nature portfolio

https://doi.org/10.1038/s41588-021-00892-1

Supplementary information

Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses

In the format provided by the authors and unedited

SUPPLEMENTARY INFORMATION

Supplementary Note

1 Imputation accuracy and imputation panel coverage 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 on an additional *N*=150K exomes subsequently released by UK Biobank (restricting to *N*=141,255 individuals who reported European ancestry, belonged to the imp_v3 data set, and had not withdrawn from the study).

In the benchmark of WES imputation accuracy, we considered all coding or splice SNPs called in both the *N*=50K exome reference panel and the subsequent exome data release. In the benchmark of imp_v3 imputation accuracy, we considered all coding or splice SNPs that were biallelic in imp_v3 and had not been directly genotyped on the UK Biobank SNP-arrays. We restricted the benchmarks to SNPs with at most a 1% missing rate in the *N*=141,255 WES validation data set, and we also required allele frequencies in the imputed vs. validation data sets to differ by at most 0.1 (to exclude likely genotyping errors in either the reference panel or the validation data set).

We further benchmarked imputation accuracy of the TOPMed imputation panel (r2) by imputing the same validation set of *N*=141,255 UK Biobank samples using the TOPMed imputation server (restricting to chromosome 19, as the server required ~4 days to complete this computation). We assessed accuracy on the SNP set above (for all such SNPs present in the TOPMed imputation results).

We assessed the coverage of all three imputation panels (HRC + UK10K/1000G, *N*=50K WES, and TOPMed) based on SNP calls in the *N*=200K WES release provided by UK Biobank. To create a robust “gold standard” SNP set suitable for assessing coverage of all panels, we restricted to chr19 SNPs that passed the following filters: (1) lifted uniquely to hg19, (2) fell within exome capture targets, (3) passed Hardy-Weinberg equilibrium (*P*>1e-9) in European-ancestry samples, (4) missing rate <1%. We assessed coverage in MAF ranges defined using MAF in the subset of European-ancestry samples (*N*=188,262 samples).

2 Assessment of novelty of genes implicated in likely-causal coding variant associations with blood traits and height

To explore the extent to which our rare coding variant association analyses implicated genes not previously associated with the analyzed traits, we compared our results to data from two recent studies of blood traits and height. In general, assessing whether a gene-trait association is novel requires extensive literature review. We therefore decided to focus our analyses on blood trait and height associations as for both of these sets of traits, we could find a recent study of a largest-to-date data set that could serve as a reasonable proxy for prior knowledge.

First, we compared our blood trait results to those of Vuckovic *et al*. (2020)[1]. This study performed genome-wide association analysis of blood traits in UK Biobank (with replication in the BCX cohort; total *N*=563,085), allowing us to directly compare, for each blood trait, the genes implicated in conditionally independent associations from their study (Table S3) to the genes involved in likely-causal rare coding associations we identified. Explicitly, for results from both Vuckovic *et al*. and our analysis, we compiled a set of genes (for each blood trait) consisting of all genes implicated by at least one reported variant. (In our analysis of rare coding variants, each variant by definition modified a gene; for associations from Vuckovic *et al*., we used the nearby gene(s) listed in Table S3 of Vuckovic *et al*., which we expect tended to produce a conservatively large set of potentially-known genes.) We then compared the two sets of genes and considered a gene identified by our analyses to be novel if it did not appear in the gene set from Vuckovic *et al*. for the same trait. If the gene was present for a similar trait, we noted this observation but still considered the gene to be novel for the trait in question. Combining across all 19 blood traits we analyzed, approximately one-quarter (86 out of 337) of the gene-trait pairs implicated by likely-causal rare coding associations in our study were not previously reported in Vuckovic *et al*. (**Supplementary Table 5**).

Next, we compared our variants associated with height to the results of Marouli *et al*. (2017), mentioned earlier in our manuscript as part of our replication analysis[2]. The set of genes that we considered to be implicated in Marouli *et al*. consisted of all genes with at least one variant reported as significantly associated with height. Of the unique genes implicated by likely-causal associations we detected for height, ~45% (23 out of 51) were novel compared to those reported in Marouli *et al*. (**Supplementary Table 6**).

3 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[3], and linear mixed model (LMM) analysis additionally corrects for confounding from sample relatedness[4]. 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)[5] 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[6]; 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[7]). 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 association strengths – 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 at least 5-fold enriched in the Ashkenazi Jewish population vs. any other population in gnomAD[8], 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 7**), these lines of evidence indicate that our rare variant association analyses were robust to effects of sample structure.

4 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 (i.e., not flagged in step 1) 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).

5 Robustness of imputation to recurrent mutation

A potential concern when attempting to impute very rare variants carried by very few individuals within a reference panel is that if a variant arises independently multiple times (rather than being shared among carriers descended from a common ancestor), imputation may not be possible.

While we did not expect this issue to substantially impact imputation for most variants (given the low genome-wide rate of *de novo* mutations), we explored this question by performing a simulation in which we merged each pair of consecutive variants on chr19 in our benchmarking set (described in **Supplementary Note 1**) into a single "recurrent variant" and attempted to impute these recurrent variants. We compared imputation accuracy for these merged variants to imputation accuracy for the original variants (using the same benchmarking procedure as in **Supplementary Note 1**) and observed only a modest decrease in imputation accuracy (**Supplementary Fig. 9**), consistent with the expectation that the difficulty of imputing a merged variant with a given MAF should roughly merge the difficulties of imputing the constituent (lower-MAF) variants.

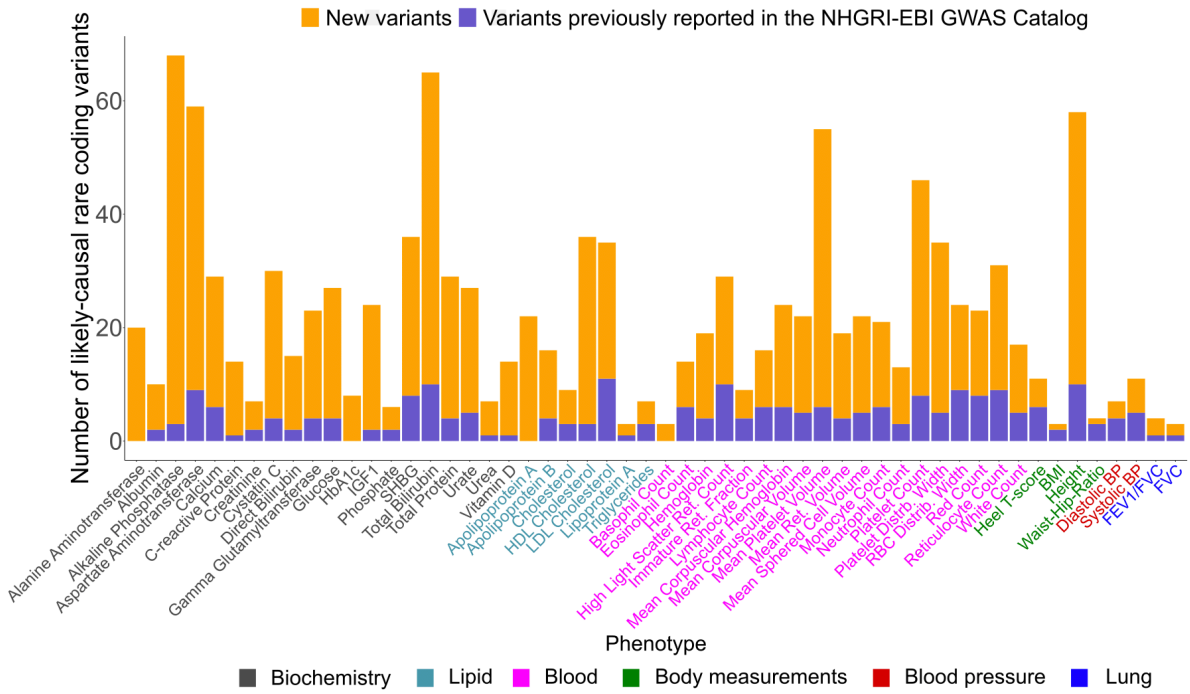
We also assessed the impact of recurrent mutation on our data empirically by using tsinfer[9] to infer coalescent trees at 9 randomly-sampled ultra-rare variants in our analysis. Specifically, using the "sample()" function in R, we randomly sampled 9 variants from the 181 ultra-rare variants (MAF < 10^{-4}) in our set of 675 unique likely-causal variants. For each of these variants, we generated a tsinfer[9] tree of genetic relatedness among both haplotypes of all exome-sequenced carriers of European ancestry (i.e., we included both the haplotypes containing the likely-causal coding variant and the haplotypes that did not, which served as "controls" randomly sampled from the population) (**Supplementary Fig. 10**). We ran tsinfer with default parameters using all SNP-array-genotyped, phased variants within ±5 Mb of the variant under consideration, and we extracted the tree at the position of the variant from the tree sequence inferred by tsinfer.

Of the 9 trees, all but one showed all haplotypes carrying the rare variant clustering together on the tree indicating their shared ancestry (**Supplementary Fig. 10**). In the remaining tree, the carrier haplotypes formed two clusters, suggesting that the rare variant may have arisen twice; however, both clusters contained at least two carriers, such that imputation was still possible.

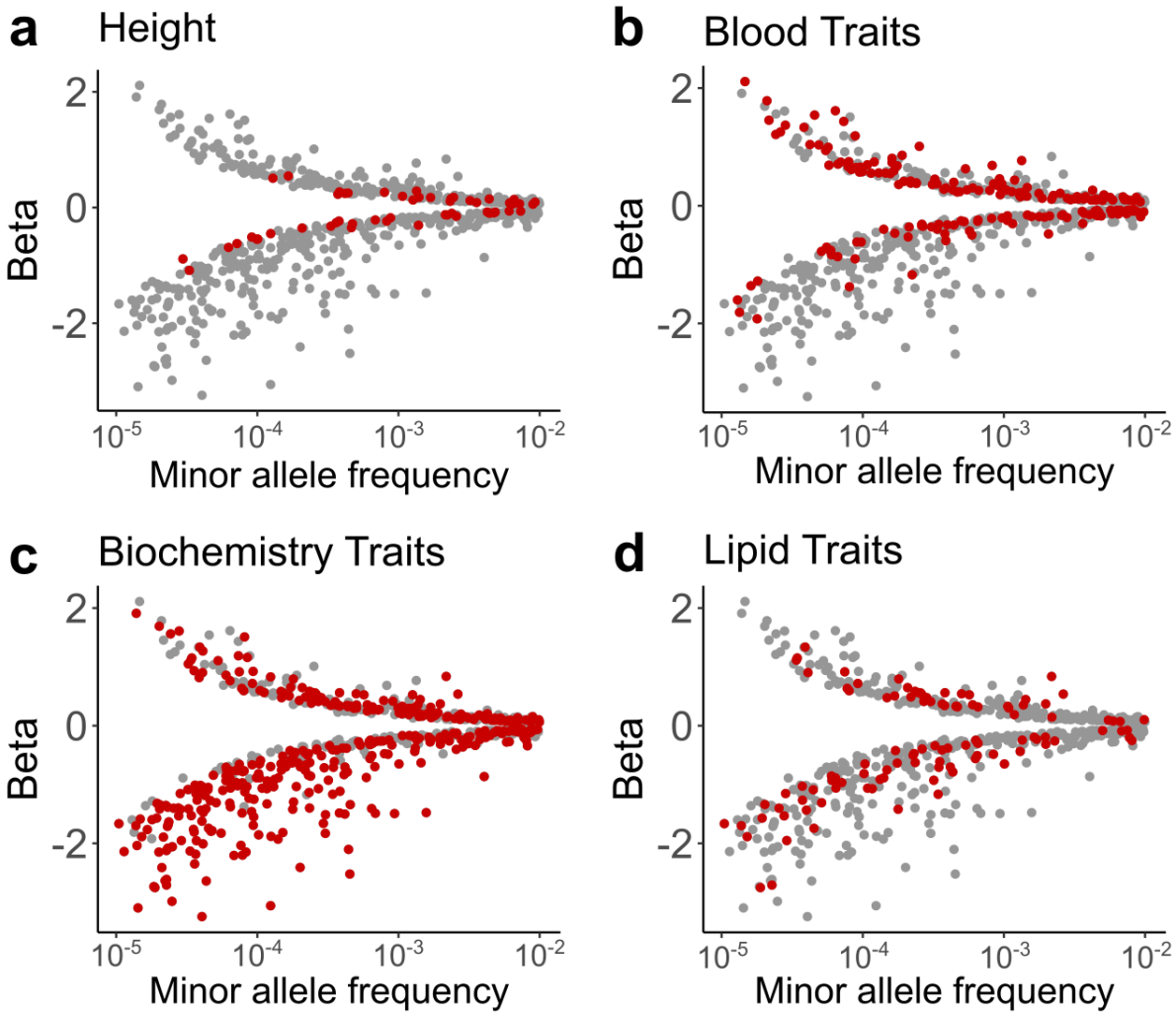
We then expanded this analysis to all 112 ultra-rare variants with 2-5 European-ancestry carriers included in the *N*=50K whole-exome sequenced cohort (**Supplementary Table 15**). Approximately 20% of variants (23 out of 112) were not clustered in a single clade by tsinfer, with the fraction increasing modestly with the number of carriers. Importantly, imputation accuracy was only moderately depressed for variants for which tsinfer suggested recurrence, consistent with the simulation above.

References

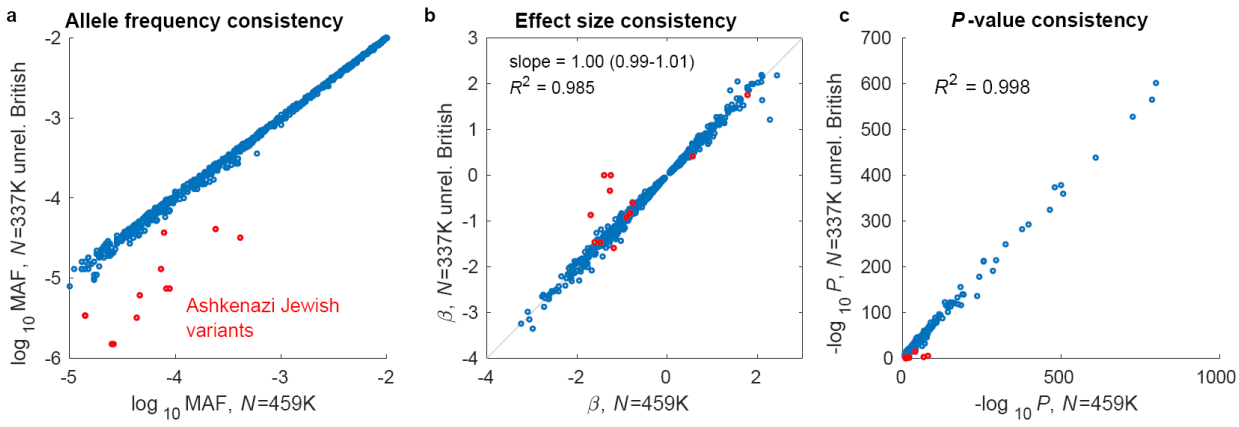
1. Vuckovic, D. *et al.* The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell* **182**, 1214-1231.e11 (2020).
2. Marouli, E. *et al.* Rare and low-frequency coding variants alter human adult height. *Nature* **542**, 186–190 (2017).
3. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
4. 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).
5. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* **44**, 243–246 (2012).
6. Haworth, S. *et al.* Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nat. Commun.* **10**, 333 (2019).
7. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
8. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
9. Kelleher, J. *et al.* Inferring whole-genome histories in large population datasets. *Nat. Genet.* **51**, 1330–1338 (2019).



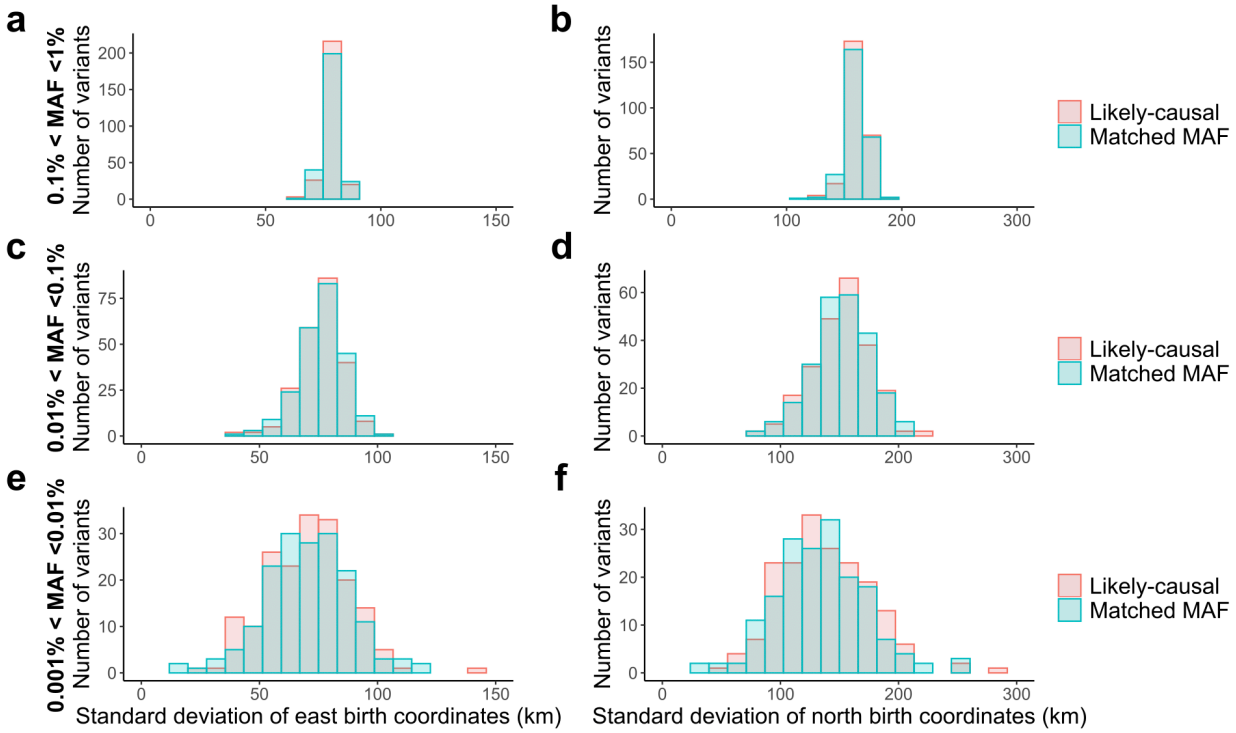
Supplementary Figure 1. Most likely-causal rare coding variant associations identified by whole-exome imputation in UK Biobank were not in the GWAS Catalog. For each trait, we tabulated whether each likely-causal variant was previously reported in the NHGRI-EBI GWAS catalog as associated with any trait (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).



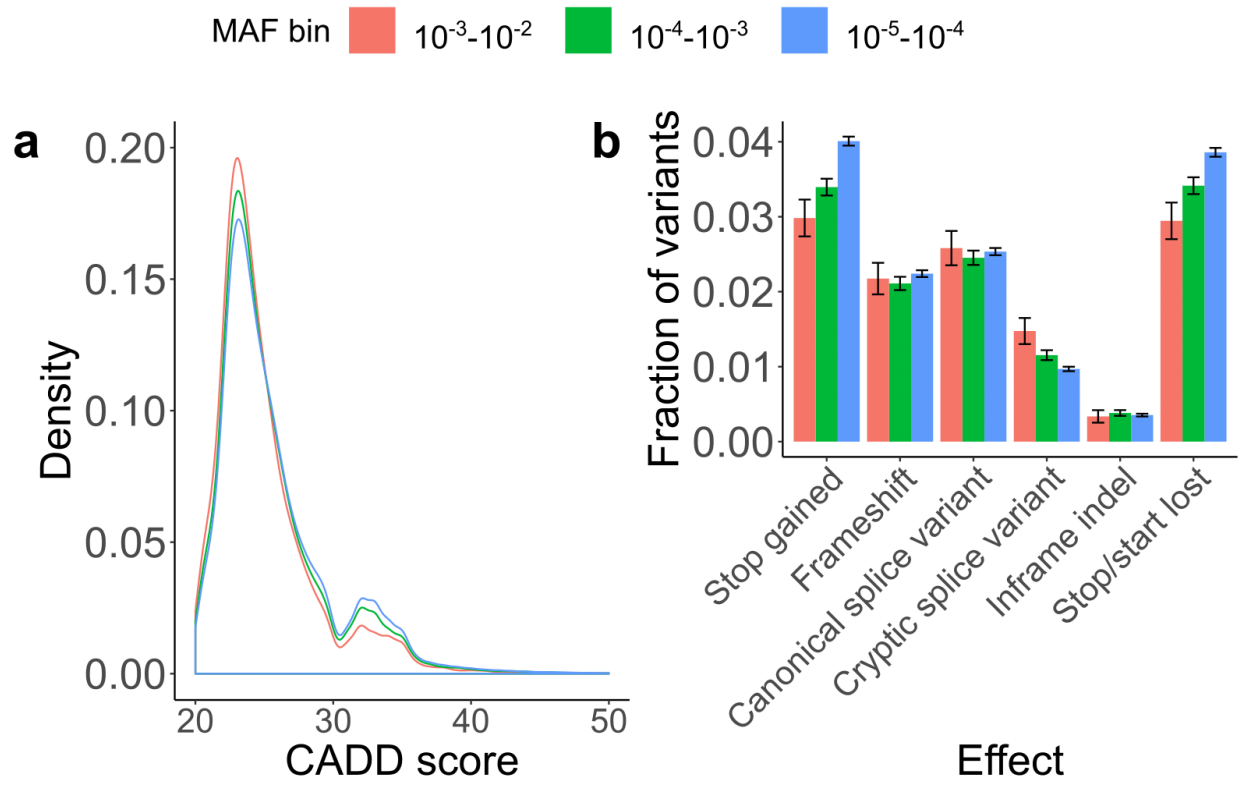
Supplementary Figure 2. Magnitudes of effect sizes of likely-causal rare coding variants generally increase with decreasing minor allele frequency. Gray dots represent the full set of 1,189 likely-causal coding associations with red dots highlighting this trend for specifically (**a**) height, (**b**)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



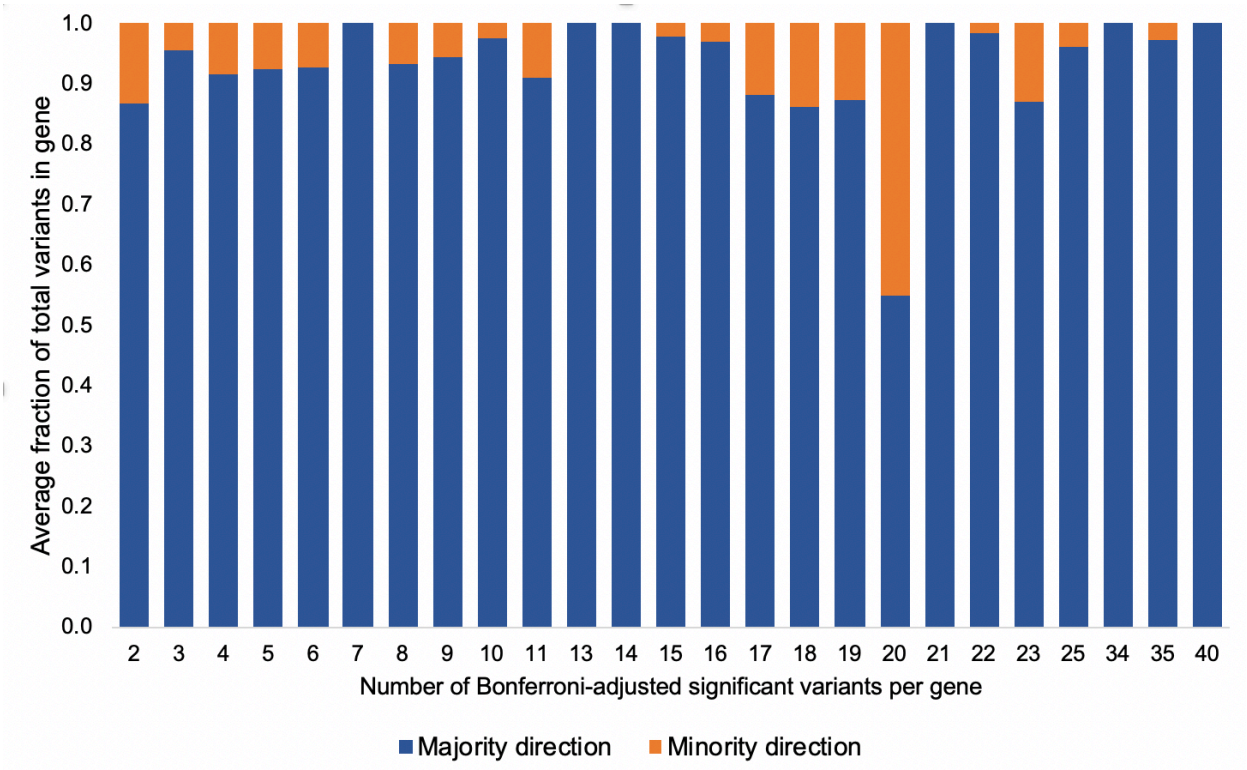
Supplementary Figure 3. Rare variant association tests using linear mixed models show no evidence of confounding from sample structure within UK Biobank. (**a**) Allele frequencies, (**b**) effect size estimates, and (**c**) association *P*-values are all highly consistent between our primary analyses, 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**).



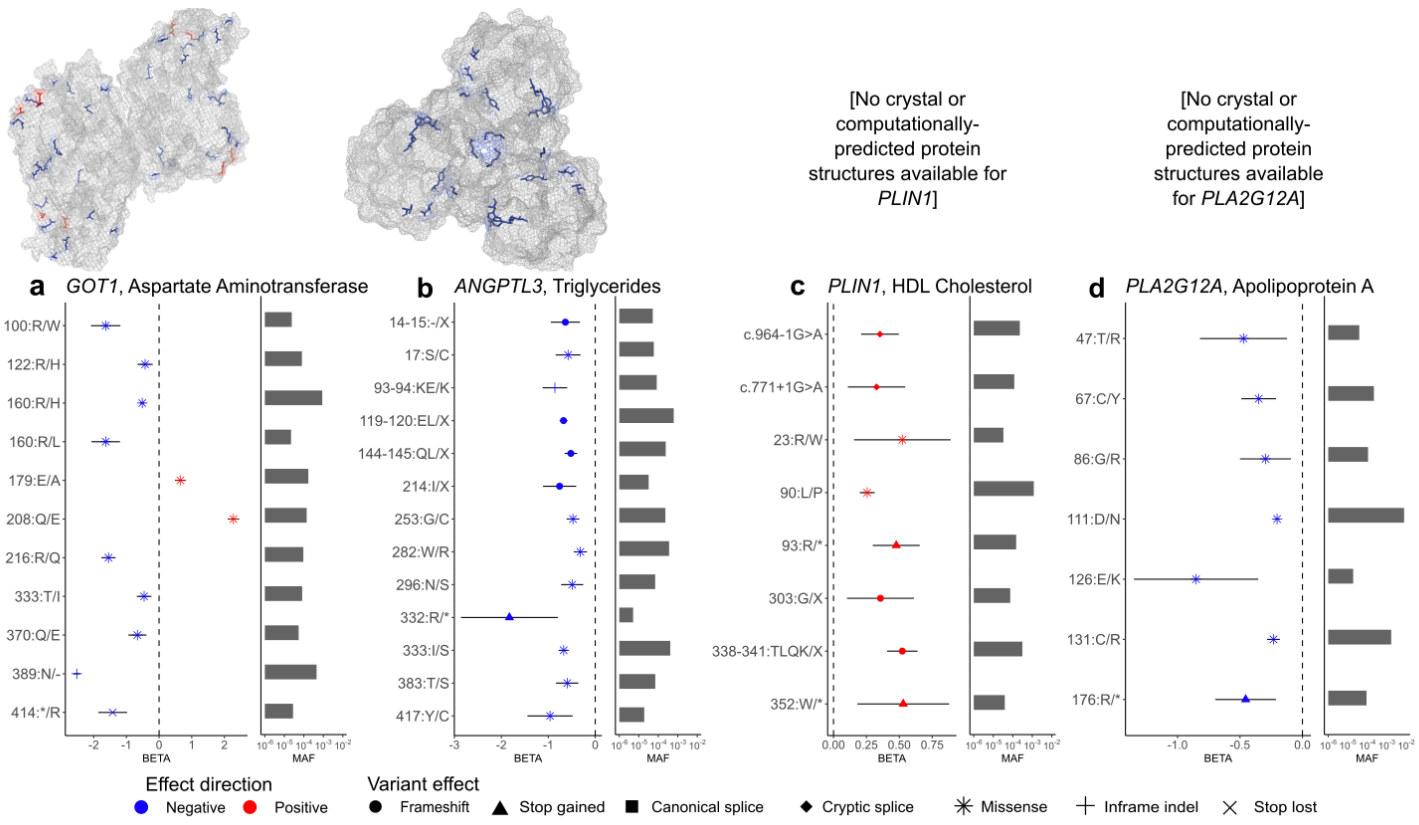
Supplementary Figure 4. Likely-causal rare coding variants are no more geographically localized than allele frequency-matched background variants. For each likely-causal variant, and for a MAF-matched set of background variants, we computed the standard deviation of east (respectively north) birth coordinates 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.



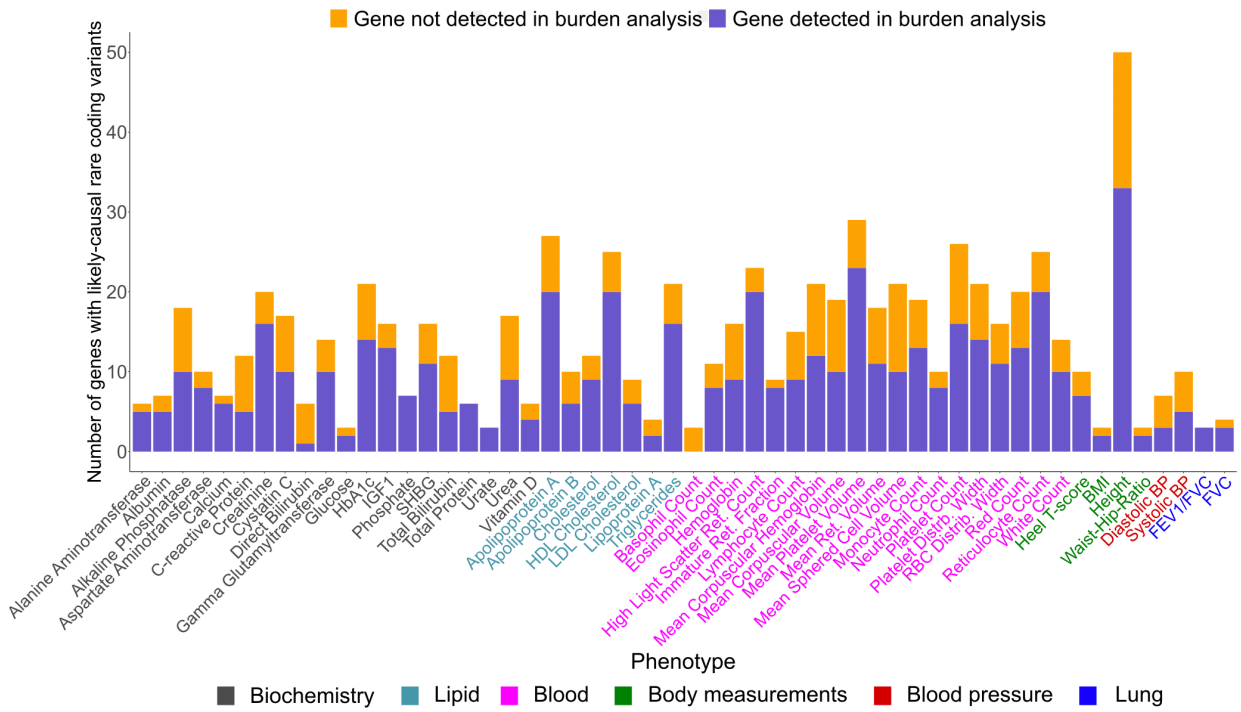
Supplementary Figure 5. Measures of deleteriousness among protein-altering variants increase modestly with decreasing minor allele frequency. (**a**) CADD score. (**b**) Predicted protein alteration from VEP or SpliceAI (for cryptic splice sites). Distributions are across all protein-altering variants present in the UKB *N*=50K whole-exome sequencing genotype call set with European minor allele frequencies within the indicated tranches. MAF 10^{-3}-10^{-2} (pink) *n*= 18,438; MAF 10^{-4}-10^{-3} (green) *n*=100,468; MAF 10^{-5}-10^{-4} (blue) *n*=410,642. Bar heights represent identified fraction. Error bars, 95% CIs.



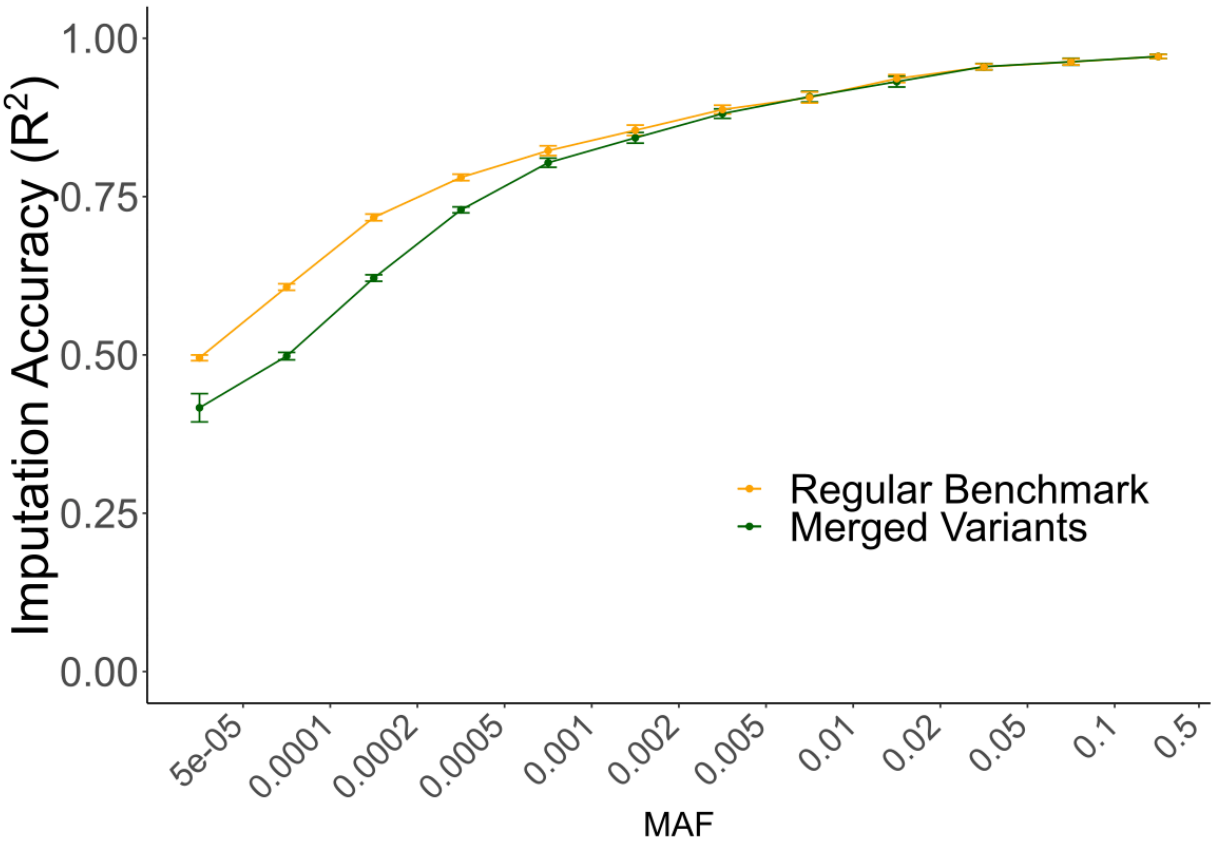
Supplementary Figure 6. Concordance in effect directions of rare coding variants within the same gene. Gene-trait pairs were stratified by the number of independent rare coding variant associations identified (in follow-up analyses that relaxed the significance threshold to Bonferroni-adjusted $P<0.05$, 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. (Under the FDR procedure, having a larger number of positive-effect variants, for example, would lower the threshold for additional positive-effect variants compared to negative-effect variants within the same gene, which would inflate directional concordance.)



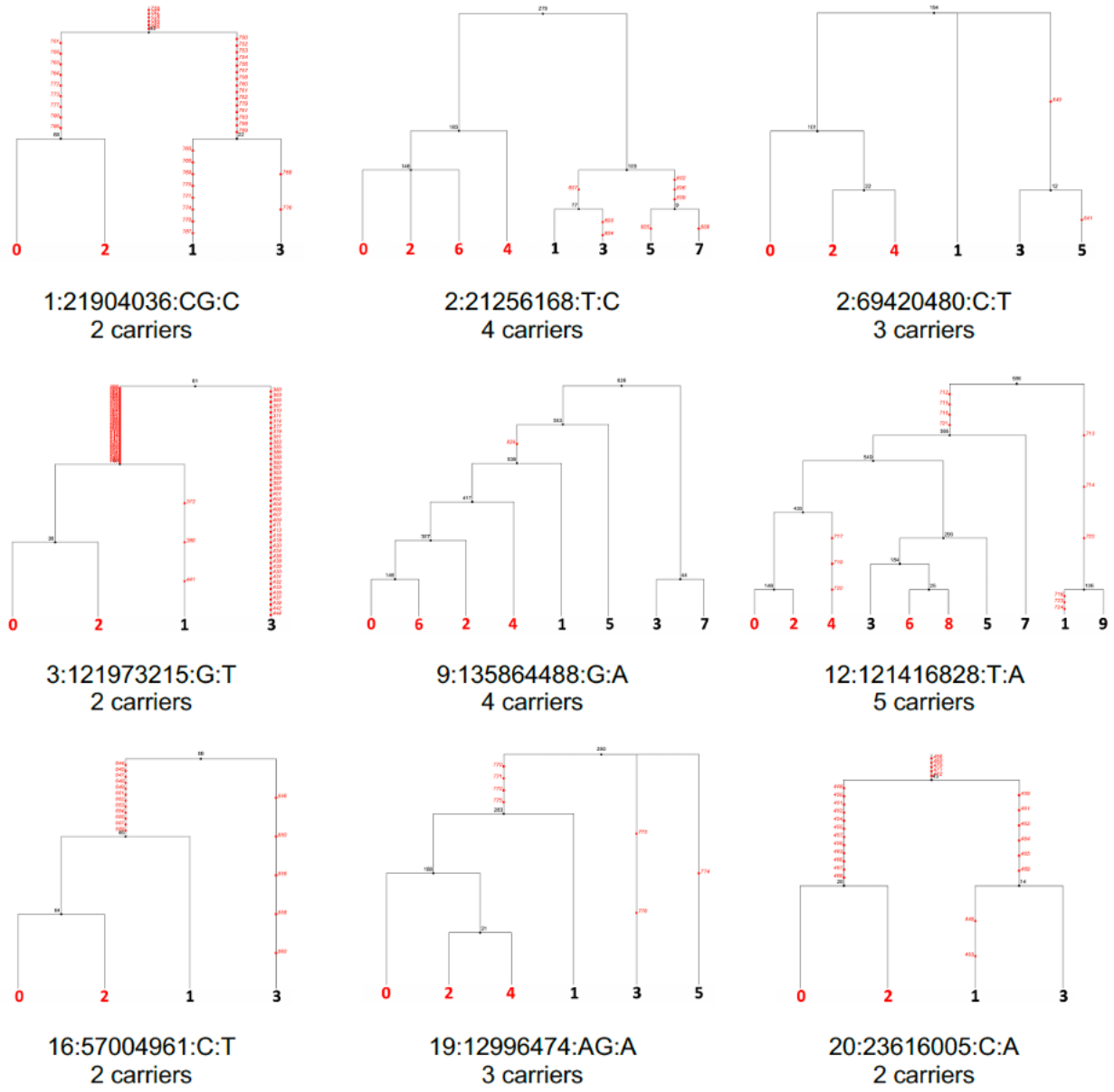
Supplementary Figure 7. Allelic series of trait-associated rare coding variants in *GOT1*, *ANGPTL3*, *PLIN1,* and *PLA2G12A*. Statistically independent associations (reaching FDR<0.05 significance) for: (**a**) *GOT1* and aspartate aminotransferase, (**b**) *ANGPTL3* and triglycerides, (**c**) *PLIN1* 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 (mean values plotted, 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.



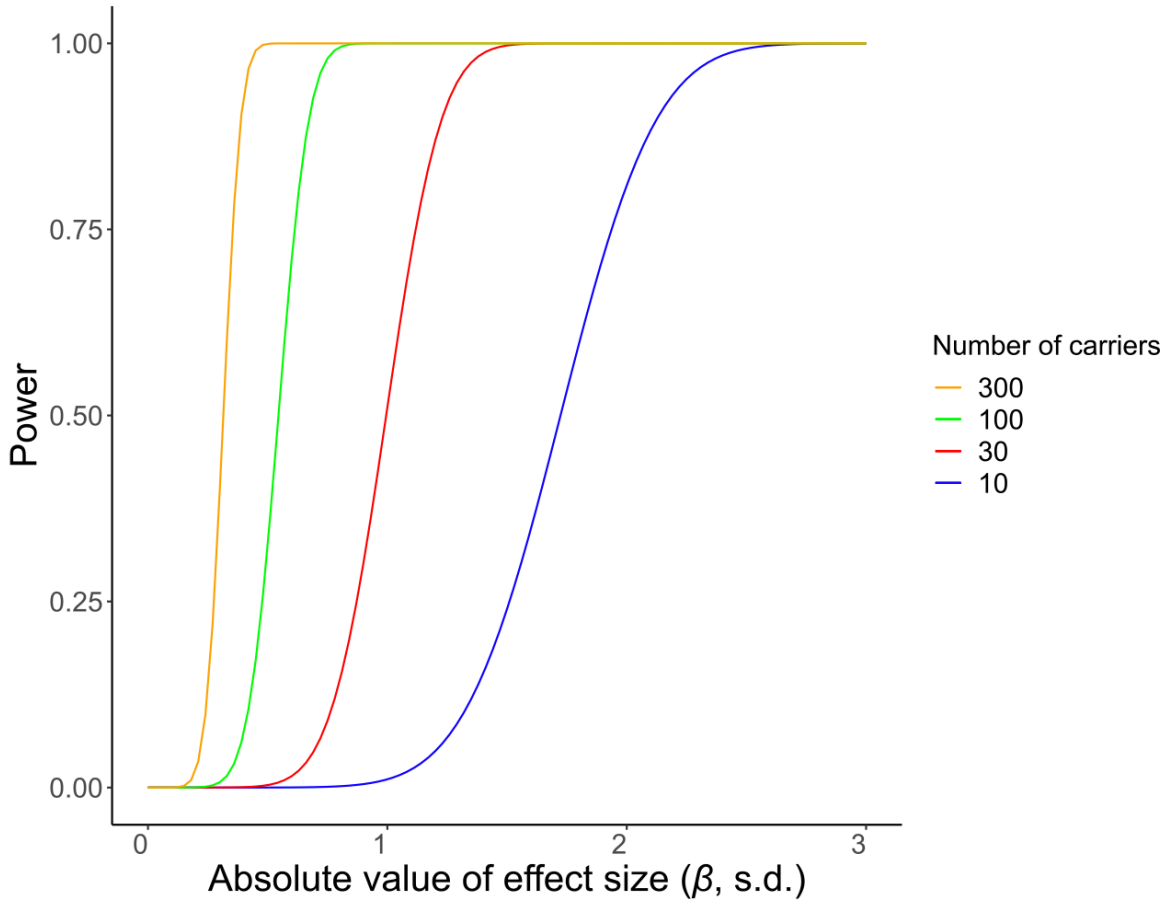
Supplementary Figure 8. Single-variant association analysis discovers likely-causal rare coding variants in genes not identified by gene burden analysis. For each trait, genes containing at least one likely-casual rare coding variant were tabulated according to whether or not they reached significance 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.



Supplementary Figure 9. Robustness of imputation to recurrent mutation. To explore the effect of recurrent mutation on imputation accuracy, we simulated this scenario by merging pairs of consecutive variants, imputing these "recurrent variants," and comparing accuracy to our regular imputation benchmark (**Fig. 1b** and **Supplementary Table 1**). We found that merged variants did not suffer a large depression in accuracy, consistent with the intuition that each constituent variant should still impute roughly as well as before (such that merging variants primarily has the effect of reducing the effective MAF of recurrent variants). Mean values plotted. Error bars, 95% CI.



Supplementary Figure 10. Coalescent trees inferred on carriers of 9 randomly sampled ultra-rare (MAF<10^{-4}) likely-causal variants. Trees were generated using tsinfer on phased SNPs 5 Mb up and downstream of the variant in carriers of that variant. Each tree includes both the haplotypes that included the variant (even numbers in red) and the opposite haplotype in each individual (odd numbers in black). Haplotypes carrying the variant nearly always cluster together, indicating descent from a common ancestor.



Supplementary Figure 11. Power curves for detection of associations between rare variants and quantitative phenotypes. In an additive model in which a rare variant has an effect size of β standard deviations on a normally distributed trait, power to detect the association is plotted as a function of β. The separate curves correspond to different numbers of carriers of the rare variant (*N*=10 in blue, *N*=30 in red, *N*=100 in green, and *N*=300 in orange). These curves assume the variant is perfectly genotyped in the data set; if the variant is instead imputed with accuracy R^2, the effective number of carriers decreases by a factor of R^2.