyab AH, Chan KHK, Reiner AP, Sofer T. Generalizing polygenic risk scores from Europeans to Hispanics/Latinos. Genet Epidemiol. 2019;43(1):50-62. Epub 20181015. doi: 10.1002/gepi.22166. PubMed PMID: 30368908; PMCID: PMC6330129.

9. Bentley AR, Callier SL, Rotimi CN. Evaluating the promise of inclusion of African ancestry populations in genomics. NPJ Genom Med. 2020;5:5. Epub 20200225. doi: 10.1038/s41525-019-0111-x. PubMed PMID: 32140257; PMCID: PMC7042246.

10. Li JH, Findley K, Pickrell JK, Blease K, Zhao J, Kruglyak S. Low-pass sequencing plus imputation using avidity sequencing displays comparable imputation accuracy to sequencing by synthesis while reducing duplicates. G3 (Bethesda). 2024;14(2). doi: 10.1093/g3journal/jkad276. PubMed PMID: 38038370; PMCID: PMC10849336.

11. Li JH, Mazur CA, Berisa T, Pickrell JK. Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays. Genome Res. 2021;31(4):529-37. Epub 20210203. doi: 10.1101/gr.266486.120. PubMed PMID: 33536225; PMCID: PMC8015847.

12. Wasik K, Berisa T, Pickrell JK, Li JH, Fraser DJ, King K, Cox C. Comparing low-pass sequencing and genotyping for trait mapping in pharmacogenetics. BMC Genomics. 2021;22(1):197. Epub 20210320. doi: 10.1186/s12864-021-07508-2. PubMed PMID: 33743587; PMCID: PMC7981957.

13. Kato GJ, Piel FB, Reid CD, Gaston MH, Ohene-Frempong K, Krishnamurti L, Smith WR, Panepinto JA, Weatherall DJ, Costa FF, Vichinsky EP. Sickle cell disease. Nat Rev Dis Primers. 2018;4:18010. Epub 20180315. doi: 10.1038/nrdp.2018.10. PubMed PMID: 29542687.

14. Bender MA, Carlberg K. Sickle Cell Disease. In: Adam MP, Feldman J, Mirzaa GM, Pagon RA, Wallace SE, Amemiya A, editors. GeneReviews((R)). Seattle (WA)1993.

15. Inusa BPD, Hsu LL, Kohli N, Patel A, Ominu-Evbota K, Anie KA, Atoyebi W. Sickle Cell Disease-Genetics, Pathophysiology, Clinical Presentation and Treatment. Int J Neonatal Screen. 2019;5(2):20. Epub 20190507. doi: 10.3390/ijns5020020. PubMed PMID: 33072979; PMCID: PMC7510211.

16. Manwani D, Frenette PS. Vaso-occlusion in sickle cell disease: pathophysiology and novel targeted therapies. Blood. 2013;122(24):3892-8. Epub 20130919. doi: 10.1182/blood-2013-05-498311. PubMed PMID: 24052549; PMCID: PMC3854110.

17. Warner EN, Ammerman RT, Glauser TA, Pestian JP, Agasthya G, Strawn JR. Developmental Epidemiology of Pediatric Anxiety Disorders. Child Adolesc Psychiatr Clin N Am. 2023;32(3):511-30. Epub 20230321. doi: 10.1016/j.chc.2023.02.001. PubMed PMID: 37201964.

18. Alegria M, Vallas M, Pumariega AJ. Racial and ethnic disparities in pediatric mental health. Child Adolesc Psychiatr Clin N Am. 2010;19(4):759-74. doi: 10.1016/j.chc.2010.07.001. PubMed PMID: 21056345; PMCID: PMC3011932.

19. Mahmood A, Kedia S, Arshad H, Mou X, Dillon PJ. Disparities in Access to Mental Health Services Among Children Diagnosed with Anxiety and Depression in the United States. Community Ment Health J. 2024;60(8):1532-46. Epub 20240622. doi: 10.1007/s10597-024-01305-3. PubMed PMID: 38907843; PMCID: PMC11579094.

20. Binney S, Flanders WD, Sircar K, Idubor O. Trends in US Pediatric Asthma Hospitalizations, by Race and Ethnicity, 2012-2020. Prev Chronic Dis. 2024;21:E71. Epub 20240919. doi: 10.5888/pcd21.240049. PubMed PMID: 39298796; PMCID: PMC11451570.

21. Beck AF, Seid M, McDowell KM, Udoko M, Cronin SC, Makrozahopoulos D, Powers T, Fairbanks S, Prideaux J, Vaughn LM, Hente E, Thurmond S, Unaka NI. Building a regional pediatric asthma learning health system in support of optimal, equitable outcomes. Learn Health Syst. 2024;8(2):e10403. Epub 20231211. doi: 10.1002/lrh2.10403. PubMed PMID: 38633017; PMCID: PMC11019385.

22. Forno E, Celedon JC. Asthma and ethnic minorities: socioeconomic status and beyond. Curr Opin Allergy Clin Immunol. 2009;9(2):154-60. doi: 10.1097/aci.0b013e3283292207. PubMed PMID: 19326508; PMCID: PMC3920741.

23. Beck AF, Huang B, Auger KA, Ryan PH, Chen C, Kahn RS. Explaining Racial Disparities in Child Asthma Readmission Using a Causal Inference Approach. JAMA Pediatr. 2016;170(7):695-703. Epub 2016/05/18. doi: 10.1001/jamapediatrics.2016.0269. PubMed PMID: 27182793; PMCID: PMC5503118.

24. Clay SM, Schoettler N, Goldstein AM, Carbonetto P, Dapas M, Altman MC, Rosasco MG, Gern JE, Jackson DJ, Im HK, Stephens M, Nicolae DL, Ober C. Fine-mapping studies distinguish genetic risks for childhood- and adult-onset asthma in the HLA region. Genome Med. 2022;14(1):55. Epub 20220524. doi: 10.1186/s13073-022-01058-2. PubMed PMID: 35606880; PMCID: PMC9128203.

25. James B, Milstien S, Spiegel S. ORMDL3 and allergic asthma: From physiology to pathology. The Journal of allergy and clinical immunology. 2019;144(3):634-40. Epub 20190731. doi: 10.1016/j.jaci.2019.07.023. PubMed PMID: 31376405; PMCID: PMC6910079.

26. Torgerson DG, Ampleford EJ, Chiu GY, Gauderman WJ, Gignoux CR, Graves PE, Himes BE, Levin AM, Mathias RA, Hancock DB, Baurley JW, Eng C, Stern DA, Celedon JC, Rafaels N, Capurso D, Conti DV, Roth LA, Soto-Quiros M, Togias A, Li X, Myers RA, Romieu I, Van Den Berg DJ, Hu D, Hansel NN, Hernandez RD, Israel E, Salam MT, Galanter J, Avila PC, Avila L, Rodriquez-Santana JR, Chapela R, Rodriguez-Cintron W, Diette GB, Adkinson NF, Abel RA, Ross KD, Shi M, Faruque MU, Dunston GM, Watson HR, Mantese VJ, Ezurum SC, Liang L, Ruczinski I, Ford JG, Huntsman S, Chung KF, Vora H, Li X, Calhoun WJ, Castro M, Sienra-Monge JJ, del Rio-Navarro B, Deichmann KA, Heinzmann A, Wenzel SE, Busse WW, Gern JE, Lemanske RF, Jr., Beaty TH, Bleecker ER, Raby BA, Meyers DA, London SJ, Mexico City Childhood Asthma S, Gilliland FD, Children's Health S, study H, Burchard EG, Genetics of Asthma in Latino Americans Study SoG-E, Admixture in Latino A, Study of African Americans AG, Environments, Martinez FD, Childhood Asthma R, Education N, Weiss ST, Childhood Asthma Management P, Williams LK, Study of Asthma P, Pharmacogenomic Interactions by R-E, Barnes KC, Genetic Research on Asthma in African Diaspora S, Ober C, Nicolae DL. Meta-analysis of genome-wide association

studies of asthma in ethnically diverse North American populations. Nat Genet. 2011;43(9):887-92. Epub 20110731. doi: 10.1038/ng.888. PubMed PMID: 21804549; PMCID: PMC3445408.

27. Yan Q, Brehm J, Pino-Yanes M, Forno E, Lin J, Oh SS, Acosta-Perez E, Laurie CC, Cloutier MM, Raby BA, Stilp AM, Sofer T, Hu D, Huntsman S, Eng CS, Conomos MP, Rastogi D, Rice K, Canino G, Chen W, Barr RG, Burchard EG, Celedon JC. A meta-analysis of genome-wide association studies of asthma in Puerto Ricans. Eur Respir J. 2017;49(5). Epub 20170501. doi: 10.1183/13993003.01505-2016. PubMed PMID: 28461288; PMCID: PMC5527708.

28. Gordon ED, Palandra J, Wesolowska-Andersen A, Ringel L, Rios CL, Lachowicz-Scroggins ME, Sharp LZ, Everman JL, MacLeod HJ, Lee JW, Mason RJ, Matthay MA, Sheldon RT, Peters MC, Nocka KH, Fahy JV, Seibold MA. IL1RL1 asthma risk variants regulate airway type 2 inflammation. JCI Insight. 2016;1(14):e87871. Epub 20160908. doi: 10.1172/jci.insight.87871. PubMed PMID: 27699235; PMCID: PMC5033813.

29. Mersha TB, Qin K, Beck AF, Ding L, Huang B, Kahn RS. Genetic ancestry differences in pediatric asthma readmission are mediated by socioenvironmental factors. J Allergy Clin Immunol. 2021;148(5):1210-8 e4. Epub 2021/07/05. doi: 10.1016/j.jaci.2021.05.046. PubMed PMID: 34217757; PMCID: PMC8578303.

30. Mandl KD, Glauser T, Krantz ID, Avillach P, Bartels A, Beggs AH, Biswas S, Bourgeois FT, Corsmo J, Dauber A, Devkota B, Fleisher GR, Heath AP, Helbig I, Hirschhorn JN, Kilbourn J, Kong SW, Kornetsky S, Majzoub JA, Marsolo K, Martin LJ, Nix J, Schwarzhoff A, Stedman J, Strauss A, Sund KL, Taylor DM, White PS, Marsh E, Grimberg A, Hawkes C, Genomics R, Innovation N. The Genomics Research and Innovation Network: creating an interoperable, federated, genomics learning system. Genet Med. 2020;22(2):371-80. Epub 20190904. doi: 10.1038/s41436-019-0646-3. PubMed PMID: 31481752; PMCID: PMC7000325.

31. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26(22):2867-73. Epub 20101005. doi: 10.1093/bioinformatics/btq559. PubMed PMID: 20926424; PMCID: PMC3025716.

32. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19(9):1655-64. Epub 20090731. doi: 10.1101/gr.094052.109. PubMed PMID: 19648217; PMCID: PMC2752134.

33. Brown DW, Myers TA, Machiela MJ. PCAmatchR: a flexible R package for optimal case-control matching using weighted principal components. Bioinformatics. 2021;37(8):1178-81. doi: 10.1093/bioinformatics/btaa784. PubMed PMID: 32926120; PMCID: PMC8599751.

34. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559-

75. Epub 20070725. doi: 10.1086/519795. PubMed PMID: 17701901; PMCID: PMC1950838.

35. Choudhury A, Aron S, Sengupta D, Hazelhurst S, Ramsay M. African genetic diversity provides novel insights into evolutionary history and local adaptations. Hum Mol Genet. 2018;27(R2):R209-R18. doi: 10.1093/hmg/ddy161. PubMed PMID: 29741686; PMCID: PMC6061870.

36. Sibomana O. Genetic Diversity Landscape in African Population: A Review of Implications for Personalized and Precision Medicine. Pharmgenomics Pers Med. 2024;17:487-96. Epub 20241111. doi: 10.2147/PGPM.S485452. PubMed PMID: 39555236; PMCID: PMC11566596.

37. Milton JN, Rooks H, Drasar E, McCabe EL, Baldwin CT, Melista E, Gordeuk VR, Nouraie M, Kato GR, Minniti C, Taylor J, Campbell A, Luchtman-Jones L, Rana S, Castro O, Zhang Y, Thein SL, Sebastiani P, Gladwin MT, Walk PI, Steinberg MH. Genetic determinants of haemolysis in sickle cell anaemia. Br J Haematol. 2013;161(2):270-8. Epub 20130214. doi: 10.1111/bjh.12245. PubMed PMID: 23406172; PMCID: PMC4129543.

38. Alshabeeb MA, Alwadaani D, Al Qahtani FH, Abohelaika S, Alzahrani M, Al Zayed A, Al Saeed HH, Al Ajmi H, Alsomaie B, Rashid M, Daly AK. Impact of Genetic Variations on Thromboembolic Risk in Saudis with Sickle Cell Disease. Genes (Basel). 2023;14(10). Epub 20231009. doi: 10.3390/genes14101919. PubMed PMID: 37895268; PMCID: PMC10606407.

39. Sakaue S, Kanai M, Tanigawa Y, Karjalainen J, Kurki M, Koshiba S, Narita A, Konuma T, Yamamoto K, Akiyama M, Ishigaki K, Suzuki A, Suzuki K, Obara W, Yamaji K, Takahashi K, Asai S, Takahashi Y, Suzuki T, Shinozaki N, Yamaguchi H, Minami S, Murayama S, Yoshimori K, Nagayama S, Obata D, Higashiyama M, Masumoto A, Koretsune Y, FinnGen, Ito K, Terao C, Yamauchi T, Komuro I, Kadowaki T, Tamiya G, Yamamoto M, Nakamura Y, Kubo M, Murakami Y, Yamamoto K, Kamatani Y, Palotie A, Rivas MA, Daly MJ, Matsuda K, Okada Y. A cross-population atlas of genetic associations for 220 human phenotypes. Nat Genet. 2021;53(10):1415-24. Epub 20210930. doi: 10.1038/s41588-021-00931-x. PubMed PMID: 34594039.

40. Zhu Z, Guo Y, Shi H, Liu CL, Panganiban RA, Chung W, O'Connor LJ, Himes BE, Gazal S, Hasegawa K, Camargo CA, Jr., Qi L, Moffatt MF, Hu FB, Lu Q, Cookson WOC, Liang L. Shared genetic and experimental links between obesity-related traits and asthma subtypes in UK Biobank. J Allergy Clin Immunol. 2020;145(2):537-49. Epub 20191024. doi: 10.1016/j.jaci.2019.09.035. PubMed PMID: 31669095; PMCID: PMC7010560.

41. Demenais F, Margaritte-Jeannin P, Barnes KC, Cookson WOC, Altmuller J, Ang W, Barr RG, Beaty TH, Becker AB, Beilby J, Bisgaard H, Bjornsdottir US, Bleecker E, Bonnelykke K, Boomsma DI, Bouzigon E, Brightling CE, Brossard M, Brusselle GG, Burchard E, Burkart KM, Bush A, Chan-Yeung M, Chung KF, Couto Alves A, Curtin JA,

Custovic A, Daley D, de Jongste JC, Del-Rio-Navarro BE, Donohue KM, Duijts L, Eng C, Eriksson JG, Farrall M, Fedorova Y, Feenstra B, Ferreira MA, Australian Asthma Genetics Consortium c, Freidin MB, Gajdos Z, Gauderman J, Gehring U, Geller F, Genuneit J, Gharib SA, Gilliland F, Granell R, Graves PE, Gudbjartsson DF, Haahtela T, Heckbert SR, Heederik D, Heinrich J, Heliovaara M, Henderson J, Himes BE, Hirose H, Hirschhorn JN, Hofman A, Holt P, Hottenga J, Hudson TJ, Hui J, Imboden M, Ivanov V, Jaddoe VWV, James A, Janson C, Jarvelin MR, Jarvis D, Jones G, Jonsdottir I, Jousilahti P, Kabesch M, Kahonen M, Kantor DB, Karunas AS, Khusnutdinova E, Koppelman GH, Kozyrskyj AL, Kreiner E, Kubo M, Kumar R, Kumar A, Kuokkanen M, Lahousse L, Laitinen T, Laprise C, Lathrop M, Lau S, Lee YA, Lehtimaki T, Letort S, Levin AM, Li G, Liang L, Loehr LR, London SJ, Loth DW, Manichaikul A, Marenholz I, Martinez FJ, Matheson MC, Mathias RA, Matsumoto K, Mbarek H, McArdle WL, Melbye M, Melen E, Meyers D, Michel S, Mohamdi H, Musk AW, Myers RA, Nieuwenhuis MAE, Noguchi E, O'Connor GT, Ogorodova LM, Palmer CD, Palotie A, Park JE, Pennell CE, Pershagen G, Polonikov A, Postma DS, Probst-Hensch N, Puzyrev VP, Raby BA, Raitakari OT, Ramasamy A, Rich SS, Robertson CF, Romieu I, Salam MT, Salomaa V, Schlunssen V, Scott R, Selivanova PA, Sigsgaard T, Simpson A, Siroux V, Smith LJ, Solodilova M, Standl M, Stefansson K, Strachan DP, Stricker BH, Takahashi A, Thompson PJ, Thorleifsson G, Thorsteinsdottir U, Tiesler CMT, Torgerson DG, Tsunoda T, Uitterlinden AG, van der Valk RJP, Vaysse A, Vedantam S, von Berg A, von Mutius E, Vonk JM, Waage J, Wareham NJ, Weiss ST, White WB, Wickman M, Widen E, Willemsen G, Williams LK, Wouters IM, Yang JJ, Zhao JH, Moffatt MF, Ober C, Nicolae DL. Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. Nat Genet. 2018;50(1):42-53. Epub 2017/12/24. doi: 10.1038/s41588-017-0014-7. PubMed PMID: 29273806; PMCID: PMC5901974.

42. Pickrell JK, Berisa T, Liu JZ, Segurel L, Tung JY, Hinds DA. Detection and interpretation of shared genetic influences on 42 human traits. Nat Genet. 2016;48(7):709-17. Epub 20160516. doi: 10.1038/ng.3570. PubMed PMID: 27182965; PMCID: PMC5207801.

43. Ferreira MAR, Mathur R, Vonk JM, Szwajda A, Brumpton B, Granell R, Brew BK, Ullemar V, Lu Y, Jiang Y, andMe Research T, e QC, Consortium B, Magnusson PKE, Karlsson R, Hinds DA, Paternoster L, Koppelman GH, Almqvist C. Genetic Architectures of Childhood- and Adult-Onset Asthma Are Partly Distinct. Am J Hum Genet. 2019;104(4):665-84. Epub 20190328. doi: 10.1016/j.ajhg.2019.02.022. PubMed PMID: 30929738; PMCID: PMC6451732.

44. Han Y, Jia Q, Jahani PS, Hurrell BP, Pan C, Huang P, Gukasyan J, Woodward NC, Eskin E, Gilliland FD, Akbari O, Hartiala JA, Allayee H. Genome-wide analysis highlights contribution of immune system pathways to the genetic architecture of asthma. Nat Commun. 2020;11(1):1776. Epub 20200415. doi: 10.1038/s41467-020-15649-3. PubMed PMID: 32296059; PMCID: PMC7160128.

45. Johansson A, Rask-Andersen M, Karlsson T, Ek WE. Genome-wide association analysis of 350 000 Caucasians from the UK Biobank identifies novel loci for asthma,

hay fever and eczema. Hum Mol Genet. 2019;28(23):4022-41. doi: 10.1093/hmg/ddz175. PubMed PMID: 31361310; PMCID: PMC6969355.

46. Olafsdottir TA, Theodors F, Bjarnadottir K, Bjornsdottir US, Agustsdottir AB, Stefansson OA, Ivarsdottir EV, Sigurdsson JK, Benonisdottir S, Eyjolfsson GI, Gislason D, Gislason T, Guethmundsdottir S, Gylfason A, Halldorsson BV, Halldorsson GH, Juliusdottir T, Kristinsdottir AM, Ludviksdottir D, Ludviksson BR, Masson G, Norland K, Onundarson PT, Olafsson I, Sigurdardottir O, Stefansdottir L, Sveinbjornsson G, Tragante V, Gudbjartsson DF, Thorleifsson G, Sulem P, Thorsteinsdottir U, Norddahl GL, Jonsdottir I, Stefansson K. Eighty-eight variants highlight the role of T cell regulation and airway remodeling in asthma pathogenesis. Nat Commun. 2020;11(1):393. Epub 20200120. doi: 10.1038/s41467-019-14144-8. PubMed PMID: 31959851; PMCID: PMC6971247.

47. Pividori M, Schoettler N, Nicolae DL, Ober C, Im HK. Shared and distinct genetic risk factors for childhood-onset and adult-onset asthma: genome-wide and transcriptome-wide studies. Lancet Respir Med. 2019;7(6):509-22. Epub 20190427. doi: 10.1016/S2213-2600(19)30055-4. PubMed PMID: 31036433; PMCID: PMC6534440.

48. Valette K, Li Z, Bon-Baret V, Chignon A, Berube JC, Eslami A, Lamothe J, Gaudreault N, Joubert P, Obeidat M, van den Berge M, Timens W, Sin DD, Nickle DC, Hao K, Labbe C, Godbout K, Cote A, Laviolette M, Boulet LP, Mathieu P, Theriault S, Bosse Y. Prioritization of candidate causal genes for asthma in susceptibility loci derived from UK Biobank. Commun Biol. 2021;4(1):700. Epub 20210608. doi: 10.1038/s42003-021-02227-6. PubMed PMID: 34103634; PMCID: PMC8187656.

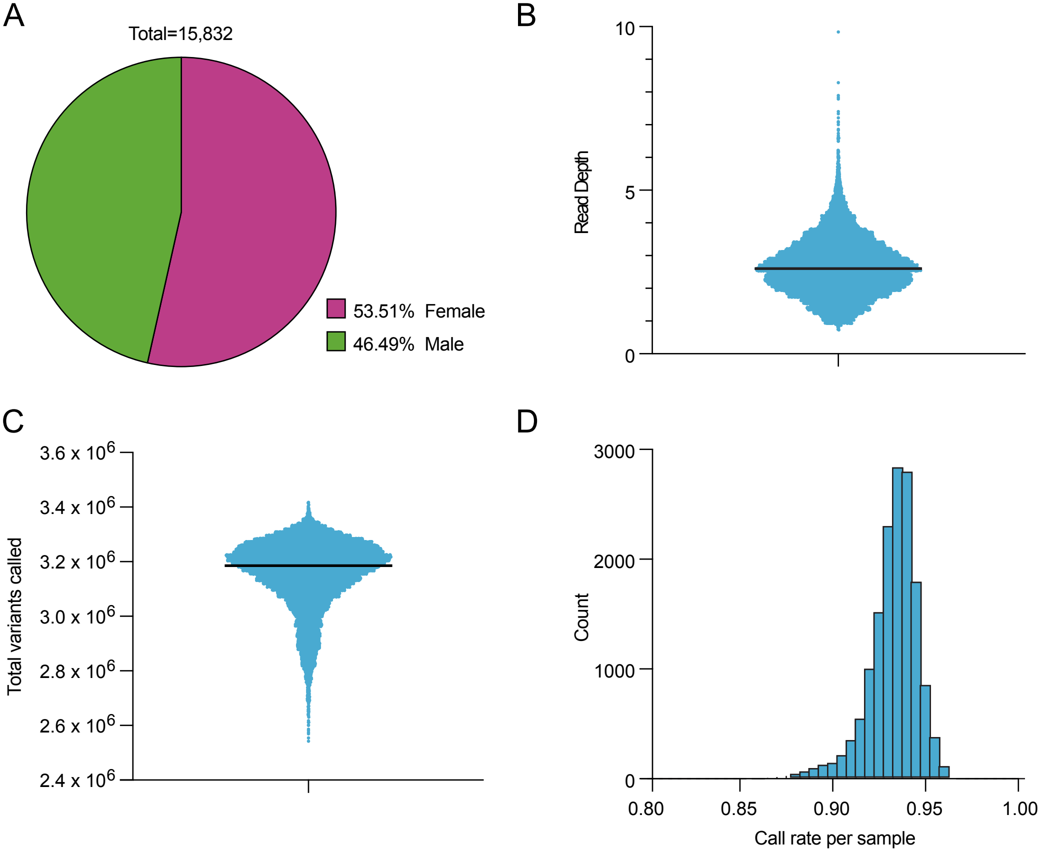
49. Zhou W, Kanai M, Wu KH, Rasheed H, Tsuo K, Hirbo JB, Wang Y, Bhattacharya A, Zhao H, Namba S, Surakka I, Wolford BN, Lo Faro V, Lopera-Maya EA, Lall K, Fave MJ, Partanen JJ, Chapman SB, Karjalainen J, Kurki M, Maasha M, Brumpton BM, Chavan S, Chen TT, Daya M, Ding Y, Feng YA, Guare LA, Gignoux CR, Graham SE, Hornsby WE, Ingold N, Ismail SI, Johnson R, Laisk T, Lin K, Lv J, Millwood IY, Moreno-Grau S, Nam K, Palta P, Pandit A, Preuss MH, Saad C, Setia-Verma S, Thorsteinsdottir U, Uzunovic J, Verma A, Zawistowski M, Zhong X, Afifi N, Al-Dabhani KM, Al Thani A, Bradford Y, Campbell A, Crooks K, de Bock GH, Damrauer SM, Douville NJ, Finer S, Fritsche LG, Fthenou E, Gonzalez-Arroyo G, Griffiths CJ, Guo Y, Hunt KA, Ioannidis A, Jansonius NM, Konuma T, Lee MTM, Lopez-Pineda A, Matsuda Y, Marioni RE, Moatamed B, Nava-Aguilar MA, Numakura K, Patil S, Rafaels N, Richmond A, Rojas-Munoz A, Shortt JA, Straub P, Tao R, Vanderwerff B, Vernekar M, Veturi Y, Barnes KC, Boezen M, Chen Z, Chen CY, Cho J, Smith GD, Finucane HK, Franke L, Gamazon ER, Ganna A, Gaunt TR, Ge T, Huang H, Huffman J, Katsanis N, Koskela JT, Lajonchere C, Law MH, Li L, Lindgren CM, Loos RJF, MacGregor S, Matsuda K, Olsen CM, Porteous DJ, Shavit JA, Snieder H, Takano T, Trembath RC, Vonk JM, Whiteman DC, Wicks SJ, Wijmenga C, Wright J, Zheng J, Zhou X, Awadalla P, Boehnke M, Bustamante CD, Cox NJ, Fatumo S, Geschwind DH, Hayward C, Hveem K, Kenny EE, Lee S, Lin YF, Mbarek H, Magi R, Martin HC, Medland SE, Okada Y, Palotie AV, Pasaniuc B, Rader DJ, Ritchie MD, Sanna S, Smoller JW, Stefansson K, van Heel DA, Walters RG, Zollner S, Biobank of the A, Biobank Japan P, BioMe, BioVu, CanPath - Ontario Health S,

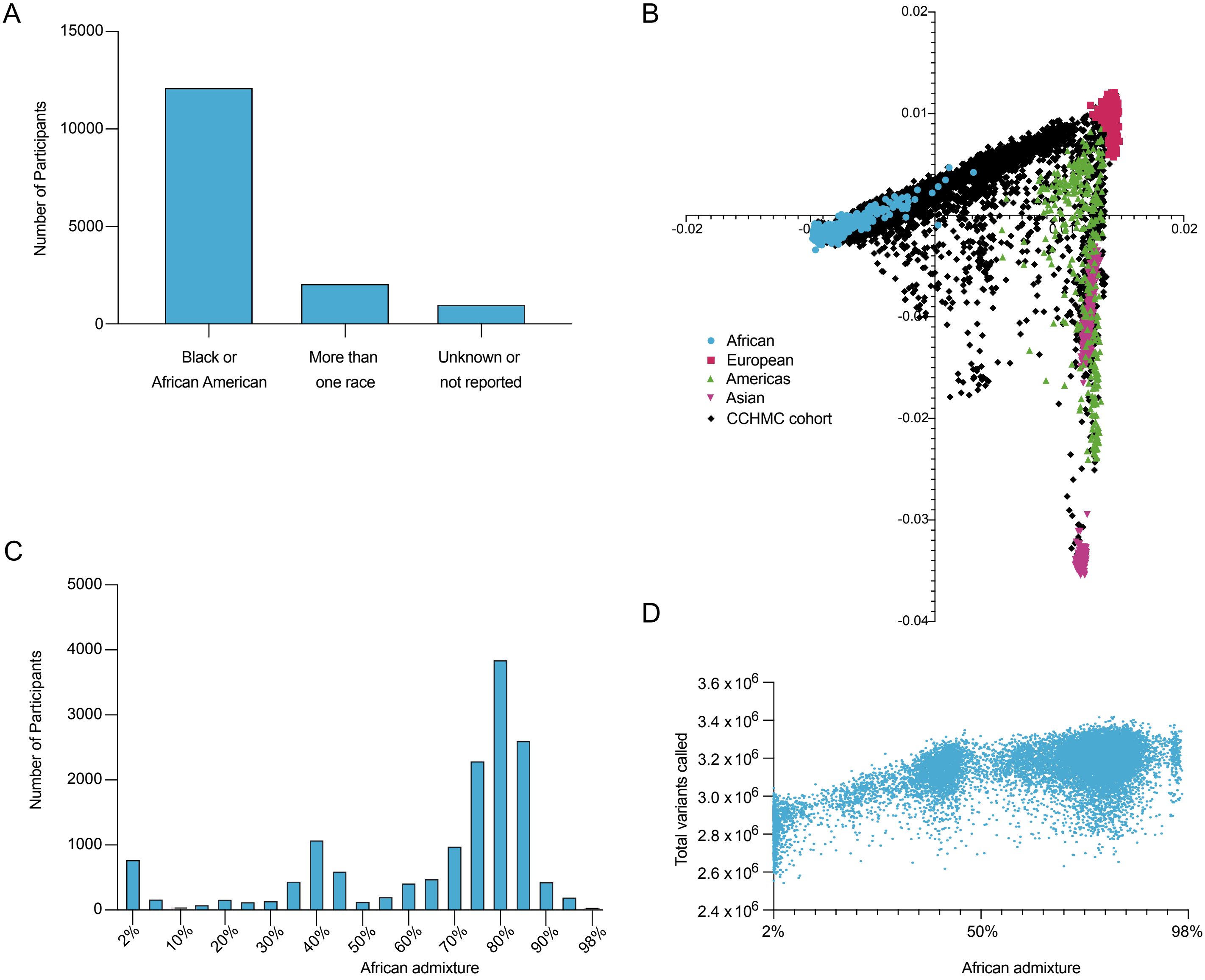
China Kadoorie Biobank Collaborative G, Colorado Center for Personalized M, de CG, Estonian B, FinnGen, Generation S, Genes, Health Research T, LifeLines, Mass General Brigham B, Michigan Genomics I, National Biobank of K, Penn Medicine B, Qatar B, Sun QS, Health S, Taiwan B, Study H, Initiative UACH, Uganda Genome R, Biobank UK, Martin AR, Willer CJ, Daly MJ, Neale BM. Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. Cell Genom. 2022;2(10):100192. Epub 20221012. doi: 10.1016/j.xgen.2022.100192. PubMed PMID: 36777996; PMCID: PMC9903716.

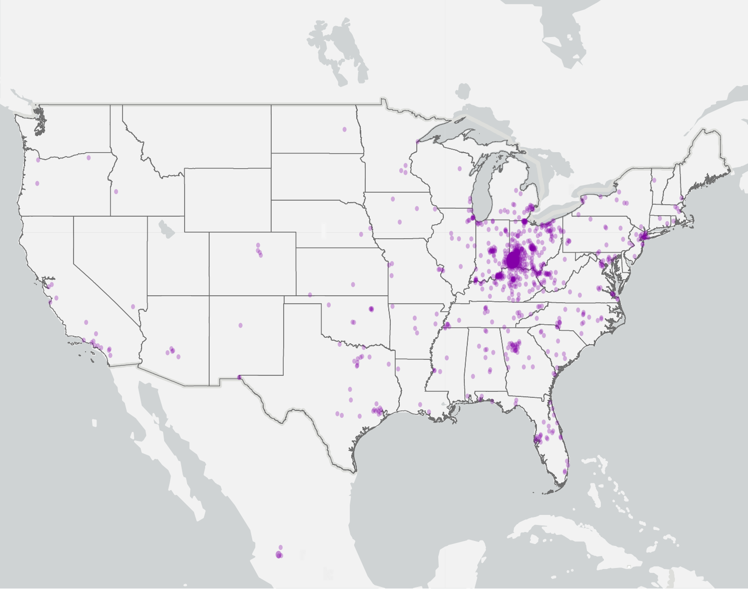
50. Zhu Z, Zhu X, Liu CL, Shi H, Shen S, Yang Y, Hasegawa K, Camargo CA, Jr., Liang L. Shared genetics of asthma and mental health disorders: a large-scale genome-wide cross-trait analysis. Eur Respir J. 2019;54(6). Epub 2019/10/18. doi: 10.1183/13993003.01507-2019. PubMed PMID: 31619474.

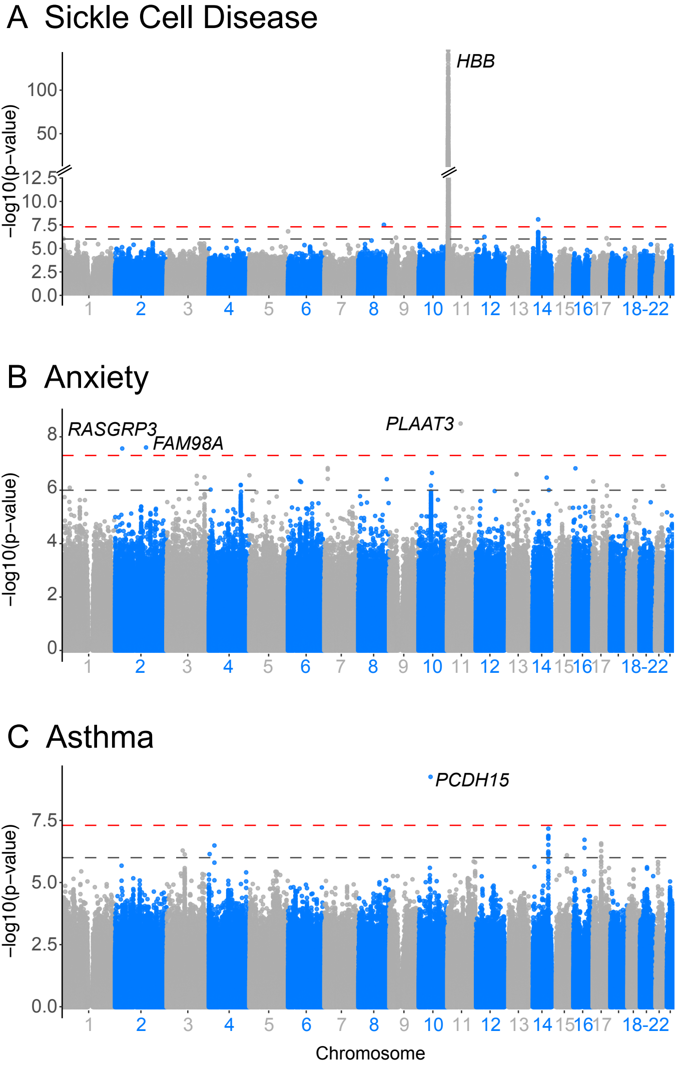
51. Namjou B, Lape M, Malolepsza E, DeVore SB, Weirauch MT, Dikilitas O, Jarvik GP, Kiryluk K, Kullo IJ, Liu C, Luo Y, Satterfield BA, Smoller JW, Walunas TL, Connolly J, Sleiman P, Mersha TB, Mentch FD, Hakonarson H, Prows CA, Biagini JM, Khurana Hershey GK, Martin LJ, Kottyan L, e MN. Multiancestral polygenic risk score for pediatric asthma. J Allergy Clin Immunol. 2022;150(5):1086-96. Epub 20220518. doi: 10.1016/j.jaci.2022.03.035. PubMed PMID: 35595084; PMCID: PMC9643615.

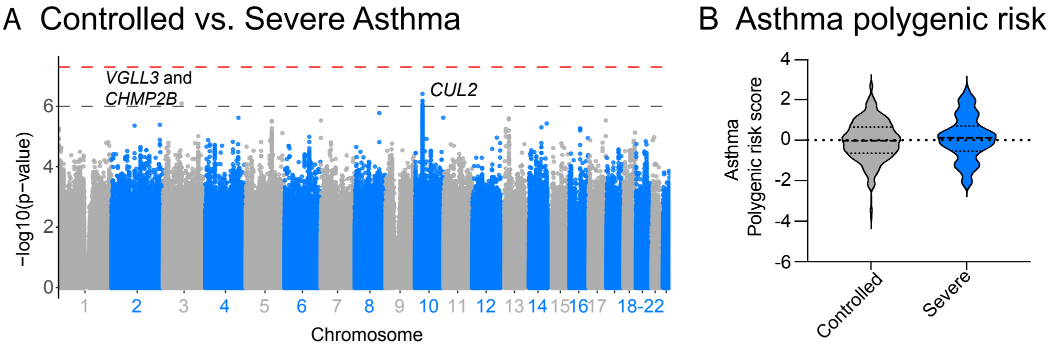
52. Medical Progress for all. Nat Med. 2025. doi: doi.org/10.1038/s41591-025-03634-6.


A
Total=15,832
53.51% Female
46.49% Male
B
Read Depth
10
5
0
C
Total variants called
3.6 x 10^6
3.4 x 10^6
3.2 x 10^6
3.0 x 10^6
2.8 x 10^6
2.6 x 10^6
2.4 x 10^6
D
Count
3000
2000
1000
0
0.80
0.85
0.90
0.95
1.00
Call rate per sample


A
15000
10000
5000
0
Number of Participants
Black or
African American
More than
one race
Unknown or
not reported
B
0.02
0.01
-0.02
-0.01
0.01
0.02
-0.01
-0.02
-0.03
-0.04
● African
■ European
▲ Americas
▼ Asian
◆ CCHMC cohort
C
5000
4000
3000
2000
1000
0
Number of Participants
2%
10%
20%
30%
40%
50%
60%
70%
80%
90%
98%
African admixture
D
3.6 x 10^6
3.4 x 10^6
3.2 x 10^6
3.0 x 10^6
2.8 x 10^6
2.6 x 10^6
2.4 x 10^6
Total variants called
2%
50%
98%
African admixture

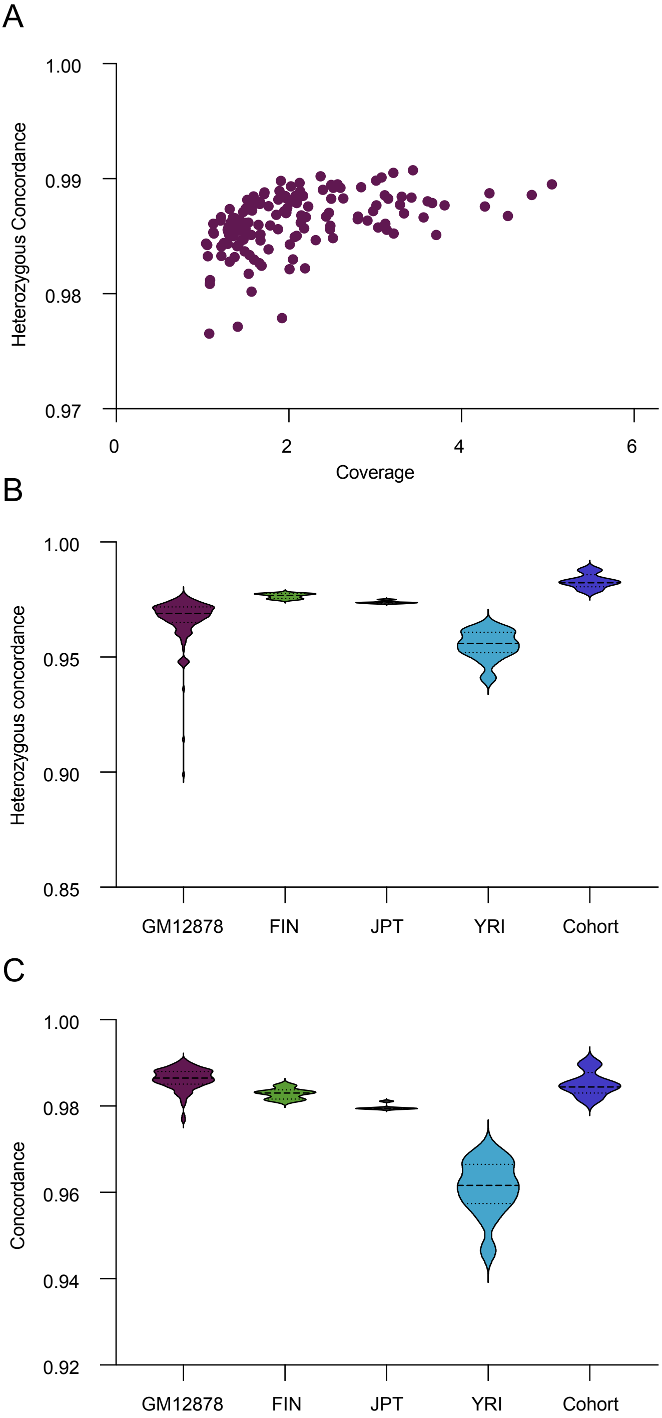


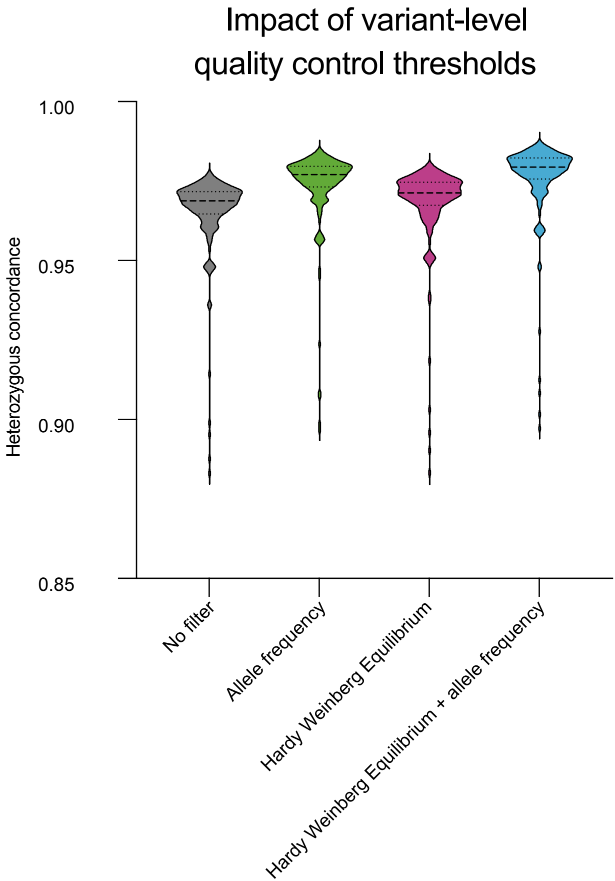

A Sickle Cell Disease
HBB
−log10(p−value)
100
50
12.5
10.0
7.5
5.0
2.5
0.0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18-22
B Anxiety
RASGRP3
FAM98A
PLAAT3
−log10(p−value)
8
6
4
2
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18-22
C Asthma
PCDH15
−log10(p−value)
7.5
5.0
2.5
0.0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18-22
Chromosome

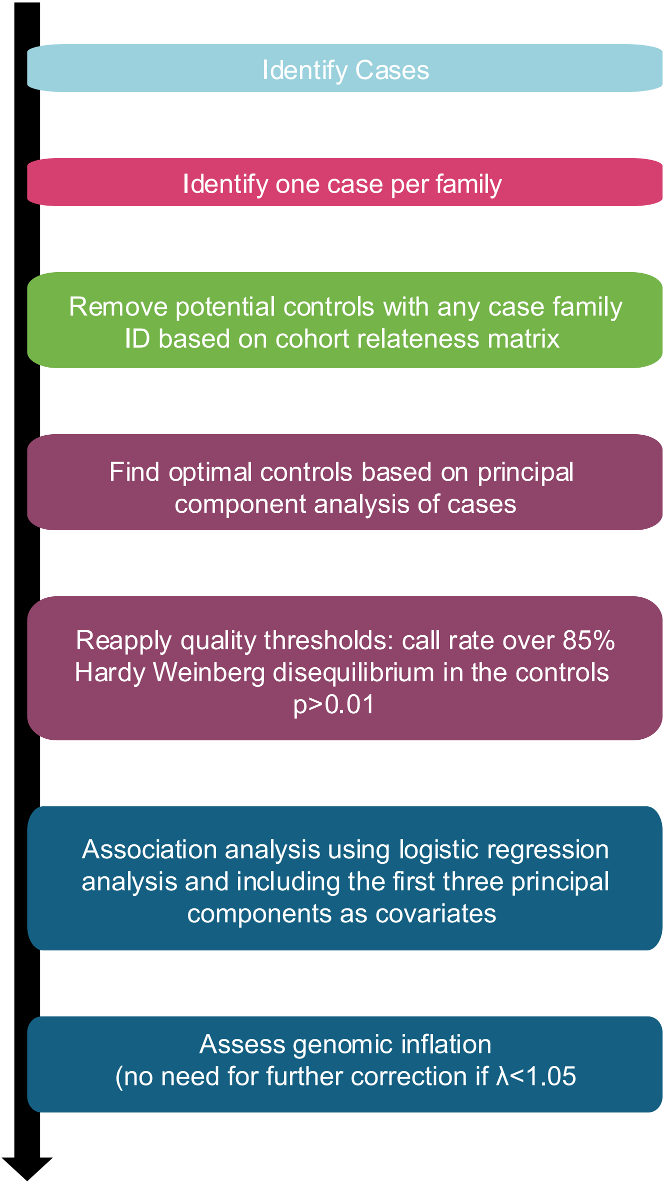


A Controlled vs. Severe Asthma

B Asthma polygenic risk


A
Heterozygous Concordance
1.00
0.99
0.98
0.97
0
2
4
6
Coverage
B
Heterozygous concordance
1.00
0.95
0.90
0.85
GM12878
FIN
JPT
YRI
Cohort
C
Concordance
1.00
0.98
0.96
0.94
0.92
GM12878
FIN
JPT
YRI
Cohort


Impact of variant-level
quality control thresholds
Heterozygous concordance
1.00
0.95
0.90
0.85
No filter
Allele frequency
Hardy Weinberg Equilibrium
Hardy Weinberg Equilibrium + allele frequency



Identify Cases

Identify one case per family

Remove potential controls with any case family ID based on cohort relateness matrix

Find optimal controls based on principal component analysis of cases

Reapply quality thresholds: call rate over 85% Hardy Weinberg disequilibrium in the controls p>0.01

Association analysis using logistic regression analysis and including the first three principal components as covariates

Assess genomic inflation
(no need for further correction if λ<1.05