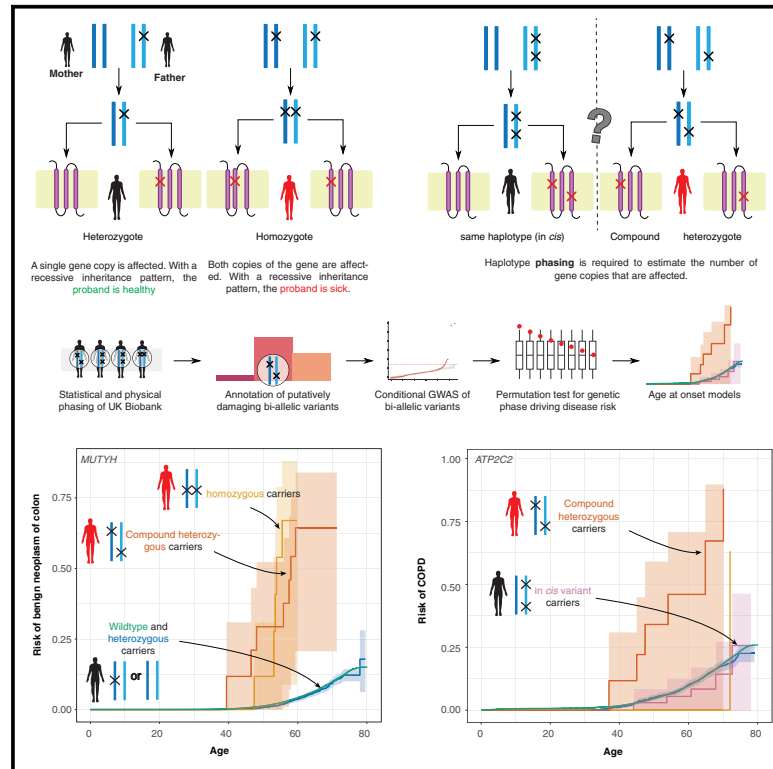# Exome-wide evidence of compound heterozygous effects across common phenotypes in the UK Biobank

Graphical abstract

Authors

Frederik H. Lassen,
Samvida S. Venkatesh, Nikolas Baya, ...,
Nicola Whiffin, Cecilia M. Lindgren,
Duncan S. Palmer

Correspondence

flassen@well.ox.ac.uk (F.H.L.),
cecilia.lindgren@wrh.ox.ac.uk (C.M.L.),
duncan.stuart.palmer@gmail.com
(D.S.P.)

In brief

Lassen et al. investigate the role of specific rare genetic variants inherited from both parents in the development of common diseases. By applying statistical methods to analyze inheritance patterns, they link such mutations to diseases ranging from asthma to cancer in the UK Biobank.

Highlights

- Biobank-scale genetic datasets can be phased and used to infer phenotypic consequences

- Incorporation of genetic phase can boost power in rare variant burden testing

- Bi-allelic variants in *ATP2C2* are associated with chronic obstructive pulmonary disease

CellPress

Article

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

Frederik H. Lassen,[1,2,*] Samvida S. Venkatesh,[1,2] Nikolas Baya,[1,2] Barney Hill,[2] Wei Zhou,[4,5,6] Alex Bloemendal,[4,7,8] Benjamin M. Neale,[4,5,6] Benedikt M. Kessler,[3] Nicola Whiffin,[1,2,4] Cecilia M. Lindgren,[1,2,9,10,*] and Duncan S. Palmer[2,4,9,10,11,*]

[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK
[2]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK
[3]Target Discovery Institute, Centre for Medicines Discovery, Nuffield Department of Medicine, University of Oxford, Oxford, UK
[4]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[5]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[6]Analytical and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA
[7]Novo Nordisk Center for Genomic Mechanisms of Disease, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[8]Data Sciences Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[9]Nuffield Department of Population Health, Medical Sciences Division, University of Oxford, Oxford, UK
[10]These authors contributed equally
[11]Lead contact
*Correspondence: flassen@well.ox.ac.uk (F.H.L.), cecilia.lindgren@wrh.ox.ac.uk (C.M.L.), duncan.stuart.palmer@gmail.com (D.S.P.)
https://doi.org/10.1016/j.xgen.2024.100602

SUMMARY

The phenotypic impact of compound heterozygous (CH) variation has not been investigated at the population scale. We phased rare variants (MAF ∼0.001%) in the UK Biobank (UKBB) exome-sequencing data to characterize recessive effects in 175,587 individuals across 311 common diseases. A total of 6.5% of individuals carry putatively damaging CH variants, 90% of which are only identifiable upon phasing rare variants (MAF < 0.38%). We identify six recessive gene-trait associations ($p < 1.68 \times 10^{-7}$) after accounting for relatedness, polygenicity, nearby common variants, and rare variant burden. Of these, just one is discovered when considering homozygosity alone. Using longitudinal health records, we additionally identify and replicate a novel association between bi-allelic variation in *ATP2C2* and an earlier age at onset of chronic obstructive pulmonary disease (COPD) ($p < 3.58 \times 10^{-8}$). Genetic phase contributes to disease risk for gene-trait pairs: *ATP2C2*-COPD ($p = 0.000238$), *FLG*-asthma ($p = 0.00205$), and *USH2A*-visual impairment ($p = 0.0084$). We demonstrate the power of phasing large-scale genetic cohorts to discover phenome-wide consequences of compound heterozygosity.

INTRODUCTION

Thousands of independent genetic variants have been robustly associated with common, complex human diseases, leading to important advancements in therapeutic development.[1] Naturally occurring variants that disrupt protein-coding sequences are of interest in the context of drug discovery as they modulate potential biological targets, with measurable effects on human physiology.[2,3] Thus, individuals who carry loss-of-function (LoF) variants on both the maternal and paternal copies of a gene are, in principle, experiments of nature, and their identification can help to determine causality between gene function and phenotype.[4–6]

Coding variants in a gene can either be homozygous, where both gene copies harbor the same variant, or compound heterozygous (CH), where both copies harbor different variants, usually at distinct genetic locations within the same gene locus. Alternatively, if two variants are located on a single gene copy, they are said to be in *cis*. Although both copies of a gene are disrupted in two-hit (CH or homozygous) carriers, analyses of the phenotypic

impact of coding variation have typically ignored genetic phase information—that is, the separation or "phasing" of an individual's genome into maternally and paternally derived alleles.[7,8] Large-scale studies of bi-allelic damaging variation have generally been restricted to homozygotes in populations with excess homozygosity, such as Icelanders,[9] Finns,[10,11] and consanguineous populations.[12] In contrast, CH is expected to be more common in outbred populations, but are understudied outside of rare disorders.[13–17] While recent efforts have resulted in the characterization of bi-allelic variation in large-scale population cohorts, researchers have so far been unable to systematically link CH variants to disease.[18,19]

A series of methods exists to infer the genetic phase of two variants. "Phasing by transmission" employs family member genotyping and Mendelian inheritance principles,[20] while "read-backed phasing" utilizes physical relationships among variants within sequencing reads.[21] In large-scale biobanks, extensively genotyping family members is impractical, and short-read sequencing technologies only allow read-backed

phasing for variants in close proximity. Therefore, "statistical phasing," which models the generative process of newly arising genetic variation subject to recombination and mutation,[20,22–25] is typically used to phase haplotypes in genetic biobank data. Obtaining high-quality statistically phased genetic data requires large sample sizes (10^5–10^6 individuals), and tends to require large reference panels.[23] Furthermore, statistical phasing is more error prone for rare variants, which are precisely the collection of variants that we would like to investigate as they are *a priori* more likely to be deleterious variants of large effect under purifying selection. This difficulty in the accurate statistical phasing of rare variation has historically deterred the analysis of CH variants in biobanks. However, recent advances in statistical phasing,[19] achieved by combining common variation across genotyping arrays and exome sequencing (ES) to create haplotype "scaffolds"[24] enables accurate phasing of rare variants. By using this new accurate phase information, which extends down into rare allele frequencies, we can identify damaging CH variants to expand the pool of identifiable two-hit carriers and screen for phenotypic consequences.

We describe and apply a systematic analytical approach to test for autosomal bi-allelic effects, gene by gene, across up to 311 traits in the UK Biobank (UKBB) 200K ES release, combining both CH and homozygous variation. We iteratively refine candidate associations by adjusting for polygenic background, nearby common variant risk, and rare variant burden within the analyzed gene. Our approach identifies both known and novel bi-allelic-trait associations that we replicate using a subset of the UKBB 450K ES release, distinct from samples in the 200K ES release.

RESULTS

Accurate phase inference and validation using parent-offspring trios and short-read sequences

We identified 13,377,336 high-quality variants in 176,935 individuals exome sequenced in the UKBB 200K release (STAR Methods). To identify variants co-occurring on the same haplotype (in *cis*) or on opposite haplotypes (in *trans*) gene by gene, we jointly phased ES and genotype array data in the UKBB using SHAPEIT5[26] (STAR Methods) following an investigation into the performance of popular phasing software (Table S1). Rare variants (minor allele frequency [MAF] <0.001) are assigned a posterior probability (PP) of true haplotype assignment, known as the phasing confidence score. Confidence in our ability to accurately statistically phase variants decreases with MAF (Figure S1). However, we *a priori* expect a disproportionate recessive damaging signal to reside in CH variants with at least one rare variant, and as a result, choosing a PP cutoff represents a trade-off in the signal-to-noise ratio. Following phasing, we restricted our study to 176,935 individuals of genetically ascertained non-Finnish European (NFE) ancestry (STAR Methods).

To assess statistical phasing quality, we benchmarked against phasing determined with parent-offspring trio data and performed read-backed phasing using short-read sequences. We first quantified phasing quality before and after filtering by PP $\geq$ 0.9 at the genotype level in 96 parent-offspring trios by calculating switch error rates (SERs), estimated using Mendelian transmission, across 2,044,234 unique variants stratified by minor allele count (MAC) (Figures 1A and S2; Tables S2 and S3). Across the 96 children, 93.1% of the protein coding genes contained variants that were phased without switch errors (Table S4). In SHAPEIT5, singletons are phased by identifying the longest shared identity by descent (IBD) segment between the two haplotypes in the target individual and those in the population. The minor allele is then assigned to the haplotype with the shortest shared IBD segment, as a short IBD segment indicates an older common ancestor providing more time for a germline variant to occur. Consistent with previously reported estimates in ES and genotyping array data,[19] we observe a singleton SER of 32.4% (95% confidence interval [CI] = 30.2%–45.6%) prior to filtering by PP. Upon restricting to PP $\geq$ 0.9, SERs among singletons (MAC = 1) reduce to 12.1% (95% CI = 8.42–17.2). By comparison, rare variants with 2 $\leq$ MAC $\leq$ 5 exhibit SERs of 0.27% (95% CI 0.13–0.53) after applying the PP filter (Figure 1A).

Although the calculation of SER using trios is the gold standard approach for phasing quality estimation,[23] it is limited by the number of parent-offspring trios available. For this reason, we also performed read-backed phasing of 453,514 unique pairs of variants using UKBB short-read sequences on autosomes in 49,756 NFE individuals using WhatsHap[27] (STAR Methods). While read-backed phasing only permits ascertainment of genetic phase among pairs of variants spanning one or a few overlapping short-read sequences (with typical lengths of up to 250 bp), read-backed phasing accuracy is independent of allele frequency, and therefore represents an orthogonal approach to evaluating the quality of statistically phased variation. Consistent with trio-SER, we observed increasing agreement between pairs of statistically and read-backed phased variants with increasing MAC (Figure S3; Table S5). For singletons, we observed that with increased PP, phasing accuracy increased, while the total proportion of retained singletons decreased (Figure S4). Filtering to phased variants with PP $\geq$ 0.9 at the genotype level, singletons and variants with 2 $\leq$ MAC $\leq$ 5, agreement between read-backed phasing, and statistical phasing was 86.34% (95% CI = 85.68%–86.98%) and 99.37% (95% CI = 99.33%–99.41%), respectively (Figure S5; Table S5).

Taken together, our benchmarking suggests that statistical phasing of the UKBB dataset is of high quality for rare to ultra-rare variants and even singletons, increasing our confidence in the identification of damaging CH variation. Given our observations of well-calibrated PP and the distribution of phasing confidence binned by MAC, we selected the empirical cutoff of PP $\geq$ 0.9 at the genotype level to retain 6,352,396 variants (12% of which are singletons) for downstream characterization and testing (Table S2).

Identification and examination of CH variation in the UKBB

To interrogate the functional role of mono- and bi-allelic variation in the population, we annotated 6,352,396 variants (762,843 of which were singletons) with PP $\geq$ 0.9 and MAF $\leq$ 5% across 17,998 autosomal protein-coding genes. We enriched our search for variants with putatively large effect sizes by restricting analyses to two categories of predicted damaging variation. First, we annotated 146,299 (22,380 singletons) high-confidence protein truncating variants, including stop-gain, essential splice,
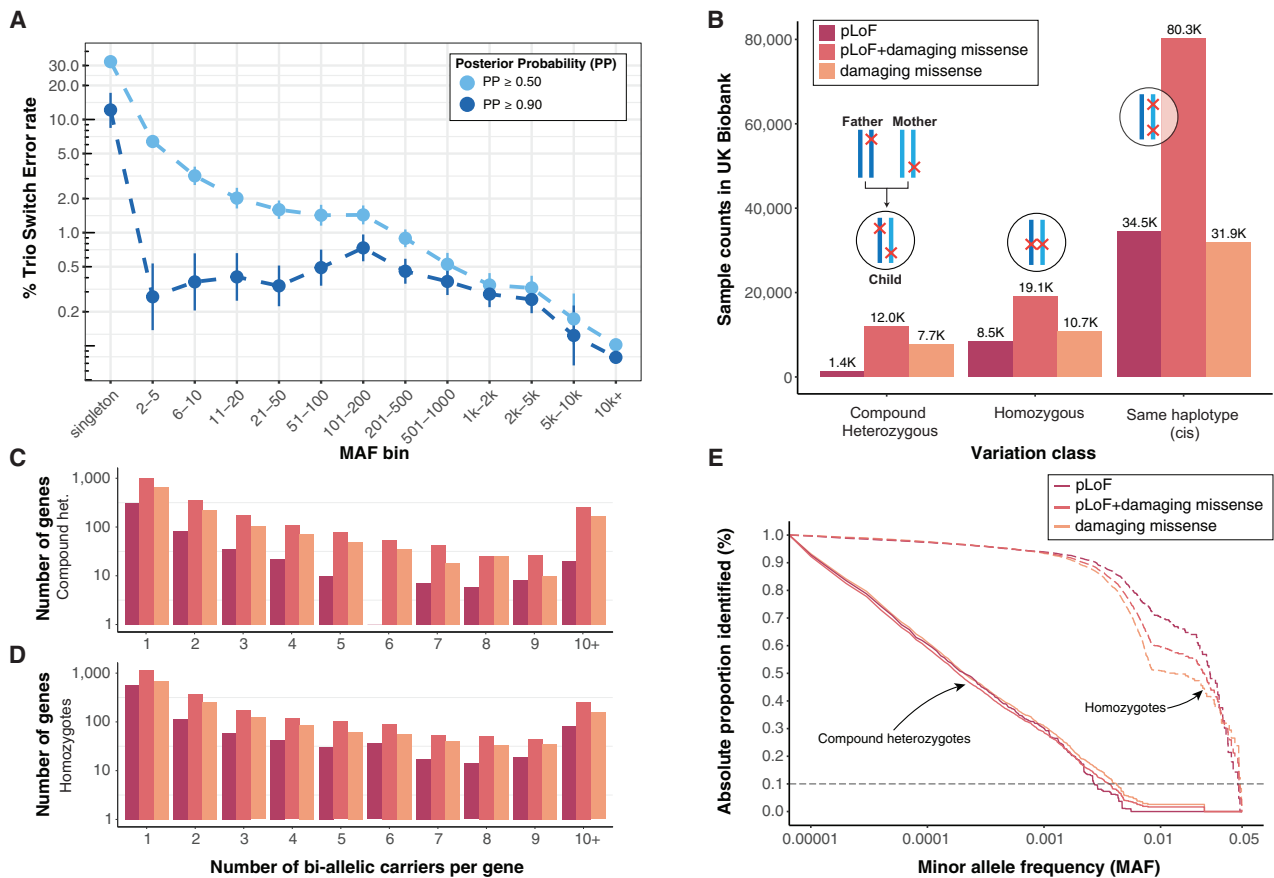
Figure 1. CH variants composed of at least 1 ultra-rare variant (MAC ≤ 10) can be robustly identified in large-scale biobanks

(A) Trio SER depicted on the y axis as a function of MAC bin (x axis) for phased variants with MAF ≤ 5%, stratified by phasing confidence score PP ≥ 0.5 or PP ≥ 0.9. Error bars display 95% binomial confidence intervals.

(B) Counts of samples harboring different classes of variation with at least 2 variants in UKBB. Each set of 3 bars depicts the number of individuals with at least 1 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 and D) Number of CH or homozygous carriers per gene.

(E) One minus 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 ≥ 0.9 (solid line), stratified by variant consequence according to the color key.

and frameshift variants identified as high confidence by LoF transcript effect estimator (LOFTEE),[28] which we refer to as putative LoF (pLoF) variants. Second, we annotated 242,859 (32,199 singletons) missense variants classified as damaging by both Rare Exome Variant Ensemble Learner (REVEL) score ≥0.6 and Phred scaled Combined Annotation Dependent Depletion (CADD) score ≥20, or LOFTEE low confidence protein-truncating variants; we refer to these variants collectively as damaging missense/protein altering (Figure S6; Table S6). For each individual, we then determined the set of genes predicted to be affected by pLoFs + damaging missense/protein-altering variants in a CH, homozygous, or in *cis* state on the same haplotype.

As we *a priori* expected that essential genes would be less permissible to bi-allelic damaging variants when compared to non-essential genes, we investigated tolerance toward predicted bi-allelic pLoF and pLoF + damaging missense/protein-altering

variants across the genome. As some genes carry bi-allelic variants more often than others (owing to a variety of factors such as gene length and baseline mutation frequency[29]), we fit counts of the number of individuals carrying bi-allelic variants per gene using a Poisson regression model, accounting for variation in gene length and mutation rate (STAR Methods; Tables S7 and S8). Both pLoF and pLoF + damaging missense/protein-altering bi-allelic variants (homozygous and CH) were significantly depleted in five of the six analyzed essential gene sets ($p < \frac{0.05}{6} \approx 0.0083$) (Figure S7). Conversely, across three non-essential gene-sets, bi-allelic pLoFs + damaging missense/protein-altering variants were enriched among LoF-tolerant genes[29] ($p \le \frac{0.05}{3} \approx 0.0167$). We found that the degree and direction of effects were consistent across CH, homozygous bi-allelic, and heterozygous variants (Figure S7) and concordant with previous study on CH depletion using the same gene set definitions.[19] Additional gene set enrichment analyses are provided in Table S7.

In founder[9] and bottle-necked[10] populations, some alleles are enriched to high frequency by chance, resulting in better-powered association studies for the subset of variant alleles that are inherited from the parental population at higher frequency.[6] To explore the diversity of bi-allelic variation in UKBB, a largely outbred population, we enumerated two-hit carriers across 176,587 individuals (Table S8). We observed complete bi-allelic knockout of 1,174 unique genes strictly due to pLoF variants, identifying 1,431 (0.8%) CH and 8,582 (4.8%) homozygous individuals with bi-allelic pLoF variants in at least one gene (Figure 1B). Across genes, 307 (26.1%) CH and 560 (47.7%) homozygous knockouts were observed only in a single individual (Figures 1C and 1D). Our simulations (STAR Methods) closely recapitulated the empirically observed patterns of bi-allelic pLoF events, with strong correlation between simulated and observed homozygous ($R = 0.996$, $p < 2.2 \times 10^{-16}$) and CH ($R = 0.932$, $p < 2.2 \times 10^{-16}$) pLoF events (Figure S8). We reasoned that the inclusion of damaging missense/protein-altering variants in addition to pLoFs would expand the number of identifiable damaging bi-allelic variants compared to assessing the two categories independently. Across 3,288 unique genes, we observed 11,491 (6.5%) CH and 17,863 (10.1%) homozygous carriers of pLoF + damaging missense/protein-altering variants. Of these, 1,112 (0.6%) CH and 436 (0.2%) homozygotes were carriers of bi-allelic pLoF + damaging missense/protein-altering variants in genes linked to traits with an autosomal recessive mode of inheritance in OMIM.[30] Consistent with previous observations,[18] we generally observed a higher prevalence of carriers with variants in *cis* compared to CH, with over one-third of individuals (64,555, 36.6%) carrying $\ngeqq 2$ pLoF + damaging missense/protein-altering variants on a single haplotype (Figure 1B).

To better understand the evolutionary dynamics giving rise to pathogenic variants in *trans*, we examined the spectrum of allele frequencies of the constituent variants among our confidently called damaging CH variants. CHs variants tend to comprise two variants where one resides on a common haplotype, while the other resides on a rare haplotype, with a median difference in MAC of 1,181 (Figures S9 and S10). Approximately 90% of CH-constituent variants have MAF ≤ 0.0038, compared to homozygotes in which 90% are detected at MAF ≥ 0.0022 (Figure 1E), suggesting that identifying deleterious bi-allelic CH variants requires reliable phasing of rare alleles (Figures S11 and S12).

Multiple studies have assessed the prospects of ascertaining bi-allelic LoF variation at larger sample sizes in consanguineous, bottle-necked, and outbred populations.[6,12] To investigate empirically how the number of unique genes with bi-allelic variants scales in an outbred population, we performed down-sampling of UKBB participants. Consistent with the previous literature, additional CH and homozygous variants can be inferred by considering both pLoF and damaging missense/protein-altering variation at even larger sample sizes (Figure S13).

Systematic evaluation of bi-allelic effects on common disease

We performed a series of association analyses using the scalable and accurate implementation of generalized mixed model (SAIGE),[31] a generalized mixed-model association testing framework that uses a saddle point approximation to provide accurate p values for traits with extreme case-control ratio imbalance. This allowed us to investigate the effects of bi-allelic variants in 176,587 individuals across 311 phenotypes with varying population prevalence identified from secondary care electronic health records (STAR Methods). We restricted to 952 protein-coding genes with at least 5 individuals carrying bi-allelic variants in the same gene, which allowed us to detect odds ratios (ORs) ≥ 10, for traits at approximately 2% population prevalence, with 80% power at exome-wide significance (Bonferroni $p < \frac{0.05}{952 \times 311} \approx 1.68 \times 10^{-7}$) (STAR Methods; Figure S14). Using simulations, we confirmed our ability to detect recessive signals of association with well-calibrated false positive rates across a range of effect sizes (STAR Methods; Figures S15A–S15C). We tested a total of 299,854 gene-trait combinations and identified 7 gene-trait associations following stringent Bonferroni correction ($p < 1.68 \times 10^{-7}$) (Figure S16; Table S9). We observed that excluding singletons had a minimal impact on the resulting gene-trait associations (Figure S17), likely due to the limited number of CH variants comprising at least one singleton (Figure S18). Finally, we performed confirmatory replication analysis in the remainder of the UKBB (STAR Methods) and found that 7 of 7 (100%) gene-trait associations with at least 5 bi-allelic carriers replicated at $p < 0.05$ (Table S10).

A recessive gene-trait association may be influenced by a variety of genetic factors unrelated to CH or homozygous status, such as polygenic background or through genetic tagging of a nearby common variant association. To mitigate these factors, we created a pipeline to condition on external genetic effects within the gene-trait regression model (STAR Methods). We trained polygenic risk scores (PRSs) for 111 significantly heritable traits (h_{snp}^2 $p < 0.05$ and $n_{eff} \ngeqq 5,000$) using LDPred2[32] (STAR Methods; Table S11), a tool that allows PRS derivation based on summary statistics and linkage information. To control for polygenic risk and potentially boost power for association,[33] we included the off-chromosome PRS as an additional covariate (Table S9). We observed that controlling for PRS had no significant influence on the binary case-control association testing, and the resulting p values were altered by less than a single order of magnitude with the incorporation of PRS (Figure S19). To capture the effects of any causal common variants in linkage disequilibrium with the pLoF or damaging missense/protein-altering variants constituting the CH or homozygous variant, we further conditioned on nearby (within 1 Mb of the associated gene) common (MAF ≥ 1) variant association signals (STAR Methods; Table S12), which abrogated ($p > 0.05$) the signal of a single gene-trait pair.

Lastly, we investigated whether any of the identified putative recessive associations could be accounted for by assuming an additive genetic architecture. To do this, we counted the number of gene copies affected by pLoF + damaging missense/protein altering variants in each individual. For each putative recessive association, we re-ran the analysis while simultaneously conditioning on the number of affected haplotypes. We also employed a complementary variant-level approach and repeated the analysis, conditioned on all low-frequency (MAC >10, MAF ≤ 5) and ultra-rare (MAC ≤ 10) damaging variants (pLoF + damaging
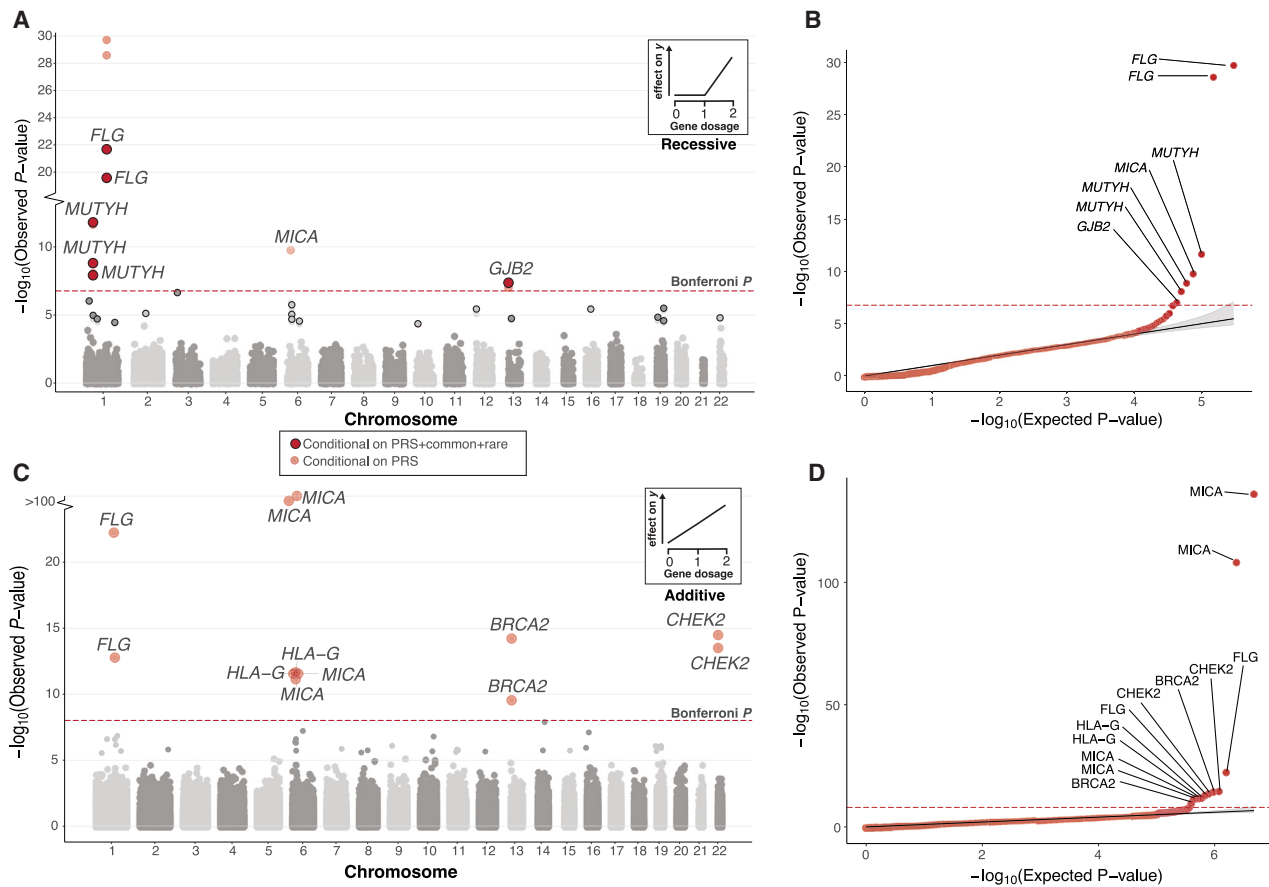
Figure 2. Conditional recessive and additive modeling of gene copy disruption in 311 phenotypes across 176,587 participants

(A) Recessive Manhattan plot depicting $\log_{10}$-transformed gene-trait association p values against chromosomal location. Associations are colored red if they are Bonferroni ($p < 1.68 \times 10^{-7}$) 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 (STAR Methods). The Bonferroni significance threshold is also displayed as a red dashed line. A gene may appear multiple times if it is associated with >1 phenotype. A qualifying example of the recessive inheritance pattern is shown at the top right of the panel: disruption of both gene copies results in an effect on the phenotype.

(B) Quantile-quantile (Q-Q) 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 and D) Additive Manhattan plot and corresponding Q-Q plot for genes with mono- and bi-allelic damaging variants associated with at least 1 phenotype after conditioning on off-chromosome PRS when applicable (STAR Methods). No additional conditioning was performed in this analysis. Gene-trait associations are colored red if they are Bonferroni ($p < 9.8 \times 10^{-9}$) significant. The additive inheritance model is depicted at the top right of the panel; each affected haplotype results in a incremental effect on the phenotype.

missense/protein-altering), including those that constitute the bi-allelic variant in question. Among the remaining six Bonferroni significant associations, none of the associations were abrogated after conditioning on additive effects (Table S9).

Together, these analyses refined the list of putative gene-trait associations to six Bonferroni associations after stringent Bonferroni correction and conditioning (Figures 2A and 2B; Table S9) comprising three unique genes and six traits. Notably, only three of the six associations remained significant (Bonferroni $p < 1.68 \times 10^{-7}$) when restricted to only CH variant carriers, and just one of six when testing homozygous variants alone, underscoring the power of jointly analyzing these variant sets (Figure S20).

We observed recessive gene-trait relationships across multiple physiological domains (respiratory, neoplasm, skin, ear, and mas-

toid). All six associations that met the significance threshold after Bonferroni correction and conditioning ($p < 1.68 \times 10^{-7}$) have been reported previously in the literature. For example, individuals with bi-allelic variants in *MUTYH*, a gene involved in oxidative DNA damage repair,[34] are at significantly increased risk of developing colorectal cancer ($\log_{10}(OR) = 4.7$ [95% CI = 3.38–6.01], $p = 2.2 \times 10^{-12}$). We also find that bi-allelic variants in *FLG* increase the risk of both asthma[35] ($\log_{10}(OR) = 3.3$ [0.26–0.39], $p = 2.09 \times 10^{-22}$) and dermatitis[36] ($\log_{10}(OR) = 0.28$ [0.22–0.33], $p = 2.65 \times 10^{-20}$). In addition, we observe that bi-allelic variants in *GJB2* increase the risk of hearing loss[30] ($\log_{10}(OR) = 1.66$ [1.05–2.26], $p = 9.93 \times 10^{-8}$).

To assess the degree to which compound heterozygosity, rather than co-occurring variants on the same haplotype, drives
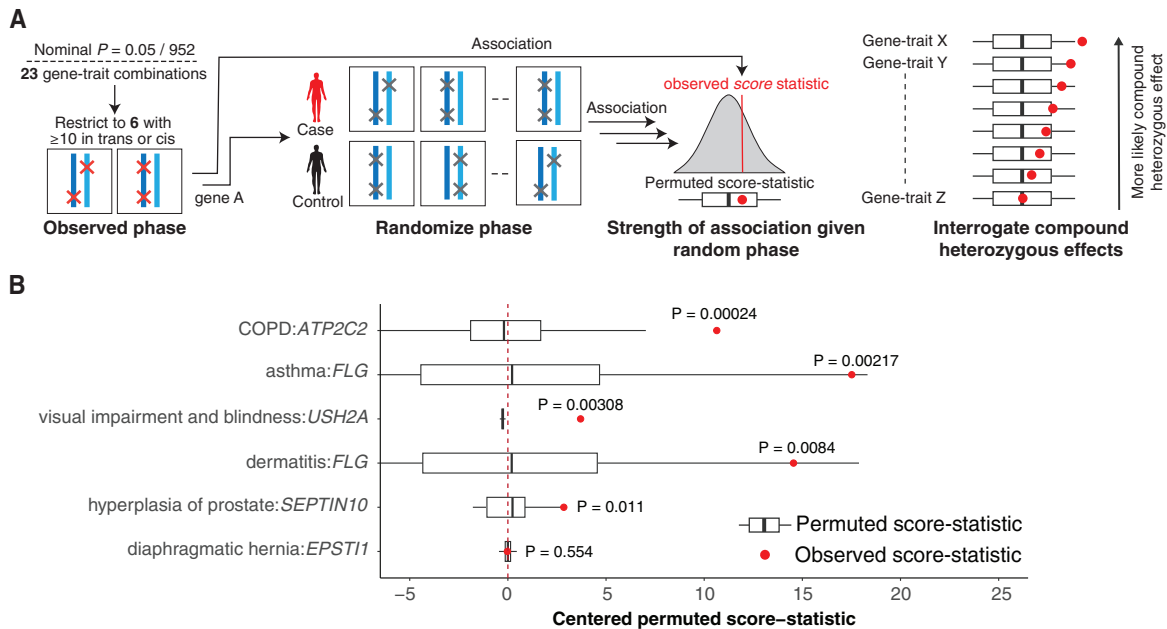
Figure 3. *In silico* permutation of genetic phase provides evidence for CH-specific effects

(A) Overview of the permutation pipeline. To be sufficiently powered to detect effects, we considered 5 significant ($p < 0.01$) gene-trait pairs from the genome-wide analysis that have at least 10 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 score statistics corresponding to the association strength in the absence of phase information. We derive the 1-tailed empirical p value by comparing the observed score statistics with the empirical null distribution.

(B) The resulting distributions of permuted (white and black boxplots) and observed score statistic (red dot) for each gene-trait and the resulting empirical p value. p values shown in bold indicate Bonferroni significance (($p < 0.05/06 = 0.0083$). Box and whisker plots display the quartiles of the empirical null distribution.

disease risk, we permuted the genetic phase of observed pLoF + damaging missense/protein-altering variants within a gene to generate an empirical distribution of score statistics corresponding to disease-association strength in the absence of phase information (Figures 3A and 3B). To ensure a sufficiently large sampling distribution, we restricted our analysis to six nominally significant ($p < \frac{0.05}{952} \approx 5.25 \times 10^{-5}$) gene-trait combinations with at least 10 individuals who are either CH variant carriers or with ≥ 2 pLoF or damaging missense/protein-altering variants on the same haplotype (STAR Methods). We found evidence that the incorporation of CH variation significantly (Bonferroni $p = 0.05/6 = 0.0083$) increased the score statistic in three of the six analyzed gene-trait combinations. CH variants in *ATP2C2* are associated with an increased risk of chronic obstructive pulmonary disease (COPD) ($p = 0.000238$), and CH variants in *FLG* are associated with an increased risk of asthma ($p = 0.002174$), while CH variants in *USH2A* are associated with an increased risk of visual impairment and blindness ($p = 0.000307$) (Figure 3B). We identified two additional gene-trait association at nominal significance ($p < 0.05$): CH variants in *FLG* that are associated with an increased risk of dermatitis ($p = 0.0092$), and CH variants in *SEPTIN10* that are associated with hyperplasia of prostate ($p = 0.011$). Of these, *FLG*-asthma, *FLG*-dermatitis, and *USH2A*-visual impairment associations have previously been linked to disease in the CH state.[37–39] These observations demonstrate on a large scale the effect of compound heterozygosity in driving disease susceptibility and, by extension, how appropriately integrating genetic phase can lead to increased power to discover gene-trait associations.

Incorporation of genetic phase boosts power in gene-level rare variant burden models

Complementary to the recessive models described above, rare variant burden testing, which involves the aggregation of rare variants within a gene, has proven to be a robust method to collectively assess the phenotypic impact of rare variation across individuals. Rare variants are aggregated due to their low allele frequency leading to lack of statistical power for the detection of single-marker associations. However, these frameworks generally ignore the genetic phase within each individual, and therefore do not differentiate between scenarios in which multiple damaging variants reside on the same (in *cis*) or opposite (in *trans*) haplotypes, despite these two forms having potentially distinct functional and phenotypic effects. We conducted additive genome-wide association analyses by testing for associations between the number of disrupted gene copies (across 16,363 protein-coding genes with at least 10 haplotypes carrying pLoF + damaging missense/protein altering variation) on an individual and case status (across 311 phenotypes) (Figures 2C and 2D; STAR Methods). After adjusting for polygenic contribution, we found 12 significant associations after stringent multiple-testing correction (Bonferroni $p < \frac{0.05}{16,363 \times 311} \approx 9.8 \times 10^{-9}$; Table S13). Among the significant hits are previously reported associations, including association between the number of

putatively damaged copies of *BRCA2* ($p = 6.16 \times 10^{-15}$) and *CHEK2* ($p = 3.34 \times 10^{-15}$) and breast cancer. We then compared our approach to an established gene-based method for rare variant burden testing, specifically SAIGE-GENE+.[40] We applied SAIGE-GENE+ to the same variant set used in our study, comparing the *p* values derived from this method with those from our haplotype-collapsing approach (Table S14). This initial comparison revealed 5 of 12 (42%) associations with insufficient evidence to reach Bonferroni significance ($p < 9.8 \times 10^{-9}$) using SAIGE-GENE+ burden testing. Moreover, among our significant additive haplotype-based gene associations, 11 of 12 (92%) were more strongly associated than in SAIGE-GENE+ burden testing (one-sided sign test, $p = 0.00317$). Recognizing that our initial analysis deviated from the regular usage of SAIGE-GENE+, which ignores phasing confidence, we also performed an analysis without filtering variants by PP. This allowed us to increase the pool of variants and compare the performance of the haplotype-collapsing method with the standard application of the burden test. We observed that while SAIGE-GENE+ now recapitulated all 12 associations at the Bonferroni cutoff ($p < 9.8 \times 10^{-9}$), we found that 12 of 12 (100%) gene-trait associations were more strongly associated using the haplotype-based collapsing method (one-sided sign test, $p = 0.00024$).

Non-additive effects of CH variants elevate lifetime risk of disease

Bi-allelic effects may be associated with earlier age at onset of disease, which is also often correlated with disease severity. We therefore explored whether CH and homozygous variants had longitudinal effects by evaluating age at diagnosis of 278 phenotypes with Cox proportional hazards models. To identify effects due to the disruption of both gene copies, as opposed to haploinsufficiency, we compared bi-allelic variant carriers against a reference group comprising carriers of a single heterozygous variant for each gene. We tested 267,400 gene-trait combinations with at least 5 bi-allelic variants (homozygotes or CH) and 100 heterozygotes (Figure 4A). After adjusting for polygenic risk via off-chromosome PRS, we identified seven gene-trait associations with significantly earlier age at diagnosis in bi-allelic variants compared to heterozygous carriers of pLoF + damaging missense/protein-altering variants (Bonferroni $p < \frac{0.05}{952 \times 278} \approx 1.89 \times 10^{-7}$; Figures 4B, 4C, and S21–S23; Tables S15 and S16). For six out of the seven Bonferroni significant gene-trait combinations, we found no evidence ($p > 0.05/7 \approx 0.00833$) that carrying a single heterozygous variant altered lifetime disease risk compared to carrying two copies of the reference allele. We also performed confirmatory replication analyses for 7 associations with at least 5 bi-allelic and 100 heterozygous carriers in the replication cohort (STAR Methods). We found that all 7 (100%) Bonferroni significant gene-trait associations replicated at ($p < 0.05$), including *ATP2C2*-COPD (replication $p = 0.013$), which has not previously been reported in the literature (Table S17).

We further sought to disentangle the effects of homozygous and CH variants on lifetime disease risk from that attributable to multiple damaging rare variant effects on a single haplotype. To do this, we analyzed these effects in the three gene-trait pairs with both (1) at least 5 CH and/or homozygous variants and (2) at least 5 individuals harboring $\ngeqq 2$ variants on the same haplotype

(Figure 4C; Table S18). Compared to individuals with a single disrupted haplotype, both homozygous and CH carriers of pLoF + damaging missense/protein-altering variants in *ATP2C2* were at an increased lifetime risk of developing COPD (homozygote hazard ratio [HR] = 6.65 [95% CI = 4.5–8.8], $p = 0.084$, CH HR = 8.98 [95% CI = 7.79–10.17]; $p = 0.00028$). Similarly, both homozygous and CH variants of *FLG* were at an increased lifetime risk of asthma (homozygote HR = 1.97 [95% CI = 1.1–2.84], $p = 0.126$, CH HR = 2.44 [95% CI = 1.61–3.26], $p = 0.033$) and dermatitis (homozygote HR = 1.7 [95% CI = 0.88–2.5], $p = 0.20$, CH HR = 1.16 [95% CI = 0.38–1.94], $p = 0.7$) (Figure 4C). For these gene-trait relationships, mono-allelic carriers have no increase in risk of disease compared to wild types, whereas bi-allelic carriers have a significant increase in lifetime risk of disease.

Biological insights into common complex disorders implicated by CH variation

Six of the seven gene-trait combinations for which we identify Bonferroni significant associations with lifetime disease risk are also significant in our cross-sectional recessive association analysis (Table S19). The six have previously been reported in the literature, albeit without age at onset effects. These include *MUTYH* and colorectal cancer, *GJB2* and hearing loss, and a pleiotropic association of *FLG* with both dermatitis and asthma (Figures 5A, 5B,[41] and S22). We find that *ATP2C2*-COPD is a novel candidate association ($p = 3.58 \times 10^{-8}$) with plausible mechanistic effects. All constituent variants and allele frequencies for Bonferroni significant associations are provided (Table S20).

ATP2C2, a calcium-transporting ATPase linked to surfactant protein D levels via an intronic variant (rs9927461, a causal risk factor for COPD),[42] is associated with COPD in our gene-trait analyses (HR = 8.3 [95% CI = 7.54–9.05], ($p < 3.56 \times 10^{-8}$). As we did not observe any nearby (1 Mb upstream or downstream) common variants in *ATP2C2* associated with cross-sectional COPD (all $p > 5 \times 10^{-6}$), the association between bi-allelic variants of *ATP2C2* and COPD potentially is driven by the unique configurations of CH damaging missense ($n = 7$) and pLoF ($n = 1$) variants that primarily reside in functional protein domains (Figures 5E and S24; Table S21). Seven of the eight (87.5%) individuals we identified as bi-allelic carriers of damaging variation in *ATP2C2* (6 CH and 2 homozygous) were diagnosed with COPD (median age of diagnosis = 54.1 [interquartile range = 46.2–67.5] years) (Figures 5C and 5D). In contrast, only 6.9% of individuals harboring multiple pLoF + damaging missense/protein-altering variants on the same *ATP2C2* haplotype were diagnosed with COPD, with the same median age of diagnosis (60.8 [53.7–67.9] years) as both heterozygote carriers (58.0 [48.5–64.1] years) and those with wild type (59.2 [51.3–65.1] years).

FLG plays a pivotal role in the differentiation and maintenance of skin barriers.[35] *FLG* variants have been selectively associated with individuals with both asthma and atopic dermatitis, but not with those who have asthma without atopic dermatitis.[36] Our findings indicate that individuals carrying a single deleterious *FLG* allele face increased risk of dermatitis (HR = 1.11 [95% CI = 1.05–1.19], $p \approx 7.2 \times 10^{-5}$), but not asthma (HR = 1.06 [95% CI = 1.01–0.12], $p = 0.018$), when compared to wild types.
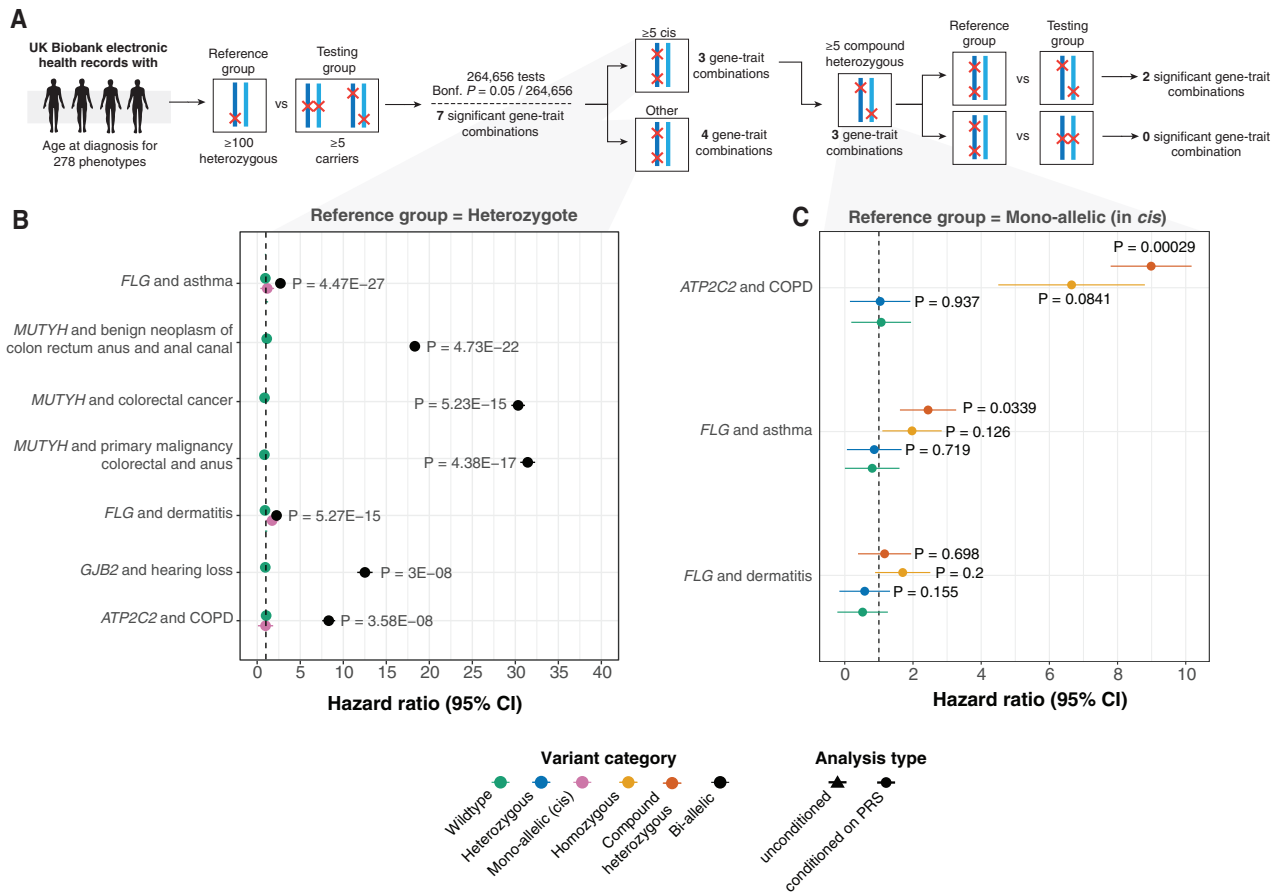
Figure 4. Age at diagnosis modeling reveals novel recessive effects driven by damaging bi-allelic variants

(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 performed Cox proportional hazards modeling for gene-trait combinations in which ≥5 samples are 2-hit carriers (CH or homozygotes) and ≥100 samples that are heterozygotes. Among Bonferroni (Bonf.) significant associations ($p < 1.89 \times 10^{-7}$), we filter to gene-trait pairs for which at least 5 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 after 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 PRSs are predictive. HRs for gene-traits with ≥2 individuals with multiple *cis* variants on the same haplotype are displayed in pink. Only associations that pass the stringent Bonferroni significance threshold ($p < 1.89 \times 10^{-7}$) cutoff are illustrated.

(C) HRs when comparing wild-type, heterozygous, CH, and homozygous carrier status against individuals that harbor ≥2 putatively damaging variants on the same haplotype. 95% CIs are shown in the figure.

In contrast, individuals carrying two variant alleles have an increased risk of developing both dermatitis (HR = 2.23 [95% CI = 2.03–2.43], $p = 5.27 \times 10^{-15}$) and asthma (HR = 2.68 [95% CI = 2.51–2.87], $p = 4.47 \times 10^{-27}$), suggesting a recessive mode of inheritance for *FLG*-related asthma and a semi-dominant inheritance mode for *FLG*-related dermatitis.[30] This implies that the loss of a single *FLG* copy can result in dermatitis, while the loss of both copies can lead to asthma. Together, this may help clarify why *FLG*-related asthma is seldom observed without the presence of *FLG*-related dermatitis.

DISCUSSION

In this large biobank-scale effort, we systematically interrogated the role of bi-allelic coding variants in genes conferring risk for

common complex diseases. In the cross-sectional and longitudinal recessive analysis, we identified six and seven significant gene-trait associations after stringent Bonferroni correction, respectively. In the cross-sectional analysis, only 2 of the 6 (33%) associations we identified would have been discovered considering homozygotes alone, whereas in the longitudinal analysis 5 of the 7 (71%) associations would have been discovered without considering compound heterozygosity. Collectively, the associations both replicate established relationships in the literature and identify a previously unreported gene-trait association for binary phenotypes across the common disease spectrum.

We show that the 90% of deleterious CH variants occur at MAF < 0.38%. Given that phasing quality is directly correlated with allele frequency, it is essential to filter to the set of variants phased at high confidence to eliminate false positive identifications.
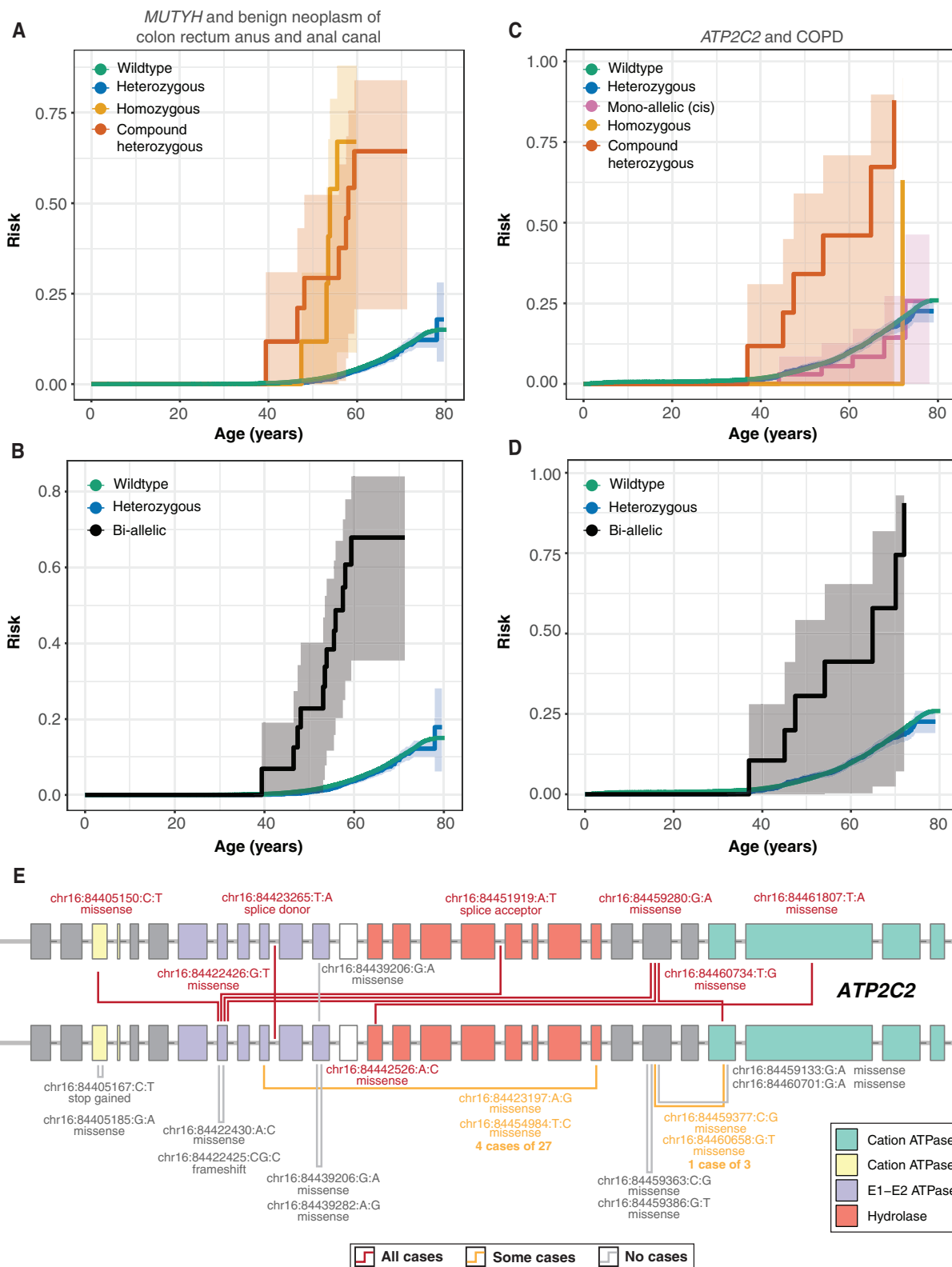
Figure 5. Trajectories of haplotype disruption in common diseases

(A and B) Kaplan-Meier survival curves for CH (red), homozygous (orange), heterozygous carriers (blue), and single disruption of haplotypes (pink) due to pLoF or damaging missense/protein-altering mutations. Shaded regions indicate 95% confidence intervals for risk estimates. Wild types and bi-allelic variants (CH or

(legend continued on next page)

Here, we quantified the increase in phasing quality using Mendelian inheritance logic in parent-offspring relationships and compared pairs of statistically phased variants to read-backed phased variants using short-read sequences. While read-backed phasing is computationally expensive and restricted to variants in close proximity, we demonstrate that it can be employed to evaluate statistical phasing quality in cohorts that lack trio relationships, with error rates comparable to those of trio SERs.

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

Many common complex traits have polygenic architectures, which should be accounted for when performing gene-trait association testing. The presence of bi-allelic variants in individuals with such diseases might be coincidental and not causally related to the trait, which may instead be a result of a high polygenic risk. However, across the significant recessive genome-wide associations, we observed that the inclusion of PRS as a covariate affected the resulting association p value by less than an order of magnitude for the binary traits we analyzed. While we were only able to account for the polygenic contribution to disease development for 111 diseases with significant common variant heritability in the UKBB, due to low case numbers, these observations suggest that the incorporation of polygenic background has limited influence on the degree of association when evaluating ultra-rare variation across binary traits.

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

The likelihood of damaging alleles occurring on the same haplotype is influenced by a complex interplay of factors, including population structure and balance between selection, drift, mutation, and recombination. We and others[18] have found that damaging CH variants occur less frequently than multiple damaging variants affecting the same haplotype, suggesting that in certain circumstances, natural selection operates on a haplotype level. Once an LoF variant occurs and expands in the population, the affected haplotype has no selection against additional acquisition of damaging mutations. This has implications for association studies investigating CH effects by counting the number of damaging variants in a gene while attributing equal probability to each of affecting each haplotype,[8] as such frameworks may overestimate the frequency of CH events.

"Human knockouts" have been extensively discussed in the context of therapeutic development. Examining both bi- and mono-allelic carriers can help one assess the safety of therapeutic interventions by analyzing how varying degrees of target modulation affect biological response.[3,6] We showcased several gene-trait relationships where the number of affected haplotypes influences the lifetime risk of disease, potentially representing the manifestation of adverse events, which are important endpoints in clinical trials. The absence of adverse events in

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 wild types.

(C and D) Kaplan-Meier survival curves for *ATP2C2* mono- and bi-allelic variant carriers. Carriers of CH variants develop COPD earlier compared to heterozygotes carriers and wild types. Moreover, individuals who harbor a single putatively disrupted haplotype due to ≥2 damaging variants develop COPD at the same frequency as heterozygotes and wild types.

(E) Gene plots for *ATP2C2*, displaying protein coding variants for samples that carry ≥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 the 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, and red lines indicating 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.[41]

mono-allelic carriers can imply that partial pharmacological inhibition of a target may be a safe and effective approach. However, adverse effects observed in bi-allelic carriers of damaging variation within the same locus could indicate potential risks associated with complete target inhibition. A natural extension of this work could involve investigating mono- and bi-allelic effects on quantitative outcomes, such as serum proteins. Changes in biomarkers (or other continuous outcomes) may reflect direct or indirect consequences of gene modulation and could serve as potential pharmacodynamic biomarkers commonly used to assess target engagement in clinical trials.

This work showcases the value of statistical phasing of damaging rare variants and that association analyses that account for compound heterozygosity can be better powered for gene-trait discovery. We showed that this approach can be employed to discover well-established and novel non-additive and additive gene-trait relationships across a wide range of disease etiologies. From a clinical perspective, we demonstrated the importance of interrogating the genetic phase when dealing with CH variants in traits with recessive modes of inheritance. This is an important step toward uncovering the phenome-wide consequences of bi-allelic disruption across the human genome.

Limitations of the study

We acknowledge several limitations to our study. First, the accuracy of statistical phasing decreases with the MAF. Our study encompasses ultra-rare variation (MAF < 0.001%), which we filtered to confidently phased heterozygous variants (PP < 0.9). This criterion excludes variants that could form CH relationships, potentially leading to an underestimation of CH events, especially among those that constitute rare variants and singletons. Second, we applied a conservative threshold for statistical significance ($p < 1.68 \times 10^{-7}$) allowing us to replicate all associations. However, caution should be exercised when interpreting findings that are less significant than this Bonferroni significance threshold, especially those involving a low frequency of bi-allelic variants and low case counts. These conditions can result in instability and lower power in the mixed-model framework, particularly for rare binary traits.[43] Third, this work is based on coding mutations within protein-coding regions. Including splicing variants and non-coding variants, such as those in untranslated regions, could enhance the power of the analyses and potentially lead to further associations.

Lastly, our study was conducted within the UKBB using individuals of European ancestry. It is possible that the number of CH events and their prevalence across the genome may vary across non-European populations. Other population structures could affect the power for association. For example, in populations such as the Finns, where bottlenecks have caused rare variants from founders to rise to appreciable frequencies, there could be increased detection power for those alleles.[10] Similarly, in consanguineous populations, the prevalence of homozygous pLoF genotypes increases due to parental relatedness. This increases the number of identifiable bi-allelic genotypes leading to a potential increase in power for association analysis. This is particularly relevant for members of the East London Genes & Health cohort[44] or other cohorts with participants reporting high degrees of consanguinity. It will be critical to apply similar approaches in other cohorts encompassing diverse populations to further explore this diversity.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Exome sequencing quality control summary
  - Exome sequencing quality control
  - Additional ES quality control
  - Sex imputation
  - Defining samples with non-Finnish European ancestry
  - Final hard filters
  - Phasing
  - Trio-switch error rates
  - Read-backed phasing
  - Phenotype curation
  - Variant annotation masks
  - Bi-allelic encoding and recessive models
  - Gene copy dosage encoding and additive models
  - Polygenic risk scores
  - Common variant association testing
  - Estimating heritability
  - Generating PRS using LDPred2
  - Validation of polygenic risk scores
  - Conditional analysis
  - Permutation of genetic phase
  - Gene-set enrichment of bi-allelic variation
  - Homozygote and CH down-sampling
  - Power analysis for bi-allelic association
  - Simulation
  - Age-at-onset analysis
  - Cox proportional-hazards modeling
  - Replication analyses
  - Simulation of CH and homozygous variants

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2024.100602.

AUTHOR CONTRIBUTIONS

DECLARATION OF INTERESTS

REFERENCES

1. Nelson, M.R., Tipney, H., Painter, J.L., Shen, J., Nicoletti, P., Shen, Y., Floratos, A., Sham, P.C., Li, M.J., Wang, J., et al. (2015). The support of human genetic evidence for approved drug indications. Nat. Genet. *47*, 856–860. https://doi.org/10.1038/ng. 3314.

2. Plenge, R.M., Scolnick, E.M., and Altshuler, D. (2013). Validating therapeutic targets through human genetics. Nat. Rev. Drug Discov. *12*, 581–594. https://doi.org/10.1038/nrd4051.

3. Whiffin, N., Armean, I.M., Kleinman, A., Marshall, J.L., Minikel, E.V., Goodrich, J.K., Quaife, N.M., Cole, J.B., Wang, Q., Karczewski, K.J., et al. (2020). The effect of LRRK2 loss-offunction variants in humans. Nat. Med. *26*, 869–877. https://doi.org/10.1038/s41591-020- 0893-5.

4. Tobert, J.A. (2003). Lovastatin and beyond: the history of the HMG-CoA reductase inhibitors. Nat. Rev. Drug Discov. *2*, 517–526. https://doi.org/10.1038/nrd1112.

5. Do, R.Q., Vogel, R.A., and Schwartz, G.G. (2013). PCSK9 Inhibitors: potential in cardiovascular therapeutics. Curr. Cardiol. Rep. *15*, 345. https://doi.org/10.1007/s11886-012-0345-z.

6. Minikel, E.V., Karczewski, K.J., Martin, H.C., Cummings, B.B., Whiffin, N., Rhodes, D., Alföldi, J., Trembath, R.C., van Heel, D.A., Daly, M.J., et al. (2020). Evaluating drug targets through human loss-of-function genetic variation. Nature *581*, 459–464. https://doi.org/10.1038/s41586-020- 2267-z.

7. Van Hout, C.V., Tachmazidou, I., Backman, J.D., Hoffman, J.D., Liu, D., Pandey, A.K., Gonzaga-Jauregui, C., Khalid, S., Ye, B., Banerjee, N., et al. (2020). Exome sequencing and characterization of 49,960 individuals in the UK Biobank. Nature *586*, 749–756. https://doi.org/10.1038/s41586- 020-2853-0.

8. DeBoever, C., Tanigawa, Y., Lindholm, M.E., McInnes, G., Lavertu, A., Ingelsson, E., Chang, C., Ashley, E.A., Bustamante, C.D., Daly, M.J., and Rivas, M.A. (2018). Medical relevance of proteintruncating variants across 337,205 individuals in the UK Biobank study. Nat. Commun. *9*, 1612. https://doi.org/10.1038/s41467-018-03910-9.

9. Sulem, P., Helgason, H., Oddson, A., Stefansson, H., Gudjonsson, S.A., Zink, F., Hjartarson, E., Sigurdsson, G.T., Jonasdottir, A., Jonasdottir, A., et al. (2015). Identification of a large set of rare complete human knock-outs. Nat. Genet. *47*, 448–452. https://doi.org/10.1038/ng.3243.

10. Heyne, H.O., Karjalainen, J., Karczewski, K.J., Lemmelä, S.M., Zhou, W., FinnGen; Havulinna, A.S., Kurki, M., Rehm, H.L., Palotie, A., and Daly, M.J. (2023). Mono- and biallelic variant effects on disease at biobank scale. Nature *613*, 519–525. https://doi.org/10.1038/s41586-022- 05420-7.

11. Lim, E.T., Würtz, P., Havulinna, A.S., Palta, P., Tukiainen, T., Rehnström, K., Esko, T., Mägi, R., Inouye, M., Lappalainen, T., et al. (2014). Distribution and Medical Impact of Loss-of-Function Variants in the Finnish Founder Population. PLoS Genet. *10*, e1004494. https://doi.org/10.1371/journal. pgen.1004494.

12. Saleheen, D., Natarajan, P., Armean, I.M., Zhao, W., Rasheed, A., Khetarpal, S.A., Won, H.H., Karczewski, K.J., O'Donnell-Luria, A.H., Samocha, K.E., et al. (2017). Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. Nature *544*, 235–239. https://doi.org/10.1038/nature22034.

13. De Rosa, M., Fasano, C., Panariello, L., Scarano, M.I., Belli, G., Iannelli, A., Ciciliano, F., and Izzo, P. (2000). Evidence for a recessive inheritance of Turcot's syndrome caused by compound heterozygous mutations within the PMS2 gene. Oncogene *19*, 1719–1723. https://doi.org/10.1038/sj. onc.1203447.

14. Hague, S., Rogaeva, E., Hernandez, D., Gulick, C., Singleton, A., Hanson, M., Johnson, J., Weiser, R., Gallardo, M., Ravina, B., et al. (2003). Early-onset Parkinson's disease caused by a compound heterozygous DJ-1 mutation. Ann. Neurol. *54*, 271–274. https://doi.org/10.1002/ana.10663.

15. Robinson, J.P., Johnson, V.L., Rogers, P.A., Houlston, R.S., Maher, E.R., Bishop, D.T., Evans, D.G.R., Thomas, H.J.W., Tomlinson, I.P.M., and Silver, A.R.J. (2005). Evidence for an Association between Compound Heterozygosity for Germ Line Mutations in the Hemochromatosis (HFE) Gene and Increased Risk of Colorectal Cancer. Cancer Epidemiol. Biomarkers Prev. *14*, 1460–1463. https://doi.org/10.1158/1055-9965.EPI-04-0672.

16. Maffei, L., Rochira, V., Zirilli, L., Antunez, P., Aranda, C., Fabre, B., Simone, M.L., Pignatti, E., Simpson, E.R., Houssami, S., et al. (2007). 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. Clin. Endocrinol. *67*, 218–224. https://doi.org/10.1111/j.1365-2265.2007.02864.x.

17. Wang, X.-H., Xie, L., Chen, S., Xu, K., Bai, X., Jin, Y., Qiu, Y., Liu, X.Z., Sun, Y., and Kong, W.J. (2021). Identification of Novel Compound Heterozygous MYO15A Mutations in Two Chinese Families with Autosomal Recessive Nonsyndromic Hearing Loss. Neural Plast. *2021*, 9957712. https:// doi.org/10.1155/2021/9957712.

18. Guo, M.H., Francioli, L.C., Stenton, S.L., Goodrich, J.K., Watts, N.A., Singer-Berk, M., Groopman, E., Darnowsky, P.W., Solomonson, M., Baxter, S., et al. (2023). Inferring compound heterozygosity from large-scale exome sequencing data. Preprint at bioRxiv. https://doi.org/10. 1101/2023.03.19.533370.

19. Hofmeister, R. J., Ribeiro, D. M., Rubinacci, S., Delaneau O. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. Nat. Genet. 55, 1243-1249. 2022.

20. Delaneau, O., Zagury, J.-F., Robinson, M.R., Marchini, J.L., and Dermitzakis, E.T. (2019). Accurate, scalable and integrative haplotype estimation. Nat. Commun. *10*, 5436. https://doi.org/10.1038/s41467-019-13225-y.

21. Maestri, S., Maturo, M.G., Cosentino, E., Marcolungo, L., Iadarola, B., Fortunati, E., Rossato, M., and Delledonne, M. (2020). A Long-Read Sequencing Approach for Direct Haplotype Phasing in Clinical Settings. Int. J. Mol. Sci. *21*, 9177. https://doi.org/10.3390/ijms21239177.

22. Li, N., and Stephens, M. (2003). Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. Genetics *165*, 2213–2233. https://doi.org/10.1093/genetics/ 165.4.2213.

23. Loh, P.-R., Danecek, P., Palamara, P.F., Fuchsberger, C., A Reshef, Y., K Finucane, H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R., et al. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. Nat. Genet. *48*, 1443–1448. https://doi.org/10.1038/ ng.3679.

24. Barton, A.R., Sherman, M.A., Mukamel, R.E., and Loh, P.R. (2021). Whole-exome imputation within UK Biobank powers rare coding variant

association and fine-mapping analyses. Nat. Genet. *53*, 1260–1269. https://doi.org/10.1038/s41588-021-00892-1.

25. Browning, S.R., and Browning, B.L. (2011). Haplotype phasing: existing methods and new developments. Nat. Rev. Genet. *12*, 703–714. https://doi.org/10.1038/nrg3054.

26. Delaneau, O., Zagury, J.-F., Robinson, M.R., Marchini, J.L., and Dermitzakis, E.T. (2019). Accurate, scalable and integrative haplotype estimation. Nat. Commun. *10*, 5436.

27. Martin, M., Patterson, M., Garg, S., Fischer, SO, Pisanti, N, Klau, GW, Schöenhuth, A, Marschall, T WhatsHap: fast and accurate read-based phasing. Preprint at: bioRxiv 10.1101/085050 Nov. 2016.

28. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443. https://doi.org/10.1038/s41586-020-2308-7.

29. Samocha, K.E., Kosmicki, J.A., Karczewski, K.J., O'Donnell-Luria, A.H., Pierce-Hoffman, E., MacArthur, D.G., Neale, B.M., and Daly, M.J. (2017). Regional missense constraint improves variant deleteriousness prediction. Preprint at bioRxiv. https://doi.org/10.1101/148353.

30. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., and McKusick, V.A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledge-base of human genes and genetic disorders. Nucleic Acids Res. *33*, D514–D517. https://doi.org/10.1093/nar/gki033.

31. Zhou, W., Nielsen, J.B., Fritsche, L.G., Dey, R., Gabrielsen, M.E., Wolford, B.N., LeFaive, J., VandeHaar, P., Gagliano, S.A., Gifford, A., et al. (2018). Efficiently controlling for casecontrol imbalance and sample relatedness in large-scale genetic association studies. Nat. Genet. *50*, 1335–1341. https://doi.org/10.1038/s41588-018-0184-y.

32. Privé, F., Arbel, J., and Vilhjálmsson, B.J. (2021). LDpred2: better, faster, stronger. Bioinformatics *36*, 5424–5431. https://doi.org/10.1093/bioinformatics/btaa1029.

33. Jurgens, S.J., Pirruccello, J.P., Choi, S.H., Morrill, V.N., Chaffin, M., Lubitz, S.A., Lunetta, K.L., and Ellinor, P.T. (2023). Adjusting for common variant polygenic scores improves yield in rare variant association analyses. Nat. Genet. *55*, 544–548. https://doi.org/10.1038/s41588-023-01342-w.

34. Kavec, M.J., Urbanova, M., Makovicky, P., Opattová, A., Tomasova, K., Kroupa, M., Kostovcikova, K., Siskova, A., Navvabi, N., Schneiderova, M., et al. (2022). Oxidative Damage in Sporadic Colorectal Cancer: Molecular Mapping of Base Excision Repair Glycosylases MUTYH and hOGG1 in Colorectal Cancer Patients. Int. J. Mol. Sci. *23*, 5704. https://doi.org/10.3390/ijms23105704.

35. Rice, N.E., Patel, B.D., Lang, I.A., Kumari, M., Frayling, T.M., Murray, A., and Melzer, D. (2008). Filaggrin gene mutations are associated with asthma and eczema in later life. J. Allergy Clin. Immunol. *122*, 834–836. https://doi.org/10.1016/j.jaci.2008.07.027.

36. Palmer, C.N.A., Irvine, A.D., Terron-Kwiatkowski, A., Zhao, Y., Liao, H., Lee, S.P., Goudie, D.R., Sandilands, A., Campbell, L.E., Smith, F.J.D., et al. (2006). Common loss-offunction variants of the epidermal barrier protein filaggrin are a major predisposing factor for atopic dermatitis. Nat. Genet. *38*, 441–446. https://doi.org/10.1038/ng1767.

37. Carlsen, B.C., Meldgaard, M., Johansen, J.D., Thyssen, J.P., Menné, T., Szecsi, P.B., and Stender, S. (2013). Filaggrin compound heterozygous patients carry mutations in trans position. Exp. Dermatol. *22*, 572–575. https://doi.org/10.1111/exd.12199.

38. Riethmuller, C., McAleer, M.A., Koppes, S.A., Abdayem, R., Franz, J., Haftek, M., Campbell, L.E., MacCallum, S.F., McLean, W.H.I., Irvine, A.D., and Kezic, S. (2015). Filaggrin breakdown products determine corneocyte conformation in patients with atopic dermatitis. J. Allergy Clin. Immunol. *136*, 1573–1580.e2. https://doi.org/10.1016/j.jaci.2015.04. 042.

39. Liu, X., Tang, Z., Li, C., Yang, K., Gan, G., Zhang, Z., Liu, J., Jiang, F., Wang, Q., and Liu, M. (2010). Novel USH2A compound heterozygous mutations cause RP/USH2 in a Chinese family. Mol. Vis. *16*, 454–461. https://doi.org/10.1167/3.9. 454.

40. Zhou, W., Bi, W., Zhao, Z., Dey, K.K., Jagadeesh, K.A., Karczewski, K.J., Daly, M.J., Neale, B.M., and Lee, S. (2021). SAIGE-GENE+ improves the efficiency and accuracy of set-based rare variant association tests. Nat. Genet. *54*, 1466–1469.

41. Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., et al. (2014). Pfam: the protein families database. Nucleic Acids Res. *42*, D222–D230. https://doi.org/10.1093/nar/gkt1223.

42. Obeidat, M., Li, X., Burgess, S., Zhou, G., Fishbane, N., Hansel, N.N., Bossé, Y., Joubert, P., Hao, K., Nickle, D.C., et al. (2017). Surfactant protein D is a causal risk factor for COPD: results of Mendelian randomisation. Eur. Respir. J. *50*, 1700657. https://doi.org/10.1183/13993003. 00657-2017.

43. Karczewski, K.J., Solomonson, M., Chao, K.R., Goodrich, J.K., Tiao, G., Lu, W., Riley-Gillis, B.M., Tsai, E.A., Kim, H.I., Zheng, X., et al. (2022). Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. Cell Genom. *2*, 100168. https://doi.org/10.1016/j.xgen.2022.100168.

44. Finer, S., Martin, H.C., Khan, A., Hunt, K.A., MacLaughlin, B., Ahmed, Z., Ashcroft, R., Durham, C., MacArthur, D.G., McCarthy, M.I., et al. (2020). Cohort Profile: East London Genes & Health (ELGH), a community-based population genomics and health study in British Bangladeshi and British Pakistani people. Int. J. Epidemiol. *49*, 20–21i. https://doi.org/10.1093/ije/dyz174.

45. Hail Team. Hail 0.2. https://github.com/hail-is/hail.

46. Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993.

47. Eastwood, S.V., Mathur, R., Atkinson, M., Brophy, S., Sudlow, C., Flaig, R., de Lusignan, S., Allen, N., and Chaturvedi, N. (2016). Algorithms for the Capture and Adjudication of Prevalent and Incident Diabetes in UK Biobank. PLoS One *11*, e0162388.

48. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235.

49. Privé, F., Aschard, H., Ziyatdinov, A., and Blum, M.G.B. (2018). Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. Bioinformatics *34*, 2781–2787.

50. Georgi, B., Voight, B.F., and Bućan, M. (2013). From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. PLoS Genet. *9*, e1003484. https://doi.org/10.1371/journal.pgen.1003484.

51. Vinceti, A., Karakoc, E., Pacini, C., Perron, U., De Lucia, R.R., Garnett, M.J., and Iorio, F. (2021). CoRe: a robustly benchmarked R package for identifying core-fitness genes in genome-wide pooled CRISPR-Cas9 screens. BMC Genom. *22*, 828.

52. Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K.R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., et al. (2015). High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell *163*, 1515–1526. https://doi.org/10.1016/j.cell.2015.11.015.

53. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience *4*, 7.

54. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.

55. Li, H. (2014). Toward better understanding of artifacts in variant calling from highcoverage samples. Bioinformatics *30*, 2843–2851.

56. Van der Auwera, G.A., and O'Connor, B.D. (2020). Genomics in the Cloud: Using Docker, GATK, and WDL in Terra en (O'Reilly Media, Inc.).

57. Zhang, D., Dey, R., and Lee, S. (2020). Fast and robust ancestry prediction using principal component analysis. Bioinformatics 36, 3439–3446.

58. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience 10, giab008. https://doi.org/10.1093/gigascience/giab008.

59. Harrell, F.E. (2019). Package 'hmisc'. CRAN2018 2019, 235–236. https://cran.r-project.org/web/packages/Hmisc/index.html.

60. Kuan, V., Denaxas, S., Gonzalez-Izquierdo, A., Direk, K., Bhatti, O., Husain, S., Sutaria, S., Hingorani, M., Nitsch, D., Parisinos, C.A., et al. (2019). A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. Lancet. Digit. Health 1, e63–e77. https://doi.org/10.1016/S2589-7500(19)30012-3.

61. Censin, J.C., Peters, S.A.E., Bovijn, J., Ferreira, T., Pulit, S.L., Mägi, R., Mahajan, A., Holmes, M.V., and Lindgren, C.M. (2019). Causal relationships between obesity and the leading causes of death in women and men. PLoS Genet. 15, e1008405.

62. Udler, M.S., McCarthy, M.I., Florez, J.C., and Mahajan, A. (2019). Genetic Risk Scores for Diabetes Diagnosis and Precision Medicine. Endocr. Rev. 40, 1500–1520.

63. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. 17, 122.

64. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443.

65. Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., et al. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am. J. Hum. Genet. 99, 877–885.

66. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 47, D886–D894.

67. Zhou, W., Nielsen, J.B., Fritsche, L.G., Dey, R., Gabrielsen, M.E., Wolford, B.N., LeFaive, J., VandeHaar, P., Gagliano, S.A., Gifford, A., et al. (2018). Efficiently controlling for casecontrol imbalance and sample relatedness in large-scale genetic association studies. Nat. Genet. 50, 1335–1341.

68. Privé, F., Aschard, H., Carmi, S., Folkersen, L., Hoggart, C., O'Reilly, P.F., and Vilhjálmsson, B.J. (2021). Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. AJHG 109, 12–23. https://doi.org/10.1101/2021.02.05.21251061.

69. International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. Nature 467, 52–58.

70. Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium; Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. 47, 291–295.

71. Hart, T., Tong, A.H.Y., Chan, K., Van Leeuwen, J., Seetharaman, A., Aregger, M., Chandrashekhar, M., Hustedt, N., Seth, S., Noonan, A., et al. (2017). Evaluation and Design of Genome-Wide CRISPR/SpCas9 Knockout Screens. G3 (Bethesda). 7, 2719–2727. https://doi.org/10.1534/g3.117.041277.

72. Gavish, A., Tyler, M., Greenwald, A.C., Hoefflin, R., Simkin, D., Tschernichovsky, R., Galili Darnell, N., Somech, E., Barbolin, C., Antman, T., et al. (2023). Hallmarks of transcriptional intratumour heterogeneity across a thousand tumours. Nature 618, 598–606. https://doi.org/10.1038/s41586-023-06130-4.

73. Pirinen, M. GWAS 3: Statistical Power Feb. 2023 https://www.mv.helsinki.fi/home/mjxpirin/GWAS_course/material/GWAS3.html.

74. Kaplan, E.L., and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. J. Am. Stat. Assoc. 53, 457–481.

75. Therneau, T. M. A Package for Survival Analysis in R R Package Version 3.2-3 (2020) https://cran.r-project.org/web/packages/survival/vignettes/survival.pdf.

76. Kassambara, A & Kosinski, M, . Biecek P. survminer: drawing survival curves using 'ggplot2'R package version 0.4. 9 https://CRAN.R-project.org/package=survminer,

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
UK Biobank Exome and Genotype data	Bycroft et al.[45]	https://www.ukbiobank.ac.uk/
Summary statistics	This paper	https://doi.org/10.5281/zenodo.10939698
Software and algorithms		
SAIGE (1.1.9)	Zhou et al.[40]	https://www.nature.com/articles/s41588-020-0621-6
HAIL (0.2.97)	Hail team[44]	https://github.com/hail-is/hail
SHAPEIT5 (5.0.0)	Delaneau et al.[26]	https://github.com/odelaneau/shapeit5
LDpred2 (1.12.1)	Prive et al.[32]	https://github.com/privefl/bigsnpr
PLINK (1.9)	Chang et al.[41]	https://www.cog-genomics.org/plink/1.9/
BCFtools (1.12)	Danecek et al.[46]	https://samtools.github.io/bcftools/
Whatshap (1.6)	Martin et al.[27]	https://whatshap.readthedocs.io/en/latest/
Variant Effect Predictor (95)	McLaren et al.[47]	https://www.ensembl.org/info/docs/tools/vep/index.html
Gene lists	Multiple authors[28,48–52]	https://github.com/macarthur-lab/gene_lists
Main pipeline code	This paper	https://github.com/frhl/wes_ko_ukbb_nexus,10.5281/zenodo.10926001

RESOURCE AVAILABILITY

Lead contact
For additional information, as well as requests regarding resources and reagents, please direct your inquiries to the lead contact, Duncan S. Palmer: duncan.stuart.palmer@gmail.com.

Materials availability
This study did not generate new unique reagents.

Data and code availability
All original code has been deposited at Zenodo and is publicly available as of the date of publication (https://doi.org/10.5281/zenodo.10926001). Summary statistics have been deposited at Zenodo (https://doi.org/10.5281/zenodo.10939698) and are publicly available as of the date of publication. DOIs are listed in the key resources table. Quality control and CH calling were performed using Hail (version between 0.2.95 and 0.2.120). Phasing was performed using SHAPEIT5 (5.0.0). All analysis pertaining to PRS were performed using the R-package bigsnpr (version between 1.10 and 1.12.1). Assocation analysis was performed usign SAIGE (1.1.9). Age-of-onset analysis and Kaplan-Meier plotting was performed using R-packages survival (3.2.13) and survminer (0.4.9). Essential and non-essential gene lists were downloaded from https://github.com/macarthur-lab/gene_lists (https://doi.org/10.5281/ZENODO.6724345). Unless otherwise indicated, analyses were performed in R (4.1.1) and python (3.6.13) and plotted using the R-package ggplot2 (3.4.0).

METHOD DETAILS

Exome sequencing quality control summary
We perform a series of hard-filters on genotype, sample, and variant metrics (Tables S22–S24). We confirm genetic sex with reported sex, and restrict analysis to genetically ascertained samples of NFE ancestry, using random forest (RF) classifiers (Figures S25 and S26). Finally, we filter based on a second collection of sample and variant filters (Tables S23 and S24). We used Hail 0.2[45] and PLINK 1.9[53] 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
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,[45] Table S22. 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),[54] and filtered out sample outliers of the residuals for each metric based on MAD (median absolute deviation) thresholds (Table S22). Samples without PC data were subject to more stringent thresholds (Table S22).

Variant-level filters

Retain variants satisfying all of the following conditions.

(1) Not in a low complexity region (LCR).[55]
(2) In sequencing target regions ± 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[56] in founders as determined by relatedness estimates and recorded ages of UKBB participants.[54] Variants were retained as recommended in genome analysis toolkit (GATK)[56]

Genotype filters

Multi-allelic variants were split into bi-allelic variants and insertions and deletions (indel) were left-aligned.[46] 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 -3P < 1 × 10
(b) Indels: alternate allele read depth/total read depth <0.3.
(4) Homozygous indel calls: alternate allele read depth/total read depth <0.7.

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 datasets. 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.[45]

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× or mean GQ < 47.8 (Figure S27).

Sex imputation

To confirm participant sex and calculate PCs, we extracted high quality common variants (allele frequency between 0.01 and 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, 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 (Figure S25). 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 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.[57] Next, we removed samples outside of the European population (EUR) using an RF classifier: we train an RF on the super-populations labels of 1KGP and predict the super-population for each of the UKBB samples (Figure S26). 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.

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 standard deviations (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 dataset consists of 176,935 samples, and 9,169,408 variants (Tables S23 and S24). A summary of sample and variant filters are provided in Tables S23 and S24 and Figure S28. 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[45] and BCFtools[58] (1.12). For variants in both datasets, 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.[45] 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, we phased overlapping chunks of 100,000 variants with $\geq 50,000$ variant overlap between consecutive chunks using Hail.[45] 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[58] (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 (Tables S23 and S24).

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[58] (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[45] and then repeated the switch error calculation step. We calculated binomial 95% confidence intervals (CIs) for SERs using the R-package HMisc[59] (4.7).

Read-backed phasing

We performed read-backed phasing with UKBB ES short paired-end read sequences using.cram files provided by UKBB. As WhatsHap[27] (1.6) 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 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,[45] and determined agreement between estimated phasing in WhatsHap and SHAPEIT5 (Figure 3).

Phenotype curation

We considered a collection of 282 binary quality controlled and publicly available common complex phenotypes for analysis.[60] 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.,[61] 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.,[47] subsequently applied by Udler et al.[62] 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 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)[63] (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 (Table S6). We then split PTVs into putative loss of function (pLoF) (HC) and LC loss-of-function variants using LOFTEE,[64] and labeled missense variants with both Rare Exome Variant Ensemble Learner (REVEL)[65] score $\geq$ 0.6 and CADD[66] 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 models

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 (1.1.9),[67] adjusting for sex, age, sex × age, age ^{2}, UKBB center, 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 significance ($P < 0.05$/gene $-$ trait pairs).

Gene copy dosage encoding and additive models

We define annotate individuals as being 'mono-allelic' for a gene if they harbor at least one pLoFs or damaging missense variant with 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,[67] adjusting for sex, age, sex × age, age ^{2}, UKBB center, 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 ($P < 0.05$/gene $-$ trait pairs).

Polygenic risk scores
Curation of array-based genetic data

We generated PRSs using imputed genotypes provided by UKBB.[54] In the following, we make the distinction between training and testing data. The first represents the samples that are used for fitting LDPred2[68] 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^2_{snp} and $n_{eff} \geq 5000$), 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.,[68] and filtered to common Haplotype Map 3 (HM3) SNPs.[69] 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:

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

where n_{cases} and $n_{controls}$ 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,[45] running logistic regression (logistic_regression_rows) adjusting for sex, age, sex × age, age ^{2}, UKBB assessment center, 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^2_{snp} and standard errors (SEs) using LD score regression (LDSC).[48,70] We evaluated PRS for phenotypes with

nominal significant h^2_{snp} estimates ($P < 0.05$) and restricted to phenotypes with nominally significant ($P < 0.05$) LDSC based SNP heritability estimates and effective sample size $n_{eff} \geq 5,000$, where:

$$n_{eff} = \frac{4}{\dfrac{1}{n_{cases}} + \dfrac{1}{n_{controls}}}.$$

(Equation 2)

Generating PRS using LDPred2

For a given phenotype, we trained a PRS predictor with LDPred2-auto,[68] 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^2_{snp} 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[49] (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[67] models while controlling for off-chromosome PRS by including it is 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 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,[67] conditioning on the lead variant and repeating the regression until the conditional P for the newly included variant dropped below 5×10^{-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 ≥ 10, MAF ≤ 0.05) and ultra-rare (MAC ≤ 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 ≤ 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 score-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:

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

(Equation 3)

where n is the number of permutations, t_i is the score-statistic under the ith random label shuffling, and $t_{observed}$ is the observed score-statistic determined using the observed genetic phase. To ensure sampling of score-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,[50] essential gnomAD,[28] essential ADaM,[51] essential in culture,[52] essential CRISPR,[71] genes with pLI > 0.9 in gnomAD,[28] non-essential in culture,[52] homozygous LoF tolerant,[28] and non-essential gnomAD[28] and Curated Cancer Cell Atlast.[72]

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.

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

where x is a pair (x_1, x_2): $x_1 \in$ {pLoF, damaging missense, pLoF and/or damaging missense, other missense, synonymous}, $x_2 \in$ {heterozygote, CH, bi-allelic variants}. For each annotation category we use the transcript-specific mutation rate.[29] 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 (Figure S13).

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[73] allowing us to determine the effective effect size on the OR scale across candidate configurations of binary case-control counts by substituting alternate allele frequencies with bi-allelic variant frequencies. We calculated effect sizes at 80% power at Bonferroni significance ($P < 1.68 \times 10^{-7}$) 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 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, comprised of variants annotated as pLoF or damaging missense). We then define a n samples $\times$ m genes matrix $\tilde{\mathbf{B}}$ with entries:

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

We then simulated liability under the following model:

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

where $\mathbf{B}_{i,j}$ is the $(i,j)^{th}$ entry of $\mathbf{B}$ 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 **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:

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

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

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 $b \in \{0, 0.5, 1, 2, 10\}$.

Age-at-onset analysis
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.,[60] excluding any codes related to "history of …" events for which accurate age-at-diagnosis could not be extracted. The remaining 27 phenotypes 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/protein-altering 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 to 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.[74] 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[75] and Kaplan-Meier plots drawn with the R package survminer 0.4.9.[76]

Replication analyses
In our initial analysis, we examined 176k individuals from the initial release of 200k exomes. The subsequent release of the full 450k exomes during the drafting of this paper presented an opportunity to replicate our bi-allelic associations. To do this, we leveraged the gnomAD joint variant called exomes which we combined with genotyping array data for the 450k participants, allowing phasing of rare variants. Before phasing, we performed the same initial quality-control using various hard filters, as in the discovery cohort. Our replication cohort consist of the 233,837 individuals not included in the discovery analysis. In the cross-sectional analysis, we performed replication for genes with at least five bi-allelic (pLoF + damaging missense/protein-altering) variant carriers were present. Similarly, in the longitudinal analysis, we performed replication for gene-trait combinations with at least five bi-allelic variants (homozygotes or CH) and 100 heterozygotes in the population. In the replication cohort, variant annotation masks were created using VEP (v105).

Simulation of CH and homozygous variants
We conducted simulations to compare the expected and empirically observed number of bi-allelic variants. We focused on 1,174 genes with at least one homozygous of CH variant in our dataset. Then we generated genotypes for 176,935 individuals. Each genotype simulation assumed that variants were independent and utilized a Bernoulli distribution, where the probability of success was set to the variant's minor allele frequency. We ran the simulations 10 times and derived the expected number of homozygous or CH events with. This average provided us with an estimate for the expected number of bi-allelic variants, including both homozygotes and CHs. We then compared these estimates to our actual empirical findings to benchmark the concordance.

Supplemental information

Exome-wide evidence of compound heterozygous

effects across common phenotypes in the UK Biobank

Frederik H. Lassen, Samvida S. Venkatesh, Nikolas Baya, Barney Hill, Wei Zhou, Alex Bloemendal, Benjamin M. Neale, Benedikt M. Kessler, Nicola Whiffin, Cecilia M. Lindgren, and Duncan S. Palmer

Supplemental information for Exome-wide evidence of compound heterozygous effects across common phenotypes in the UK Biobank

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

*Correspondence to: flassen@well.ox.ac.uk, cecilia.lindgren@wrh.ox.ac.uk and
duncan.stuart.palmer@gmail.com

This file includes:

- Tables **S22**-**S24**
- Figure **S1**-**S24**

Metric	Metric residual (w/ PCs)	Raw (w/o PCs)
call_rate	[24, ∞)	[4, ∞)
n_insertion	[8, 8]	[4, 4]
n_deletion	[8, 8]	[4, 4]
r_insertion_deletion	[8, 8]	[4, 4]
n_het	[12, 12]	[4, 4]
n_hom_var	[12, 12]	[4, 4]
r_het_hom_var	[16, 16]	[4, 4]
n_non_ref	[8, 8]	[4, 4]
n_singleton	(-∞, 16]	(-∞, 4]
n_snp	[8, ∞)	[4, 4]
n_transition	[8, 8]	[4, 4]
n_transversion	[8, 8]	[4, 4]
r_ti_tv	[8, 8]	[4, 4]

Table S22: **Sample filtering: MAD Intervals, related to Star Methods**. 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.

Filter	Samples	Batch 1	Batch 2	%
Initial samples in raw UKBB variant call format (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 genotype quality (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
n singletons >175	19	2	17	0.0
Samples after all sample filters	**176,935**	**41,371**	**135,564**	**88.2**

Table S23: **Summary of sample filters, related to Star Methods**. Moving down through the rows of the table, we move through QC filtering steps.

Filter	Variants	%
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
Variants remaining after all filters	**9,169,408**	**57.6**

Table S24: **Summary of variant filters, related to Star Methods**. Moving down through the rows of the table, we move through QC filtering steps
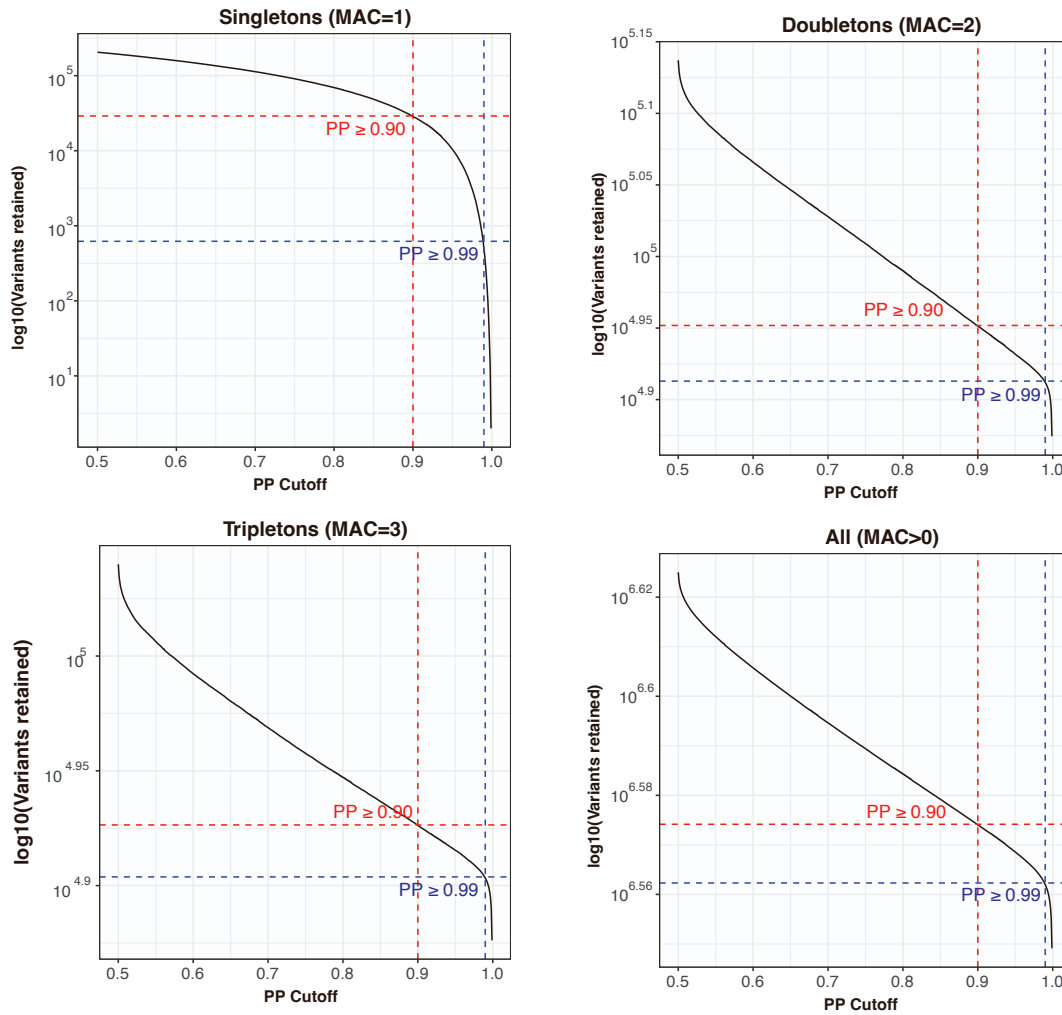
Figure S1: **Phased variants retained as a function of phasing confidence score, related to Figure 1**. Each subplot displays the number of variants retained on the $\log_{10}$ scale as the PP is increased, split by rarity of variants described in the subplot title. Dotted red and blue lines highlight the number of variants retained after imposing PP cut-offs of 0.9 and 0.99, respectively.
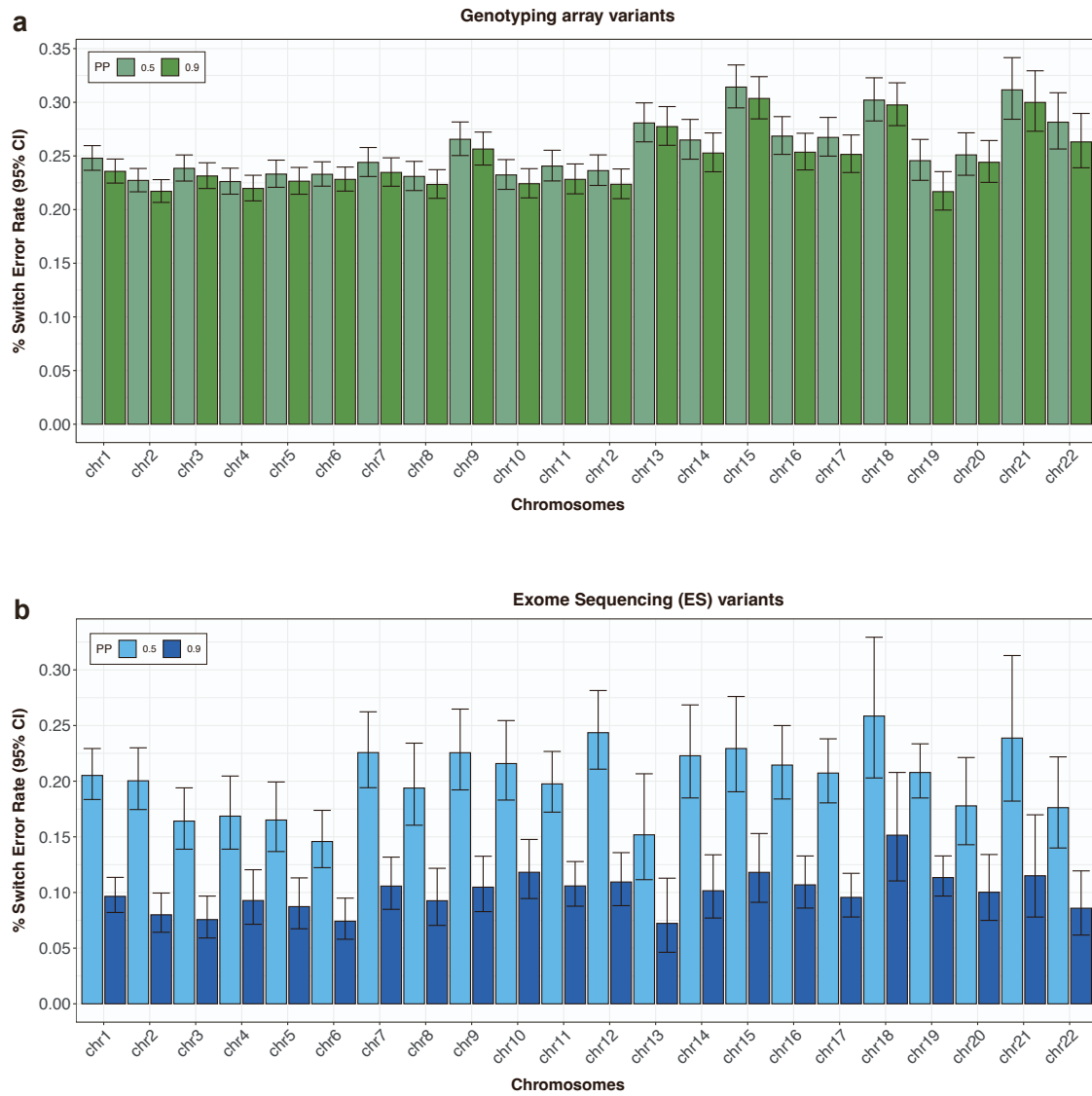
Figure S2: **Trio switch error rates by chromosome, related to Figure 1**. Parent-offspring trios are used to determine switch error rates for variants that originate from the genotyping array **(a)** and exome sequencing data **(b)**. We stratify by phasing confidence (PP) according to the color legends. Mean switch error rates are plotting, with whiskers enclosing the 95% binomial CI.

6
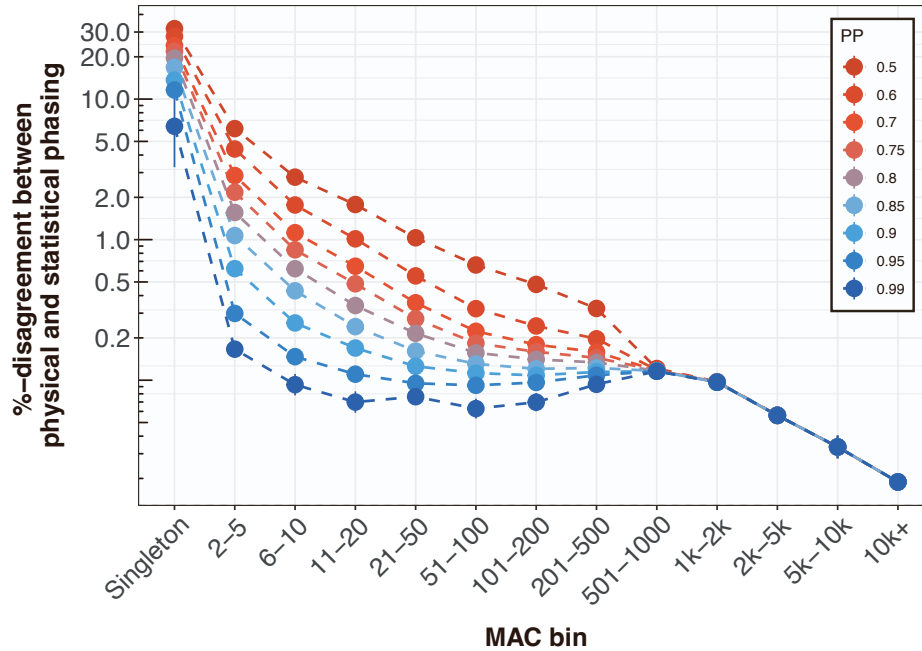
Figure S3: **Agreement between read-backed and statistical phase estimation, related to Star Methods**. Genetic phase was estimated using WhatsHap (read-backed phasing) and SHAPEIT5 (statistical phasing) in 49,756 individuals across all autosomes. We only carried forward pairs of variants close proximity 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
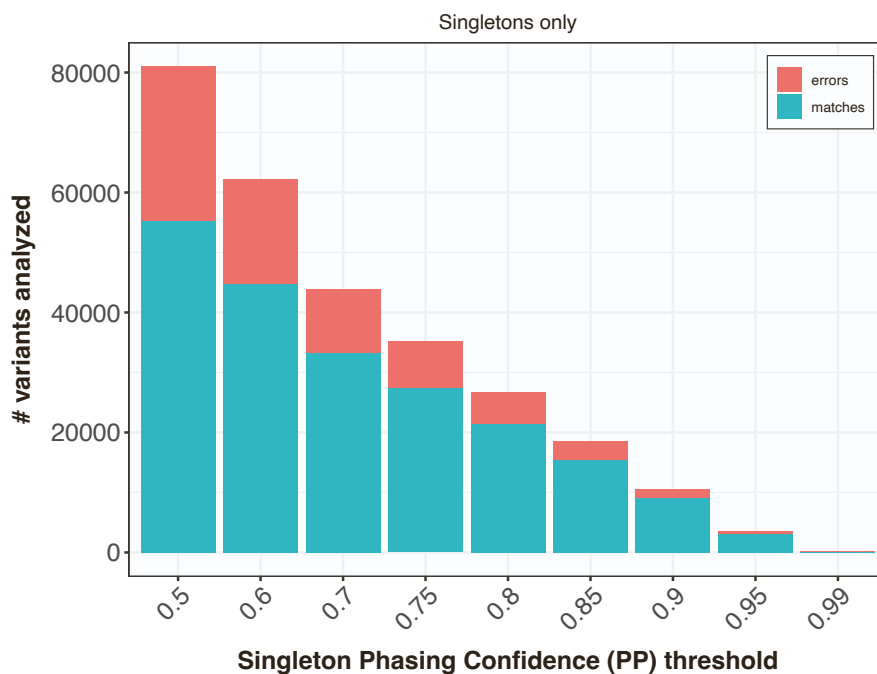
Figure S4: **Agreement between read-backed phased and statistically phased singleton variants, related to Figure 1**. Across the 49,756 samples, we identified 80,978 reads with at least one singleton (MAC= 1) variant irrespective of phasing quality. For each PP-bin we determine the agreement between read-backed phased variants and statistically phased variants with red indicating disagreement and blue indicating agreement. With higher PP-score, the proportion of correctly phased variants increases, while the total number of variants assessed decreases.
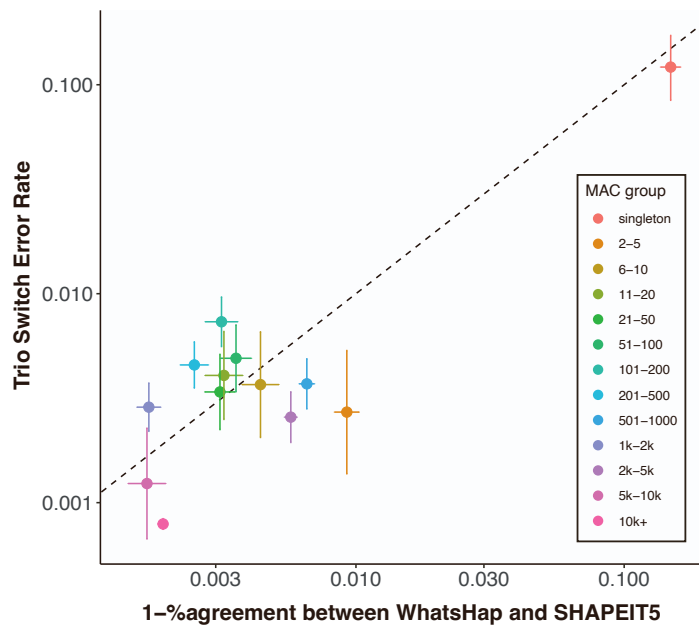
Figure S5: **Agreement between read-backed phasing and statistical phasing, related to Star Methods**. We plot the disagreement between WhatsHap (read-backed phasing) and SHAPEIT5 (statistical phasing) in UKBB on the *x*-axis against switch error rate in SHAPEIT5 phase estimates implied by trio-based 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 $y = x$.
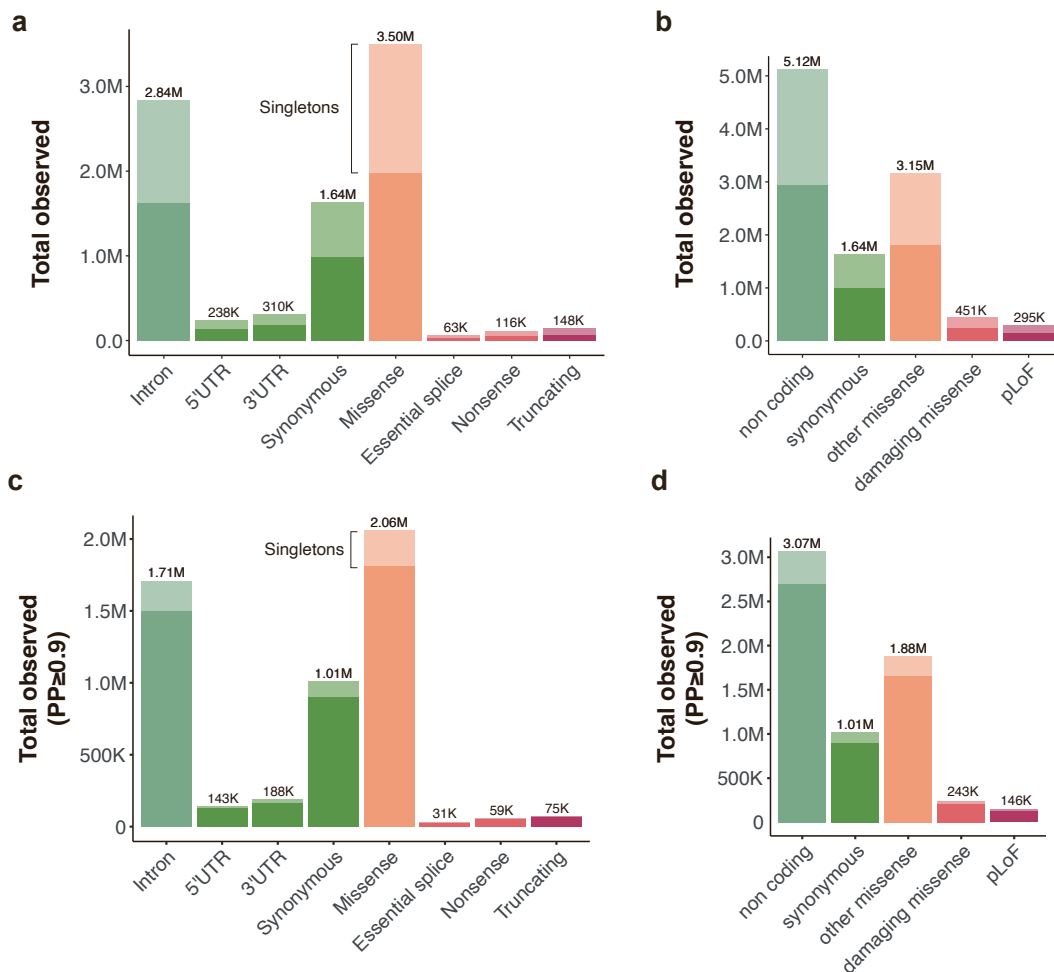
Figure S6: **Distribution of variant annotation categories before and after broad conse-quence categorization and before and after filtering by PP$\geq$ 0.9, related to Figure 1**. We annotate variants using VEP and by the most severe consequence in the canonical transcript. Panels **(a)** and **(b)** 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. Panels **(c)** and **(d)** depicts the same as above but after restricting to variants with PP$\geq$ 0.9. 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.

10
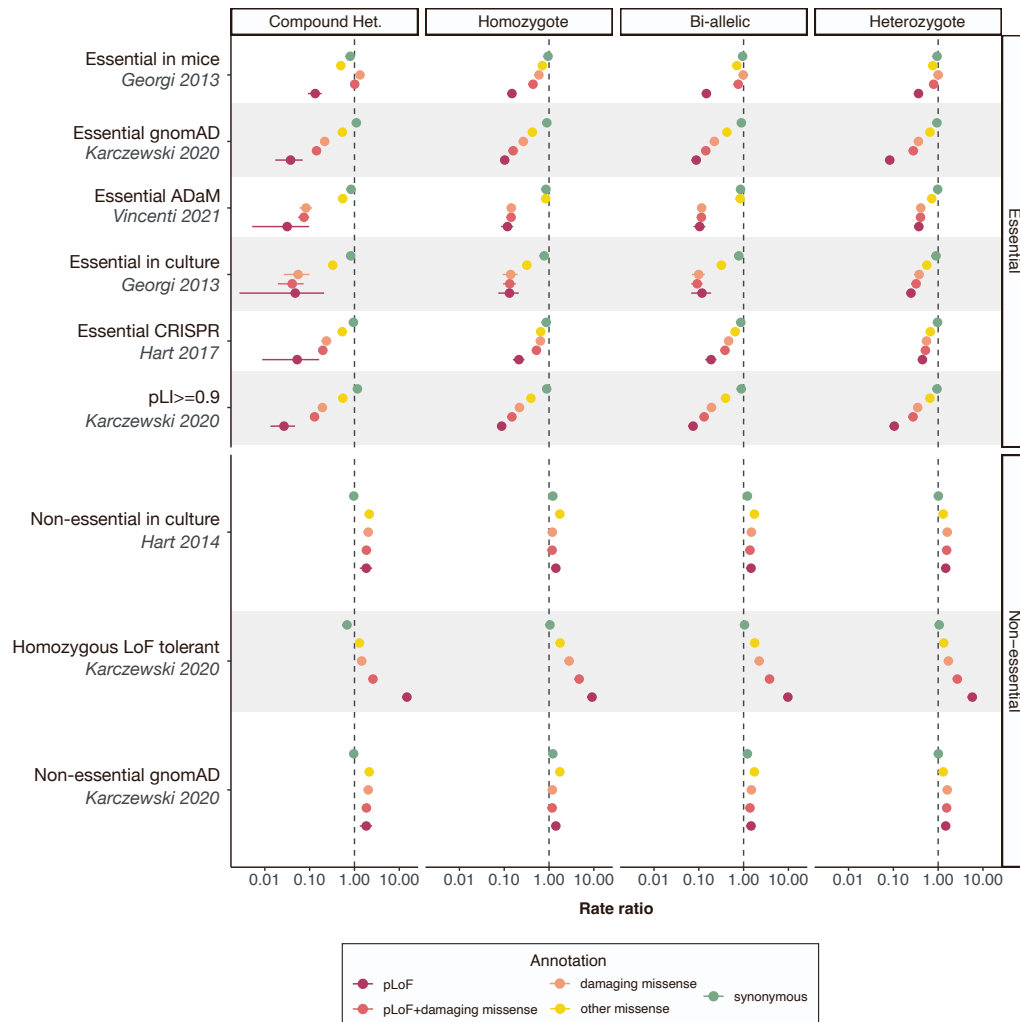
Figure S7: **Gene-set depletion/enrichment modeling, related to Star Methods**. Poisson regression to model mono- and bi-allelic variant (heterozygous, CH, homozygous or both) depletion and enrichment across essential and non-essential gene-sets. Rate ratios are shown for synonymous (green), other missense (yellow), damaging missense (orange) and pLoF (red) variants. The dashed line depicts a rate ratio of 1.
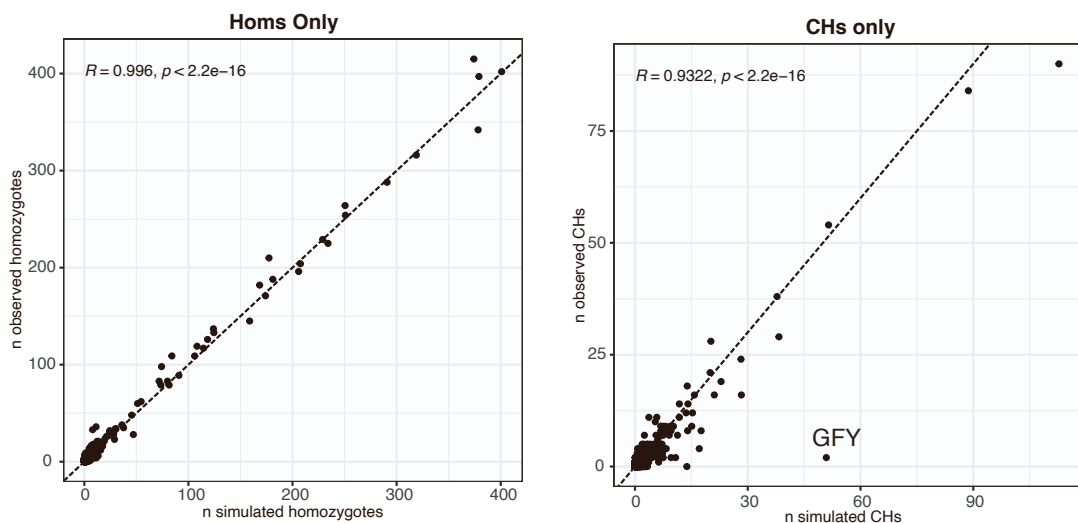
Figure S8: **Simulation of CH and homozygous events in an outbred population, related to Star Methods**. We generated genotypes for 1174 genes using allele frequencies derived from observed pLoF variants. For each gene, we simulated genotypes for 176,935 individuals. We averaged the number of bi-allelic variants across 10 simulations. This served as an estimate for the expected count (*x*-axis) of bi-allelic variants, against which we compared the empirically observations (*y*-axis). The first panel (left) is the comparison between observed and simulated homozygous variants. The second panel (right) is the comparison between observed and simulated CH variants.
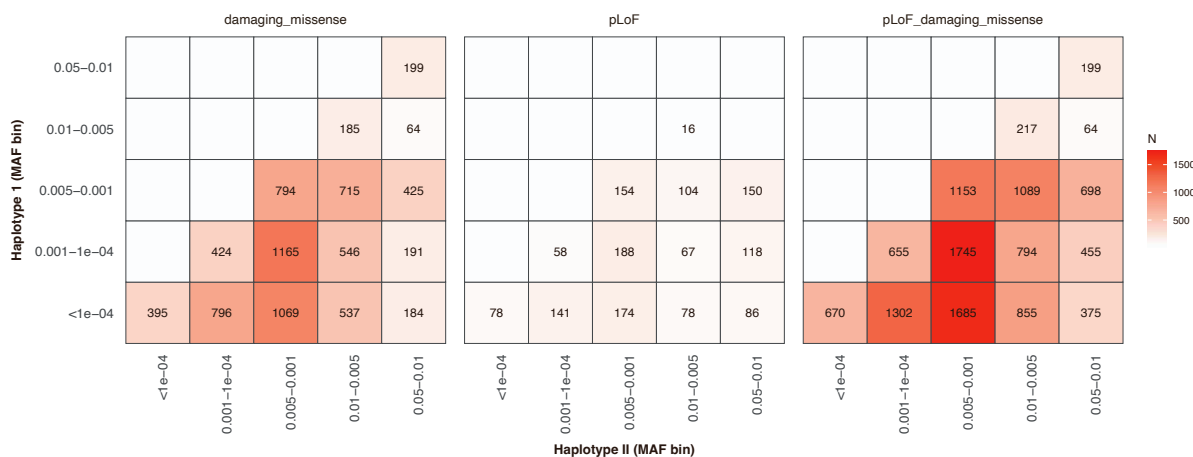
Figure S9: **Allele frequencies of variants in the CH state, related to Star Methods**. Heatmap of allele counts for variants in CH state stratified by predicted variant consequence (damaging missense, pLoF or pLoF+damaging missense). We plot the MAC for variants residing on the most common haplotype (y-axis) versus the rarest haplotype (x-axis).
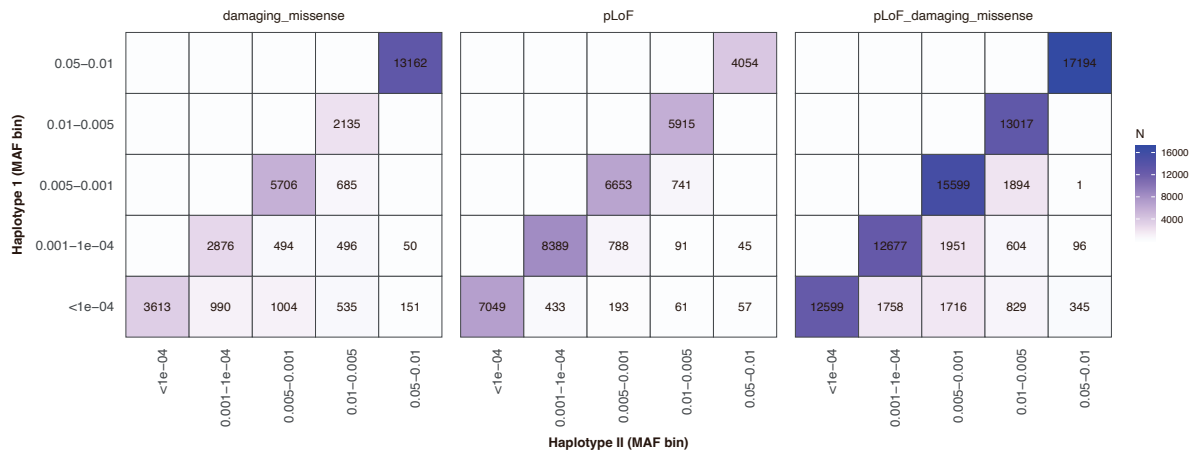
Figure S10: **Allele frequencies of variants in *cis*, related to Star Methods**. Heatmap of allele counts for co-occurring variants on the same haplotype stratified by predicted variant consequence (damaging missense, pLoF or pLoF+damaging missense). The most common variant on the haplotype versus the least common are plotted.
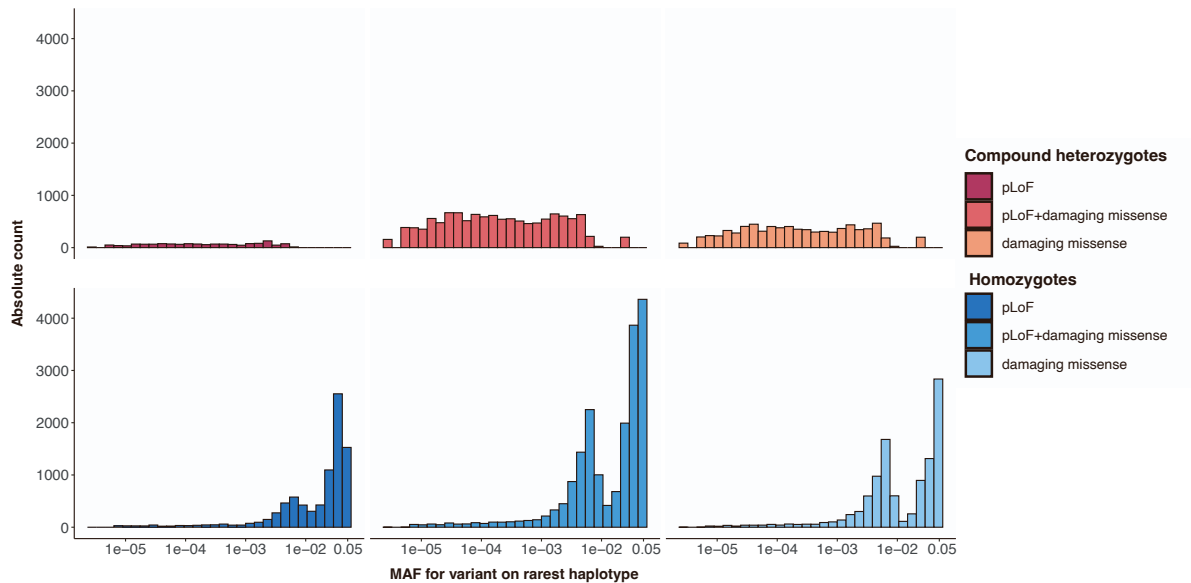
Figure S11: **Distribution of observed variants across samples by allele frequency, related to Star Methods**. Histogram of unique bi-allelic variant (CH and homozygotes) prevalence across the allele frequency spectrum. For a qualifying CH variant, the allele frequency corresponding to the alternate allele on the rarest haplotype are plotted.

13
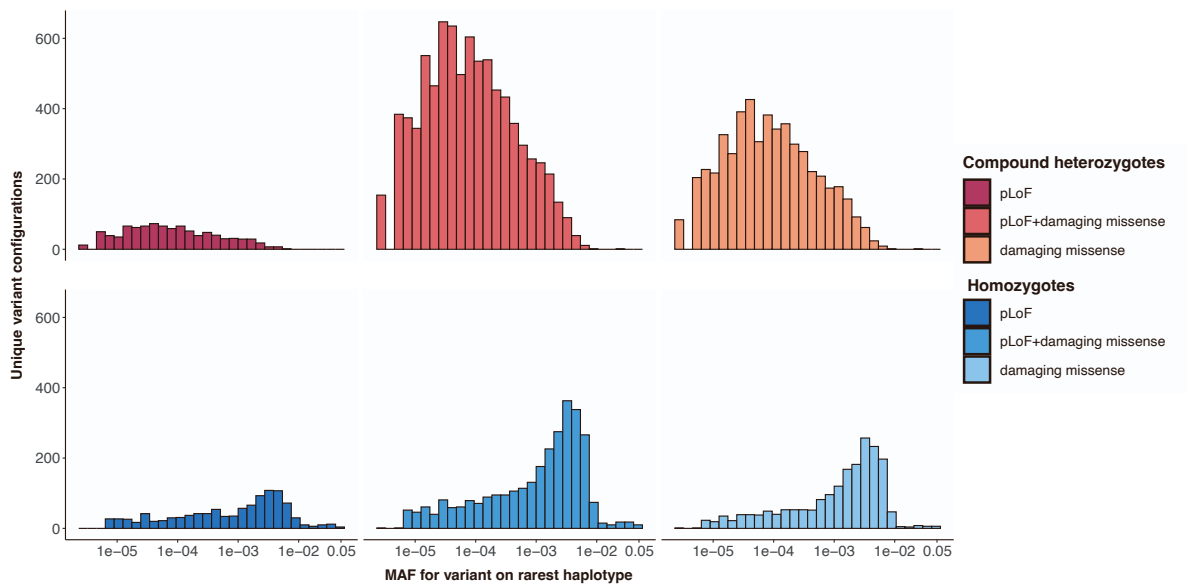
Figure S12: **Distribution of unique variants observed by allele frequency, related to Star Methods**. Histogram of bi-allelic variant (CH and homozygotes) count for all gene-samples pairs in the analysis. For a qualifying CH variant, the allele frequency corresponding to the alternate allele on the rarest haplotype are plotted.
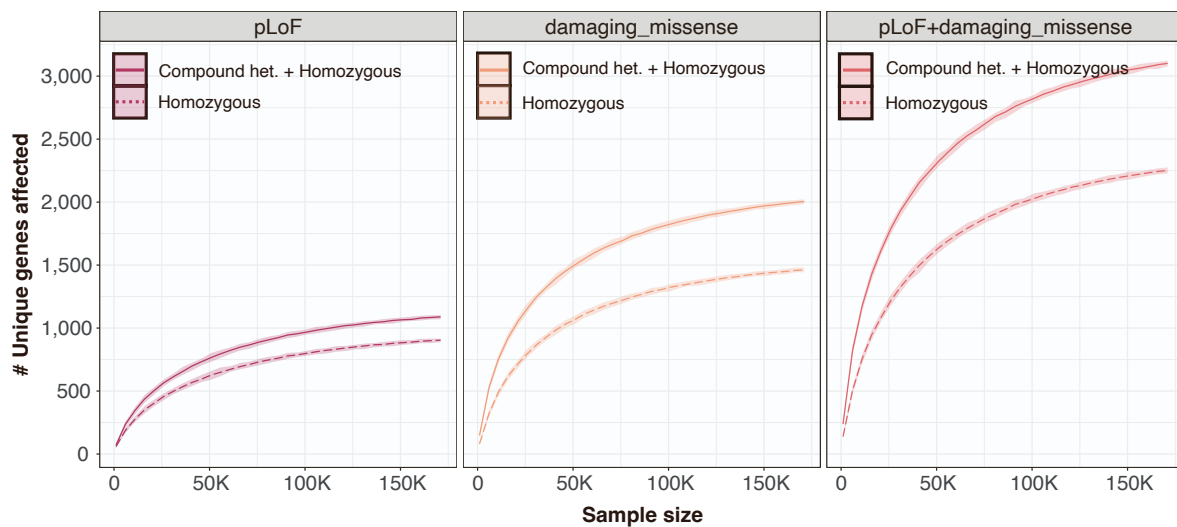
Figure S13: **Count of unique genes affected by a homozygous and homozygous or compound heterozygous variants as a function of sample size, related to Star Methods**. Starting with the full data, and down-sampling, we plot counts number of unique genes harboring homozygous and homozygous or compound 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.
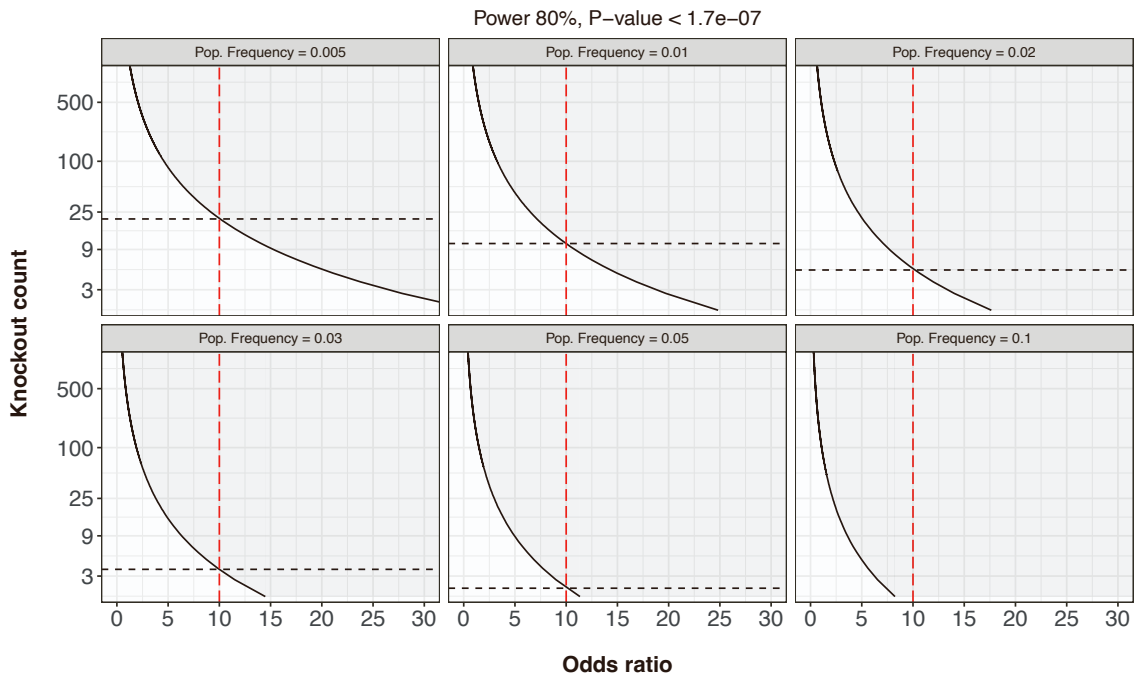
Figure S14: **Power analysis to determine the required number of bi-allelic variants to detect specific ORs at 80% power at bonferroni significance** ($P < 1.7 \times 10^{-7}$)**, related to Star Methods.** We repeat the analysis while varying trait population prevalence assuming 823 (0.5%), 1766 (1%), 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.

Figure S15:

17
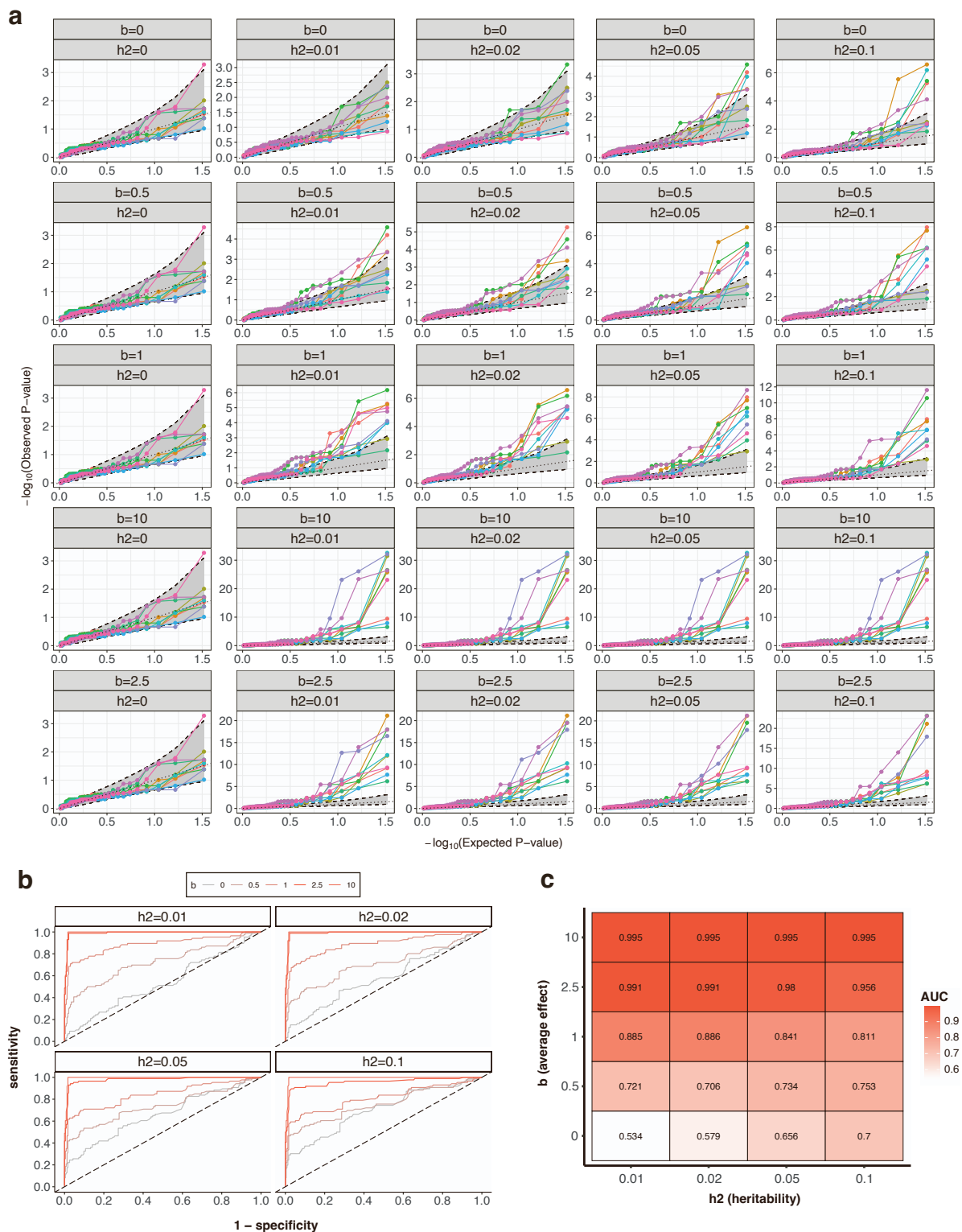
Figure S15 *(previous page)*: **Simulation study to test our ability to detect bi-allelic effects in the presence of true effects, related to Star Methods**. We simulate phenotypic data applied to 100,000 genetically-ascertained NFE on chromosome 22 (Methods) under the liability-threshold model assuming 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.
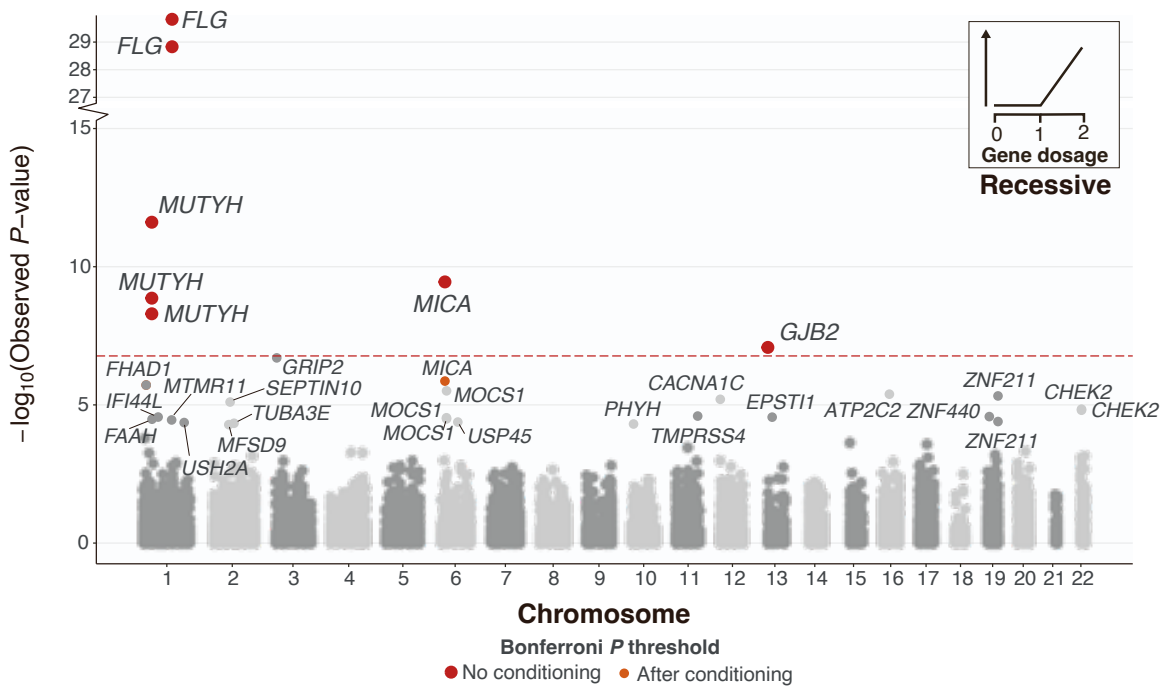
Figure S16: **Recessive association analysis without accounting for PRS, related to Figure 2**. Recessive Manhattan plot depicting $\log_{10}$-transformed gene-trait association P-values versus chromosomal location. Associations are colored red if they are Bonferroni significant ($P < 1.68 \times 10^{-7}$). Any gene-trait association with $P < 3.05 \times 10^{-6}$ (nominal significance) has been labeled with gene symbol. No additional conditioning was carried out in this analysis.
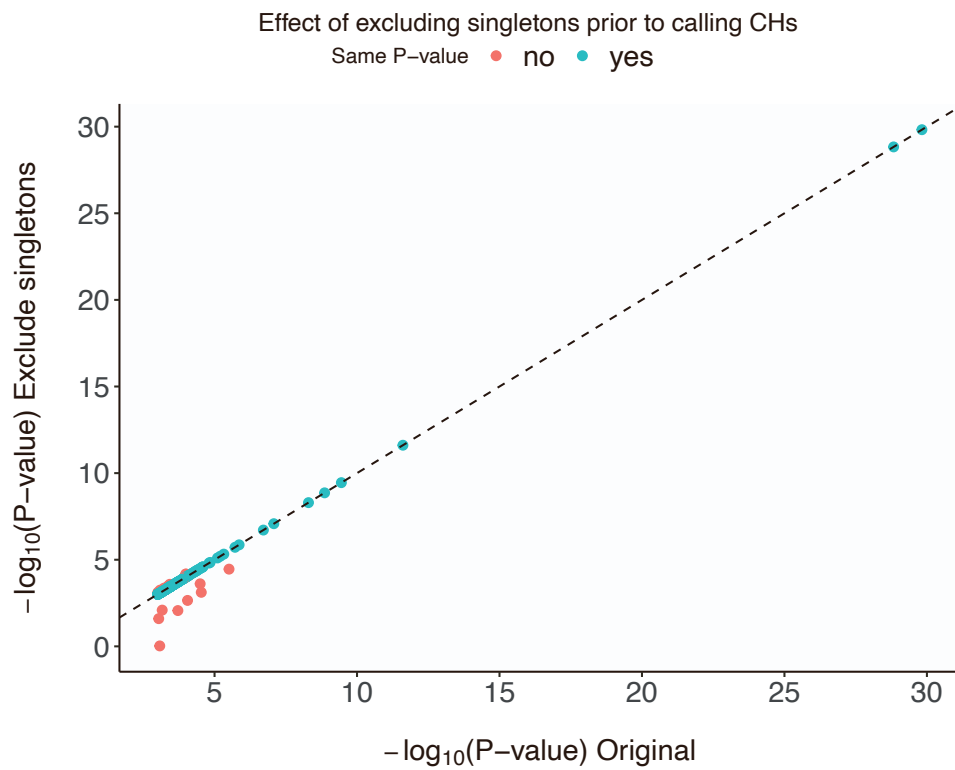
Figure S17: **Effect of excluding singletons prior to calling CH variants for any gene-traits with a $P < 0.001$ from the initial analysis, related to Figure 2**. The P-value from the original cross-sectional analysis is shown on the x-axis, while the P-value for the same associated gene-traits without singletons is depicted on the y-axis. The dots are colored by whether the degree of significance changes after excluding singletons.
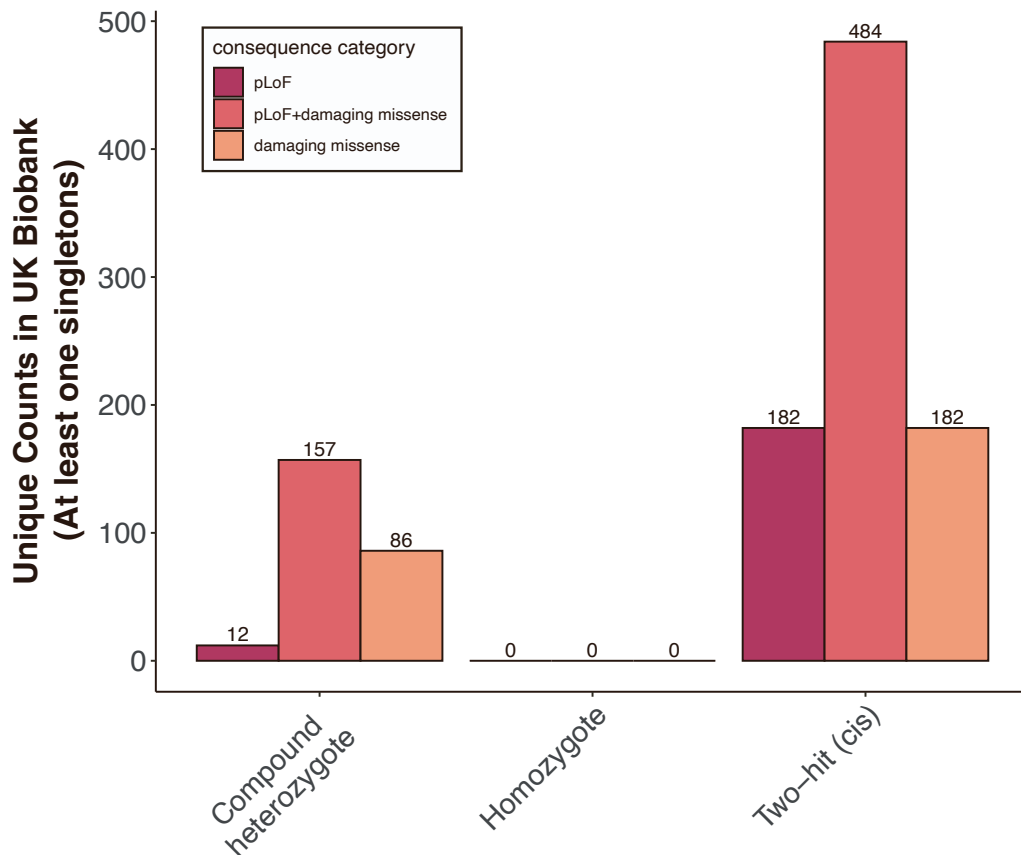
Figure S18: **Number of CH variants after filtering to hose that consist of at least one singleton, related to Figure 1**. We filter to CH variants that are comprised of at least one singleton across compound heterozygotes, homozygotes and carriers with variants *in cis*. The actual number of empirically observed carriers for each category is displayed on top of the each bar.
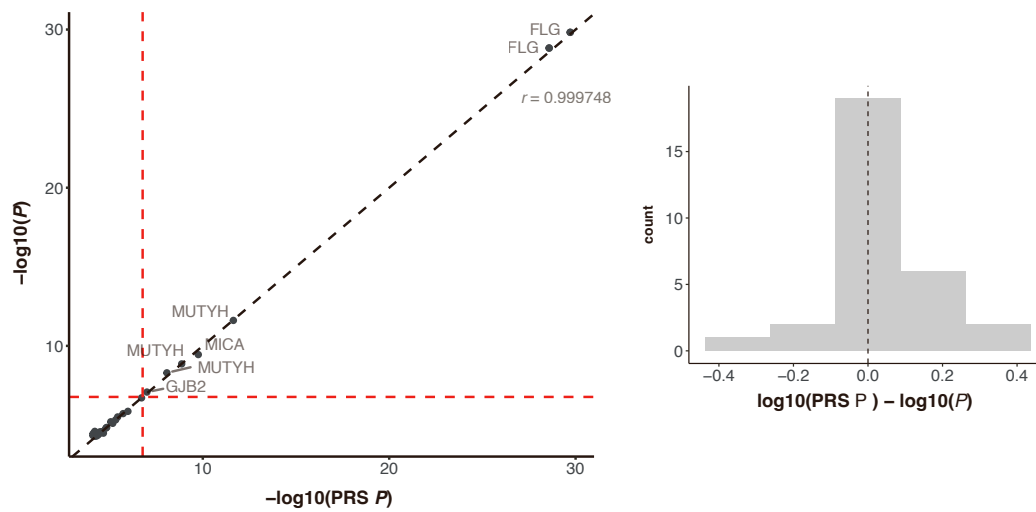
Figure S19: **Association *P*-values before and after inclusion of PRS as a covariate, related to Star Methods**. The scatter plot depicts the association P-values both before and after PRS was included as a covariate. The y-axis represents the P-value prior to PRS adjustment, while the x-axis demonstrates the P-value afterPRS adjustment. On the right, the difference in log-transformed P-values before and after PRS adjustment is displayed. The plot highlights gene-trait associations that were considered Bonferroni significant in the recessive analysis.
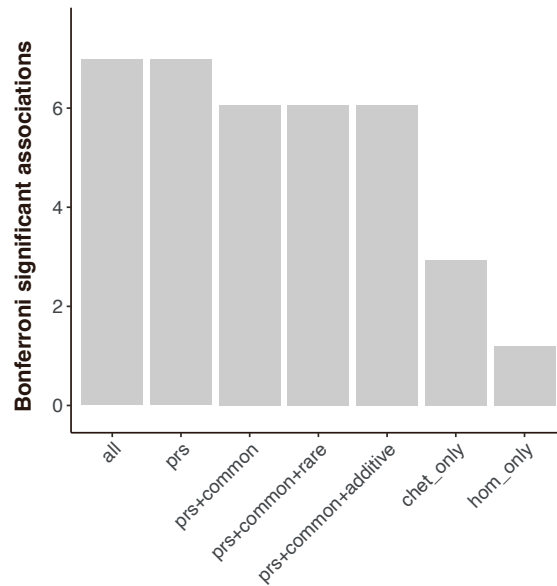
Figure S20: **Overview of attrition for Bonferroni significant associations after successive conditioning steps or filters, related to Star Methods**. This bar chart presents the number of Bonferroni significant associations that remain after successive conditioning steps in a gene-trait regression model or variant filters. The first bar represents all initial Bonferroni significant associations. The second bar shows the impact of conditioning on off-chromosome PRS. The third bar accounts for nearby common variants, which eliminates two gene-trait pairs. Subsequent bars indicate the effect of further conditioning on rare variants in the gene and an additive model of affected haplotypes, neither of which reduces the number of associations. The last two bars separate the associations those that remain after filtering to compound heterozygous or homozygous variants, respectively.
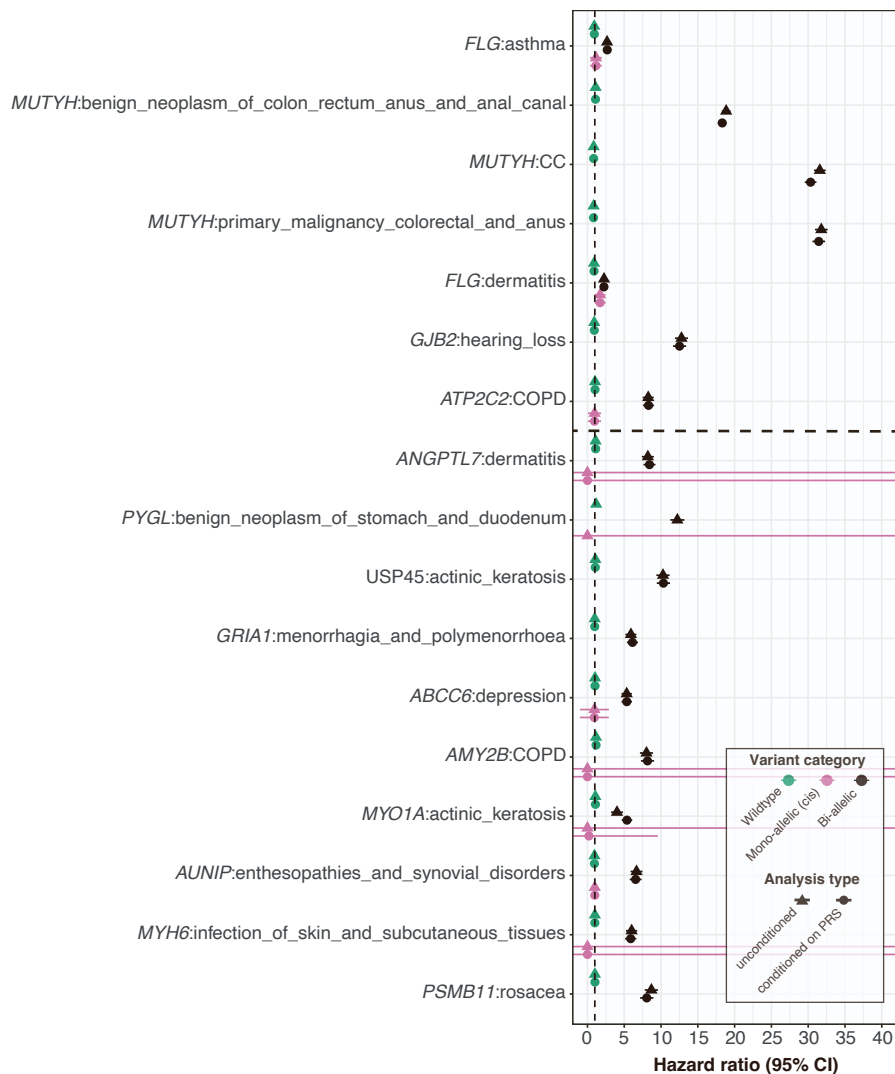
Figure S21: **Cox proportional hazards modeling with and without polygenic effects, related to Figure 4**. HRs when comparing CH and homozygous status versus heterozygous carrier status. Throughout, we display hazard ratios with (circles) and without (triangles) taking the polygenic contribution into 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 either Bonferroni significance ($P < 1.89 \times 10^{-7}$) or FDRs < 0.1 cutoff are demarcated by the dashed line in the top and bottom half respectively. Abbreviations: CC (colorectal cancer), COPD (chronic obstructive pulmonary disease).

24
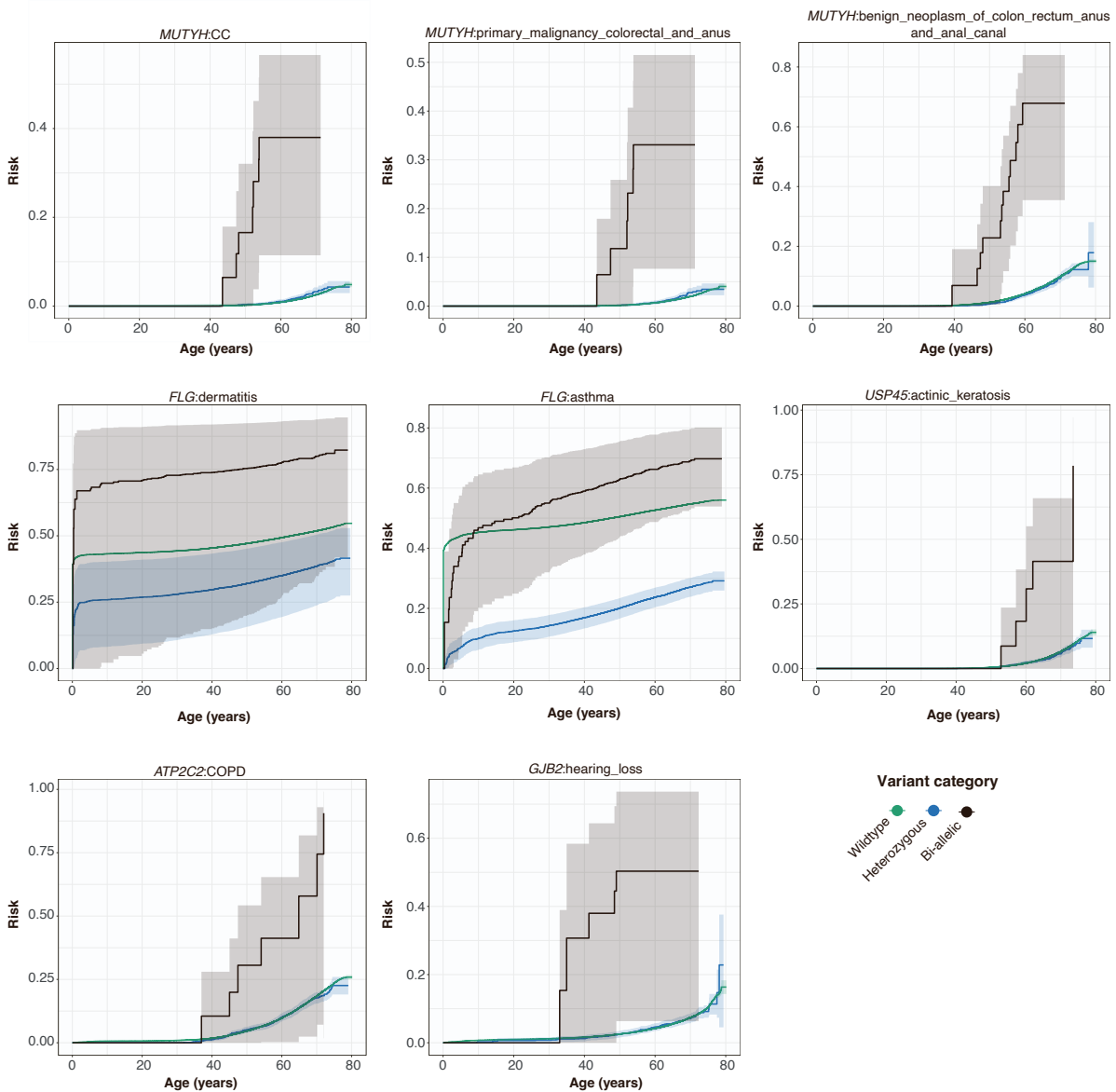
Figure S22: **Kaplan-Meier survival curves for carriers of bi-allelic variants, related to Figure 5**. Trajectories for wildtypes and bi-allelic (CH or homozygous) carriers of damaging missense/protein-altering mutations are shown with green and black lines respectively. For traits where over 50% of 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.

25
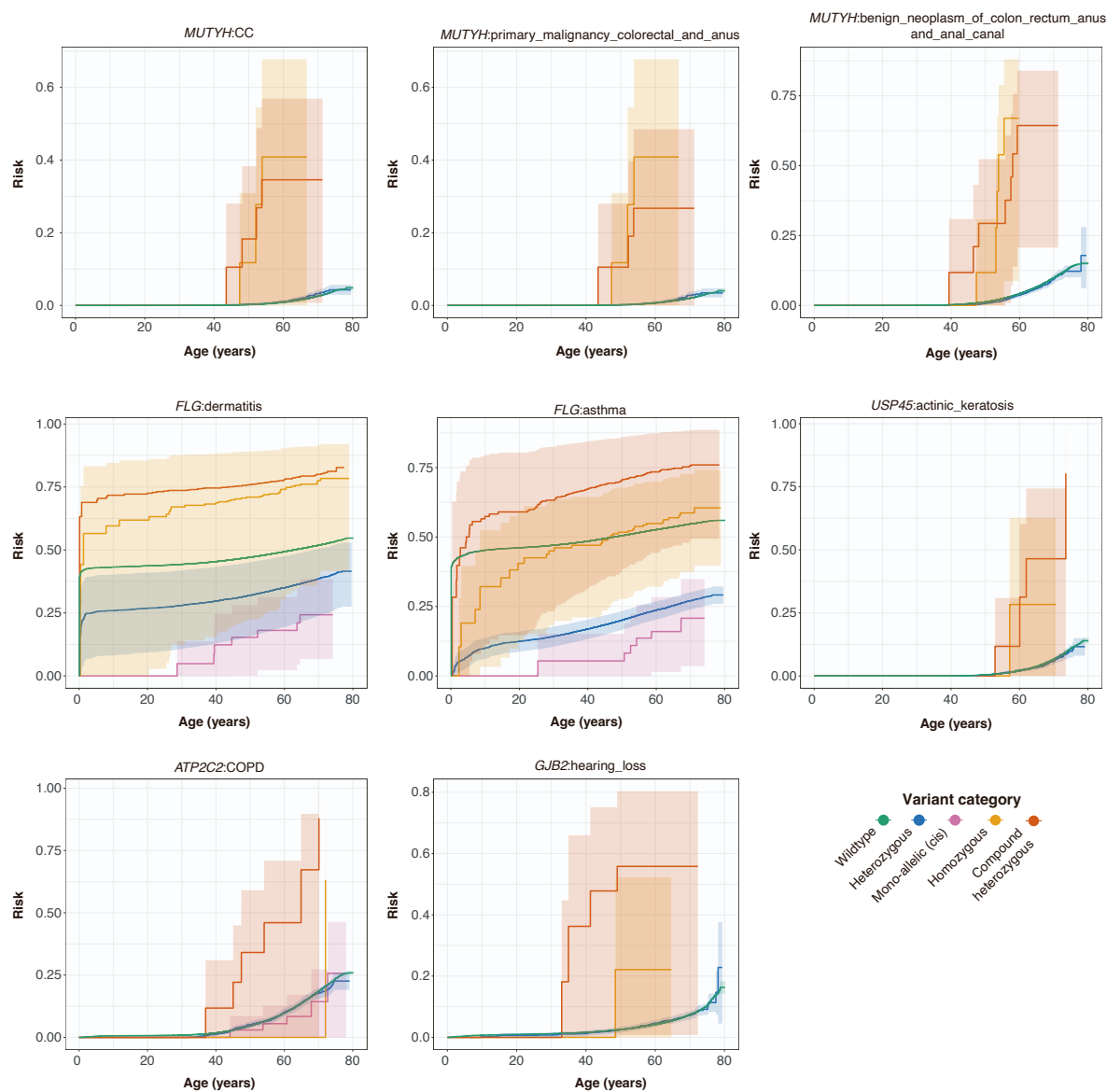
Figure S23: **Kaplan-Meier survival curves for carriers of CH, homozygous, heterozygous variants, related to Figure 5**. Kaplan-Meier survival curves for CH (red), homozygous (orange), heterozygous carriers (blue), single disruption of haplotypes (pink) owed to pLoF or damaging missense/protein-altering 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.
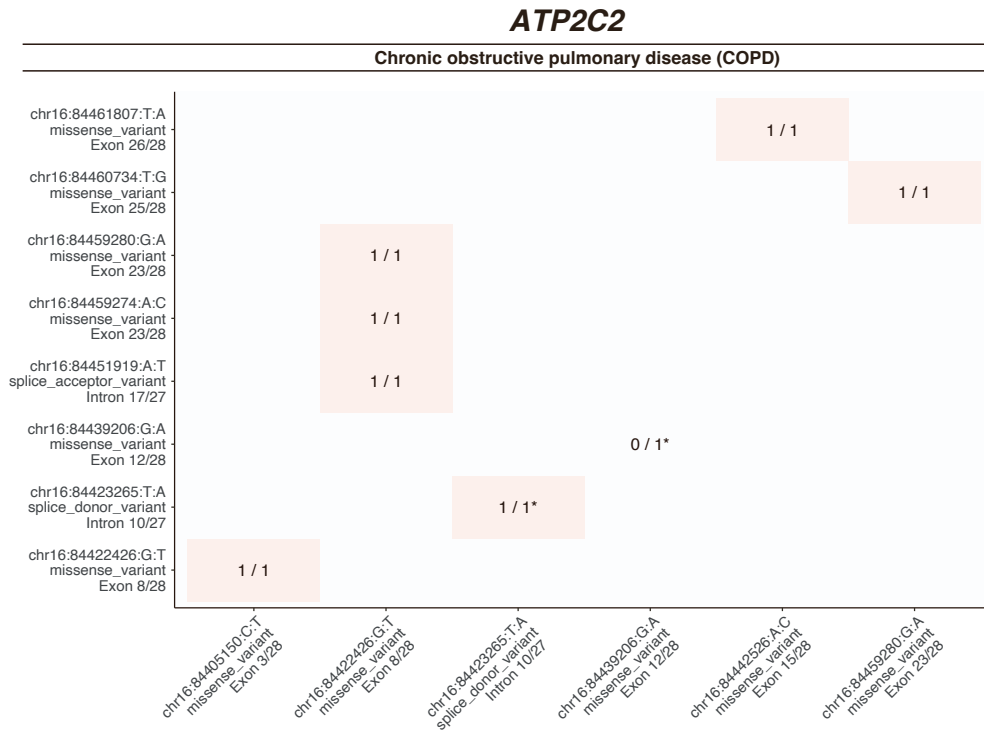
26

Figure S24: **Co-occurrence of deleterious *ATP2C2* variants by COPD status, related to Figure 5**. 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.
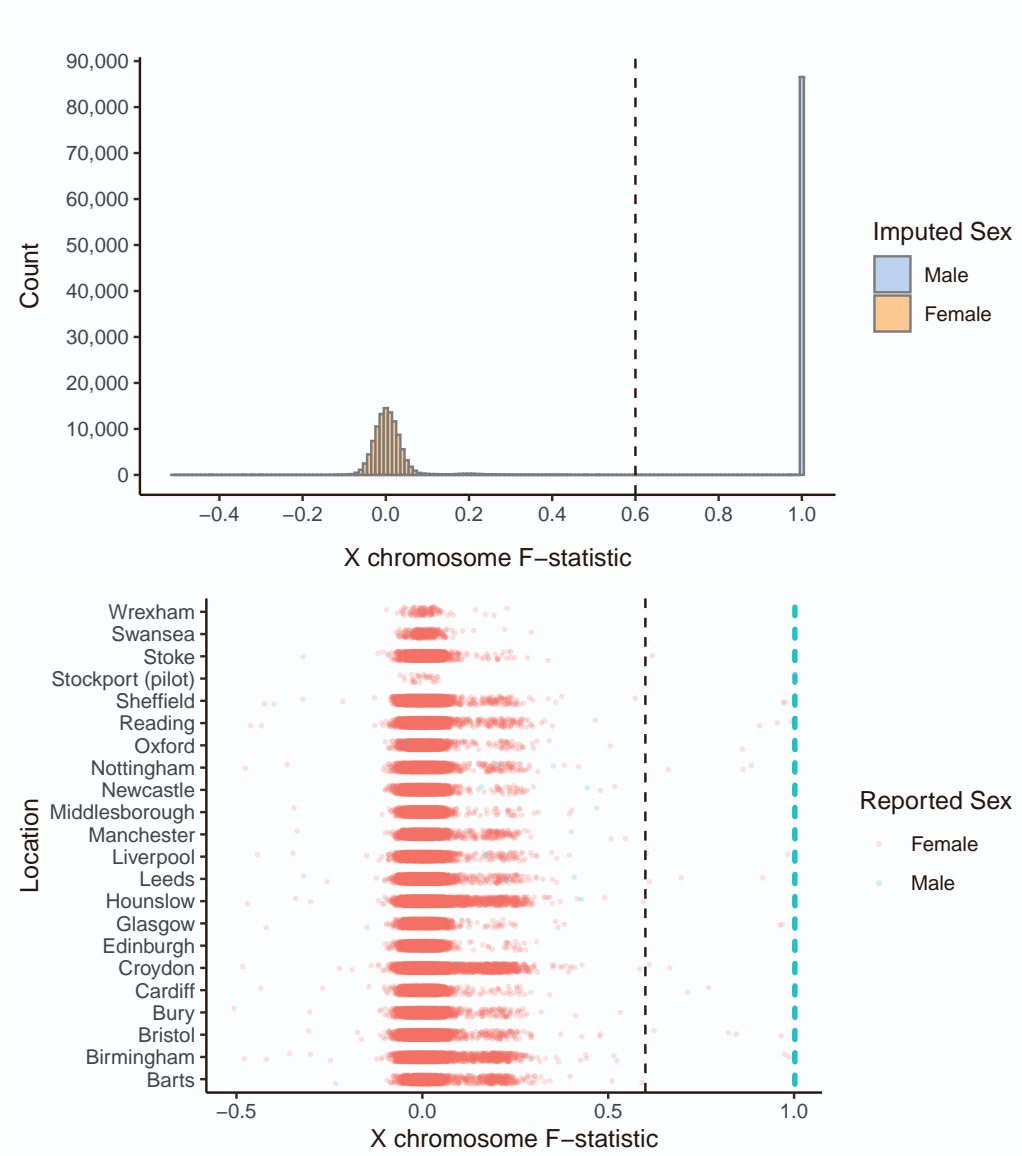
Figure S25: **Histogram and scatter-plots of X chromosome F-statistic by collection, related to Star Methods**. Samples lying to the left and right of the dashed line were called as female and male respectively, according to the imputed sex colorings in the upper histogram. Reported sex, split by UKBB 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.
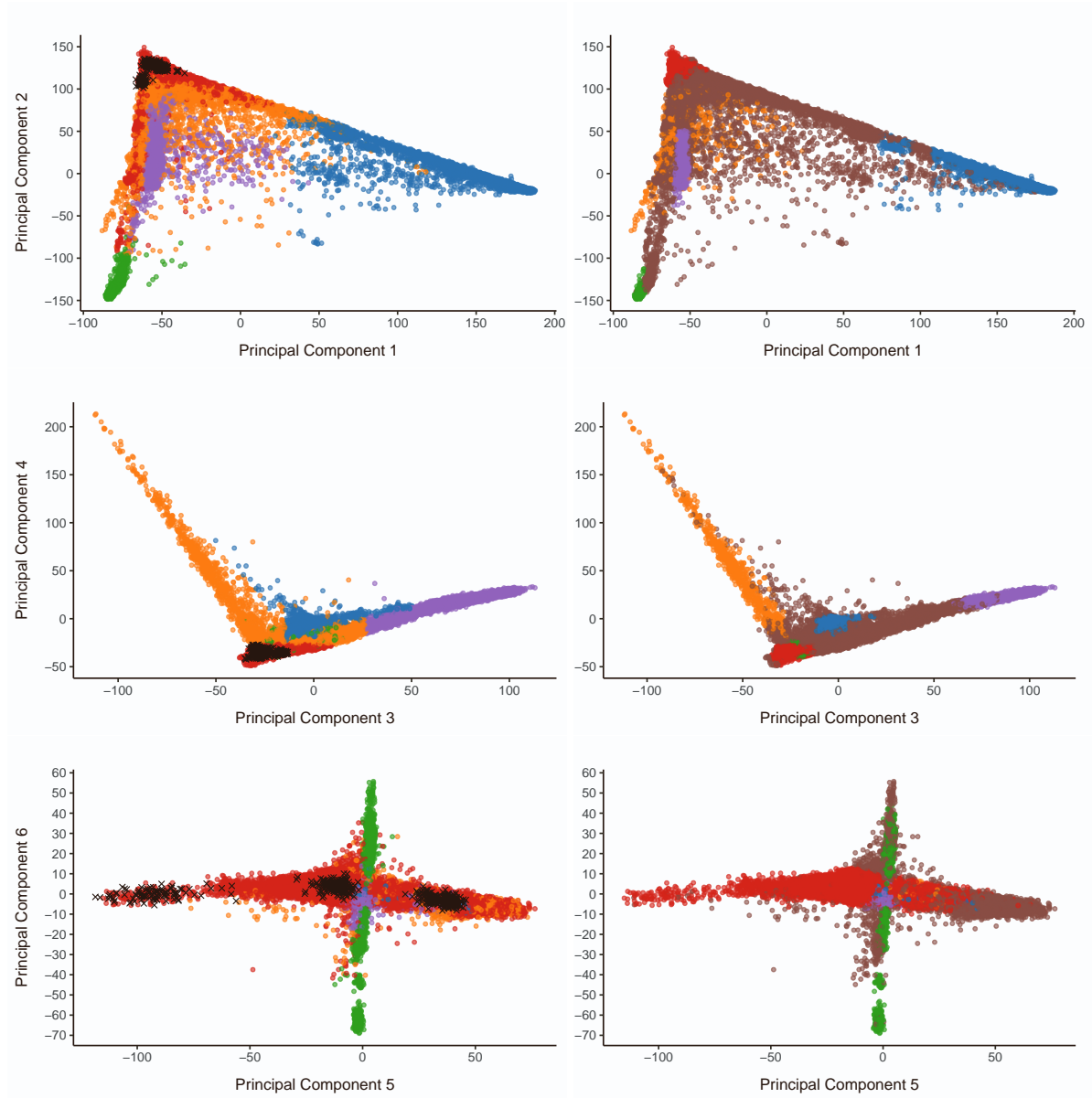
Figure S26: **Scatter-plots of PCs of UKBB genotype data projected into the PC space defined by 1KGP samples, related to Star Methods**. Points are colored according to sample collection, with 1KGP samples colored in blue. 1KGP super-populations labels were used to train a random forest classifier.
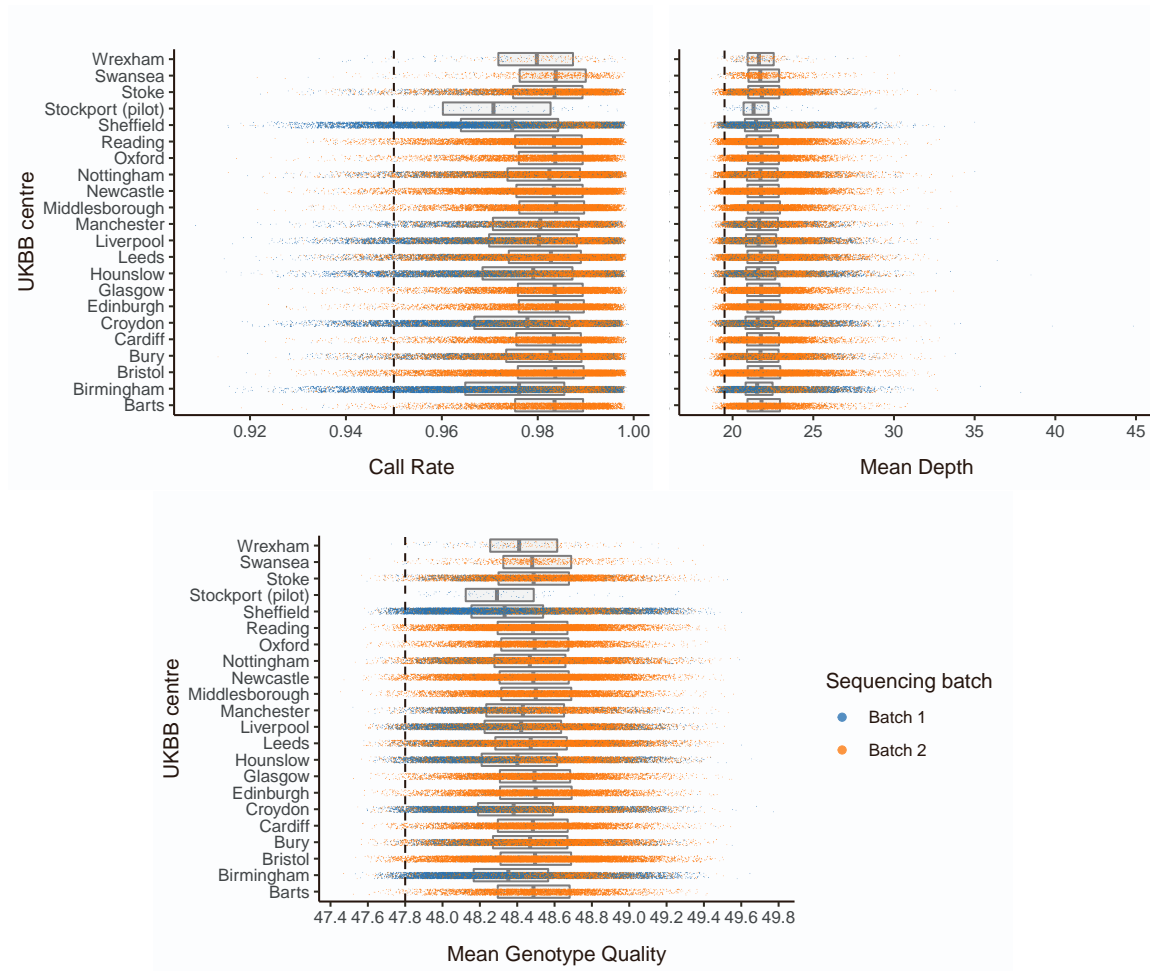
Figure S27: **Distributions of sample metrics following initial restriction to variants, lying outside LCRs and inside the padded (50 bp) target intervals, and prior to the initial hard sample filters (call rate > 0.95, mean depth > 19.5, mean GQ > 47.8), related to Star Methods**. In each plot, jittered 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.
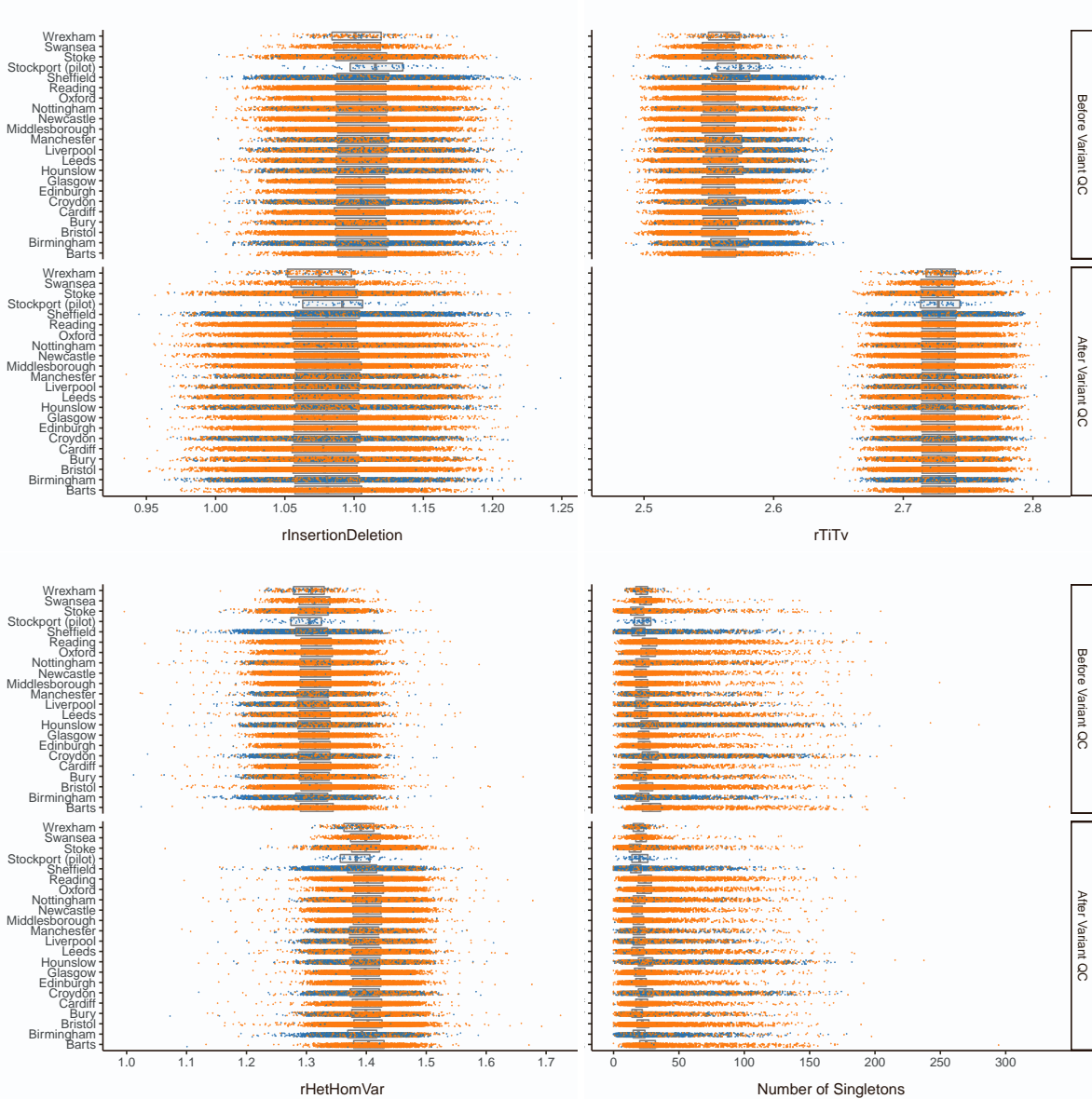
Figure S28: **Distributions of variant metrics before and after the removal of invariant sites, variants with call rate < 0.97, and variants out of HWE ($P < 1 \times 10^{-6}$), related to Star Methods**. In each plot, jittered scatters display the distribution for each sequencing batch colored by sequencing 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.