

1 **Title**

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

3

4 **Authors**

5 Mariko Isshiki PhD,<sup>1</sup> Anthony Griffen BSc,<sup>2</sup> Paul Meissner MSPH,<sup>3, 4</sup> Paulette Spencer MPH,<sup>5</sup>

6 Michael D. Cabana MD,<sup>6</sup> Susan D. Klugman MD,<sup>4</sup> Mirtha Colón MSW,<sup>7</sup> Zoya Maksumova MD,<sup>8</sup>

7 Shakira Suglia ScD,<sup>10</sup> Carmen Isasi MD,<sup>6,9</sup> John M. Grealley DMed,<sup>1,6,†</sup> Srilakshmi M. Raj PhD<sup>1,†</sup>

8

9 **Affiliations**

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

13 <sup>5</sup>Bronx Community Health Network, One Fordham Plaza, Suite 1108, Bronx, NY 10458

14 <sup>7</sup>Hondurans Against AIDS/Casa Yurumein, 324 E 151st St, Bronx, NY 10451

15 <sup>8</sup>10310 Fuerte Drive, La Mesa, CA 91941

16 <sup>10</sup>Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA  
17 30322

18 †co-corresponding authors

19 Address reprint requests to: Srilakshmi Raj [srilakshmi.raj@einsteinmed.edu](mailto:srilakshmi.raj@einsteinmed.edu)

20 Contact information for corresponding authors: Srilakshmi Raj [srilakshmi.raj@einsteinmed.edu](mailto:srilakshmi.raj@einsteinmed.edu)

21 John Grealley [john.grealley@einsteinmed.edu](mailto:john.grealley@einsteinmed.edu)

22

23

24 **Abstract**

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

44

45

46

47

48 **Author Summary**

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

61

## 62 **Introduction**

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

87 genetic ancestry groups, particularly those with a founder effect. We were able to define groups  
88 with increased 'genetic similarity'(2) and characterize population structure by identifying segments  
89 of DNA that are shared among individuals due to inheritance from a common ancestor, also called  
90 identity-by-descent (IBD). This has previously been performed successfully in cohorts of different  
91 genetic ancestries (10–14) with implications for understanding population-specific disease risk  
92 (9,10,13) Here we explicitly test the association between population structure and disease risk by  
93 focusing on population-specific enrichment of variants curated as disease-causing in the ClinVar  
94 database (15) The results show how health systems and providers can benefit from recognizing  
95 rare diseases in the populations they serve, including the potential benefits of early detection of  
96 rare diseases as well as prenatal carrier screening in these communities, and the targeted use of  
97 specific therapies.

98

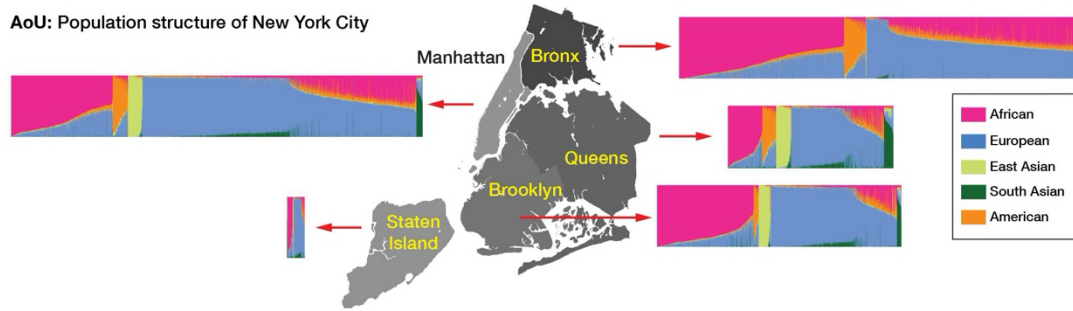
## 99 **Results**

### 100 **The population structure of NYC participants of the All of Us Research Program**

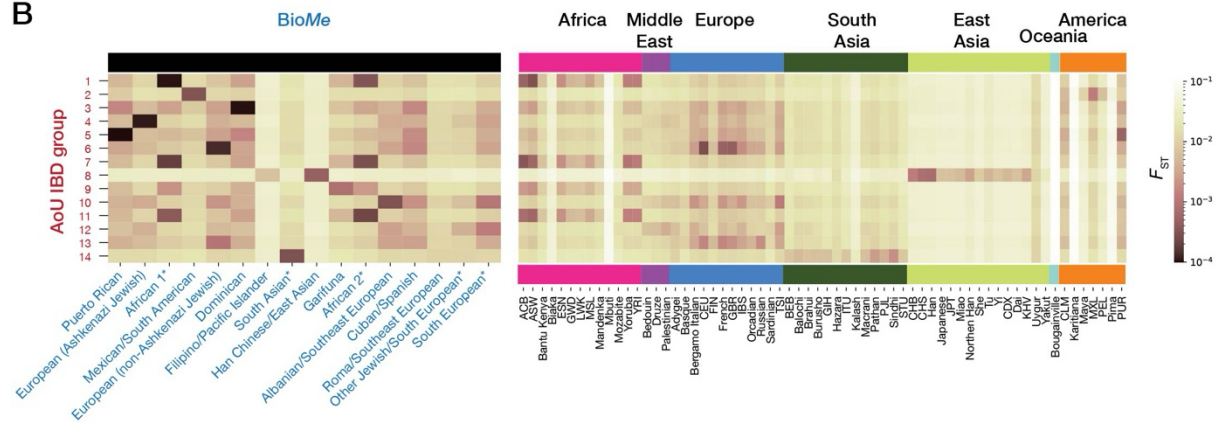
101 We studied the genetic diversity of NYC residents using genetic data from 13,817 participants of  
102 the AoU Research Program. This dataset excludes 'related' individuals, those who are second  
103 cousins or closer. We found that the AoU cohort in NYC is diverse across the five boroughs (**Fig.**  
104 **1a, Figure S1**), and that the proportions of self-reported race/ethnicity information for each  
105 borough are comparable to those from census data (**Figure S2a**) (16). Of the boroughs,  
106 Manhattan and the Bronx are over-represented (**Figure S2b**). Admixed individuals accounted for

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

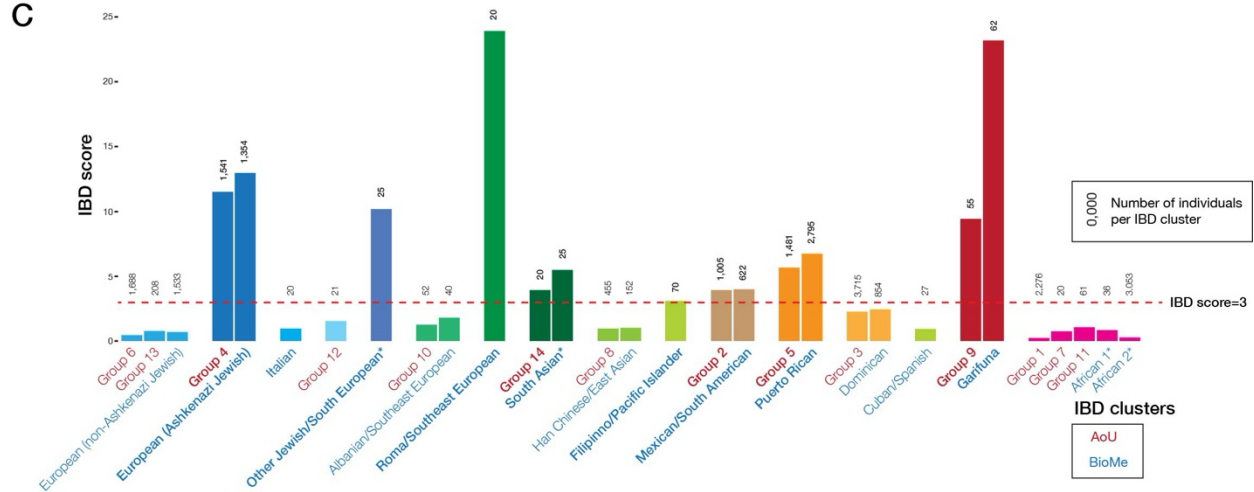
A



B



C



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, BioMe 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

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

118

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

130

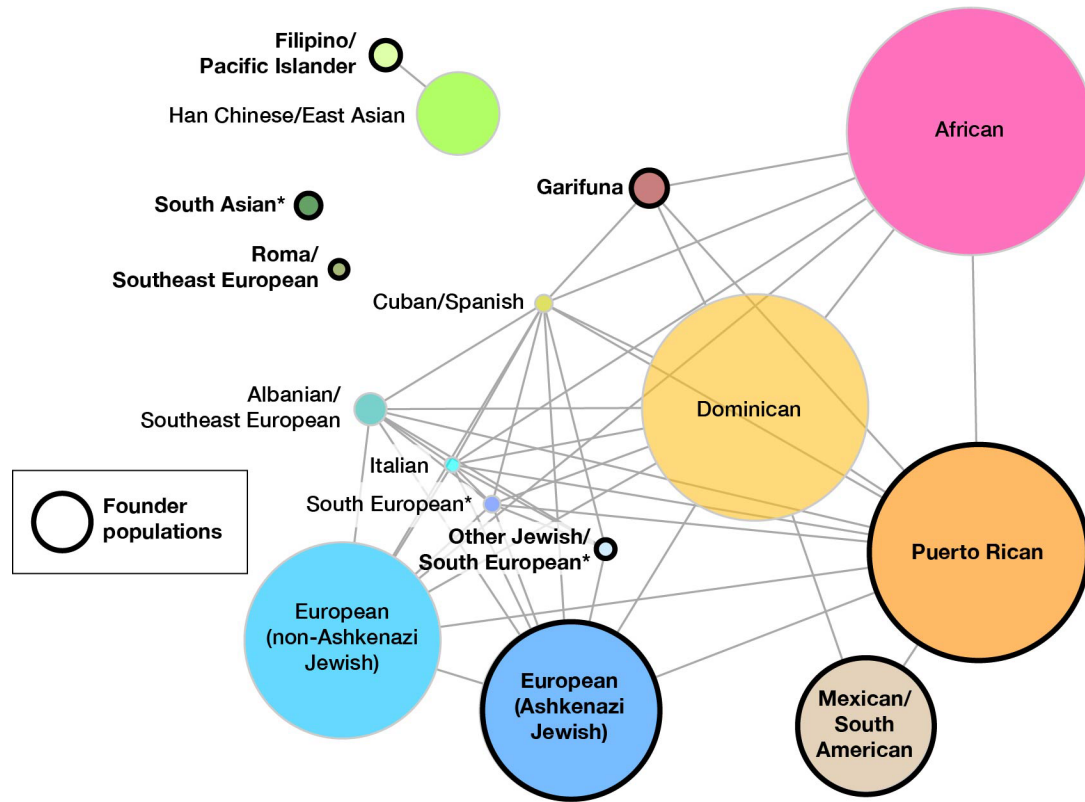
131 Of the IBD groups, there were five from AoU and eight from BioMe with IBD scores  $> 3$ , defining  
132 them as founder populations (**Fig. 1c**). Of the eight founder groups from BioMe, founder effects  
133 in Ashkenazi Jewish, Puerto Rican and Garifuna were reported previously (9,10). We also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

137

### 138 **Detection of founder populations in NYC**

139 Since our AoU dataset and BioMe are both NYC cohorts with shared genetic features (**Figs. 1b-**  
140 **c, Figure S3**), we combined the IBD groups from AoU and BioMe with pairwise Hudson's  $F_{ST}$   
141 values  $< 0.001$ , resulting in 16 IBD groups which we annotated based on inferred ancestry (**Fig.**  
142 **2**). Eight groups were identified as founder populations (IBD score  $>3$ ). The populations were  
143 named based on self-defined ancestry as provided by the reference datasets, when available  
144 (**Figs. 1b-c, 2, 3, Table S1**). Genetic ancestry also acts on a continuum(17), therefore some IBD  
145 groups appeared to be more discrete (*e.g.* Garifuna, Puerto Rican), whereas others include  
146 individuals across a broader geographic range (*e.g.* Filipino/Pacific Islander, Mexican/South  
147 American, Han Chinese/East Asian). In situations where groups could not be confidently labeled,  
148 the closest associated population group was used as a placeholder label until more genomic  
149 reference datasets become available. These populations are labeled with an asterisk (**Figs. 1b-**  
150 **c, 2, 3**).





151

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

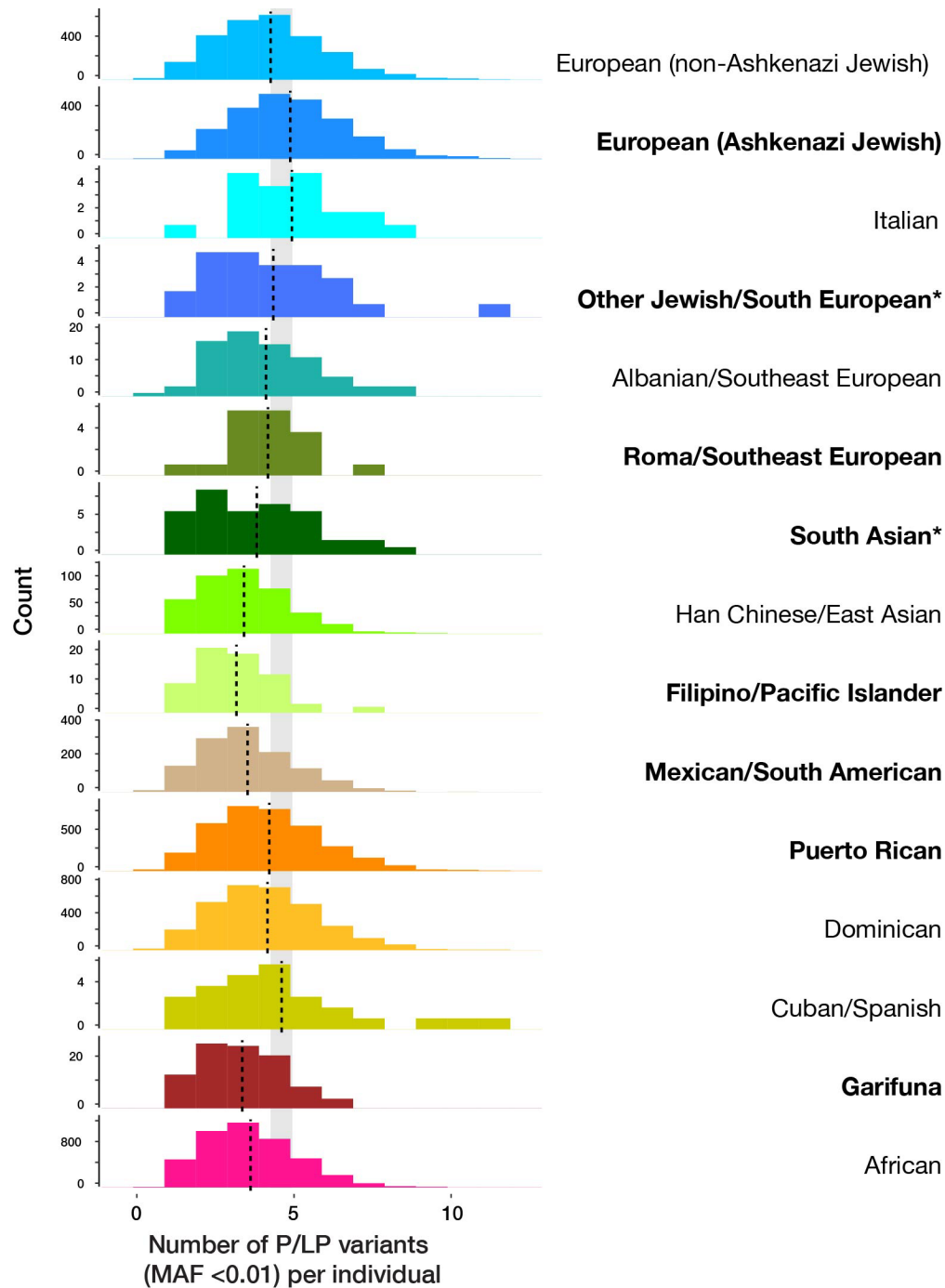
153 A network depiction of the IBD groups based on  $F_{ST}$  between IBD groups. Edges were weighted  
154 by logarithm of  $F_{ST}$ . Only edges representing  $F_{ST} < 0.01$  are shown, with founder populations  
155 circled in black. Circle sizes reflect the number of individuals in each IBD group.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

157

### 158 Identification of pathogenic founder variants

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

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



168

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

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

173 The 'South European' IBD group only included WGS data for fewer than 20 individuals and was  
174 removed due to the All of Us Data and Statistics Dissemination Policy. The shaded gray rectangle  
175 represents the range of mean values for the uppermost three European groups, highlighting the  
176 lower number of ClinVar variants annotated in those of Asian, American and African ancestries.  
177 Asterisk next to labels represent populations with inadequate reference information for annotation.  
178

179 We detected 674 unique P/LP variants significantly enriched across the 8 founder populations  
180 (Fisher's Exact  $p > 0.05$ ) (**Table S2**) with 202 of these variants passing Bonferroni correction. Of  
181 the 674 variants, 478 variants have two or more ClinVar gold stars, meaning variants are from  
182 practice guidelines, reviewed by expert panel, or from multiple submitters with evidence and no  
183 classification conflicts. **Table S** Those variants with no ClinVar gold stars should in general be  
184 interpreted with caution, such as the *KRT18* variant not previously described to be common in  
185 Ashkenazi Jewish individuals. The *CD55* variant associated with protein-losing enteropathy (19)  
186 and shown in cell studies to cause loss of CD55 on the cell surface (20) also lacks a star rating,  
187 illustrating how the absence of this rating should not be used to exclude variants as disease-  
188 causing. This *CD55* variant also does not pass Bonferroni correction, nor do known founder effect  
189 variants in *HBB* in the South Asian group [REF] and a *SLC26A4* variant in the Filipino/Pacific  
190 Islander group [REF], prompting us to include variants that do not pass multiple testing correction  
191 in **Table 1** as candidate founder disease-causing variants in the IBD groups. We identified 51  
192 variants from this broader list that have minor allele frequencies of  $> 0.005$  in one or more IBD  
193 groups, the Tier 2 threshold for inclusion into prenatal screening panels (21). Of these, 25 are  
194 new, previously unrecognized founder effect variants (**Table 1**). The results shown in **Fig. 3**  
195 support the likelihood that the numbers of P/LP variants in non-European groups are likely to  
196 represent an underestimate, and that more disease-causing variants remain to be discovered in  
197 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)
<b>European (Ashkenazi Jewish)</b>						
<i>F11</i>	VCV000011892	NM_000128.4(F11):c.901T>C (p.Phe301Leu)	2	Hereditary factor XI deficiency	<b>0.0248</b>	2813350
<i>GJB2</i>	VCV000017010	NM_004004.6(GJB2):c.167del (p.Leu56fs)	3	Autosomal recessive nonsyndromic hearing loss	<b>0.0157</b>	9819448
<i>ELP1</i>	VCV000006085	NM_003640.5(ELP1):c.2204+6T>C	2	Familial dysautonomia	<b>0.0154</b>	12116234
<i>HEXA</i>	VCV000003889	NM_000520.6(HEXA):c.1274_1277dup (p.Tyr427fs)	2	Tay-Sachs disease	<b>0.0142</b>	2848800
<i>F11</i>	VCV000011891	NM_000128.4(F11):c.403G>T (p.Glu135Ter)	2	Hereditary factor XI deficiency	<b>0.0140</b>	2813350
<i>KRT18</i>	VCV000014585	NM_000224.3(KRT18):c.383A>T (p.His128Leu)	none	Cirrhosis	<b>0.0098</b>	Not described
<i>CFTR</i>	VCV000007129	NM_000492.4(CFTR):c.3846G>A (p.Trp1282Ter)	4	Cystic fibrosis	<b>0.0094</b>	1384328
<i>MPL</i>	VCV000135563	NM_005373.3(MPL):c.79+2T>A	2	Congenital amegakaryocytic thrombocytopenia	<b>0.0067</b>	21489838
<i>DDX11</i>	VCV000252749	NM_030653.4(DDX11):c.1763-1G>C	2	Warsaw breakage syndrome	<b>0.0063</b>	31287223
<i>ASPA</i>	VCV000002605	NM_000049.4(ASPA):c.854A>C (p.Glu285Ala)	2	Canavan disease	<b>0.0058</b>	8252036
<i>SLC3A1</i>	VCV0000336195	NM_000341.4(SLC3A1):c.808C>T (p.Arg270Ter)	2	Cystinuria	<b>0.0054</b>	7539209
<i>FKTN</i>	VCV000003203	NM_001079802.2(FKTN):c.1167dup (p.Phe390fs)	2	Walker-Warburg congenital muscular dystrophy	<b>0.0054</b>	19266496
<i>CLRN1</i>	VCV000004395	NM_174878.3(CLRN1):c.144T>G (p.Asn48Lys)	2	Usher syndrome type 3A	<b>0.0052</b>	12145752
<i>PAH</i>	VCV000102706	NM_000277.3(PAH):c.506G>A (p.Arg169His)	3	Phenylketonuria	<b>0.0050</b>	29144512
<i>DLD</i>	VCV000011966	NM_000108.5(DLD):c.685G>T (p.Gly229Cys)	2	Pyruvate dehydrogenase E3 deficiency	<b>0.0050</b>	14765544
<i>FANCC</i>	VCV000012045	NM_000136.3(FANCC):c.456+4A>T	2	Fanconi anemia complementation group C	<b>0.0050</b>	8348157
<b>Other Jewish/South European*</b>						
<i>FMO3</i>	VCV000985096	NM_001002294.3(FMO3):c.1499G>A (p.Arg500Gln)	1	Trimethylaminuria	<b>0.0600</b>	Not described
<i>CD55</i>	VCV000431759	NM_000574.5(CD55):c.596C>T (p.Ser199Leu)	none	Complement hyperactivation, angiopathic thrombosis, and protein-losing enteropathy	0.0400	35314883
<i>KLKB1</i>	VCV000012033	NM_000892.5(KLKB1):c.337C>T (p.Arg113Ter)	none	Prekallikrein deficiency	0.0400	Not described
<b>Puerto Rican</b>						
<i>HPS1</i>	VCV000005277	NM_000195.5(HPS1):c.1472_1487dup (p.His497fs)	2	Hermansky-Pudlak syndrome 1	<b>0.0097</b>	8896559
<i>RSPH4A</i>	VCV000088863	NM_001010892.3(RSPH4A):c.921+3_921+6del	2	Primary ciliary dyskinesia	<b>0.0096</b>	23798057
<i>COL27A1</i>	VCV000143245	NM_032888.4(COL27A1):c.2089G>C (p.Gly697Arg)	2	Steel syndrome	<b>0.0091</b>	24986830
<i>TBCK</i>	VCV000225235	NM_001163435.3(TBCK):c.376C>T (p.Arg126Ter)	2	Infantile hypotonia, infantile with psychomotor retardation and characteristic facies	<b>0.0089</b>	29283439

It is made available under a [CC-BY-NC-ND 4.0 International license](https://creativecommons.org/licenses/by-nc-nd/4.0/) .

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

201

## 202 **TABLE LEGEND**

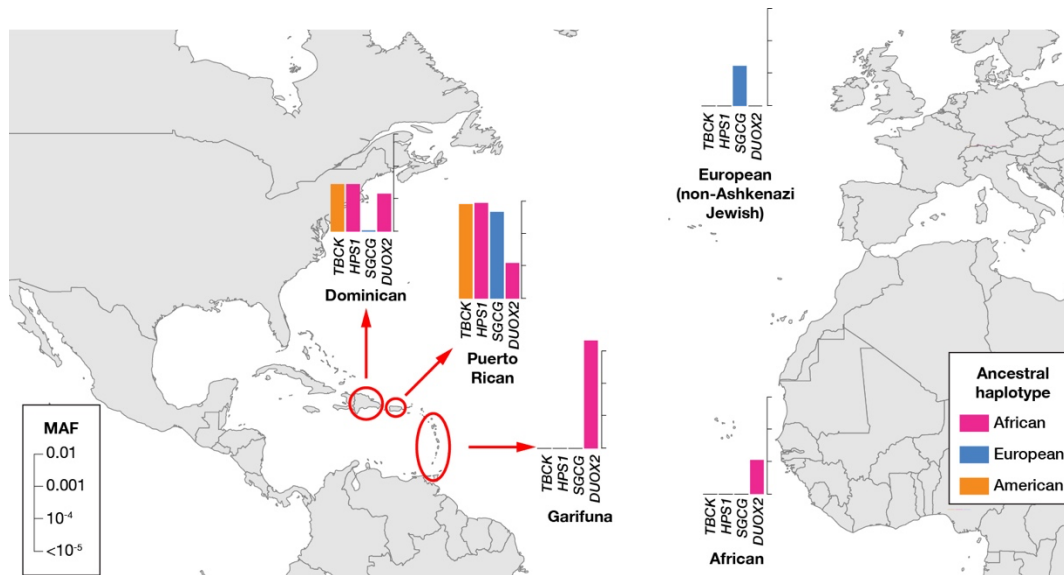
203 The table shows the P/LP variants that occur within each population at a frequency of at least  
204 1/200 alleles, the threshold for inclusion in prenatal testing. Those with allele frequencies in bold  
205 type pass Bonferroni correction. Of these 51 variants, 26 have already been published as founder  
206 mutations, especially in the European (Ashkenazi Jewish) and Puerto Rican populations. The  
207 other 25 are new, previously unrecognized founder effect variants. Abbreviations: PMID, PubMed  
208 reference number. An asterisk next to labels represent populations with inadequate reference  
209 information for annotation. Some of the allele frequency counts displayed here represent <20 AoU  
210 participants. The AoU Resource Access Board reviewed this work and granted an exception to  
211 their Data and Statistics Dissemination policy to report these frequencies.

212

### 213 **Ancestry analysis for shared founder variants in individuals of Caribbean ancestry**

214 We detected 12 and 9 founder P/LP variants for Puerto Rican and Garifuna IBD groups, including  
215 variants that did not pass Bonferroni correction, respectively (**Table 1**). Of these 21 variants, 15  
216 were also detected at lower frequencies in other IBD groups (**Table S3**). To test whether this was  
217 due to shared ancestry, we inferred local ancestry (the origin of the DNA containing each P/LP  
218 variant) in the Caribbean individuals to identify the ancestral population in which each P/LP variant  
219 arose originally. We found the majority of the founder variants in Puerto Ricans to be located on  
220 haplotypes of European ancestry, with the remaining founder variants located on African and  
221 Indigenous American haplotypes (**Table S3**). Of these, the origin of the *COL27A1* Steel  
222 syndrome variant on chromosome 9 has also previously been characterized as Native American  
223 ancestry (22) We illustrate the sharing of founder effect mutations across Caribbean groups and  
224 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**



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

261

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

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

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

286 The potential that the recognition of the presence of a rare disease in a founder population can  
287 lead to targeted therapies is exemplified by the *CD55* variant in the Bukharian Jewish community,  
288 which can cause a spectrum of presentations from mild abdominal discomfort following a high-fat  
289 meal to a severe syndrome including protein-losing enteropathy, and is effectively treated by the  
290 complement C5-inhibitor eculizumab (26) Implementation in health care systems of information  
291 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

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

328

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

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

346

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

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

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

369

## 370 **Materials and Methods**

### 371 **Research participants and dataset preparation**

372 Participants living in NYC in the AoU Program version 6 curated data repository (8) were identified  
373 by the first three digits of their Zip Code, allowing borough-level resolution of geographic  
374 residence. Microarray genotype data were used to assess the population structure of NYC  
375 participants by principal component analysis (PCA), by global ancestry analysis as performed by  
376 SCOPE (36) and by Identity-by-descent (IBD) analysis as described below. Samples were QCed  
377 by call rate and kinship coefficient using PLINK v2.00a2.3LM (37) No individuals were filtered out  
378 by the call rate threshold of 0.9 (`---mind 0.1`). To remove close relatives, either of the pairs of  
379 individuals who showed king kinship coefficients  $> 0.125$  were removed using `--king-cutoff 0.125`  
380 in PLINK2.0 (37). Variants of the array data were filtered with the following conditions using  
381 PLINK2.0: minor allele frequency  $> 0.01$ , genotyping rate per site  $> 0.95$ , and p-value for the  
382 departures from Hardy Weinberg Equilibrium (HWE)  $> 1 \times 10^{-6}$  (`--maf 0.01 --geno 0.05 --snps-only`  
383 `--hwe 1e-06`). After QC steps, 13,817 participants and 720,630 SNPs remained for downstream  
384 analysis.

385 Out of these individuals, 10,381 individuals had whole genome sequence (WGS) data available.  
386 We used this subset of individuals and whole genome sequence data from an independent NYC  
387 biobank, the Mount Sinai BioMe biobank (9,10), to identify founder pathogenic variants. Approval  
388 to study these de-identified data was granted by the Albert Einstein College of Medicine Internal  
389 Review Board (Protocol 2016-7099). All analyses on AoU participants included in the manuscript  
390 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/richmondcountynyork,newyorkcountynyork,q>  
398 [ueenscountynyork,kingscountynyork,bronxcountynyork,newyorkcitynyork/PST04522](https://www.census.gov/quickfacts/fact/table/richmondcountynyork,newyorkcountynyork,q)  
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  $\geq 12$  cM and  $\leq 72$  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**



439 Since AoU did not provide the detailed ethnicity information that is particularly useful for defining  
440 founder groups, we inferred population ancestry (e.g. Puerto Rican, Dominican, Ashkenazi  
441 Jewish) in AoU IBD groups by estimating Hudson's  $F_{ST}$  between each group and populations in  
442 genomic reference panels using PLINK2.0. The reference panel included 14,985 individuals and  
443 140,952 biallelic SNPs from global populations with sample sizes > 10 individuals in 1KGP, HGDP  
444 and SGDP together with IBD groups in BioMe. We also conducted PCA for the merged dataset.  
445 IBD groups in AoU and BioMe with  $F_{ST}$  values < 0.001 were combined in further analyses.

446

#### 447 **Detection of pathogenic founder variants for rare diseases**

448 We extracted variants categorized as pathogenic or likely pathogenic (P/LP) in the ClinVar  
449 database (15) (version ClinVarFullRelease\_2023-01.xml) from WGS data of 10,381 NYC AoU  
450 participants and 11,549 BioMe participants. Of the 193,935 P/LP variants registered in ClinVar  
451 as of Jan 7, 2023, we detected 27,125 variants in our NYC cohort. We removed close relatives and  
452 excluded variants that appeared only once in the NYC dataset. We then filtered variants with  
453 genotype rate < 0.9 and p-values for departures from Hardy-Weinberg Equilibrium (HWE) <  $1 \times 10^{-16}$ .  
454 We set a small HWE threshold anticipating that rare variants may likely diverge from HWE  
455 due to high heterozygosity. After filtering, 3,616 P/LP variants were observed in NYC individuals.  
456 HGVS description and review status (gold stars) were obtained from variant\_summary.txt.gz in  
457 <https://ftp.ncbi.nlm.nih.gov/pub/clinvar> last updated on March 30, 2024. The variants which were  
458 not classified as P/LP as of March 30, 2024 were removed from results.

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

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

#### 469 **Ancestry analysis for the founder variant in the Caribbean IBD groups**

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

476

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

485

## 486 **ACKNOWLEDGEMENTS**

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

496

497 The All of Us Research Program is supported by the National Institutes of Health, Office of the  
498 Director: Regional Medical Centers: 1 OT2 ODO26549; 1 OT2 ODO26554; 1 OT2 ODO26557;  
499 1 OT2 ODO26556; 1 OT2 ODO26550; 1 OT2 ODO26552; 1 OT2 ODO26553; 1 OT2 ODO26548;  
500 1 OT2 ODO26548; 1 OT2 ODO2551; 1 OT2 ODO26555; IAA #: AOD 16037; Federally  
501 Qualified Health Centers: HHSN 263201600085U; Data and Research Center: 5 U2C  
502 ODO23196; Biobank: 1 U24 OD023121; The Participant Center: U24 ODO23176; Participant  
503 Technology Systems Center: 1 U24 ODO23163; Communications and Engagement: 3 OT2  
504 ODO23205; 3 OT2 ODO23206; and Community Partners: 1 OT2 ODO25277; 3 OT2  
505 ODO25315; 1 OT2 ODO25337; 1 OT2 ODO25276.

506

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

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

517

#### 518 **Data availability**

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

523

524

## 525 References

- 526 1. Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, et al.  
527 Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet  
528 database. *European Journal of Human Genetics* 2019 28:2. 2019 Sep 16;28(2):165–73.
- 529 2. National Academies of Sciences, Engineering, and Medicine. *Using Population*  
530 *Descriptors in Genetics and Genomics Research*. Washington, D.C.: National Academies  
531 Press; 2023.
- 532 3. Scott SA, Edelman L, Liu L, Luo M, Desnick RJ, Kornreich R. Experience with carrier  
533 screening and prenatal diagnosis for 16 Ashkenazi Jewish genetic diseases. *Hum Mutat*.  
534 2010 Nov;31(11):1240–50.
- 535 4. Gross SJ, Pletcher BA, Monaghan KG. Carrier screening in individuals of Ashkenazi  
536 Jewish descent. *Genetics in Medicine*. 2008 Jan;10(1):54.
- 537 5. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies.  
538 *Cell*. 2019;177(1):26–31.
- 539 6. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current  
540 polygenic risk scores may exacerbate health disparities. *Nat Genet*. 2019 Apr  
541 29;51(4):584–91.
- 542 7. Fox K. The Illusion of Inclusion — The “All of Us” Research Program and Indigenous  
543 Peoples’ DNA. *New England Journal of Medicine*. 2020 Jul 30;383(5):411–3.
- 544 8. All of Us Research Program Investigators, Denny JC, Rutter JL, Goldstein DB,  
545 Philippakis A, Smoller JW, et al. The “All of Us” Research Program. *N Engl J Med*. 2019  
546 Aug 15;381(7):668–76.
- 547 9. Belbin GM, Rutledge S, Dodatko T, Cullina S, Turchin MC, Kohli S, et al. Leveraging  
548 health systems data to characterize a large effect variant conferring risk for liver disease in  
549 Puerto Ricans. *Am J Hum Genet*. 2021 Nov 4;108(11):2099–111.
- 550 10. Belbin GM, Cullina S, Wenric S, Soper ER, Glicksberg BS, Torre D, et al. Toward a fine-  
551 scale population health monitoring system. *Cell [Internet]*. 2021 Apr 15 [cited 2022 Jul  
552 14];184(8):2068-2083.e11. Available from:  
553 <https://linkinghub.elsevier.com/retrieve/pii/S0092867421003652>
- 554 11. Dai CL, Vazifeh MM, Yeang CH, Tachet R, Wells RS, Vilar MG, et al. Population  
555 Histories of the United States Revealed through Fine-Scale Migration and Haplotype  
556 Analysis. *The American Journal of Human Genetics*. 2020 Mar 5;106(3):371–88.
- 557 12. Nakatsuka N, Moorjani P, Rai N, Sarkar B, Tandon A, Patterson N, et al. The promise of  
558 discovering population-specific disease-associated genes in South Asia. *Nature Genetics*  
559 2017 49:9 [Internet]. 2017 Jul 17 [cited 2023 Dec 25];49(9):1403–7. Available from:  
560 <https://www.nature.com/articles/ng.3917>
- 561 13. Nait Saada J, Kalantzis G, Shyr D, Cooper F, Robinson M, Gusev A, et al. Identity-by-  
562 descent detection across 487,409 British samples reveals fine scale population structure  
563 and ultra-rare variant associations. *Nature Communications* 2020 11:1 [Internet]. 2020  
564 Nov 30 [cited 2022 Jul 17];11(1):1–15. Available from:  
565 <https://www.nature.com/articles/s41467-020-19588-x>
- 566 14. Han E, Carbonetto P, Curtis RE, Wang Y, Granka JM, Byrnes J, et al. Clustering of  
567 770,000 genomes reveals post-colonial population structure of North America. *Nature*  
568 *Communications* 2017 8:1. 2017 Feb 7;8(1):1–12.

- 569 15. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar:  
570 public archive of relationships among sequence variation and human phenotype. *Nucleic*  
571 *Acids Res.* 2014 Jan;42(Database issue):D980-5.
- 572 16. U.S. Census Bureau. QuickFacts [Internet]. 2022 Jul [cited 2023 Feb 14]. Available from:  
573 [https://www.census.gov/quickfacts/fact/table/richmondcountynyork,newyorkcountynyork,](https://www.census.gov/quickfacts/fact/table/richmondcountynyork,newyorkcountynyork,queenscountynyork,kingscountynyork,bronxcountynyork,newyorkcitynewyork/PST045222)  
574 [queenscountynyork,kingscountynyork,bronxcountynyork,newyorkcitynew](https://www.census.gov/quickfacts/fact/table/richmondcountynyork,newyorkcountynyork,queenscountynyork,kingscountynyork,bronxcountynyork,newyorkcitynewyork/PST045222)  
575 [york/PST045222](https://www.census.gov/quickfacts/fact/table/richmondcountynyork,newyorkcountynyork,queenscountynyork,kingscountynyork,bronxcountynyork,newyorkcitynewyork/PST045222)
- 576 17. Lewis ACF, Molina SJ, Appelbaum PS, Dauda B, Di Rienzo A, Fuentes A, et al. Getting  
577 genetic ancestry right for science and society. *Science* (1979). 2022 Apr  
578 15;376(6590):250–2.
- 579 18. Kessler MD, Yerges-Armstrong L, Taub MA, Shetty AC, Maloney K, Jeng LJB, et al.  
580 Challenges and disparities in the application of personalized genomic medicine to  
581 populations with African ancestry. *Nature Communications* 2016 7:1 [Internet]. 2016 Oct  
582 11 [cited 2023 Dec 25];7(1):1–8. Available from:  
583 <https://www.nature.com/articles/ncomms12521>
- 584 19. Kurolap A, Hagin D, Freund T, Fishman S, Zunz Henig N, Brazowski E, et al. CD55-  
585 deficiency in Jews of Bukharan descent is caused by the Cromer blood type Dr(a-)   
586 variant. *Hum Genet.* 2023 May 1;142(5):683–90.
- 587 20. Lublin DM, Thompson ES, Green AM, Levene C, Telen MJ. Dr(a-) polymorphism of  
588 decay accelerating factor. Biochemical, functional, and molecular characterization and  
589 production of allele-specific transfectants. *J Clin Invest.* 1991 Jun;87(6):1945–52.
- 590 21. Gregg AR, Aarabi M, Klugman S, Leach NT, Bashford MT, Goldwaser T, et al.  
591 Screening for autosomal recessive and X-linked conditions during pregnancy and  
592 preconception: a practice resource of the American College of Medical Genetics and  
593 Genomics (ACMG). *Genetics in Medicine.* 2021 Oct 1;23(10):1793–806.
- 594 22. Belbin GM, Odgis J, Sorokin EP, Yee MC, Kohli S, Glicksberg BS, et al. Genetic  
595 identification of a common collagen disease in puerto ricans via identity-by-descent  
596 mapping in a health system. *Elife.* 2017;6:1–28.
- 597 23. Ramirez AH, Sulieman L, Schlueter DJ, Halvorson A, Qian J, Ratsimbazafy F, et al. The  
598 All of Us Research Program: Data quality, utility, and diversity. *Patterns.* 2022 Aug  
599 12;3(8):100570.
- 600 24. Baskovich B, Hiraki S, Upadhyay K, Meyer P, Carmi S, Barzilai N, et al. Expanded  
601 genetic screening panel for the Ashkenazi Jewish population. *Genet Med.* 2016 May  
602 1;18(5):522–8.
- 603 25. Chiong CM, Reyes-Quintos RTM, Yarza TKL, Tobias-Grasso CAM, Acharya A, Leal  
604 SM, et al. The SLC26A4 c.706C>G (p.Leu236Val) Variant is a Frequent Cause of  
605 Hearing Impairment in Filipino Cochlear Implantees. *Otology and Neurotology.* 2018 Sep  
606 1;39(8):E726–30.
- 607 26. Hagin D, Lahav D, Freund T, Shamai S, Brazowski E, Fishman S, et al. Eculizumab-  
608 Responsive Adult Onset Protein Losing Enteropathy, Caused by Germline CD55-  
609 Deficiency and Complicated by Aggressive Angiosarcoma. *J Clin Immunol.* 2021 Feb  
610 1;41(2):477–81.
- 611 27. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and  
612 guidelines for the interpretation of sequence variants: a joint consensus recommendation  
613 of the American College of Medical Genetics and Genomics and the Association for  
614 Molecular Pathology. *Genet Med.* 2015 May;17(5):405–24.



- 615 28. Yang S, Lincoln SE, Kobayashi Y, Nykamp K, Nussbaum RL, Topper S. Sources of  
616 discordance among germ-line variant classifications in ClinVar. *Genet Med*. 2017  
617 Oct;19(10):1118–26.
- 618 29. Manrai AK, Funke BH, Rehm HL, Olesen MS, Maron BA, Szolovits P, et al. Genetic  
619 Misdiagnoses and the Potential for Health Disparities. *N Engl J Med*. 2016 Aug  
620 18;375(7):655–65.
- 621 30. Shah N, Hou YCC, Yu HC, Sainger R, Caskey CT, Venter JC, et al. Identification of  
622 Misclassified ClinVar Variants via Disease Population Prevalence. *Am J Hum Genet*  
623 [Internet]. 2018 Apr 5 [cited 2023 Dec 25];102(4):609–19. Available from:  
624 <http://www.cell.com/article/S0002929718300879/fulltext>
- 625 31. Whiffin N, Minikel E, Walsh R, O'Donnell-Luria AH, Karczewski K, Ing AY, et al.  
626 Using high-resolution variant frequencies to empower clinical genome interpretation.  
627 *Genetics in Medicine*. 2017 Oct;19(10):1151–8.
- 628 32. Gusmano MK, Rodwin VG, Weisz D. Persistent Inequalities in Health and Access to  
629 Health Services: Evidence From New York City. *World Med Health Policy*. 2017 Jun  
630 12;9(2):186–205.
- 631 33. Caraballo C, Ndumele CD, Roy B, Lu Y, Riley C, Herrin J, et al. Trends in Racial and  
632 Ethnic Disparities in Barriers to Timely Medical Care Among Adults in the US, 1999 to  
633 2018. *JAMA Health Forum*. 2022 Oct 7;3(10):e223856.
- 634 34. All of Us Research Program. Policy on Stigmatizing Research [Internet]. 2020. Available  
635 from: [https://www.researchallofus.org/wp-content/themes/research-hub-wordpress-](https://www.researchallofus.org/wp-content/themes/research-hub-wordpress-theme/media/2020/05/AoU_Policy_Stigmatizing_Research_508.pdf)  
636 [theme/media/2020/05/AoU\\_Policy\\_Stigmatizing\\_Research\\_508.pdf](https://www.researchallofus.org/wp-content/themes/research-hub-wordpress-theme/media/2020/05/AoU_Policy_Stigmatizing_Research_508.pdf)
- 637 35. Using Population Descriptors in Genetics and Genomics Research. Washington, D.C.:  
638 National Academies Press; 2023.
- 639 36. Chiu AM, Molloy EK, Tan Z, Talwalkar A, Sankararaman S. Inferring population  
640 structure in biobank-scale genomic data. *Am J Hum Genet*. 2022 Apr 7;109(4):727–37.
- 641 37. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation  
642 PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015 Dec  
643 25;4(1):7.
- 644 38. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang  
645 HM, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68–74.
- 646 39. Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, et al. Insights into  
647 human genetic variation and population history from 929 diverse genomes. *Science*  
648 (1979). 2020 Mar 20;367(6484).
- 649 40. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons  
650 Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*.  
651 2016;538(7624):201–6.
- 652 41. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in  
653 unrelated individuals. *Genome Res*. 2009;19(9):1655–64.
- 654 42. Browning BL, Tian X, Zhou Y, Browning SR. Fast two-stage phasing of large-scale  
655 sequence data. *Am J Hum Genet*. 2021 Oct 7;108(10):1880–90.
- 656 43. Freyman WA, Mcmanus KF, Shringarpure SS, Jewett EM, Bryc K, Auton A. Fast and  
657 Robust Identity-by-Descent Inference with the Templated Positional Burrows–Wheeler  
658 Transform. *Mol Biol Evol*. 2021 May 4;38(5):2131–51.
- 659 44. Csardi G, Nepusz T. The igraph software package for complex network research.  
660 *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