SUPPLEMENTARY INFORMATION

Supplementary Note

1 Imputation accuracy benchmarks

To compare the accuracy of genotype imputation using the UK Biobank *N*=50K exome sequencing call set to the accuracy of the latest UK Biobank imputation release (imp_v3, which used the Haplotype Reference Consortium (HRC) and UK10K/1000G reference panels), we computed squared correlations between imputed genotype dosages and direct genotype calls from the UK Biobank genotyping array (on the full cohort) and from whole-exome sequencing (on *N*=50K samples).

To benchmark the accuracy of the HRC+UK10K/1000G-based imp_v3 data set, we computed R^2 between imp_v3 dosages and WES-based genotype calls at biallelic SNPs with very high genotyping rates in the WES call set (missingness<0.0005, indicating high confidence in the WES genotypes). We restricted this benchmark to *N*=46,476 exome-sequenced individuals who reported European ancestry.

To benchmark imputation accuracy using the *N*=50K WES samples as a reference panel, we computed R^2 between imputed dosages and array-based genotype calls at well-typed biallelic SNPs (which were not included in the imputation scaffold). We identified a subset of well-typed SNPs to use as a gold standard by requiring R^2>0.999 concordance between the array-based and WES-based genotypes among the exome-sequenced individuals; we then benchmarked imputation R^2 in the *N*=412,764 remaining (unsequenced) individuals of European ancestry.

Restricting to a subset of well-typed SNPs was necessary to perform this benchmark because the UK Biobank genotyping array has imperfect accuracy, particularly for rarer SNPs[1]. For common SNPs, our requirement of R^2>0.999 vs. WES (within the exome-sequenced cohort) was enough to ensure high genotyping accuracy for common and somewhat rare SNPs; however, for ultra-rare SNPs with MAF<0.0001 (i.e., MAC<10 among the WES cohort), an imperfectly typed SNP could still achieve perfect concordance within the WES cohort by chance (despite a non-negligible level of genotyping error in the rest of the cohort). This phenomenon should produce a slight downward bias in our imputation accuracy benchmark for very rare SNPs due to error in the array-based gold standard. Consistent with this expectation, when we further restricted our benchmark to a small subset of SNPs previously determined (by inspection of SNP cluster plots) to have high-quality array genotypes[1], we observed evidence for a slight increase in estimated imputation accuracy (**Supplementary Table 1**), suggesting that our primary benchmark was indeed slightly conservative.

Finally, for reference, we also benchmarked accuracy of the array-based genotypes against WES-based genotypes at high-confidence biallelic SNPs. As in our benchmark of the imp_v3 data, we restricted to SNPs with WES missingness<0.0005 and computed R^2 among *N*=46,476 exome-sequenced individuals who reported European ancestry (**Supplementary Table 1**).

2 Robustness of rare variant association analyses to population stratification

Genome-wide association analysis of common and low-frequency variants using regression with genetic principal component (PC) covariates is generally accepted to be robust to population stratification[2], and linear mixed model (LMM) analysis additionally corrects for confounding from sample relatedness[3]. However, the extent to which these now-standard approaches produce robust associations when used to analyze very rare variants is less well-understood, with some concerns arising from a key paper of Mathieson and McVean (2012)[4] that simulated scenarios of extreme stratification in which rare variants escaped correction from analyses that used either PCs or LMMs. Recent work exploring subtle population structure in the UK Biobank cohort has also caused some general concern about potential uncorrected effects of stratification on epidemiological analyses[5]; however, this work focused on aggregate effects of common variants in analytical frameworks very different from rare variant association analysis.

Given our focus on identifying very rare coding variants influencing quantitative traits, we revisited the theoretical basis for rare variant stratification that could escape PC/LMM correction, and we also performed additional supporting analyses to verify that our association results were robust to potential confounding structure.

First, on a theoretical level, the type of population stratification necessary to produce false positive rare variant associations is very different from the type of stratification that confounds naïve common variant association analyses. The latter form of stratification commonly manifests as a weak correlation between genetic ancestry and environmental effects on a phenotype; when GWAS sample size is sufficiently large, such correlations create significant (false-positive) associations at ancestry-informative common variants. In contrast, rare variant stratification requires environmental effects that are both much stronger in magnitude and highly localized in a manner that matches geographical localization of rare alleles (because effect size estimates for rare variants have much wider error bars, so strong environmental deviations are needed to appreciably inflate significance). Indeed, Mathieson and McVean (2012) observed exactly this behavior: in the context of broad, smoothly varying environmental confounding – which is typically observed in GWAS – rare variants actually exhibited less confounding than common variants. The simulations that produced rare variant confounding involved sharp, highly localized effects in which mean phenotypes were locally shifted by 1 to 2 s.d. Such extreme effects (which would correspond to environmental effects that modify height by ~5 inches, for example) seem unlikely to exist for most phenotypes in most cohorts. Moreover, even if such strong, sharp stratification were to exist in UK Biobank, it would be ameliorated by geographical covariates (such as assessment center) that we included in our analyses.

To confirm the above intuition, we repeated our association analyses using the same statistical approach (BOLT-LMM with covariates including 20 PCs and assessment center) but restricting to a genetically homogeneous, unrelated (at third-degree or closer) subset of 337,539 white British participants (with ancestry confirmed by principal component analysis[6]). While this subset of participants is not completely free of population structure, we reasoned that any effects of uncorrected confounding would at least begin to manifest as differences in analytical results

between our primary analyses (which included all 459,327 self-reported white individuals) and the restricted analyses.

Across the 1,189 rare coding variant associations our primary analyses identified as likely-causal, the key statistical properties of these associations – minor allele frequencies, estimated effect sizes, and *P*-values (adjusted for sample size) – were all extremely consistent between the full and restricted analyses. Nearly all variants had similar minor allele frequencies in the two sample subsets: only 9 variants (involved in 12 associations) exhibited >2-fold differences in MAF (**Supplementary Fig. 3a**). All 9 of these variants were enriched in the Ashkenazi Jewish population[7], explaining their much lower allele frequencies in the white British sub-cohort, and 8 of the 9 variants modified genes clearly related to the associated traits (*GPT*, *ALPL*, *ABCA1*, *SCARB1*, *SHBG*, *PDZK1*, and *TUBB1*). Across all associations, estimated effect sizes were highly consistent between the full and restricted analyses (R^2 = 0.985), showing no evidence of diminished effects within the restricted cohort (regression slope = 1.00 (0.99 – 1.01); **Supplementary Fig. 3b**) (which would be expected if some associations were driven by confounding structure). Association *P*-values were also highly consistent between the full and restricted analyses (R^2 = 0.998 for $-\log_{10}$ *P*-values), with most associations (79%) still reaching genome-wide significance ($P < 5 \times 10^{-8}$) in the restricted cohort and nearly all associations (94%) reaching $P < 3 \times 10^{-6}$, the sample-size-adjusted threshold corresponding to our $P < 5 \times 10^{-8}$ threshold in the full cohort (**Supplementary Fig. 3c**).

We further assessed the extent to which likely-causal rare coding variants exhibited geographical localization by comparing the birth coordinate distributions of carriers of likely-causal rare coding variants to the birth coordinate distributions of carriers of rare coding variants from an allele-frequency-matched distribution of "background variants" (**Methods**). We determined that likely-causal rare coding variants were no more geographically localized than background variants (**Supplementary Fig. 4**): among likely-causal variants, the mean of the standard deviation of east (respectively, north) birth coordinates of carriers was 75.6 km (respectively, 151.2 km), which almost exactly matched the corresponding measures of geographical localization for the allele-frequency matched background variants (75.7 km and 150.6 km, respectively).

Together with our replication analyses showing that the effect signs of associations we identified replicated in previous exome array data sets (**Supplementary Table 4**), these lines of evidence indicate that our rare variant association analyses were robust to effects of sample structure.

3 Identification of additional independently-associated rare coding variants in genes containing multiple likely-causal variants

In each gene in which our primary analysis pipeline identified multiple likely-causal rare coding variants for a trait, we searched for additional rare protein-altering variants that did not reach the stringent significance thresholds used in our primary analyses but nonetheless exhibited good evidence for being trait-altering. As summarized in **Methods**, this secondary analysis pipeline involved two runs of FINEMAP followed by evaluation of statistical significance at an FDR<0.05 threshold. The details of these steps are as follows.

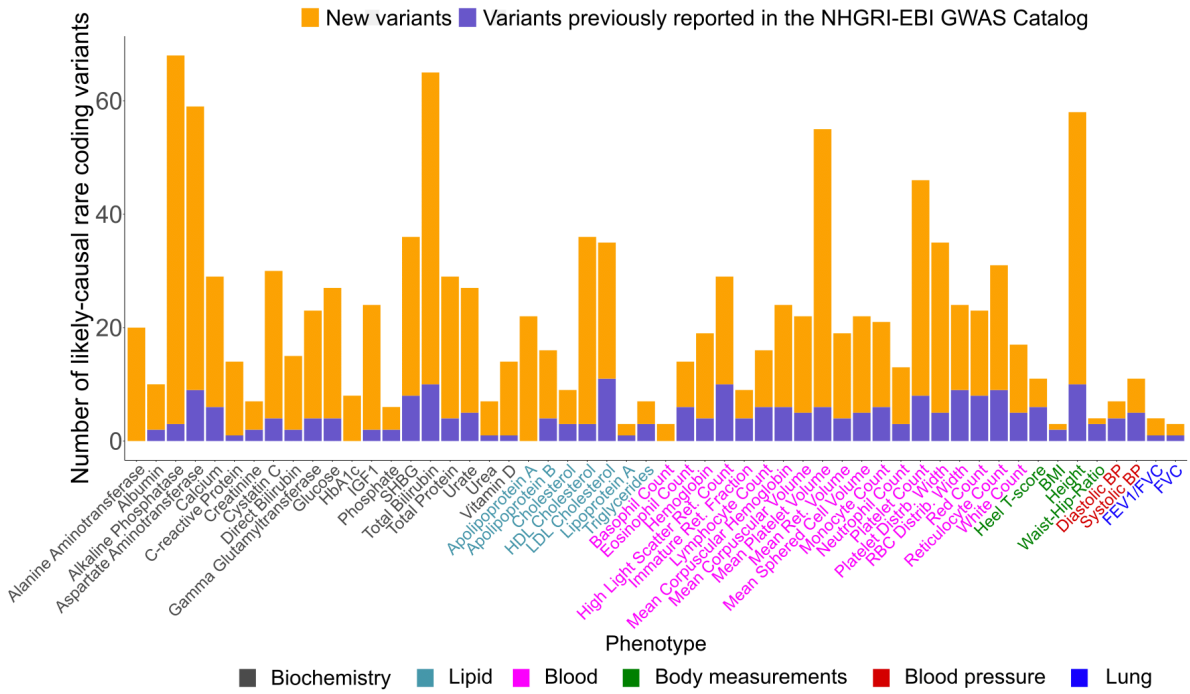
1. Run a first round of FINEMAP on (i) common and low-frequency (MAF>0.001) HRC/UK10K-imputed variants within 1Mb of the gene that associated with the trait at genome-wide significance; together with (ii) all WES-imputed coding variants in the gene (with no restrictions on CADD or significance). We allowed FINEMAP to select up to 15 causal variants (to keep computational cost reasonable; the largest job took ~10h and ~90GB RAM). This round of analysis was primarily intended to identify a subset of variants that captured the bulk of the common variant association signal so that we could evaluate rare coding variant association signals after conditioning on these variants (in round 2 below). Round 1 also sometimes identified rare coding variant associations that clearly become non-significant after conditioning on other variants (i.e., P>0.05 in the top configuration in which the variant appeared); these variants were flagged to drop from round 2.

2. Run a second round of FINEMAP on (i) putatively causal variants selected from round 1; together with (ii) non-dropped WES-imputed coding variants (with no restrictions on CADD or significance), using stepwise conditional analysis (FINEMAP --cond instead of --sss) and using a flat prior on whether or not variants are causal (via --prior-k). The top configuration in the output of this analysis represented a series of conditionally independent associations, and the joint-model betas and standard errors for this configuration provided conditional *P*-values for variants in this series.

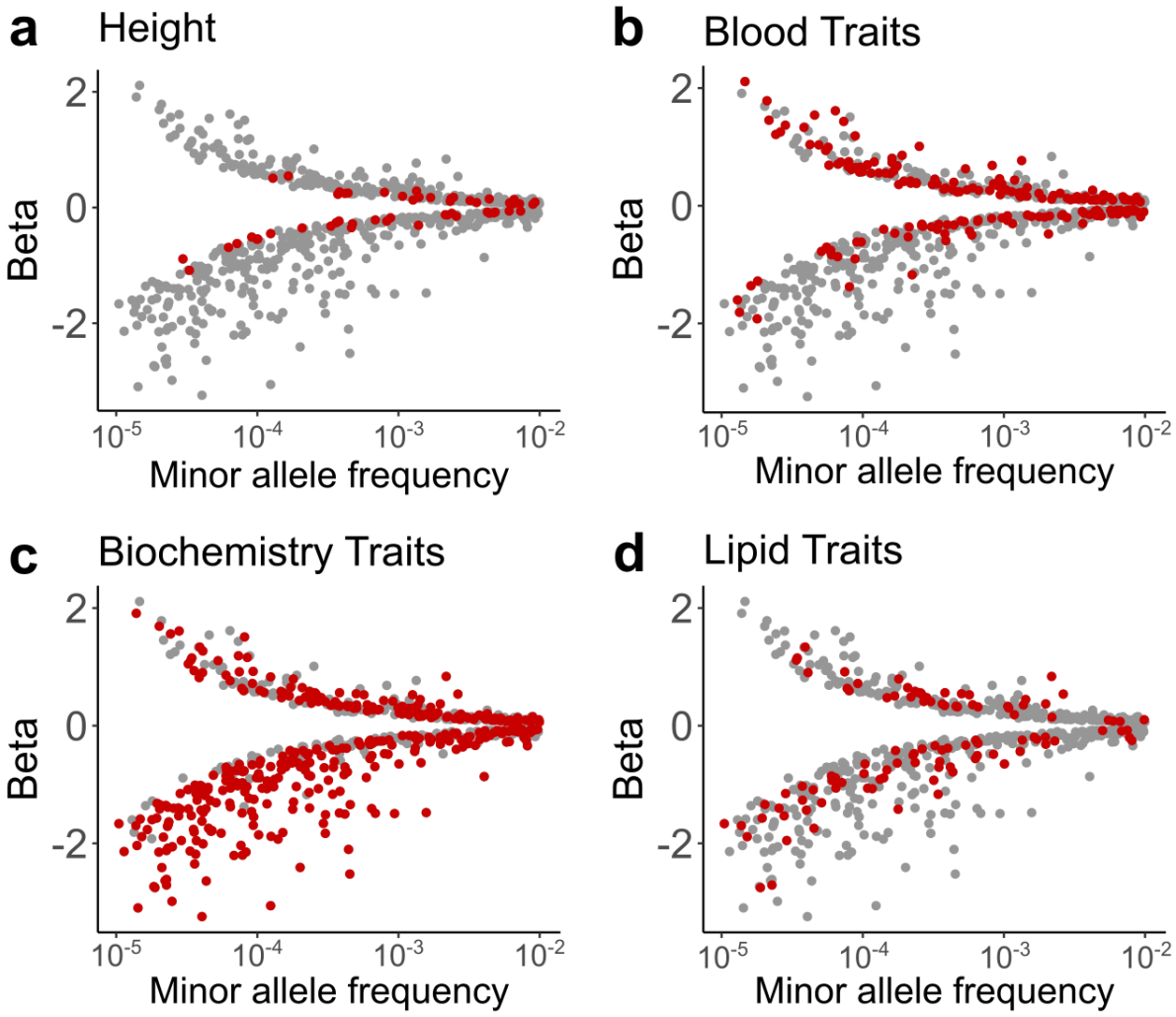
3. To determine which variants pass FDR<0.05, compute *q*-values for all WES-imputed coding variants for the gene after (i) setting the *P*-value of each variant in the configuration to the maximum of its conditional *P*-value and original (marginal) *P*-value (to be conservative); and (ii) setting the *P*-value of each variant not in the top configuration to 1 (since these variants were eliminated from consideration for causality). Because most allelic series involved variants with trait-modifying effects predominantly in one direction (either positive or negative), we assessed FDR<0.05 independently for variants with effects in each direction (so as not to allow the existence of many associations in one direction to reduce the significance threshold for associations in the opposite direction).

References

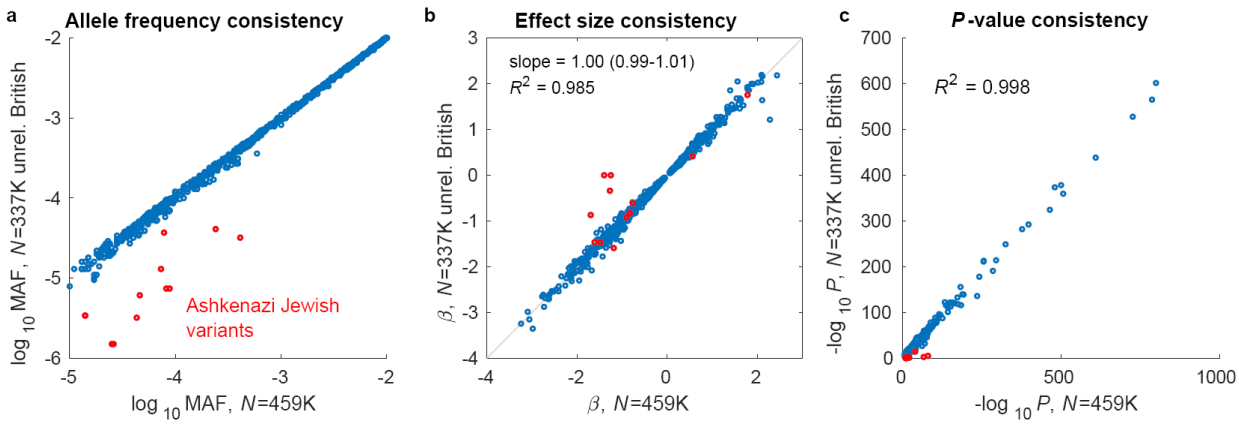
1. Wright, C. F. *et al.* Assessing the Pathogenicity, Penetrance, and Expressivity of Putative Disease-Causing Variants in a Population Setting. *Am. J. Hum. Genet.* **104**, 275–286 (2019).
2. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
3. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**, 203–208 (2006).
4. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* **44**, 243–246 (2012).
5. Haworth, S. *et al.* Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nat. Commun.* **10**, 333 (2019).
6. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
7. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).



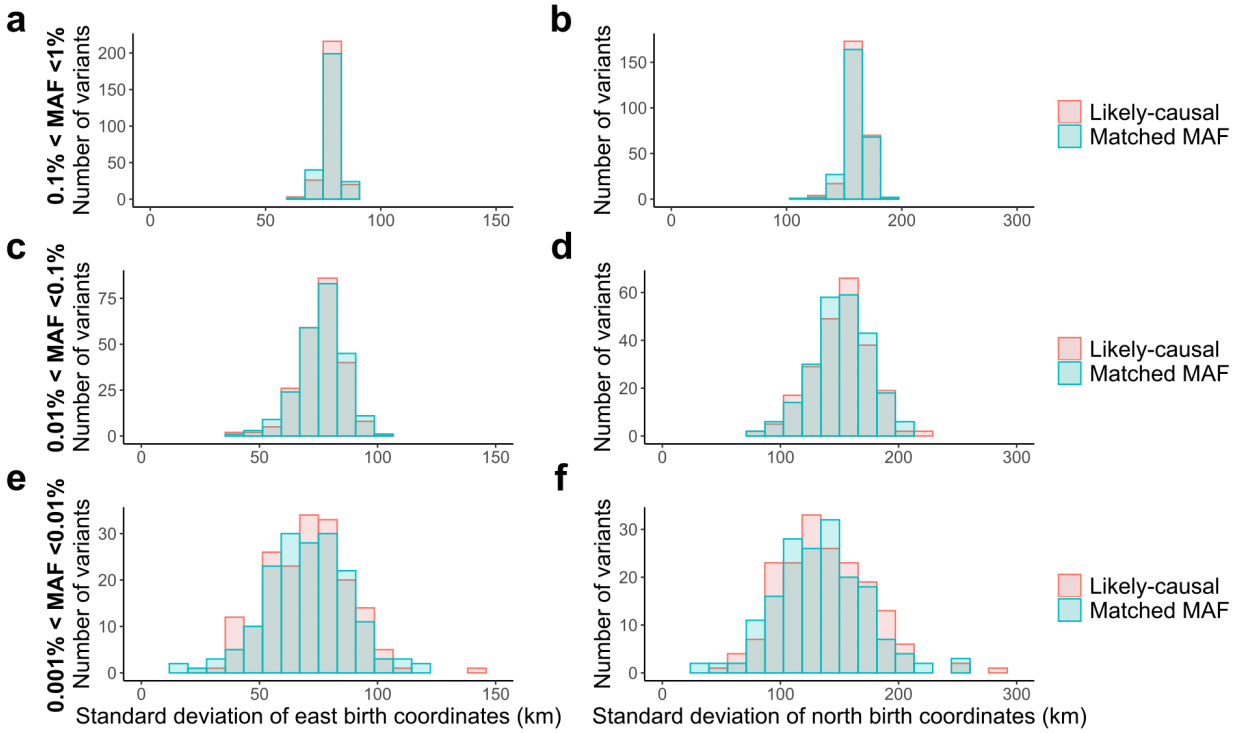
Supplementary Figure 1. Most likely-causal rare coding variant associations identified by whole-exome imputation in UK Biobank were not in the GWAS Catalog. For each trait, we tabulated whether each likely-causal variant was previously reported in the NHGRI-EBI GWAS catalog as associated with any trait (so as to be maximally conservative with respect to the possibility that trait names in the GWAS Catalog might differ from those in UK Biobank).



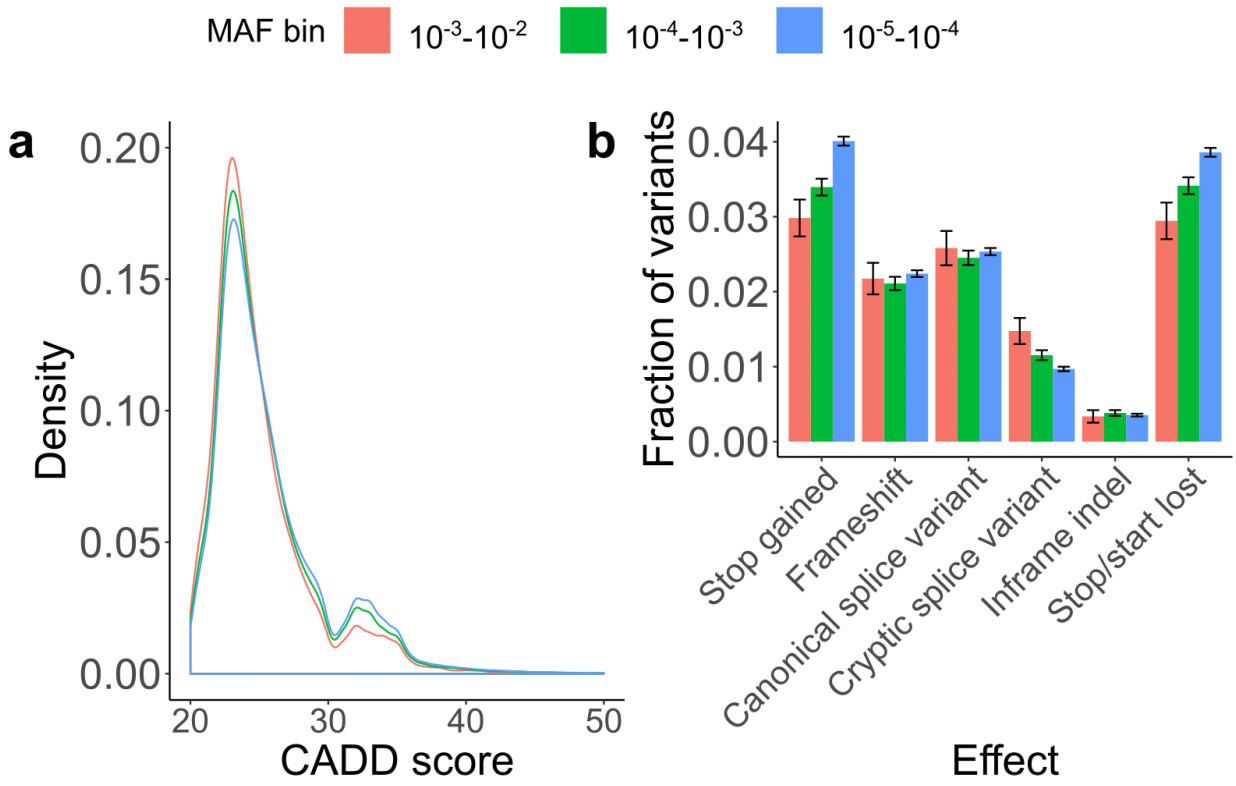
Supplementary Figure 2. Magnitudes of effect sizes of likely-causal rare coding variants generally increase with decreasing minor allele frequency. Gray dots represent the full set of 1,189 likely-causal coding associations with red dots highlighting this trend for specifically (**a**) height, (**b**) all blood traits, (**c**) all biochemistry traits, and (**d**) all lipid traits.



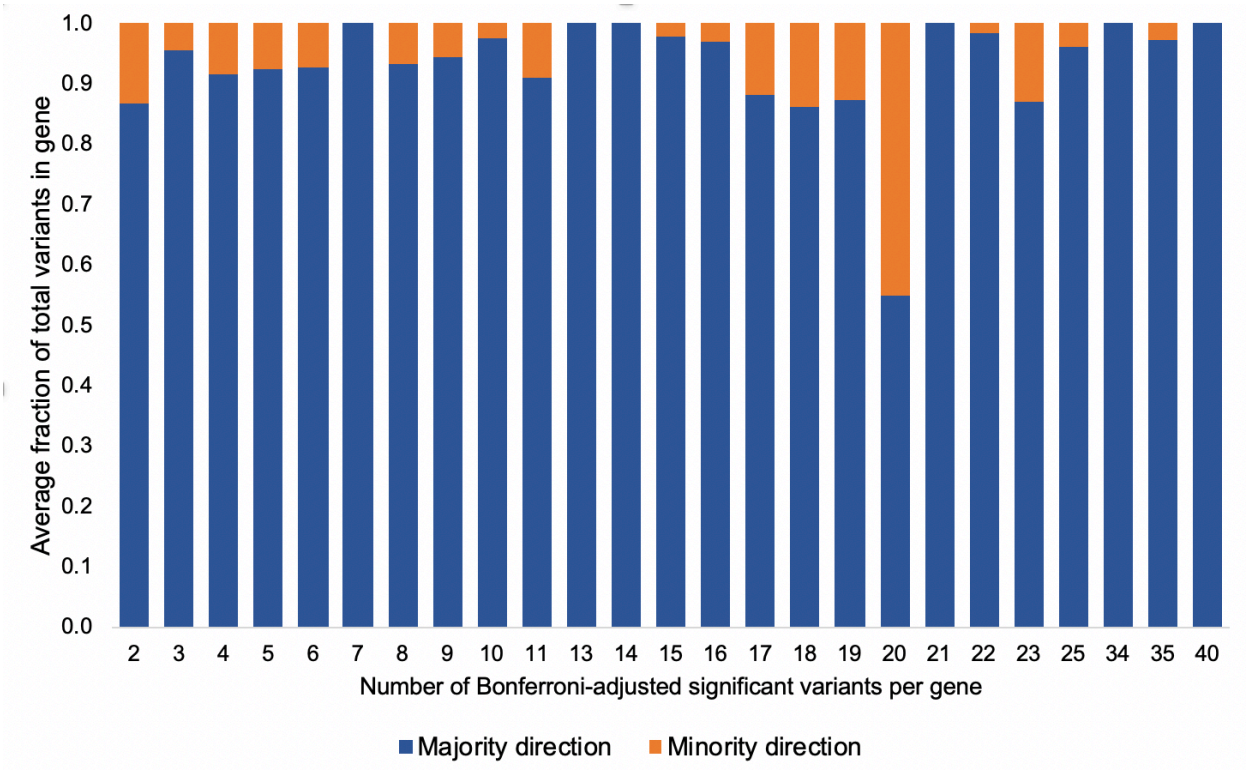
Supplementary Figure 3. Rare variant association tests using linear mixed models show no evidence of confounding from sample structure within UK Biobank. (**a**) Allele frequencies, (**b**) effect size estimates, and (**c**) association *P*-values are all highly consistent between our primary analyses, which included all *N*=459,327 UK Biobank participants of European ancestry, and analyses restricted to a subset of *N*=337,539 unrelated British participants. The only notable outliers were a few very rare variants found much more frequently in Ashkenazi Jewish individuals than the rest of the UK Biobank cohort; nearly all of these variants affected genes known to be relevant to the associated traits (**Supplementary Note**).



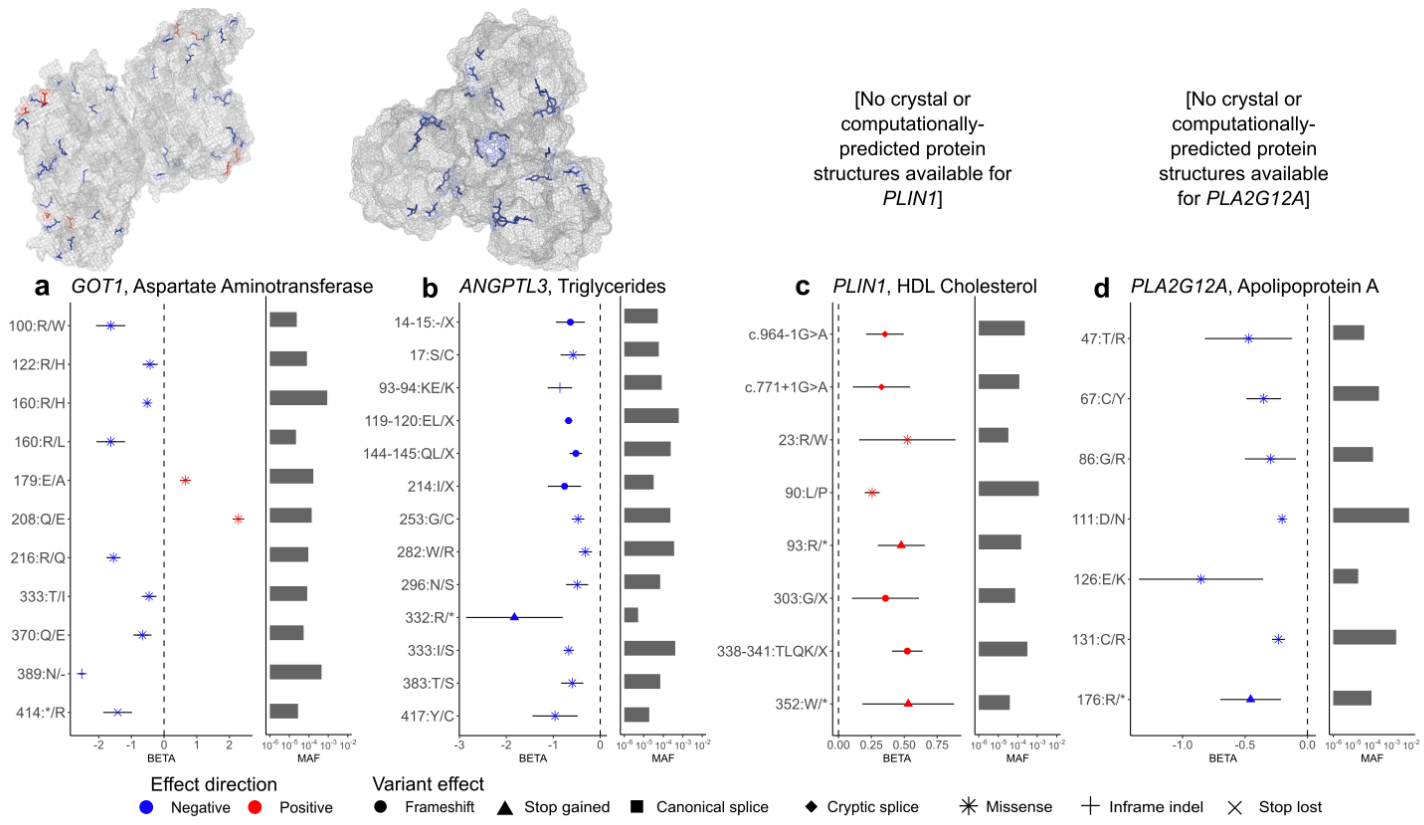
Supplementary Figure 4. Likely-causal rare coding variants are no more geographically localized than allele frequency-matched background variants. For each likely-causal variant, and for a MAF-matched set of background variants, we computed the standard deviation of east (respectively north) birth coordinates among carriers of the rare allele. The plotted histograms compare birth coordinate variation between likely-causal variants vs. background variants, stratified by MAF range: (**a**,**b**) 0.1% < MAF < 1%, (**c**,**d**) 0.01% < MAF < 0.1%, and (**e**,**f**) 0.001% < MAF < 0.01%. Both likely-causal and background variants exhibit the expected trend of decreasing birth coordinate variation (i.e., increasing geographical localization) with decreasing minor allele frequency.



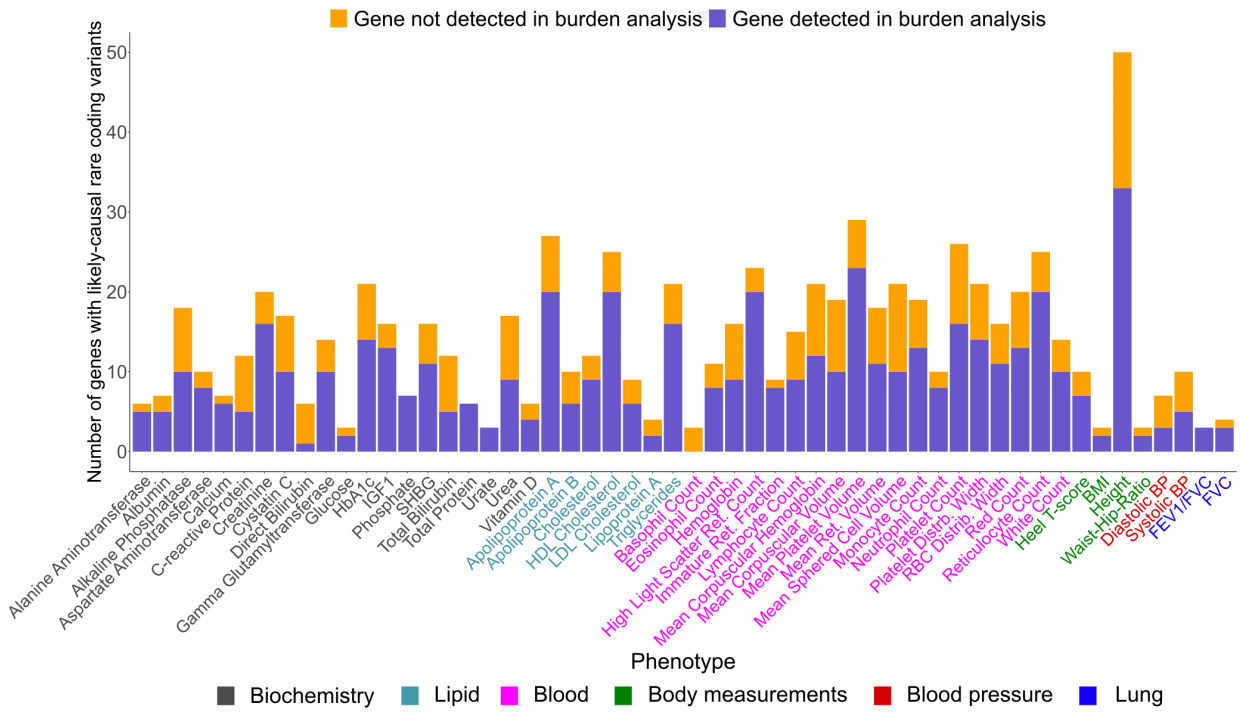
Supplementary Figure 5. Measures of deleteriousness among protein-altering variants increase modestly with decreasing minor allele frequency. (**a**) CADD score. (**b**) predicted protein alteration from VEP or SpliceAI (for cryptic splice sites). Distributions are across all protein-altering variants present in the UKB *N*=50K whole-exome sequencing genotype call set with European minor allele frequencies within the indicated tranches. Error bars, 95% CIs.



Supplementary Figure 6. Concordance in effect directions of rare coding variants within the same gene. Gene-trait pairs were stratified by the number of independent rare coding variant associations identified (in follow-up analyses that relaxed the significance threshold to Bonferroni-adjusted $P<0.05$, correcting for the number of coding variants within each gene; **Methods**); these strata are indicated in the *x*-axis of the figure. Across gene-trait pairs with an allelic series of a given length, we computed the average fraction of variants with effect directions in the majority effect direction. For this assessment we analyzed allelic series determined using a Bonferroni-adjusted $P<0.05$ significance threshold rather than FDR<0.05 (which we had applied independently to determine significance thresholds for variants in each gene with positive vs. negative effects) to avoid bias in directional concordance due to differing significance thresholds for positive vs. negative effects.



Supplementary Figure 7. Allelic series of trait-associated rare coding variants in *GOT1*, *ANGPTL3*, *PLIN1,* and *PLA2G12A*. Statistically independent associations (reaching FDR<0.05 significance) for: (**a**) *GOT1* and aspartate aminotransferase, (**b**) *ANGPTL3* and triglycerides, (**c**) *PLIN1* and HDL cholesterol, and (**d**) *PLA2G12A* and apolipoprotein A. Top, protein structures with altered amino acids (modified by missense variants) color-coded by effect direction (red for trait-increasing variants and blue for trait-decreasing variants). Bottom, per-variant effect sizes (error bars, 95% CIs) and allele frequencies. Protein structures were previously determined experimentally (for GOT1 and ANGPTL3). The structure for GOT1 represents a homodimer and for ANGPTL3 a homotrimer of the fibrinogen-like domain only.



Supplementary Figure 8. Single-variant association analysis discovers likely-causal rare coding variants in genes not identified by gene burden analysis. For each trait, genes containing at least one likely-casual rare coding variant were tabulated according to whether or not they reached significance in a gene burden analysis using filtering criteria of CADD $\geq$ 20 and MAF $\leq$ 0.01 for inclusion of variants in the burden test.