

# Exome-wide evidence of compound heterozygous effects across common phenotypes in the UK Biobank

Frederik H. Lassen<sup>1,2,\*</sup>, Samvida S. Venkatesh<sup>1,2</sup>, Nikolas Baya<sup>1,2</sup>, Wei Zhou<sup>4,5,6</sup>, Alex Bloemendal<sup>4,7,8</sup>, Benjamin M. Neale<sup>4,5,6</sup>, Benedikt M. Kessler<sup>3</sup>, Nicola Whiffin<sup>1,2,4</sup>, Cecilia M. Lindgren<sup>1,2,9†</sup> and Duncan S. Palmer<sup>2,†,\*</sup>

<sup>1</sup> Wellcome Trust Centre for Human Genetics,  
University of Oxford, Oxford, United Kingdom.

<sup>2</sup> Big Data Institute, Li Ka Shing Centre for Health Information and Discovery,  
University of Oxford, Oxford, United Kingdom.

<sup>3</sup> Target Discovery Institute, Centre for Medicines Discovery, Nuffield Department of Medicine,  
University of Oxford, Oxford, United Kingdom.

<sup>4</sup> Program in Medical and Population Genetics  
Broad Institute of MIT and Harvard, Cambridge, MA, USA.

<sup>5</sup> Stanley Center for Psychiatric Research  
Broad Institute of MIT and Harvard, Cambridge, MA, USA.

<sup>6</sup> Analytical and Translational Genetics Unit, Department of Medicine  
Massachusetts General Hospital, Boston, MA, USA.

<sup>7</sup> Novo Nordisk Center for Genomic Mechanisms of Disease  
Broad Institute of MIT and Harvard, Cambridge, MA, USA.

<sup>8</sup> Data Sciences Platform  
Broad Institute of MIT and Harvard, Cambridge, MA, USA.

<sup>9</sup> Nuffield Department of Population Health Health, Medical Sciences Division  
University of Oxford, Oxford, United Kingdom.

\*Correspondence to: [flassen@well.ox.ac.uk](mailto:flassen@well.ox.ac.uk), [cecilia.lindgren@wrh.ox.ac.uk](mailto:cecilia.lindgren@wrh.ox.ac.uk)  
and [duncan.stuart.palmer@gmail.com](mailto:duncan.stuart.palmer@gmail.com)

†These authors contributed equally.

## 1 Abstract

2 Exome-sequencing association studies have successfully linked rare protein-coding variation to  
3 risk of thousands of diseases. However, the relationship between rare deleterious compound  
4 heterozygous (CH) variation and their phenotypic impact has not been fully investigated. Here,  
5 we leverage advances in statistical phasing to accurately phase rare variants ( $MAF \sim 0.001\%$ )  
6 in exome sequencing data from 175,587 UK Biobank (UKBB) participants, which we then  
7 systematically annotate to identify putatively deleterious CH coding variation. We show that  
8 6.5% of individuals carry such damaging variants in the CH state, with 90% of variants occurring  
9 at  $MAF < 0.34\%$ . Using a logistic mixed model framework, systematically accounting for  
10 relatedness, polygenic risk, nearby common variants, and rare variant burden, we investigate  
11 recessive effects in common complex diseases. We find six exome-wide significant ( $P <$   
12  $1.68 \times 10^{-7}$ ) and 17 nominally significant ( $P < 5.25 \times 10^{-5}$ ) gene-trait associations. Among  
13 these, only four would have been identified without accounting for CH variation in the gene. We  
14 further incorporate age-at-diagnosis information from primary care electronic health records,  
15 to show that genetic phase influences lifetime risk of disease across 20 gene-trait combinations  
16 ( $FDR < 5\%$ ). Using a permutation approach, we find evidence for genetic phase contributing to  
17 disease susceptibility for a collection of gene-trait pairs, including *FLG*-asthma ( $P = 0.00205$ )  
18 and *USH2A*-visual impairment ( $P = 0.0084$ ). Taken together, we demonstrate the utility of  
19 phasing large-scale genetic sequencing cohorts for robust identification of the phenome-wide  
20 consequences of compound heterozygosity.

## 21 Main

22 Thousands of independent genetic variants have been robustly associated with common, com-  
23 plex human diseases, leading to important advancements in therapeutic development<sup>1</sup>. Naturally

24 occurring variants that disrupt protein-coding sequences are of interest in the context of drug  
25 discovery as they modulate potential biological targets with measurable effects on human phys-  
26 iology<sup>2,3</sup>. Thus, individuals who carry loss-of-function (LoF) variants on both the maternal and  
27 paternal copy of a gene, are in principle ‘experiments of nature’ and their identification can help  
28 to determine causality between gene function and phenotype<sup>4–6</sup>.

29 Coding variants in a gene can either be homozygous, when both gene copies harbor the same  
30 variant, or CH when both copies harbor different variants, usually at distinct genetic locations  
31 within the same gene locus. Alternatively, when two variants are located on a single gene  
32 copy, they are said to be ‘in *cis*’. Although both copies of a gene are disrupted in two-  
33 hit (CH or homozygous) carriers, analyses of the phenotypic impact of coding variation have  
34 typically ignored genetic phase information, that is, the separation or ‘phasing’ of an individual’s  
35 genome into maternally and paternally derived alleles<sup>7,8</sup>. Large-scale studies of bi-allelic  
36 damaging variation have generally been restricted to homozygotes in populations with excess  
37 homozygosity, such as Icelanders<sup>9</sup>, Finns<sup>10,11</sup>, and consanguineous populations<sup>12</sup>. In contrast,  
38 CH are expected to be more common in outbred populations, but are largely under-studied  
39 outside of rare disorders<sup>13–17</sup>.

40 Various methods exist to infer the genetic phase of two variants. ‘Phasing by transmission’ em-  
41 ploys family member genotyping and Mendelian inheritance principles<sup>18</sup>, while ‘read-backed  
42 phasing’ utilizes physical relationships among variants within sequencing reads<sup>19</sup>. In large-scale  
43 biobanks, extensively genotyping family members is impractical, and short-read sequencing  
44 technologies only allow read-backed phasing for variants in close proximity. Therefore, ‘sta-  
45 tistical phasing’, which models the generative process of newly arising genetic data subject to  
46 recombination and mutation<sup>18,20–23</sup>, is typically used to phase haplotypes in genetic biobank  
47 data. Obtaining high-quality statistically phased genetic data requires large sample sizes, (10<sup>5</sup>-

48  $10^6$  individuals), and tends to require large reference panels<sup>21</sup>. Furthermore, statistical phasing  
49 is more error prone for rare variants, which are precisely the collection of variants that we  
50 would like to investigate as they are *a priori* more likely to be deleterious variants of large effect  
51 under purifying selection. This difficulty in the accurate statistical phasing of rare variation has  
52 historically deterred the analysis of CH variants in biobanks. However, recent advancements  
53 in statistical phasing<sup>24</sup>, achieved by combining common variation across genotyping arrays and  
54 exome sequencing to create haplotype ‘scaffolds’<sup>22</sup>, enables accurate phasing of rare variants.  
55 By using this new accurate phase information which extends down into rare allele frequencies,  
56 we can identify damaging CH variants to expand the pool of identifiable two-hit carriers and  
57 screen for phenotypic consequences.

58 We describe and apply a systematic analytical approach to test for autosomal bi-allelic effects,  
59 gene-by-gene, across 311 traits in the UKBB 200k exome sequencing (ES) release, combining  
60 both CH and homozygous variation. We iteratively refine the candidate associations by adjusting  
61 for polygenic background, nearby common variant risk, and rare variant burden within the  
62 analyzed gene. Our approach identifies both known and novel bi-allelic-trait associations,  
63 providing important insights into the phenotypic impact of gene knockout in humans.

## 64 **Results**

### 65 **Accurate phase inference and validation using parent-offspring trios and** 66 **short-read sequences**

67 We identified 13,377,336 high-quality variants in 176,935 individuals exome sequenced in the  
68 UKBB (Methods). To identify variants co-occurring on the same haplotype (in *cis*) or on  
69 opposite haplotypes (in *trans*) gene-by-gene, we jointly phased ES and genotype array data in  
70 the UKBB using SHAPEIT5<sup>25</sup> (Methods) following an investigation into the performance of  
71 current phasing software (Supplementary Table 4). Rare variants (MAF < 0.001) are assigned

72 a posterior probability (PP) of true haplotype assignment, known as the phasing confidence  
73 score. Confidence in our ability to accurately statistically phase variants decreases with MAF  
74 (Supplementary Fig. 5). However, we *a priori* expect a disproportionate recessive damaging  
75 signal to reside in CH variants with at least one rare variant, and as a result, choosing a PP cutoff  
76 represents a trade-off in the signal to noise ratio. Following phasing, we restrict to 176,587  
77 individuals of genetically-ascertained non-Finnish European (NFE) ancestry (Methods).

78 To assess statistical phasing quality, we benchmarked against phasing determined with parent-  
79 offspring trio data and read-backed phasing. We quantified phasing quality before and after  
80 filtering by  $PP \geq 0.9$  in 96 parent-offspring trios by calculating switch error rates (SER),  
81 estimated using Mendelian transmission, across 2,044,234 unique variants stratified by minor  
82 allele count (MAC) (Fig. 1a, Supplementary Fig. 6, Supplementary Tables 5-6). Across the  
83 96 children, 93.1% of protein coding genes contained variants that were phased without switch  
84 errors (Supplementary Table 7). SERs among singletons ( $MAC=1$ ) and variants with  $2 \leq MAC$   
85  $\leq 5$  were 12.1% (95% CI = 8.42 – 17.2) and 0.27% (0.13–0.53), respectively (Fig. 1a).

86 Although calculation of SER using trios is the gold-standard approach for phasing quality  
87 estimation<sup>21</sup>, it is limited by the number of parent-offspring trios available. For this reason, we  
88 also performed read-backed phasing of 62,762 unique pairs of variants using UKBB short read  
89 sequences on chromosomes 20-22 in 176,586 NFE individuals using WhatsHap<sup>26</sup> (Methods).  
90 While read-backed phasing only permits ascertainment of genetic phase among pairs of variants  
91 spanning one or a few overlapping short read sequences (with typical lengths of 150-300 bp),  
92 read-backed phasing accuracy is independent of allele frequency, and therefore represents an  
93 orthogonal approach for evaluating the quality of statistically phased variation. Consistent  
94 with trio-SER, we observed increasing agreement between pairs of statistically and read-backed  
95 phased variants with increasing MAC (Supplementary Fig. 7, Supplementary Table 8). Filtering

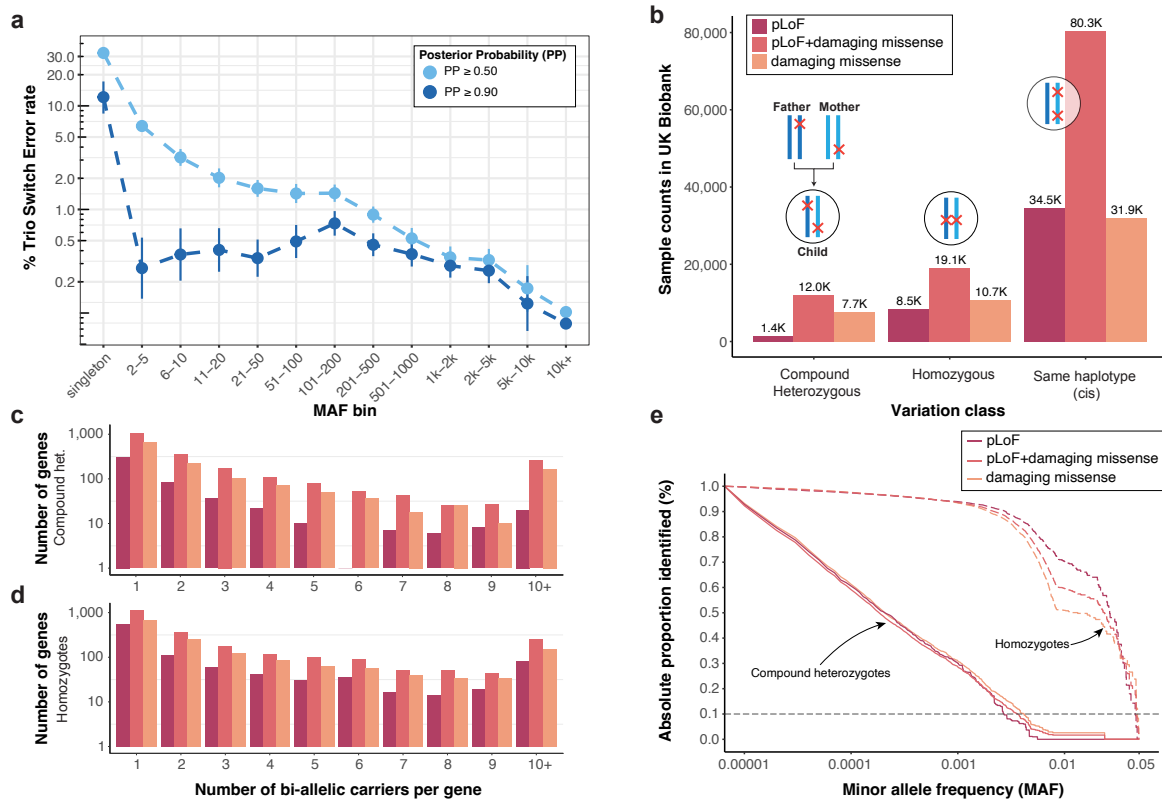
96 to phased variants with  $PP \geq 0.9$ , singletons and variants with  $2 \leq MAC \leq 5$ , agreement between  
97 read-backed phasing and statistical phasing was 85.1% (95% CI = 83.7 – 86.3%) and 99.1%  
98 (95% CI = 98.98 – 99.16%) respectively (Supplementary Table 8, Supplementary Fig. 8).

99 Taken together, our benchmarking suggests that statistical phasing of the UKBB dataset is of  
100 high quality for rare to ultra-rare variants, increasing our confidence in the identification of  
101 damaging CH variation. Given our observations of well-calibrated PP, and the distribution of  
102 phasing confidence binned by MAC, we selected the empirical cutoff of  $PP \geq 0.9$  to retain  
103 8,616,236 variants (43% of which are singletons) for downstream characterization and testing  
104 (Supplementary Fig. 5).

## 105 **Identification and examination of CH variation in the UKBB**

106 To interrogate the functional role of mono- and bi-allelic variation in the population, we anno-  
107 tated 8,616,236 variants with  $PP \geq 0.9$  and  $MAF \leq 5\%$  across 17,998 autosomal protein-coding  
108 genes. We enriched our search for variants with putatively large effect sizes by restricting analyses  
109 to two categories of predicted damaging variation: first, we annotated 289,981 high-confidence  
110 protein truncating variants, including stop-gain, essential splice and frame-shift variants iden-  
111 tified as high-confidence by Loss-Of-Function Transcript Effect Estimator (LOFTEE)<sup>27</sup>, which  
112 we refer to as ‘putative loss-of-function (pLoF) variants’. Second, we annotated 444,804 mis-  
113 sense variants classified as damaging by both REVEL score  $\geq 0.6$  and Phred scaled Combined  
114 Annotation Dependent Depletion (CADD) score  $\geq 20$ , or LOFTEE low confidence (LC) protein  
115 truncating variants; we refer to these variants collectively as ‘damaging missense/protein alter-  
116 ing’ (Supplementary Fig. 9, Supplementary Table 9). For each individual, we then determined  
117 the set of genes predicted to be affected by pLoFs+damaging missense/protein-altering variants  
118 in a CH, homozygous or in *cis* state on the same haplotype.

119 As we *a priori* expected that essential genes would be less permissible to bi-allelic damaging vari-



**Fig. 1: CH variants composed of at least one ultra-rare variant ( $MAC \leq 10$ ) can be robustly identified in large scale biobanks.** a) Trio SER depicted on y-axis as a function of MAC bin (x-axis) for phased variants with  $MAF \leq 5\%$ , stratified by phasing confidence score  $PP \geq 0.5$  or  $PP \geq 0.9$ . b) Counts of samples harboring different classes of variation with at least two variants in UKBB. Each set of three bars depicts the number of individuals with at least one CH variant, homozygous variant, or multi-hit (*cis*) variant, respectively. Here, we define a CH pLoF+damaging missense variant as any combination of pLoF and/or damaging missense variation on opposite haplotypes. A qualifying carrier for each bar occurs according to the configuration displayed above the bars, and is grouped by variant consequence according to the color legend. c-d) Number of CH or homozygous carriers per gene. e) 1 - cumulative fraction (y-axis) of homozygous (dashed line) and CH carriers as a function of lowest MAF (x-axis) in bi-allelic variant pairs for which both variants phased at  $PP \geq 0.9$  (solid line), stratified by variant consequence according to the color key.

120 ants when compared to non-essential genes, we investigated tolerance towards predicted bi-allelic  
121 pLoF and pLoF+damaging missense/protein-altering variants across the genome. As some genes  
122 carry bi-allelic variants more often than others (owed to a variety of factors such as gene length  
123 and baseline mutation frequency<sup>28</sup>), we fit counts of the number of individuals carrying bi-allelic  
124 variants per gene using a Poisson regression model accounting for variation in gene length  
125 and mutation rate (Methods, Supplementary Tables 10-11). Both pLoF and pLoF+damaging  
126 missense/protein-altering bi-allelic variants (homozygous and CH) were significantly depleted  
127 in five of the six analyzed essential gene-sets ( $P < \frac{0.05}{6} \approx 0.0083$ ) (Supplementary Fig. 12).  
128 Conversely, across three non-essential gene-sets, bi-allelic pLoFs+damaging missense/protein-  
129 altering variants were enriched among LoF tolerant genes<sup>27</sup> ( $P \leq \frac{0.05}{3} \approx 0.0167$ ). We found  
130 that the degree and direction of effects were consistent across CH, homozygous bi-allelic, and  
131 heterozygous variants (Supplementary Fig. 12).

132 In founder<sup>9</sup> and bottle-necked<sup>10</sup> populations, some alleles are enriched to high frequency by  
133 chance, resulting in better powered association studies for the subset of variant alleles that are  
134 inherited from the parental population<sup>6</sup> at higher frequency. To explore the diversity of bi-allelic  
135 variation in UKBB, a largely outbred population, we enumerated two-hit carriers across 176,587  
136 individuals. We observed complete bi-allelic ‘knockout’ of 1,174 unique genes strictly owed  
137 to pLoF variants, identifying 1,431 (0.8%) CH and 8,582 (4.8%) homozygous individuals with  
138 bi-allelic pLoF variants in at least one gene (Fig. 1b). Across genes, 307 (26.1%) CH and 560  
139 (47.7%) homozygous ‘knockouts’ were observed only in a single individual (Fig. 1c-d). We  
140 reasoned that inclusion of damaging missense/protein-altering variants in addition to pLoFs,  
141 would expand the number of identifiable damaging bi-allelic variants compared to assessing the  
142 two categories independently. Across 3,288 unique genes, we observed 11,491 (6.5%) CH and  
143 17,863 (10.1%) homozygous carriers of pLoF+damaging missense/protein-altering variants. Of  
144 these, 1,112 (0.6%) CH and 436 (0.2%) homozygotes were carriers of bi-allelic pLoF+damaging



145 missense/protein-altering variants in genes linked to traits with autosomal recessive mode of  
146 inheritance in Online Mendelian Inheritance in Man (OMIM)<sup>29</sup>. We generally observed a higher  
147 prevalence of carriers with variants in *cis* compared to CH, with over a third of individuals  
148 (64,555, 36.6%) carrying  $\geq 2$  pLoF+damaging missense/protein-altering variants on a single  
149 haplotype (Fig. 1b).

150 To better understand the evolutionary dynamics giving rise to pathogenic variants in *trans*, we  
151 examined the spectrum of allele frequencies of the constituent variants among our confidently  
152 called damaging CHs variants. CHs variants tend to comprise of two variants where one resides  
153 on a common haplotype, while the other on a rare haplotype, with a median difference in MAC  
154 of 1,181 (Supplementary Fig. 13-14). Approximately 90% of CH-constituent variants have  
155  $MAF \leq 0.0038$ , compared to homozygotes in which 90% are detected at  $MAF \geq 0.0022$  (Fig.  
156 1d), suggesting that identifying deleterious bi-allelic CH variants requires phasing of rare alleles  
157 (Supplementary Fig. 15-16).

158 Multiple studies have assessed the prospects of ascertaining bi-allelic LoF variation at larger  
159 sample sizes in consanguineous, bottle-necked, and outbred populations<sup>6,12</sup>. To investigate em-  
160 pirically how the number of unique genes with bi-allelic variants scales in an outbred population,  
161 we performed down-sampling of UKBB participants. Consistent with previous literature, ad-  
162 ditional CH and homozygous variants can be inferred by considering both pLoF and damaging  
163 missense/protein-altering variation at even larger sample sizes (Supplementary Fig. 19).

## 164 **Systematic evaluation of bi-allelic effects on common disease**

165 We performed a series of association analyses using Scalable and Accurate Implementation  
166 of GEneralized mixed model (SAIGE)<sup>30</sup>, a generalized mixed model that uses a saddle-point  
167 approximation to provide accurate *P*-values for traits with extreme case-control ratio imbalance.  
168 This allowed us to investigate the effects of bi-allelic variants in 176,587 individuals across

169 311 phenotypes with varying population prevalence identified from primary and secondary care  
170 electronic health records (EHRs) (Methods). We restricted to 952 protein-coding genes with  
171 at least 5 individuals carrying bi-allelic variants in the same gene, which allowed us to detect  
172 odds ratio (OR)  $\geq 10$ , for traits at approximately 2% population prevalence, with 80% power  
173 at exome-wide significance (Bonferroni  $P < \frac{0.05}{952 \times 311} \approx 1.68 \times 10^{-7}$ ) (Methods, Supplementary  
174 Fig. 20). Using simulations, we confirmed our ability to detect recessive signals of association  
175 with well-calibrated false positive rates across a range of effect sizes (Methods, Supplementary  
176 Fig. 21a-c). We tested a total of 299,854 gene-trait combinations, and identified 30 gene-  
177 trait associations at nominal significance ( $P < \frac{0.05}{952} \approx 5.25 \times 10^{-5}$ ), of which seven remained  
178 significant following stringent Bonferroni correction ( $P < 1.68 \times 10^{-7}$ ) (Supplementary Table  
179 14, Supplementary Fig. 10).

180 A recessive gene-trait association may be driven by a variety of genetic factors unrelated to CH  
181 or homozygous status, such as polygenic background or through genetic tagging of a nearby  
182 common variant association. To mitigate these factors, we created a pipeline to condition on  
183 external genetic effects within the gene-trait regression model (Methods). First, we trained  
184 polygenic risk scores (PRS) for 111 significantly heritable traits ( $h_{\text{snp}}^2 P < 0.05$  and  $n_{\text{eff}} \geq 5000$ )  
185 using LDpred2<sup>31</sup> (Methods, Supplementary Table 12), a tool that allows PRS derivation based  
186 on summary statistics and linkage information. To control for polygenic risk and potentially  
187 boost power for association<sup>32</sup>, we included the off-chromosome PRS as an additional covariate  
188 (Supplementary Table 14). While the resulting  $P$ -values were altered by less than a single order of  
189 magnitude with the incorporation of PRS (Supplementary Fig. 11), controlling for PRS resulted  
190 in the abrogation of four nominally significant ( $P < 5.25 \times 10^{-5}$ ) gene-trait associations. To  
191 capture the effects of any causal common variants in linkage disequilibrium (LD) with the pLoF  
192 or damaging missense/protein-altering variants constituting the CH or homozygous variant, we  
193 further conditioned on nearby (within 1 mega base pairs (Mb) of the associated gene) common

194 (MAF > 1%) variant association signals (Methods, Supplementary Table 13), which abrogated  
195 ( $P > 0.05$ ) the signal of two gene-trait pairs.

196 Lastly, we investigated whether any of the identified, putative recessive, associations could  
197 be accounted for by assuming an additive genetic architecture. To do this, we counted the  
198 number of gene copies affected by pLoF+damaging missense/protein altering variants in each  
199 individual. For each putative recessive association, we re-ran the analysis while simultaneously  
200 conditioning on the number of affected haplotypes. We also employed a complementary variant-  
201 level approach and repeated the analysis, conditioned on all low-frequency (MAC > 10, MAF <  
202 5%) and ultra-rare (MAC  $\leq$  10) damaging variants (pLoF+damaging missense/protein-altering),  
203 including those that constitute the bi-allelic variant in question. Conditioning on the additive  
204 effects abrogated the signal of a single nominally significant gene-trait pair ( $P < 5.25 \times 10^{-5}$ )  
205 (Supplementary Table 14).

206 Together, these analyses refined the list of putative gene-trait associations to 23 nominally  
207 significant associations out of which six are significant after correcting for multiple testing  
208 (conservative Bonferroni  $P < 1.68 \times 10^{-7}$ ) (Fig. 2a-2b, Supplementary Table 14) comprising  
209 17 unique genes and 22 traits. Notably, only six of the 23 associations remained nominally  
210 significant ( $P < 5.25 \times 10^{-5}$ ) when restricted to only CH variant-carriers, and just four of  
211 23 remained nominally significant when testing homozygous variants alone, underscoring the  
212 power of jointly analyzing these variant sets.

213 We observed recessive gene-trait relationships across multiple organ systems (nervous, respira-  
214 tory, circulatory, and genitourinary among others). All six associations that met the significance  
215 threshold after Bonferroni correction ( $P < 1.68 \times 10^{-7}$ ) have previously been reported in the  
216 literature. For example, individuals with bi-allelic variants in *MUTYH*, a gene involved in ox-  
217 idative DNA-damage repair<sup>33</sup>, are at significantly increased risk of developing colorectal cancer

218  $(\log_{10}(\text{OR}) = 4.7$  (95% CI = 3.38 – 6.01),  $P = 2.2 \times 10^{-12}$ ). We also find that bi-allelic variants  
219 in *FLG* increase risk of both asthma<sup>34</sup> ( $\log_{10}(\text{OR}) = 0.33$  (0.26 – 0.39),  $P = 2.09 \times 10^{-22}$ ) and  
220 dermatitis<sup>35</sup> ( $\log_{10}(\text{OR}) = 0.28$  (0.22 – 0.33),  $P = 2.65 \times 10^{-20}$ ). In addition, we observe that  
221 bi-allelic variants in *GJB2* increase the risk of hearing loss<sup>29</sup> ( $\log_{10}(\text{OR}) = 1.66$  (1.05 – 2.26),  
222  $P = 9.93 \times 10^{-8}$ ). At nominal significance ( $P < 5.25 \times 10^{-5}$ ), 10 of 23 associations have previ-  
223 ously been reported. For example, bi-allelic variants in *USH2A*, linked to retina homeostasis<sup>36</sup>,  
224 increase risk of visual impairment ( $\log_{10}(\text{OR}) = 5.77$  (2.93 – 8.62),  $P = 3.5 \times 10^{-5}$ ). For the re-  
225 maining unreported hits, we observe gene-trait associations with plausible mechanistic insights.  
226 For example, we observe that putatively damaging bi-allelic variation in *FAAH*, a fatty acid amide  
227 hydrolase<sup>37</sup>, are associated with increased risk of dementia ( $\log_{10}(\text{OR}) = 22.92$  (12.35 – 33.48),  
228  $P = 1.06 \times 10^{-5}$ ), consequently offering evidence supporting the hypothesis that lipid metabolism  
229 dysfunction is central to dementia pathogenesis<sup>38</sup>.

## 230 **Boosting power in gene-level regression models through rare variant hap-** 231 **lotype collapsing**

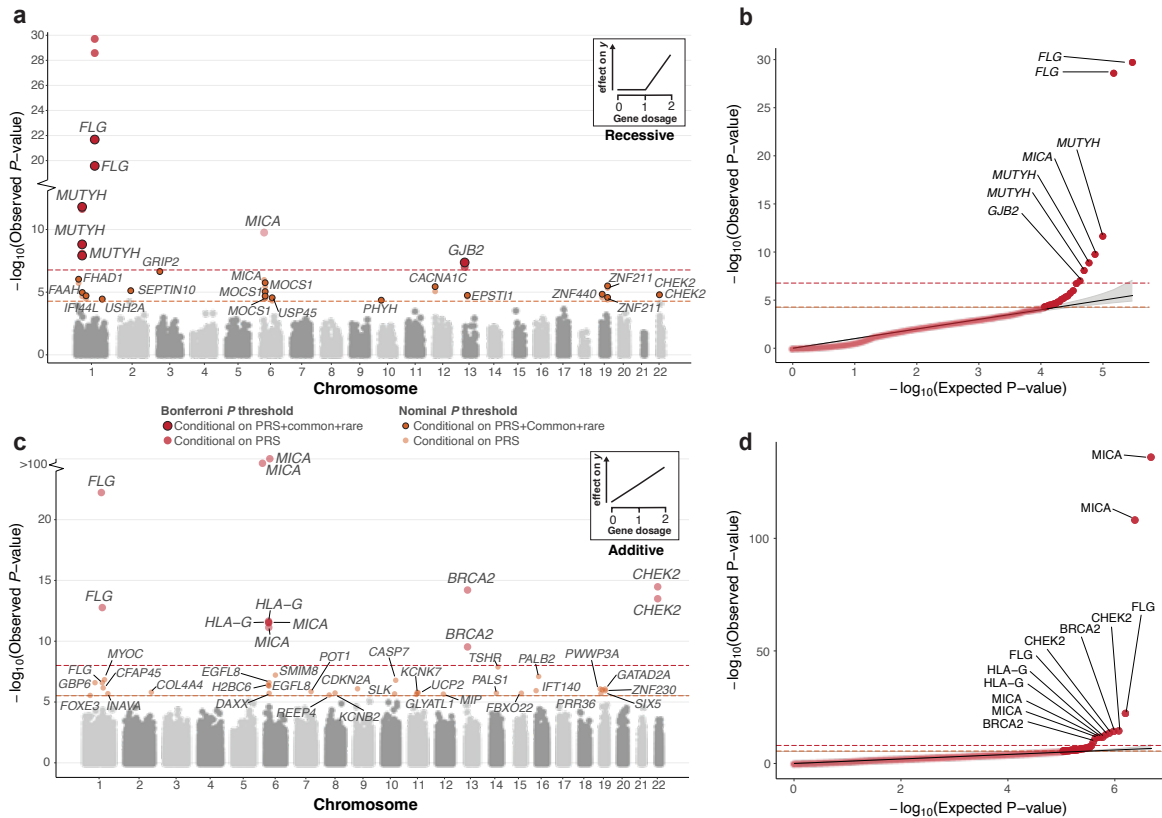
232 Complementary to the recessive models described above, rare variant burden testing, which  
233 involves the aggregation of rare variants within a gene, has proven to be a robust method to  
234 collectively assess the phenotypic impact of rare variation across individuals. Rare variants are  
235 aggregated due to their low allele frequency leading to lack of statistical power for detection  
236 of single-marker associations. However, these frameworks generally ignore the genetic phase  
237 within each individual, and therefore do not differentiate between scenarios in which multi-  
238 ple damaging variants reside on the same (*in cis*) or opposite (*in trans*) haplotypes, despite  
239 these two forms having potentially distinct functional and phenotypic effects. We conducted  
240 additive genome-wide association analyses by testing for associations between the the num-  
241 ber of disrupted gene copies (across 16,363 protein-coding genes with at least 10 haplotypes

242 carrying pLoF+damaging missense/protein altering variation) in an individual and case status  
243 (across 311 phenotypes) (Methods, Fig. 2c-2d). After adjusting for polygenic contribution,  
244 we found 38 nominally significant gene-trait associations (Nominal  $P < \frac{0.05}{16363} \approx 3.05 \times 10^{-6}$ ),  
245 among which 12 were significant associations after multiple-testing correction (Bonferroni  
246  $P < \frac{0.05}{16363 \times 311} \approx 9.8 \times 10^{-9}$ , Supplementary Table 15). Among the significant hits are previ-  
247 ously reported associations, including association between the number of putatively damaged  
248 copies of *BRCA2* ( $P = 6.16 \times 10^{-15}$ ) and *CHEK2* ( $P = 3.34 \times 10^{-15}$ ), and breast cancer.

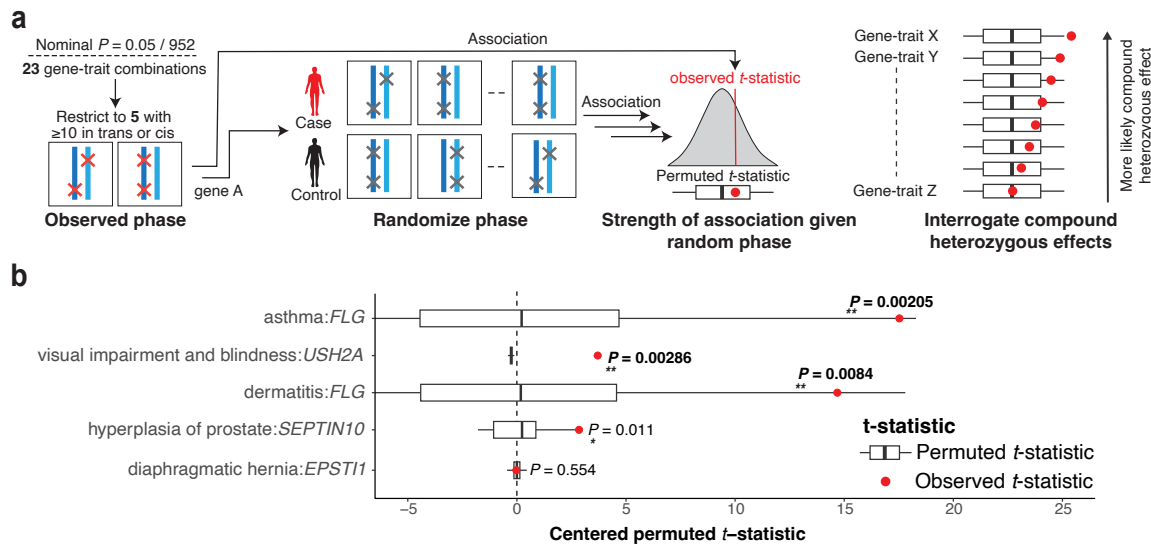
### 249 **Permutation testing to establish the impact of genetic phase on disease risk**

250 It is commonly accepted that compound heterozygosity drives recessive disease risk by disruption  
251 of both copies of an implicated gene<sup>13–15</sup>. However, this notion has not been well studied in a  
252 large-scale population cohort. To assess the degree to which compound heterozygosity, rather  
253 than co-occurring variants on the same haplotype, drives disease risk, we permuted the genetic  
254 phase of observed pLoF+damaging missense/protein-altering variants within a gene to generate  
255 an empirical distribution of *t*-statistics corresponding to disease-association strength in the  
256 absence of phase information (Fig. 3a-b). To ensure a sufficiently large sampling distribution,  
257 we restricted our analysis to 5 nominally significant ( $P < 5.25 \times 10^{-5}$ ) gene-trait combinations  
258 with at least ten individuals that are either CH variant-carriers or with two or more pLoF or  
259 damaging missense/protein-altering variants on the same haplotype (Methods).

260 We found evidence for significant (Bonferroni  $P = 0.05/5 = 0.01$ ) compound-heterozygous  
261 specific effects in three of the five analyzed gene-trait combinations: CH variants in *FLG* are  
262 associated with increased risk of both asthma ( $P = 0.00205$ ) and dermatitis ( $P = 0.0084$ ), while  
263 CH variants in *USH2A* are associated with increased risk of visual impairment and blindness ( $P =$   
264  $0.00286$ ) (Fig. 3c). We identified an additional gene-trait association at nominal significance  
265 ( $P < 0.05$ ), namely CH variants in *SEPTIN10* associated with hyperplasia of prostate ( $P =$



**Fig. 2: Conditional recessive and additive modeling of gene copy disruption in 311 phenotypes across 176,587 participants.** a) Recessive Manhattan plot depicting  $\log_{10}$ -transformed gene-trait association  $P$ -values against chromosomal location. Associations are colored red or orange based on whether they are Bonferroni ( $P < 1.68 \times 10^{-7}$ ) or nominally ( $P < 5.25 \times 10^{-5}$ ) significant. Transparent coloring represents the resulting  $P$ -value when conditioning only on PRS, whereas solid coloring with black outline represents the  $P$ -value derived after conditioning on off-chromosome PRS, nearby (500 kb) common variant association signal, and rare variants within the gene when applicable (methods). The Bonferroni and nominal significance thresholds are also displayed as orange and red dashed lines respectively. A gene may appear multiple times if it is associated with more than one phenotype. A qualifying example of the recessive inheritance pattern is shown in the top right: disruption of both gene copies result in an effect on the phenotype ( $y$ ). b) QQ-plot for genes with bi-allelic damaging variants after conditioning on off chromosome PRS. The shaded area depicts the 95%CI under the null. Gene-trait associations passing Bonferroni significance are labeled accordingly. c-d) Additive Manhattan plot and corresponding QQ-plot for genes with mono and bi-allelic damaging variants associated with at least one phenotype after conditioning on off chromosome PRS when applicable (methods). No additional conditioning was performed in this analysis. Gene-trait associations are colored red and orange based on whether they are respectively Bonferroni ( $P < 9.8 \times 10^{-9}$ ) or nominally ( $P < 3.05 \times 10^{-6}$ ) significant. The additive inheritance model is depicted in the top right: each affected haplotype result in a incremental effect on the phenotype ( $y$ ).



**Fig. 3: *In-silico* permutation of genetic phase provides evidence for CH-specific effects.** a) Overview of the permutation pipeline. To be sufficiently powered to detect effects, we considered five significant ( $P < 0.01$ ) gene-trait pairs from the genome-wide analysis that have at least ten individuals harboring pLoF or damaging missense/protein-altering variants on the same haplotypes or CH carriers. Then, we shuffled CH *trans* and *cis* labels across samples and re-ran the association analysis, resulting in a null distribution of permuted  $t$ -statistics corresponding to the association strength in the absence of phase information. We derive the one-tailed empirical  $P$ -value by comparing the observed  $t$ -statistics with the empirical null distribution. b) The resulting distributions of permuted (white and black box plots) and observed  $t$ -statistic (red dot) for each gene-trait and the resulting empirical  $P$ -value.  $P$ -values shown in bold indicate Bonferroni significance ( $P < 0.05/05 = 0.01$ ). Box and whisker plots display the quartiles of the empirical null distribution.

266 0.011). Of these, *FLG*-asthma, *FLG*-dermatitis, and *USH2A*-visual impairment associations  
 267 have previously been linked to disease in the CH state<sup>39–41</sup>. These observations demonstrate,  
 268 on a large scale, the effect of compound heterozygosity in driving disease susceptibility, and by  
 269 extension, how appropriately integrating genetic phase can lead to increased power to discover  
 270 gene-trait associations.

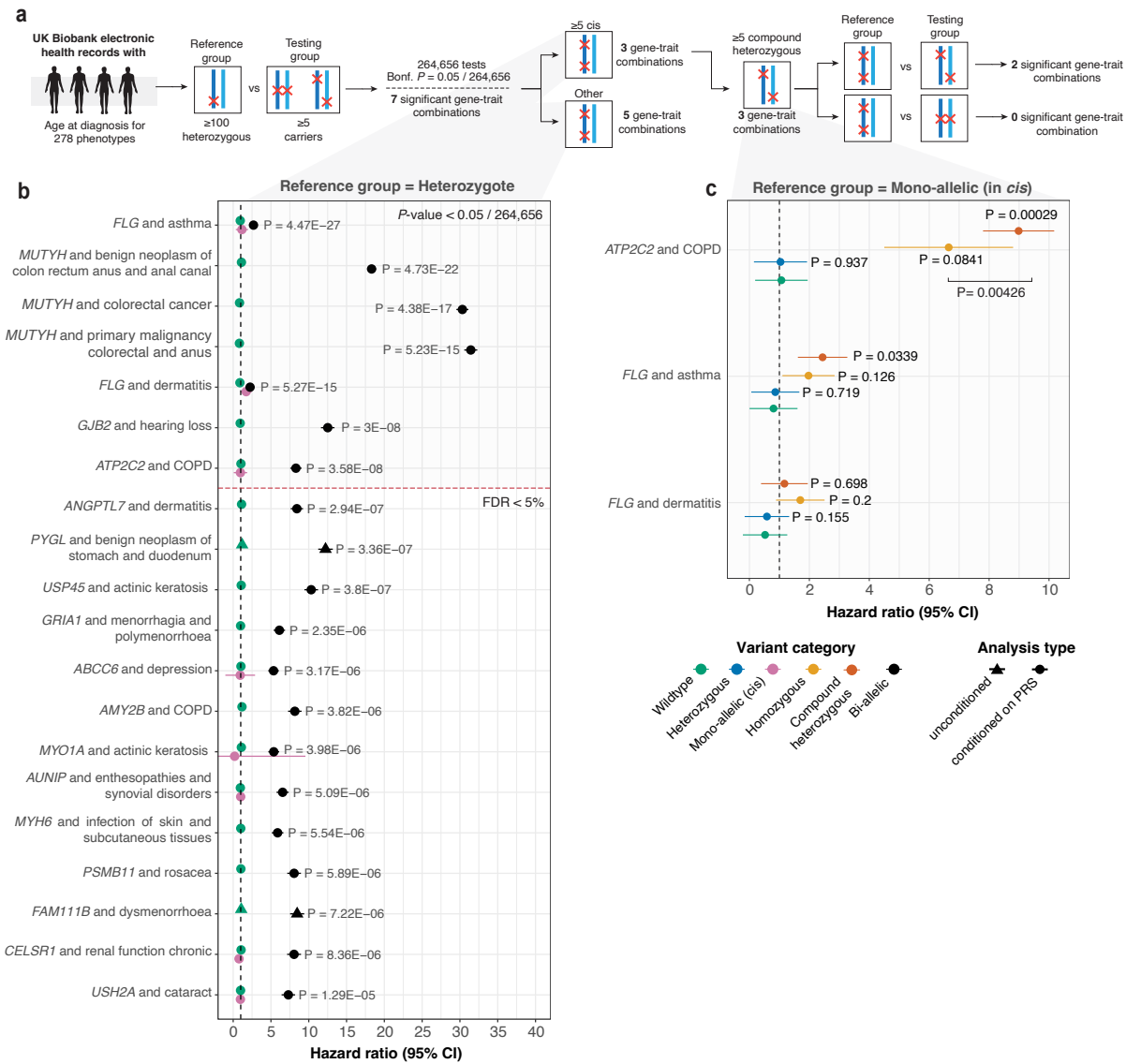
## 271 **Non-additive effects of compound heterozygous variants elevate lifetime risk** 272 **of disease**

273 Bi-allelic effects may be associated with earlier age at onset of disease, which is also often cor-  
274 related with disease severity. We therefore explored whether CH and homozygous variants had  
275 longitudinal effects by evaluating age-at-diagnosis of 278 phenotypes with Cox proportional-  
276 hazards models. To identify effects owed to disruption of both gene copies, as opposed to  
277 haploinsufficiency, we compared bi-allelic variant carriers against a reference group comprising  
278 carriers of a single heterozygous variant for each gene. We tested 267,400 gene-trait combina-  
279 tions with at least five bi-allelic variants (homozygotes or CH) and 100 heterozygotes (Fig. 4a).  
280 After adjustment for polygenic risk via off-chromosome PRS, we identified seven gene-trait asso-  
281 ciations with significantly earlier age-at-diagnosis in bi-allelic variants compared to heterozygous  
282 carriers of pLoF+damaging missense/protein-altering variants ( $P < \frac{0.05}{952 \times 278} \approx 1.89 \times 10^{-7}$ , Fig.  
283 4b-c, Supplementary Tables 16-17). Beyond these seven associations, we also identified 13  
284 additional gene-trait relationships at a false discovery rate (FDR) < 5% (4b, Supplementary  
285 Fig. 22). For six out of the seven Bonferroni significant gene-trait combinations, we found no  
286 evidence ( $P > 0.05/7 \approx 0.00833$ ) that carrying a single heterozygous variant altered lifetime  
287 disease risk compared to carrying two copies of the reference allele.

288 We further sought to disentangle the effects of homozygous and CH variants on lifetime disease  
289 risk from that attributable to multiple damaging rare variant effects on a single haplotype.  
290 To do this, we analyzed such effects in the three gene-trait pairs with both (1) at least five  
291 CH and/or homozygous variants and (2) at least five individuals harboring  $\geq 2$  variants on the  
292 same haplotype (Fig. 4d, Supplementary Table 18). Compared to individuals with a single  
293 disrupted haplotype, both homozygous and CH carriers of pLoF+damaging missense/protein-  
294 altering variants in *ATP2C2* were at increased lifetime risk of developing chronic obstructive  
295 pulmonary disease (COPD) (homozygote HR = 6.65 (95% CI = 4.5–8.8);  $P = 0.084$ , CH



296 HR = 8.98 (7.79–10.17);  $P = 0.00028$ ). Similarly, both homozygous and CH variants of  
297 *FLG* were at increased lifetime risk of asthma (homozygote HR = 1.97 (1.1–2.84);  $P = 0.126$ ,  
298 CH HR = 2.44 (1.61 – 3.26);  $P = 0.033$ ) and dermatitis (homozygote HR = 1.7 (0.88–2.5);  
299  $P = 0.20$ , CH HR = 1.16 (0.38–1.94);  $P = 0.7$ ) (Fig. 4c). For these gene-trait relationships,  
300 information encoded in genetic phase influences the risk of disease, with mono-allelic disruption  
301 leading to virtually unaltered risk while bi-allelic disruption may result in dramatic increase in  
302 lifetime risk of disease.



**Fig. 4 (previous page): Age-at-diagnosis modeling reveals novel recessive effects driven by damaging bi-allelic variants.** a) Flow diagram of our approach. To investigate whether homozygous and/or CH effects are associated with a difference in lifetime risk of disease development, we perform Cox proportional-hazards modeling for gene-trait combinations in which  $\geq 5$  samples are two-hit carriers (CH or homozygotes) and  $\geq 100$  samples that are heterozygotes. Among Bonferroni significant associations ( $P < 1.89 \times 10^{-7}$ ), we filter to gene-trait pairs for which at least five samples carry multiple variants disrupting the same haplotype, and test for an association between CH or homozygous carrier status and lifetime disease risk (corresponding to HRs $>1$ ). b) HRs when comparing CH and homozygous status versus heterozygous carrier status. Throughout, we display hazard ratios and corresponding  $P$ -values with (circles) and without (triangles) taking the polygenic contribution into account by conditioning on off-chromosome PRSs for heritable traits that pass our quality control cutoffs.  $P$ -values following inclusion of polygenic contribution to disease status are provided where PRS are predictive. HRs for gene-traits with two or more individuals with multiple *cis* variants on the same haplotype are displayed in pink. Associations that pass Bonferroni significance ( $P < 1.89 \times 10^{-7}$ ) and FDRs  $< 5\%$  cutoff are illustrated in the top and bottom respectively. c) HRs when comparing bi-allelic status versus heterozygous carrier status for gene-trait pairs with  $\geq 3$  individuals harboring variants disrupting the same haplotype, allowing ascertainment of confidence intervals. c) HRs when comparing wildtype, heterozygous, CH and homozygous status against individuals that harbor two damaging variants on the same haplotype. 95% CIs are shown in the figure. Abbreviations: CC (colorectal cancer), COPD (chronic obstructive pulmonary disease).

### 303 **Biological insights into common complex disorders implicated by CH vari-** 304 **ation**

305 Six of the seven gene-trait combinations for which we identify Bonferroni significant associations  
306 with lifetime disease risk are also significant in our cross-sectional recessive association analysis  
307 (Supplementary table 19). All six have previously been reported in the literature, albeit without  
308 age-at-onset effects. These include *MUTYH* and colorectal cancer, *GJB2* and hearing loss, and  
309 a pleiotropic association of *FLG* with both dermatitis and asthma. *ATP2C2*-COPD is a novel  
310 candidate association with plausible mechanistic effects.

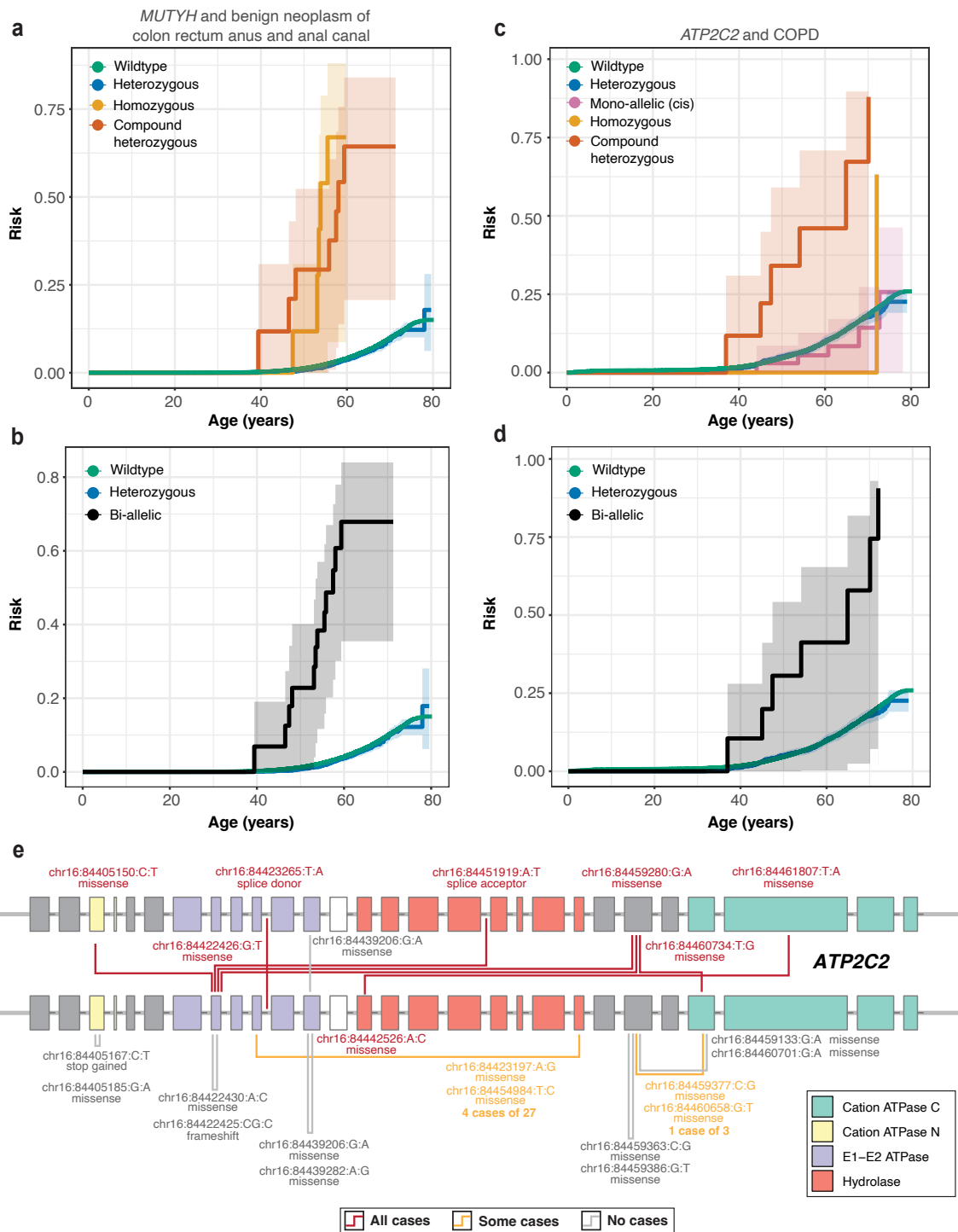
311 *MUTYH*-associated polyposis is considered a highly penetrant Mendelian cancer syndrome  
312 leading to adenomatous polyposis<sup>42</sup>. We link bi-allelic variants of *MUTYH* to elevated risk

313 of benign neoplasms of the colon, with bi-allelic carriers having a median age of diagnosis at  
314 age 53.7 years (interquartile range (IQR) = 47.9 - 56.3 years), as compared to heterozygotes  
315 (median age-at-diagnosis = 61.7 (56.2 - 66.7) years) and wildtypes (61.1 (54.6 - 66.5) years); as  
316 well as malignant neoplasms of the colon (median age at diagnosis for bi-allelic carriers = 52.1  
317 (IQR = 48.6 - 53.4) years, heterozygotes = 63.2 (57.7 - 67.0) years, and wildtypes = 62.9 (57.2  
318 - 67.9) years) (Fig. 5a-b). Because benign growths can be precursors to malignant neoplasms,  
319 and since risk of both disorders was elevated in *MUTYH* bi-allelic carriers (benign HR = 18  
320 (95% CI = 17.72–18.9);  $P = 4.7 \times 10^{-22}$ , malignant HR = 31.4 (95% CI = 30.57–32.3);  
321  $P = 5.2 \times 10^{-15}$ ), we examined the co-occurrence of variants across colorectal cancer outcomes.  
322 The same set of CH and homozygous variants are involved in the pathophysiology of benign  
323 and malignant neoplasms of the colon, suggesting that *MUTYH*-variant composition alone in  
324 insufficient to explain the dichotomy between malignant and benign polyposis (Supplementary  
325 Fig. 17).

326 *ATP2C2*, a calcium-transporting ATPase linked to surfactant protein D levels<sup>43</sup> (a causal risk  
327 factor for COPD), is associated with COPD in our gene-trait analyses (HR = 8.3 (95% CI =  
328 7.54–9.05);  $P = 3.56 \times 10^{-8}$ ). As we did not observe any nearby (1 Mb upstream or downstream)  
329 common variants in *ATP2C2* associated with cross-sectional COPD (all  $P > 5 \times 10^{-6}$ ), the  
330 association between bi-allelic variants of *ATP2C2* and COPD is potentially driven by the unique  
331 configurations of damaging-missense ( $n = 7$ ) and pLoF ( $n = 1$ ) variants that primarily reside in  
332 functional protein domains (Fig. 5e, Supplementary Fig. 18, Supplementary Table 20). 7 of 8  
333 (87.5%) identified bi-allelic carriers of *ATP2C2* (6 CH and 2 homozygous) were diagnosed with  
334 COPD (median age of diagnosis = 54.1 (IQR = 46.2 - 67.5) years) (Fig. 5c-d). In contrast, only  
335 6.9% of individuals harboring multiple pLoF+damaging missense/protein-altering variants on  
336 the same *ATP2C2* haplotype were diagnosed with COPD, and at the same median age (60.8  
337 (53.7 - 67.9) years) as heterozygotes (58.0 (48.5 - 64.1) years) and wildtypes (59.2 (51.3 - 65.1)

338 years).

339 *FLG* plays a pivotal role in the differentiation and maintenance of skin barriers<sup>34</sup>. *FLG* variants  
340 have been selectively associated with individuals with both asthma and atopic dermatitis, but not  
341 with those who have asthma without atopic dermatitis<sup>35</sup>. The exact nature of this relationship  
342 remains unclear. Our findings indicate that individuals carrying a single deleterious *FLG* allele  
343 face increased risk of dermatitis ( $P \approx 7.2 \times 10^{-5}$ ), but not asthma ( $P = 0.018$ ), when compared  
344 to wildtypes. In contrast, individuals carrying two variant alleles have an increased risk of  
345 developing both dermatitis ( $P = 5.27 \times 10^{-15}$ ) and asthma ( $P = 4.47 \times 10^{-27}$ ), suggesting a  
346 recessive mode of inheritance for *FLG*-related asthma and a semi-dominant inheritance pattern  
347 for *FLG*-related dermatitis<sup>29</sup>. This implies that the loss of a single *FLG* copy can result in  
348 dermatitis, while the loss of both copies can lead to asthma. Together, this may help clarify why  
349 *FLG*-related asthma is seldom observed without the presence of *FLG*-related dermatitis.



**Fig. 5 (previous page): Trajectories of haplotype disruption in common disease.** a-b) Kaplan-Meier survival curves for CH (red), homozygous (orange), heterozygous carriers (blue), single disruption of haplotypes (pink) owed to pLoF or damaging missense/protein-altering mutations. Wildtypes and bi-allelic variants (CH or homozygous) are shown with green and black lines respectively. Both CH and homozygous *MUTYH*-variant carriers are at elevated lifetime risk of developing benign neoplasm of the colon compared to heterozygous carriers and wildtypes. c-d) Kaplan-Meier survival curves for *ATP2C2* mono and bi-allelic variant carriers. Carriers of CH variants develop COPD more early compared to heterozygotes carriers and wildtypes. Moreover, individuals who harbor a single putatively disrupted haplotype owed to  $\geq 2$  damaging variants develop COPD at the same frequency as heterozygotes and wildtypes. e) Gene plots for *ATP2C2*, displaying protein coding variants for samples that carry  $\geq 2$  pLoF or damaging missense/protein-altering variants stratified by exon or intron. CH variants, multiple variants in *cis*, and homozygous variants are highlighted by lines joining the positions of co-occurring variants in a sample. Lines are colored by number of cases for the shown variant configurations, with gray lines indicating no observed samples are cases; orange lines indicating some some samples are cases; red lines indicate that all observed samples are cases. Variants are labeled by position (GRCh38) and according to inferred consequence (missense, stop gain, splice acceptor/donor). Protein domains are highlighted accordingly<sup>44</sup>.

## 350 Discussion

351 In this large biobank-scale effort, we systematically interrogate the role of bi-allelic coding  
352 variants in genes conferring risk for common complex diseases. In the cross-sectional and  
353 longitudinal analysis we identify 20 nominally significant ( $P < 5.25 \times 10^{-5}$ ) and 23 significant  
354 (FDR < 5%) gene-trait associations, respectively. Together, we find 36 unique gene-trait  
355 associations, that both replicate established relationships and identify previously unreported  
356 gene-trait associations for a range of binary phenotypes across the common disease spectrum.

357 We show that the 90% of deleterious CH variants occur at MAF < 0.34%. Given that phasing  
358 quality is directly correlated with allele frequency, it is essential to filter to the set of variants  
359 phased at high confidence to eliminate false positive identifications. Here, we quantified the  
360 increase in phasing quality using Mendelian inheritance logic in parent-offspring relationships  
361 and compared pairs of statistically phased variants to read-backed phased variants using short

362 read sequences. While read-backed phasing is computationally expensive and restricted to  
363 variants in close proximity, we demonstrate that it can be employed to evaluate statistical  
364 phasing quality in cohorts that lack trio relationships, with error rates comparable to that of trio  
365 switch error rates.

366 CH disease associations have mainly been explored in rare disorders<sup>13–17</sup>, but are seldom  
367 investigated in the study of common disease. This is due to the low prevalence of variants in the  
368 CH state and the genetic architecture of common complex traits, which are typically influenced by  
369 environmental factors and numerous loci with low to modest contribution to risk. In this study, we  
370 address these challenges and offer multiple lines of evidence to demonstrate the role of CH effects  
371 in driving disease risk for common traits. We employed two complementary analyses to detect  
372 gene-trait associations: a genome-wide logistic association analysis and a time-to-event model.  
373 Through these methods, we identified associations in which variants in the homozygous or CH  
374 state resulted in increased disease risk compared to wildtypes and individuals carrying multiple  
375 pathogenic variants on the same haplotype. Our findings show that for certain gene-trait pairs,  
376 individuals with a single disrupted gene copy have a risk of developing disease that is virtually  
377 indistinguishable from that of wildtypes, suggesting non-additive gene dosage effects. Further,  
378 by permuting the genetic phase, we find evidence that incorporation of confidently phased CH  
379 variants can boost power to detect recessive associations in common disease. Collectively, our  
380 results emphasize the importance of considering each individual's specific genetic context when  
381 assessing their genetic risk in a clinical setting. Simply identifying the presence of multiple  
382 pathogenic variants in a gene, disregarding the phase, may not be sufficient to fully understand  
383 an individual's risk profile.

384 Many common complex traits have polygenic architectures, which should be accounted for when  
385 performing gene-trait association testing. The presence of bi-allelic variants in individuals with



386 such diseases might be coincidental and not causally related to the trait, which may instead  
387 be a result of a high polygenic risk. However, across the significant recessive genome-wide  
388 associations, we observed that inclusion of PRS as a covariate, affected the resulting association  
389 *P*-value by less than single order of magnitude for the binary traits we analyzed. While we  
390 were only able to account for the polygenic contribution to disease development for 111 diseases  
391 with significant common variant heritability in the UKBB, due to low case numbers, these  
392 observations suggest that incorporation of polygenic background has limited influence on the  
393 degree of association when evaluating ultra-rare variation across binary traits.

394 We found that the majority of bi-allelic gene-disease associations are driven by variant com-  
395 binations containing at least one missense variant, which would have been excluded under a  
396 stricter high-confidence pLoF criterion. Although our less stringent inclusion threshold enabled  
397 us to identify a greater number of bi-allelic variants, it is likely that some damaging missense  
398 or protein-altering variants would incorrectly be predicted as damaging, or may exhibit gain-  
399 of-function rather than loss-of-function effects, consequently reducing the signal-to-noise ratio  
400 in our analyses. Even ‘knockouts’ by *bona fide* pLoF variants may only result in partial gene  
401 inactivation, and not necessarily complete gene knockdown. Additionally, pLoF variants may  
402 be ‘rescued’ and not lead to complete or even partial loss-of-function. While we show that  
403 including damaging missense/protein-altering variants to define bi-allelic variants can improve  
404 power for certain phenotypic associations, further manual curation and experimental validation  
405 will be required to demonstrate that these variants truly result in loss-of-function.

406 The likelihood of damaging alleles occurring on the same haplotype is influenced by a com-  
407 plex interplay of factors, including population structure and balance between selection, drift,  
408 mutation, and recombination. We and others<sup>45</sup> find that damaging CH variants occur less fre-  
409 quently than multiple damaging variants affecting the same haplotype, suggesting that in certain

410 circumstances, natural selection operates on a haplotype level. Once a LoF variant occurs and  
411 expands in the population, the affected haplotype has no selection against additional acquisition  
412 of damaging mutations. This has implications for association studies investigating CH effects  
413 by counting the number of damaging variants in a gene while attributing equal probability to  
414 each of affecting each haplotype<sup>8</sup>, as such frameworks may overestimate the frequency of CH  
415 events.

416 ‘Human knockouts’ have been extensively discussed in the context of therapeutic development.  
417 Examining both bi-allelic and mono-allelic carriers can help assess the safety of therapeutic in-  
418 terventions by analyzing how varying degrees of target modulation affect biological response<sup>3,6</sup>.  
419 We showcase several gene-trait relationships where the number of affected haplotypes influ-  
420 ences the lifetime risk of disease, potentially representing the manifestation of ‘adverse events’  
421 which are important endpoint in clinical trials. The absence of adverse events in mono-allelic  
422 carriers can potentially imply that partial pharmacological inhibition of a target may be a safe  
423 and effective approach. However, adverse effects observed in bi-allelic carriers of damaging  
424 variation within the same locus could indicate potential risks associated with complete target  
425 inhibition. A natural extension of this work could involve investigating mono and bi-allelic  
426 effects on quantitative outcomes, such as serum proteins. Changes in biomarkers (or other  
427 continuous outcomes) may reflect direct or indirect consequences of gene modulation and could  
428 serve as potential pharmacodynamic biomarkers commonly used to assess target engagement in  
429 clinical trials.

430 This work showcases the value of statistical phasing of damaging rare variants, and that asso-  
431 ciation analyses that account for compound heterozygosity can be better-powered for gene-trait  
432 discovery. We show that this approach can be employed to discover well-established and novel  
433 non-additive and additive gene-trait relationships across a wide range of disease etiologies. From

434 a clinical perspective, we demonstrate the importance of interrogating the genetic phase when  
435 dealing with CH variants in traits with recessive mode of inheritance. This is an important step  
436 towards uncovering the phenome-wide consequences of bi-allelic disruption across the human  
437 genome.

### 438 **Code availability**

439 The code required to reproduce our analyses are publicly available at [https://github.com/](https://github.com/frhl/wes_ko_ukbb)  
440 `frhl/wes_ko_ukbb`. Data produced in the present study are available upon reasonable request  
441 to the authors.

### 442 **Acknowledgments**

443 F.H.L. is supported by the Wellcome Trust (award 224894/Z/21/Z), and the Medical Sciences  
444 Doctoral Training Centre at the University of Oxford. S.S.V. is supported by the Rhodes Schol-  
445 arship, Clarendon Fund, and the Medical Sciences Doctoral Training Centre at the University of  
446 Oxford. N.B. is supported by the Clarendon Fund, and the Medical Sciences Doctoral Training  
447 Centre at the University of Oxford. W.Z. is supported by the National Human Genome Re-  
448 search Institute of the National Institutes of Health under award number K99HG012222. A.B.  
449 is supported by the Novo Nordisk Center for Genomic Mechanisms of Disease at the Broad  
450 Institute (NNF21SA0072102). N.W. is supported by a Sir Henry Dale Fellowship jointly funded  
451 by the Wellcome Trust and the Royal Society (220134/Z/20/Z) and research grant funding from  
452 the Rosetrees Trust (PGL19-2/10025). C.M.L. is supported by the Li Ka Shing Foundation,  
453 NIHR Oxford Biomedical Research Centre, Oxford, NIH (1P50HD104224-01), Gates Founda-  
454 tion (INV-024200), and a Wellcome Trust Investigator Award (221782/Z/20/Z). This research  
455 has been conducted using the UK Biobank resource under application Number 10844.

## 456 **Competing interests**

457 B.M.N. is a member of the scientific advisory board at Deep Genomics and Neumora. All other  
458 authors declare no competing interests.

## **References**

1. Nelson, M. R., Tipney, H., Painter, J. L., *et al.* The support of human genetic evidence for approved drug indications. en. *Nature Genetics* **47**, 856–860 (Aug. 2015).
2. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human genetics. en. *Nature Reviews Drug Discovery* **12**. Number: 8 Publisher: Nature Publishing Group, 581–594 (Aug. 2013).
3. Whiffin, N., Armean, I. M., Kleinman, A., *et al.* The effect of LRRK2 loss-of-function variants in humans. en. *Nature Medicine* **26**, 869–877 (June 2020).
4. Tobert, J. A. Lovastatin and beyond: the history of the HMG-CoA reductase inhibitors. en. *Nature Reviews Drug Discovery* **2**. Number: 7 Publisher: Nature Publishing Group, 517–526 (July 2003).
5. Do, R. Q., Vogel, R. A. & Schwartz, G. G. PCSK9 Inhibitors: potential in cardiovascular therapeutics. eng. *Current Cardiology Reports* **15**, 345 (Mar. 2013).
6. Minikel, E. V., Karczewski, K. J., Martin, H. C., *et al.* Evaluating drug targets through human loss-of-function genetic variation. en. *Nature* **581**, 459–464 (May 2020).
7. Van Hout, C. V., Tachmazidou, I., Backman, J. D., *et al.* Exome sequencing and characterization of 49,960 individuals in the UK Biobank. en. *Nature* **586**, 749–756 (Oct. 2020).
8. DeBoever, C., Tanigawa, Y., Lindholm, M. E., *et al.* Medical relevance of protein-truncating variants across 337,205 individuals in the UK Biobank study. eng. *Nature Communications* **9**, 1612 (Apr. 2018).
9. Sulem, P., Helgason, H., Oddson, A., *et al.* Identification of a large set of rare complete human knockouts. en. *Nature Genetics* **47**. Number: 5 Publisher: Nature Publishing Group, 448–452 (May 2015).
10. Heyne, H. O., Karjalainen, J., Karczewski, K. J., *et al.* Mono- and biallelic variant effects on disease at biobank scale. eng. *Nature* **613**, 519–525 (Jan. 2023).
11. Lim, E. T., Würtz, P., Havulinna, A. S., *et al.* Distribution and Medical Impact of Loss-of-Function Variants in the Finnish Founder Population. en. *PLoS Genetics* **10** (ed Cutler, D.) e1004494 (July 2014).
12. Saleheen, D., Natarajan, P., Armean, I. M., *et al.* Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. en. *Nature* **544**, 235–239 (Apr. 2017).
13. De Rosa, M., Fasano, C., Panariello, L., *et al.* Evidence for a recessive inheritance of Turcot’s syndrome caused by compound heterozygous mutations within the PMS2 gene. en. *Oncogene* **19**, 1719–1723 (Mar. 2000).

14. Hague, S., Rogaeva, E., Hernandez, D., *et al.* Early-onset Parkinson's disease caused by a compound heterozygous DJ-1 mutation. en. *Annals of Neurology* **54**, 271–274 (Aug. 2003).
15. Robinson, J. P., Johnson, V. L., Rogers, P. A., *et al.* Evidence for an Association between Compound Heterozygosity for Germ Line Mutations in the Hemochromatosis ( *HFE* ) Gene and Increased Risk of Colorectal Cancer. en. *Cancer Epidemiology, Biomarkers & Prevention* **14**, 1460–1463 (June 2005).
16. Maffei, L., Rochira, V., Zirilli, L., *et al.* A novel compound heterozygous mutation of the aromatase gene in an adult man: reinforced evidence on the relationship between congenital oestrogen deficiency, adiposity and the metabolic syndrome. en. *Clinical Endocrinology* **67**, 218–224 (Aug. 2007).
17. Wang, X.-H., Xie, L., Chen, S., *et al.* Identification of Novel Compound Heterozygous MYO15A Mutations in Two Chinese Families with Autosomal Recessive Nonsyndromic Hearing Loss. eng. *Neural Plasticity* **2021**, 9957712 (2021).
18. Delaneau, O., Zagury, J.-F., Robinson, M. R., *et al.* Accurate, scalable and integrative haplotype estimation. en. *Nature Communications* **10**. Number: 1 Publisher: Nature Publishing Group, 5436 (Nov. 2019).
19. Maestri, S., Maturo, M. G., Cosentino, E., *et al.* A Long-Read Sequencing Approach for Direct Haplotype Phasing in Clinical Settings. en. *International Journal of Molecular Sciences* **21**. Number: 23 Publisher: Multidisciplinary Digital Publishing Institute, 9177 (Jan. 2020).
20. Li, N. & Stephens, M. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics* **165**, 2213–2233 (Dec. 2003).
21. Loh, P.-R., Danecek, P., Palamara, P. F., *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. en. *Nature Genetics* **48**. Number: 11 Publisher: Nature Publishing Group, 1443–1448 (Nov. 2016).
22. Barton, A. R., Sherman, M. A., Mukamel, R. E., *et al.* Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. en. *Nature Genetics* **53**. Number: 8 Publisher: Nature Publishing Group, 1260–1269 (Aug. 2021).
23. Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. en. *Nature Reviews Genetics* **12**. Number: 10 Publisher: Nature Publishing Group, 703–714 (Oct. 2011).
24. Hofmeister, R. J., Ribeiro, D. M., Rubinacci, S., *et al.* *Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank* en. Oct. 2022.
25. Delaneau, O., Zagury, J.-F., Robinson, M. R., *et al.* Accurate, scalable and integrative haplotype estimation. en. *Nat. Commun.* **10**, 5436 (Nov. 2019).
26. Martin, M., Patterson, M., Garg, S., *et al.* *WhatsHap: fast and accurate read-based phasing* en. Nov. 2016.

27. Karczewski, K. J., Francioli, L. C., Tiao, G., *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. en. *Nature* **581**. Number: 7809 Publisher: Nature Publishing Group, 434–443 (May 2020).
28. Samocha, K. E., Kosmicki, J. A., Karczewski, K. J., *et al.* Regional missense constraint improves variant deleteriousness prediction. *BioRxiv* (2017).
29. Hamosh, A., Scott, A. F., Amberger, J. S., *et al.* Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research* **33**, D514–D517 (Jan. 2005).
30. Zhou, W., Nielsen, J. B., Fritsche, L. G., *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. en. *Nature Genetics* **50**, 1335–1341 (Sept. 2018).
31. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. en. *Bioinformatics* **36** (ed Schwartz, R.) 5424–5431 (Apr. 2021).
32. Jurgens, S. J., Pirruccello, J. P., Choi, S. H., *et al.* Adjusting for common variant polygenic scores improves yield in rare variant association analyses. en. *Nature Genetics*. Publisher: Nature Publishing Group, 1–5 (Mar. 2023).
33. Kavec, M. J., Urbanova, M., Makovicky, P., *et al.* Oxidative Damage in Sporadic Colorectal Cancer: Molecular Mapping of Base Excision Repair Glycosylases MUTYH and hOGG1 in Colorectal Cancer Patients. *International Journal of Molecular Sciences* **23**, 5704 (May 2022).
34. Rice, N. E., Patel, B. D., Lang, I. A., *et al.* Filaggrin gene mutations are associated with asthma and eczema in later life. en. *The Journal of allergy and clinical immunology* **122**. Publisher: NIH Public Access, 834 (Oct. 2008).
35. Palmer, C. N. A., Irvine, A. D., Terron-Kwiatkowski, A., *et al.* Common loss-of-function variants of the epidermal barrier protein filaggrin are a major predisposing factor for atopic dermatitis. en. *Nature Genetics* **38**. Number: 4 Publisher: Nature Publishing Group, 441–446 (Apr. 2006).
36. Sandberg, M. A., Rosner, B., Weigel-DiFranco, C., *et al.* Disease course in patients with autosomal recessive retinitis pigmentosa due to the USH2A gene. eng. *Investigative Ophthalmology & Visual Science* **49**, 5532–5539 (Dec. 2008).
37. Bajaj, S., Zameer, S., Jain, S., *et al.* Effect of the MAGL/FAAH Dual Inhibitor JZL-195 on Streptozotocin-Induced Alzheimer’s Disease-like Sporadic Dementia in Mice with an Emphasis on A, HSP-70, Neuroinflammation, and Oxidative Stress. eng. *ACS chemical neuroscience* **13**, 920–932 (Apr. 2022).
38. Chew, H., Solomon, V. A. & Fonteh, A. N. Involvement of Lipids in Alzheimer’s Disease Pathology and Potential Therapies. *Frontiers in Physiology* **11** (2020).
39. Carlsen, B. C., Meldgaard, M., Johansen, J. D., *et al.* Filaggrin compound heterozygous patients carry mutations in trans position. eng. *Experimental Dermatology* **22**, 572–575 (Sept. 2013).

40. Riethmuller, C., McAleer, M. A., Koppes, S. A., *et al.* Filaggrin breakdown products determine corneocyte conformation in patients with atopic dermatitis. en. *Journal of Allergy and Clinical Immunology* **136**, 1573–1580.e2 (Dec. 2015).
41. Liu, X., Tang, Z., Li, C., *et al.* Novel USH2A compound heterozygous mutations cause RP/USH2 in a Chinese family. eng. *Molecular Vision* **16**, 454–461 (Mar. 2010).
42. Weren, R. D., Ligtenberg, M. J., Geurts van Kessel, A., *et al.* NTHL1 and MUTYH polyposis syndromes: two sides of the same coin? en. *The Journal of Pathology* **244**. [eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/path.5002](https://onlinelibrary.wiley.com/doi/pdf/10.1002/path.5002), 135–142 (2018).
43. Obeidat, M., Li, X., Burgess, S., *et al.* Surfactant protein D is a causal risk factor for COPD: results of Mendelian randomisation. en. *European Respiratory Journal* **50**. Publisher: European Respiratory Society Section: Original articles (Nov. 2017).
44. Finn, R. D., Bateman, A., Clements, J., *et al.* Pfam: the protein families database. eng. *Nucleic Acids Research* **42**, D222–230 (Jan. 2014).
45. Guo, M. H., Francioli, L. C., Stenton, S. L., *et al.* *Inferring compound heterozygosity from large-scale exome sequencing data* en. Pages: 2023.03.19.533370 Section: New Results. Mar. 2023.
46. Hail Team. *Hail* Oct. 2022.
47. Chang, C. C., Chow, C. C., Tellier, L. C., *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. en. *Gigascience* **4**, 7 (Feb. 2015).
48. Bycroft, C., Freeman, C., Petkova, D., *et al.* The UK Biobank resource with deep phenotyping and genomic data. en. *Nature* **562**, 203–209 (Oct. 2018).
49. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. en. *Bioinformatics* **27**, 2987–2993 (Nov. 2011).
50. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. en. *Bioinformatics* **30**, 2843–2851 (Oct. 2014).
51. Van der Auwera, G. A. & O’Connor, B. D. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra* en (“O’Reilly Media, Inc.”, Apr. 2020).
52. Zhang, D., Dey, R. & Lee, S. Fast and robust ancestry prediction using principal component analysis. en. *Bioinformatics* **36**, 3439–3446 (June 2020).
53. Danecek, P., Bonfield, J. K., Liddle, J., *et al.* Twelve years of SAMtools and BCFtools. eng. *GigaScience* **10**, giab008 (Feb. 2021).
54. Harrell Jr, F. E. & Harrell Jr, M. F. E. Package ‘hmisc’. *CRAN2018* **2019**, 235–236 (2019).
55. Kuan, V., Denaxas, S., Gonzalez-Izquierdo, A., *et al.* A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. English. *The Lancet Digital Health* **1**. Publisher: Elsevier, e63–e77 (June 2019).
56. Censin, J. C., Peters, S. A. E., Bovijn, J., *et al.* Causal relationships between obesity and the leading causes of death in women and men. en. *PLoS Genet.* **15**, e1008405 (Oct. 2019).
57. Eastwood, S. V., Mathur, R., Atkinson, M., *et al.* Algorithms for the Capture and Adjudication of Prevalent and Incident Diabetes in UK Biobank. en. *PLoS One* **11**, e0162388 (Sept. 2016).

58. Udler, M. S., McCarthy, M. I., Florez, J. C., *et al.* Genetic Risk Scores for Diabetes Diagnosis and Precision Medicine. en. *Endocr. Rev.* **40**, 1500–1520 (Dec. 2019).
59. McLaren, W., Gil, L., Hunt, S. E., *et al.* The Ensembl Variant Effect Predictor. en. *Genome Biol.* **17**, 122 (June 2016).
60. Karczewski, K. J., Francioli, L. C., Tiao, G., *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. en. *Nature* **581**, 434–443 (May 2020).
61. Ioannidis, N. M., Rothstein, J. H., Pejaver, V., *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. en. *Am. J. Hum. Genet.* **99**, 877–885 (Oct. 2016).
62. Rentzsch, P., Witten, D., Cooper, G. M., *et al.* CADD: predicting the deleteriousness of variants throughout the human genome. en. *Nucleic Acids Res.* **47**, D886–D894 (Jan. 2019).
63. Zhou, W., Nielsen, J. B., Fritsche, L. G., *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. en. *Nat. Genet.* **50**, 1335–1341 (Sept. 2018).
64. Privé, F., Aschard, H., Carmi, S., *et al.* High-resolution portability of 245 polygenic scores when derived and applied in the same cohort. *medRxiv* (2021).
65. International HapMap 3 Consortium, Altshuler, D. M., Gibbs, R. A., *et al.* Integrating common and rare genetic variation in diverse human populations. en. *Nature* **467**, 52–58 (Sept. 2010).
66. Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. en. *Nat. Genet.* **47**, 291–295 (Mar. 2015).
67. Finucane, H. K., Bulik-Sullivan, B., Gusev, A., *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. en. *Nat. Genet.* **47**, 1228–1235 (Nov. 2015).
68. Privé, F., Aschard, H., Ziyatdinov, A., *et al.* Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. en. *Bioinformatics* **34**, 2781–2787 (Aug. 2018).
69. Georgi, B., Voight, B. F. & Bućan, M. From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. eng. *PLoS genetics* **9**, e1003484 (May 2013).
70. Vinceti, A., Karakoc, E., Pacini, C., *et al.* CoRe: a robustly benchmarked R package for identifying core-fitness genes in genome-wide pooled CRISPR-Cas9 screens. en. *BMC Genomics* **22**, 828 (Nov. 2021).
71. Hart, T., Chandrashekar, M., Aregger, M., *et al.* High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. en. *Cell* **163**, 1515–1526 (Dec. 2015).
72. Hart, T., Tong, A. H. Y., Chan, K., *et al.* Evaluation and Design of Genome-Wide CRISPR/SpCas9 Knockout Screens. *G3 Genes|Genomes|Genetics* **7**, 2719–2727 (Aug. 2017).
73. Pirinen, M. *GWAS 3: Statistical power* Feb. 2023.



74. Kaplan, E. L. & Meier, P. Nonparametric Estimation from Incomplete Observations. *J. Am. Stat. Assoc.* **53**, 457 (June 1958).
75. Therneau, T. M. *A Package for Survival Analysis in R* R package version 3.2-3 (2020).
76. Kassambara, A., Kosinski, M. & Biecek, P. *survminer: Drawing Survival Curves using 'ggplot2'* R package version 0.4.9 (2021).

## Methods

### Exome sequencing quality control summary

We perform a series of hard-filters on genotype, sample, and variant metrics (Table 1, Table 2-3). We confirm genetic sex with reported sex, and restrict analysis to genetically ascertained samples of NFE ancestry, using random forest (RF) classifiers (Fig. 2-3). Finally, we filter based on a second collection of sample and variant filters (Tables 2-3). We used Hail 0.2<sup>46</sup> and PLINK 1.9<sup>47</sup> to perform all QC steps, and use R (4.0.2) scripts for plotting and filtering. Data was manipulated in R using `data.table` (1.14.2) and `dplyr` (1.0.7), random forest classifiers were trained using the `randomForest` (4.6-14) library, and plotting was performed using a `ggplot2` (3.3.5).

### Exome sequencing quality control, full details

**Sample filters** We evaluated sample-level quality control (QC) metrics on the 200,643 UKBB ES multi-sample project level variant call format (VCF) call-set files<sup>46</sup>, Supplementary Table 1. All metrics were calculated for bi-allelic single nucleotide polymorphisms (SNPs), except for metrics involving insertions and deletions. We regressed out the first 21 principal components (PCs)<sup>48</sup>, and filtered out sample outliers of the residuals for each metric based on MAD (median absolute deviation) thresholds (Supplementary Table 1). Samples without PC data were subject to more stringent thresholds (Supplementary Table 1).

<b>Metric</b>	<b>Metric residual (w/ PCs)</b>	<b>Raw (w/o PCs)</b>
<i>call_rate</i>	[24, $\infty$ )	[4, $\infty$ )
<i>n_insertion</i>	[8, 8]	[4, 4]
<i>n_deletion</i>	[8, 8]	[4, 4]
<i>r_insertion_deletion</i>	[8, 8]	[4, 4]
<i>n_het</i>	[12, 12]	[4, 4]
<i>n_hom_var</i>	[12, 12]	[4, 4]
<i>r_het_hom_var</i>	[16, 16]	[4, 4]
<i>n_non_ref</i>	[8, 8]	[4, 4]
<i>n_singleton</i>	$(-\infty, 16]$	$(-\infty, 4]$
<i>n_snp</i>	[8, $\infty$ )	[4, 4]
<i>n_transition</i>	[8, 8]	[4, 4]
<i>n_transversion</i>	[8, 8]	[4, 4]
<i>r_ti_tv</i>	[8, 8]	[4, 4]

**Supplementary Table 1: Sample filtering: MAD Intervals.** The interval  $[a, b]$  represents  $\text{median}(X) + \text{MAD}(X)[-a, b]$  for the metric,  $X$ . Samples with metrics outside these intervals were removed.

**Genotype filters** Multi-allelic variants were split into bi-allelic variants and insertions and deletions (indel) were left-aligned<sup>49</sup>. Genotype calls meeting any of the following criteria were set to missing:

1. Genotype quality (GQ)  $\leq 20$
2. Total sequencing depth (DP)  $\leq 10$
3. Heterozygous calls:
  - (a) SNPs: 1-sided binomial test of alternate allele depth related to total read depth  

$$P < 1 \times 10^{-3}$$
  - (b) Indels: alternate allele read depth / total read depth  $< 0.3$
4. Homozygous indel calls: alternate allele read depth / total read depth  $< 0.7$

**Variant-level filters** Retain variants satisfying all of the following conditions:

1. Not in a low complexity region (LCR)<sup>50</sup>.
2. In sequencing target regions  $\pm 50$  base pairs.
3. MAF  $> 0$  following genotype QC.
4. Excess heterozygosity (ExcessHet  $< 54.69$ ) filter: Phred-scaled  $P$ -value for exact test of excess heterozygosity<sup>51</sup> in founders as determined by relatedness estimates and recorded ages of UKBB participants<sup>48</sup>. Variants were retained as recommended in genome analysis toolkit (GATK)<sup>51</sup>

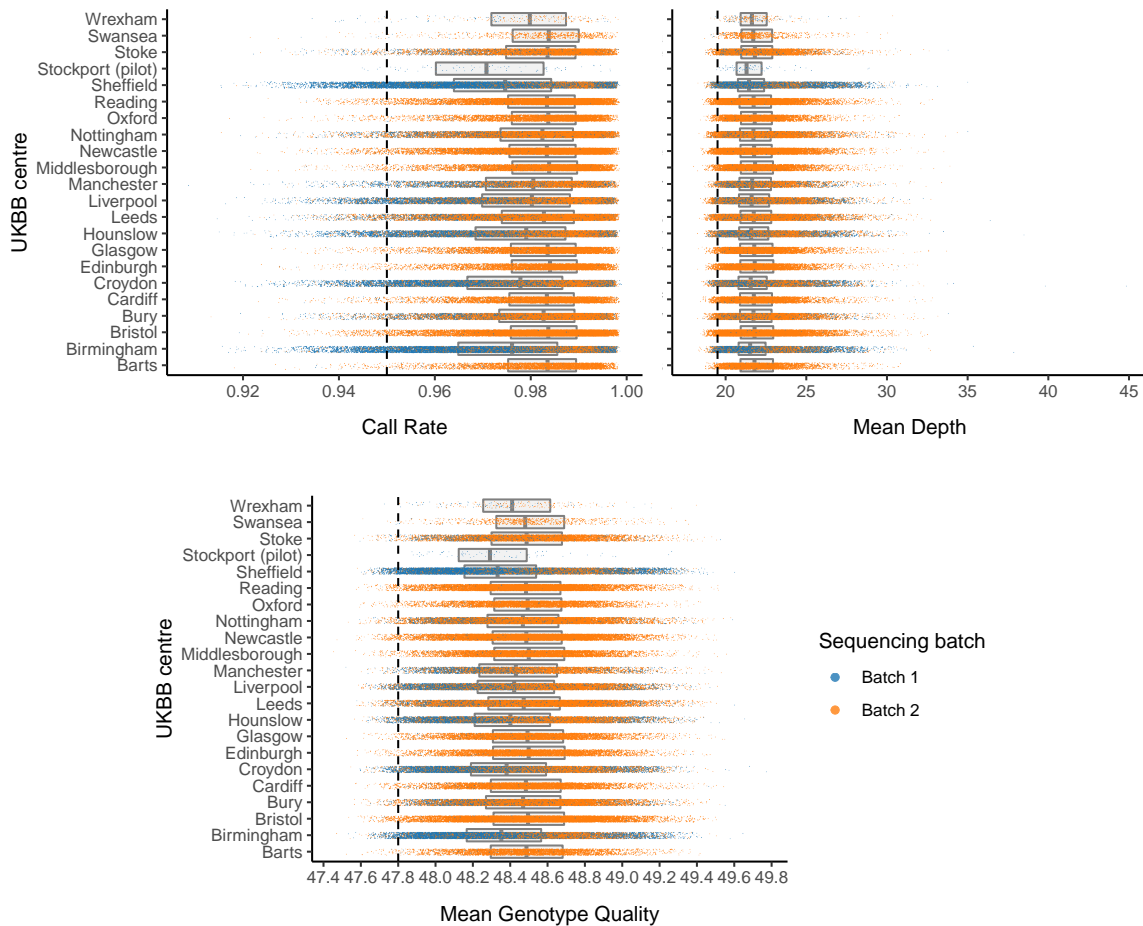
### **Additional ES quality control**

To perform further QC we use Hail, an open-source Python library which focuses on the analysis of large-scale genetic data sets. We used Hail to create our own methods, and we take advantage of the functionality that has been rewritten to enable fast and scalable analysis of large exome and genome sequencing projects. Unless otherwise stated, all of the following the data curation and quality control steps were performed in Hail<sup>46</sup>.

Briefly, we apply a collection of hard-filters on sample metrics. We confirm genotypic sex with reported sex, remove samples with excess glsplurv, and restrict analysis to samples of genetically ascertained NFE ancestry. Finally we apply a second collection of sample and variant hard filters. As an initial pass to remove low quality and contaminated samples, we filter out samples with call rate  $< 0.95$ , mean DP  $< 19.5\times$  or mean GQ  $< 47.8$  (Fig. 1).

### **Sex imputation**

To confirm participant sex and calculate PCs, we extracted high quality common variants (allele frequency between 0.01 to 0.99 with high call rate ( $> 0.98$ )) and LD prune to pseudo-independent SNPs using `--indep 50 5 2` in PLINK 1.9. When reported sex does not match genotypic sex,



**Supplementary Fig. 1: Distributions of sample metrics following initial restriction to variants, lying outside LCRs and inside the padded (50 bp) target intervals, and prior to the initial hard sample filters (call rate > 0.95, mean depth > 19.5, mean GQ > 47.8).** In each plot, jittered scatters display the distribution for each UKBB recruitment center, colored according to sequencing batch. Box-plots behind the scatter display the median and interquartile range for each sequencing batch. Hard-filtering thresholds are denoted by the dashed vertical line.

<b>Filter</b>	<b>Samples</b>	<b>Batch 1</b>	<b>Batch 2</b>	<b>%</b>
Initial samples in raw UKBB VCF	200,643	NA	NA	100.0
Initial samples in filtered VCF	199,795	49,759	150,036	99.6
Sample call rate <0.95	7,400	4,780	2,620	3.7
Mean DP <19.5	3,253	511	2,742	1.6
Mean GQ <47.8	2,123	834	1,289	1.1
Samples with sex swap	85	24	61	0.0
Samples with excess ultra-rare variants (URVs)	76	6	70	0.0
PCA based filters	13,537	3,390	10,147	6.7
Within batch Ti/Tv ratio outside 4 standard deviations (SDs)	13	3	10	0.0
Within batch Het/HomVar ratio outside 4 SDs	251	46	205	0.1
Within batch Insertion/Deletion ratio outside 4 SDs	9	4	5	0.0
<i>n</i> singletons >175	19	2	17	0.0
<b>Samples after all sample filters</b>	<b>176,935</b>	<b>41,371</b>	<b>135,564</b>	<b>88.2</b>

**Supplementary Table 2: Summary of sample filters.** Moving down through the rows of the table, we move through QC filtering steps.

<b>Filter</b>	<b>Variants</b>	<b>%</b>
Initial variants in raw UKBB VCF	15,922,704	100.0
Variants removed in initial filters	2,883,660	18.1
Invariant sites after sample filters	2,744,044	17.2
Overall variant call rate <0.97	1,122,987	7.1
Variants failing HWE filter	5,237	0.0
<b>Variants remaining after all filters</b>	<b>9,169,408</b>	<b>57.6</b>

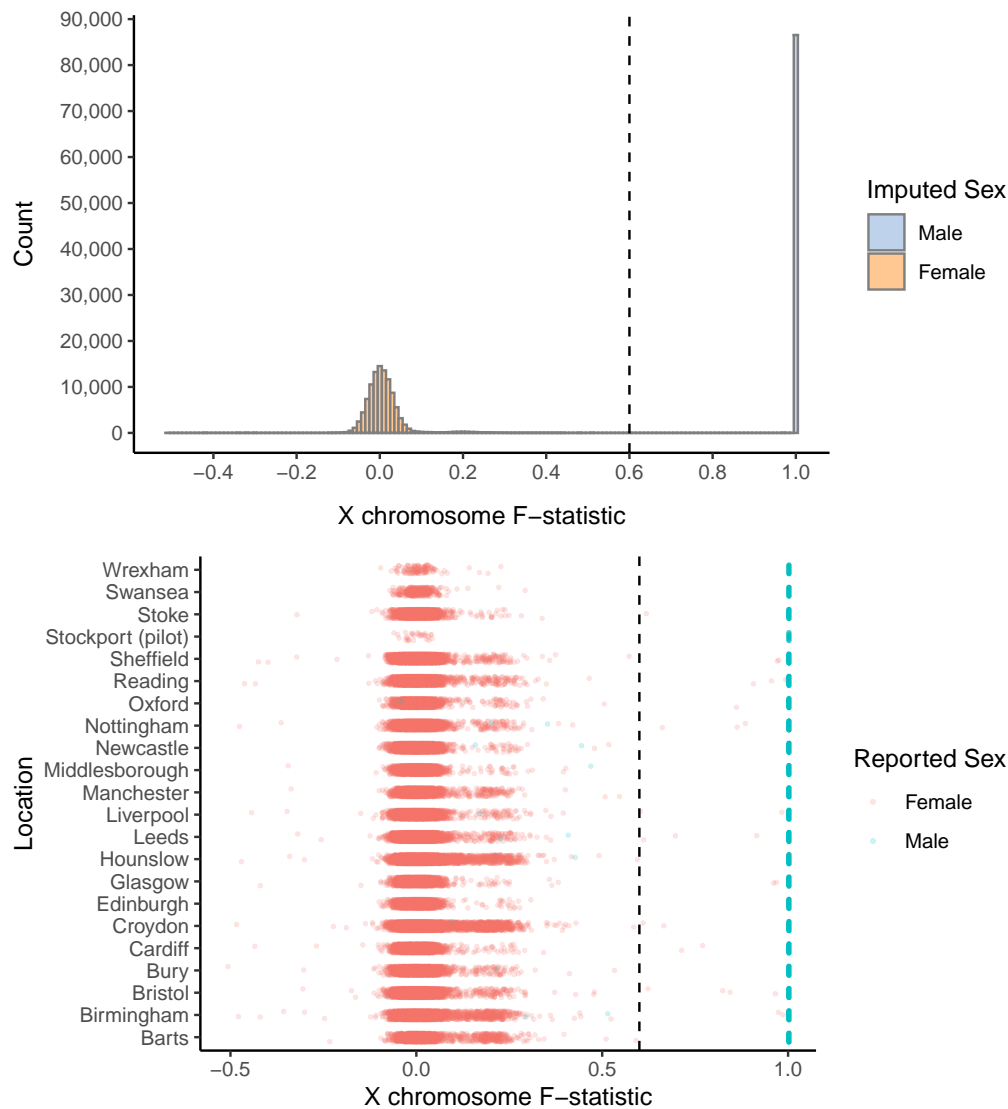
**Supplementary Table 3: Summary of variant filters.** Moving down through the rows of the table, we move through QC filtering steps.

it may signal potential sample swaps in the data. Using the  $F$ -statistic for each sample using the subset of the non-pseudo autosomal region on chromosome X, we identify and remove samples where reported sex information is not confirmed in the sequence data (Fig. 2). Specifically, we remove samples satisfying at least one of the following criteria:

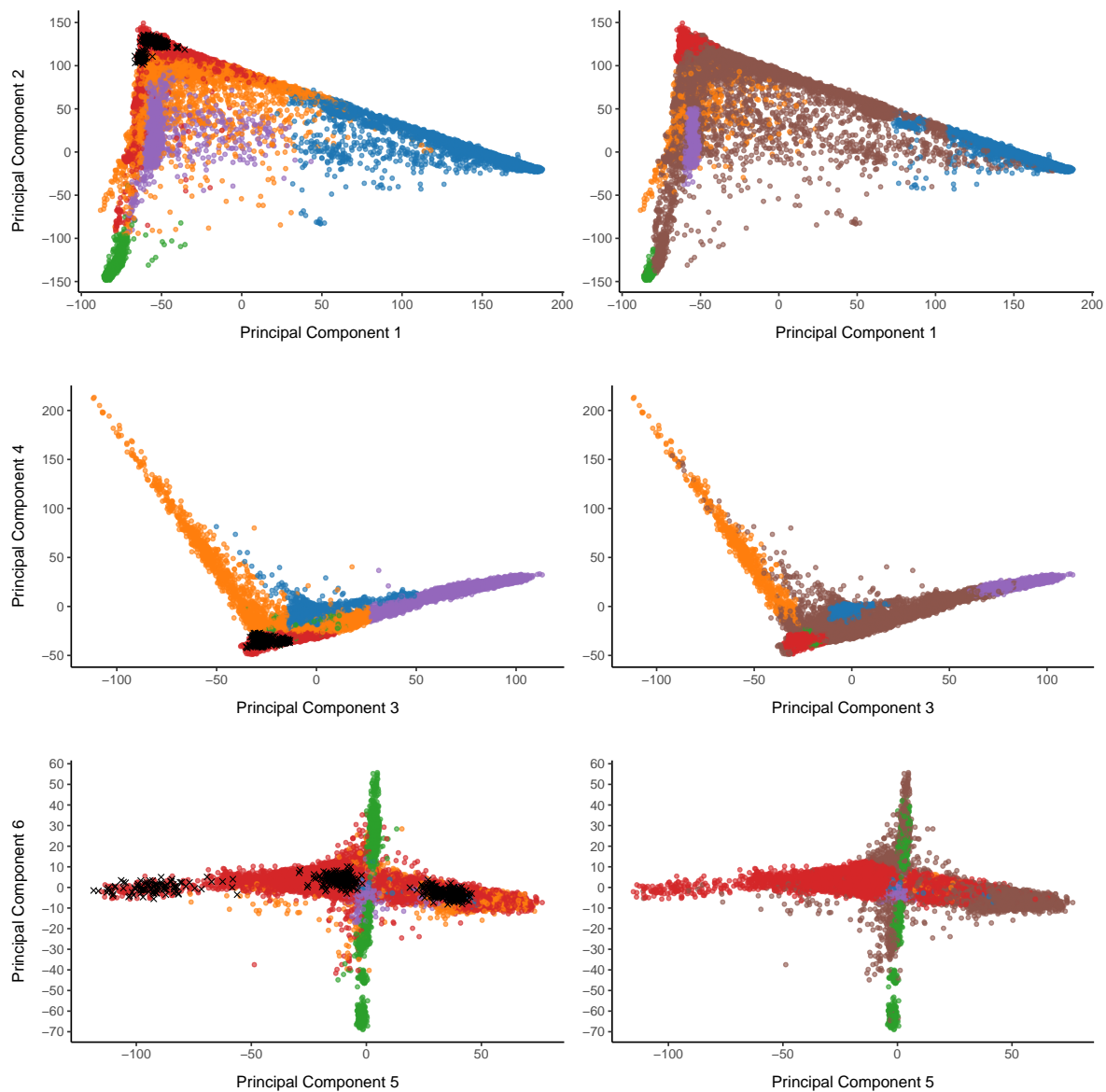
- Sex is unknown in the phenotype files.
- $F$ -statistic  $> 0.6$  and the sex is female in the phenotype file.
- $F$ -statistic  $< 0.6$  and the sex is male in the phenotype file.
- $F$ -statistic  $> 0.6$  and number of calls on the Y chromosome is  $< 100$ .

### **Defining a set of samples with non-Finnish European ancestry**

To ensure adequate case-control for as many traits as possible, we restricted our analysis to a set of genetically ascertained NFE samples. To do this, we perform a number of principal component analysis (PCA) steps to ensure that we have subset down to NFE. We first run PCA on the 1000 Genomes (1KGP) samples (minus the small subset of related individuals within the 1KGP) using subsetting to LD pruned autosomal variants. We then project in the UKBB samples, ensuring that we correctly account for shrinkage bias in the projection<sup>52</sup>. Next, we removed samples outside of the European population (EUR) using a RF classifier: we train a RF on the super-populations labels of 1KGP and predict the super-population for each of the UKBB samples (Fig. 3). We denote strictly defined European subset as those with probability  $> 0.99$  of being European according to the classifier. Another RF classifier is trained following restriction of the 1KGP samples to Europeans to determine NFE, using a classifier probability of 0.95. RF classifiers were trained using the randomForest (4.6) library in R. Samples not assigned to the NFE cluster were removed from downstream analysis.



**Supplementary Fig. 2: Histogram and scatter-plots of X chromosome  $F$ -statistic by collection.** Samples lying to the left and right of the dashed line were called as female and male respectively, according to the imputed sex colorings in the upper histogram. Reported sex, split by UKBB recruitment center are shown in the lower jittered scatter-plots: red if the sample is reported as female, and blue if the sample is reported as male.



**Supplementary Fig. 3: Scatter-plots of PCs of UKBB genotype data projected into the PC space defined by 1KGP samples.** Points are colored according to sample collection, with 1KGP samples colored in blue. 1KGP super-populations labels were used to train a random forest classifier.



## **Final hard filters**

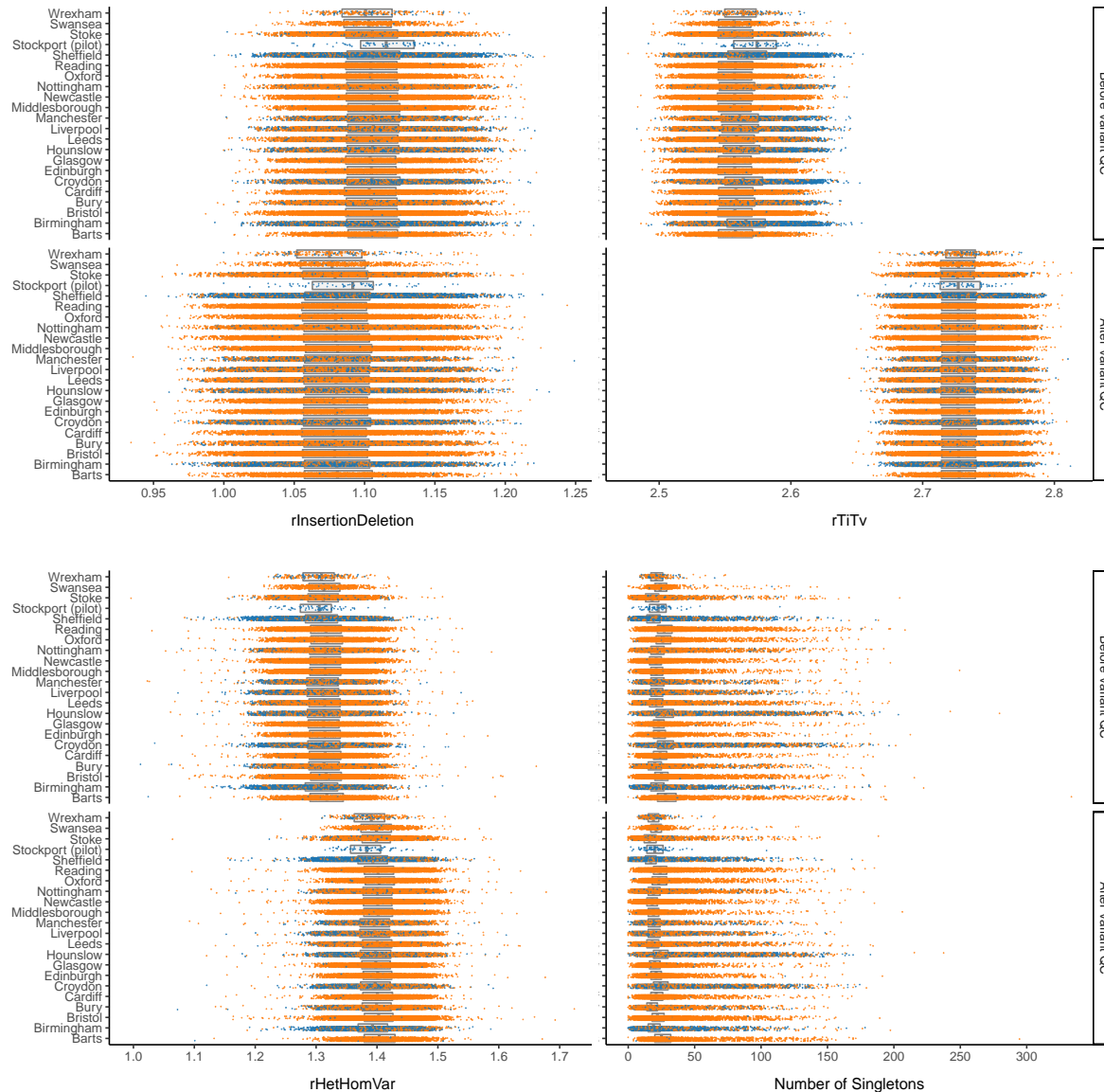
For our final variant filtering step, following restriction to the NFE subset, and removal of incorrectly defined sex or unknown sex, and run variant QC. We then filter out variants with call rate  $< 0.97$ , variants out of Hardy-Weinberg equilibrium (HWE) ( $P < 1 \times 10^{-6}$ ), and remove invariant sites following the previous sample based filters. After restricting to these high quality variants, we perform a final set of sample filters to finalize the quality controlled data. We evaluate a collection of sample metrics and remove samples falling outside four SDs of the sequencing batch mean (Ti/Tv, Het/HomVar, Insertion/Deletion ratios), and remove the collection of samples with over 175 singletons. The resultant curated analysis ready data set consists of 176,935 samples, and 9,169,408 variants (Supplementary Table 2-3).

A summary of sample and variant filters are provided in Supplementary Tables 2-3. The high quality ES call-set consisted of 176,935 samples and 9,169,408 variants.

## **Phasing**

### **Combining ES data with genotype array data**

We combined genotyping array (UK BiLEVE Axiom array and UKBB Axiom array) and exome chip (IDT xGen Exome Research Panel v1.0) variants after general ES quality control using Hail<sup>46</sup> and BCFtools<sup>53</sup> (1.12). For variants in both data sets, we preferentially retained those on the ES data. For variants on the genotyping array we excluded variants missingness  $> 5\%$  after performing a liftover to GRCh38 using Hail<sup>46</sup>. To avoid biasing the phasing quality estimates, we excluded parents among trio relationships prior to phasing. We first created a common variant scaffold by phasing variants in the combined (exome sequencing and genotyping array) data with MAF  $> 0.1\%$  and otherwise default parameters using SHAPEIT5\_PHASE\_COMMON module. We then phased the remaining rare variants using the common variant scaffold using the SHAPEIT5\_PHASE\_RARE with recommended parameters. To ensure computational tractability,

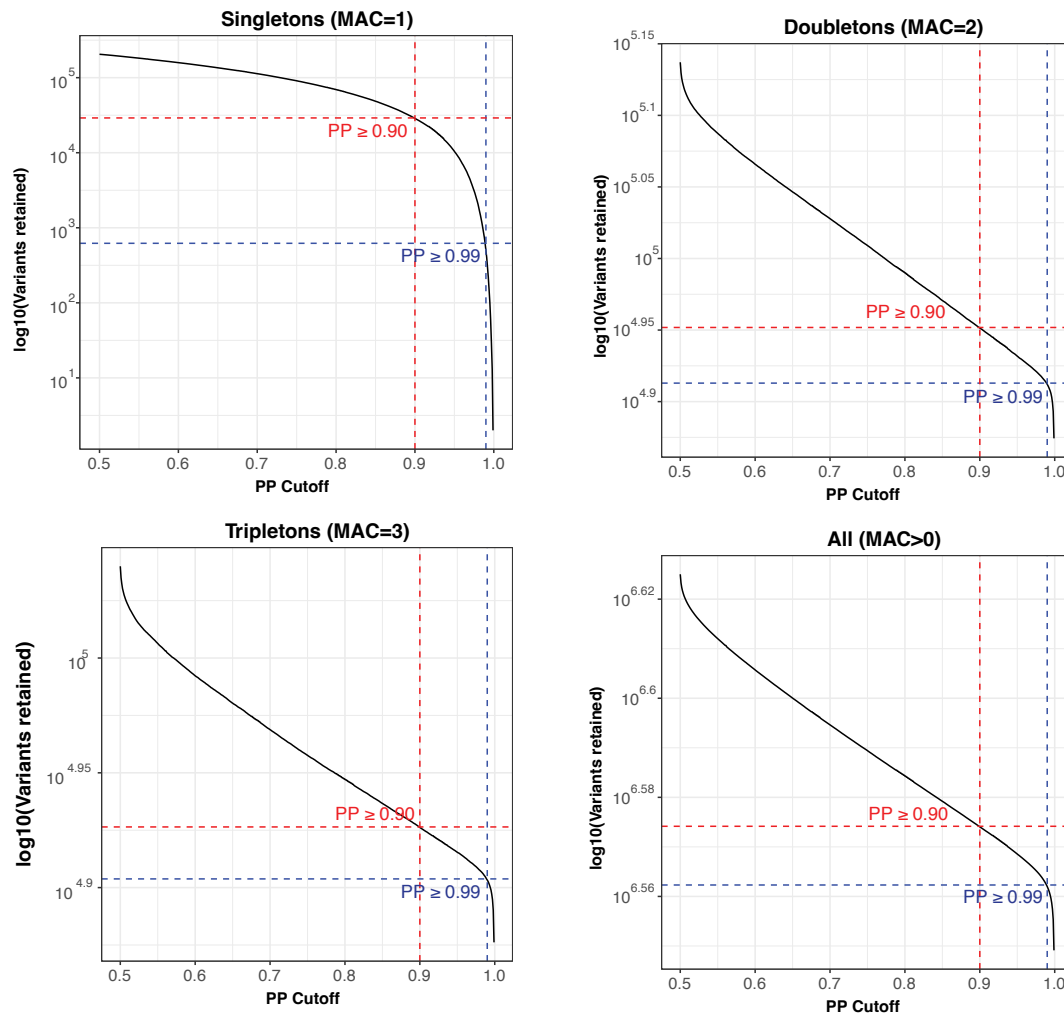


**Supplementary Fig. 4: Distributions of variant metrics before and after the removal of invariant sites, variants with call rate  $< 0.97$ , and variants out of HWE ( $P < 1 \times 10^{-6}$ ).** In each plot, jittered scatters display the distribution for each sequencing batch colored by sequencing batch. Box-plots behind the scatter display the median and interquartile range for each sequencing batch. Points shown are following variants hard-filters and prior to removal of variants with metrics outside four standard deviations of the sequencing batch mean.

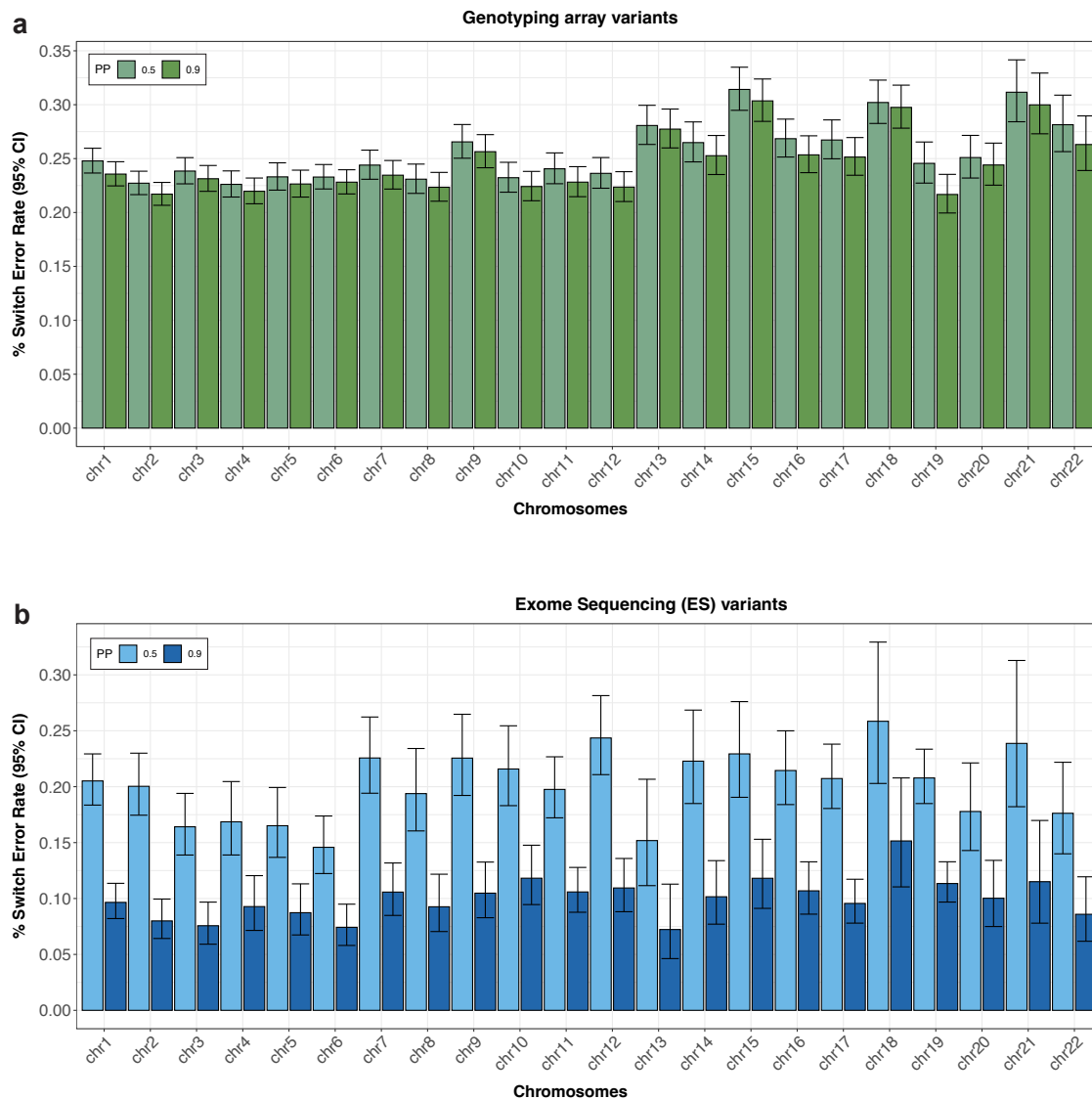
we phased overlapping chunks of 100,000 variants with  $\geq 50,000$  variant overlap between consecutive chunks using Hail<sup>46</sup>. Following chunk phasing, we then removed the initial and final 22,500 variants from each chunk, so that 5,000 overlapping variants remained between contiguous phased chunks. We then combined the phased chunks, matching haplotype phase using bcftools<sup>53</sup> (1.12) with the `--ligate` option. We then restrict this phased genetic dataset to the set of samples and variants present in the analysis ready NFE subset (Supplementary Tables 2-3).

### **Trio-switch error rates**

We assessed phasing quality by comparing statistically phased genotypes to those implied in 96 trios using Mendelian inheritance logic. Switch errors are determined by traversing the statistically phased and parent-offspring transmitted haplotypes simultaneously and scanning for inconsistencies in phase between pairs of contiguous variants. This method only allows us to consider sites in which the one parent is heterozygous and the other is homozygous for the reference or alternate allele, and thus do not consider *de novo* variants or Mendelian inconsistencies in the trio data. To assess switch error in a site-specific manner, we modified and recompiled bcftools<sup>53</sup> (1.12) to output errors by genomic position. We then used the modified version to assess switch by variant categories, for example by genetic data modality (genotyping array or ES), or by MAF bins. To evaluate switch errors across different phasing confidence thresholds, we filtered VCF using Hail<sup>46</sup> and then repeated the switch error calculation step. We calculated binomial 95% confidence intervals (CIs) for SERs using the R-package HMisc<sup>54</sup> (4.7).



**Supplementary Fig. 5: Phased variants retained as a function of phasing confidence score.** Each subplot displays the number of variants retained on the log<sub>10</sub> scale as the PP is increased, split by rarity of variants described in the subplot title. Dotted red and blue lines highlight the number of variants retained after imposing PP cut-offs of 0.9 and 0.99, respectively.



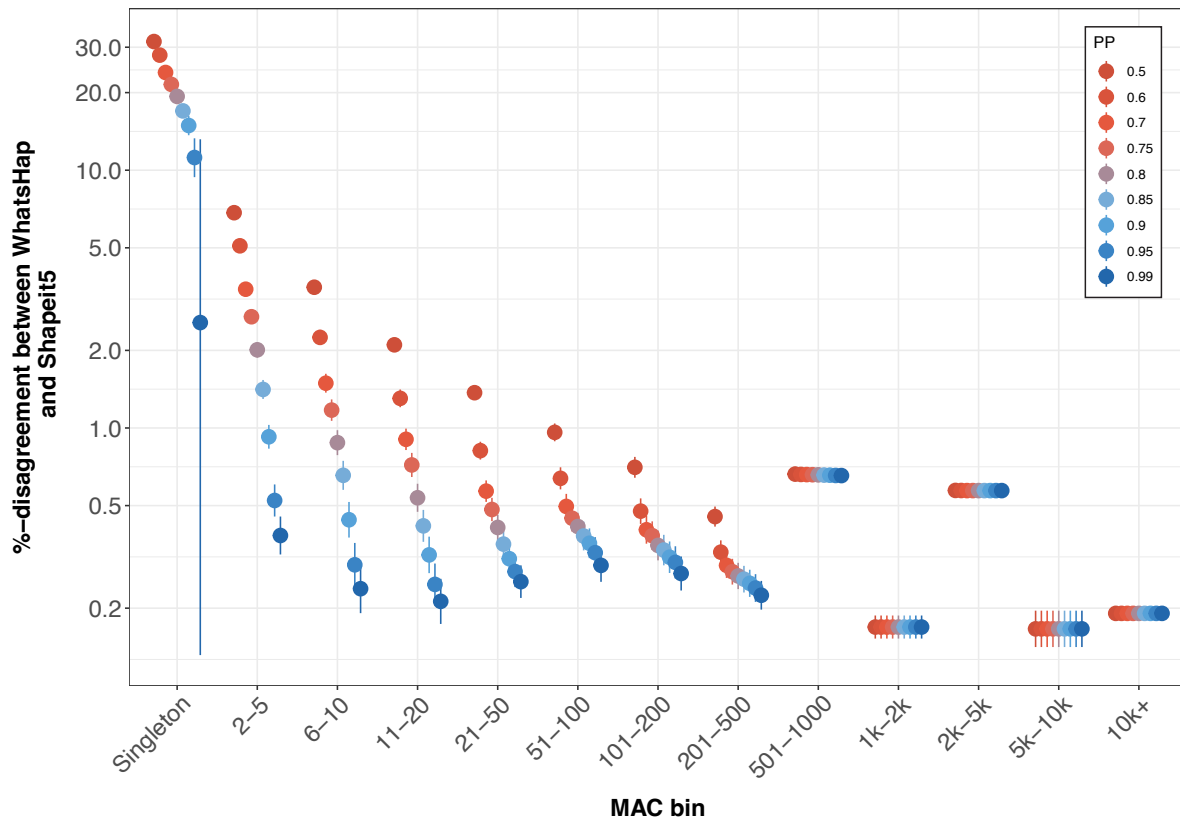
**Supplementary Fig. 6: Trio switch error rates by chromosome.** Parent-offspring trios are used to determine switch error rates for variants that originate from the genotyping array (a) and exome sequencing data (b). We stratify by phasing confidence (PP) according to the color legends. Mean switch error rates are plotting, with whiskers enclosing the 95% binomial CI.

## Read-backed phasing

We performed read-backed phasing with UKBB ES short paired-end read sequences using .cram files provided by UKBB. As WhatsHap is computationally expensive, we restricted our analysis to pairs of variants on chromosomes 20-22 in 176,586 genetically ascertained NFEs. We phased both single nucleotide polymorphism (SNV) and indel with WhatsHap<sup>26</sup> using the default recommended parameters. WhatsHap outputs lists of phased variants within ‘phased sets’. We carried forward reads overlapping no more than two variants, for which phase could be inferred. We combined these phased variants with statistically phased variants from SHAPEIT5 using Hail<sup>46</sup>, and determined agreement between estimated phasing in WhatsHap and SHAPEIT5 (Fig. 7).

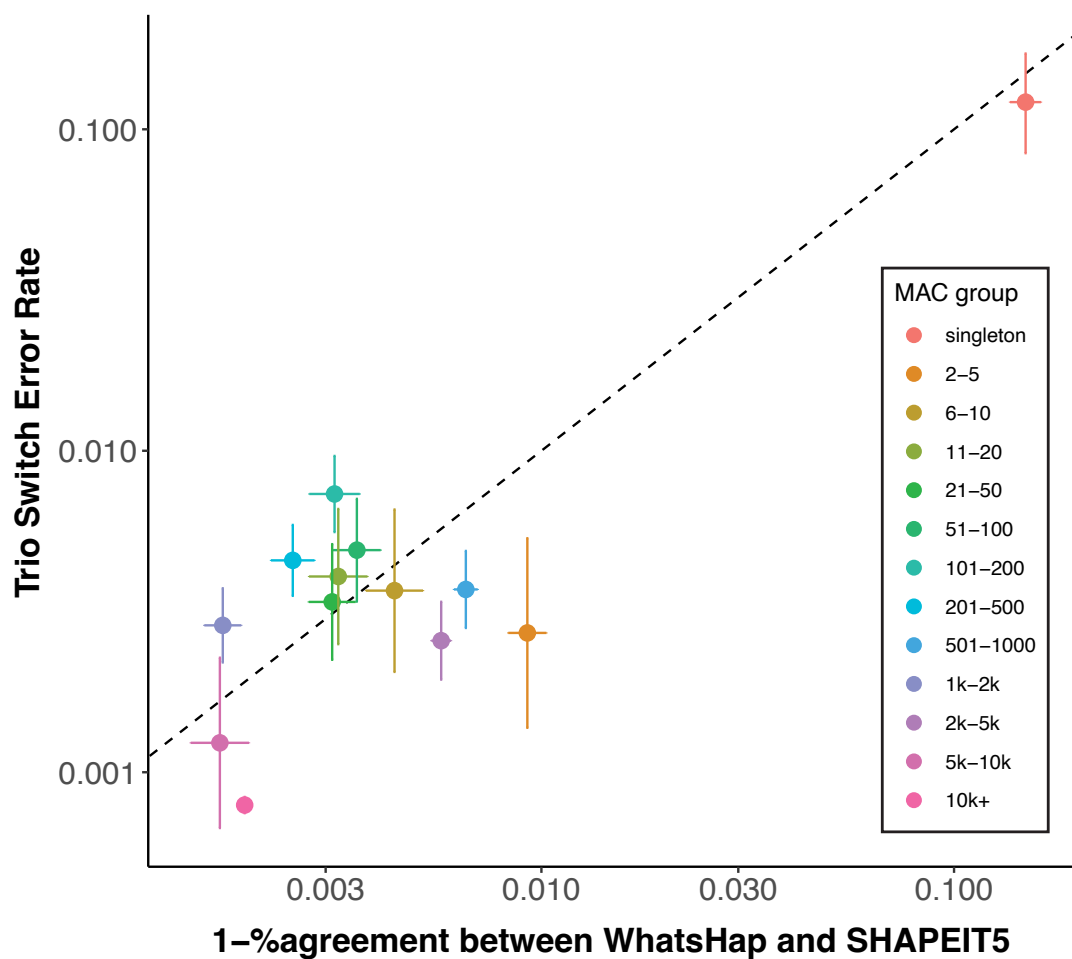
## Phenotype curation

We considered a collection of 282 binary quality controlled and publicly available common complex phenotypes for analysis<sup>55</sup>. To complement these, we also considered 28 common complex phenotypes that were obtained through manual curation, resulting in a total of 311 binary phenotypes for analysis. To increase our power for analyses for binary traits, we amalgamated a collection of phenotypes where possible: combining the phenotype curation of Censin et al.<sup>56</sup>, with the primary care mappings file provided by UK Biobank `all_lkps_map_v3.xlsx` and our own manual curation. We aggregated across ICD-10, ICD-9, operating codes, nurses interview reports, and self-reported diagnosis by doctor from the main phenotype file, as well as v2 and v3 read codes in the primary care data. As in Censin et al., we made use of the careful definitions of Eastwood *et al.*<sup>57</sup>, subsequently applied by Udler *et al.*<sup>58</sup> for diabetes subtype curation. Briefly, the algorithm developed in Eastwood *et al.* bins individuals into putative diabetes status using a collection of phenotypes in the UK Biobank data including self-reported diabetes diagnosis, age of diagnosis, medications, start of insulin within a year of diagnosis. We defined cases as



**Supplementary Fig. 7: Agreement between read-backed and statistical phase estimation.**

Genetic phase was estimated using WhatsHap (read-backed phasing) and SHAPEIT5 (statistical phasing) in 176,586 individuals on chromosomes 20-22. We only carried forward pairs of variants close proximity in which phase could be inferred using WhatsHap. We combined with statistically phased counterparts derived from SHAPIET5 and determine % disagreement of phase estimation of variant pairs on the y-axis, when filtering to phased pairs of variants where the minimum PP >  $p$  for  $p \in \{0.5, 0.6, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 0.99\}$  according to the color legend. We stratify pairs of variants into bins based on the minimum MAC in the variant pair, on the x-axis. Mean disagreement rates are plotted on y-axis with whiskers enclosing the 95% binomial CI



**Supplementary Fig. 8: Agreement between read-backed phasing and statistical phasing.**

We plot the disagreement between WhatsHap (read-backed phasing) and SHAPEIT5 (statistical phasing) in UKBB on the  $x$ -axis against switch error rate in SHAPEIT5 phase estimates implied by trio-based phasing in UKBB on the  $y$ -axis. For each comparison, bin pairs of variants according to the minimum MAC in the variant pair according to the color legend. Horizontal and vertical lines enclose 95% binomial CIs around mean estimates. The dotted line is included to display  $y = x$ .



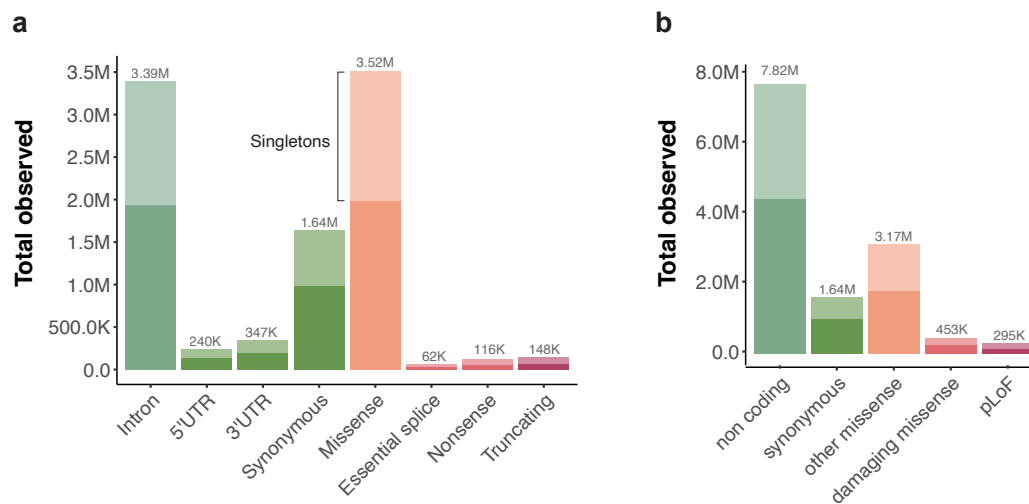
those placed in the probable and possible case categories in the algorithms output. Controls were defined as samples labeled as ‘diabetes unlikely’ by the algorithm.

## **Variant annotation masks**

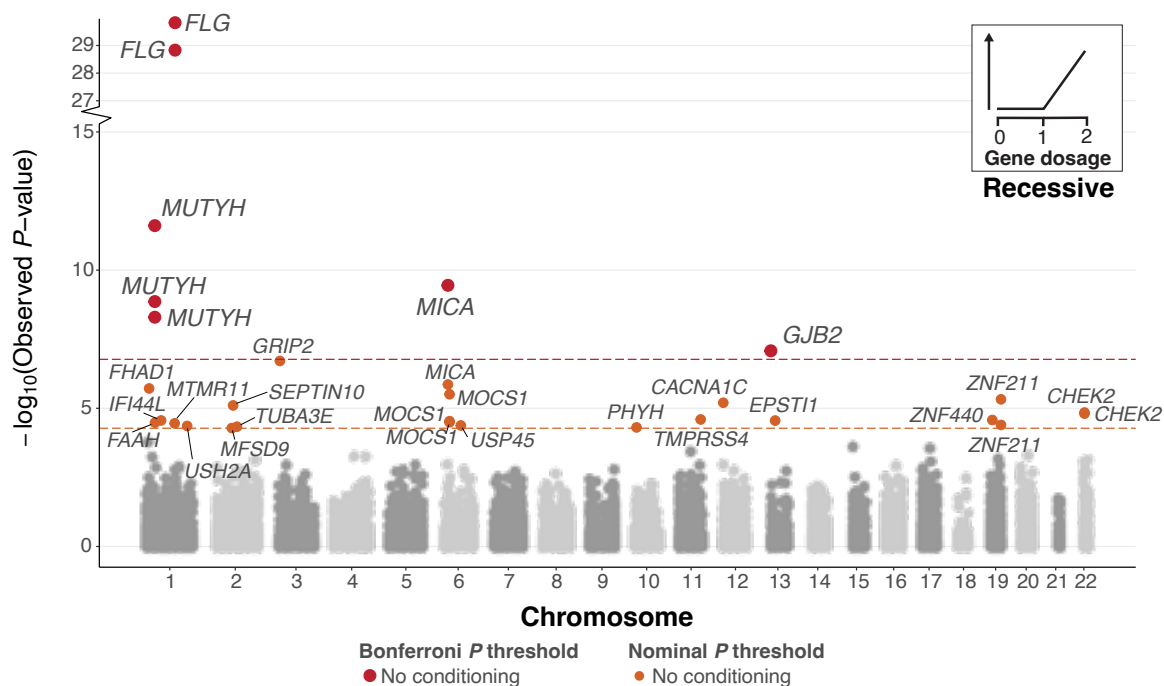
We annotated coding variation using Variant Effect Predictor (VEP)<sup>59</sup> (v95) using the worst consequence by gene within ‘canonical’ transcripts. We classified variants into four categories: protein truncating variants (PTVs), missense variants, synonymous variants, and other variants (Supplementary Table 9). We then split PTVs into putative loss of function (pLoF) (HC) and LC loss-of-function variants using LOFTEE<sup>60</sup>, and labeled missense variants with both Rare Exome Variant Ensemble Learner (REVEL)<sup>61</sup> score  $\geq 0.6$  and CADD<sup>62</sup> score  $\geq 20$  as ‘damaging missense’ or otherwise as ‘other missense’. Finally, we combine the resultant ‘damaging missense’ category with LC loss-of-function variants, which we denote as ‘damaging missense/protein-altering’.

## **Bi-allelic encoding and recessive association modeling**

Using custom Hail scripts, we define and annotate individuals as being ‘bi-allelic’ for a gene if they harbor at least one pLoFs or damaging missense variant with MAF  $< 5\%$  on both inherited copies of the gene. For each sample, we encoded the presence and absence of a damaging bi-allelic variant for each gene as zero and two, respectively. We encode this information in a .vcf file and test for an association between presence of a damaging bi-allelic variant in a gene and a trait using SAIGE<sup>63</sup>, adjusting for sex, age, sex  $\times$  age, age<sup>2</sup>, UKBB centre, genotyping batch and the first 10 PCs. We took relatedness into account using a sparse genetic relatedness matrix (GRM) fitted on NFE. We restrict analysis to (gene, trait) pairs with at least five bi-allelic variants in the curated ES with non-missing corresponding phenotype data (corresponding to a minimum MAC  $\geq 10$ ), and adjust for multiple testing at Bonferroni



**Supplementary Fig. 9: Distribution of variant annotation categories before and after broad consequence categorization.** We annotate variants using VEP and by the most severe consequence in the canonical transcript. Panels (a) and (b) display the total number of unique variants observed across a set of variant consequences colored by degree of predicted impact, before and after broad variant consequence categorization. In each panel, green, orange and red colored bars indicate low, medium and high impact respectively, according to the color legends. Singleton variation within the variant class is stacked and displayed in a lighter shade. Counts of variant within each annotation category are displayed above the bars. Note that all counts shown here are before filtering to accurately phased variants.



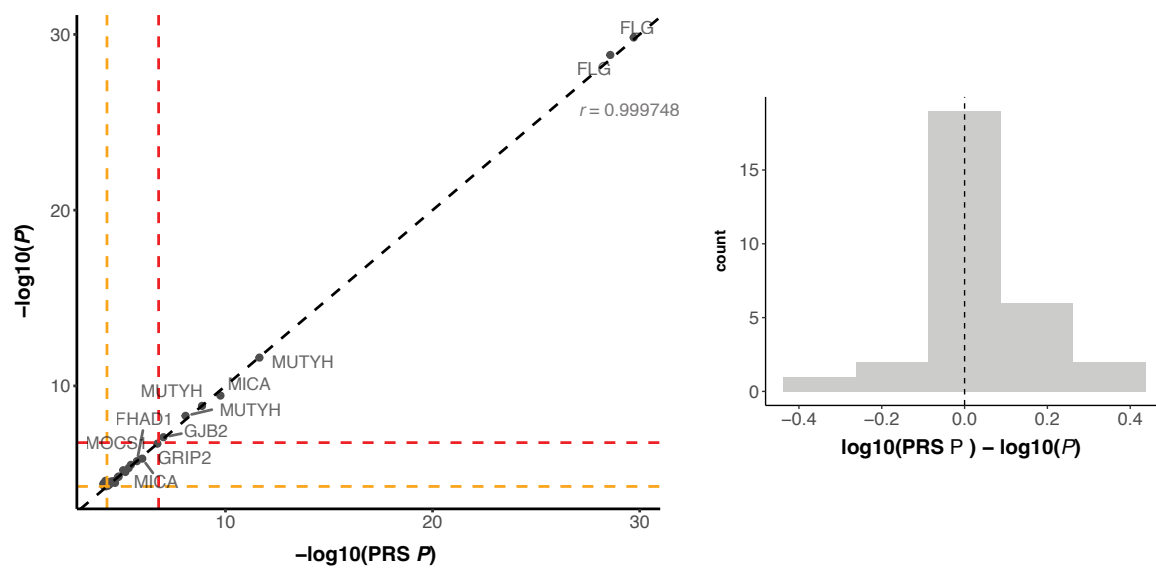
### Supplementary Fig. 10: Recessive analysis of 311 phenotype without accounting for PRS.

Recessive Manhattan plot depicting  $\log_{10}$ -transformed gene-trait association  $P$ -values versus chromosomal location. Associations are colored red or orange based on whether they are Bonferroni ( $P < 1.68 \times 10^{-7}$ ) or nominally ( $P < 5.25 \times 10^{-5}$ ) significant. No additional conditioning was carried out in this analysis.

significance ( $P < 0.05/\text{gene-trait pairs}$ ).

### Gene copy dosage encoding and additive association modeling

We define annotate individuals as being ‘mono-allelic’ for a gene if they harbor at least one pLoFs or damaging missense variant with  $\text{MAF} < 5\%$  on a single copy of the gene. Furthermore, if they harbor at least one pLoF or damaging missense variant on both inherited copies of the gene, we annotate them as ‘bi-allelic’. Using custom Hail scripts, we encode wildtypes, mono-allelic and bi-allelic carriers as 0, 1 and 2 respectively, thus representing the number of affected gene copies in an individual. We test for association using SAIGE<sup>63</sup>, adjusting for sex, age,



**Supplementary Fig. 11: Association  $P$ -values before and after inclusion of PRS as a covariate.** The scatter plot depicts the association  $P$ -values both before and after PRS was included as a covariate. The y-axis represents the  $P$ -value prior to PRS adjustment, while the x-axis demonstrates the  $P$ -value after PRS adjustment. On the right, the difference in log-transformed  $P$ -values before and after PRS adjustment is displayed. The plot exclusively showcases gene-trait associations that were considered nominally significant in the recessive analysis.

sex  $\times$  age, age<sup>2</sup>, UKBB centre, genotyping batch and the first 10 PCs. Again, we took relatedness into account using a sparse GRM fitted on NFE. We restricted to gene-pairs with at least 10 disrupted haplotypes (corresponding to a minimum MAC  $\geq 10$ ), and adjust for multiple testing at Bonferroni significance ( $P < 0.05/\text{gene-trait pairs}$ ).

## **Polygenic risk scores**

### **Curation of array-based genetic data**

We generated PRSs using imputed genotypes provided by UKBB<sup>48</sup>. In the following, we make the distinction between training and testing data. The first represents the samples that are used for fitting LDpred2<sup>64</sup> weights and parameters while the latter represent the samples with bi-allelic variant (with homozygous or CH status) information in which we use to assess the predictive accuracy the fitted LDpred based PRS. For the training data, we took the genetically ascertained NFE and filtered to 246,152 unrelated samples (kinship coefficient  $< 2^{-4.5}$ ) that did not have quality controlled ES data available. NFE samples with high quality ES and imputed genotype data available were used for testing. Where predictive (nominal significant  $h_{\text{snp}}^2$  and  $n_{\text{eff}} \geq 5000$ ), we include PRS as a covariate for downstream biallelic association testing to account for common variant polygenic risk for the trait under investigation.

### **Genotype variant filtering**

We followed best practices from Privé *et al.*<sup>64</sup>, and filtered to common Haplotype Map 3 (HM3) SNPs<sup>65</sup>. Additionally, we exclude any variants with genotyping proportion  $< 1\%$  and MAF  $< 1\%$ , resulting in a total of 1,165,296 common autosomal variants for fitting PRS weights. To reduce the likelihood of spurious correlations between low-frequency variants in traits with low case or control count, we restricted to binary phenotypes with at least 1,250 cases and controls. Additionally, we imposed a phenotype specific MAF filter based on the number of cases and

controls in a trait, specifically:

$$\text{MAF} > \max(0.01, 2 \times \min(n_{\text{cases}}, n_{\text{controls}})), \quad (1)$$

where  $n_{\text{cases}}$  and  $n_{\text{controls}}$  are the numbers of cases and controls with high quality imputed sequence data available, respectively, to guard against non-causal variants that are overrepresented in cases or controls leading to false positive associations.

### **Common variant association testing**

We tested for associations between the 1,165,296 common autosomal HM3 variants and phenotypes using Hail<sup>46</sup>, running logistic regression (`logistic_regression_rows`) adjusting for sex, age, sex  $\times$  age, age<sup>2</sup>, UKBB assessment centre, genotyping batch and the first 10 PCs, using a Wald test.

### **Estimating heritability**

We generated LD-scores for HM3 variants in sample, using a random subset 10,000 of 246,152 unrelated genetically ascertained NFEs without haplotype information. Using the genome-wide association study (GWAS) summary statistics and LD-scores, we estimated SNP heritability  $h_{\text{snp}}^2$  and standard errors (SEs) using LD score regression (LDSC)<sup>66,67</sup>. We evaluated PRS for phenotypes with nominal significant  $h_{\text{snp}}^2$  estimates ( $P < 0.05$ ) and restricted to phenotypes with nominally significant ( $P < 0.05$ ) LDSC based SNP heritability estimates and effective sample size  $n_{\text{eff}} \geq 5,000$ , where:

$$n_{\text{eff}} = \frac{4}{\frac{1}{n_{\text{cases}}} + \frac{1}{n_{\text{controls}}}}. \quad (2)$$

### **Generating PRS using LDPred2**

For a given phenotype, we trained a PRS predictor with LDPred2-auto<sup>64</sup>, using marginal effect size estimates evaluated on the 246,152 unrelated NFE samples (defined by kinship coefficient

$< 2^{-4.5}$ ) without ES data in the 200k ES UKBB release),  $h_{\text{SNP}}^2$  as estimated by LDSC, and in-sample reference panel to evaluate local LD, as input. We removed any invariant sites and mean-imputed missing genotypes, before training the predictor. Following PRS training, we then predict into the 176,266 samples with ES and high-quality imputed genotype data.

### **Validation of polygenic risk scores**

We assessed the ability of the resulting PRS to discriminate between case status by evaluating area under the curve (AUC) on the held-out unrelated set of samples with both HM3 SNPs and phased exome data. We used the function AUCBoot from the R package bigstatsr<sup>68</sup> (1.5.6) to extract 10,000 bootstrap replicates of individuals and compute the 95% CIs for AUC.

## **Conditional analysis**

### **Off-chromosome PRS conditional analysis**

For each chromosome,  $C$ , we evaluated ‘off-chromosome’ PRS by setting weights on chromosome  $C$  to 0. We repeated this for each phenotype with PRS available and fit SAIGE<sup>63</sup> models while controlling for off-chromosome PRS by including it as a covariate in the null SAIGE model.

### **Common variant conditional analysis**

To assess whether a putative signal in a gene is driven by nearby common variation, we filtered to samples that have both ES and imputed genotypes with  $\text{MAF} > 1\%$  and imputation INFO score  $> 0.5$ . Then, for each gene that passed exome wide significance in the primary analysis ( $P < 5 \times 10^{-6}$ ), we tested for common variant associations in the region (1 Mb upstream and downstream of the gene). For each of these regions, we took an iterative approach, testing for common variant associations using SAIGE<sup>63</sup>, conditioning on the lead variant and repeating the regression until the conditional  $P$  for the newly included variant dropped below  $5 \times 10^{-6}$ ,

allowing up to 25 ‘independent’ associations in the region. We used the same covariates as in the primary analysis. For every variant that passed exome-wide significance ( $P < 5 \times 10^{-6}$ ), we encoded the genotypes as dosages and embedded them alongside pseudo variants (bi-allelic variants) in a VCF. We then re-ran the primary analysis twice (with and without controlling for off-chromosome PRS), while conditioning on any nearby common variant signals of association with the phenotype of interest.

### **Rare and ultra-rare variant conditional analysis**

For each significant ( $P < 1.68 \times 10^{-7}$ ) gene-trait associations in the genome-wide analysis after conditioning on PRS and nearby common variants association signals, we considered a further conditioning step. We sought to determine whether the residual signal of association could be explained by additive rare variant effects within the associated gene. To do this, ran further conditioned on rare ( $MAC \geq 10$ ,  $MAF \leq 0.05$ ) and ultra-rare ( $MAC \leq 10$ ) variants annotated as either pLoF or damaging missense within each gene. Because conditioning on ultra-rare variation can lead to convergence issues, we performed a gene-wide collapsing of ultra-rare ( $MAF \leq 10$ ) variants, thus aggregating them into a single ‘super’ variant to represent burden of ultra-rare damaging variation in the gene. Following this collapsing, we were able to condition on the ultra rare and rare variant contribution using SAIGE, while also conditioning on PRS and nearby common variant association signals when applicable.

### **Permutation of genetic phase**

To test whether a putative gene-trait association is driven by compound heterozygosity, we designed a permutation-based pipeline that could be systematically applied and scaled across phenotypes and genes. To do this, we label samples that are either CH variants or heterozygous *cis* carriers and then randomly shuffle these labels a series of times. For each permutation, we



re-run the association analysis conditioning on covariates as previously discussed (including off chromosome PRS and nearby common variants), and determine the resultant association strength under this label shuffling. Applying this permutation procedure multiple times, we can determine an empirical null for the association strength in the absence of phase information. The result is an empirical distribution of  $t$ -statistics and corresponding  $P$ -values that reflect the degree of association that would be expected given that the phase is random. We evaluate the one-sided empirical  $P$ -value, specifically:

$$P_{\text{empirical}} = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(t_i \geq t_{\text{observed}}) \quad (3)$$

where  $n$  is the number of permutations,  $t_i$  is the  $t$ -statistic under the  $i^{\text{th}}$  random label shuffling, and  $t_{\text{observed}}$  is the observed  $t$ -statistic determined using the observed genetic phase. To ensure sampling of  $t$ -statistics at a sufficiently large number of configurations of the genetic phase, we analyzed gene-trait pairs with at least ten compound heterozygotes and/or samples with multiple variants on the same haplotype. We permuted up to 100,000 times. To control for multiple testing, we corrected for 5 gene-traits tested (Bonferroni significance threshold  $P < 0.05/5 = 0.01$ ).

## Gene-set enrichment of bi-allelic variation

### Analyzed gene-sets

We included the following gene lists in our gene-set enrichment analyses: essential in mice<sup>69</sup>, essential gnomAD<sup>27</sup>, essential ADaM<sup>70</sup>, essential in culture<sup>71</sup>, essential CRISPR<sup>72</sup>, genes with  $pLI > 0.9$  in gnomAD<sup>27</sup>, non-essential in culture<sup>71</sup>, homozygous LoF tolerant<sup>27</sup>, and non-essential gnomAD<sup>27</sup>.

## Poisson regression to assess enrichment of CH variants in gene-sets

We test for depletion and gene-set enrichment using poisson regression. We model the count of bi-allelic variants across samples as a function of gene-set and mutation frequency using the `glm` function in R.

$$|\text{samples with } > 1 \text{ variant of class } x \text{ in gene}| \sim \mathbb{I}(\text{gene-set}) + \text{mutation rate} \quad (4)$$

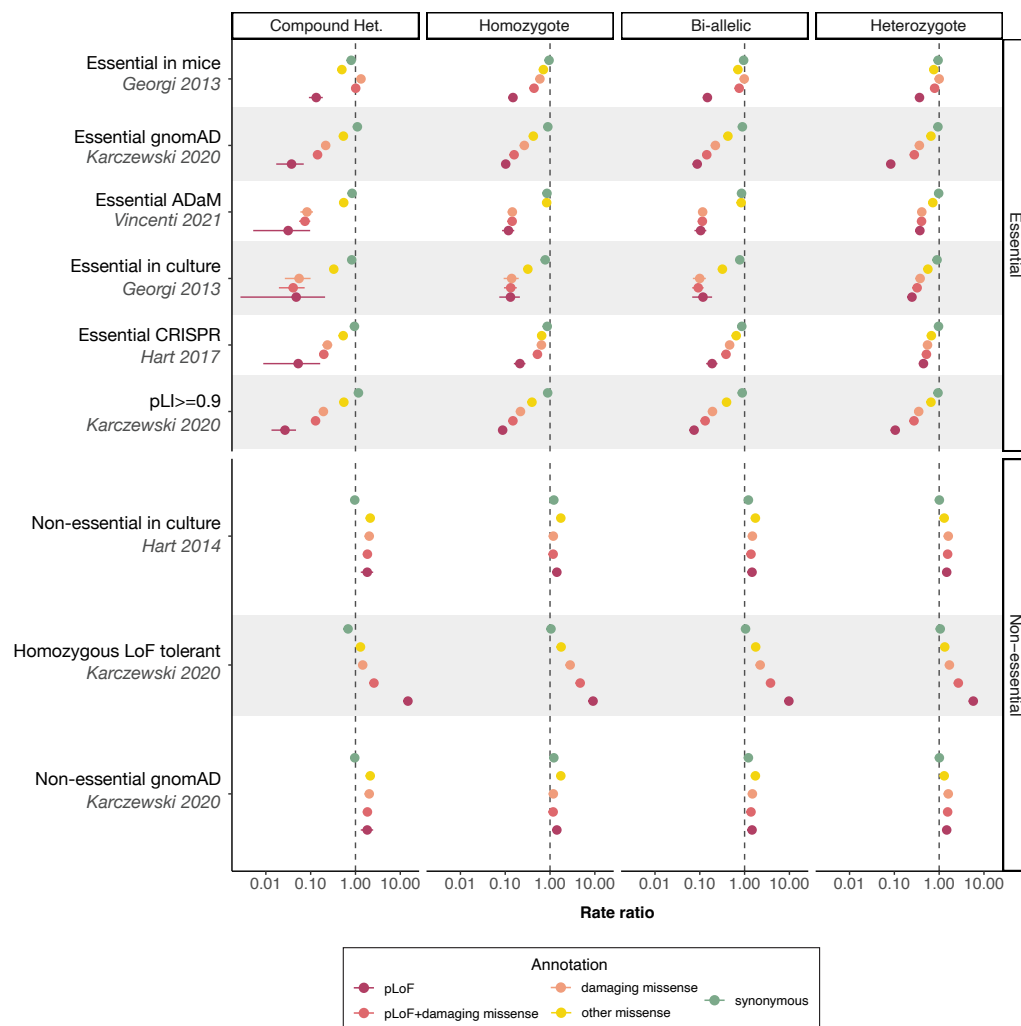
where  $x$  is a pair  $(x_1, x_2)$ :  $x_1 \in \{\text{pLoF, damaging missense, pLoF and/or damaging missense, other missense, synonymous}\}$ ,  $x_2 \in \{\text{heterozygote, CH, bi-allelic variants}\}$ . For each annotation category we use the transcript-specific mutation rate<sup>28</sup>. 95% confidence intervals are determined using `confint.glm` from the MASS-package (v7.3-58.1).

## Homozygote and CH down-sampling

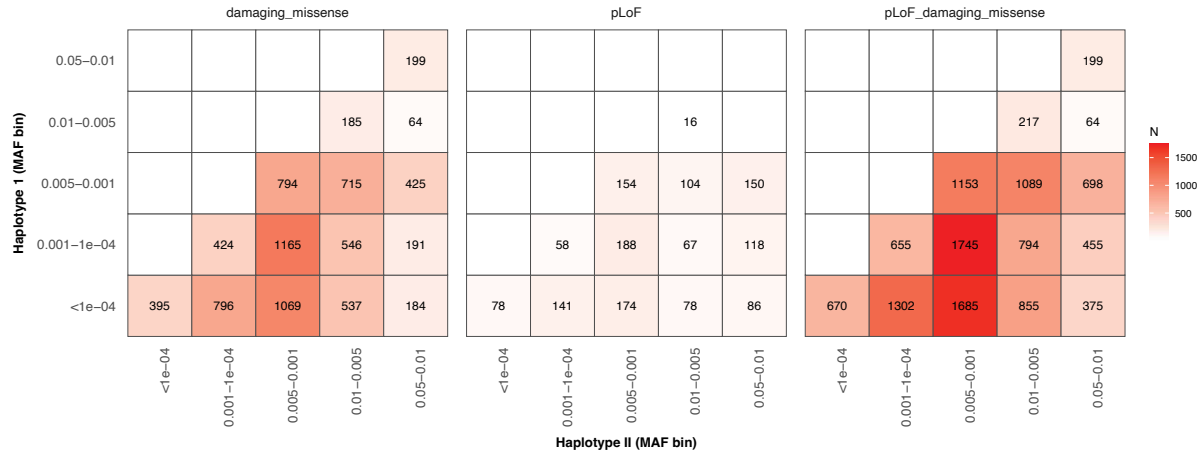
To investigate the number of identifiable CH or homozygous events across varying sample sizes and variant annotations, we performed down-sampling across the total population of 176,587 individuals. To do this, we defined a set of 35 regularly spaced cutoffs between 1,000 and 176,587 samples using increments of 5000. To determine uncertainty in our estimates of the number of unique genes implicated as a homozygote and/or CH, we randomly sampled individuals for each down-sampling 100 times, with replacement. We calculated the 95% CI by taking the 2.5% and 97.5% quantiles for the number of unique genes affected at a given sample size, and repeated across annotations (Fig. 19).

## Power analysis for bi-allelic association

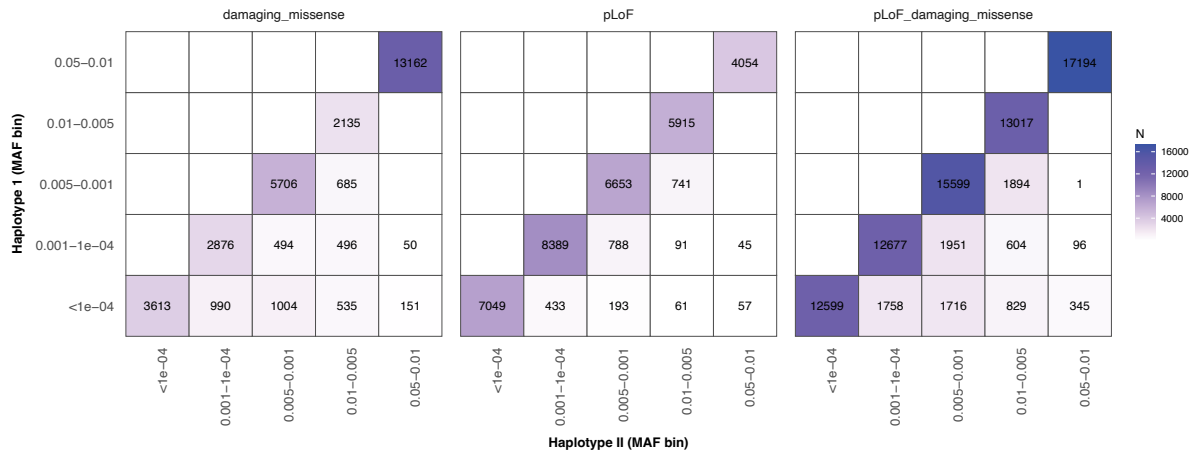
We perform a power analysis based on bi-allelic (including both CH and homozygous) variant frequencies in the population. To do this, we adopted `code`<sup>73</sup> allowing us to determine the effective effect size on the OR scale across candidate configurations of binary case-control counts



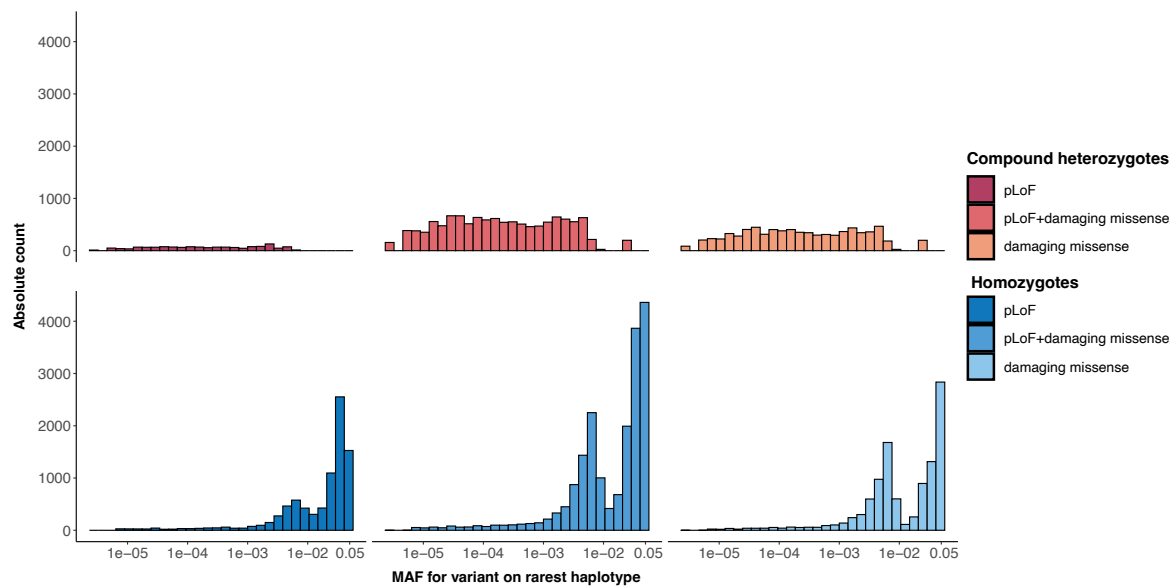
**Supplementary Fig. 12: Gene-set depletion/enrichment modeling.** Poisson regression to model mono- and bi-allelic variant (heterozygous, CH, homozygous or both) depletion and enrichment across essential and non-essential gene-sets. Rate ratios are shown for synonymous (green), other missense (yellow), damaging missense (orange) and pLoF (red) variants. The dashed line depicts a rate ratio of 1.



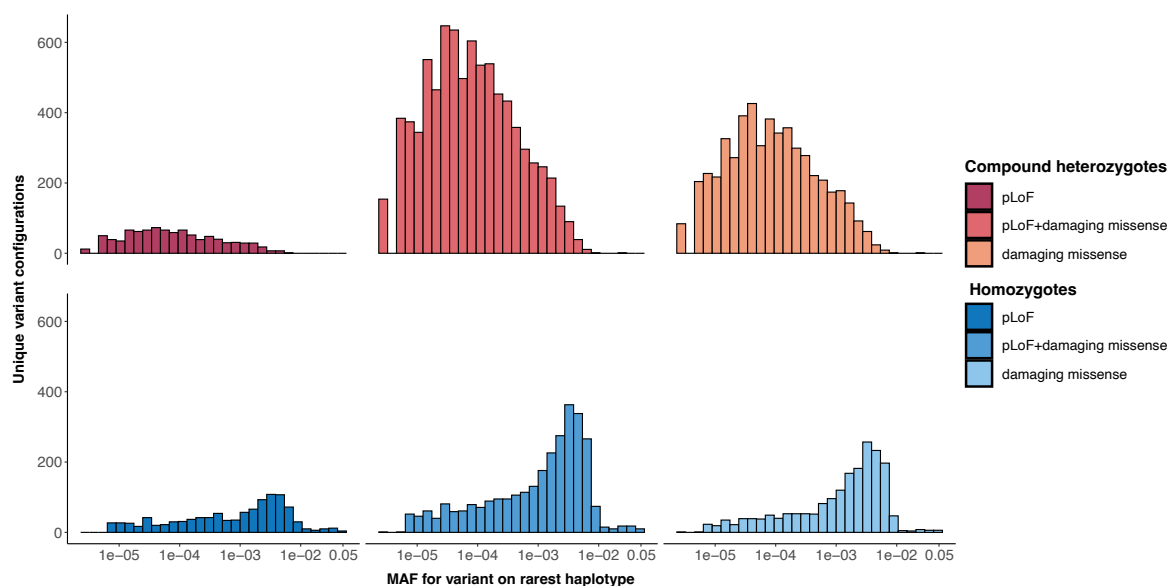
**Supplementary Fig. 13: Allele frequencies of variants in the CH state.** Heatmap of allele counts for variants in CH state stratified by predicted variant consequence (damaging missense, pLoF or pLoF+damaging missense). We plot the MAC for variants residing on the most common haplotype (y-axis) versus the rarest haplotype (x-axis).



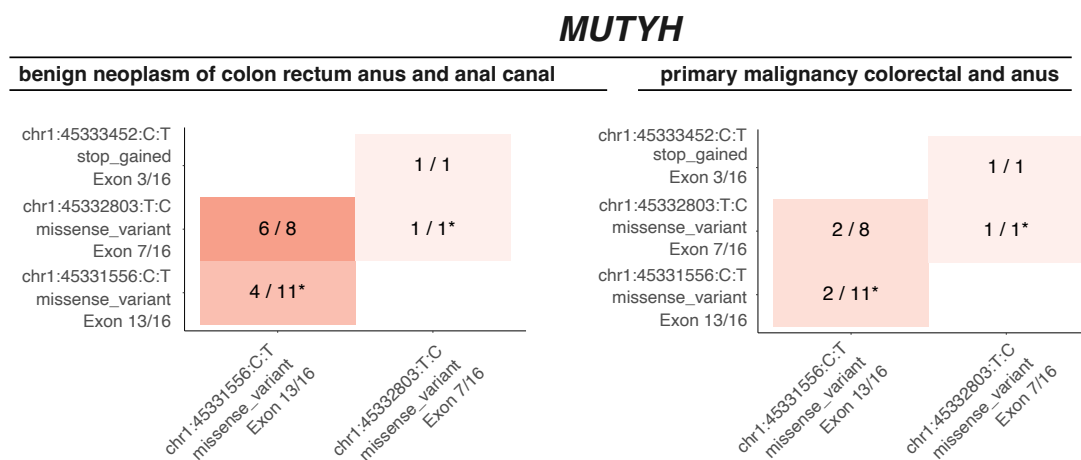
**Supplementary Fig. 14: Allele frequencies of variants in cis.** Heatmap of allele counts for co-occurring variants on the same haplotype stratified by predicted variant consequence (damaging missense, pLoF or pLoF+damaging missense). The most common variant on the haplotype versus the least common are plotted.



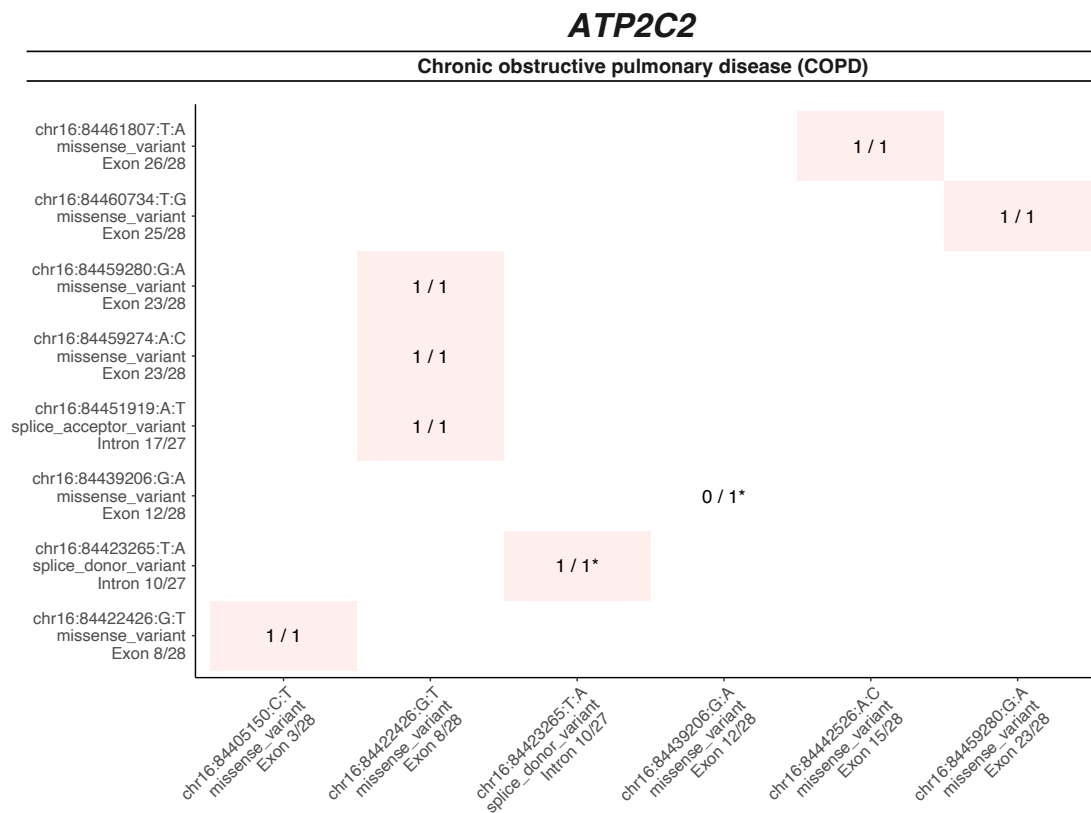
**Supplementary Fig. 15: Distribution of observed variants across samples by allele frequency.** Histogram of unique bi-allelic variant (CH and homozygotes) prevalence across the allele frequency spectrum. For a qualifying CH variant, the allele frequency corresponding to the alternate allele on the rarest haplotype are plotted.



**Supplementary Fig. 16: Distribution of unique variants observed by allele frequency.** Histogram of bi-allelic variant (CH and homozygotes) count for all gene-samples pairs in the analysis. For a qualifying CH variant, the allele frequency corresponding to the alternate allele on the rarest haplotype are plotted.

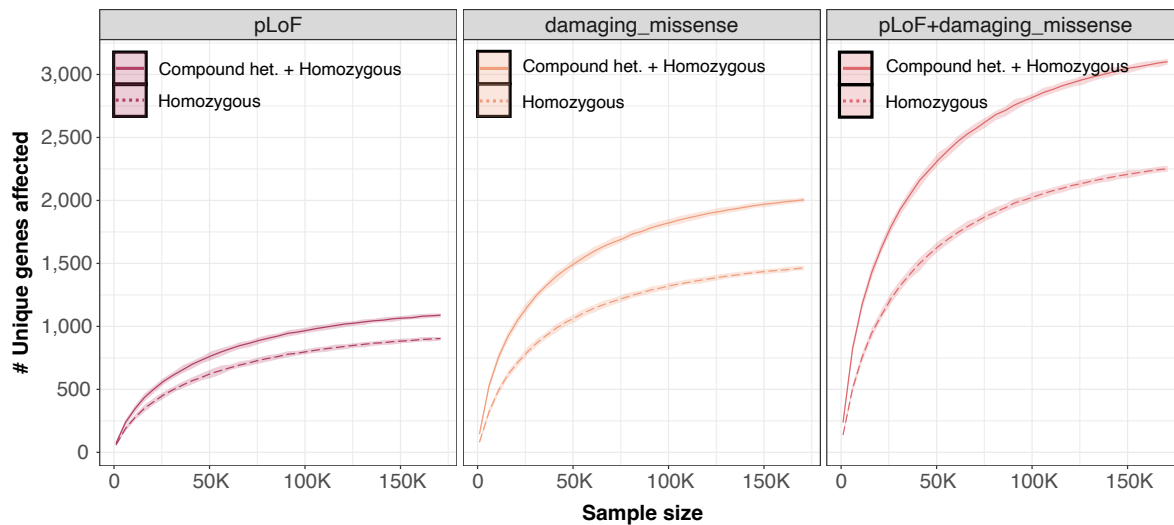


**Supplementary Fig. 17: Co-occurrence of deleterious *MUTYH*-variants across colorectal cancer outcomes.** Bi-allelic variant occurrence in *MUTYH* for benign neoplasm of the colon, rectum and anal canal (left) and primary malignancy of the colon (right). The constituent variants are shown alongside the variant consequence and involved exon or intron. Each cell indicates that number of individuals that are cases out of the total bi-allelic carriers identified. Homozygous cases and carriers are indicated with a star (\*)



**Supplementary Fig. 18: Co-occurrence of deleterious *ATP2C2* variants by COPD status.** Bi-allelic variant occurrence in *ATP2C2* for chronic obstructive pulmonary disease (COPD). The constituent variants are shown alongside the variant consequence and involved exon or intron. Each tile indicates that number of individuals are cases out of the total bi-allelic carriers identified. Only the variants that affect both gene copies are shown. Stars (\*) are included in the label to indicate homozygosity.





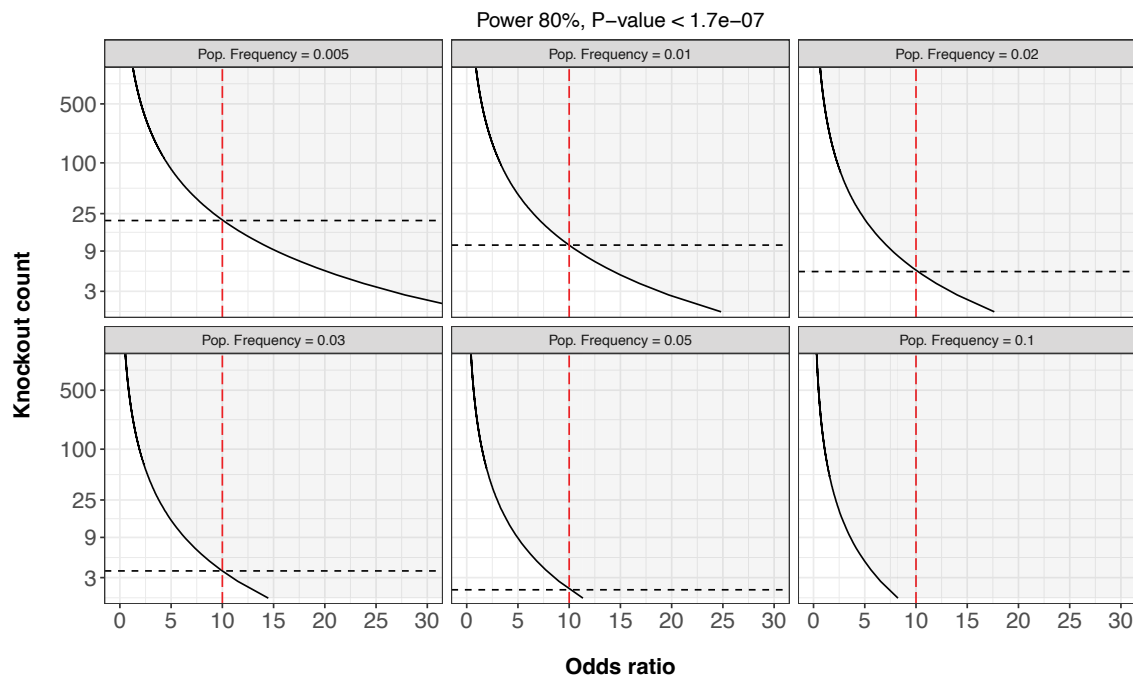
**Supplementary Fig. 19: Count of unique genes affected by a homozygous and homozygous or compound heterozygous variants as a function of sample size.** Starting with the full data, and down-sampling, we plot counts number of unique genes harboring homozygous and homozygous or compound heterozygous variants as a sample size is decreased. Class of variants in each count are denoted according to the key. Each facet indicate a specific variant annotation.

by substituting alternate allele frequencies with bi-allelic variant frequencies. We calculated effect sizes at 80% power at Bonferroni significance ( $P < 1.7 \times 10^7$ ) for a hypothetical traits with 823 (0.5%), 1766 (1%), 3532 (2%), 5298 (3%), 8829 (5%) cases of 176,587 total samples.

## Simulation

### Simulation of synthetic phenotypes using real genotypes

We performed a series of simulations to test that our pipeline would detect a CH effect in the presence of a true signal. We sampled 100,000 genetically-ascertained NFEs in the UKBB data, and extract chromosome 22 which we then use to simulate phenotypic data with a recessive genetic architecture. To emulate a scenario in which defects in protein coding genes lead to disease, we annotated the filtered UKBB genetic data and determined the collection of samples harboring damaging bi-allelic variants in each gene (compound heterozygous and homozygous,



**Supplementary Fig. 20: Power analysis to determine the required number of bi-allelic variants to detect specific ORs at 80% power at bonferroni significance ( $P < 1.7 \times 10^{-7}$ ).** We repeat the analysis while varying trait population prevalence assuming 823 (0.5%), 1766 (1%), 3532 (2%), 5298 (3%), 8829 (5%) cases out of 176,587 total individuals. The dashed red lines in the plot demonstrate the required number of bi-allelic variants to detect an OR  $\geq 10$ .

comprised of variants annotated as pLoF or damaging missense). We then define a  $n$  samples  $\times$   $m$  genes matrix  $\tilde{\mathbf{B}}$  with entries:

$$\tilde{\mathbf{B}}_{i,j} = \begin{cases} 1, & \text{if a damaging bi-allelic variant is present in sample } i \text{ at gene } j \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

We then simulated liability under the following model:

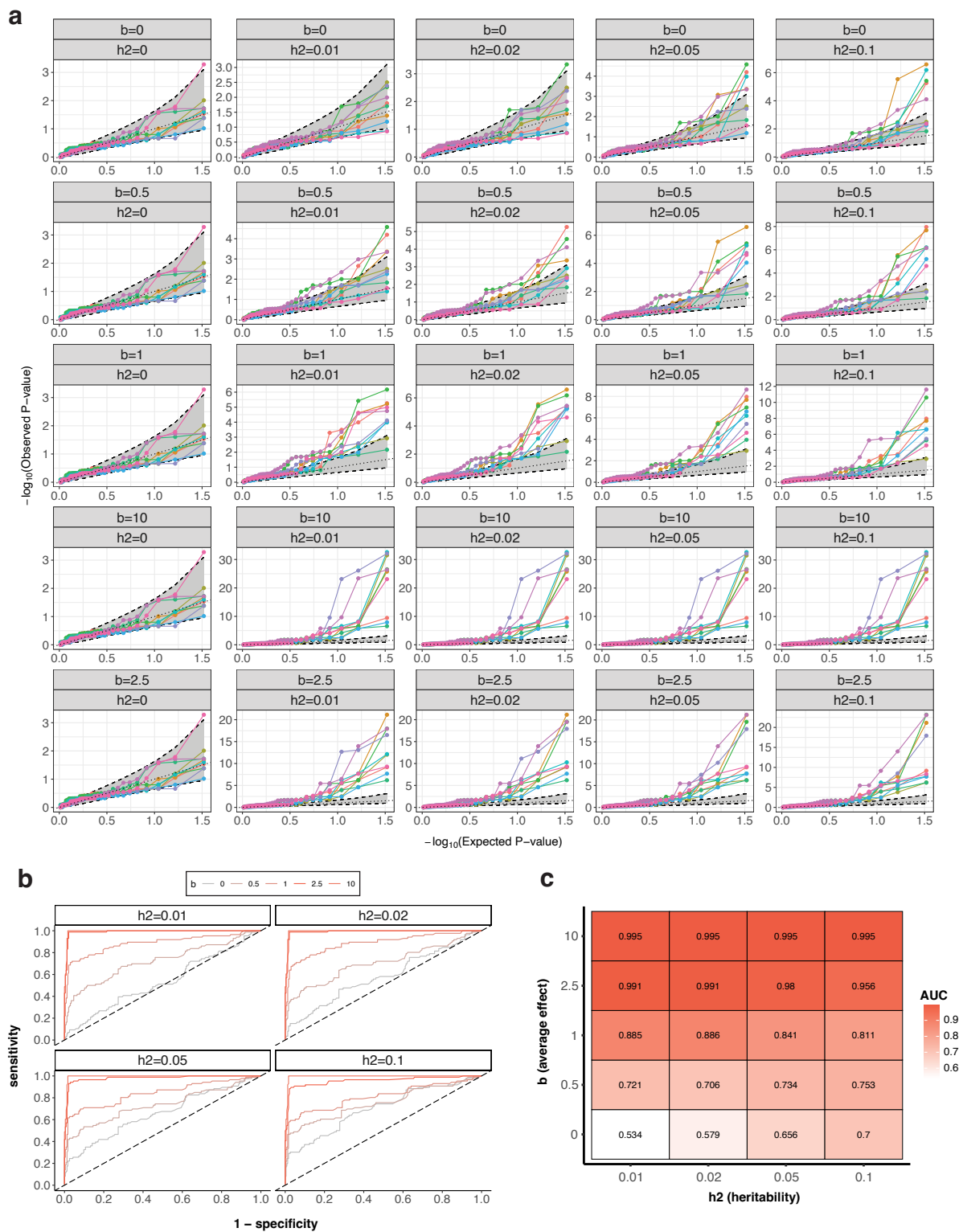
$$y_i = \sum_{j=1}^m \tilde{\mathbf{B}}_{i,j} \theta_j + \varepsilon_i \quad (6)$$

where  $\tilde{\mathbf{B}}_{i,j}$  is the  $(i, j)$ <sup>th</sup> entry of  $\tilde{\mathbf{B}}$  after standardizing the columns of  $\tilde{\mathbf{B}}$ ,  $E[\theta_j] = \frac{b}{m}$ ,  $\text{Var}[\theta_j] = \frac{h^2}{m}$ , and  $\varepsilon_i \sim \mathcal{N}(0, 1 - h^2)$ . Here, we implicitly assume that presence of at least one homozygous or CH variant of any type within a given gene contributes the same risk to disease, whose average across genes is set by the parameter  $b$ . The resultant liability  $y_i$  has mean 0 and variance 1. Note that the standardization of  $\mathbf{B}$  imposes a frequency dependent relationship between prevalence of bi-allelic damaging variants in a gene and variance explained. We simulated under the spike-and-slab model:

$$\theta_j \sim \begin{cases} \mathcal{N}\left(\frac{b}{m\pi_\theta}, \frac{h^2}{m\pi_\theta}\right), & \text{if } p_j < \pi_\theta \\ 0, & \text{otherwise} \end{cases}$$

$$p_j \sim \text{Bernoulli}(\pi_\theta)$$

in which  $\pi_\theta \in [0, 1]$  is the proportion of causal genes with a recessive contribution to the phenotype. Finally, to obtain binary traits we used the liability threshold model assuming a case prevalence of 10%. In the following simulations, we set  $\pi_\theta = 0.25$ , and considered  $h^2$  values of  $h^2 \in \{0, 0.01, 0.02, 0.05, 0.10\}$  and  $b$  values of  $b \in \{0, 0.5, 1, 2, 10\}$ .



Supplementary Fig. 21:

---

**Supplementary Fig. 21 (previous page): Simulation study to test our ability to detect bi-allelic effects in the presence of true effects.** We simulate phenotypic data applied to 100,000 genetically-ascertained NFE on chromosome 22 (Methods) under the liability-threshold model assuming a spike and slab genetic architecture. We assume a 10% disease prevalence and 25% causal genes, and consider varying levels of phenotypic variance explained by these effects  $\in \{0, 0.01, 0.02, 0.05, 0.10\}$ . We then apply SAIGE to the simulated phenotypes, testing for an association between presence of a bi-allelic variant in each gene and case status. **a)** Each panel indicates a set of simulations assuming varying levels of heritability and average effect as labeled in the subtitles. In each panel, we plot the true effect size in the simulation for a given gene on  $x$ -axis against the corresponding  $-\log_{10}(P)$  value of association. Areas of circles correspond to the number of samples harboring bi-allelic damaging variants in the 100,000 samples according to the legend. **b)** To assess the sensitivity and specificity of our approach, we created ROC-AUC curves for each combination of increasing phenotypic variance explained (facet) and increasing average affect (red lines). **c)** For each ROC-AUC curve from b, we calculate the AUC. White indicates low AUC and red indicates higher AUC.

## Longitudinal effects

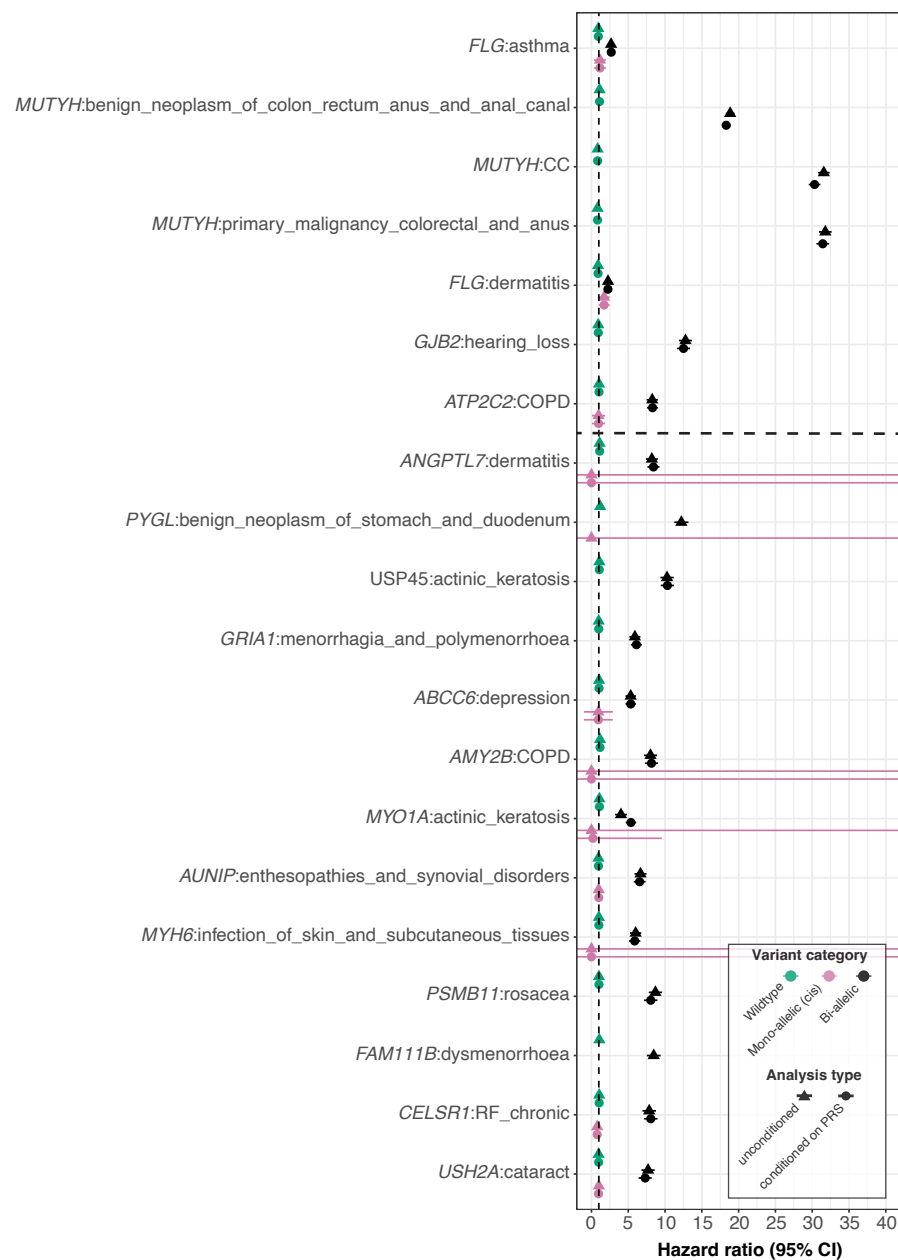
### Time-to-event data curation

We curated age-at-diagnosis for 278 binary phenotypes from the UKBB-linked primary care and hospital record data. 251 phenotypes were curated using the mapping tables generated by Kuan *et al.*<sup>55</sup>, excluding any codes related to “history of...” events for which accurate age-at-diagnosis could not be extracted. The remaining 27 phenotypes individuals’ records were left-truncated at the age of first record (of any code) in either the primary care or hospital data, and right-censored at the age of the last record.

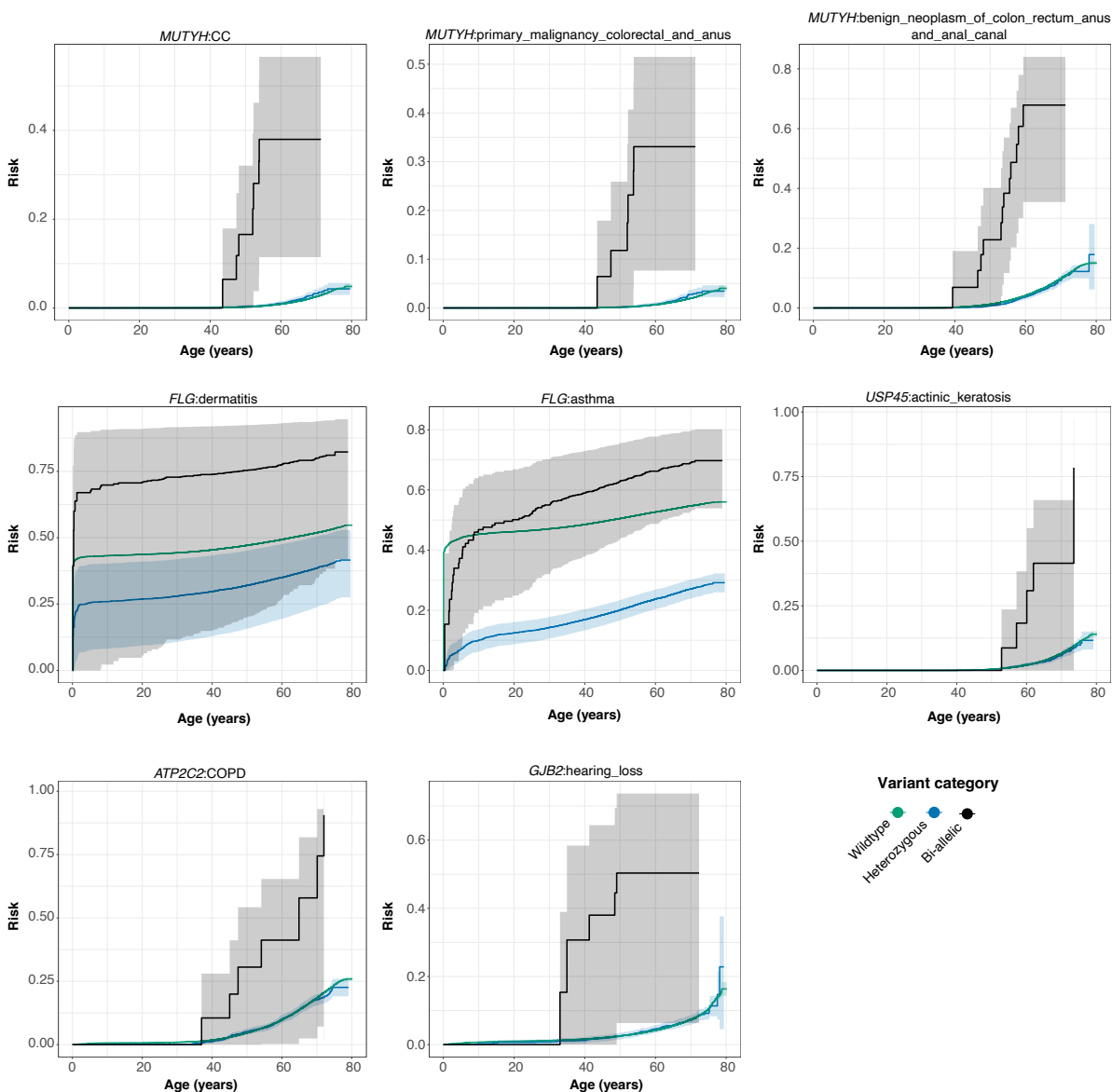
### Cox proportional-hazards modeling

For each gene-trait combination to test, we performed Cox-proportional hazards modeling to estimate differences in lifetime risk of developing the phenotype between heterozygous carriers of pLoF + damaging missense variants in the gene (reference group) and individuals who are bi-allelic carriers (compound-heterozygous or homozygous), multi-hit *cis*-heterozygous carriers, and wildtypes. All effects were adjusted for sex, the first 10 genetic PCs, birth cohort (in

ten-year intervals from 1930-1970), and UKBB assessment center. For phenotypes with a significantly heritable PRS, we additionally adjusted for off-chromosome PRS. We visualized survival probabilities using Kaplan-Meier curves<sup>74</sup>. Finally, for gene-trait combinations where we were powered to detect differences between compound-heterozygous and multi-hit *cis* heterozygous carriers of variants, i.e. where each group contained at least five cases of the phenotype, we repeated the above analysis with multi-hit *cis* heterozygous carriers as the reference group. Cox proportional-hazards regression was performed using the R package survival 3.3.1<sup>75</sup> and Kaplan-Meier plots drawn with the R package survminer 0.4.9<sup>76</sup>.

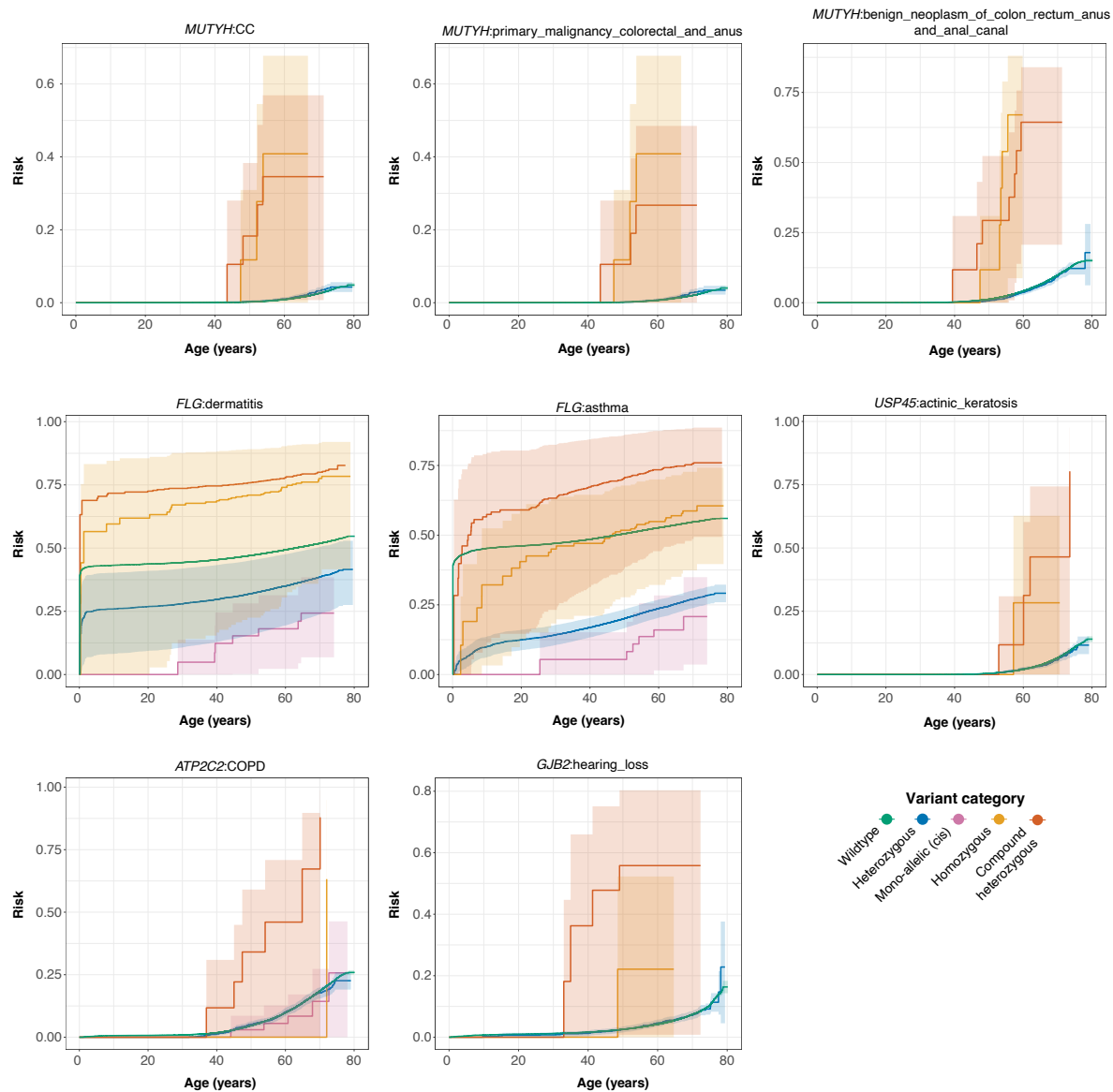


**Supplementary Fig. 22: Cox proportional hazards modeling with and without polygenic effects.** HRs when comparing CH and homozygous status versus heterozygous carrier status. Throughout, we display hazard ratios with (circles) and without (triangles) taking the polygenic contribution into account by conditioning on off-chromosome PRSs for heritable traits that pass our quality control cutoffs. HRs for gene-traits with one or more individuals with multiple *cis* variants on the same haplotype are also displayed in pink. Associations that pass Bonferroni significance ( $P < 1.89 \times 10^{-7}$ ) and FDRs  $< 5\%$  cutoff are demarcated by the dashed line in the top and bottom half respectively. Abbreviations: CC (colorectal cancer), COPD (chronic obstructive pulmonary disease).



**Supplementary Fig. 23: Kaplan-Meier survival curves for carriers of bi-allelic variants.** Trajectories for wildtypes and bi-allelic (CH or homozygous) carriers of damaging missense/protein-altering mutations are shown with green and black lines respectively. For traits where over 50% of cases are left-censored, the confidence interval estimates cannot be accurately determined using Kaplan-Meier curves, and thus, these should be disregarded. Consequently, wildtype confidence intervals for *FLG*-Asthma and *FLG*-Dermatitis are not displayed in the figure.





**Supplementary Fig. 24: Kaplan-Meier survival curves for carriers of CH, homozygous, heterozygous variants.** Kaplan-Meier survival curves for CH (red), homozygous (orange), heterozygous carriers (blue), single disruption of haplotypes (pink) owed to pLoF or damaging missense/protein-altering mutations. Wildtypes are shown in green. For traits where over 50% of cases are left-censored, the confidence interval estimates cannot be accurately determined using Kaplan-Meier curves, and thus, these should be disregarded. For this reason, wildtype confidence intervals for *FLG*-Asthma are not displayed in the figure. Wildtype and CH confidence intervals are also not shown for *FLG*-Dermatitis.

## Supplementary Tables

**Supplementary Table 4: Comparative analysis of SER point estimates for phasing Methods: SHAPEIT4, SHAPEIT5 (pre and post-filtering by phasing confidence), and Eagle2.**

This table presents a comparison of the SER point estimates for various phasing methods, including SHAPEIT4, SHAPEIT5 (both before and after filtering by phasing confidence), and Eagle2. It is important to note that we employed the full phased set of autosomes for SHAPEIT5, while for SHAPEIT4 and Eagle2, we limited our analysis to chromosomes 20, 21, and 22. The table presents 95% confidence intervals for each method.

**Supplementary Table 5: Trio-SER across final reconstituted chromosomes for variants originating from ES and genotyping array data before and after filtering by  $PP \geq 0.90$ .** All confidence intervals are 95% binomial confidence intervals.

**Supplementary Table 6: Trio-SER by MAC bin for 96 parent-offspring trio relationships before and after filtering by  $PP \geq 0.90$ .** All confidence intervals represent the 95% binomial confidence interval.

**Supplementary Table 7: Number of trio-switch errors binned by genes before and after filtering by  $PP \geq 0.90$  using 93 trio-offspring relationships.**

**Supplementary Table 8: Comparative analysis of predicted phase Using SHAPEIT5 vs. read-backed phasing with Whatshap, subsetting by  $PP \geq 0.90$ .** This comparison focuses on the predicted phase generated using SHAPEIT5 after subsetting by  $PP \geq 0.90$  against the read-backed phased variants determined through Whatshap. The comparison is limited to pairs of variants within a read. The analysis was conducted across chromosomes 20, 21, and 22, utilizing approximately 176,000 genetically ascertained NFE samples. In the 'Errors' column, errors are defined as discrepancies between the genetic phase of statistically phased variants (processed with SHAPEIT5) and read-backed phased variants (determined with Whatshap). Binomial confidence intervals are used throughout.

**Supplementary Table 9: Annotation of the most severe consequence for variants in canonical transcripts across 22 autosomes for quality-controlled variants and samples, pre- and post-filtering to  $PP \geq 0.90$ .** The table enumerates the most severe predicted consequences of variants located in canonical transcripts, analyzed across 22 autosomes for variants and samples that underwent stringent quality control, both before and after filtering to  $PP \geq 0.90$ .

**Supplementary Table 10: Results of the Poisson regression analysis used to evaluate the enrichment (rate ratios) of both mono- and bi-allelic variants within gene-sets.** We applied this model to assess both the depletion and enrichment of gene-sets using Poisson regression. The count of bi-allelic variants across samples is modeled as a function of the gene-set and mutation frequency.

**Supplementary Table 11: Tally of predicted carriers among 176k individuals classified as CH, homozygous, on the same haplotypes (*cis*), or bi-allelic (either CH and/or homozygous)**

**Supplementary Table 12: Estimation of heritability and computation of polygenic risk by applying LDSC and LDPred2 to HapMap3 SNPs to a dataset of 246k samples without phase information.** This is followed by prediction of polygenic risk for each individual within a subset of 176k phased samples. PRS that satisfy our filtering criteria (based on LDSC  $P$ -value and effective sample size) are subsequently incorporated (Methods). Evaluation of accuracy is performed through a non-parametric bootstrap method, involving calculation of the AUC and its associated standard errors.

**Supplementary Table 13: Systematic search and conditional analysis of common variants around significant gene-trait associations.** This table presents the results of our systematic search for common variants ( $MAF > 1\%$ ) located within 1 MB upstream and downstream of any significant gene-trait associations ( $P < 5.25 \times 10^{-5}$ ). Upon identifying a significant common variant ( $P < 5 \times 10^{-6}$ ), a conditional analysis was performed using that variant. If other significant variants remained, they were included in subsequent iterations of the analysis until either the signal was exhausted, or 25 iterations were completed. The table enumerates all the resulting gene-traits and variants upon which we have conditioned.

**Supplementary Table 14: Overview of significant results obtained before and after conditioning on off-chromosome PRS, nearby common variation, and the burden of rare variation.** The analysis was also carried out for compound heterozygotes and homozygotes independently while conditioning on PRS when applicable (methods). In addition, we aggregated rare variants by haplotype and modeled the number of affected haplotypes in each individual. Subsequently, we performed two analyses: (1) an additive association analysis using the haplotype encoding, and (2) a recessive association analysis, conditioned on the additive encoding of haplotypes.

**Supplementary Table 15: Overview of nominally significant ( $P < 3.05 \times 10^{-6}$ ) and Bonferroni-corrected significant hits ( $P < 9.8 \times 10^{-9}$ ) hits from additive association analysis by modeling the number of putatively disrupted haplotypes per individual.** We restrict to genes in which there are at least 10 total disrupted haplotypes in the population. Throughout the analysis, we condition on off-chromosome PRS when applicable.

**Supplementary Table 16: Significant associations ( $FDR < 5\%$ ) in Cox proportional-hazards models when comparing compound heterozygous and homozygous carriers against heterozygous carrier status.** We take the polygenic contribution into account by conditioning on off-chromosome PRS for heritable traits that pass our quality control cutoffs (Methods).

**Supplementary Table 17: Cox proportional-hazards ratios when comparing compound heterozygous and homozygous multiple variants on the same haplotype and wildtype status versus heterozygous carrier status.** We take the polygenic contribution into account by conditioning on off-chromosome PRS for heritable traits that pass our quality control cutoffs (Methods).

**Supplementary Table 18: Cox proportional-hazards ratios when comparing compound heterozygous and homozygous, heterozygous and wildtype status against carriers of multiple variants on the same haplotype.** We take the polygenic contribution into account by conditioning on off-chromosome PRS for heritable traits that pass our quality control cutoffs (Methods).

**Supplementary Table 19: Median age of diagnosis across for pLoF+damaging missense carriers that are heterozygous, homozygous, compound heterozygous, bi-allelic (homozygous and/or compound heterozygous) and have multiple variants on the same haplotype (*cis*).**

**Supplementary Table 20: Overview of variants co-occurring in significant (FDR < 5%) Cox-proportional hazards gene-trait combinations**