

1 **Whole exome sequencing and characterization of coding variation in**  
2 **49,960 individuals in the UK Biobank**  
3

**Authors:** Cristopher V. Van Hout<sup>1</sup>, Ioanna Tachmazidou<sup>2</sup>, Joshua D. Backman<sup>1</sup>, Joshua X. Hoffman<sup>2</sup>, Bin Ye<sup>1</sup>, Ashutosh K. Pandey<sup>2</sup>, Claudia Gonzaga-Jauregui<sup>1</sup>, Shareef Khalid<sup>1</sup>, Daren Liu<sup>1</sup>, Nilanjana Banerjee<sup>1</sup>, Alexander H. Li<sup>1</sup>, Colm O'Dushlaine<sup>1</sup>, Anthony Marcketta<sup>1</sup>, Jeffrey Staples<sup>1</sup>, Claudia Schurmann<sup>1</sup>, Alicia Hawes<sup>1</sup>, Evan Maxwell<sup>1</sup>, Leland Barnard<sup>1</sup>, Alexander Lopez<sup>1</sup>, John Penn<sup>1</sup>, Lukas Habegger<sup>1</sup>, Andrew L. Blumenfeld<sup>1</sup>, Ashish Yadav<sup>1</sup>, Kavita Praveen<sup>1</sup>, Marcus Jones<sup>3</sup>, William J. Salerno<sup>1</sup>, Wendy K. Chung<sup>4</sup>, Ida Surakka<sup>5</sup>, Cristen J. Willer<sup>5</sup>, Kristian Hveem<sup>6</sup>, Joseph B. Leader<sup>7</sup>, David J. Carey<sup>7</sup>, David H. Ledbetter<sup>7</sup>, Geisinger-Regeneron DiscovEHR Collaboration<sup>7</sup>, Lon Cardon<sup>2</sup>, George D. Yancopoulos<sup>3</sup>, Aris Economides<sup>3</sup>, Giovanni Coppola<sup>1</sup>, Alan R. Shuldiner<sup>1</sup>, Suganthi Balasubramanian<sup>1</sup>, Michael Cantor<sup>1</sup>, Matthew R. Nelson<sup>2,\*</sup>, John Whittaker<sup>2,\*</sup>, Jeffrey G. Reid<sup>1,\*</sup>, Jonathan Marchini<sup>1,\*</sup>, John D. Overton<sup>1,\*</sup>, Robert A. Scott<sup>2,\*</sup>, Gonçalo Abecasis<sup>1,\*</sup>, Laura Yerges-Armstrong<sup>2,\*</sup>, Aris Baras<sup>1,\*</sup> on behalf of the Regeneron Genetics Center

4 \* These authors jointly supervised this work

**Affiliations:**

1. Regeneron Genetics Center
2. GlaxoSmithKline
- 5 3. Regeneron Pharmaceuticals
4. Departments of Pediatrics and Medicine, Columbia University Irving Medical Center
5. University of Michigan
6. Norwegian University of Science and Technology
7. Geisinger

6  
7 **Corresponding authors**

Correspondence and requests for materials should be addressed to Aris Baras, [aris.baras@regeneron.com](mailto:aris.baras@regeneron.com), Cristopher Van Hout, [cristopher.vanhout@regeneron.com](mailto:cristopher.vanhout@regeneron.com), Laura Yerges-Armstrong, [laura.m.yerges-armstrong@gsk.com](mailto:laura.m.yerges-armstrong@gsk.com), Robert Scott, [robert.a.scott@gsk.com](mailto:robert.a.scott@gsk.com).

8

## 9 SUMMARY

10           **The UK Biobank is a prospective study of 502,543 individuals, combining extensive**  
11 **phenotypic and genotypic data with streamlined access for researchers around the world. Here we**  
12 **describe the first tranche of large-scale exome sequence data for 49,960 study participants, revealing**  
13 **approximately 4 million coding variants (of which ~98.4% have frequency < 1%). The data includes**  
14 **231,631 predicted loss of function variants, a >10-fold increase compared to imputed sequence for**  
15 **the same participants. Nearly all genes (>97%) had  $\geq 1$  predicted loss of function carrier, and most**  
16 **genes (>69%) had  $\geq 10$  loss of function carriers. We illustrate the power of characterizing loss of**  
17 **function variation in this large population through association analyses across 1,741 phenotypes. In**  
18 **addition to replicating a range of established associations, we discover novel loss of function variants**  
19 **with large effects on disease traits, including *PIEZO1* on varicose veins, *COL6A1* on corneal**  
20 **resistance, *MEPE* on bone density, and *IQGAP2* and *GMPR* on blood cell traits. We further**  
21 **demonstrate the value of exome sequencing by surveying the prevalence of pathogenic variants of**  
22 **clinical significance in this population, finding that 2% of the population has a medically actionable**  
23 **variant. Additionally, we leverage the phenotypic data to characterize the relationship between rare**  
24 ***BRCA1* and *BRCA2* pathogenic variants and cancer risk. Exomes from the first 49,960 participants**  
25 **are now made accessible to the scientific community and highlight the promise offered by genomic**  
26 **sequencing in large-scale population-based studies.**

27

28

## 29 INTRODUCTION

30 The UK Biobank (UKB) is a prospective population-based study of over 500,000 individuals with  
31 extensive and readily accessible phenotypic and genetic data<sup>1</sup>. The release of genome-wide genotyping  
32 array data<sup>2</sup> for study participants has accelerated genomic discovery through association studies, and  
33 enabled advances in population genetic analyses, the exploration of genetic overlap between traits, and  
34 Mendelian randomization studies<sup>3,4</sup>. While array data in combination with genotype imputation capture the  
35 spectrum of common genetic variants, rare variation that is more likely to modify protein sequences and  
36 have large phenotypic consequences is less well captured through these approaches.

37 Here, we extend the UKB resource with the first tranche of whole exome sequencing (WES) for  
38 49,960 UK Biobank participants, generated by the Regeneron Genetics Center, as part of a collaboration  
39 with GlaxoSmithKline. These data are available to approved researchers through the UKB Data Showcase  
40 (see **URLs**). Exome sequencing allows direct assessment of protein-altering variants, whose functional  
41 consequences are more readily interpretable than non-coding variants, providing a clearer path towards  
42 mechanistic and therapeutic insights, as well as potential utility in therapeutic target discovery and  
43 validation<sup>5-8</sup> and in precision medicine<sup>9,10</sup>. Here, we provide an overview of sequence variation in UKB  
44 exomes, review predicted damaging variants and their consequences in the general population and perform  
45 comprehensive loss of function (LOF) burden testing with 1,741 phenotypes, illustrating its utility in studies  
46 of common and rare phenotypes with a focus on deleterious coding variation.

47

## 48 RESULTS

### 49 Demographics and Clinical Characteristics of Sequenced Participants

50 A total of 50,000 participants were selected, prioritizing individuals with more complete phenotype  
51 data: those with whole body MRI imaging data from the UK Biobank Imaging Study, enhanced baseline  
52 measurements, hospital episode statistics (HES), and/or linked primary care records. Additionally, we  
53 selected one disease area for enrichment, including individuals with admission to hospital with a primary

54 diagnosis of asthma (ICD10 codes J45 or J46). This resulted in 8,250 participants with asthma, or ~16%  
 55 among sequenced participants, compared to ~13% among all 502,543 UKB participants (Table 1 and Ext.  
 56 Data HESinWESvs500k\_V1.xlsx). During data generation, samples from 40 participants were excluded  
 57 due to failed quality control measures or participant withdrawal (Supplemental Methods), resulting in a  
 58 final set of 49,960 individuals. The sequenced participants are representative of the overall 502,543 UKB  
 59 participants (Table 1) for age, sex and ancestry. Due to the ascertainment strategy, sequenced participants  
 60 were more likely to have HES diagnosis codes (84.2% among WES vs. 78.0% overall), were enriched for  
 61 asthmatics, and many enhanced physical measures (eye measures, hearing test, electrocardiogram, Table  
 62 1). The exome sequenced participants did not differ from all participants in the median number of primary  
 63 and secondary ICD10 codes. The sequenced subset includes 194 parent-offspring pairs, including 26  
 64 mother-father-child trios, 613 full-sibling pairs, 1 monozygotic twin pair and 195 second-degree genetically  
 65 determined relationships<sup>11</sup> based on identity-by-descent (IBD) estimates between pairs of individuals in  
 66 UKB WES is included in Sup. Figure 1.  
 67

<b>Demographic and Clinical Characteristics</b>	<b>UKB 50k WES Participants</b>	<b>UKB 500k Participants</b>
N	49,960	502,543
Female, n(%)	27,243 (54.5)	273,460 (54.4)
Age at assessment, years (1st-3rd Quartiles)	58 (45-71)	58 (45-71)
Body mass index, kg/m <sup>2</sup> (1st-3rd Quartiles)	26 (21-31)	26 (21-31)
Number of imaged participants (%)	12,075 (24.1) <sup>a</sup>	21,407 (4.3) <sup>ab</sup>
Number of current/past smokers, n(%)	17,515 (35.0)	216,482 (43.1)
Townsend Deprivation Index (1st-3rd Quartiles)	-2.0 (-6.1, -2.1)	-2.13 (-6.2, -1.9)
Inpatient ICD10 codes per patient	5	5
Patients with >=1 ICD10 diagnoses, n(%)	42,066 (84.2)	391,983 (78.0)
<b>Genetic Ancestry Assignment<sup>c</sup></b>		
African (%)	1.49	1.24
East Asian (%)	0.54	0.51
European (%)	93.6	94.5
<b>Cardiometabolic phenotypes</b>		
Coronary Disease, n(%)	3,340 (6.7)	35,879 (7.1)
Heart Failure, n(%)	300 (0.6)	4,399 (0.8)
Type 2 Diabetes, n(%)	1,541 (3.0)	17,261 (3.4)

<b>Respiratory and immunological phenotypes</b>		
Asthma, n(%)	8,250 (16.5)	68,149 (13.5)
COPD, n(%)	741 (1.4)	7,438 (1.4)
Rheumatoid Arthritis, n(%)	710 (1.4)	7,337 (1.4)
Inflammatory Bowel Disease n(%)	543 (1.0)	5,783 (1.1)
<b>Neurodegenerative phenotypes</b>		
Alzheimer's Disease, n(%)	13 (0.05)	320 (0.06)
Parkinson's Disease, n(%)	65 (0.13)	1,043 (0.21)
Multiple Sclerosis, n(%)	126 (0.25)	1,352 (0.26)
Myasthenia Gravis, n(%)	14 (0.02)	217 (0.04)
<b>Oncology phenotypes</b>		
Breast Cancer, n(%)	1,667 (3.3)	16,887 (3.3)
Ovarian Cancer, n(%)	162 (0.3)	1,777 (0.3)
Pancreatic Cancer, n(%)	602 (1.2)	4,611 (0.9)
Prostate Cancer, n(%)	848 (1.6)	8,855 (1.7)
Melanoma, n(%)	598 (1.1)	5,715 (1.1)
Lung Cancer, n(%)	172 (0.3)	2,581 (0.5)
Colorectal Cancer, n(%)	368 (0.7)	3,971 (0.8)
Cutaneous squamous cell carcinoma, n(%)	1,316 (2.6)	12,969 (2.6)
<b>Enhanced measures</b>		
Hearing test, n(%)	40,546 (81.1)	167,011 (33.2)
Pulse Rate, n(%)	40,548 (34.2)	170,761 (33.9)
Visual Acuity Measured, n(%)	39,461 (78.9)	117,092 (23.2)
IOP measured (left), n(%)	37,940 (75.9)	111,942 (22.2)
Autorefracton, n(%)	36,067 (72.1)	105,989 (21.0)
Retinal OCT, n(%)	32,748 (65.5)	67,708 (13.4)
ECG at rest, n(%)	10,829 (27.1)	13,572 (2.1)
Cognitive Function, n(%)	9,511 (19.0)	96,362 (19.1)
Digestive Health, n(%)	13,553 (28.1)	142,310 (28.3)
Physical Activity Measurement, n(%)	10,684 (21.3)	101,117 (20.1)

68  
69 **Table 1 | Clinical characteristics in whole exome sequenced and all UK Biobank participants**  
70 Demographics and clinical characteristics of UKB 50K sequenced participants and overall 500K  
71 participants. See Supplemental Methods for definition of UKB clinical phenotype definitions. Unless  
72 otherwise noted, values are expressed as median (1<sup>st</sup> and 3<sup>rd</sup> quartile). <sup>a</sup>The number of samples with exome  
73 sequencing data and at least one non-missing image derived phenotype value from data downloaded from  
74 UK Biobank in November 2018. <sup>b</sup>The number of samples with at least one non-missing image derived  
75 phenotype value from data downloaded from UK Biobank in November 2018. <sup>c</sup>Number of samples in 3

76 pre-defined regions of a plot of the first two genetic principal component scores, where the regions are  
77 selected to represent African, East Asian, and European ancestry (Sup. Figure 2).

78

## 79 **Summary and Characterization of Coding Variation from WES**

80 Exomes were captured using a slightly modified version of the IDT xGen Exome Research Panel  
81 v1.0. The basic design targets 39 megabases of the human genome (19,396 genes among autosomes and  
82 sex chromosomes) and was supplemented with additional probes to boost coverage at under-performing  
83 loci. In each sample and among targeted bases, coverage exceeds 20X at 94.6% of sites on average  
84 (standard deviation 2.1%). We observe 4,735,722 variants within targeted regions (Table 2). These variants  
85 include 1,229,303 synonymous (97.9% with minor allele frequency, MAF<1%), 2,498,947 non-  
86 synonymous (98.9% with MAF<1%), and 231,631 predicted LOF variants affecting at least one coding  
87 transcript (initiation codon loss, premature stop codons, splicing, and frameshifting indel variants; 99.6%  
88 with MAF<1%) (Fig. 1a). Our tally of the median number of variants per individual includes 9,619  
89 synonymous (IQR 128), 8,781 missense (IQR 137) and 219 LOF variants (IQR 16) and is comparable to  
90 previous exome sequencing studies<sup>12,13</sup>; the increasing proportion of rare variants in the LOF and missense  
91 categories is consistent with purifying selection. If we restrict analysis to LOF variants that affect all  
92 ENSEMBL 85 transcripts for a gene, the number of LOF variants drops to 153,903 overall and 111 per  
93 individual (a reduction of ~33.5% and ~49.3%, respectively), consistent with previous studies. In addition  
94 to variants in targeted regions, we also capture exon adjacent variation. Including non-targeted regions, we  
95 observe 9,693,526 indel and single nucleotide variants (SNVs) after quality control, 98.5% with MAF<1%.  
96 These additional variants can be helpful aids for population genetic analyses and for applications such as  
97 phasing and IBD segment detection.

98

99

100

	Variants in WES, n=49,960 Participants		Median Per Participant (IQR)	
	# Variants	# Variants MAF<1%	# Variants	# Variants MAF<1%
Total	9,693,526	9,547,730	48,982 (627)	1,626 (133)
Targeted Regions <sup>1</sup>	4,735,722	4,665,684	24,332 (283)	780 (63)
<b>Variant Type<sup>1</sup></b>				
SNVs	4,520,754	4,453,941	23,529 (276)	739 (61)
Indels	214,968	211,743	803 (29)	42 (10)
Multi-Allelic	591,340	580,728	3,388 (63)	117 (18)
<b>Functional Prediction</b>				
Synonymous	1,229,303	1,203,043	9,619 (128)	228 (28)
Missense	2,498,947	2,472,384	8,781 (137)	380 (39)
LOF (any transcript)	231,631	230,790	219 (16)	24 (8)
LOF (all transcripts)	153,903	153,441	111 (12)	15 (6)

101

102 **Table 2 | Summary statistics for variants in sequenced exomes of 49,960 UKB participants.** Counts of

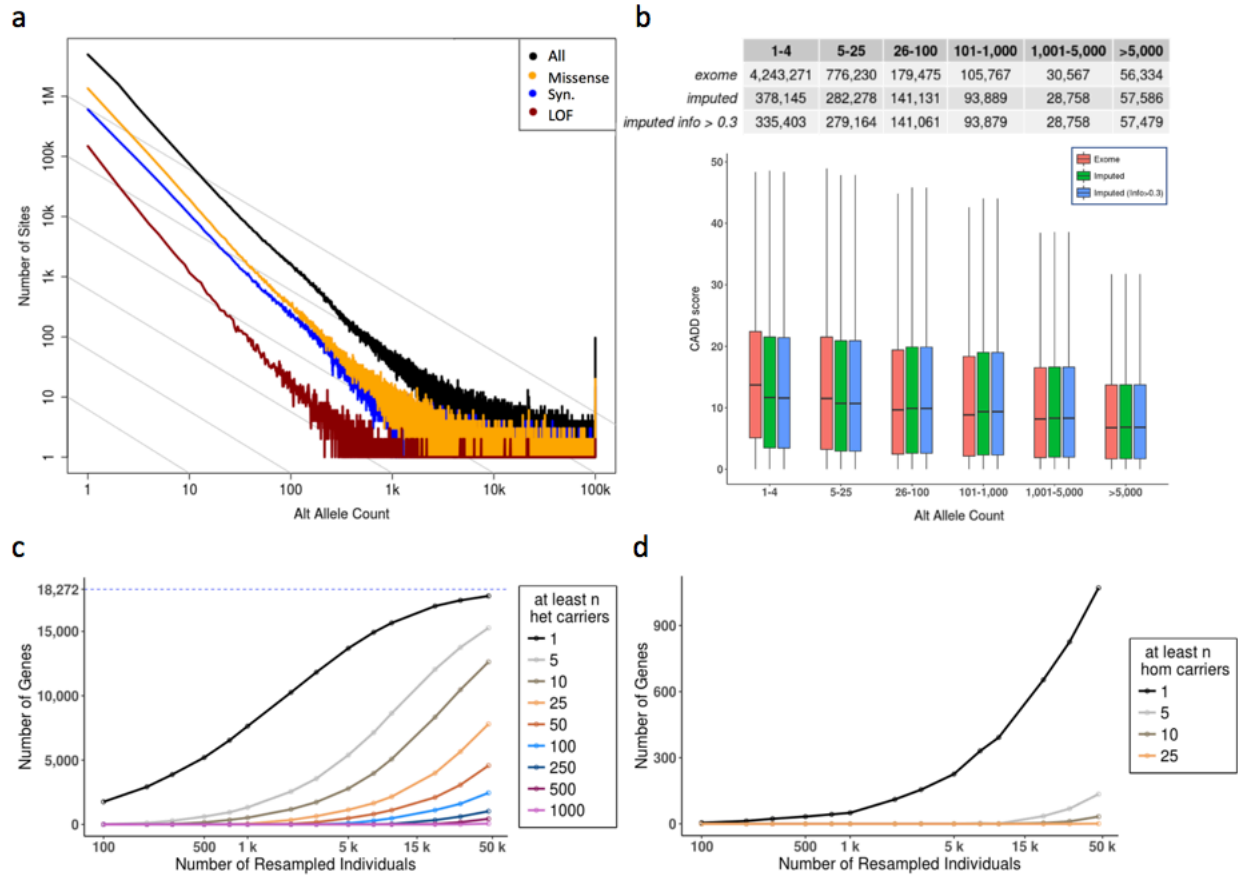
103 autosomal variants observed across all individuals by type/functional class for all and for MAF<1%

104 frequency. The number of targeted bases by the exome capture design was n=38,997,831. All variants

105 passed quality control (QC) criteria (See Supplemental Methods), individual and variant missingness <10%,

106 and Hardy Weinberg p-value>10<sup>-15</sup>. Median count and interquartile range (IQR) per individual for all

107 variants, and for MAF<1%.<sup>1</sup> Counts restricted to WES targeted regions.



108

109

110

111

112

113

114

115

116

117

118

119

120

121

**Figure 1 | Summary statistics for variation in WES and imputed sequence** **a**, Observed site frequency spectrum for all autosomal variants and by functional prediction in 49,960 UKB participants. **b**, Distribution of CADD scores for variant allele counts in regions consistently covered by WES (90% of individuals with >20X depth) in WES and imputed data in 49,797 UKB participants with WES and imputed sequence. **c**, The number of autosomal genes with at least 1, 5, 10, etc. heterozygous and homozygous non-reference genotypes **d**, LOF carriers increases with sample size. LOFs passed GL (Goldilocks) QC (see Supplemental Methods for GL QC filtering definition), genotype missingness < 10%, and HWE p-value > 10<sup>-15</sup>. 46,808 UKB participants of European ancestry with WES were down-sampled at random to the number of individuals specified on the horizontal axis. The number of genes containing at least the indicated count of LOFs MAF < 1% carriers as in the legend are plotted on the vertical axis. The maximum number of autosomal genes is 18,272 in this analysis (See Supplemental Methods for gene model).

## Enhancement of Coding Variation from WES Compared to Imputed Sequence



122 To evaluate enhancements to the UKB genetic variation resource through WES, we compared the  
123 number of coding variants observed in 49,797 individuals in whom both WES and imputed sequence were  
124 available; this is a subset of the 49,960 individuals with WES. To capture as many variants in the overlap  
125 of WES and imputed sequence as possible, only minimal filtering was applied (WES sites were filtered  
126 using genotype likelihoods, with no filtering for imputed data), thus variant counts are greater than in Table  
127 2 and Sup. Table 1, respectively. Among all autosomal variants, we observed increases in the total number  
128 of coding variants (3,995,794 to 707,124); synonymous (1,241,804 to 267,479), missense (2,518,075 to  
129 420,194), and LOF (235,915 to 19,451) in WES compared to imputed sequence, respectively (Sup. Table  
130 2). This represents a >10-fold increase in the number of LOF variants identified by WES compared to those  
131 in the imputed sequence.

132 Amongst nearly four million coding variants observed in the exome sequence data, only 13.7%  
133 were also observed in the imputed sequence, highlighting the added value of exome sequencing for  
134 ascertainment of rare coding variation. Similarly, among LOFs observed in either dataset, 92.0%  
135 (n=223,427) were unique to WES and absent in the imputed sequence data. We observed 12,488 LOFs  
136 present in both datasets, meaning that only 5.3% of the >235,915 LOFs identified by WES were present in  
137 the imputed sequence. Since LOFs are especially informative for human genetics and medical sequencing  
138 studies, this enhancement clearly emphasizes the value of exome sequencing.

139 There were 6,963 LOFs seen only in the imputed sequence. These represent both variants within  
140 regions not targeted or captured in WES and errors in imputation, which are especially pronounced at low  
141 allele frequencies. Amongst the 6,963 LOFs (5,939 SNVs) observed only in the imputed sequence, we  
142 identified 1,730 LOF variants (24.8%, including 1,348 SNVs) in regions that were not targeted, 4,438 LOF  
143 variants (63.7%, including 3,804 SNVs) that fell within consistently covered regions of WES (>20x  
144 coverage in 90% of samples), and 885 LOF variants (12.7%, including 787 SNVs) in inconsistently covered  
145 exome-captured target or buffer regions. Amongst the set of 4,438 LOF variants in consistently covered  
146 regions seen solely in the imputed sequence, we selected 363 variant sample-sites from across the MAF  
147 spectrum and manually reviewed the underlying read data in the Integrative Genomics Viewer<sup>14</sup> (IGV) (see

148 Supplemental Methods). We observed that 76% of the selected sample sites had no sequencing evidence to  
149 support the imputed LOF call. Approximately 21% had some evidence for the presence of any variation  
150 (e.g. multi-nucleotide polymorphism). Only ~3% had any clear evidence of the LOF variant called in the  
151 imputed data. Sites that validated were more likely to be common than rare, as expected for imputed  
152 variants.

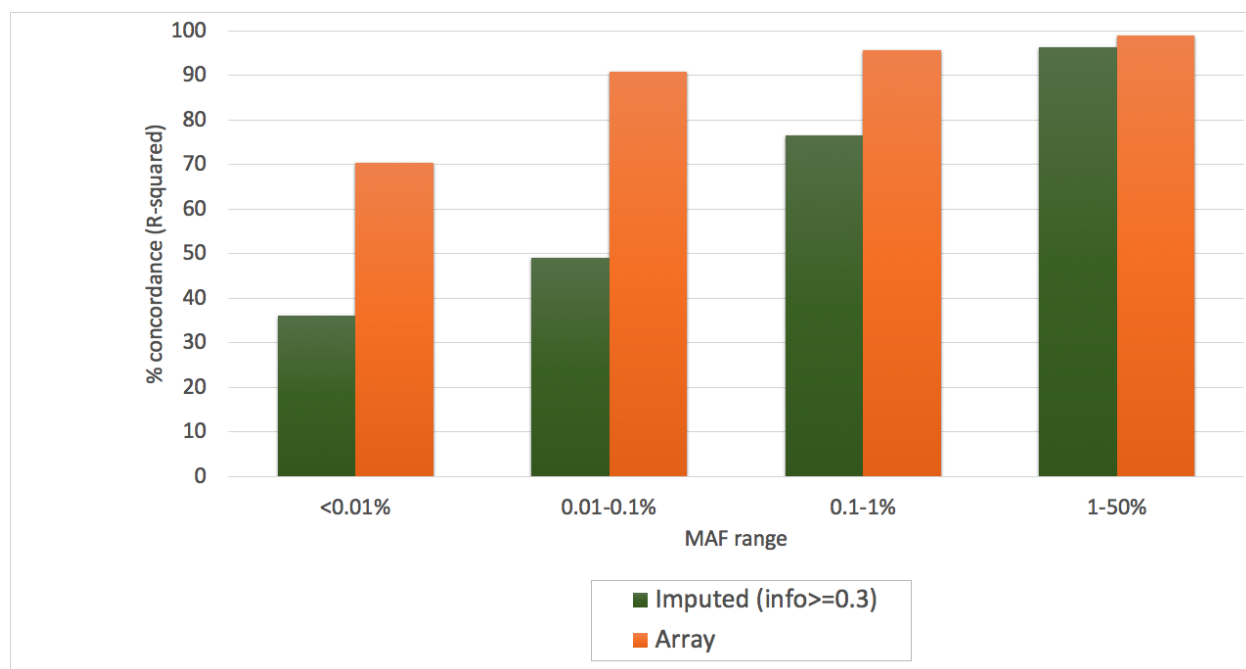
153 We also noted that amongst all the 707,124 imputed coding variants, 22.6% of them were not  
154 observed in the exome sequence data; a large portion of these will similarly suffer from poor imputation  
155 accuracy as observed in WES and imputed sequence concordance (Figure 2). As expected, common  
156 variants across functional prediction classes were more likely to be captured by both WES and imputed  
157 sequence, whereas rare variants were more likely unique to WES (Sup. Table 2). As an expected result of  
158 purifying selection, we observed that lower frequency variants were predicted to be more deleterious as  
159 measured by CADD<sup>15</sup> score distributions in both datasets (Figure 1b). Interestingly, among rare variants,  
160 those identified by WES were typically classified as more deleterious (Figure 1b) – likely because rare  
161 variants that can be imputed may often be common in other populations even when rare in UKB.

162

### 163 **Concordance of WES, Directly Genotyped Array, and Imputed Sequence**

164 We measured concordance between WES and array genotypes in individuals with both datasets  
165 (Supplemental Methods), using the squared correlation ( $R^2$ ) of allele counts in sequencing and array  
166 genotyping. This measure facilitates interpretation of assessments of accuracy for both rare and common  
167 variation<sup>16,17</sup>. Using the same approach, we also measured concordance between WES and imputed  
168 sequence. As expected, concordance between genomic measures declines with decreasing MAF (Figure 2,  
169 Sup. Figures 3,4). Concordance between WES and imputed sequence ranged from 35% for MAF <0.01%  
170 to 96% for allele frequencies >1%, averaging 54% across all allele frequencies. Concordance between WES  
171 and array genotyped variants was substantially higher, ranging from 70% for MAF <0.01% to 99% for  
172 MAF >1% (Figure 2), and averaging 99% across all allele frequencies. As expected, WES performs much

173 better in terms of concordance with array genotypes, since both directly assay the variation rather than make  
174 a computational prediction. This is particularly true in the rarest allele frequencies where the accuracy of  
175 imputation is limited using current imputation reference panels.  
176



177  
178 **Figure 2 | Concordance between WES, imputed sequence, and array genotypes.** R-squared correlation  
179 coefficients between variants in WES and imputed sequence (green) and array genotypes and WES  
180 (orange), calculated per variant and binned by minor allele frequency in WES. n=46,912 individuals and  
181 n=75,334 variants were represented in the array-WES comparison (n=4,261, n=17,780, n=23,087 and  
182 n=30,206 variants in <0.01%, 0.01-0.1%, 0.1-1%, and 1-50% MAF bins respectively). n=46,860  
183 individuals and n=899,455 variants were represented in the imputed-WES comparison (n=346,826,  
184 n=304,524, n=126,554, and n=121,551 variants in <0.01%, 0.01-0.1%, 0.1-1%, and 1-50% MAF bins  
185 respectively).

186

187 **Comparison of LOF Variants Ascertained Through WES and Imputed Sequence**

188 LOF variants constitute a major class of genetic variation of great interest due to their disruption  
189 of gene function, their causal role in many Mendelian disorders, and the success of leveraging protective  
190 LOF variants to identify novel drug targets<sup>5,6,18</sup>. Rare LOF variation is best captured by direct sequencing  
191 approaches, such as WES, and we sought to quantify the improved yield of LOF variation from WES  
192 compared to array genotyping and imputed sequence. We compared the number of LOF variants ascertained  
193 through WES and imputed sequence among 49,797 UKB participants. Notably, not all individuals with  
194 WES have imputed sequence available. We observed a larger number of LOF variants impacting any  
195 transcript in WES vs imputed sequence, 235,915 and 19,451, respectively (See Sup. Table 2). Further, we  
196 observed a greater number of genes with  $\geq 1$  heterozygous LOF variant carrier (17,751 genes from WES,  
197 8,763 from imputation (info >0.3) and genes with  $\geq 1$  homozygous LOF variant carrier (1,071 from WES,  
198 789 from imputation) (Table 3). The number of genes with LOFs at different thresholds of imputation  
199 accuracy for 50k and 500k resources are included in Sup. Table 3. At equivalent sample sizes (n=46,827  
200 European ancestry individuals with both WES and imputed sequence), WES data included a greater number  
201 of genes with LOF variants, across all carrier count thresholds. Even more striking was that WES in 46,827  
202 individuals yielded more genes (17,751) with heterozygous LOFs than imputed sequence (info>0.3) in all  
203 462,427 UKB participants of European ancestry (8,724 genes). Tracking the increase in the number of  
204 genes with heterozygous LOF variant carriers with the increase in the number of sequenced samples  
205 suggests that we are approaching saturation for this metric, having likely observed at least one heterozygous  
206 LOF variant carrier in most of the genes that tolerate these variants, and most genes overall (Figure 1c). In  
207 contrast, the number of genes for which homozygous LOF variants are observed still appears to increase  
208 rapidly as more samples in UKB are sequenced, suggesting that homozygous instances of LOF variants for  
209 many more genes can be identified by sequencing additional individuals (Figure 1d).

210

211

212

		# Autosomal genes containing at least N LOFs, MAF < 1%					
Zygosity	Genomic resource	1	5	10	25	50	100
Het	50k exome	17,751	15,269	12,629	7,799	4,573	2,453
	50k imputed (info > 0.3)	8,763	6,999	5,933	4,297	2,965	1,911
	500k imputed (info > 0.3)	8,724	7,711	7,267	6,539	5,847	4,916
Hom	50k exome	1,071	135	33	1	0	0
	50k imputed (info > 0.3)	789	74	7	0	0	0
	500k imputed (info > 0.3)	1,752	597	351	120	21	3

213 **Table 3 | Number of autosomal genes with heterozygous, homozygous LOF variants.** Count of genes  
 214 with at least the specified number of LOFs (MAF < 1%) impacting any transcript in European ancestry in  
 215 approximately 50k (n=46,827 with WES and imputed sequence), and 462,427 individuals for 500k imputed  
 216 sequence.

217

218 Extrapolating, we can estimate the number of genes where multiple loss of function carriers will  
 219 be observed once our experiment is complete and all ~500,000 participants are sequenced. Cautiously, we  
 220 currently predict that >17,000, >15,000, and >12,000 genes will have  $\geq 10$ ,  $\geq 50$ , and  $\geq 100$  heterozygous  
 221 LOF carriers in the full dataset (see Sup. Methods, Sup. Methods Fig. 1 & Table 1).

222 Our WES results are consistent with those of a recent large-scale survey<sup>19</sup> of genetic variation in  
 223 141,456 individuals from the Genome Aggregation Database (gnomAD). When we annotate both exome  
 224 variant lists with the same annotation pipelines and subset results to similar numbers of individuals and  
 225 ancestry, we observe 17,751 genes with LOFs in any transcript in 46,979 European individuals in UKB  
 226 with WES vs 17,946 genes with LOFs in any transcript in 56,885 Non-Finnish European (NFE) individuals  
 227 in gnomAD exomes. Further subsetting to high confidence LOF variants (with LOFTEE, see Sup.  
 228 Methods), we obtain 17,640 genes with LOFs in UKB and 17,856 in gnomAD (Sup. Methods Table 2).

229

### 230 **Survey of Medically Actionable Pathogenic Variants**

231 To date, more than 5,000 genetic disorders have been described for which the molecular causes and  
 232 associated genes have been defined. The American College of Medical Genetics (ACMG) has proposed 59

233 genes, ACMG SF2.0,<sup>20,21</sup> which are associated with highly penetrant disease phenotypes and for which  
234 available treatments and/or prevention guidelines can significantly reduce the morbidity and mortality in  
235 genetically susceptible individuals. Large-scale human genomic sequencing efforts coupled with EHR data  
236 provide opportunities to assess penetrance and prevalence of pathogenic (P) and likely pathogenic (LP)  
237 variants in known monogenic disease genes, as well as investigate the phenotypic effects of variants of  
238 unknown significance (VUS). Additionally, phenotypically agnostic population sampling provides  
239 opportunities to better characterize the phenotypic spectrum of these disorders and estimate the associated  
240 disease risk in the population. Furthermore, these efforts enable the implementation of precision medicine  
241 by identifying individuals carrying medically actionable pathogenic variants and providing medical care  
242 and surveillance preemptively.

243 We interrogated variant data for the 49,960 individuals with WES from the UKB to identify a  
244 “strict” set of reported pathogenic missense and LOF variants (that is, those with  $\geq 2$  stars in ClinVar and  
245 no conflicting interpretations) as well as a set of likely pathogenic LOF variants (that is, those in genes  
246 where truncating mutations are known to cause disease) according to the current ACMG 59 gene set<sup>21</sup> (see  
247 Supplemental Methods). We identified a total of 555 such variants (316 in the reported pathogenic set, 239  
248 in the likely pathogenic set) in 1,000 unique individuals (Table 4, Ext. Data ACMG59Variants.xlsx, 9  
249 individuals carried variants in 2 genes). Of note, 47 of the likely pathogenic variants would qualify as  
250 previously reported pathogenic variants using a broader definition (Sup. Table 4). Variants were observed  
251 in 48 of 59 ACMG genes, with a median number of 5 variants per gene and a median 2 carriers per gene.  
252 Overall, 2.0% of the sequenced individuals carry a flagged variant in one of the ACMG59 genes. Using the  
253 same methodology in 91,514 participants from the Geisinger-Regeneron DiscovEHR study<sup>22</sup> sequenced to  
254 date, we observed a slightly higher percentage of individuals, 2.76%, carrying a potentially actionable rare  
255 pathogenic variant in the ACMG59 genes (Sup. Table 5). This difference may reflect differences between  
256 a study of individuals seeking clinical care (DiscovEHR) and a population-based study not ascertained in  
257 the context of active clinical care (UKB).

258

Category	#Variants	% of Total Known ACMG59 Variants	#Carriers	% of individuals with reportable variants
Pathogenic (P)	316	4.23	694	1.39
Likely Pathogenic (LP)	239	-	315	0.63
P + LP	555	-	1,009	2.0 <sup>1</sup>

259

260 **Table 4 | Medically actionable variants in ACMG59 genes in UKB participants.** Of the 49,960 UKB

261 participants with WES data, 2.0% are carriers of pathogenic (P) and likely pathogenic (LP) variants in

262 ACMG59 v2.0 genes based on strict variant filtering criteria. LP variant counts include LOF variants

263 passing QC criteria in ACMG59 genes that are not reported in ClinVar ( $\geq 2$  star). Amongst all P+LP

264 variants, 384 variants were observed in only one individual, 165 were observed in 2-10 individuals, and 6

265 were observed in >10 individuals. <sup>1</sup>Percent of individuals with P or LP variants is not additive, as the 2.0%

266 represents non-redundant carriers; 9 individuals were found to have 2 medically actionable variants.

267

268 Among the 48 of the reportable ACMG59 genes, variants in cancer associated genes were the most

269 prevalent in UKB with WES; *BRCA2* (93 variants, 166 carriers), *BRCA1* (40 variants, 60 carriers), *PMS2*

270 (21 variants, 59 carriers) and *MSH6* (35 variants, 52 carriers); while variants in *LDLR*, associated with

271 familial hypercholesterolemia [MIM #143890], were the second most prevalent (35 variants, 68 carriers).

272 Cardiac dysfunction disorders were also highly represented mainly by variants in *KCNQ1* (25 variants, 55

273 carriers), *PKP2* (22 variants, 55 carriers), and *MYBPC3* (25 variants, 50 carriers) (Figure 3a). The majority

274 of variants were identified in genes responsible for dominant conditions, for which heterozygous carriers

275 are at risk (994 carriers); however, we also identified 6 individuals homozygous for pathogenic variants in

276 genes associated with autosomal recessive (familial adenomatous polyposis-2 [FAP2, MIM #608456] and

277 mismatch repair cancer syndrome [MMRCS, MIM #276300] due to biallelic pathogenic variants in

278 *MUTYH* and *PMS2*, respectively) or hemizygous for X-linked conditions (Fabry disease [MIM #301500]

279 due to pathogenic variants in *GLA*). Indeed, a male hemizygous for the c.335G>A; p.Arg112His pathogenic

280 variant in *GLA* has diagnoses of angina pectoris, atrial fibrillation, chest pain, and chronic ischemic heart

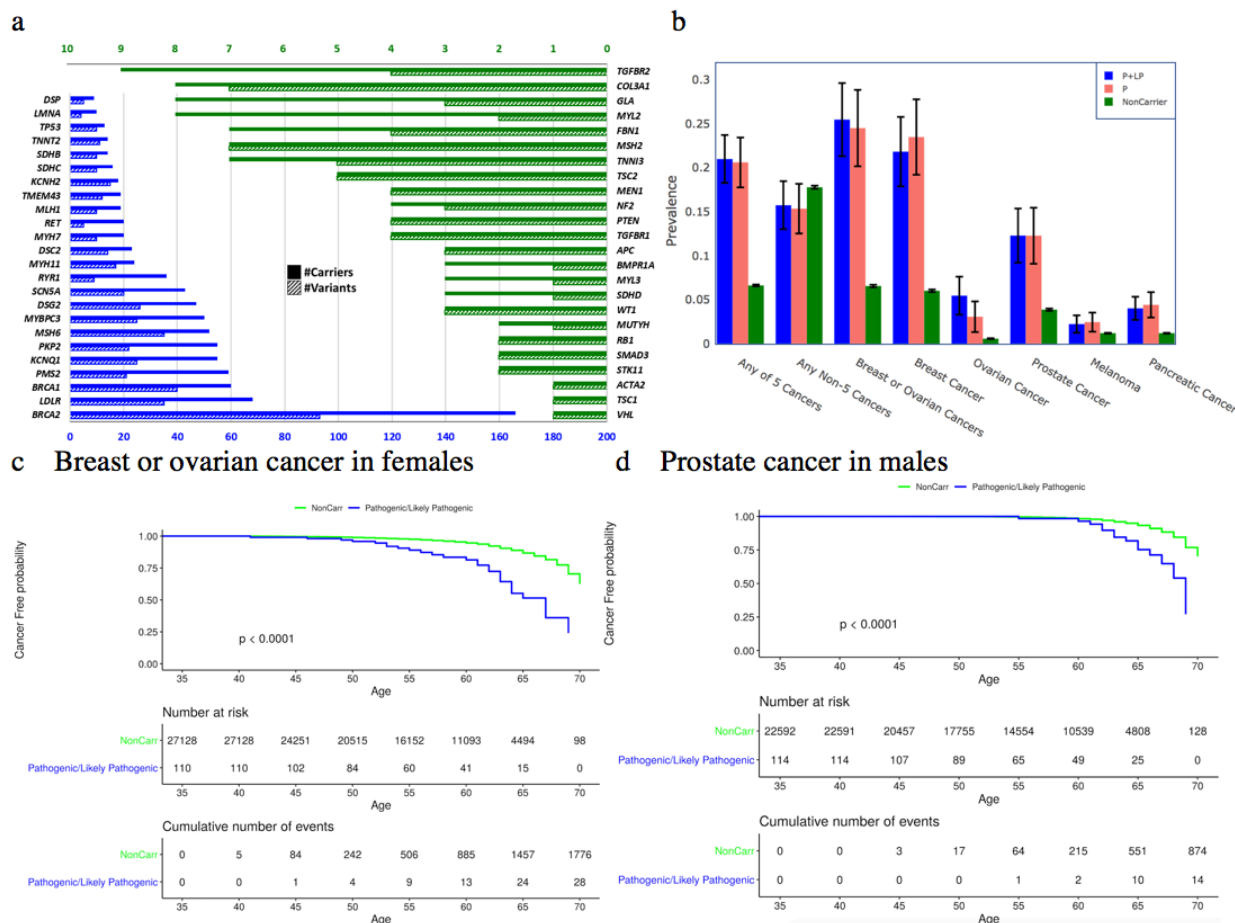
281 disease. Similarly, one individual homozygous for a pathogenic missense variant (c.1145G>A;

282 p.Gly382Asp) in *MUTYH* has a history of benign neoplasm of colon, diverticular disease of the intestine,  
283 colonic polyps, and intestinal obstruction. These examples illustrate how the extensive health data available  
284 for UK Biobank participants provide a valuable resource to assess variant pathogenicity and disease risk at  
285 the population level, and the potential to model outcomes for individuals harboring pathogenic variants.

286 As an example, we evaluated the risk of cancer in individuals carrying pathogenic variants in  
287 *BRCA1* or *BRCA2* (Figure 3b) to compare across five previously implicated *BRCA1/2* associated cancers<sup>23</sup>  
288 as well as to explore whether risk was conferred for other cancers. While *BRCA1* and *BRCA2* have  
289 differences in mechanism and risk among cancer subtypes, due to low sample sizes, we analyzed summed  
290 counts in 114 males and 110 females with *BRCA1/2* reported pathogenic and likely pathogenic variants.  
291 We found the prevalence of cancers to be elevated in carriers of pathogenic variants in *BRCA1/2* for the 5  
292 cancers (breast in females, ovarian, prostate, melanoma, and pancreatic) previously associated with *BRCA1*  
293 or *BRCA2* (cancer status was derived from self-report, HES, and cancer prevalence for any of the 5 cancers  
294 was 21.0% in carriers vs 6.6% in non-carriers, OR = 3.75 (95%CI 2.71,5.18), p-value =  $2 \times 10^{-12}$ , (3337  
295 cases, 46599 controls), and there was no significant difference between carriers and non-carriers when all  
296 other cancers, excluding these 5, were combined; 15.7% in carriers vs 17.7% in non-carriers, OR = 0.86  
297 (95%CI 0.58,1.29), p-value = 0.55, (8300 cases, 38424 controls). The most prevalent cancers in this group  
298 were C443 and C449 unspecified malignant neoplasms of skin, and D069 carcinoma in situ of cervix.  
299 Increased risk was observed in these *BRCA1/2* variant carriers for each of the five previously associated  
300 cancers analyzed individually (ovarian in females OR=9.97 (95%CI 4.32,23.06), p= $5.2 \times 10^{-5}$ , (162 cases,  
301 27,076 controls); breast in females OR=4.37 (95%CI 2.77,6.88), p= $3.6 \times 10^{-8}$ , (1,654 cases, 25,584 controls);  
302 prostate in males OR=3.48 (95%CI 1.98, 6.11), p= $1.5 \times 10^{-4}$ , (888 cases, 21,818 controls); melanoma  
303 OR=1.9 (95%CI 0.78,4.62), p=0.20, (599 cases, 49,345 controls); and pancreatic cancer OR 3.47 (95%CI  
304 1.77,6.79), p= $1.7 \times 10^{-3}$ , (602 cases, 49,342) controls; (p values by Fisher's exact test). These differences in  
305 overall cancer risk also manifest as clearly different disease onset and rates of cancer-free survival between  
306 carriers and non-carriers (see Figure 3c for breast and ovarian cancer in females; Figure 3d for prostate



307 cancer in males). Comparing the cumulative proportion of female participants free of breast and ovarian  
 308 cancer, we estimate a hazard ratio of 4.3 (95%CI 2.96,6.24,  $p < 0.0001$ ). Comparing the cumulative  
 309 proportion of male participants free of prostate cancer, the hazard ratio was 3.68 (95%CI 2.17, 6.24,  
 310  $p < 0.0001$ ). With the UKB resource, cancer risk can be more deeply explored across broader sets of variants  
 311 as well as with larger exome sequence datasets and continued accrual of incident cancers. Our results  
 312 corroborate those of another recent population-based application of WES linked to health records to  
 313 evaluate cancer risk in individuals with pathogenic variants in *BRCA1/2*<sup>10</sup>, demonstrating the value of WES  
 314 to identify high-penetrance rare alleles associated with clinical phenotypes; such efforts can be applied  
 315 across other genetic disorders, enabling the implementation of precision medicine at the population level.



317

318 **Figure 3 | Summary of observed actionable ACMG59 variants, and pathogenicity of *BRCA1/2***  
319 **variants. a,** Counts of variants and carriers in 48 ACMG59 genes with pathogenic (P) or likely pathogenic  
320 (LP) variants. **b,** Prevalence of cancers in carriers of P, P or LP, and no P or LP variants in *BRCA1* or  
321 *BRCA2*. Five major cancers related to *BRCA1/2* risk include breast in females, ovarian, prostate, melanoma,  
322 and pancreatic cancers. Cases are aggregated from Cancer Registry, HES, and Self Report. **c,** Cumulative  
323 proportion of female participants free of breast and ovarian cancer, and **d,** male participants free of prostate  
324 cancer with P or LP variants in *BRCA1/2* compared to non-carriers by age at interview.

325

### 326 **Phenotypic Associations with LOF Variation**

327 The combination of WES, allowing comprehensive capture of LOF variants, with rich health information  
328 allows for broad investigation of the phenotypic consequences of human genetic variation. We conducted  
329 burden tests for rare (MAF < 1%) LOF variants in autosomal genes with >3 LOF variant carriers and in  
330 1,741 traits (1,073 discrete traits with at least 50 cases defined by HES and self-report data, 668  
331 quantitative, anthropometric, and blood traits) in 46,979 individuals of European ancestry. For each gene-  
332 trait association, we also evaluated signal for the single variants included in the burden test. We identified  
333 25 unique gene burden-trait associations with  $p < 10^{-7}$ ; among these 21 were more significant than any  
334 single LOF variant included in the burden test. The results include several well-established associations  
335 (Table 5). For example, we observe that carriers of *MLH1* LOFs, associated with Muir-Torre and Lynch  
336 syndromes<sup>24</sup> [MIM #158320, #609310], were at increased risk of malignant neoplasms of the digestive  
337 organs (OR=84,  $P=3.5 \times 10^{-11}$ ). Carriers of *PKDI* LOFs, the major cause of autosomal dominant polycystic  
338 kidney disease<sup>25</sup> [MIM #173900], were at increased risk of chronic kidney disease (OR=91,  $P=2.9 \times 10^{-10}$ ).  
339 Carriers of *TTN* LOFs are at increased risk for cardiomyopathy (OR=11.9,  $P=1.4 \times 10^{-8}$ ), consistent  
340 with prior reports<sup>26</sup>. In addition to Mendelian disorders, other findings with strong support in the literature  
341 include *HBB* with red blood cell phenotypes<sup>27</sup>, *IL33* with eosinophils (driven by rs146597587)<sup>28</sup>, *KALRN*

342 with platelet volume (driven by rs56407180)<sup>29</sup>, *TUBB1* with multiple platelet phenotypes<sup>30</sup>, and *CALR*  
 343 with hematopoietic neoplasms<sup>31</sup>. In some cases, we see patterns of association with traits that may be  
 344 secondary to known phenotypic associations. For example, *ASXL1* and *CHEK2* are genes involved in  
 345 myeloproliferative disorders<sup>32</sup> and cancer<sup>33</sup>, respectively, which may explain the observed associations  
 346 with hematologic traits (which may be secondary to myelodysplastic disease or chemotherapy). Many  
 347 other known phenotypic associations are supported by the data at more modest significance thresholds  
 348 (Sup. table 6). These include, for example, associations between LOF variants in *LDLR* with coronary  
 349 artery disease, *GPIBB* with platelet count, *PALB2* with cancer, and *BRCA2* with cancer risk.

350

Gene	ICD10 Code, Binary Phenotype	RRIRAIAA	OR (95% CI)	WES Burden P	N SNV	Lowest P SNV	Imputed 50k Burden P
<i>MLH1</i>	Z85.0, Personal history of malignant neoplasm of digestive organs	Ctrl:39724 910 Cases:319 610	84.4 (31.0,229.5)	3.5x10 <sup>-11</sup>	9	0.88	NA
<i>PKD1</i>	N18, Chronic kidney disease	Ctrl:46389 1510 Cases:210 610	91.2 (36.3,229.1)	2.9x10 <sup>-10</sup>	15	NA	NA
<i>CALR</i>	D47, Other neoplasms of uncertain behavior of lymphoid, hematopoietic and related tissue	Ctrl:46552 310 Cases:52 310	866.1 (194.5,3857.1)	4.1x10 <sup>-8</sup>	4	NA	NA
<i>TTN</i>	I42, Cardiomyopathy	Ctrl:44341 585 3 Cases:67 1110	11.9 (6.5,21.9)	1.4x10 <sup>-8</sup>	302	1.2x10 <sup>-3</sup>	0.015

Gene	Quantitative Phenotype	RRIRAIAA	Beta (95% CI)	WES Burden P	N SNV	Lowest P SNV	Imputed 50k Burden P
<i>IL33</i>	Eosinophil percentage	44428 497 0	-0.3 (-0.4,-0.2)	5.4x10 <sup>-12</sup>	9	5.6x10 <sup>-12</sup>	1.4x10 <sup>-11</sup>
<i>IL33</i>	Eosinophil count	44423 500 0	-0.3 (-0.4,-0.2)	3.3x10 <sup>-10</sup>	9	3.7x10 <sup>-11</sup>	7.6x10 <sup>-9</sup>
<i>GPIBA</i>	Mean platelet thrombocyte volume	45509 98 1	0.5 (0.3,0.7)	6.4x10 <sup>-8</sup>	12	3.0x10 <sup>-5</sup>	0.50
<i>TUBB1</i>	Platelet distribution width	45540 28 0	1.8 (1.5,2.2)	2.5x10 <sup>-23</sup>	14	2.3x10 <sup>-6</sup>	5.4x10 <sup>-3</sup>
<i>TUBB1</i>	Mean platelet thrombocyte volume	45577 31 0	1.0 (0.6,1.3)	2.4x10 <sup>-8</sup>	14	2.1x10 <sup>-3</sup>	0.028

<i>TUBB1</i>	Platelet count	45422 30 0	-1.1 (-1.4,-0.7)	$2.1 \times 10^{-9}$	14	$1.1 \times 10^{-7}$	0.029
<i>HBB</i>	Red blood cell erythrocyte distribution width	45084 4 0	2.7 (1.7,3.6)	$5.8 \times 10^{-8}$	2	NA	0.76
<i>HBB</i>	Red blood cell erythrocyte count	45609 4 0	3 (2.0,3.9)	$1.7 \times 10^{-9}$	2	NA	0.23
<i>KLF1</i>	Red blood cell erythrocyte distribution width	45063 25 0	1.5 (1.1,1.8)	$1.5 \times 10^{-13}$	6	$4.4 \times 10^{-10}$	0.43
<i>KLF1</i>	Mean corpuscular haemoglobin	45350 27 0	-1.5 (-1.9,-1.2)	$1.7 \times 10^{-16}$	6	$8.8 \times 10^{-13}$	0.77
<i>KLF1</i>	Mean corpuscular volume	45448 27 0	-1.4 (-1.8,-1.1)	$4.0 \times 10^{-14}$	6	$1.1 \times 10^{-10}$	0.63
<i>ASXLI</i>	Platelet distribution width	45451 11 17 0	0.5 (0.34,0.7)	$4.7 \times 10^{-9}$	63	$4.1 \times 10^{-5}$	NA
<i>ASXLI</i>	Red blood cell erythrocyte distribution width	44974 11 14 0	0.6 (0.4,0.8)	$2.4 \times 10^{-11}$	63	$7.4 \times 10^{-6}$	NA
<i>CHEK2</i>	Platelet crit	45147 295 2	0.3 (0.2,0.4)	$7.9 \times 10^{-8}$	21	$1.2 \times 10^{-7}$	0.039
<i>KALRN</i>	Mean platelet thrombocyte volume	45374 233 1	-0.6 (-0.7,-0.5)	$2.7 \times 10^{-23}$	22	$1.1 \times 10^{-23}$	$1.9 \times 10^{-20}$

351  
352 **Table 5 | LOF gene burden results with previously known genetic associations.** LOF gene burden  
353 association with available clinical and continuous traits in 46,979 UKB participants of European ancestry  
354 with WES.

### 355 356 **LOF Associations and Novel Gene Discovery**

357 Our LOF gene burden association analysis identified five gene-trait LOF associations with  $p < 10^{-7}$   
358 that have not previously been reported (Table 6). We identified a novel association between *PIEZO1* LOFs  
359 (cumulative allele frequency = 0.2%) and increased risk for varicose veins (OR=4.8,  $P=2.7 \times 10^{-8}$ ). This  
360 finding is driven by a burden of rare LOF variants, with the most significant *PIEZO1* single variant LOF  
361 association achieving a p-value of  $2.3 \times 10^{-3}$ . ‘Leave-one-out’ (LOO) analyses indicated no single variant  
362 accounted for the entire signal and step-wise regression analyses indicated that 11 separate variants (5 of  
363 which had minor allele count (MAC)>1) were contributing to the overall burden signal (Sup. Table 8 and

364 Ext. Data SingleVariantLOFs.xlsx). We replicated this finding for varicose veins (1,572 cases, 75,704  
 365 controls) in WES data from the DiscovEHR study (OR= 3.8, p= 1.5x10<sup>-6</sup>) (Sup. Table 10). This region had  
 366 previously been implicated by common non-coding variants with small effects<sup>34</sup> (rs2911463 , OR = 0.996;  
 367 Supplemental Table 9). *PIEZO1* encodes a 36 transmembrane domain cation channel that is highly  
 368 expressed in the endothelium and plays a critical role in the development and adult physiology of the  
 369 vascular system, where it translates shear stress into electrical signals<sup>35</sup>. This previous report of the  
 370 rs2911463 variant mapped the association to *PIEZO1* through evidence of gene function and analysis using  
 371 DEPICT, but did not find strong eQTL evidence that would clarify mechanism to modulate the target  
 372 therapeutically<sup>34</sup>. Rare missense variants have been reported<sup>34</sup> in families segregating autosomal dominant  
 373 dehydrated hereditary stomatocytosis (DHS, MIM #194380) characterized by hemolytic anemia with  
 374 primary erythrocyte dehydration due to decreased osmotic fragility. Whereas biallelic loss of function  
 375 variants in *PIEZO1* have been reported in families with lymphatic malformation syndrome [MIM #616843],  
 376 a rare autosomal recessive disorder characterized by generalized lymphedema, intestinal and/or pulmonary  
 377 lymphangiectasia and pleural and/or pericardial effusions. Interestingly, some of the reported patients  
 378 presented with varicose veins and deep vein thrombosis<sup>36</sup>. Our data provide compelling support for *PIEZO1*  
 379 as the causal gene in this locus, but also clarifies direction of effect in that loss of gene function in  
 380 heterozygous carriers leads to increased risk of developing varicose veins.  
 381

Gene	ICD10 Code, Binary Phenotype	RRIRAIAA	OR (95% CI)	WES Burden P	N SNV	Lowest P SNV	Imputed 50k Burden P
<i>PIEZO1</i>	I83.9, Asymptomatic varicose veins of lower extremities	Ctrl:43285114210 Cases:126712010	4.8 (3.1,7.8)	2.7x10 <sup>-8</sup>	65	2.3x10 <sup>-3</sup>	0.083
Gene	Quantitative Phenotype	RRIRAIAA	Beta (95% CI)	WES Burden P	N SNV	Lowest P SNV	Imputed 50k Burden P
<i>MEPE</i>	Heel bone mineral density	42898115010	-0.5 (-0.6,-0.3)	1.4x10 <sup>-8</sup>	16	6.2x10 <sup>-5</sup>	0.033
<i>COL6A1</i>	Corneal resistance factor mean	3562111710	-1.5 (-2.0,-1.0)	3.6x10 <sup>-10</sup>	11	1.5x10 <sup>-4</sup>	NA <sup>1</sup>

<i>COL6A1</i>	Corneal hysteresis mean	3562011710	-1.3 (-1.8,-0.9)	$2.1 \times 10^{-8}$	11	$4.6 \times 10^{-4}$	NA <sup>1</sup>
<i>IQGAP2</i>	Mean platelet thrombocyte volume	45443116510	0.7 (0.5,0.8)	$1.1 \times 10^{-19}$	53	$9.3 \times 10^{-8}$	0.10
<i>GMPR</i>	Mean corpuscular haemoglobin	45195118210	0.4 (0.3,0.6)	$1.1 \times 10^{-8}$	13	$6.0 \times 10^{-8}$	$6.2 \times 10^{-6}$

382 **Table 6 | Novel LOF gene burden associations.** LOF gene burden association with available clinical and  
 383 continuous traits in 46,979 UKB participants of European ancestry with WES. <sup>1</sup>No *COL6A1* LOFs were  
 384 observed in imputed sequence.

385

386 We also identified a novel LOF burden association with *MEPE* (cumulative MAF 0.18%) and  
 387 decreased bone density (as measured by heel bone mineral density, -0.50 standard deviations (SD), p-value  
 388 =  $1.4 \times 10^{-8}$ ). LOO analyses (Sup. Table 11) suggested that the aggregate signal is driven by multiple variants,  
 389 one of which could be imputed (rs753138805, encoding a frameshift predicted to result in early protein  
 390 truncation). In analysis of all 500k UKB participants, rs753138805 was significantly associated with  
 391 decreased BMD (-0.4 SD,  $P=8.1 \times 10^{-19}$ ) and showed a trend for increased risk of osteoporosis (N=3,495  
 392 cases, OR=1.9,  $P=0.10$ , Sup. Table 16). These findings are corroborated by a query of the HUNT study<sup>37</sup>,  
 393 where rs753138805 was associated with decreased BMD (-0.5 SD,  $P=2.1 \times 10^{-18}$ ) and increased risk of any  
 394 fracture (N=24,155 cases,  $P=1.6 \times 10^{-5}$ , OR=1.4) (Sup. Table 16). In DiscovEHR, we observed a  
 395 directionally consistent, but nonsignificant, association for *MEPE* LOFs with femoral neck BMD T-score  
 396 ( $B = -0.19$ ,  $P = 0.22$ ). The *MEPE* locus was previously implicated in GWAS<sup>38</sup>, with six independent signals  
 397 with modest effects on bone mineral density. While two of the previously reported non-coding variants are  
 398 in moderate ( $r^2=0.5$ ) or high ( $r^2=0.78$ ) linkage disequilibrium (LD) with two variants contributing to the  
 399 burden test (Sup. Table 12), the burden association is only partially attenuated in conditional analyses  
 400 ( $p \sim 2 \times 10^{-4}$  with all 6 variants together). *MEPE* encodes a secreted calcium-binding phosphoprotein with a  
 401 key role in osteocyte differentiation and bone homeostasis<sup>39,40</sup>. Studies in *Mepe*<sup>-/-</sup> knockout mice  
 402 (*Mepe*<sup>tm1Tbrw</sup>) have yielded inconsistent results, with two groups reporting an increase in BMD<sup>40,41</sup> and  
 403 another reporting no change<sup>42</sup>.

404 In another novel signal, *COL6A1* LOFs (cumulative allele frequency = 0.03%) are associated with  
405 a 2.7 mmHg decrease in corneal resistance factor (CRF) (-1.5 SD, p-value =  $3.6 \times 10^{-10}$ ) and corneal  
406 hysteresis (CH) (-1.3 SD, p-value =  $2.1 \times 10^{-8}$ ), which are measures of corneal biomechanics<sup>43</sup>. *COL6A1*  
407 encodes a component of collagen type VI microfibrils, which play important roles in maintaining structure  
408 and function of the extracellular matrix, and which are major components of the human cornea<sup>44</sup>. This  
409 locus has previously been implicated in ocular traits; rs73157695 and nearby common variants have been  
410 associated with myopia<sup>45</sup> (OR = 0.94; p =  $\sim 10^{-13}$ ) and intraocular pressure<sup>46</sup>. LOO analyses indicate  
411 multiple variants are driving the association with both corneal traits (Sup. Tables 13, 14) and that these are  
412 not in strong LD with previously reported variants (Sup. Table 15). COL6A1 protein levels were reduced  
413 in eyes from patients with keratoconus<sup>47</sup>, and individuals with keratoconus and other corneal diseases such  
414 as Fuchs' corneal dystrophy have reduced CH and CRF<sup>48</sup>. Measures of CH and CRF were not available in  
415 DiscovEHR for replication analyses.

416 The remaining novel LOF associations in *IQGAP2* and *GMPR* (driven by rs147049568) are for  
417 hematologic traits in which variants in/near each gene have previously been implicated by GWAS<sup>29,49</sup>. Our  
418 results for these two genes, each of which replicated in DiscovEHR (Sup. Table 7), provide additional  
419 evidence for causal roles for these genes and establishes direction of effect with respect to gene function on  
420 hematologic traits. In equivalent sample sizes, LOF burden results from imputed sequence would not have  
421 uncovered the novel LOF associations at  $p < 10^{-7}$  as identified by WES (Table 6). Further, for 19 of 25 LOF  
422 burden results described herein, gene burden results from WES in 50k were more significant than burden  
423 results from imputed sequence in all 500k UKB participants (Sup. table 10), demonstrating the value of  
424 rare variants captured by WES to power these associations.

425

## 426 DISCUSSION

427 Integration of large scale genomic and precision medicine initiatives offer the potential to  
428 revolutionize medicine and healthcare. Such initiatives provide a foundation of knowledge linking genomic

429 and molecular data to health-related data at population scale, allowing for the ability to more completely  
430 and systematically study genetic variation and its functional consequences on health and disease. Here, we  
431 describe the initial tranche of large-scale exome sequencing of 49,960 UK Biobank participants, which to  
432 our knowledge is currently the largest open access resource of exome sequence data linked to health records  
433 and extensive longitudinal study measures. These data greatly extend the current genetic resource,  
434 particularly in ascertainment of rare coding variation, which we demonstrate has utility in resolving variant  
435 to gene links and directionality of gene to phenotype associations.

436         After quality control, we observed nearly four million single nucleotide and indel coding variants.  
437 Only approximately 14% of coding variants identified by WES were observed in the imputed sequence of  
438 49,797 participants with both WES and imputed sequence, highlighting the added value of exome  
439 sequencing. This enrichment was even more pronounced with LOF variation where WES identified  
440 >230,000 LOF variants and only approximately 5% of these were present in the imputed sequence. Further,  
441 22.6% of the coding variants in the imputed sequence were not observed in the exome sequence data which  
442 may represent a large proportion of rare variants that have poor imputation accuracy, as observed in our  
443 concordance and visual validation analyses. A small proportion of these variants, seen only in the imputed  
444 sequence, also represents variants not in the regions targeted by exome capture design and sequencing, a  
445 limitation of the targeted capture approach. Increasing numbers of individuals and ancestral diversity in  
446 imputation reference panels are expected to improve imputation accuracy for rare variants.

447         As with previous studies of this size<sup>22</sup>, we observed a large number of LOF variants, including at  
448 least one rare heterozygous LOF variant carrier in >97% of autosomal genes (compared to >36% of  
449 autosomal genes in the imputed sequence for the same participants). It is important to note that our LOF  
450 annotation strategy is geared towards increasing sensitivity for identification of LOF variants and novel  
451 downstream association discovery. While the number of genes with heterozygous instances of LOF variants  
452 is approaching saturation at this sample size, exome sequencing of the entirety of the UKB resource will  
453 dramatically increase the number of LOF carriers and the ability to detect phenotypic associations. We also



454 observe 1,071 autosomal genes with homozygous instances of rare LOF variants, and this number of genes  
455 will also increase with continued sequencing of all UKB participants; however, studies in populations with  
456 a high degree of parental relatedness<sup>50,51</sup> will provide yet more genes with homozygous LOFs and  
457 complement efforts such as UKB. LOF variation is an extremely important class of variation for identifying  
458 drivers of high genetic risk, novel disease genes, and therapeutic targets. Very large samples sizes are  
459 needed to detect novel LOF associations given their collective rare allele frequencies. The exome sequence  
460 data provides a substantial enhancement to the number of LOF variants identified and power for detecting  
461 novel associations, which will only improve with continued sequencing of all UKB participants.

462 We illustrate the unique value of this expanded exome sequence resource in the UKB to assess  
463 pathogenic and likely pathogenic variants in an unascertained large-scale population-based study with  
464 longitudinal follow up. We conducted a survey of pathogenic and likely pathogenic variants in the  
465 medically actionable ACMG59 genes. Using stringent variant filtering criteria, we arrived at an estimated  
466 prevalence of 2% of individuals in this study population having a clinically actionable finding. This  
467 resource allows us to characterize disease risk profiles for individuals who carry pathogenic and likely  
468 pathogenic variants in medically relevant disease genes, including cancer susceptibility genes, such as  
469 *BRCA1* and *BRCA2*. We observed that pathogenic and likely pathogenic variant carriers had 3.75-fold  
470 greater odds of any of the 5 cancers previously associated with *BRCA1/2*; prevalence for any of the 5 cancers  
471 was 21.0% in carriers vs 6.6% in non-carriers. We further explored whether these variants conferred risk to  
472 any other cancers and did not observe any such associations. This resource will be valuable for assessment  
473 of variant pathogenicity, particularly for variants of unknown significance and novel variants, and in  
474 exploring the full spectrum of disease risk and phenotypic expression. One limitation of the resource for  
475 such purposes is limited ancestral diversity. This and other similar studies also highlight the value and  
476 potential to apply large scale sequencing at the population scale to identify a meaningful proportion of  
477 individuals who are at high risk of diseases where effective interventions are available that can significantly

478 reduce the morbidity and mortality of genetically susceptible individuals; such precision medicine  
479 approaches could substantially reduce the burden of many diseases.

480 We conducted gene burden association testing for LOF variants across all genes and encompassing  
481 greater than 1,700 binary and quantitative traits. In addition to replication of numerous positive controls,  
482 we also identified a handful of significant novel LOF associations highlighting novel biology and genetics  
483 of large effect on disease traits of interest; this included *PIEZO1* for varicose veins, *MEPE* for bone density,  
484 and *COL6A1* for corneal thickness, amongst others. We identified a novel burden association in *PIEZO1*,  
485 a mechanosensing ion channel present in endothelial cells in vascular walls, that confers a nearly five-fold  
486 increased odds of varicose veins in heterozygous LOF carriers. We also identified a novel LOF burden  
487 association in *MEPE* with decreased BMD and an approximately 2-fold increased odds of osteoporosis and  
488 1.5-fold increased risk of fractures. Overall, through WES and gene burden tests of association for LOF  
489 variants, we identified 25 unique gene-trait associations exceeding a  $p < 10^{-7}$  of which 21 were substantially  
490 more significant than any single LOF variant included in the burden test, highlighting the value of WES  
491 and the ability to detect novel associations driven by rare coding variation. While these regions had  
492 previously been identified in genome-wide association studies of  $>10x$  the sample size, a key strength of  
493 the current approach is compelling identification of likely causal genes and the direction of effect: two key  
494 pieces of information required for translation towards novel therapeutics. This survey of rare LOF  
495 associations was limited by sample size for most binary traits but was well powered for many quantitative  
496 traits. While surveys of LOF variation in the entire UKB study using array and imputed sequence have  
497 identified LOF associations in previous reports<sup>52,53</sup>, WES identifies novel associations, unique to exome  
498 sequence and detected in only approximately one tenth of the sample size; this highlights the considerable  
499 power of exome sequencing for LOF and rare variant association discovery and the further promise of novel  
500 biological insights through sequencing all participants in the UKB resource.

501 Efforts are underway to sequence the exomes of all 500,000 UKB participants; these efforts will  
502 greatly expand the total amount of rare coding variation ascertained, including the number of heterozygous

503 LOF instances that can now be observed in nearly all genes and the number of genes for which naturally  
504 occurring homozygous knockouts can be observed. Coupled with rich laboratory, biomarker, health record,  
505 imaging, and other health related data continually added to the UKB resource, exome sequencing will  
506 enhance the power for discovery and will continue to yield many important findings and insights. The WES  
507 data is available to approved researchers through similar access protocols as existing UK Biobank data (see  
508 URLs).

509

## 510 **ACKNOWLEDGEMENTS**

511 The authors thank the UKBiobank participants and researchers for creating an open scientific  
512 resource for the research community; the MyCode Community Health Initiative participants for taking  
513 part in the DiscovEHR collaboration; and the participants of the Nord-Trøndelag Health Study (HUNT), a  
514 collaboration between the HUNT Research Centre (Faculty of Medicine, Norwegian University of  
515 Science and Technology NTNU), the Nord-Trøndelag County Council, the Central Norway Health  
516 Authority and the Norwegian Institute of Public Health. This work was referring UKB application 26041.

517

## 518 **URLs**

519 **UK Biobank website and data access** <http://ukbiobank.ac.uk/>

520

## 521 **REFERENCES**

- 522 1 Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range  
523 of complex diseases of middle and old age. *PLoS Med* **12**, e1001779,  
524 doi:10.1371/journal.pmed.1001779 (2015).
- 525 2 Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature*  
526 **562**, 203-209, doi:10.1038/s41586-018-0579-z (2018).
- 527 3 Tyrrell, J. *et al.* Height, body mass index, and socioeconomic status: mendelian randomisation  
528 study in UK Biobank. *BMJ* **352**, i582, doi:10.1136/bmj.i582 (2016).

- 529 4 Lyall, D. M. *et al.* Association of Body Mass Index With Cardiometabolic Disease in the UK  
530 Biobank: A Mendelian Randomization Study. *JAMA Cardiol* **2**, 882-889,  
531 doi:10.1001/jamacardio.2016.5804 (2017).
- 532 5 Abul-Husn, N. S. *et al.* A Protein-Truncating HSD17B13 Variant and Protection from Chronic  
533 Liver Disease. *N Engl J Med* **378**, 1096-1106, doi:10.1056/NEJMoa1712191 (2018).
- 534 6 Dewey, F. E. *et al.* Genetic and Pharmacologic Inactivation of ANGPTL3 and Cardiovascular  
535 Disease. *N Engl J Med* **377**, 211-221, doi:10.1056/NEJMoa1612790 (2017).
- 536 7 Cohen, J. C., Boerwinkle, E., Mosley, T. H., Jr. & Hobbs, H. H. Sequence variations in PCSK9,  
537 low LDL, and protection against coronary heart disease. *N Engl J Med* **354**, 1264-1272,  
538 doi:10.1056/NEJMoa054013 (2006).
- 539 8 Scott, R. A. *et al.* A genomic approach to therapeutic target validation identifies a glucose-  
540 lowering GLP1R variant protective for coronary heart disease. *Sci Transl Med* **8**, 341ra376,  
541 doi:10.1126/scitranslmed.aad3744 (2016).
- 542 9 Abul-Husn, N. S. *et al.* Genetic identification of familial hypercholesterolemia within a single  
543 U.S. health care system. *Science* **354**, doi:10.1126/science.aaf7000 (2016).
- 544 10 Manickam, K. *et al.* Exome Sequencing-Based Screening for BRCA1/2 Expected Pathogenic  
545 Variants Among Adult Biobank Participants. *JAMA Netw Open* **1**, e182140,  
546 doi:10.1001/jamanetworkopen.2018.2140 (2018).
- 547 11 Staples, J. *et al.* PRIMUS: rapid reconstruction of pedigrees from genome-wide estimates of  
548 identity by descent. *Am J Hum Genet* **95**, 553-564, doi:10.1016/j.ajhg.2014.10.005 (2014).
- 549 12 MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding  
550 genes. *Science* **335**, 823-828, doi:10.1126/science.1215040 (2012).
- 551 13 Gonzaga-Jauregui, C., Lupski, J. R. & Gibbs, R. A. Human Genome Sequencing in Health and  
552 Disease. *Annual Review of Medicine* **63**, 35-61, doi:10.1146/annurev-med-051010-162644  
553 (2012).

- 554 14 Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-  
555 performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178-192,  
556 doi:10.1093/bib/bbs017 (2013).
- 557 15 Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the  
558 deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47**, D886-D894,  
559 doi:10.1093/nar/gky1016 (2019).
- 560 16 Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for  
561 genome-wide association studies by imputation of genotypes. *Nat Genet* **39**, 906-913,  
562 doi:10.1038/ng2088 (2007).
- 563 17 Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype  
564 data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* **34**, 816-834,  
565 doi:10.1002/gepi.20533 (2010).
- 566 18 Cohen, J. *et al.* Low LDL cholesterol in individuals of African descent resulting from frequent  
567 nonsense mutations in PCSK9. *Nat Genet* **37**, 161-165, doi:10.1038/ng1509 (2005).
- 568 19 Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the  
569 spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*,  
570 doi:10.1101/531210 (2019).
- 571 20 Green, R. C. *et al.* ACMG recommendations for reporting of incidental findings in clinical exome  
572 and genome sequencing. *Genet Med* **15**, 565-574, doi:10.1038/gim.2013.73 (2013).
- 573 21 Kalia, S. S. *et al.* Recommendations for reporting of secondary findings in clinical exome and  
574 genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College  
575 of Medical Genetics and Genomics. *Genet Med* **19**, 249-255, doi:10.1038/gim.2016.190 (2017).
- 576 22 Dewey, F. E. *et al.* Distribution and clinical impact of functional variants in 50,726 whole-exome  
577 sequences from the DiscovEHR study. *Science* **354**, doi:10.1126/science.aaf6814 (2016).
- 578 23 Buchanan, A. H. *et al.* Early cancer diagnoses through BRCA1/2 screening of unselected adult  
579 biobank participants. *Genet Med* **20**, 554-558, doi:10.1038/gim.2017.145 (2018).

- 580 24 Dowty, J. G. *et al.* Cancer risks for MLH1 and MSH2 mutation carriers. *Hum Mutat* **34**, 490-497,  
581 doi:10.1002/humu.22262 (2013).
- 582 25 Halvorson, C. R., Bremmer, M. S. & Jacobs, S. C. Polycystic kidney disease: inheritance,  
583 pathophysiology, prognosis, and treatment. *Int J Nephrol Renovasc Dis* **3**, 69-83 (2010).
- 584 26 Herman, D. S. *et al.* Truncations of titin causing dilated cardiomyopathy. *N Engl J Med* **366**, 619-  
585 628, doi:10.1056/NEJMoa1110186 (2012).
- 586 27 Rund, D. & Rachmilewitz, E. Beta-thalassemia. *N Engl J Med* **353**, 1135-1146,  
587 doi:10.1056/NEJMra050436 (2005).
- 588 28 Smith, D. *et al.* A rare IL33 loss-of-function mutation reduces blood eosinophil counts and  
589 protects from asthma. *PLoS Genet* **13**, e1006659, doi:10.1371/journal.pgen.1006659 (2017).
- 590 29 Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to  
591 Common Complex Disease. *Cell* **167**, 1415-1429 e1419, doi:10.1016/j.cell.2016.10.042 (2016).
- 592 30 Johnson, B., Fletcher, S. J. & Morgan, N. V. Inherited thrombocytopenia: novel insights into  
593 megakaryocyte maturation, proplatelet formation and platelet lifespan. *Platelets* **27**, 519-525,  
594 doi:10.3109/09537104.2016.1148806 (2016).
- 595 31 Nangalia, J. *et al.* Somatic CALR mutations in myeloproliferative neoplasms with nonmutated  
596 JAK2. *N Engl J Med* **369**, 2391-2405, doi:10.1056/NEJMoa1312542 (2013).
- 597 32 Carbuccia, N. *et al.* Mutations of ASXL1 gene in myeloproliferative neoplasms. *Leukemia* **23**,  
598 2183-2186, doi:10.1038/leu.2009.141 (2009).
- 599 33 Meijers-Heijboer, H. *et al.* Low-penetrance susceptibility to breast cancer due to  
600 CHEK2(\*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nat Genet* **31**, 55-59,  
601 doi:10.1038/ng879 (2002).
- 602 34 Shadrina, A. S., Sharapov, S. Z., Shashkova, T. I. & Tsepilov, Y. A. Varicose veins of lower  
603 extremities: insights from the first large-scale genetic study. *BioRxive*, doi:10.1101/368365  
604 (2018).

- 605 35 Li, J. *et al.* Piezo1 integration of vascular architecture with physiological force. *Nature* **515**, 279-  
606 282, doi:10.1038/nature13701 (2014).
- 607 36 Fotiou, E. *et al.* Novel mutations in PIEZO1 cause an autosomal recessive generalized lymphatic  
608 dysplasia with non-immune hydrops fetalis. *Nat Commun* **6**, 8085, doi:10.1038/ncomms9085  
609 (2015).
- 610 37 Krokstad, S. *et al.* Cohort Profile: the HUNT Study, Norway. *Int J Epidemiol* **42**, 968-977,  
611 doi:10.1093/ije/dys095 (2013).
- 612 38 Estrada, K. *et al.* Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals  
613 14 loci associated with risk of fracture. *Nat Genet* **44**, 491-501, doi:10.1038/ng.2249 (2012).
- 614 39 Plotkin, L. I. & Bellido, T. Osteocytic signalling pathways as therapeutic targets for bone  
615 fragility. *Nat Rev Endocrinol* **12**, 593-605, doi:10.1038/nrendo.2016.71 (2016).
- 616 40 Zelenchuk, L. V., Hedge, A. M. & Rowe, P. S. Age dependent regulation of bone-mass and renal  
617 function by the MEPE ASARM-motif. *Bone* **79**, 131-142, doi:10.1016/j.bone.2015.05.030  
618 (2015).
- 619 41 Gowen, L. C. *et al.* Targeted disruption of the osteoblast/osteocyte factor 45 gene (OF45) results  
620 in increased bone formation and bone mass. *J Biol Chem* **278**, 1998-2007,  
621 doi:10.1074/jbc.M203250200 (2003).
- 622 42 Liu, S. *et al.* Role of matrix extracellular phosphoglycoprotein in the pathogenesis of X-linked  
623 hypophosphatemia. *J Am Soc Nephrol* **16**, 1645-1653, doi:10.1681/ASN.2004121060 (2005).
- 624 43 Garcia-Porta, N. *et al.* Corneal biomechanical properties in different ocular conditions and new  
625 measurement techniques. *ISRN Ophthalmol* **2014**, 724546, doi:10.1155/2014/724546 (2014).
- 626 44 Zimmermann, D. R., Trüeb, B., Winterhalter, K. H., Witmer, R. & Fischer, R. W. Type VI  
627 collagen is a major component of the human cornea. *FEBS Letters* **197**, 55-58, doi:10.1016/0014-  
628 5793(86)80297-6 (1986).
- 629 45 Pickrell, J. K. *et al.* Detection and interpretation of shared genetic influences on 42 human traits.  
630 *Nat Genet* **48**, 709-717, doi:10.1038/ng.3570 (2016).

- 631 46 Choquet, H. *et al.* A large multi-ethnic genome-wide association study identifies novel genetic  
632 loci for intraocular pressure. *Nat Commun* **8**, 2108, doi:10.1038/s41467-017-01913-6 (2017).
- 633 47 Chaerkady, R. *et al.* The keratoconus corneal proteome: loss of epithelial integrity and stromal  
634 degeneration. *J Proteomics* **87**, 122-131, doi:10.1016/j.jprot.2013.05.023 (2013).
- 635 48 del Buey, M. A., Cristobal, J. A., Ascaso, F. J., Lavilla, L. & Lanchares, E. Biomechanical  
636 properties of the cornea in Fuchs' corneal dystrophy. *Invest Ophthalmol Vis Sci* **50**, 3199-3202,  
637 doi:10.1167/iovs.08-3312 (2009).
- 638 49 Eicher, J. D. *et al.* Platelet-Related Variants Identified by Exomechip Meta-analysis in 157,293  
639 Individuals. *Am J Hum Genet* **99**, 40-55, doi:10.1016/j.ajhg.2016.05.005 (2016).
- 640 50 Finer, S. *et al.* Cohort Profile: East London Genes & Health (ELGH), a community based  
641 population genomics and health study of British-Bangladeshi and British-Pakistani people.  
642 *bioRxiv*, doi:10.1101/426163 (2019).
- 643 51 Saleheen, D. *et al.* Human knockouts and phenotypic analysis in a cohort with a high rate of  
644 consanguinity. *Nature* **544**, 235-239, doi:10.1038/nature22034 (2017).
- 645 52 McInnes, G. *et al.* Global Biobank Engine: enabling genotype-phenotype browsing for biobank  
646 summary statistics. *Bioinformatics*, doi:10.1093/bioinformatics/bty999 (2018).
- 647 53 Emdin, C. A. *et al.* Analysis of predicted loss-of-function variants in UK Biobank identifies  
648 variants protective for disease. *Nat Commun* **9**, 1613, doi:10.1038/s41467-018-03911-8 (2018).

649

650 Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints)

651

## 652 **AUTHOR INFORMATION**

### 653 **Contributions**

654 C.V.H., M.R.N., J.W., J.G.R., J.M., J.D.O., R.A.S., L.Y-A, G.A., A.B., directed and designed

655 research; J.B., B.Y., D.L., A.H.L., A.M., C.G-J., C.O., C.V.H., I.T., J.M. contributed to statistical



656 analyses; C.G.-J., W.C., S.B., B.Y., S.K., J.S., A.L.B., C.V.H. contributed to the medically actionable  
657 variants survey and cancers analysis, N.B., J.X.H., B.Y., A.Y., S.K., A.H., S.B., M.C., contributed to  
658 the preparation of genetic and phenotype data; E.M., L.B., A.L., L.H., J.P., W.J.S., J.G.R., J.D.O.  
659 contributed to exome sequencing and variant calling; I.S., C.J.W., K.H., J.B.L., D.J.C., D.H.L.  
660 contributed data or results for replication; C.V.H., J.B., I.T., L.Y-A., C.G.J., C.O., S.B., N.B., R.A.S.,  
661 J.M., J.R., G.A., A.B. co-wrote the manuscript. All authors reviewed the manuscript.

## 662 **Competing interests**

663 C.V.H., J.D.B., B.Y., C.G.J., S.K., D.L., N.B., A.H.L., C.O., A.M., J.S., C.S., A.H., E.M., L.B., A.L., J.P.,  
664 L.H., A.L.B., A.Y., K.P., M.J., W.J.S., G.D.Y., A.E., G.C., A.R.S., S.B., M.C., J.G.R., J.M., J.D.O., G.A.,  
665 A.B. are current or former employees and/or stockholders of Regeneron Genetics Center or Regeneron  
666 Pharmaceuticals.

667 I.T., J.X.H., A.P., L.C., M.R.N., J.W., R.A.S., L.Y-A are current or former employees and/or stockholders  
668 of GlaxoSmithKline.

669 L.C. is a current employee of BioMarin.

670 No other authors declare a competing interest.

671

## 672 **Regeneron Genetics Center Banner Author List and Contribution Statements**

673  
674 All authors/contributors are listed in alphabetical order.

675

### 676 RGC Management and Leadership Team:

677 Goncalo Abecasis, Ph.D., Aris Baras, M.D., Michael Cantor, M.D., Giovanni Coppola, M.D., Aris  
678 Economides, Ph.D., John D. Overton, Ph.D., Jeffrey G. Reid, Ph.D., Alan Shuldiner, M.D.

679

680 Contribution: All authors contributed to securing funding, study design and oversight, and review and  
681 interpretation of data and results. All authors reviewed and contributed to the final version of the  
682 manuscript.

683

### 684 Sequencing and Lab Operations:

685 Christina Beechert, Caitlin Forsythe, M.S., Erin D. Fuller, Zhenhua Gu, M.S., Michael Lattari, Alexander  
686 Lopez, M.S., John D. Overton, Ph.D., Thomas D. Schleicher, M.S., Maria Sotiropoulos Padilla, M.S.,  
687 Karina Toledo, Louis Widom, Sarah E. Wolf, M.S., Manasi Pradhan, M.S., Kia Manoochehri, Ricardo H.  
688 Ulloa.

689

690 Contribution: C.B., C.F., K.T., A.L., and J.D.O. performed and are responsible for sample genotyping.  
691 C.B, C.F., E.D.F., M.L., M.S.P., K.T., L.W., S.E.W., A.L., and J.D.O. performed and are responsible for  
692 exome sequencing. T.D.S., Z.G., A.L., and J.D.O. conceived and are responsible for laboratory  
693 automation. M.P., K.M., R.U., and J.D.O are responsible for sample tracking and the library information  
694 management system.

695

696 Genome Informatics:

697 Xiaodong Bai, Ph.D., Suganthi Balasubramanian, Ph.D., Leland Barnard, Ph.D., Andrew Blumenfeld,  
698 Yating Chai, Ph.D., Gisu Eom, Lukas Habegger, Ph.D., Young Hahn, Alicia Hawes, B.S., Shareef  
699 Khalid, Jeffrey G. Reid, Ph.D., Evan K. Maxwell, Ph.D., John Penn, M.S., Jeffrey C. Staples, Ph.D.,  
700 Ashish Yadav, M.S.

701

702 Contribution: X.B., A.H., Y.C., J.P., and J.G.R. performed and are responsible for analysis needed to  
703 produce exome and genotype data. G.E., Y.H., and J.G.R. provided compute infrastructure development  
704 and operational support. S.K., S.B., and J.G.R. provide variant and gene annotations and their functional  
705 interpretation of variants. E.M., L.B., J.S., A.B., A.Y., L.H., J.G.R. conceived and are responsible for  
706 creating, developing, and deploying analysis platforms and computational methods for analyzing genomic  
707 data.

708

709 Clinical Informatics:

710 Nilanjana Banerjee, Ph.D., Michael Cantor, M.D.

711

712 Contribution: All authors contributed to the development and validation of clinical phenotypes used to  
713 identify study participants and (when applicable) controls.

714

715 Analytical Genomics and Data Science:

716 Goncalo Abecasis, Ph.D., Amy Damask, Ph.D., Manuel Allen Revez Ferreira, Ph.D., Lauren Gurski,  
717 Alexander Li, Ph.D., Nan Lin, Ph.D., Daren Liu, Jonathan Marchini Ph.D., Anthony Marcketta, Shane  
718 McCarthy, Ph.D., Colm O'Dushlaine, Ph.D., Charles Paulding, Ph.D., Claudia Schurmann, Ph.D., Dylan  
719 Sun, Cristopher Van Hout, Ph.D., Bin Ye

720

721 Contribution: Development of statistical analysis plans. QC of genotype and phenotype files and  
722 generation of analysis ready datasets. Development of statistical genetics pipelines and tools and use  
723 thereof in generation of the association results. QC, review and interpretation of result. Generation and  
724 formatting of results for manuscript figures. Contributions to the final version of the manuscript.

725

726 Therapeutic Area Genetics:

727 Jan Freudenberg, M.D., Nehal Gosalia, Ph.D., Claudia Gonzaga-Jauregui, Ph.D., Julie Horowitz, Ph.D.,  
728 Kavita Praveen, Ph.D.

729

730 Contribution: Development of study design and analysis plans. Development and QC of phenotype  
731 definitions. QC, review, and interpretation of association results. Contributions to the final version of the  
732 manuscript.

733

734 Planning, Strategy, and Operations:

735 Paloma M. Guzzardo, Ph.D., Marcus B. Jones, Ph.D., Lyndon J. Mitnaul, Ph.D.

736

737 Contribution: All authors contributed to the management and coordination of all research activities,  
738 planning and execution. All authors managed the review of data and results for the manuscript. All  
739 authors contributed to the review process for the final version of the manuscript.