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

nature genetics

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

Alison R. Barton[1,2,3] ✉, Maxwell A. Sherman[1,2,4], Ronen E. Mukamel[1,2] and Po-Ru Loh[1,2] ✉

Exome association studies to date have generally been underpowered to systematically evaluate the phenotypic impact of very rare coding variants. We leveraged extensive haplotype sharing between 49,960 exome-sequenced UK Biobank participants and the remainder of the cohort (total $n \approx 500{,}000$) to impute exome-wide variants with accuracy $R^2 > 0.5$ down to minor allele frequency (MAF) ~0.00005. Association and fine-mapping analyses of 54 quantitative traits identified 1,189 significant associations ($P < 5 \times 10^{-8}$) involving 675 distinct rare protein-altering variants (MAF < 0.01) that passed stringent filters for likely causality. Across all traits, 49% of associations (578/1,189) occurred in genes with two or more hits; follow-up analyses of these genes identified allelic series containing up to 45 distinct 'likely-causal' variants. Our results demonstrate the utility of within-cohort imputation in population-scale genome-wide association studies, provide a catalog of likely-causal, large-effect coding variant associations and foreshadow the insights that will be revealed as genetic biobank studies continue to grow.

Exome association studies have shown that rare coding variants tend to have larger phenotypic effects than common variants and collectively contribute an important component of complex trait heritability[1–4]. However, the phenotypic effects of very rare coding variants have been difficult to assess comprehensively, as exome sequencing studies have only begun to reach the sample sizes needed to power such analyses ($n > 100{,}000$)[5–10], and imputation of rare variants into cohorts of this scale has been insufficiently accurate[11]. The largest exome-wide association studies published to date have analyzed cohorts of $n \approx 50{,}000$ exome-sequenced individuals, and, while these studies have identified modest numbers of variants and genes associated with phenotypes, they have largely been underpowered to evaluate the effects of individual very rare coding variants[7–10].

The UK Biobank (UKB) is a powerful resource for genetic association analyses because of its large sample size ($n \approx 500{,}000$) and deep phenotyping[12]. Previous studies of UKB have examined disease associations of protein-truncating variants genotyped on the UKB array, which was designed to include most predicted loss-of-function (LoF) variants with MAF > 0.02% and missense variants with MAF > 0.2% (refs. [13,14]). However, most LoF variants are ultrarare (MAF < 0.01%), such that only ~14% of rare LoF variants detected in whole-exome sequencing (WES) of 49,960 UKB participants had been genotyped on the UKB array[8].

We reasoned that, although exome sequencing of ~10% of the UKB cohort provided insufficient power to assess directly the effects of ultrarare variants (which have <10 carriers in $n \approx 50{,}000$ sequenced participants), we could leverage the extensive haplotype sharing in the UKB cohort[15,16] to accurately impute these variants into up to ~100 carriers in the full cohort, thereby powering association analysis. (This strategy is distinct from a recent analysis of 'putative LoF-segments' determined based on identity-by-descent sharing, which did not consider LoF phase[17].) By combining this exome-wide imputation strategy with careful fine mapping of significant associations to identify causal effects of rare coding variants on 54 quantitative traits, we identified hundreds of new 'likely-causal' variant–trait associations and obtained insights into widespread allelic heterogeneity and pleiotropy.

Results

Exome-wide imputation, association and fine mapping. We leveraged WES of 49,960 UKB participants together with single nucleotide polymorphism (SNP)-array genotyping in the full cohort to impute exome-wide variants into all UKB participants as follows (full details in Methods). First, we created an imputation reference panel by phasing WES genotype calls together with SNP-array genotypes in the WES cohort using Eagle2 (ref. [16]), restricting to 4.9 million variants with minor allele count (MAC) ≥2. Second, we used Minimac4 (ref. [11]) to impute these variants into phased SNP-array haplotypes we had previously generated for 487,409 UKB participants[18]. This strategy achieved imputation accuracy (R^2) >0.5 for rare variants down to MAF ~0.00005 (Fig. 1a,b, Supplementary Table 1 and Supplementary Note), consistent with previous simulations[19] and roughly one order of magnitude deeper into the rare allele frequency spectrum than the current UKB imputation release (v.3)[12], which used the Haplotype Reference Consortium (HRC) and UK10K/1000 Genomes reference panels[20,21]. Compared with imputation using $n = 97{,}256$ genomes in the TOPMed reference panel[22], within-cohort imputation using the $n = 49{,}960$ UKB WES panel achieved substantially greater coverage of very rare variants while maintaining similar accuracy per imputed variant (Fig. 1a,b).

We tested the imputed variants for association with 54 heritable quantitative traits (measuring anthropometric traits, blood pressure, lung function, bone mineral density, blood cell indices and serum biomarkers; Supplementary Table 2) by running linear mixed model association analysis on $n = 459{,}259$ participants of European

[1]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [2]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [3]Bioinformatics and Integrative Genomics Program, Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [4]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA.
✉e-mail: alisonbarton@g.harvard.edu; poruloh@broadinstitute.org

ancestry by using BOLT–LMM[23,24], which we verified was robust to potential population stratification in rare variant association analysis (Methods and Supplementary Note). This procedure identified tens of thousands of associations between coding variants and traits that reached nominal genome-wide significance ($P < 5 \times 10^{-8}$); however, we expected that most of these associations were not causal but rather reflected linkage disequilibrium (LD) with nearby causal variants.

To filter detected associations to a high-confidence subset containing primarily causal variants, we developed a stringent filtering pipeline that combined variant annotation filters (to increase the prior on causality) with statistical fine mapping (Fig. 1c and Methods). First, we restricted to rare (MAF < 1%) variants predicted to have high protein-altering impact based on either of the following criteria: (1) combined annotation-dependent depletion (CADD)[25] score ≥ 20 (for coding variants annotated by variant effect predictor (VEP)[26], including canonical splice variants); or (2) SpliceAI[27] score ≥ 0.5 (for noncanonical splice variants). In our primary analyses, we further restricted to variants with high estimated imputation accuracy (INFO > 0.5) and with imputed MAF > 10^{-5}. These filters left 529,602 rare coding variants under consideration, of which 440,253 (83%) either were not present or were poorly imputed (INFO < 0.5) in the HRC-based UKB imputation release. Among the 529,602 variants, 1,647 distinct variants associated with at least one phenotype ($P < 5 \times 10^{-8}$), accounting for a total of 2,706 variant–trait associations (Fig. 1c) (with 1.4 false discoveries expected across all 54 traits).

We combined our variant annotation filters with a statistical fine-mapping filter to exclude associations that could be explained by LD with other variants. Our primary filter required that each association remain significant ($P < 5 \times 10^{-8}$, slightly conservative for 529,620 variants tested) after conditioning on any other more strongly associated variant within 3 Mb (considering in turn each variant from either our WES imputation or the UKB imputation v.3 release; Methods). This filter was more robust for our rare variant analyses than standard fine-mapping software packages, which aim to find small sets of variants that explain maximal phenotypic variance, making configurations that include rare variants less likely to be considered the most probable[28,29]. Fine-mapping algorithms do have the advantage of accounting for the possibility of variants tagging combinations of multiple nearby causal variants (which our pairwise conditional filter did not consider); to account for this possibility, we applied a second filtering pipeline based on iterative runs of the FINEMAP software[28] (Methods). Together, these filters reduced the set of associations to a final likely-causal set of 1,189 associations involving 675 unique variants (Fig. 1c and Supplementary Table 3). Both the variant annotation filters and the fine-mapping filters were designed to be very stringent, with the goal of producing a conservative set of associations with high confidence of causality for downstream analysis. Association data for all variants (including those that failed filters) are also available (Data availability).

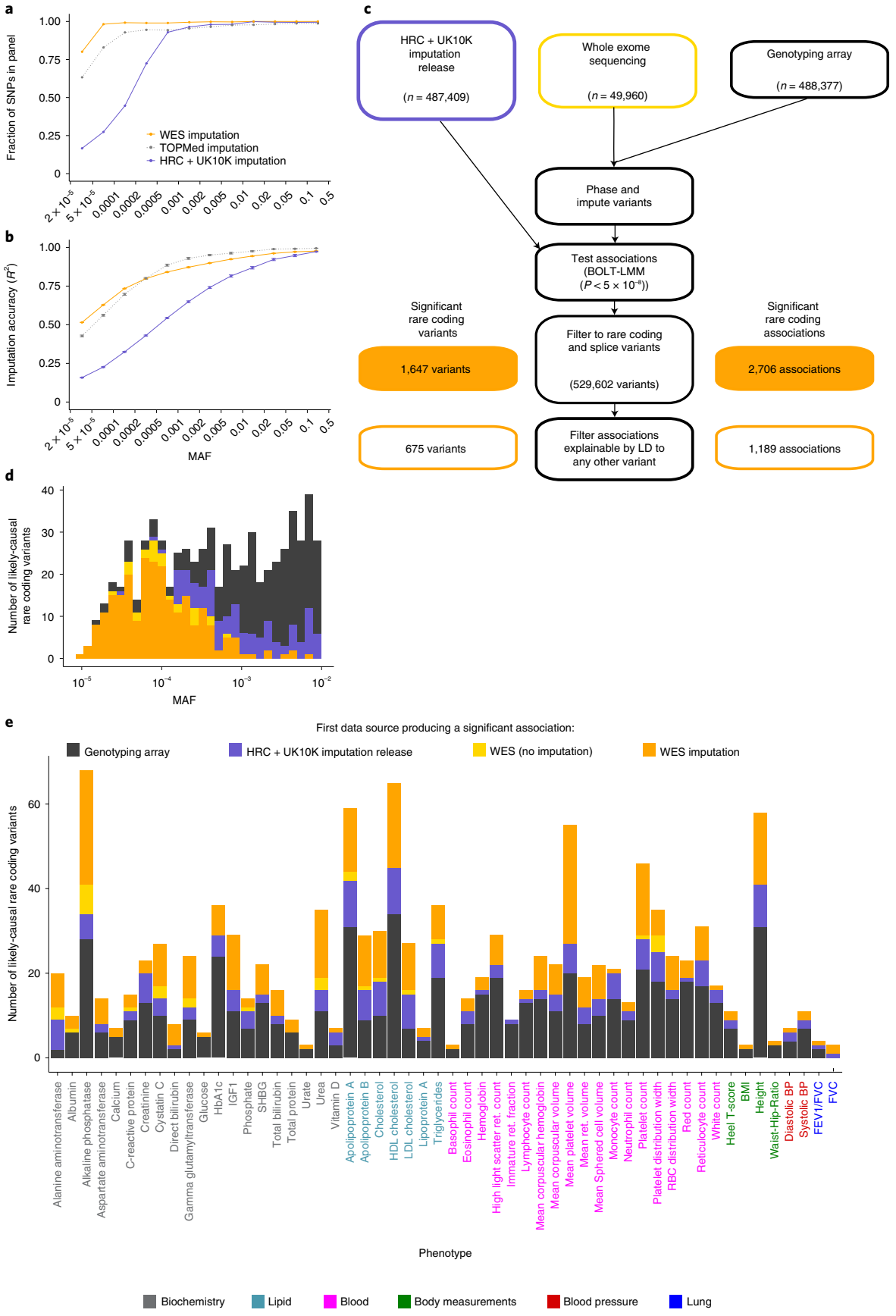
Among the 1,189 likely-causal associations, 30% could be discovered only using imputation from UKB exome sequencing data, demonstrating the power of this approach for causal variant discovery (Fig. 1d,e). The remaining associations could have been discovered previously using either the UKB SNP-array (51% of likely-causal associations, reflecting the inclusion of rare coding variants on the array), the HRC-based UKB imputation v.3 release (an additional 16%) or association analysis in the WES cohort (an additional 3%). Furthermore, among likely-causal associations involving ultrarare variants (MAF < 0.01%), most (197 of 253 associations; 78%) were discoverable only using imputation from the UKB WES cohort (Fig. 1d). Roughly half (576 of 1,189; ~48%) of all likely-causal associations were still not discoverable in the subsequent release of 200,643 UKB exomes[30,31] (Fig. 2 and Supplementary Table 4). Most likely-causal variants (600 of 675; 89%) were not reported in the NHGRI–EBI GWAS catalog for association with any trait, underscoring the power of exome imputation in UKB to detect new rare coding associations (Supplementary Fig. 1). Effect sizes generally increased with decreasing MAF among likely-causal rare coding variants (Supplementary Fig. 2), which collectively explained an average of 0.6% of variance per trait (Supplementary Table 2).

We further attempted to assess the extent to which the likely-causal variants we identified implicated new genes influencing traits. This determination is challenging and generally requires substantial literature review, so we focused our assessment on two types of traits—blood cell traits and height—for which recent, largest-to-date ($n > 500{,}000$) association studies could serve as proxies for previous knowledge (Supplementary Note). For blood cell traits, we found that ~26% (86 out of 337) of the unique gene–trait pairs implicated by our likely-causal associations did not appear among conditionally independent associations reported by Vuckovic et al.[32] (Supplementary Table 5). For height, ~45% (23 out of 51) of the unique genes implicated by our associations were new compared with genes reported by Marouli et al.[2] (Supplementary Table 6).

We expected that the linear mixed models we used for association tests had adequately controlled any potential confounding from population stratification or relatedness[24]. To verify robustness of our results, we performed multiple confirmatory analyses. First, we attempted to replicate associations with traits for which large-scale exome array studies (not including UKB participants) had previously been published. For height, 28 variants we identified as likely-causal had been analyzed in a previous ExomeChip study of height[2]; for all 28 variants, the direction of effect replicated, and 21 of the 28 variants reached nominal significance ($P < 0.05$) in the replication data set (Table 1). Similarly, effect directions replicated for 75 out of 75 lipid associations for which association statistics were available from the Global Lipids Genomics Consortium (GLGC)[3] and for 9 out of 10 blood pressure associations for which data were available from the CHARGE-BP Consortium[4] (Supplementary Table 7). Second, we verified that associations were robust to restricting analysis to a genetically homogeneous subset of unrelated British UKB participants ($n = 337{,}539$): effect sizes (Pearson $R^2 = 0.985$), association strengths (Pearson $R^2 = 0.998$) and allele frequencies (Pearson $R^2 = 0.999$) were all very consistent in this subset (Methods and Supplementary Fig. 3). Third, we verified that likely-causal rare alleles had geographical distributions nearly

Fig. 1 | Whole-exome imputation, association and fine mapping identify rare coding variants likely to causally associate with 54 quantitative traits. **a**,**b**, Imputation panel coverage (**a**) and imputation accuracy (**b**) assessed using SNP calls from the second release of UKB WES data ($n = 200{,}643$; accuracy benchmarks excluded individuals in the initial release). Data are presented as mean values. Error bars, 95% confidence intervals (CIs). **c**, Schematic of our analytical pipeline, which combined UKB WES with SNP-array genotypes to impute exome-wide genotypes into the full cohort. We analyzed imputed exome variants together with the genome-wide UKB imputation release to find significant variant–trait associations independent of neighboring variants, and we restricted to rare (MAF < 0.01) protein-altering variants with CADD ≥ 20 or SpliceAI support to form a final list of likely-causal variants. **d**, Distribution of first UKB genetic data set in which each association could have been detected. Roughly one-third of all likely-causal variants (and nearly all very rare likely-causal variants) were discoverable only using WES imputation. **e**, WES imputation enabled identification of new rare coding variants for all but one trait (immature reticulocyte (ret.) fraction) among 54 quantitative traits analyzed.

identical to MAF-matched background variants (Supplementary Fig. 4 and Supplementary Note). These results indicate that, while subtle stratification in large genetic analyses may affect some types of epidemiological studies[33], the strong, highly localized stratification required to confound rare variant association analyses[34] is unlikely to be present in UKB.



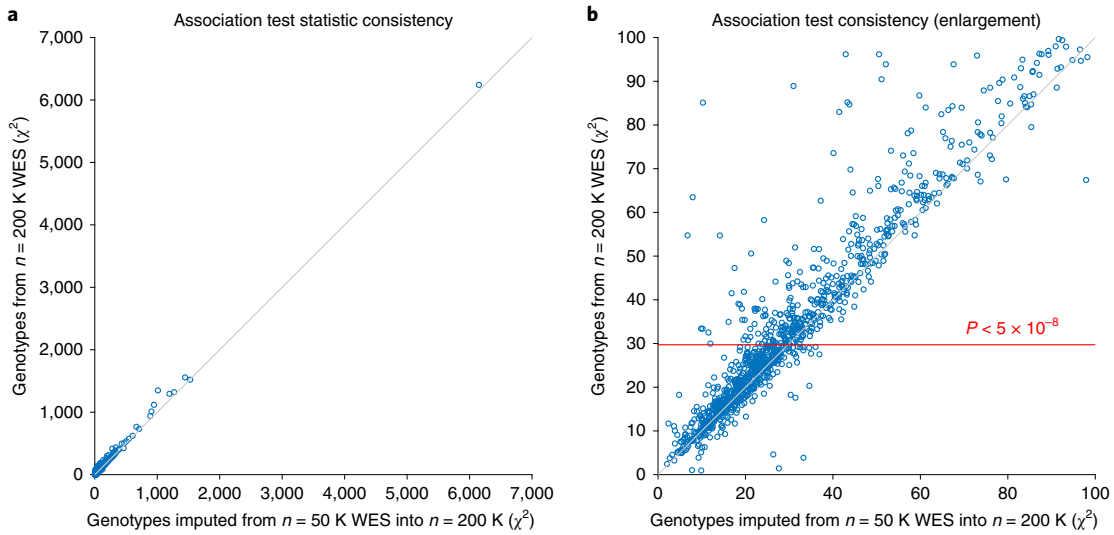


Fig. 2 | Association analyses of the subsequent *n* = 200,643 UKB exome release demonstrate robustness of likely-causal variant–trait associations ascertained using genotypes imputed from *n* = 49,960 exomes. a,**b**, For each likely-causal association, we repeated the association analysis restricting to (1) the *n* = 200,643 cohort, but still using imputed genotypes (*x* axis), or (2) to the *n* = 200,643 cohort and using genotypes directly derived from exome sequencing (*y* axis). Only 613 of 1,189 likely-causal associations from the imputed *n* = 487,409 data set reached significance (BOLT-LMM $P < 5 \times 10^{-8}$; red line in **b**) using the *n* = 200,643 exomes alone. Association test statistics were highly correlated (Pearson $R = 0.96$) between these two approaches. Only six associations involving five distinct variants (1:120463017:C:T, 2:174130918:G:A, 11:48285468:G:A, 16:2287866:G:A and 20:30610469:G:T) decreased in strength by more than twofold in the direct analysis, potentially due to inaccurate imputation or inaccurate genotyping. Panel **b** is an enlargement of the bottom-left corner of panel **a**.

Likely-causal variants are enriched for deleteriousness. The 675 rare coding variants that we identified as likely-causal were distributed roughly evenly across the full range of allele frequencies we considered (MAF = 10^{-5} to 10^{-2}; Fig. 3a). In contrast, the 972 rare coding variants that were annotated as high impact and associated significantly with at least one trait but were filtered after considering LD with other associated variants were enriched for more-common variants (MAF = 10^{-3} to 10^{-2}), suggesting that many of these filtered variants, which constituted most trait-associated rare coding variants, merely tagged causal common variants (Fig. 3a).

To assess enrichment of measures of deleteriousness among the 675 likely-causal variants while controlling for MAF (which is modestly negatively correlated with deleteriousness; Supplementary Fig. 5), we compared features of these variants to a MAF-matched background distribution that we generated by subsampling the 529,602 predicted-high-impact variants we tested (Methods). The average CADD score of likely-causal variants was +1.6 higher than in the background distribution (mean CADD = 27.3 versus 25.3; $P = 1.6 \times 10^{-23}$, two-sample *t*-test) (Fig. 3b). Furthermore, predicted loss-of function mutations (including frameshifts, stop gains and canonical splice variants) were enriched 2.1-fold ($P = 3.7 \times 10^{-16}$, Fisher's exact test) among likely-causal variants (comprising 19.1% of likely-causal variants versus 8.9% of variants from the background distribution; Fig. 3c). In contrast, variants that failed our fine-mapping filters had CADD and variant type distributions similar to background, providing further evidence against causality of most of these variants (Fig. 3b,c). Missense variants, which comprised most likely-causal and background variants, produced broadly more severe amino acid substitutions (as measured by BLOSUM62 scores) across likely-causal variants compared with background (mean BLOSUM62 score = −0.78 versus −0.57; $P = 0.003$, two-sample *t*-test) (Fig. 3d). Cryptic splice variants (computationally predicted by SpliceAI) accounted for 11 of the 675 likely-causal variants and were slightly depleted relative to background, suggesting that these variants were, on average, slightly less likely to affect function than missense variants with CADD ≥ 20 (Fig. 3c); however, our statistical power here was limited.

Rare coding variants form long allelic series. Among the 1,189 likely-causal variant–trait associations we identified, roughly half (578 out of 1,189; 49%) occurred in genes containing multiple likely-causal rare coding variants for the same trait. The observation of two or more rare coding hits in the same gene strengthened our evidence for these associations, and suggested the possibility of longer allelic series in these genes containing very rare causal coding variants that either had not reached genome-wide significance or had been excluded by our stringent filters. To increase our power to detect additional independently associated rare coding variants in these genes, we performed follow-up analyses in which we relaxed the significance threshold (to a 5% false discovery rate (FDR) in each gene–trait pair) and relaxed our fine-mapping filter (conditioning only on a set of associated variants selected by FINEMAP) and annotation-based filter (considering all protein-altering variants regardless of CADD score; Methods).

These analyses revealed very long allelic series of rare coding variants likely to alter phenotypes: for 56 gene–trait pairs, the allelic series contained 10 or more variants on distinct haplotypes, and eight distinct genes contained allelic series of 30 or more variants (Fig. 4 and Supplementary Table 8). In the longest allelic series, 45 rare coding variants in *ALPL* (out of 76 such variants tested) independently associated with serum alkaline phosphatase levels, all with negative effect directions for the rare minor allele. This consistency in effect directions was broadly displayed across the allelic series we identified (93% mean concordance with the main effect direction; Supplementary Fig. 6). Somewhat surprisingly, the amino acid residues modified by missense variants in these allelic series tended not to cluster in specific protein domains (Fig. 4a–d and Supplementary Fig. 7); instead, they appeared to be distributed throughout protein structures, suggesting that protein structures may often contain many domains that are sensitive to mutation.

Table 1 | Replication of likely-causal associations between rare coding variants and height

SNP	Cytoband	Gene	Variant effect	MAF (UKB)	*P* value (UKB)	*P* value (Marouli et al.)[2]	Effect size (s.e.) (UKB)	Effect size (s.e.) (Marouli et al.)[2]	Effect sign agreement	Gene reported
Height variants reaching exome-wide significance in Marouli et al.[2]										
rs143365597	1p34.2	*SCMH1*	Missense	4.4×10^{-3}	6.1×10^{-38}	1.6×10^{-25}	0.149 (0.012)	0.190 (0.018)	✓	✓
rs144673025	1q41	*DISP1*	Missense	7.3×10^{-3}	1.6×10^{-13}	1.1×10^{-9}	−0.063 (0.009)	−0.078 (0.013)	✓	✓
rs142036701	2q35	*IHH*	Missense	3.8×10^{-4}	9.3×10^{-10}	1.1×10^{-15}	−0.234 (0.044)	−0.320 (0.040)	✓	✓
rs149385790	4q26	*PDE5A*	Missense	1.1×10^{-3}	1.8×10^{-14}	7.5×10^{-17}	0.194 (0.026)	0.260 (0.031)	✓	✓
rs778920303	5p13.3	*NPR3*	Missense	2.5×10^{-3}	3.6×10^{-29}	1.1×10^{-8}	0.177 (0.016)	0.130 (0.022)	✓	✓
rs61736454	5q12.3	*ADAMTS6*	Missense	2.6×10^{-3}	5.3×10^{-24}	7.8×10^{-9}	−0.151 (0.016)	−0.150 (0.026)	✓	✓
rs78727187	5q23.3	*FBN2*	Missense	6.6×10^{-3}	1.2×10^{-49}	2.5×10^{-33}	0.139 (0.010)	0.180 (0.015)	✓	✓
rs148833559	5q35.2	*STC2*	Missense	1.3×10^{-3}	9.1×10^{-40}	5.7×10^{-15}	0.285 (0.022)	0.290 (0.037)	✓	✓
rs75596750	8q24.22	*ZFAT*	Cryptic splice	8.0×10^{-4}	6.5×10^{-23}	1.5×10^{-12}	0.264 (0.028)	0.250 (0.036)	✓	✓
rs138273386	11p14.2	*FIBIN*	Missense	4.3×10^{-3}	2.4×10^{-10}	5.8×10^{-12}	−0.078 (0.013)	−0.120 (0.017)	✓	✓
rs141308595	15q26.1	*HAPLN3*	Missense	1.4×10^{-3}	7.4×10^{-46}	2.8×10^{-13}	−0.307 (0.022)	−0.27 (0.037)	✓	✓
Height variants not reaching exome-wide significance in Marouli et al.[2]										
rs121908188	1p36.11	*SEPN1*	Missense	8.8×10^{-4}	4.8×10^{-11}	1.1×10^{-1}	−0.178 (0.028)	−0.088 (0.055)	✓	
rs200496074	1p35.2	*COL16A1*	Missense	6.5×10^{-3}	1.1×10^{-10}	2.9×10^{-1}	0.065 (0.010)	0.063 (0.015)	✓	
rs201166538	3q13.33	*LRRC58*	Missense	1.7×10^{-4}	6.7×10^{-20}	1.0×10^{-1}	0.545 (0.061)	0.097 (0.092)	✓	
rs143137713	3q24	*GYG1*	Missense	2.4×10^{-3}	1.0×10^{-10}	1.7×10^{-1}	−0.101 (0.017)	−0.049 (0.036)	✓	
rs73181210	3q26.2	*PHC3*	Missense	7.2×10^{-3}	9.0×10^{-11}	1.1×10^{-5}	0.066 (0.009)	0.056 (0.013)	✓	✓
rs149437411	3q27.3-q28	*LPP*	Missense	3.2×10^{-3}	2.0×10^{-15}	1.7×10^{-5}	0.122 (0.015)	0.088 (0.020)	✓	
rs147927477	6p21.32	*COL11A2*	Missense	6.8×10^{-4}	9.3×10^{-12}	9.9×10^{-1}	−0.215 (0.031)	−0.001 (0.049)	✓	
rs146458902	7p14.1	*GLI3*	Missense	6.1×10^{-3}	1.1×10^{-15}	3.5×10^{-4}	0.080 (0.010)	0.060 (0.017)	✓	✓
rs121912974	7q11.23	*POR*	Missense	3.8×10^{-4}	5.7×10^{-10}	1.2×10^{-3}	0.271 (0.042)	0.180 (0.057)	✓	
rs140870470	9p13.3	*NPR2*	Missense	4.4×10^{-4}	1.8×10^{-11}	1.4×10^{-1}	0.251 (0.039)	0.160 (0.110)	✓	✓
rs143836544	9q34.11	*LRRC8A*	Missense	6.1×10^{-3}	3.2×10^{-9}	2.0×10^{-5}	−0.060 (0.011)	−0.065 (0.015)	✓	✓
rs200733908	11q13.1	*LTBP3*	Missense	4.7×10^{-4}	8.1×10^{-12}	4.0×10^{-1}	−0.281 (0.042)	−0.059 (0.070)	✓	✓
rs202116412	12p13.1	*APOLD1*	Splice donor	1.3×10^{-3}	3.8×10^{-9}	3.4×10^{-6}	0.131 (0.024)	0.110 (0.024)	✓	✓
rs142153001	14q11.2	*LRP10*	Missense	9.0×10^{-3}	1.2×10^{-11}	5.0×10^{-3}	0.057 (0.009)	0.035 (0.013)	✓	✓
rs201029932	14q22.2	*SAMD4A*	Missense	4.8×10^{-3}	8.6×10^{-13}	6.8×10^{-4}	−0.080 (0.012)	−0.057 (0.017)	✓	✓
rs35816944	16p13.3	*SPSB3*	Missense	6.7×10^{-3}	1.1×10^{-18}	2.9×10^{-5}	−0.083 (0.010)	−0.067 (0.016)	✓	✓
rs141510764	16q23.1	*CLEC3A*	Missense	3.7×10^{-4}	3.2×10^{-9}	9.1×10^{-4}	0.230 (0.042)	0.260 (0.077)	✓	✓

P values and effect sizes are compared for the 28 height-associated variants that were included in the ExomeChip analysis previously performed by Marouli et al.[2] Effect directions were replicated for all 28 variants, most of which had not previously reached exome-wide significance. The last column indicates whether any variants in the affected gene had previously reached significance; several implicated genes were new relative to Marouli et al.[2]

Most of the allelic series we identified extended previously described allelic series (such as in *PCSK9* and *IQGAP2*; Fig. 4a,b); however, several genes contained long allelic series in which most or all variants represented new associations. At *IFRD2* (interferon-related developmental regulator 2, which has an unknown function), 24 rare coding variants associated independently with high-light-scatter reticulocyte count (Fig. 4c and Supplementary Table 8), suggesting an important role of *IFRD2* in red blood cell development; these associations were specific to reticulocyte indices and did not extend to red blood cell count. A common *IFRD2* eQTL variant (rs1076872, which is synonymous in one *IFRD2* transcript and in the 5′ UTR of another transcript) exhibited the strongest association with reticulocyte indices ($P=1.8\times10^{-545}$), and variants in LD with rs1076872 have been reported by many association studies of blood cell indices. However, *IFRD2* has no common protein-altering variants, such that its apparent sensitivity to coding mutations had not been observable previously; among the 24 variants we identified, only two had MAF > 0.1%. Of the remaining 22 very rare *IFRD2* variants, 19 had positive, large effects on high-light-scatter reticulocyte count (median + 0.61 s.d.); homozygotes and compound heterozygotes for these variants exhibited extreme phenotypes (mean + 2.52 s.d.; s.e.m., 0.25 s.d.).

At *NPR2*, which encodes a natriuretic peptide receptor involved in bone growth regulation[35], 11 rare coding variants associated independently with height (Fig. 4d and Supplementary Table 8). Loss-of-function and gain-of-function variants in *NPR2* have been implicated previously in Mendelian skeletal disorders with very strong, mirror effects on stature; however, well-powered exome array studies have not linked *NPR2* polymorphisms to height in the general population[2]. Our exome-imputation approach uncovered many more *NPR2* alleles that appear to exert milder (but still

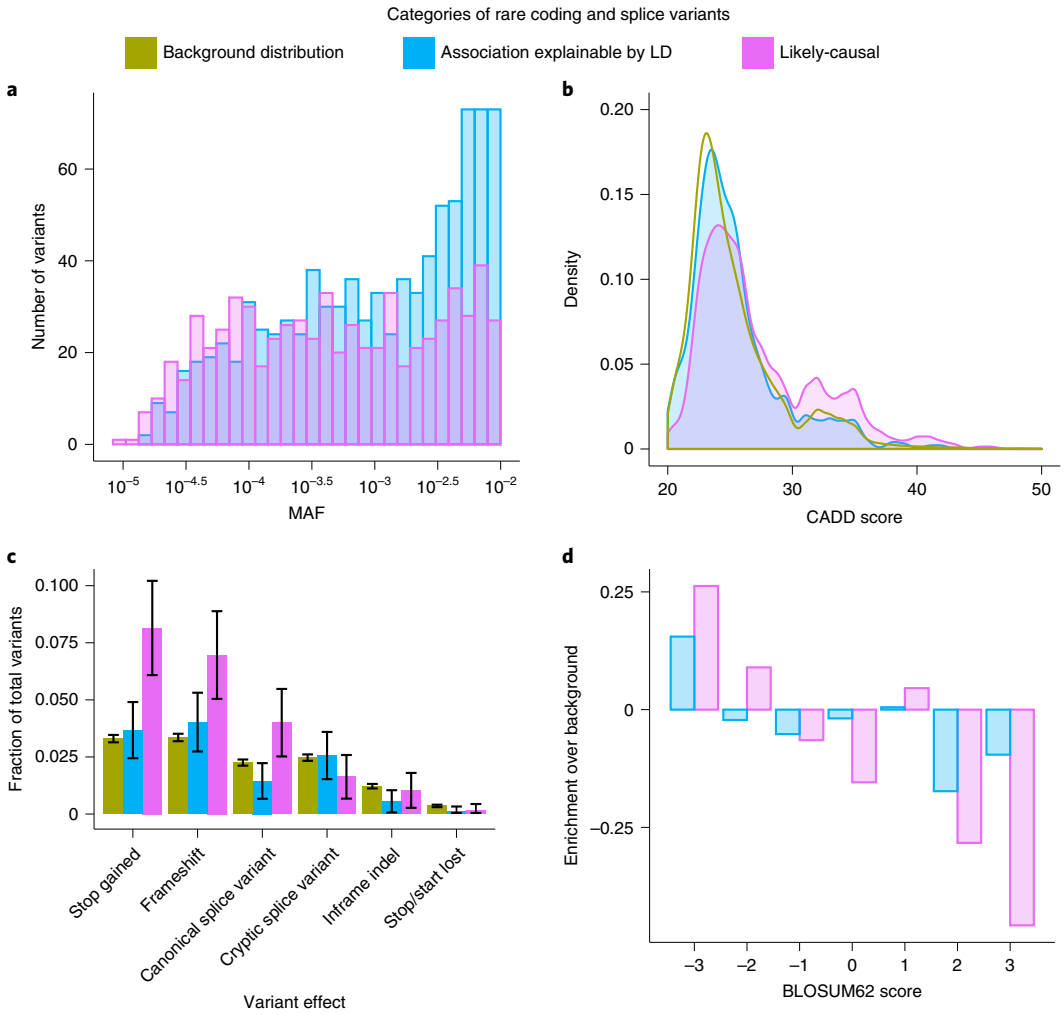


Fig. 3 | Likely-causal coding variants are rare and enriched for deleteriousness. a, Likely-causal variants (pink, $n=675$) had minor allele frequencies distributed relatively evenly across the range under consideration (MAF $=10^{-5}$ to 10^{-2}), whereas variants that failed LD-based filters (blue, $n=898$) tended to be less rare. **b**, Likely-causal variants had elevated CADD scores compared with those that failed LD-based filters and compared with a randomly sampled background distribution of rare coding variants (green, $n=47{,}002$). **c**, Likely-causal variants were enriched for predicted loss-of-function mutations. Bar height represents identified fraction. Error bars estimate sampling uncertainty based on a binomial model, 95% CIs. **d**, Likely-causal missense variants were enriched for higher-impact amino acid substitutions (as measured by more negative BLOSUM62 scores).

strong) effects on height in the UK population, with estimated effect sizes ranging from −1.09 (0.18) s.d. to +0.25 (0.04) s.d.

At *PLA2G12A* and *PLIN1*, allelic series containing up to seven rare coding variants in *PLA2G12A* and eight in *PLIN1* associated with serum lipid levels (Supplementary Fig. 7 and Supplementary Table 8), and the lead association in each series replicated in GLGC data (*PLA2G12A* missense SNP rs41278045: $P=3.3\times10^{-4}$ for HDL and $P=2.3\times10^{-6}$ for triglycerides; *PLIN1* missense SNP rs139271800: $P=1.2\times10^{-4}$ for HDL). *PLA2G12A* encodes a secretory phospholipase that liberates arachidonic acid for eicosanoids with many downstream effects; *PLIN1* encodes a protein that coats lipid droplets. While frameshift variants in *PLIN1* have been implicated in Mendelian lipodystrophies[36], the contribution of rare variants in each gene to population variation in blood lipid levels has been largely unexplored.

Rare coding variants often exhibit pleiotropic effects. Of the 371 genes involved in at least one variant–trait association, 151 genes contained likely-causal variants for two or more traits. These associations often involved related traits, or traits connected by pathways known to involve the gene in question. For example, the cell cycle regulators *CHEK2* and *JAK2* both contained likely-causal variants associated with white blood cell, red blood cell and platelet traits; a *JAK2* missense variant also associated with IGF-1 measurements (Supplementary Table 9). Additionally, three genes that regulate Rho GTPases (*DENND2C*, *DOCK8* and *KALRN*) contained likely-causal variants associated with multiple platelet traits, consistent with the key role of Rho GTPases in platelet function[37]. Other genes associated with more distinct sets of traits (Supplementary Table 9). *APOC3* exhibited the widest variety of likely-causal associations, with the splice donor variant rs138326449 associating with 13 distinct traits, including lipid levels, white- and red blood cell traits and kidney biomarkers. In *PDE3B*, the stop gain variant rs150090666 associated likely-causally with ten distinct traits, including expected associations with waist–hip ratio and lipid measurements[38], but also associations with red blood cell traits, sex hormone-binding globulin levels and height. Further work will be required to determine which of these associations represent direct biological effects versus downstream effects of perturbed regulatory networks (as posited by the omnigenic model)[39].

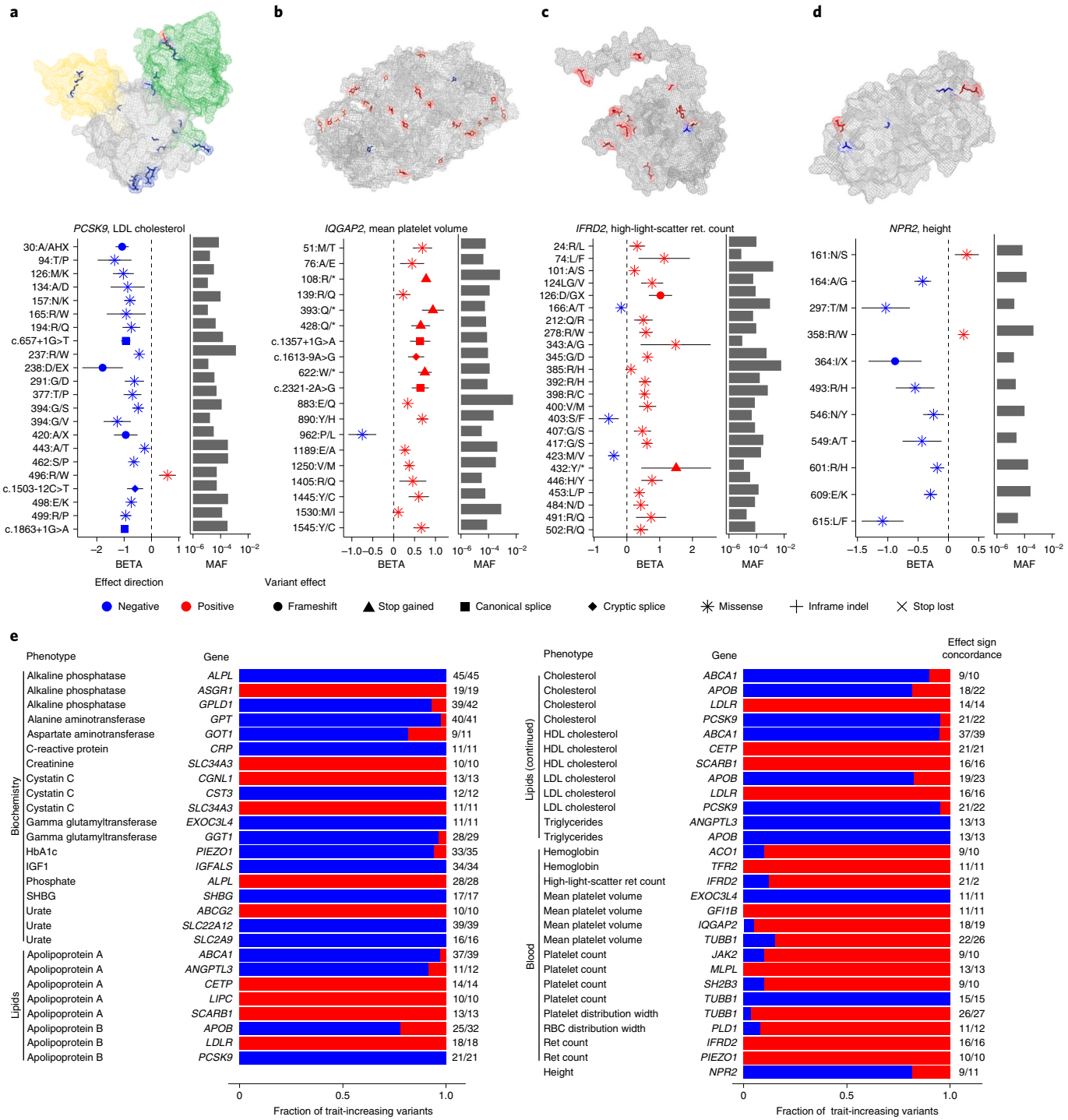


Fig. 4 | Many genes contain long allelic series of rare coding variants with consistent effect directions. a–d, Allelic series of rare coding variants with statistically independent phenotype associations (reaching FDR < 0.05 significance) for *PCSK9* and LDL cholesterol (**a**), *IQGAP2* and mean platelet volume (**b**), *IFRD2* and high-light-scatter reticulocyte count (**c**) and *NPR2* and height (**d**). Top, protein structures with amino acids with codons containing protein-altering variants color-coded by effect direction (red for trait-increasing variants and blue for trait-decreasing variants). Bottom, per variant effect sizes (data points, mean values; error bars, 95% CIs) and allele frequencies. Protein structures were determined previously experimentally (for PCSK9 and IQGAP2) or predicted computationally (for IFRD2 and NPR2). Functional domains of PCSK9 are shaded in different colors. IQGAP2 is represented as a homodimer in its crystal structure. **e**, Distributions of effect directions for all gene–trait pairs with ten or more variants in an allelic series.

Exome imputation uncovers new large-effect variants. Our ability to probe the effects of ultrarare variants revealed ten variants in ten different genes with very large estimated effects on height (≥0.5 s.d.; Supplementary Table 10); in contrast, the largest effect sizes detected in a recent exome array study of height were ~0.3 s.d. (ref. [2]). Four of these genes (*NPR2*, *COL2A1*, *HERC1* and *PCNA*) have been implicated in Mendelian diseases manifesting short stature or skeletal disorder phenotypes; however, the specific variants we identified were not reported previously in ClinVar[40], consistent with their rarity as well as their effects being less extreme, and

contributing to complex genetic variation in height. We also detected one very large effect variant for body mass index in *MC4R* (+0.62 (0.12) s.d.; Supplementary Table 10); this variant had been associated previously with obesity in a Mendelian fashion[41].

Rare coding variants with more moderate effects on height also yielded new insights into the genetic basis of height. Among the 28 height-associated likely-causal variants for which we could replicate effect directions in the ExomeChip study of Marouli et al.[2] (Table 1), seven altered genes that did not contain any variants that had previously reached significance, representing potentially new height loci. Many of these genes had functions suggestive of their association with height, including two collagen genes, *COL16A1* and *COL11A2*. Gene Ontology analysis of all genes containing likely-causal height variants implicated numerous biological processes relating to skeletal system development and extracellular matrix organization (Supplementary Table 11)[42,43].

Biomarker-associated variants confer downstream disease risk. Many phenotypes we analyzed measured blood cell indices or biomarkers for liver, kidney, cardiovascular or endocrine function, suggesting the possibility that rare coding variants affecting these molecular or cellular phenotypes might have downstream impacts on diseases of the corresponding systems. To test this hypothesis, we analyzed likely-causal variants from our blood and biomarker association analyses for association with disease status for related disorders (Methods). A total of 17 associations involving 12 distinct variants reached FDR < 0.05 significance ($P < 1.5 \times 10^{-4}$; Supplementary Table 12), all of which either replicated previous results[44] or added to allelic series at known disease genes (for example, a MAF = 0.1% splice donor in *SLC34A3* that conferred threefold-increased risk of kidney stones ($P = 2.0 \times 10^{-5}$, odds ratio (OR) = 3.1 (2.0–4.8)). In contrast to our analyses of quantitative traits, in which nearly one-third of the associations we identified were discoverable only through exome imputation, 11 of the 12 disease-associated variants had either been genotyped on the UKB SNP-array or accurately imputed from the HRC panel (the only exception being a MAF = 0.04% *LDLR* missense variant implicated previously in familial hypercholesterolemia; Supplementary Table 12). This behavior was consistent with the greater difficulty of identifying robust statistical associations with disease traits (for which causal variants tend to have low penetrance) as compared to molecular or cellular traits (for which causal variants can have much more direct effects). The rarest of the 12 disease-associated variants we identified had MAF = 0.04%; to identify ultrarare variants that influence disease in population cohorts, even larger sample sizes will be needed.

Single-variant tests implicate genes missed by burden tests. Most exome association analyses conducted to date have used gene-based association tests to aggregate signal from very rare variants in the same gene[8,9,45], motivating a comparison between results from our single-variant analyses and a gene-based test using imputed coding variants. In light of our observation that most likely-causal variants from our single-variant analyses had consistent effect directions (Fig. 4e), we aggregated our whole-exome imputed variants in a burden-test framework (rather than using a kernel test that trades off power in this scenario to account for bidirectional effects[46]). A key consideration in performing burden tests is deciding which variants to include as potentially deleterious; as such, we considered two possible functional criteria (protein-altering with CADD ≥ 20 versus predicted LoF) and three possible MAF cutoffs (MAF <1%, <0.1%, or <0.01%) for variants to include (Methods). Of these six parameter combinations, the least stringent option (CADD ≥ 20 and MAF < 1%) seemed to be the most powerful (Supplementary Table 13) and was used for subsequent analyses.

Among gene–trait pairs implicated by our single-variant association tests, 32% were not detected by burden analysis, indicating that single-variant analysis can often be more powerful than gene-based tests for discovering new loci associated with complex traits (Supplementary Table 14). Conversely, most gene–trait associations identified by burden analysis (1130 of 1572; 71% of associations) involved at least one variant that reached significance in single-variant analysis. A sizable minority of these variant associations (414 of 1130; 37% of top-associated variants) had failed our LD-based filters that detected potential tagging of other causal variants, suggesting that many statistically significant results from the burden analysis could potentially represent false-positive associations due to the presence of a very strong causal signal present in a nearby linked gene or regulatory region. The confounding effects of LD were apparent in several large clusters of gene–trait associations near large-effect loci (for example, eight genes within 1 Mb of *APOE* associated with apoB levels; Supplementary Table 14). While burden analyses are somewhat less susceptible to confounding from LD because they aggregate signal across several variants, approximately half of the burden-test associations that reached significance (51%) were dominated by one variant that accounted for most alleles collapsed in the burden analysis, such that the collapsed 'carrier genotype' shared strong, potentially confounding, LD with all variants linked to the dominating variant. These results highlight the need to account for LD, even in the context of burden analysis.

Discussion

These results demonstrate the power of using a large, well-matched reference panel to impute very rare variants into biobank data. Whereas exome sequencing on ~50,000 UKB samples offered limited power to detect associations between coding variants and phenotypes[8,9], imputation into the remainder of the UKB cohort enabled a comprehensive survey of the effects of rare coding variation on 54 quantitative phenotypes (with adequate power even for ultrarare, MAF < 0.01% variants). In combination with fine-mapping analyses, this strategy uncovered many new large-effect coding variants, revealed long allelic series in core genes for many traits, and produced a resource of likely-causal rare coding variant associations for future study. More broadly, our results suggest that sequencing 10% of a cohort and imputing into the remaining 90% can be a cost-efficient strategy for designing genetic association studies. Accurate imputation tends to be possible for variants with at least 5–10 carriers in a reference panel[11,19,20] (assuming most mutations are not highly recurrent (Supplementary Fig. 9), which we verified empirically; Supplementary Fig. 10, Supplementary Table 15 and Supplementary Note); when the panel represents 10% of a cohort, this frequency corresponds to 50–100 carriers in the full cohort, which matches well with the minimum number of carriers typically needed to detect a moderate-effect association (Supplementary Fig. 11).

Our results also have several implications for the analysis of exome association studies. First, single-variant analysis is a viable strategy for extremely large exome association studies. Second, linear mixed model association analysis is robust to population stratification for rare variants as well as for common variants. Third, careful fine mapping is critical for identifying causal associations even when analyzing rare coding variants predicted to have high impact (CADD ≥ 20): even for such variants, most associations appear not to be causal but rather to tag associations of other variants in LD (Fig. 3).

Our study does have important limitations. First, while we observed broad agreement between association statistics computed using genotypes derived from imputation versus direct sequencing (Fig. 2), this agreement was imperfect; some associations (~3%) increased in strength by more than twofold and a few associations

(<1%) decreased in strength by more than twofold in the direct analysis, demonstrating the limitations of rare variant imputation. Second, we restricted our primary analyses to quantitative traits; a comprehensive study of rare coding variant effects on UKB disease traits will require a separate analytical pipeline designed to handle unbalanced binary traits[47]. Third, while we could filter associations potentially explained by LD with other variants imputed from exome sequencing or the HRC reference panel, we could not account for potential tagging of variants unavailable to us (for example, very rare noncoding variants or structural variants). This limitation is shared by all fine-mapping studies conducted to date; here, we expect that our annotation-based filters (requiring that likely-causal coding variants be rare and have high predicted impact) ameliorate this concern. This intuition appears to be borne out by our replication analysis of height variants (in a pan-European meta-analysis that presumably contained different LD patterns) and qualitatively by the large proportion of likely-causal associations that involved genes with clear biological relevance (Supplementary Table 3).

Our study of UKB exome data also gives an indication of the analyses that will become feasible as exome association studies grow even larger. Very large exome-sequenced cohorts provide a natural genetic perturbation experiment. The 49,960 UKB exomes we studied here contained ~7 million missense variants that modified ~3.7 million different amino acids—a sizable fraction of the ~9 million amino acids encoded by all genes in the human genome[26]. Most of these variants were singletons or doubletons and were therefore difficult or impossible to impute; however, when exome sequencing of the full UKB cohort is complete, whole-exome imputation into even larger cohorts will enable characterization of the effects of much of the viable coding variation in the genome.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-021-00892-1.

Received: 21 August 2020; Accepted: 28 May 2021;
Published online: 5 July 2021

References

1. International Multiple Sclerosis Genetics Consortium. Low-frequency and rare-coding variation contributes to multiple sclerosis risk. *Cell* **175**, 1679–1687.e7 (2018).
2. Marouli, E. et al. Rare and low-frequency coding variants alter human adult height. *Nature* **542**, 186–190 (2017).
3. Liu, D. J. et al. Exome-wide association study of plasma lipids in >300,000 individuals. *Nat. Genet.* **49**, 1758–1766 (2017).
4. Liu, C. et al. Meta-analysis identifies common and rare variants influencing blood pressure and overlapping with metabolic trait loci. *Nat. Genet.* **48**, 1162–1170 (2016).
5. Fu, W. et al. Analysis of 6,515 exomes reveals a recent origin of most human protein-coding variants. *Nature* **493**, 216–220 (2013).
6. Tennessen, J. A. et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
7. Dewey, F. E. et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **354**, aaf6814 (2016).
8. Van Hout, C. V. et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, 749–756 (2020).
9. Cirulli, E. T. et al. Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. *Nat. Commun.* **11**, 542 (2020).
10. Flannick, J. et al. Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* **570**, 71–76 (2019).
11. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
12. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
13. DeBoever, C. et al. Medical relevance of protein-truncating variants across 337,205 individuals in the UK Biobank study. *Nat. Commun.* **9**, 1612 (2018).
14. Emdin, C. A. et al. Analysis of predicted loss-of-function variants in UK Biobank identifies variants protective for disease. *Nat. Commun.* **9**, 1–8 (2018).
15. Loh, P.-R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet.* **48**, 811–816 (2016).
16. Loh, P.-R. et al. Reference-based phasing using the haplotype reference consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
17. Nait Saada, J. et al. Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations. *Nat. Commun.* **11**, 6130 (2020).
18. Loh, P.-R., Genovese, G. & McCarroll, S. A. Monogenic and polygenic inheritance become instruments for clonal selection. *Nature* **584**, 136–141 (2020).
19. Browning, B. L. & Browning, S. R. Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* **98**, 116–126 (2016).
20. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
21. Huang, J. et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6**, 8111 (2015).
22. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
23. Loh, P.-R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
24. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).
25. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
26. McLaren, W. et al. The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
27. Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548.e24 (2019).
28. Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinforma. Oxf. Engl.* **32**, 1493–1501 (2016).
29. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).
30. Szustakowski, J. D. et al. Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. Preprint at *medRxiv* https://doi.org/10.1101/2020.11.02.20222232 (2020).
31. Wang, Q. et al. Surveying the contribution of rare variants to the genetic architecture of human disease through exome sequencing of 177,882 UK Biobank participants. Preprint at *bioRxiv* https://doi.org/10.1101/2020.12.13.422582 (2020).
32. Vuckovic, D. et al. The polygenic and monogenic basis of blood traits and diseases. *Cell* **182**, 1214–1231.e11 (2020).
33. Haworth, S. et al. Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nat. Commun.* **10**, 333 (2019).
34. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* **44**, 243–246 (2012).
35. Yasoda, A. et al. Natriuretic peptide regulation of endochondral ossification: Evidence for possible roles of the C-type natriuretic peptide/guanylyl cyclase-B pathway. *J. Biol. Chem.* **273**, 11695–11700 (1998).
36. Gandotra, S. et al. Perilipin deficiency and autosomal dominant partial lipodystrophy. *N. Engl. J. Med.* **364**, 740–748 (2011).
37. Aslan, J. E. & McCarty, O. J. T. Rho GTPases in platelet function. *J. Thromb. Haemost.* **11**, 35–46 (2013).
38. Zhao, A. Z., Huan, J.-N., Gupta, S., Pal, R. & Sahu, A. A phosphatidylinositol 3-kinase–phosphodiesterase 3B–cyclic AMP pathway in hypothalamic action of leptin on feeding. *Nat. Neurosci.* **5**, 727–728 (2002).
39. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
40. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
41. Ahituv, N. et al. Medical sequencing at the extremes of human body mass. *Am. J. Hum. Genet.* **80**, 779–791 (2007).
42. The Gene Ontology Consortium. The Gene Ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).
43. Mi, H., Muruganujan, A., Ebert, D., Huang, X. & Thomas, P. D. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* **47**, D419–D426 (2019).

44. Sinnott-Armstrong, N. et al. Genetics of 38 blood and urine biomarkers in the UK Biobank. *Nat. Genet.* **53**, 185–194 (2021).
45. Povysil, G. et al. Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nat. Rev. Genet.* **20**, 747–759 (2019).
46. Wu, M. C. et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
47. Zhou, W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).

Methods

UKB genetic data. Data from the UKB Resource were accessed under application number 10438. All data were collected and made available by the UKB under North West – Haydock Research Ethics Committee reference 16/NW/0274. The UKB cohort was previously genotyped using genome-wide SNP-arrays that produced genotype data for 488,377 UKB participants at 784,256 autosomal SNPs passing quality control[12]. We analyzed these data together with WES data available for 49,960 participants[8]. We analyzed WES genotype calls at 10.2 million autosomal variants from the SPB pipeline[8], filtering to a subset of 9.8 million variants that unambiguously lifted to hg19 using UCSC liftOver, among which 4.9 million had minor allele count of ≥2. We also analyzed imputed genotypes available for 487,409 participants from the UKB imp_v.3 data release, which consisted of 93 million variants imputed using the Haplotype Reference Consortium and UK10K/1000 Genomes reference panels[12].

We restricted our primary analyses to individuals who reported European ancestry (459,327 participants comprising 94% of the cohort). In supplementary analyses, to ensure that our association analyses were not affected by confounding sample structure, we further restricted to a genetically homogeneous, unrelated (at third-degree or closer) subset of 337,539 white British participants[12] (Supplementary Note). We excluded a small number of participants who withdrew from UKB (up to a maximum of 149 withdrawals by the time we completed our study).

UKB phenotype data. We analyzed 54 heritable quantitative traits measured by UKB for most participants. These traits included body measurements (3 anthropometric traits and 1 bone mineral density trait), blood pressure (2 traits), lung function (2 traits), blood cell indices (19 traits) and serum biomarker levels (7 lipid traits and 20 other biomarkers for liver, kidney or endocrine function; Supplementary Table 2). We analyzed all available blood cell traits except for nucleated red blood cell count and percentage (which were mostly zero) and blood cell percentage traits (which were highly correlated with the corresponding blood cell counts). We analyzed all available serum biomarker traits except for estradiol, testosterone and rheumatoid factor (which had measurable levels in fewer than half the cohort). We performed basic quality control on serum biomarker traits by masking extreme outliers (>1,000 times the interquartile range), stratifying by sex and menopause status, applying inverse–normal transformation, regressing out covariates (ancestry group, alcohol use, smoking status, age, height and body mass index) and reapplying inverse–normal transformation. Quality control and normalization of the other quantitative traits was previously described[24].

We also analyzed disease traits affecting organ systems corresponding to molecular and cellular traits above. We analyzed health outcomes in the 'first occurrence' data fields that UKB generated by aggregating information from self-report, inpatient hospital data, primary care or death record data.

Phasing and imputation of WES variants. To generate an imputation reference panel from the WES cohort, we phased the 4.9 million nonsingleton autosomal variants from WES together with variants genotyped on the UKB array (using Eagle2 (ref. [16]) with --Kpbwt = 20,000). We phased the data in chunks of 50,000 variants with an overlap of at least 5,000 variants between consecutive chunks, resulting in a total of 126 chunks across all autosomes. We then imputed the WES-derived variants into the phased haplotypes we had previously generated[18] for 487,409 participants in the full cohort (using Minimac4 (ref. [11])) with noncoding variants from the UKB array used as the imputation scaffold, that is, matching target and reference haplotypes based on SNP alleles at noncoding variants on the array). We benchmarked the accuracy of this imputation approach by computing correlations between imputed genotype dosages and direct genotype calls from exome sequencing of $n = 141,255$ additional individuals subsequently released by UKB (Supplementary Note).

Association tests. We tested variants for association with each of the 54 quantitative traits using the noninfinitesimal linear mixed model association test implemented in BOLT–LMM[23] (--lmmforceNonInf) with assessment center, genotyping array, sex, age, age squared and 20 genetic principal components included as covariates. We fit the mixed model on directly genotyped autosomal variants with MAF > 10^{-4} and missingness < 0.1 and computed association test statistics for WES-imputed variants and variants from the UKB imp_v.3 release. In our primary analyses, we included all participants with nonmissing phenotypes who reported European ancestry (and had not withdrawn from the study). We also performed association analyses that further restricted the sample set to the WES cohort to determine which associations were detectable in the WES cohort alone.

Filtering associations using coding variant annotations. To focus our analyses on variants likely to have protein-altering effects, we filtered significant associations to those involving variants predicted (by genome annotation algorithms) to impact function. For variants modifying protein-coding sequence or canonical splice sites, we required a CADD v.1.3 score ≥20 and a VEP annotation of missense, inframe deletion, inframe insertion, start lost, stop lost, splice acceptor, splice donor, frameshift or stop gained[25,26]. For variants that affected multiple transcripts (for one or more genes), we assigned the most severe VEP annotation (in the order listed above) across all affected transcripts. We also included potential cryptic splice variants predicted by SpliceAI v.1.2 (specifically, variants with a delta score ≥0.5 for at least one of the four splice modifier categories: gain or loss of a splice acceptor or a splice donor)[27].

Filtering associations in LD with nearby variants. To further filter significant associations to a high-confidence set of likely-causal associations, we analyzed LD between pairs of associated variants to identify and remove any associations potentially attributable to tagging of another variant in LD. We took this approach because, while many algorithms have been developed for fine mapping common variant associations, these methods are not optimized for rare variants. Intuitively, they maximize the heritable variance that can be explained by a configuration of causal variants, making configurations which include rare variants, which typically account for very little heritability even though they can have large effect sizes, less likely to be considered probable[28,29].

Our filter, which was equivalent to requiring that each association remain significant ($P < 5 \times 10^{-8}$) after conditioning on any other more strongly associated variant nearby, proceeded as follows. For each rare coding variant i significantly associated with a phenotype, we calculated its correlation r_{ij} (that is, insample LD) with each other more strongly associated variant j (including both WES-imputed variants and variants from the HRC-based imputation release) using plink '--r'[48]. We then computed the approximate chi-square statistic that would be obtained for variant i in a model including variant j as a covariate:

$$\chi^2_{i|j} = \chi^2_i (1 - r_{ij}\sqrt{\chi^2_j/\chi^2_i})^2$$

where χ^2_i and χ^2_j denote the chi-square test statistics computed by BOLT–LMM for variants i and j (and the sign of the square root reflects whether the effect directions are the same or opposite)[49]. To retain variant i association as likely-causal, we required the conditional chi-square statistic $\chi^2_{i|j}$ to exceed 29.7168 (corresponding to $P < 5 \times 10^{-8}$) for every variant j with $\chi^2_j > \chi^2_i$.

Filtering associations in LD with multiple variants. The filter described above was designed to eliminate associations involving variants that primarily tagged one other variant in LD; however, in theory, noncausal variants could escape this filter by tagging a combination of multiple other variants. To account for this possibility, we used the FINEMAP software[28] to determine, for each gene harboring a rare coding variant of interest, whether the local genetic architecture appeared to involve multiple causal variants, and, if so, to assess whether the rare coding variant(s) under consideration remained significantly associated after conditioning on the variants selected by FINEMAP.

We performed this analysis using a two-step procedure. First, we ran FINEMAP's shotgun stochastic search algorithm ('--sss') to identify up to five putatively causal variants among all significantly associated variants within 500 kb of the gene under consideration. This run produced a most probable configuration containing one to five variants, most of which were typically common. We then ran FINEMAP a second time, adjusting the number of allowed causal variants to be one greater than the number selected for the top configuration in the first run, and limiting the set of potential causal variants to those variants in the top configuration from the first run along with all significantly associated rare coding variants in the gene under consideration. The purpose of this second run was to ascertain whether each rare coding variant remained significant in a model conditioning on multiple common variants. Specifically, we extracted the conditional z-scores output by FINEMAP in its '.snp' files and dropped variants with z-score ≤ 4. This filter removed only 20 variants involved in 36 associations, suggesting that most rare variants that tagged other causal variants were tagging primarily just one neighboring variant. We set the z-score threshold to ≤ 4 after exploring other cutoffs such as $z \leq 5.45$, the equivalent of a genome-wide significance threshold. The $z \leq 5.45$ threshold filtered an additional 54 variants; however, several associations with z-scores around 5 that failed this filter appeared to be real (for example, high-CADD or stop gain mutations in genes known to alter lipid levels). In light of this observation and the stringent filtering we had already performed using pairwise tests, we decided to set a threshold of z-score ≤ 4, which appeared to filter primarily spurious associations. Applying this filter together with the previous two filters left us with the final list of 1,189 significant rare coding variant associations involving 675 unique variants for 54 quantitative traits.

Variant lookup in the NHGRI–EBI GWAS Catalog. We compared the variants we identified to those reported in the NHGRI–EBI GWAS Catalog (accessed 15 January 2020)[50]. Each variant was checked to see if it was reported in the catalog for any phenotype to exclude the possibility that the variant was previously reported for a related phenotype.

Replication analyses. Several traits we analyzed had previously been studied in large-scale meta-analyses using exome arrays, providing the opportunity for replication of likely-causal associations that involved variants assayed on the exome arrays. We compared the associations of likely-causal variants we identified for height, blood pressure and lipid measurements (low-density lipoprotein

(LDL) cholesterol, high-density lipoprotein (HDL) cholesterol, triglycerides and total cholesterol) to association statistics previously published by the GIANT Consortium[2] ($n = 381{,}625$), the CHARGE-BP Consortium[4] ($n = 120{,}473$) and the Global Lipids Genetics Consortium ($n \approx 300{,}000$), respectively[3]; all of these meta-analyses studied participants of predominantly European ancestry, and none included UKB. While most variants were too rare to attain statistical significance in these replication datasets (probably due to allele frequency differences between the UK and other European populations), 112 out of 113 associations exhibited the same effect direction in UKB and the replication data set (Table 1 and Supplementary Table 7). We also compared our height associations with association statistics reported from exome sequencing of the FinMetSeq cohort[51] ($n = 19{,}241$), which provided replication support for a few additional variants that happened to have higher allele frequencies in Finns (Supplementary Table 7).

Background distribution for assessing functional enrichment. To identify trends in the deleteriousness of likely-causal rare coding variants as compared with all rare coding variants, we generated a background distribution of rare coding variants with a MAF distribution matching that of the likely-causal variants (to account for the tendency of rarer variants to have higher deleteriousness scores). We first stratified likely-causal variants into three MAF bins: 10^{-5}–10^{-4}, 10^{-4}–10^{-3} and 10^{-3}–10^{-2}. We then subsampled the set of all rare coding variants considered in our analyses (regardless of whether or not they had a significant association) using the R 'sample' function to generate a set of variants with the same fraction of variants in each MAF bin as in the likely-causal set. We included all variants in the highest MAF bin (as this bin contained the fewest variants), which set the total number of variants in the background distribution at 47,002 variants.

Allelic series analyses. As our primary analysis pipeline for identifying likely-causal rare coding variant associations implemented strict filters on statistical significance (in both single-variant analysis and conditional analyses), we applied a secondary analysis pipeline that relaxed these filters to identify additional rare coding variant associations with good statistical support in genes with two or more likely-causal variants for a trait (indicating strong evidence for the gene–trait association). This pipeline applied a two-step approach (detailed in the Supplementary Note) using FINEMAP in a manner somewhat similar to the approach we used to filter associations that could be explained by combinations of other variants. Here, we again performed a first run of FINEMAP to allow it to select a multiple-causal-variant model (this time containing up to 15 causal variants chosen from common and low-frequency variants as well as rare coding variants), and we then ran FINEMAP a second time to perform an iterative conditional analysis using the selected variants together with rare coding variants. We used conditional P values from the second FINEMAP run to assess the extent to which each rare coding variant exhibited a trait association independent of previous variants. Finally, we converted P values to q values to determine the set of rare coding variants that reached significance at a FDR of 5%.

The expanded allelic series we identified at FDR < 0.05 significance often contained many variants. (For genes with multiple transcripts, we counted the lengths of the allelic series for the transcript that contained the most FDR < 0.05 significant variants, treating cryptic splice variants as belonging to all transcripts.) To visualize the effects of these variants, we plotted the amino acids corresponding to protein-altering variants on previously generated protein structures where possible. Experimentally derived protein structures for PCSK9 (2P4E)[52], ANGPTL3 (6EUA)[53], IQGAP2 (5CJP)[54] and GOT1 (3II0) were retrieved from PDB[55]. Computationally predicted structures for NPR2 (P20594 monomer) and IFRD2 (Q12894 monomer) were retrieved from SWISS-MODEL[56].

Associations with health outcomes. We tested likely-causal variants we identified for cellular and molecular phenotypes (blood cell traits, liver biomarkers, diabetes biomarkers, renal biomarkers and cardiovascular biomarkers) for associations with corresponding disease outcomes coded by UKB using ICD-10 codes (blood disorders, D50–D77; liver diseases, K70–K77; type 2 diabetes, E11; gout and kidney diseases, M10 and N00–N29; cardiovascular diseases, I20–I25 and I63). To further reduce multiple testing burden, we further restricted to diseases with at least 500 reported cases. These criteria left 40 phenotypes under consideration (that is, an average of 8 phenotypes tested for each likely-causal variant for each of the five classes of cellular/molecular phenotypes) and resulted in 5,508 separate tests. Setting a FDR threshold of 5% across the 5,508 tests resulted in a significance threshold of $P < 1.5 \times 10^{-4}$.

Gene-based burden tests. We assessed the performance of gene-based association analyses using burden tests that collapsed the genotypes of imputed rare coding variants in each gene. We considered six different criteria for inclusion of rare coding variants in the burden. These six criteria were defined by three different allele frequency thresholds (MAF ≤1%, ≤0.1% and ≤0.01%) and two different variant annotation criteria (protein-altering with CADD ≥ 20 or predicted loss-of-function as annotated by VEP). Collapsed genotypes were coded as 0 (if an individual had no variants meeting these requirements) or 1 (if the individual carried at least one of these variants). We performed association tests against the 54 quantitative traits using BOLT–LMM with the same settings as in our single-variant analyses, and we applied a Bonferroni-corrected P-value threshold of $P < 2.7 \times 10^{-6}$ to account for 18,530 genes tested. We compared the results of these analyses to those previously reported in burden analyses of $n = 49{,}960$ exome-sequenced UKB participants[8,9]. Among phenotypes in common between our analyses and the previous analyses, we replicated 13/15 associations from Van Hout et al.[8] and 48/58 associations from Cirulli et al.[9]. Nonreplicated results might arise from different selection criteria for variants and to a lesser extent from singletons that were included in the previous analyses but excluded from our imputation.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Access to the UKB Resource is available by application (http://www.ukbiobank.ac.uk/). Exome-wide summary association statistics for the 54 quantitative traits we analyzed are available at https://data.broadinstitute.org/lohlab/UKB_exomeWAS/ and data files containing allelic series for all gene–trait associations with multiple likely-causal variants are also available at this website.

Code availability

The following publicly available software packages were used to perform analyses: Eagle2 (v.2.3.5), https://data.broadinstitute.org/alkesgroup/Eagle/; Minimac4 (v.1.0.1), https://genome.sph.umich.edu/wiki/Minimac4; BOLT–LMM (v.2.3.4), https://data.broadinstitute.org/alkesgroup/BOLT-LMM/; FINEMAP (v.1.3.1), http://www.christianbenner.com/; plink (v.1.9 and v.2.0), https://www.cog-genomics.org/plink2/ and tsinfer (v.0.1.4), https://tsinfer.readthedocs.io/en/latest/. Information from the following databases were also used: VEP (v.95 on GRCh37 with GENCODE 19), https://www.ensembl.org/vep; CADD (v.1.5), https://cadd.gs.washington.edu/download; SpliceAI (v.1.2.1) https://github.com/Illumina/SpliceAI; NHGRI–EBI GWAS Catalog (v.1.0), https://www.ebi.ac.uk/gwas/home; TOPMed (v.r2, 97,256 TOPMed samples), https://imputation.biodatacatalyst.nhlbi.nih.gov/#!pages/about; Protein Data Bank, https://www.rcsb.org/; SWISS-MODEL, https://swissmodel.expasy.org/ and PANTHER (v.15.0), http://www.pantherdb.org/. Scripts used to perform the downstream analyses described above are available at https://data.broadinstitute.org/lohlab/UKB_exomeWAS/ (https://doi.org/10.5281/zenodo.4771214).

References

48. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
49. Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–S3 (2012).
50. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
51. Locke, A. E. et al. Exome sequencing of Finnish isolates enhances rare-variant association power. *Nature* **572**, 323–328 (2019).
52. Cunningham, D. et al. Structural and biophysical studies of PCSK9 and its mutants linked to familial hypercholesterolemia. *Nat. Struct. Mol. Biol.* **14**, 413–419 (2007).
53. Biterova, E., Esmaeeli, M., Alanen, H. I., Saaranen, M. & Ruddock, L. W. Structures of Angptl3 and Angptl4, modulators of triglyceride levels and coronary artery disease. *Sci. Rep.* **8**, 6752 (2018).
54. LeCour, L. et al. The structural basis for Cdc42-induced dimerization of IQGAPs. *Structure* **24**, 1499–1508 (2016).
55. Berman, H. M. et al. The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
56. Bienert, S. et al. The SWISS-MODEL repository—new features and functionality. *Nucleic Acids Res.* **45**, D313–D319 (2017).

Acknowledgements

We thank A. Gusev, M. Hujoel, P. Palamara, A. Price and S. Sunyaev for helpful discussions. This research was conducted using the UKB Resource under application no. 10438. A.R.B. was supported by US NIH grant T32 HG229516 and fellowship F31 HL154537. M.A.S. was supported by the MIT John W. Jarve (1978) Seed Fund for Science Innovation and US NIH Fellowship F31 MH124393. R.E.M. was supported by US NIH grant K25 HL150334 and NSF grant DMS-1939015. P.-R.L. was supported by US NIH grant DP2 ES030554, a Burroughs Wellcome Fund Career Award at the Scientific Interfaces, the Next Generation Fund at the Broad Institute of MIT and Harvard, and a Sloan Research Fellowship. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. Computational analyses were performed on the O2 High Performance Compute Cluster, supported by the Research Computing Group, at Harvard Medical School (http://rc.hms.harvard.edu).

Author contributions

A.R.B. and P.-R.L. performed statistical analyses and wrote the manuscript. M.A.S. and R.E.M. provided substantial input on all analyses and on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00892-1.

Correspondence and requests for materials should be addressed to A.R.B. or P.-R.L.

Peer review information *Nature Genetics* thanks S. Petrovski and S. Carmi for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Alison Barton
Po-Ru Loh

Last updated by author(s): 2021-5-28

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used in data collection.
Data analysis	Software used to analyze the data included: Eagle2 (v2.3.5); Minimac4 (v1.0.1); BOLT-LMM (v2.3.4); FINEMAP (v1.3.1); plink (v1.9 and v2.0); tsinfer (v0.1.4). Scripts used to perform the downstream analyses described above are available at https://data.broadinstitute.org/lohlab/UKB_exomeWAS/ (DOI: 10.5281/zenodo.4771214.)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Access to the UK Biobank Resource is available by application (http://www.ukbiobank.ac.uk/). Exome-wide summary association statistics for the 54 quantitative traits we analyzed are available at https://data.broadinstitute.org/lohlab/UKB_exomeWAS/, and data files containing allelic series for all gene-trait associations with multiple likely-causal variants are also available at this website. Databases used to annotate this data included: PDB, SWISS-MODEL, PANTHER, VEP (v95 on GRCh37 with GENCODE 19), CADD (v 1.5), NHGRI-EBI GWAS catalog (v1.0), SpliceAI (v1.2.1), TOPMed (v r2 97,256 TOPMed samples).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was determined by using all available samples from the UK Biobank cohort including the phenotype and genotype array data for 487,409 participants and additional whole-exome sequencing for an initial subset of 49,960 of those participants and secondary set of an additional 150,683 participants .
Data exclusions	For the serum biomarker traits, extreme outliers (>1000 times the interquartile range) were masked. This was done during the data preparation stage prior to any association tests being run.
Replication	Replication was performed by comparing to previous exome-wide and genome-wide association studies. 28 of 28 height-associated variants tested in both our study and Marouli et al. (2017) agreed in effect direction; 75 of 75 lipid trait associations from the Global Lipids Genomics Consortium agreed in effect direction; and 9 out of 10 blood pressure associations replicated in data from the CHARGE-BP Consortium.
Randomization	In our mixed model association analysis, we regressed out covariates including ethnic group, alcohol use, smoking status, age, height, and BMI and stratified individuals by sex and menopause status.
Blinding	Blinding was not applicable to the genome-wide association study conducted. Data collection was performed previously by the UK Biobank.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging