Title

Genetic disease risks of under-represented founder populations in New York City

Authors

Mariko Isshiki PhD,[1] Anthony Griffen BSc,[2] Paul Meissner MSPH,[3, 4] Paulette Spencer MPH,[5] Michael D. Cabana MD,[6] Susan D. Klugman MD,[4] Mirtha Colón MSW,[7] Zoya Maksumova MD,[8] Shakira Suglia ScD,[10] Carmen Isasi MD,[6,9] John M. Greally DMed,[1,6,†] Srilakshmi M. Raj PhD[1,†]

Affiliations

[1]Departments of Genetics, [2]Cell Biology, [3]Family and Social Medicine, [4]Obstetrics and Gynecology & Women's Health, [6]Pediatrics and [9]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, NY 10461

[5]Bronx Community Health Network, One Fordham Plaza, Suite 1108, Bronx, NY 10458

[7]Hondurans Against AIDS/Casa Yurumein, 324 E 151st St, Bronx, NY 10451

[8]10310 Fuerte Drive, La Mesa, CA 91941

[10]Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA 30322

[†]co-corresponding authors

Address reprint requests to: Srilakshmi Raj srilakshmi.raj@einsteinmed.edu

Contact information for corresponding authors: Srilakshmi Raj srilakshmi.raj@einsteinmed.edu

John Greally john.greally@einsteinmed.edu

Abstract

The detection of founder pathogenic variants, those observed in high frequency only in a group of individuals with increased inter-relatedness, can help improve delivery of health care for that community. We identified 16 groups with shared ancestry, based on genomic segments that are shared through identity by descent (IBD) , in New York City using the genomic data of 25,366 residents from the All Of Us Research Program and the Mount Sinai Bio*Me* biobank. From these groups we defined 8 as founder populations, mostly communities currently under-represented in medical genomics research, such as Puerto Rican, Garifuna and Filipino/Pacific Islanders. The enrichment analysis of ClinVar pathogenic or likely pathogenic (P/LP) variants in each group identified 202 of these damaging variants across the 8 founder populations. We confirmed disease-causing variants previously reported to occur at increased frequencies in Ashkenazi Jewish and Puerto Rican genetic ancestry groups, but most of the damaging variants identified have not been previously associated with any such founder populations, and most of these founder populations have not been described to have increased prevalence of the associated rare disease. Twenty-five of 51 variants meeting Tier 2 clinical screening criteria (1/100 carrier frequency within these founder groups) have never previously been reported. We show how population structure studies can provide insights into rare diseases disproportionately affecting under-represented founder populations, delivering a health care benefit but also a potential source of stigmatization of these communities, who should be part of the decision-making about implementation into health care delivery.

Author Summary

It is well recognized that genomic studies have been biased towards individuals of European ancestry, and that obtaining medical insights for populations under-represented in medical genomics is crucial to achieve health equity. Here, we use genomic information to identify networks of individuals in New York City who are distinctively related to each other, allowing us to define populations with common genetic ancestry based on genetic similarities rather than by self-reported race or ethnicity. In our study of >25,000 New Yorkers, we identified eight highly-interrelated founder populations, with 202 likely disease-causing variants with increased frequencies in specific founder populations. Many of these population-specific variants are new discoveries, despite their high frequency in founder populations. Studying recent genetic ancestry can help reveal population-specific disease insights that can help with early diagnosis, carrier screening, and opportunities for targeted therapies that all help to reduce health disparities in genomic medicine.

Introduction

Rare diseases collectively occur in 3.5-5.9% of the population(1) They involve significant morbidity and mortality, risk to family members and socio-economic consequences, and thus have the characteristics typical of a public health priority. Responding to this public health issue by studying rare diseases on a population scale is challenging because of the difficulty identifying individuals and families with uncommon conditions that are often refractory to diagnosis. An eventual solution will involve widespread application of sequencing of patients' entire genomes in health care with sensitive and high-confidence prediction of damaging DNA sequence variants, but this remains a remote goal at present. In the interim, a typical approach in clinical practice is to use a person's 'genetic ancestry group' (2) to highlight the rare diseases that are more common in that community and could be affecting the patient presenting for care. Populations that experienced small population size in the past tend to have enrichment for otherwise rare genetic conditions due to a 'founder effect', as exemplified in Ashkenazi Jewish individuals for their well-characterized set of genetic conditions (3,4) By identifying other populations with founder effects, the genetic conditions more likely to occur in individuals from those communities can also be defined, and clinicians who serve these communities can be prepared to look out for these conditions. Extending the insights into rare disease risks for genetic ancestry groups other than White Europeans has been limited by the failure to include non-European populations in genomics research (5,6) This bias magnifies health disparities and impedes effective delivery of medical care to marginalized groups and underserved populations. Recognizing this neglect, the All of Us (AoU) Research Program in the United States has been designed to represent the country's diversity (7,8) In this study, we focused on the genomes of individuals in New York City (NYC), representing a diverse and admixed urban population studied extensively through AoU as well as the separate Bio*Me* biobank(9,10). We show how population genetics approaches using these data resources are able to reveal previously undiscovered rare disease susceptibilities in diverse

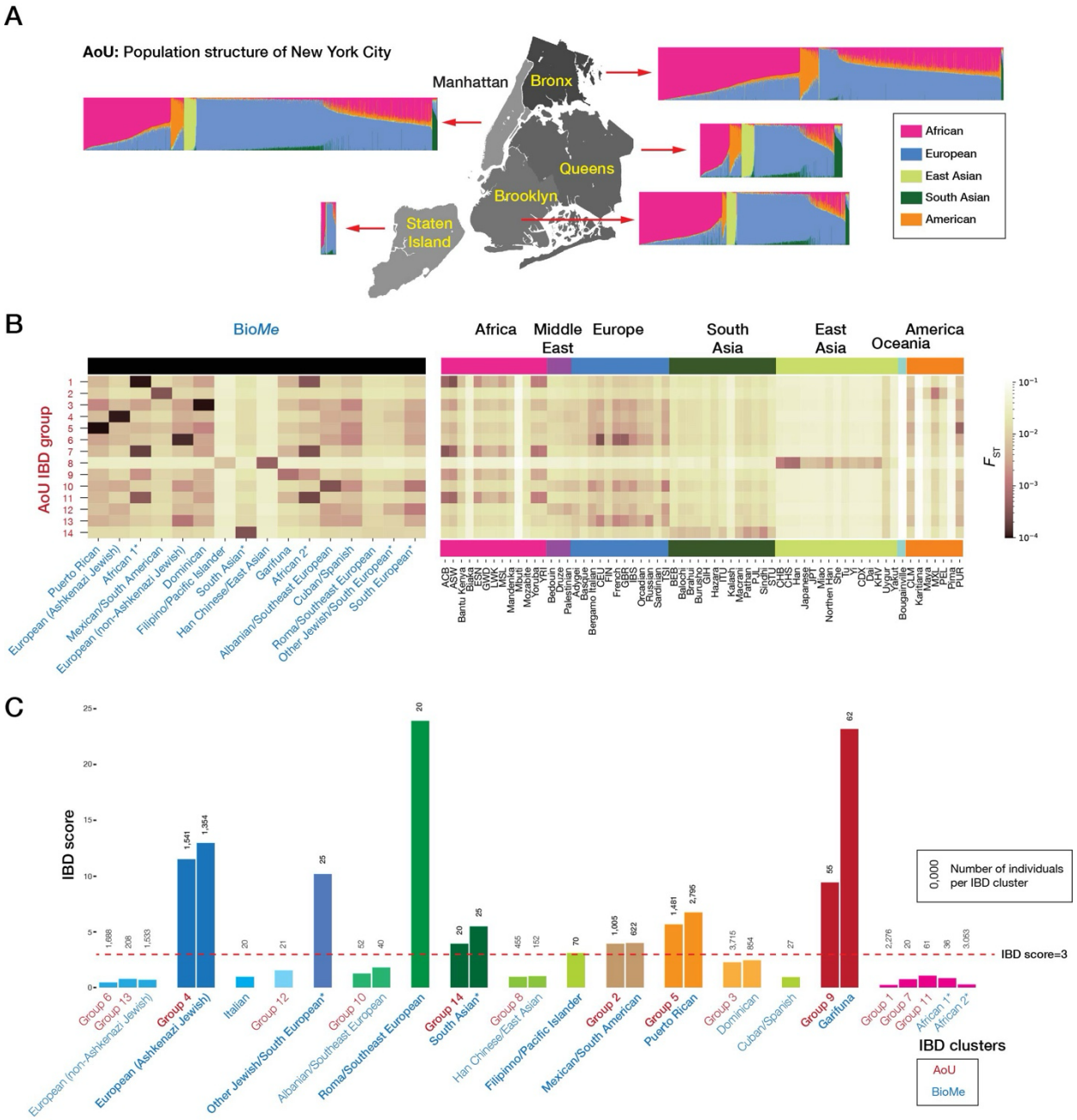
genetic ancestry groups, particularly those with a founder effect. We were able to define groups with increased 'genetic similarity'(2) and characterize population structure by identifying segments of DNA that are shared among individuals due to inheritance from a common ancestor, also called identity-by-descent (IBD). This has previously been performed successfully in cohorts of different genetic ancestries (10–14) with implications for understanding population-specific disease risk (9,10,13) Here we explicitly test the association between population structure and disease risk by focusing on population-specific enrichment of variants curated as disease-causing in the ClinVar database (15) The results show how health systems and providers can benefit from recognizing rare diseases in the populations they serve, including the potential benefits of early detection of rare diseases as well as prenatal carrier screening in these communities, and the targeted use of specific therapies.

Results

The population structure of NYC participants of the All of Us Research Program

We studied the genetic diversity of NYC residents using genetic data from 13,817 participants of the AoU Research Program. This dataset excludes ‘related’ individuals, those who are second cousins or closer. We found that the AoU cohort in NYC is diverse across the five boroughs (**Fig. 1a, Figure S1**), and that the proportions of self-reported race/ethnicity information for each borough are comparable to those from census data (**Figure S2a**) (16). Of the boroughs, Manhattan and the Bronx are over-represented (**Figure S2b**). Admixed individuals accounted for

107 a large proportion of the dataset (**Figure S1**).



108

109 **Figure 1: Population membership and geospatial distribution of All of Us IBD groups.**

110 **(A)** The geospatial_distribution of all AoU individuals across NYC boroughs; **(B)** Pairwise F_{ST}

111 comparisons_between AoU IBD groups, Bio*Me* IBD groups and global reference populations; **(C)**

112 IBD scores for AoU and BioMe IBD groups, with sample sizes above each bar. IBD group with

F_{ST} < 0.001 are indicated by the same color, bold fonts identify founder populations. The red dotted line indicates an IBD score of 3, which is the cut-off value we used to define a founder population (**Fig. 2**); Only IBD groups with 20 and more individuals are shown to protect participants' privacy based on the AoU Data and Statistics Dissemination Policy. An asterisk next to labels represents populations with inadequate reference information for annotation.

We constructed a network based on identity-by-descent (IBD) sharing to capture fine-scale recent population structure in the AoU NYC participants. After filtering edges to only reflect recent shared ancestry and exclude close familial ties, 98.6 % of the cohort was included in the network. Among these individuals, we identified 14 IBD groups with a minimum of 20 individuals each (**Figure S1b**), representing 91% of the AoU NYC cohort. To allow comparison of our AoU results with results using the independent NYC Bio*Me* biobank, we replicated the AoU IBD analysis on Bio*Me.* The network included 95.6% of the Bio*Me* cohort. We identified 16 groups with ≥20 individuals representing 92.5% of the Bio*Me* cohort (**Fig. 1b**), consistent with their published results (9,10). We found that AoU and Bio*Me* have several similar populations with F_{ST} < 0.001 between them, even after removing related individuals across both datasets, reflecting their shared NYC recruitment area (**Fig. 1b**).

Of the IBD groups, there were five from AoU and eight from Bio*Me* with IBD scores >3, defining them as founder populations (**Fig. 1c**). Of the eight founder groups from BioMe, founder effects in Ashkenazi Jewish, Puerto Rican and Garifuna were reported previously (9,10). We also found the founder populations in AoU showed major geospatial differences, with the IBD group 9 (Garifuna) in particular over-represented in the Bronx compared to other boroughs (not shown to comply with AoU Data and Statistics Dissemination Policy).

Detection of founder populations in NYC

Since our AoU dataset and Bio*Me* are both NYC cohorts with shared genetic features (**Figs. 1b-c, Figure S3**), we combined the IBD groups from AoU and Bio*Me* with pairwise Hudson's F_{ST} values < 0.001, resulting in 16 IBD groups which we annotated based on inferred ancestry (**Fig. 2**). Eight groups were identified as founder populations (IBD score >3). The populations were named based on self-defined ancestry as provided by the reference datasets, when available (**Figs. 1b-c,2,3, Table S1**). Genetic ancestry also acts on a continuum(17), therefore some IBD groups appeared to be more discrete (*e.g.* Garifuna, Puerto Rican), whereas others include individuals across a broader geographic range (*e.g.* Filipino/Pacific Islander, Mexican/South American, Han Chinese/East Asian). In situations where groups could not be confidently labeled, the closest associated population group was used as a placeholder label until more genomic reference datasets become available. These populations are labeled with an asterisk (**Figs. 1b-c, 2,3**).

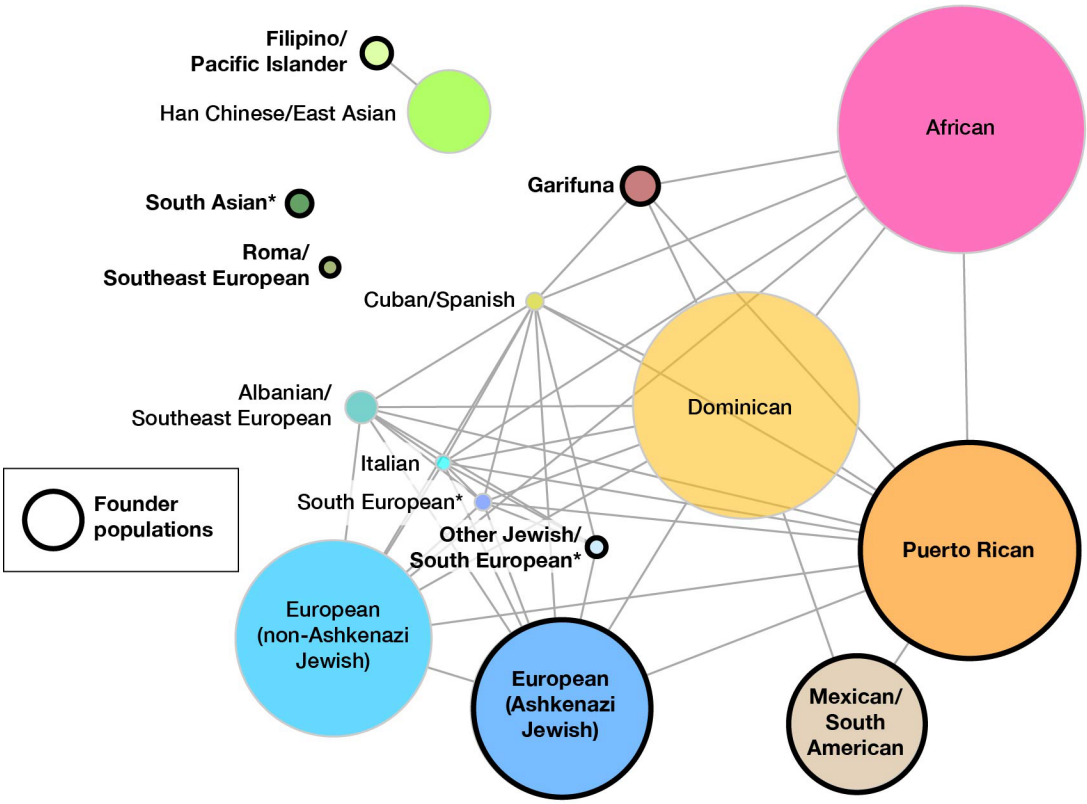


Figure 2: IBD groups identified in NYC from the combined AoU and Bio*Me* datasets.

A network depiction of the IBD groups based on F_{ST} between IBD groups. Edges were weighted by logarithm of F_{ST}. Only edges representing F_{ST} <0.01 are shown, with founder populations circled in black. Circle sizes reflect the number of individuals in each IBD group. An asterisk next to labels represent populations with inadequate reference information for annotation.

Identification of pathogenic founder variants

We then studied the genomes of the members of the founder populations to identify pathogenic variants characterizing each group. We used the ClinVar resource (15) as a curated source of disease-causing variants, while recognizing that the bias of ClinVar towards documentation of

variants in individuals of European ancestry (18). We focused on recurrent, rare, disease-causing variants, given our focus on founder effects. By requiring the pathogenic/likely pathogenic (P/LP) variant to occur in at least 2 individuals not from the same family, we identified 3,616 P/LP recurrent variants in NYC individuals. Consistent with the known ClinVar bias(18), European ancestry IBD groups showed more pathogenic variants than other IBD groups, especially when compared with individuals with African ancestry (**Fig. 3**).

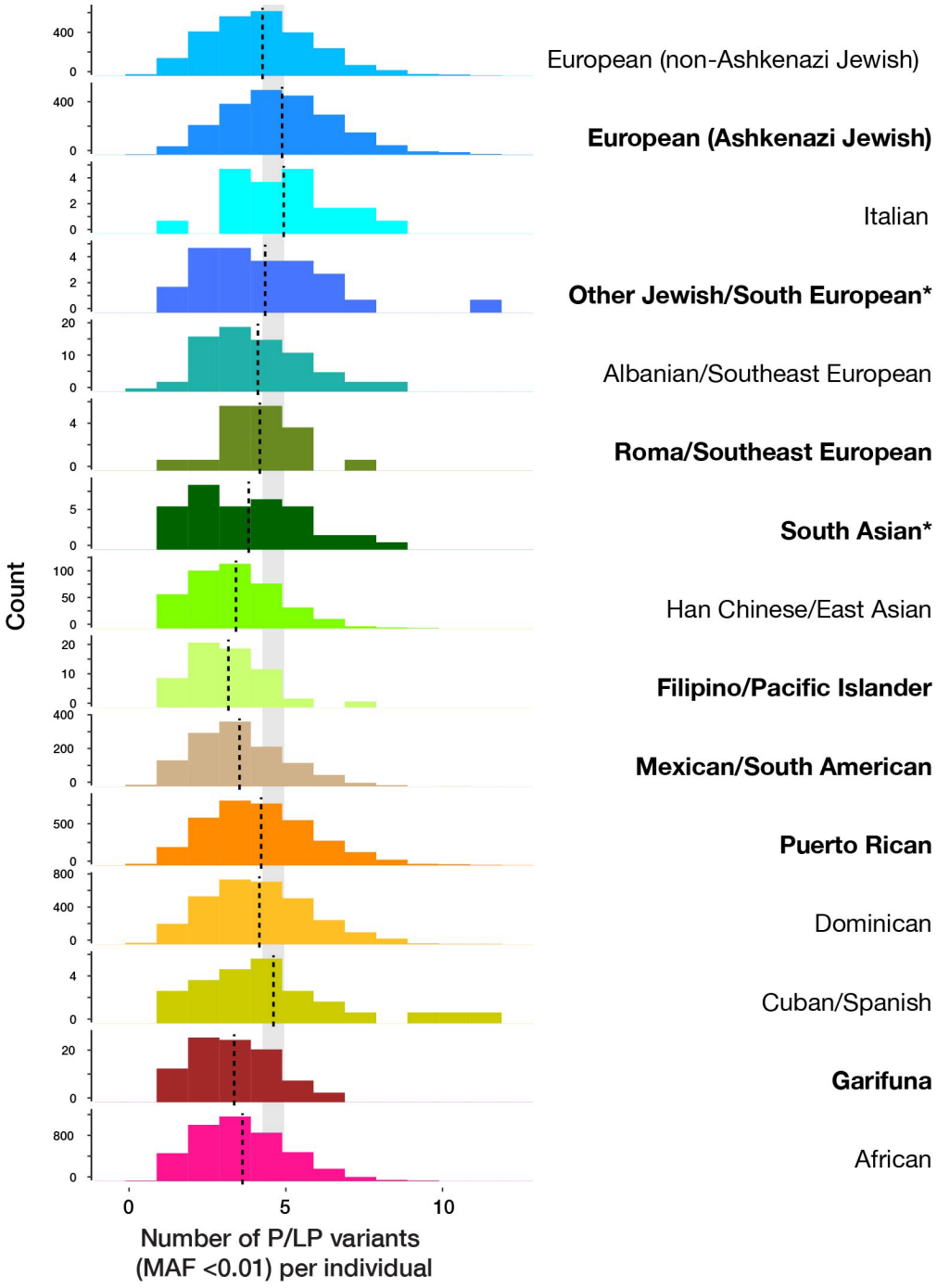


Figure 3: Number of P/LP variants per individual for each NYC IBD group identified from the AoU and Bio*Me* datasets.

Each histogram shows the number of P/LP variants with minor allele frequency < 0.01 per individual for 15 IBD groups. The vertical dashed line indicates the mean value in each group.

The 'South European' IBD group only ~~i~~included WGS data for fewer than 20 individuals and was removed due to the All of Us Data and Statistics Dissemination Policy. The shaded gray rectangle represents the range of mean values for the uppermost three European groups, highlighting the lower number of ClinVar variants annotated in those of Asian, American and African ancestries. Asterisk next to labels represent populations with inadequate reference information for annotation.

We detected 674 unique P/LP variants significantly enriched across the 8 founder populations (Fisher's Exact $p > 0.05$) (**Table S2**) with 202 of these variants passing Bonferroni correction. Of the 674 variants, 478 variants have two or more ClinVar gold stars, meaning variants are from practice guidelines, reviewed by expert panel, or from multiple submitters with evidence and no classification conflicts. **Table S**Those variants with no ClinVar gold stars should in general be interpreted with caution, such as the *KRT18* variant not previously described to be common in Ashkenazi Jewish individuals. The *CD55* variant associated with protein-losing enteropathy (19) and shown in cell studies to cause loss of CD55 on the cell surface (20) also lacks a star rating, illustrating how the absence of this rating should not be used to exclude variants as disease-causing. This *CD55* variant also does not pass Bonferroni correction, nor do known founder effect variants in *HBB* in the South Asian group [REF] and a *SLC26A4* variant in the Filipino/Pacific Islander group [REF], prompting us to include variants that do not pass multiple testing correction in **Table 1** as candidate founder disease-causing variants in the IBD groups. We identified 51 variants from this broader list that have minor allele frequencies of > 0.005 in one or more IBD groups, the Tier 2 threshold for inclusion into prenatal screening panels (21). Of these, 25 are new, previously unrecognized founder effect variants (**Table 1).** The results shown in **Fig. 3** support the likelihood that the numbers of P/LP variants in non-European groups are likely to represent an underestimate, and that more disease-causing variants remain to be discovered in these under-studied groups.

198

199 **Table 1: ClinVar P/LP variants with allele frequencies exceeding 1/200 within seven**
200 **founder populations in NYC.**

Gene	ClinVar accession	HGVS description	ClinVar star rating	Disease	Allele frequency (Bold: significant following Bonferroni correction)	Published founder variant (PMID)
European (Ashkenazi Jewish)						
F11	VCV000011892	NM_000128.4(F11):c.901T>C (p.Phe301Leu)	2	Hereditary factor XI deficiency	**0.0248**	2813350
GJB2	VCV000017010	NM_004004.6(GJB2):c.167del (p.Leu56fs)	3	Autosomal recessive nonsyndromic hearing loss	**0.0157**	9819448
ELP1	VCV000006085	NM_003640.5(ELP1):c.2204+6T>C	2	Familial dysautonomia	**0.0154**	12116234
HEXA	VCV000003889	NM_000520.6(HEXA):c.1274_1277dup (p.Tyr427fs)	2	Tay-Sachs disease	**0.0142**	2848800
F11	VCV000011891	NM_000128.4(F11):c.403G>T (p.Glu135Ter)	2	Hereditary factor XI deficiency	**0.0140**	2813350
KRT18	VCV000014585	NM_000224.3(KRT18):c.383A>T (p.His128Leu)	none	Cirrhosis	**0.0098**	Not described
CFTR	VCV000007129	NM_000492.4(CFTR):c.3846G>A (p.Trp1282Ter)	4	Cystic fibrosis	**0.0094**	1384328
MPL	VCV000135563	NM_005373.3(MPL):c.79+2T>A	2	Congenital amegakaryocytic thrombocytopenia	**0.0067**	21489838
DDX11	VCV000252749	NM_030653.4(DDX11):c.1763-1G>C	2	Warsaw breakage syndrome	**0.0063**	31287223
ASPA	VCV000002605	NM_000049.4(ASPA):c.854A>C (p.Glu285Ala)	2	Canavan disease	**0.0058**	8252036
SLC3A1	VCV000336195	NM_000341.4(SLC3A1):c.808C>T (p.Arg270Ter)	2	Cystinuria	**0.0054**	7539209
FKTN	VCV000003203	NM_001079802.2(FKTN):c.1167dup (p.Phe390fs)	2	Walker-Warburg congenital muscular dystrophy	**0.0054**	19266496
CLRN1	VCV000004395	NM_174878.3(CLRN1):c.144T>G (p.Asn48Lys)	2	Usher syndrome type 3A	**0.0052**	12145752
PAH	VCV000102706	NM_000277.3(PAH):c.506G>A (p.Arg169His)	3	Phenylketonuria	**0.0050**	29144512
DLD	VCV000011966	NM_000108.5(DLD):c.685G>T (p.Gly229Cys)	2	Pyruvate dehydrogenase E3 deficiency	**0.0050**	14765544
FANCC	VCV000012045	NM_000136.3(FANCC):c.456+4A>T	2	Fanconi anemia complementation group C	**0.0050**	8348157
Other Jewish/South European*						
FMO3	VCV000985096	NM_001002294.3(FMO3):c.1499G>A (p.Arg500Gln)	1	Trimethylaminuria	**0.0600**	Not described
CD55	VCV000431759	NM_000574.5(CD55):c.596C>T (p.Ser199Leu)	none	Complement hyperactivation, angiopathic thrombosis, and protein-losing enteropathy	0.0400	35314883
KLKB1	VCV000012033	NM_000892.5(KLKB1):c.337C>T (p.Arg113Ter)	none	Prekallikrein_deficiency	0.0400	Not described
Puerto Rican						
HPS1	VCV000005277	NM_000195.5(HPS1):c.1472_1487dup (p.His497fs)	2	Hermansky-Pudlak syndrome 1	**0.0097**	8896559
RSPH4A	VCV000088863	NM_001010892.3(RSPH4A):c.921+3_921+6del	2	Primary ciliary dyskinesia	**0.0096**	23798057
COL27A1	VCV000143245	NM_032888.4(COL27A1):c.2089G>C (p.Gly697Arg)	2	Steel syndrome	**0.0091**	24986830
TBCK	VCV000225235	NM_001163435.3(TBCK):c.376C>T (p.Arg126Ter)	2	Infantile hypotonia, infantile with psychomotor retardation and characteristic facies 3	**0.0089**	29283439

ABCB4	VCV000802326	NM_000443.4(ABCB4):c.2784-12T>C	1	Progressive familial intrahepatic cholestasis type 3	**0.0080**	34678161
ERCC6L2	VCV000421974	NM_020207.7(ERCC6L2):c.19C>T (p.Gln7Ter)	2	Bone marrow failure syndrome 2	**0.0076**	Not described
MYO15A	VCV000164548	NM_016239.4(MYO15A):c.7226del (p.Pro2409fs)	2	Deafness, autosomal recessive 3	**0.0063**	Not described
LTBP2	VCV001515466	NM_000428.3(LTBP2):c.2908+1G>A	1	Microspherophakia and/or megalocornea, with ectopia lentis and with or without secondary glaucoma	**0.0059**	Not described
ECE1	VCV000009133	NM_001397.3(ECE1):c.2260C>T (p.Arg754Cys)	none	Hirschsprung disease, cardiac defects, and autonomic dysfunction	**0.0056**	Not described
MRPS34	VCV000438633	NM_023936.2(MRPS34):c.322-10G>A	2	Leigh Syndrome	**0.0054**	28777931
GDAP1	VCV000004202	NM_018972.4(GDAP1):c.692C>T (p.Pro231Leu)	2	Charcot-Marie-Tooth disease	**0.0053**	34057104
SGCG	VCV000002009	NM_000231.3(SGCG):c.787G>A (p.Glu263Lys)	2	Limb-girdle muscular dystrophy type 2C	**0.0051**	16832103
Roma/Southeast European						
TSEN54	VCV000620188	NM_207346.3(TSEN54):c.1039A>T (p.Lys347Ter)	2	Pontocerebellar hypoplasia	**0.0526**	Not described
South Asian*						
PAH	VCV000092741	NM_000277.3(PAH):c.355C>T (p.Pro119Ser)	3	Phenylketonuria	**0.0385**	Not described
ABCC2	VCV000426249	NM_000392.5(ABCC2):c.3337del (p.Val1114fs)	2	Dubin-Johnson syndrome	0.0256	Not described
HBB	VCV000015437	NM_000518.5(HBB):c.92+1G>T	2	Beta thalassemia	0.0256	2064964
CEP152	VCV000158223	NM_001194998.2(CEP152):c.1155del (p.Thr386fs)	2	Primary microcephaly 9	0.0256	Not described
TGFBI	VCV001175370	NM_000358.3(TGFBI):c.1406G>A (p.Arg469His)	none	Granular corneal dystrophy	0.0256	Not described
FANCE	VCV001696377	NM_021922.3(FANCE):c.2_7del (p.Met1_Ala2del)	1	Fanconi anemia	0.0256	Not described
Garifuna						
MYBPC3	VCV000164113	NM_000256.3(MYBPC3):c.1484G>A (p.Arg495Gln)	2	Hypertrophic cardiomyopathy	**0.0245**	Not described
DUOX2	VCV000004065	NM_001363711.2(DUOX2):c.1126C>T (p.Arg376Trp)	2	Congenital hypothyroidism	**0.0196**	Not described
CNGB1	VCV001031963	NM_001297.5(CNGB1):c.1217G>A (p.Trp406Ter)	1	Retinitis pigmentosa	**0.0147**	Not described
COL18A1	VCV001484134	NM_001379500.1(COL18A1):c.2214+1G>A	1	Glaucoma, primary closed-angle; Knobloch syndrome	**0.0147**	Not described
GBE1	VCV000478912	NM_000158.4(GBE1):c.993-1G>T	2	Glycogen storage disease, type IV; Polyglucosan body disease, adult form	**0.0147**	Not described
BBS12	VCV000550386	NM_152618.3(BBS12):c.1151del (p.Ser384fs)	2	Bardet-Biedl syndrome 12	**0.0147**	Not described
GALC	VCV000429982	NM_000153.4(GALC):c.379C>T (p.Arg127Ter)	2	Krabbe disease	0.0098	Not described
COL7A1	VCV001454264	NM_000094.4(COL7A1):c.7244dup (p.Met2415fs)	2	Dystrophic epidermolysis bullosa	0.0098	Not described
EOGT	VCV000523593	NM_001278689.2(EOGT):c.78_81del (p.His27fs)	2	Adams-Oliver syndrome 4	0.0098	Not described
Filipino/Pacific Islanders						
LMBR1L	VCV001285608	NM_018113.4(LMBR1L):c.863G>A (p.Arg288Gln)	none	Differences of sex development	**0.0214**	Not described
FLG	VCV000280218	NM_002016.2(FLG):c.7487del (p.Thr2496fs)	2	Ichthyosis vulgaris	0.0143	Not described
SLC26A4	VCV000043565	NM_000441.2(SLC26A4):c.706C>G (p.Leu236Val)	3	Pendred syndrome; Deafness, autosomal recessive 4, with enlarged vestibular aqueduct	0.0143	30113565

CD36	VCV000225309	NM_001001548.3(CD36):c.332_333del (p.Thr111fs)	2	Platelet glycoprotein IV deficiency	0.0143	Not described

TABLE LEGEND

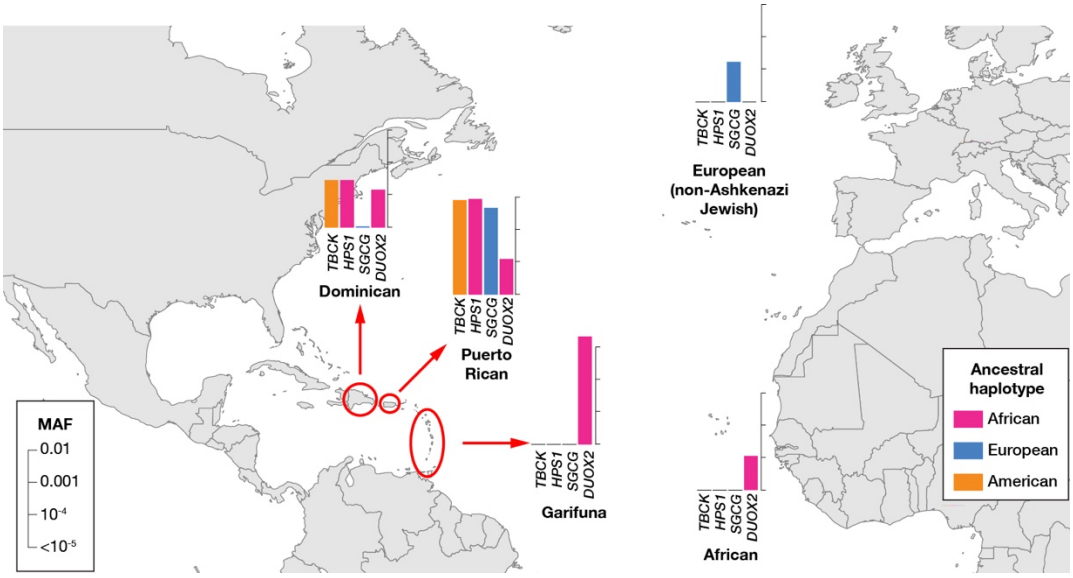
The table shows the P/LP variants that occur within each population at a frequency of at least 1/200 alleles, the threshold for inclusion in prenatal testing. Those with allele frequencies in bold type pass Bonferroni correction. Of these 51 variants, 26 have already been published as founder mutations, especially in the European (Ashkenazi Jewish) and Puerto Rican populations. The other 25 are new, previously unrecognized founder effect variants. Abbreviations: PMID, PubMed reference number. An asterisk next to labels represent populations with inadequate reference information for annotation. Some of the allele frequency counts displayed here represent <20 AoU participants. The AoU Resource Access Board reviewed this work and granted an exception to their Data and Statistics Dissemination policy to report these frequencies.

Ancestry analysis for shared founder variants in individuals of Caribbean ancestry

We detected 12 and 9 founder P/LP variants for Puerto Rican and Garifuna IBD groups, including variants that did not pass Bonferroni correction, respectively (**Table 1**). Of these 21 variants, 15 were also detected at lower frequencies in other IBD groups **(Table S3)**. To test whether this was due to shared ancestry, we inferred local ancestry (the origin of the DNA containing each P/LP variant) in the Caribbean individuals to identify the ancestral population in which each P/LP variant arose originally. We found the majority of the founder variants in Puerto Ricans to be located on haplotypes of European ancestry, with the remaining founder variants located on African and Indigenous American haplotypes (**Table S3**). Of these, the origin of the *COL27A1* Steel syndrome variant on chromosome 9 has also previously been characterized as Native American ancestry (22) We illustrate the sharing of founder effect mutations across Caribbean groups and with European and African groups in **Fig. 4**. The *MYBPC3* variant that occurs frequently in the

225 Garifuna was in an African haplotype, but was also found in a European haplotype in an individual
226 unassigned to any IBD group, indicating that the same mutation arose independently in African
227 and European individuals. This finding demonstrates how local ancestry analysis can be used to
228 reveal the evolutionary history of founder effect mutations, and how the presence of a founder
229 effect mutation does not by itself indicate a person is part of a known founder population.

230



231

232 **Figure 4: Shared founder variant frequencies in Caribbean IBD groups.**

233 Four examples of founder mutations are illustrated with comparisons of frequencies in other IBD
234 groups. Local ancestry analysis reveals African origin of the *HPS1* and *DUOX2* pathogenic
235 variants, the Indigenous American origin of the *TBCK* and European origin of the *SGCG*
236 pathogenic variants. No variant is exclusive to one IBD group but occurs across multiple
237 Caribbean groups, reflecting the complex ancestries of these populations and the weakness of
238 demographic categories such as race and ethnic origin as the sole predictors of genetic disease
239 risks.

240

241 **Discussion**

In this study, we showed that founder populations exist in the megacity of New York, and that the individuals from these genetic ancestry groups have distinctive increased risks of rare genetic diseases. The deliberate inclusiveness of the AoU Research Program(23) has ensured that insights extend to communities with genetic ancestries other than those included in the non-Hispanic White demographic. By defining genetic similarity using IBD sharing network as opposed to crude demographic or continental groupings, and focusing on DNA sequence variants with strong prior evidence for causing genetic diseases, we rediscovered many known rare disease-causing variants common in the better-studied Ashkenazi Jewish and Puerto Rican communities, while revealing new founder mutations in these and other founder populations in NYC. The value of this recognition is not only in terms of otherwise rare diseases entering into the differential diagnosis of a patient being evaluated clinically, but also in terms of inclusion of these genes and conditions in prenatal carrier screening. The utility of information about rare genetic conditions in a founder population is exemplified in the Ashkenazi Jewish community, where genetic testing panels for prenatal use is expanding to include presymptomatic testing for conditions affecting the parents (24) The American College of Medical Genetics has recommended that "carrier screening paradigms should be ethnic and population neutral and more inclusive of diverse populations to promote equity and inclusion" (21) This study demonstrates how current carrier screening panels can be expanded to serve a broader set of under-represented communities, using the example of the diverse population of NYC.

To identify genetic similarities between individuals, we used an IBD sharing network, which can identify groups of individuals who share recent ancestry in an unbiased manner (10–14) Due to the nature of IBD segments, this approach works well to identify founder populations, who are expected to share higher genetic ancestry as well as population-specific disease-causing variants. However, the assignment of individuals to a group within the network is highly dependent on who

is included in the network. This becomes more complicated when there are individuals who lie at the boundaries between groups, because they have multiple ancestry components due to admixture. The inclusion of admixed individuals with shared ancestry allows us to capture more population-specific pathogenic variants but also leads to underestimations of allele frequencies. Varying the length thresholds of IBD segments, using different community detection algorithms, or combining or annotating IBD groups based on different F_{ST} thresholds will also change the resolution of population structure that is captured.

To describe these groups, we used information about how the individuals from these groups described themselves as well as genetic similarity with reference population (F_{ST}), defining the genetic ancestry groups for this study (**Table S1)**. Prior reports of disease-causing DNA sequence variants allowed inference of the origins of some of the founder populations, with the *SLC26A4* variant in the Filipino/Pacific Islander group known to be common in Filipinos (25), and the *CD55* variant in the Other Jewish/South European group previously identified in Bukharian Jewish individuals (19) While genetic ancestry group information is not routinely captured in health records, a clinical encounter seeking to understand a patient's rare disease will typically involve a detailed family history, seeking evidence of consanguinity that involves asking about the origins of grandparents and prior generations. Genetic ancestry group information is therefore much more likely to be captured in rare disease care and can be used to raise the possibility that known founder mutations in that community may be the cause of the affected individual's rare disease.

The potential that the recognition of the presence of a rare disease in a founder population can lead to targeted therapies is exemplified by the *CD55* variant in the Bukharian Jewish community, which can cause a spectrum of presentations from mild abdominal discomfort following a high-fat meal to a severe syndrome including protein-losing enteropathy, and is effectively treated by the complement C5-inhibitor eculizumab (26) Implementation in health care systems of information about rare disease susceptibility for founder populations can therefore encompass prenatal

292 screening, clinical decision support to prompt clinicians to be aware of an otherwise rare disease
293 in a patient from a defined community, and can lead to therapeutic interventions.

294

295 For a variant to be categorized in ClinVar as Pathogenic or Likely Pathogenic, it has to fulfil a
296 number of stringent criteria (27) The degree of confidence about the variant categorization is
297 represented by a star system, reflecting the degree of expert curation of the variant. We note that
298 the *KRT18* variant that meets the criteria for inclusion in **Table 1** is rated with zero stars, and may
299 not be a true risk allele in the European (Ashkenazi Jewish) genetic ancestry group. There are
300 other reasons why categorizations of likely pathogenic or pathogenic variants in ClinVar (15) may
301 not be reliable. For example, older submissions to the database are prone to subsequent
302 conflicting interpretations (28) sometimes because of the failure to appreciate the variant to be
303 relatively common in one understudied population at the time of submission (29). Our use of
304 ClinVar has revealed many new variants causing rare genetic diseases in under-represented
305 populations of NYC, but we recognize that ClinVar's bias towards P/LP variants in Europeans
306 implies that there remains even more to be discovered about rare diseases in the non-European
307 founder populations studied.

308 Another general problem with pathogenicity classifications is the assumption that the presence of
309 a variant at a frequency higher than the prevalence of the associated disease should lead to the
310 variant being reclassified as non-causative. This by itself is a reasonable general assumption,
311 but in the context of the groups we are studying here we find two reasons for concern. One is
312 that founder populations can have high frequencies of a variant and should be excluded from this
313 filtering approach (30) This approach is implemented in Grpmax FAF, the filtering allele
314 frequencies offered by gnomAD (31) which excludes founder populations like the Amish,
315 Ashkenazi Jewish and Finnish from frequency calculations. By identifying other founder
316 populations, approaches like Grpmax FAF can be refined with variant frequency information from

these additional groups. The other concern is that disease prevalence measurements may vary between communities depending on access to care, which is a concern in NYC (32) and in US health care more generally (33) If a community lacks access to care, the prevalence of a rare disease in that community may go unrecognized, with the further failure to recognize an association with a disease-causing variant that may then be misclassified as benign. As we demonstrate here, large-scale population studies like AoU will make it increasingly feasible to gain insights into variant frequencies in different genetic ancestry groups, including founder populations, but some of these genetic ancestry groups will also be defined by limited access to health care. There is the potential to worsen health equity by applying exclusive variant frequency thresholds that fail to recognize the genetic disease burden of founder populations through adequate phenotyping.

We have to balance the value of identifying a genetic disease risk in a community with the risk of stigmatizing that group. The AoU Research Program notes this potential for biased interpretation promoting negative stereotypes (34) We therefore emphasize how pathogenic variants occur in everyone, regardless of demographic categorization or genetic ancestry (**Fig. 3**). What distinguishes founder effect groups is not likely to be the overall burden of genetic damage, but instead the over-representation of specific genetic diseases (**Table 1**) within that burden of damaging variants. We also stress how differences in the numbers of damaging variants in the genomes of people from different parts of the world have more to do with incompleteness of information about and genomic annotations of damage. We demonstrate the shortcomings of crude demographic categories such as race and ethnic origin in predicting genetic disease risks. We identified a strong founder effect in the Garifuna ancestry group, but when they self-identify their race and ethnicity they include African American/Black, Hispanic and Latino, and in some cases diverse countries of origin, illustrating the weakness of these categorizations as proxies for

genetic variation (17,35) Similarly, we found multiple self-identified ancestries in some of the non-founder groups, who were not the focus of this manuscript. For example, the 'African' group included individuals who align with African, African-American, and African-Caribbean ancestry groups, reflected in the wide spectrum of ancestry in the PCA analysis (Figure 1b, Figure S1b).

We find that some of the risk alleles from the founder Caribbean populations in NYC also exist at lower frequencies in other Caribbean New Yorkers, because of the complex history of pre-colonial civilization, colonization, slavery and migration. In some cases, individuals of non-Caribbean origin appeared to have founder effect variants that appear in the Caribbean, but these variants were mostly located on shared haplotypes derived from the same continental ancestry, with the exception of one variant that appears to have independently arisen in Garifuna and European individuals (**Table S3**). Demography is therefore only modestly informative in predicting disease risk, making any associated stigma tenuous. Instead we followed the guidelines of the National Academies of Sciences, Engineering, and Medicine (NASEM) on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research (2) to quantify objectively 'genetic similarity' using IBD sharing, and 'genetic ancestry' labels using those provided by members of each group (**Table S1**). We also worked with members of the genetic ancestry groups highlighted in the results to discuss and prepare this report, following recommendation 5 of the NASEM report. These best practice guidelines are clearly of value in using population descriptors in ways that enhance the application of genomic insights in medical care delivery.

This study shows how genetic variants that cause diseases that are rare globally can be common locally within a population, and can influence the spectrum of diseases of patients served by individual health systems. Our focus was on NYC, but the same approaches can be extended nationally using AoU data and comparable international data resources. The insights gained are essential for better health care provision, while highlighting the need to gain insights into the

phenotypic manifestations of disease-causing variants in marginalized populations with less access to health care.

Materials and Methods

Research participants and dataset preparation

Participants living in NYC in the AoU Program version 6 curated data repository (8) were identified by the first three digits of their Zip Code, allowing borough-level resolution of geographic residence. Microarray genotype data were used to assess the population structure of NYC participants by principal component analysis (PCA), by global ancestry analysis as performed by SCOPE (36) and by Identity-by-descent (IBD) analysis as described below. Samples were QCed by call rate and kinship coefficient using PLINK v2.00a2.3LM (37) No individuals were filtered out by the call rate threshold of 0.9 (---mind 0.1). To remove close relatives, either of the pairs of individuals who showed king kinship coefficients > 0.125 were removed using --king-cutoff 0.125 in PLINK2.0 (37). Variants of the array data were filtered with the following conditions using PLINK2.0: minor allele frequency >0.01, genotyping rate per site > 0.95, and p-value for the departures from Hardy Weinberg Equilibrium (HWE) > $1x10^{-6}$ (--maf 0.01 –geno 0.05 --snps-only --hwe 1e-06). After QC steps, 13,817 participants and 720,630 SNPs remained for downstream analysis.

Out of these individuals, 10,381 individuals had whole genome sequence (WGS) data available. We used this subset of individuals and whole genome sequence data from an independent NYC biobank, the Mount Sinai Bio*Me* biobank (9,10), to identify founder pathogenic variants. Approval to study these de-identified data was granted by the Albert Einstein College of Medicine Internal Review Board (Protocol 2016-7099). All analyses on AoU participants included in the manuscript were also approved by the All of Us Resource Access Board.

391

392

393 **Comparison of demography between US Census and AoU NYC participants**

394 To show the extent to which our dataset represents the demography of NYC, we compared the
395 proportion of four major self-described race and ethnicities per borough between census data and
396 AoU NYC participants. We obtained census data from the following source:
397 https://www.census.gov/quickfacts/fact/table/richmondcountynewyork,newyorkcountynewyork,q
398 ueenscountynewyork,kingscountynewyork,bronxcountynewyork,newyorkcitynewyork/PST04522
399 (16).

400 For AoU NYC participants, self-identified race/ethnicity was obtained using a questionnaire.
401 Participants answered the question: “Which categories describe you? Select all that apply. Note,
402 you may select more than one group.” in the Basics Survey. Borough residence was defined
403 based on the first three digits of the zip code of residence, provided by AoU.

404

405 **PCA and global ancestry analysis**

406 PCA and global ancestry analysis were conducted using PLINK 2.0 and SCOPE (36) in
407 supervised mode, respectively, on a combined dataset comprising 13,817 AoU participants and
408 3,584 individuals from the 1000 Genomes Project (1KGP) (38) the Human Genome Diversity
409 Project (HGDP) (39) and the Simons Genome Diversity Project (SGDP) (40) using a total of
410 150,213 SNPs. Prior to global ancestry analysis using SCOPE, we conducted ADMIXTURE
411 analysis (41) with K=5 on this assembled reference panel, and further identified individuals within
412 this panel for whom >95% of their genomes appeared to originate in any of five continental
413 ancestries: African, European, South Asian, East Asian and Native American (38) The supervised
414 SCOPE analysis was run based on this subset of reference panel participants.

415

416 **Identity-by-descent (IBD) analysis**

417 IBD groups, the sets of individuals who share ancestry as defined by shared IBD segments, were
418 constructed from the microarray genotypes. Phasing of the genotypes was conducted with Beagle
419 v5.4 (42) using all populations from 1KGP (38) as references. We used Templated Positional
420 Burrows-Wheeler Transform (TPBWT) (43) on the phased dataset to infer IBD segments >3 cM
421 across all pairs of individuals. The total length of IBD sharing for all pairs of individuals was used
422 to construct an undirected network using the iGraph package(44) in R. To focus on recent
423 demography and to reduce clustering of extended families, we filtered for edges with cumulative
424 IBD sharing ≥12 cM and ≤72 cM, as previously described (11,14) IBD groups were detected
425 using the infomap.community() (45) function on the constructed network using default parameters.
426 To assess the strength of the founder effect for each IBD group, we estimated the 'IBD score',
427 the average length of IBD segments between 3–20 centimorgans (cM) shared between two
428 genomes normalized to sample size, as previously described (46) We also estimated the IBD
429 score per group and per borough. To confirm the robustness of our approach and to obtain a
430 reliable reference for population labels, we conducted IBD sharing network analysis for an
431 independent NYC cohort, the Mount Sinai BioMe Biobank9(9,10) (dbGaP Accession number
432 phs001644). After QC, the BioMe dataset consisted of 11,549 individuals and 982,770 SNPs.
433 We performed IBD analysis as above and named each BioMe IBD group based on
434 individuals' detailed self-reported ethnicity provided separately by BioMe leadership (Alexander
435 Charney, personal communication). IBD groups in BioMe and AoU with IBD score >3 were
436 defined as founder populations (**Fig. 1c**).

437

438 **Inferring ancestral background of individuals in IBD groups**

Since AoU did not provide the detailed ethnicity information that is particularly useful for defining founder groups, we inferred population ancestry (*e.g.* Puerto Rican, Dominican, Ashkenazi Jewish) in AoU IBD groups by estimating Hudson's F_{ST} between each group and populations in genomic reference panels using PLINK2.0. The reference panel included 14,985 individuals and 140,952 biallelic SNPs from global populations with sample sizes > 10 individuals in 1KGP, HGDP and SGDP together with IBD groups in Bio*Me*. We also conducted PCA for the merged dataset. IBD groups in AoU and Bio*Me* with F_{ST} values < 0.001 were combined in further analyses.

Detection of pathogenic founder variants for rare diseases

We extracted variants categorized as pathogenic or likely pathogenic (P/LP) in the ClinVar database (15) (version ClinVarFullRelease_2023-01.xml) from WGS data of 10,381 NYC AoU participants and 11,549 Bio*Me* participants. Of the 193,935 P/LP variants registered in ClinVar as of Jan 7, 2023, we detected 27,125 variants in our NYC cohort. We removed close relatives and excluded variants that appeared only once in the NYC dataset. We then filtered variants with genotype rate < 0.9 and p-values for departures from Hardy-Weinberg Equilibrium (HWE) $<1x10^{-16}$. We set a small HWE threshold anticipating that rare variants may likely diverge from HWE due to high heterozygosity. After filtering, 3,616 P/LP variants were observed in NYC individuals. HGVS description and review status (gold stars) were obtained from variant_summary.txt.gz in https://ftp.ncbi.nlm.nih.gov/pub/clinvar last updated on March 30, 2024. The variants which were not classified as P/LP as of March 30, 2024 were removed from results.

We defined eight IBD groups with IBD scores >3 as founder populations. To identify founder variants, we set a conservative threshold, including only those that: a) were significantly enriched in a certain founder population compared with other NYC individuals (Fisher's Exact p < 0.05), b) occurred at a MAF of <0.0001 in NYC individuals not assigned to that group, and c) appeared more than once in that group. We applied the Bonferroni correction (p value < 0.05 / (3,616 x 8)),

but all results are listed in **Table 1** and **Table S2** since it is too strict for populations with small sample size. The minor allele frequencies of founder variants were extracted from gnomAD v3.1.225(47) using gnomAD_DB (https://github.com/KalinNonchev/gnomAD_DB) to compare frequencies in NYC dataset. The number of P/LP variants per individual was also counted for each IBD group.

Ancestry analysis for the founder variant in the Caribbean IBD groups

We identified multiple IBD groups that appeared to have Caribbean ancestry, based on F_{ST} analysis against reference populations. Since Caribbean populations have three different continental ancestries in their genomes (African, Native American and European) due to their complex history, we inferred local ancestry around the founder variants detected in Caribbean populations shown in **Table 1** in order to reveal the ancestral background of those founder variants.

Genotype datasets for the carriers were generated by combining the genotype dataset used for IBD analysis and genotype data of each ClinVar variant, and phased by Beagle v5.4 without reference genomes. We then used RFMix(48) version 2 to infer local ancestry ±20 Mb of the variant, with 3 expectation-maximization steps. To assess local ancestry, we assembled a reference panel by identifying individuals from 1KG, HGDP and SGDP for whom >95% of their genomes appeared to have either African, American or European ancestry based on ADMIXTURE analysis with K=5 as reference (the same reference individuals in the SCOPE analysis).

ACKNOWLEDGEMENTS

The authors thank Drs. Alex Charney, Gillian Belbin, and Eimear Kenny for providing Bio*Me* self-described population ancestry labels. The advice of Humberto Brown (Director of Health Disparities, Arthur Ashe Institute for Urban Health) and of Karen Blanco and Katherine Oliva Blanco (Hondurans Against AIDS/Casa Yurumein) is also gratefully acknowledged. This work was supported by OT2OD031919 from the Office of the Director (NIH) to MS and SR and R01AG057422 from the National Institute on Aging (NIH) to JMG. We thank the All of Us Resource Access Board for their thoughtful review and final approval of this manuscript prior to publication. Finally, the authors thank the participants of AoU and Bio*Me*, without whom this research would not be possible.

The All of Us Research Program is supported by the National Institutes of Health, Office of the Director: Regional Medical Centers: 1 OT2 ODO26549; 1 OT2 ODO26554; 1 OT2 ODO26557; 1 OT2 ODO26556; 1 OT2 ODO26550; 1 OT2 OD26552; 1 OT2 ODO26553; 1 OT2 ODO26548; 1 OT2 ODO26548; 1 OT2 ODO2551; 1 OT2 ODO26555; IAA #: AOD 16037; Federally Qualified Health Centers: HHSN 263201600085U; Data and Research Center: 5 U2C ODO23196; Biobank: 1 U24 OD023121; The Participant Center: U24 ODO23176; Participant Technology Systems Center: 1 U24 ODO23163; Communications and Engagement: 3 OT2 ODO23205; 3 OT2 ODO23206; and Community Partners: 1 OT2 ODO25277; 3 OT2 ODO25315; 1 OT2 ODO25337; 1 OT2 ODO25276.

Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). Genome sequencing for “NHLBI TOPMed:phs001644 was performed at MGI (3UM1HG008853-01S2). Core support including

centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393;U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

Data availability

The data included in this manuscript was entirely obtained from publicly available resources. We have reported on our findings in accordance with the terms of the respective data sources (*e.g.* AoU requires that reporting on groups of individuals be restricted to ≥ 20 individuals). We have not generated data that requires sharing.

References

1. Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. European Journal of Human Genetics 2019 28:2. 2019 Sep 16;28(2):165–73.
2. National Academies of Sciences, Engineering, and Medicine. Using Population Descriptors in Genetics and Genomics Research. Washington, D.C.: National Academies Press; 2023.
3. Scott SA, Edelmann L, Liu L, Luo M, Desnick RJ, Kornreich R. Experience with carrier screening and prenatal diagnosis for 16 Ashkenazi Jewish genetic diseases. Hum Mutat. 2010 Nov;31(11):1240–50.
4. Gross SJ, Pletcher BA, Monaghan KG. Carrier screening in individuals of Ashkenazi Jewish descent. Genetics in Medicine. 2008 Jan;10(1):54.
5. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. Cell. 2019;177(1):26–31.
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 Apr 29;51(4):584–91.
7. Fox K. The Illusion of Inclusion — The "All of Us" Research Program and Indigenous Peoples' DNA. New England Journal of Medicine. 2020 Jul 30;383(5):411–3.
8. All of Us Research Program Investigators, Denny JC, Rutter JL, Goldstein DB, Philippakis A, Smoller JW, et al. The "All of Us" Research Program. N Engl J Med. 2019 Aug 15;381(7):668–76.
9. Belbin GM, Rutledge S, Dodatko T, Cullina S, Turchin MC, Kohli S, et al. Leveraging health systems data to characterize a large effect variant conferring risk for liver disease in Puerto Ricans. Am J Hum Genet. 2021 Nov 4;108(11):2099–111.
10. Belbin GM, Cullina S, Wenric S, Soper ER, Glicksberg BS, Torre D, et al. Toward a fine-scale population health monitoring system. Cell [Internet]. 2021 Apr 15 [cited 2022 Jul 14];184(8):2068-2083.e11. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867421003652
11. Dai CL, Vazifeh MM, Yeang CH, Tachet R, Wells RS, Vilar MG, et al. Population Histories of the United States Revealed through Fine-Scale Migration and Haplotype Analysis. The American Journal of Human Genetics. 2020 Mar 5;106(3):371–88.
12. Nakatsuka N, Moorjani P, Rai N, Sarkar B, Tandon A, Patterson N, et al. The promise of discovering population-specific disease-associated genes in South Asia. Nature Genetics 2017 49:9 [Internet]. 2017 Jul 17 [cited 2023 Dec 25];49(9):1403–7. Available from: https://www.nature.com/articles/ng.3917
13. Nait Saada J, Kalantzis G, Shyr D, Cooper F, Robinson M, Gusev A, et al. Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations. Nature Communications 2020 11:1 [Internet]. 2020 Nov 30 [cited 2022 Jul 17];11(1):1–15. Available from: https://www.nature.com/articles/s41467-020-19588-x
14. Han E, Carbonetto P, Curtis RE, Wang Y, Granka JM, Byrnes J, et al. Clustering of 770,000 genomes reveals post-colonial population structure of North America. Nature Communications 2017 8:1. 2017 Feb 7;8(1):1–12.

15. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014 Jan;42(Database issue):D980-5.
16. U.S. Census Bureau. QuickFacts [Internet]. 2022 Jul [cited 2023 Feb 14]. Available from: https://www.census.gov/quickfacts/fact/table/richmondcountynewyork,newyorkcountynewyork,queenscountynewyork,kingscountynewyork,bronxcountynewyork,newyorkcitynewyork/PST045222
17. Lewis ACF, Molina SJ, Appelbaum PS, Dauda B, Di Rienzo A, Fuentes A, et al. Getting genetic ancestry right for science and society. Science (1979). 2022 Apr 15;376(6590):250–2.
18. Kessler MD, Yerges-Armstrong L, Taub MA, Shetty AC, Maloney K, Jeng LJB, et al. Challenges and disparities in the application of personalized genomic medicine to populations with African ancestry. Nature Communications 2016 7:1 [Internet]. 2016 Oct 11 [cited 2023 Dec 25];7(1):1–8. Available from: https://www.nature.com/articles/ncomms12521
19. Kurolap A, Hagin D, Freund T, Fishman S, Zunz Henig N, Brazowski E, et al. CD55-deficiency in Jews of Bukharan descent is caused by the Cromer blood type Dr(a−) variant. Hum Genet. 2023 May 1;142(5):683–90.
20. Lublin DM, Thompson ES, Green AM, Levene C, Telen MJ. Dr(a-) polymorphism of decay accelerating factor. Biochemical, functional, and molecular characterization and production of allele-specific transfectants. J Clin Invest. 1991 Jun;87(6):1945–52.
21. Gregg AR, Aarabi M, Klugman S, Leach NT, Bashford MT, Goldwaser T, et al. Screening for autosomal recessive and X-linked conditions during pregnancy and preconception: a practice resource of the American College of Medical Genetics and Genomics (ACMG). Genetics in Medicine. 2021 Oct 1;23(10):1793–806.
22. Belbin GM, Odgis J, Sorokin EP, Yee MC, Kohli S, Glicksberg BS, et al. Genetic identification of a common collagen disease in puerto ricans via identity-by-descent mapping in a health system. Elife. 2017;6:1–28.
23. Ramirez AH, Sulieman L, Schlueter DJ, Halvorson A, Qian J, Ratsimbazafy F, et al. The All of Us Research Program: Data quality, utility, and diversity. Patterns. 2022 Aug 12;3(8):100570.
24. Baskovich B, Hiraki S, Upadhyay K, Meyer P, Carmi S, Barzilai N, et al. Expanded genetic screening panel for the Ashkenazi Jewish population. Genet Med. 2016 May 1;18(5):522–8.
25. Chiong CM, Reyes-Quintos RTM, Yarza TKL, Tobias-Grasso CAM, Acharya A, Leal SM, et al. The SLC26A4 c.706C>G (p.Leu236Val) Variant is a Frequent Cause of Hearing Impairment in Filipino Cochlear Implantees. Otology and Neurotology. 2018 Sep 1;39(8):E726–30.
26. Hagin D, Lahav D, Freund T, Shamai S, Brazowski E, Fishman S, et al. Eculizumab-Responsive Adult Onset Protein Losing Enteropathy, Caused by Germline CD55-Deficiency and Complicated by Aggressive Angiosarcoma. J Clin Immunol. 2021 Feb 1;41(2):477–81.
27. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015 May;17(5):405–24.

28. Yang S, Lincoln SE, Kobayashi Y, Nykamp K, Nussbaum RL, Topper S. Sources of discordance among germ-line variant classifications in ClinVar. Genet Med. 2017 Oct;19(10):1118–26.
29. Manrai AK, Funke BH, Rehm HL, Olesen MS, Maron BA, Szolovits P, et al. Genetic Misdiagnoses and the Potential for Health Disparities. N Engl J Med. 2016 Aug 18;375(7):655–65.
30. Shah N, Hou YCC, Yu HC, Sainger R, Caskey CT, Venter JC, et al. Identification of Misclassified ClinVar Variants via Disease Population Prevalence. Am J Hum Genet [Internet]. 2018 Apr 5 [cited 2023 Dec 25];102(4):609–19. Available from: http://www.cell.com/article/S0002929718300879/fulltext
31. Whiffin N, Minikel E, Walsh R, O'Donnell-Luria AH, Karczewski K, Ing AY, et al. Using high-resolution variant frequencies to empower clinical genome interpretation. Genetics in Medicine. 2017 Oct;19(10):1151–8.
32. Gusmano MK, Rodwin VG, Weisz D. Persistent Inequalities in Health and Access to Health Services: Evidence From New York City. World Med Health Policy. 2017 Jun 12;9(2):186–205.
33. Caraballo C, Ndumele CD, Roy B, Lu Y, Riley C, Herrin J, et al. Trends in Racial and Ethnic Disparities in Barriers to Timely Medical Care Among Adults in the US, 1999 to 2018. JAMA Health Forum. 2022 Oct 7;3(10):e223856.
34. All of Us Research Program. Policy on Stigmatizing Research [Internet]. 2020. Available from: https://www.researchallofus.org/wp-content/themes/research-hub-wordpress-theme/media/2020/05/AoU_Policy_Stigmatizing_Research_508.pdf
35. Using Population Descriptors in Genetics and Genomics Research. Washington, D.C.: National Academies Press; 2023.
36. Chiu AM, Molloy EK, Tan Z, Talwalkar A, Sankararaman S. Inferring population structure in biobank-scale genomic data. Am J Hum Genet. 2022 Apr 7;109(4):727–37.
37. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015 Dec 25;4(1):7.
38. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.
39. Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, et al. Insights into human genetic variation and population history from 929 diverse genomes. Science (1979). 2020 Mar 20;367(6484).
40. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016;538(7624):201–6.
41. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19(9):1655–64.
42. Browning BL, Tian X, Zhou Y, Browning SR. Fast two-stage phasing of large-scale sequence data. Am J Hum Genet. 2021 Oct 7;108(10):1880–90.
43. Freyman WA, Mcmanus KF, Shringarpure SS, Jewett EM, Bryc K, Auton A. Fast and Robust Identity-by-Descent Inference with the Templated Positional Burrows–Wheeler Transform. Mol Biol Evol. 2021 May 4;38(5):2131–51.
44. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal. 2006;Complex Sy:1695.

661 45. Rosvall M, Bergstrom CT. Maps of random walks on complex networks reveal
662 community structure. Proceedings of the National Academy of Sciences. 2008 Jan
663 29;105(4):1118–23.
664 46. Nakatsuka N, Moorjani P, Rai N, Sarkar B, Tandon A, Patterson N, et al. The promise of
665 discovering population-specific disease-associated genes in South Asia. Nat Genet. 2017
666 Sep 17;49(9):1403–7.
667 47. Chen S, Francioli LC, Goodrich JK, Collins RL, Wang Q, Alföldi J, et al. A genome-wide
668 mutational constraint map quantified from variation in 76,156 human genomes. bioRxiv
669 [Internet]. 2022 Oct 10 [cited 2023 May 31];2022.03.20.485034. Available from:
670 https://www.biorxiv.org/content/10.1101/2022.03.20.485034v2
671 48. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: A Discriminative Modeling
672 Approach for Rapid and Robust Local-Ancestry Inference. The American Journal of
673 Human Genetics. 2013 Aug 8;93(2):278–88.
674

675

676 **Supporting information captions**

677

678 **Text S1: Supplementary methods**

679 **Figure S1: Ancestry background of AoU IBD clusters.**

680 **Figure S2: Comparison of NYC Census data in July 2022 and AoU NYC participants.**

681 **Figure S3: PCA plot for AoU NYC participants (a), BioMe (b) and global reference**
682 **populations.**

683 **Figure S4: 16 IBD groups in the combined dataset of NYC.**

684 Included in TextS1_FigureSs.docx

685

686 **Table S1: ancestry background assignment to IBD groups**

687 **Table S2: All candidate founder P/LP variants in the eight founder populations in NYC**

688 **Table S3: Frequencies of Caribbean founder variants from Table 1 in other shared ancestry**
689 **IBD groups**

690 Included in SupTable.xlsx

691