nature genetics

Article

https://doi.org/10.1038/s41588-024-01913-5

Population-specific putative causal variants shape quantitative traits

Received: 17 May 2023

Accepted: 14 August 2024

Published online: 3 October 2024

Check for updates

Satoshi Koyama[1,2,3,27], Xiaoxi Liu[4,27], Yoshinao Koike[4,5,6,27], Keiko Hikino[7], Masaru Koido[4,8,9], Wei Li[10], Kotaro Akaki[10], Kohei Tomizuka[4], Shuji Ito[4,5,11], Nao Otomo[4,5,12], Hiroyuki Suetsugu[4,5,13], Soichiro Yoshino[4,13], Masato Akiyama[4,14], Kohei Saito[15], Yuki Ishikawa[4], Christian Benner[16], Pradeep Natarajan[2,3,17,18,19], Patrick T. Ellinor[2,3], Taisei Mushiroda[7], Momoko Horikoshi[20], Masashi Ikeda[21], Nakao Iwata[21], Koichi Matsuda[22], Biobank Japan Project[23,*], Shumpei Niida[24], Kouichi Ozaki[1,24], Yukihide Momozawa[25], Shiro Ikegawa[4,5], Osamu Takeuchi[10], Kaoru Ito[1] & Chikashi Terao[4,15,26]

Human genetic variants are associated with many traits through largely unknown mechanisms. Here, combining approximately 260,000 Japanese study participants, a Japanese-specific genotype reference panel and statistical fine-mapping, we identified 4,423 significant loci across 63 quantitative traits, among which 601 were new, and 9,406 putatively causal variants. New associations included Japanese-specific coding, splicing and noncoding variants, exemplified by a damaging missense variant rs730881101 in *TNNT2* associated with lower heart function and increased risk for heart failure ($P = 1.4 \times 10^{-15}$ and odds ratio = 4.5, 95% confidence interval = 3.1–6.5). Putative causal noncoding variants were supported by state-of-art in silico functional assays and had comparable effect sizes to coding variants. A plausible example of new mechanisms of causal variants is an enrichment of causal variants in 3′ untranslated regions (UTRs), including the Japanese-specific rs13306436 in *IL6* associated with pro-inflammatory traits and protection against tuberculosis. We experimentally showed that transcripts with rs13306436 are resistant to mRNA degradation by regnase-1, an RNA-binding protein. Our study provides a list of fine-mapped causal variants to be tested for functionality and underscores the importance of sequencing, genotyping and association efforts in diverse populations.

Genome-wide association studies (GWAS) have identified thousands of loci associated with diseases and traits and have contributed to our molecular understanding of human phenotypes[1–9]. However, for most of these loci, we still do not fully understand the causal mechanisms of the associations. This is partly because of insufficient resolution of associations and limited population sources of genetic associations. Non-European large-scale association studies with sufficient resolution of variants would expand the causal mechanisms implicated by population-specific associations and variants. Additionally, the limited availability of sensitive fine-mapping strategies has hindered our understanding of causal variants[10,11]. Furthermore, a substantial fraction of the lead variants and their linked variants exist in noncoding regions, where functional interpretation is still challenging. Enrichment of causal variants in functional annotations would provide clues about the underlying mechanisms[12].

To overcome these challenges and improve our understanding of causal genetic relationships, we adopted the following strategies. First, using 3,256 high-depth whole-genome sequencing (WGS) data from individuals of Japanese ancestry combined with the 1000 Genomes Project[13], we developed a new genotype imputation reference panel. This

A full list of affiliations appears at the end of the paper. *A list of authors and their affiliations appears at the end of the paper. e-mail: chikashi.terao@riken.jp

https://doi.org/10.1038/s41588-024-01913-5

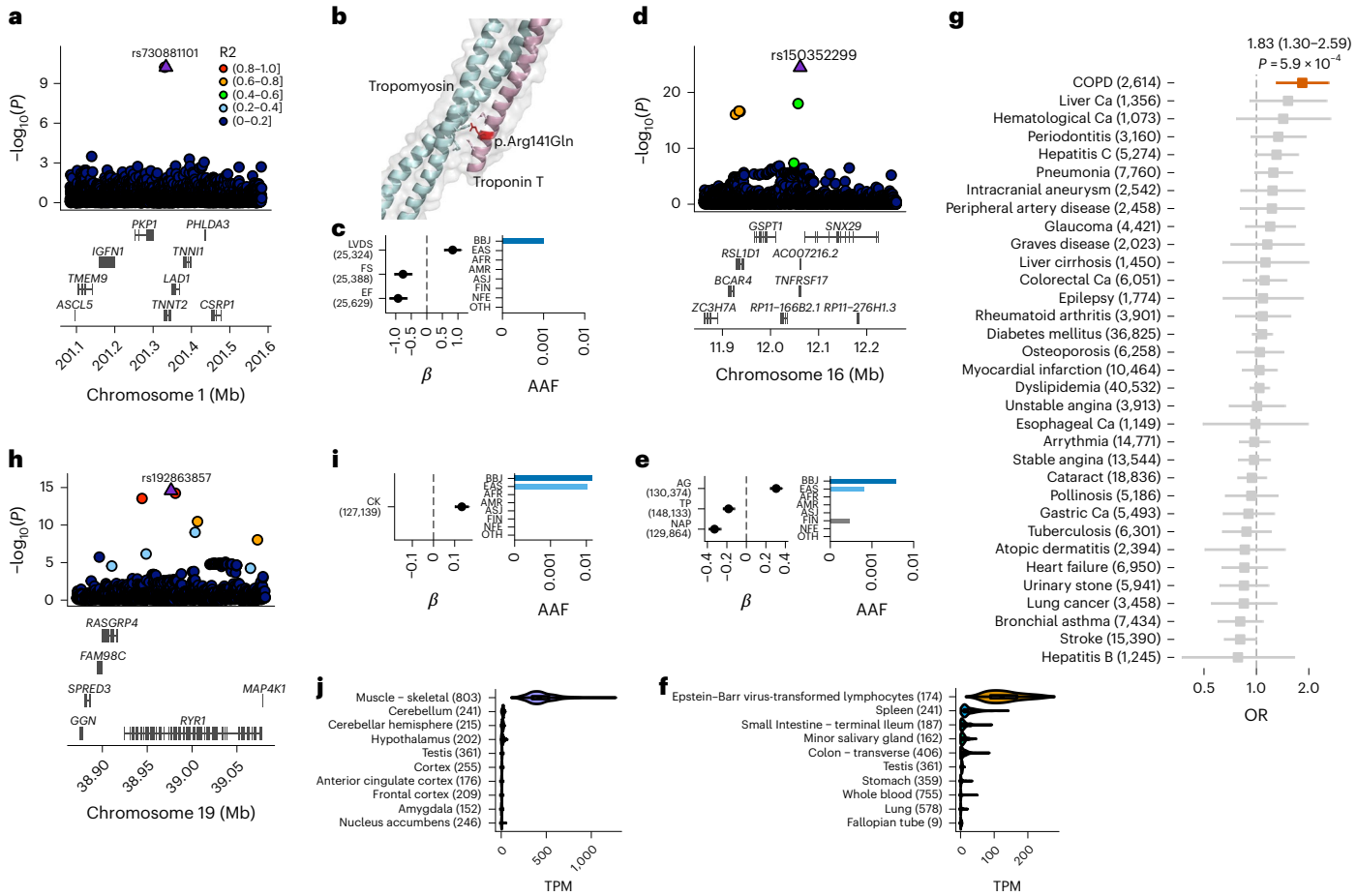


Fig. 1 | New rare putative causal coding variants associated with human quantitative traits implicate candidate causal genes. a, Deleterious coding variant in *TNNT2* (rs730881101) showing strong associations with cardiac functions. The horizontal axis indicates the genomic coordinates; the vertical axis indicates the negative $\log_{10}(P)$. Statistical significance was tested using a linear mixed model. The displayed *P* values are two-sided and not adjusted for multiple testing. **b**, Three-dimensional structure of Troponin-T and putative effect of the coding variant. **c**, β estimates, PPI and alternative allele frequency (AAF) of rs730881101. The error bar for the β estimates indicates the 95% CI. The number of individuals included in the analysis is shown after the trait names. **d**, A deleterious coding variant in *TNFRSF17* (rs150352299) showing strong associations with AG ratio and non-ALB protein levels. The horizontal axis indicates the genomic coordinates; the vertical axis indicates the negative $\log_{10}(P)$. **e**, β estimates, PPI and AAF of rs150352299. **f**, Bulk tissue expression of *TNFRSF17* in the GTEx. The number of samples is shown after the organ name. The violin plots show the distribution of gene expression in transcripts per million (TPM). The box plot shows the median value as the centerline; the box boundaries show the first and third quartiles and the whiskers extend 1.5 times the interquartile range. **g**, OR for 29 diseases of rs150352299 in unrelated Biobank Japan (BBJ) participants. Case counts are shown after the outcomes (n_{Total} = 169,020). The squares indicate the OR; the error bars indicate the 95% CI. Statistical significance was tested using a logistic regression with two-sided test at $P < 0.05/29$. The displayed *P* values were not adjusted for multiple testing. **h**, Deleterious coding variant in *RYR1* (rs192863857) associated with CK levels. **i**, β estimate, PPI and AAF of rs192863857. **j**, Bulk tissue expression of *RYR1* in the GTEx. AFR, African; AMR, Admixed American; ASJ, Ashkenazi Jewish; Ca, cancer; FIN, Finnish; NFE, non-Finnish European; OTH, others. The AAF was obtained from the Genome Aggregation Database (gnomAD) dataset. The number of individuals included in the association analysis is found in Supplementary Table 1; the abbreviations for the phenotypes are found in Supplementary Table 2.

high-quality reference panel enabled us to impute population-specific rare coding and noncoding variants with high accuracy at the population scale. Second, we performed GWAS analyses in up to 260,000 Japanese individuals. Third, we applied statistical fine-mapping to decompose the observed associations into independent causal signals, leveraging the precise linkage disequilibrium (LD) determined by our dense WGS reference panel. Lastly, we conducted comprehensive in silico analyses and follow-up biological experiments for functional interpretation of noncoding variants.

Results

GWAS for 63 quantitative traits

We compiled a new genotype imputation reference panel and used the WGS data to impute the genotypes of 203,216 Japanese individuals; we then performed GWAS analyses for 63 quantitative traits and up to 15,907,072 variants. To replicate the results and maximize statistical power to find new associations, we additionally analyzed 53,083 individuals for 26 traits in another Japanese dataset (Methods, Extended Data Fig. 1, Supplementary Tables 1 and 2, and Supplementary Note 1). We observed a calibrated distribution of test statistics according to the polygenicity of these traits (median LD Score intercept = 1.06; Supplementary Table 3) and high replication rates (Supplementary Note 2). We identified 4,423 genome-wide significant associated loci, including 601 previously unreported loci (Supplementary Tables 4 and 5 and Supplementary Note 2). Statistical fine-mapping revealed 826 phenotype–variant pairs (associations) with a marginal posterior probability of inclusion (PPI) greater than 0.9 and 9,406 with a PPI greater than 0.1 (Supplementary Note 3 and Supplementary Tables 6–12), which, as shown in previous studies[14–16], included loci with multiple signals.

New associations with rare functional coding variants

We found rare Japanese-specific coding variants driving new associations and directly implicating probable causal genes. One such example is a very rare missense variant rs730881101 in *TNNT2*

https://doi.org/10.1038/s41588-024-01913-5

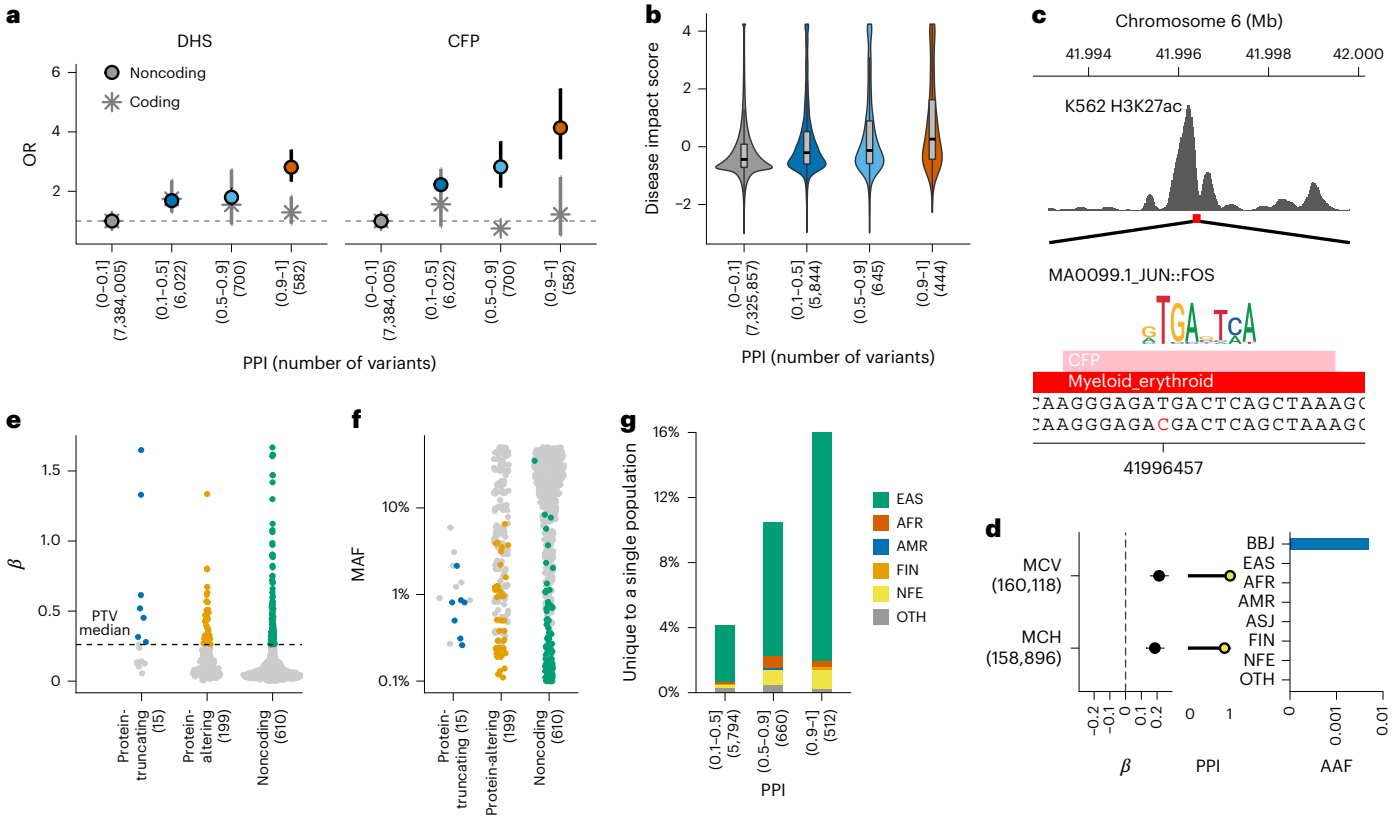


Fig. 2 | Noncoding rare variants associated with human quantitative traits represent a substantial fraction of putative causal variants. a, Enrichment of variants within the regulatory region in variants with high PPI. The vertical axis indicates the OR of variants in each PPI bin within the DHS/CFP or not in comparison with the variants with the lowest PPI bin (0–0.1). The error bars indicate the 95% CIs. The circles and stars indicate noncoding and coding variants, respectively. **b**, Higher predicted pathogenicity of noncoding putative causal variants. The vertical axis indicates the disease impact score predicted from its sequence changes (Methods). The box plot shows the median value as the centerline; the box boundaries show the first and third quartiles and the whiskers extend 1.5 times the interquartile range. **c**, A rare Japanese-specific noncoding variant rs146018792 in *CCND3* strongly associated with MCV and MCH is in the CFP of the myeloid cell line K562. **d**, β estimates, PPI and AAF of rs146018792. The error bar for the β estimates indicates the 95% CI. The number of individuals included in the analysis is shown after the trait names. **e**, Distribution of the absolute β estimates of associations with a PPI > 0.9. The dashed line shows the median absolute β estimate of protein-truncating associations (median $|\beta_{PTV}| = 0.261$). The colored dots indicate large effect associations with $|\beta| > 0.261$. **f**, Distribution of the MAFs of associations with a PPI > 0.9. The colored dots indicate the large effect associations defined in **e**. **g**, Proportion of population-specific variants within each PPI bin. The *y* axis indicates the fraction of variants found in only one population in each indicated PPI bin. The color indicates the population in which the variants were found. The AAF was obtained from the gnomAD dataset. The number of individuals included in the association analysis is found in Supplementary Table 1; the abbreviations for the phenotypes are found in Supplementary Table 2.

(ENST00000509001:c.422G>A, p.R141Q), minor allele frequency (MAF) = 0.1%) associated with decreased systolic heart function (reduced ejection fraction (EF) and increased left ventricular end-systolic diameter (LVDS)) (MAF = 0.1%, $\beta_{EF} = -0.925$, $P_{EF} = 5.9 \times 10^{-11}$, $PPI_{EF} = 0.50$, $\beta_{LVDS} = 0.830$, $P_{LVDS} = 2.7 \times 10^{-9}$, $PPI_{LVDS} = 0.50$; Fig. 1a–c). Notably, the effect size of this variant was more than 80% of the s.d. *TNNT2* is a causal gene for dilated cardiomyopathy and has not been reported for its association with cardiac function in a population-scale GWAS. We also found that this variant was strongly associated with the prevalence of heart failure with a large effect size (odds ratio (OR) = 4.5 (3.1–6.5), $P = 1.4 \times 10^{-15}$).

Another example is rs150352299 in *TNFRSF17* (ENST00000053243: 457G>A, p.A153T; Fig. 1d–g and Supplementary Table 5). This rare (MAF = 0.38%) Japanese-specific missense variant was significantly associated with a higher albumin:globulin ratio (AG) ($\beta_{AG} = 0.306$, $P_{AG} = 3.9 \times 10^{-22}$), lower non-albumin protein (NAP) ($\beta_{NAP} = -0.327$, $P_{NAP} = 3.3 \times 10^{-25}$) and lower total protein (TP) ($\beta_{TP} = -0.183$, $P_{TP} = 6.5 \times 10^{-10}$). These associations suggested decreased globulin concentration in the blood. *TNFRSF17* encodes B cell maturation antigen (BMA), which is specifically expressed in mature B cells and is responsible for antibody production (Extended Data Fig. 2a). Furthermore, we identified an increased risk of chronic obstructive pulmonary disease with rs150352299, which is consistent with several reports of primary immunodeficiency as an underlying cause of chronic obstructive pulmonary disease[17]. BMA is known to interact with B cell activating factor encoded by *TNFRSF13B*, in which we also identified a Japanese-specific rare loss-of-function variant, rs769165409, associated with the AG with high PPI (MAF = 0.1%, $\beta_{AG} = 0.353$, $P_{AG} = 3.9 \times 10^{-7}$; Supplementary Note 4.1). These results provide genetic evidence for critical roles of BMA–B cell activating factor interaction in the immunoglobulin production of B cells.

Other examples include associations with creatine kinase (CK) levels. *RYR1* encodes the ryanodine receptor, a crucial calcium channel in muscle. rs192863857, a rare missense substitution in *RYR1* (ENST00000359596:c.5317C>T, p.P1773S), was associated with CK (MAF = 1.48%, $\beta_{CK} = 0.134$, $P_{CK} = 2.5 \times 10^{-15}$, $PPI_{CK} = 0.67$; Fig. 1h–j and Supplementary Table 5). We also identified a new missense variant associated CK levels in *CACNA1S*, which encodes the main subunit of the calcium channel (MAF = 3.5%, $\beta_{CK} = -0.064$, $P_{CK} = 1.5 \times 10^{-9}$, $PPI_{CK} = 1.00$; Extended Data Fig. 2). These genes are specifically expressed in skeletal muscle (Extended Data Fig. 2b,c) and are involved in malignant hyperthermia (MH), a disease characterized by massive CK elevations precipitated by exposure to certain anesthetics. The results suggest that high serum CK levels in the absence of the causative stressors of MH may reflect the effects of variants in these causal genes.

https://doi.org/10.1038/s41588-024-01913-5

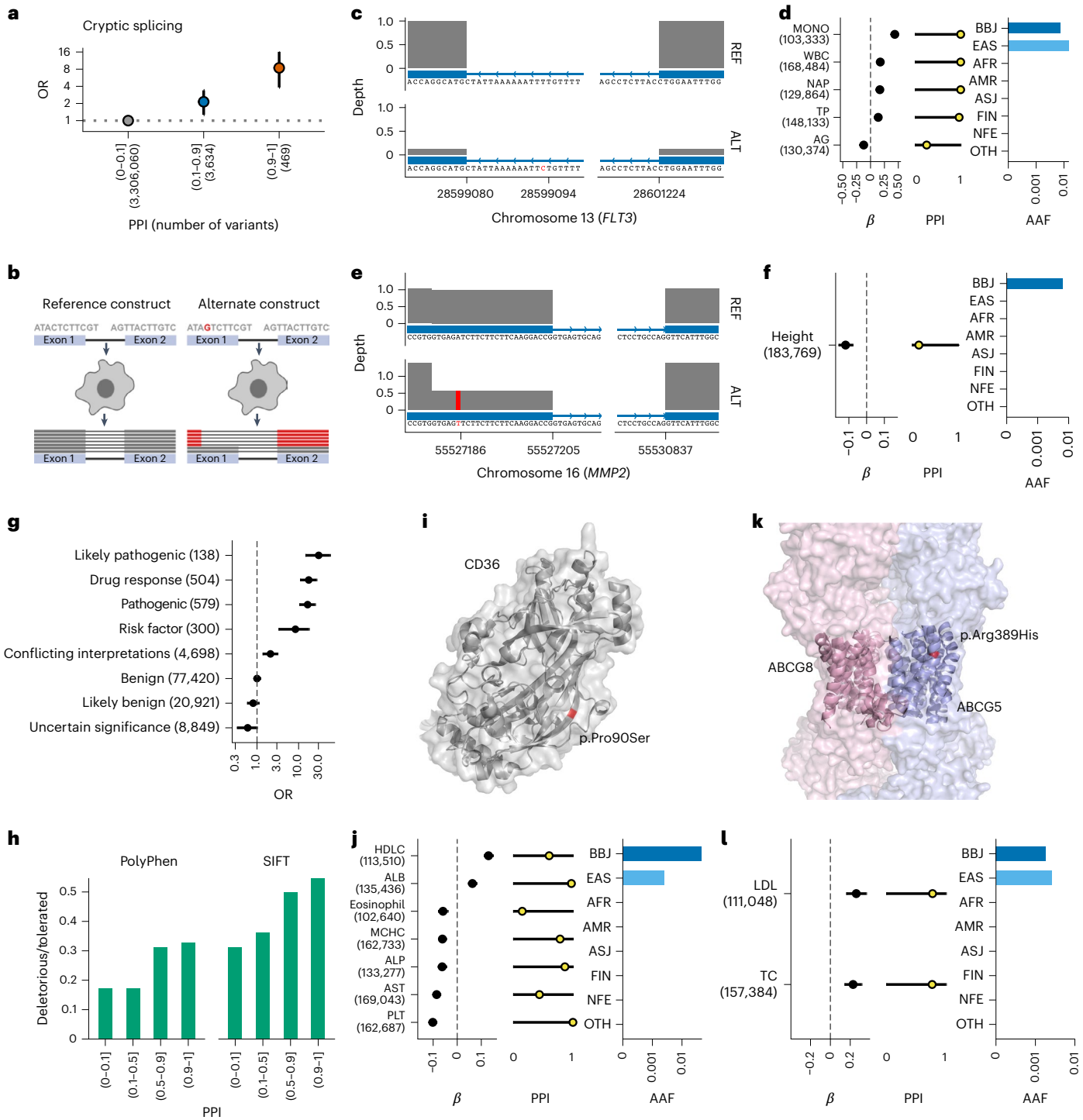


Fig. 3 | Rare population-specific putative causal splice variants and pathogenic variants associated with human quantitative traits. **a**, Enrichment of putative cryptic splice variants among variants with high PPI. The vertical axis indicates the OR and 95% CI of the cryptic splice variants (Splice-AI delta score > 0.2) for each PPI bin (the horizontal axis) to the lowest PPI bin. The OR and 95% CI were estimated using a Fisher's exact test. The number of variants included in the analysis is shown after the PPI bins. **b**, Schematic representation of the in vitro splicing assay. **c**–**f**, Schematic representation of alternative splicing, effect size, PPI and population frequency of the cryptic splice variant rs76080105 (*FLT3*, **c**,**d**) and rs141440582 (*MMP2*, **e**,**f**). The error bar for the β estimates indicates the 95% CI. The number of individuals included in the analysis is shown after the trait names. The horizontal axes indicate the genomic coordinate. The vertical axes indicate the exon coverage of the RNA sequence from the reference construct (top) and the alternate construct (bottom). Variant sites are indicated in red. **g**, Enrichment of ClinVar variants among variants with a high PPI. The vertical axis indicates the categories in ClinVar. The horizontal axis indicates the OR of a high PPI using benign variants as reference and the 95% CI estimated using a Fisher's exact test. The number of variants included in the analysis is shown after the variant annotations. **h**, Fraction of deleterious to tolerated variants evaluated using PolyPhen or sorting intolerant from tolerant (SIFT) in each PPI bin (horizontal axis). **i**, Schematic representation of the *CD36* locus where rs75326924 is located. **j**, β estimates, PPI and AAF of rs75326924. **k**, Schematic representation of the *ABCG5* locus where rs119480069 is located. **l**, β estimates, PPI and AAF of rs119480069. The AAF was obtained from the gnomAD dataset. The number of individuals included in the association analysis is found in Supplementary Table 1; the abbreviations for the phenotypes are found in Supplementary Table 2.

Another plausible example is a Japanese-specific rare deleterious missense variant of *USP47* associated with glucose levels (Extended Data Fig. 3 and Supplementary Table 4). USP47 was reported to be associated with several cancers in humans, but was not reported in the context of glucose levels. To support this finding, knockout of *Usp47* in mice resulted in increased glucose levels (Supplementary Note 4.1).

We also found other new associations between quantitative traits and missense variants that are rare and specific to, or more prevalent in, East Asians (EAS). These include associations of *ARHGAP36* with sodium levels, *RFWD2* with basophil counts, *S1PR4* with segmented neutrophil counts, *EVC* with estimated glomerular filtration rate (eGFR), *MYCT1* with red blood cell count, *EGLN1* with eGFR, *STAB2* with activated partial thromboplastin time and *SLC12A3* with chloride levels (Supplementary Note 4.1 and Extended Data Fig. 3). These new associations deepen our understanding of mechanisms underlying complex traits by providing highly likely causal genes and variants. In line with these findings, putative causal variants (PPI > 0.9) were significantly enriched in protein-altering or protein-truncating variants (PTVs) ($OR_{protein\text{-}altering} = 45$ (95% CI = 37–55), $P = 5.4 \times 10^{-147}$; $OR_{protein\text{-}truncating} = 106$ (95% CI = 56–184), $P = 2.7 \times 10^{-22}$; Extended Data Fig. 3 and Supplementary Tables 7 and 8).

New population-specific noncoding associations

We also found other new associations with noncoding variants, including Japanese-specific rare variants. New associations of noncoding variants included genes whose functions are largely unknown or known in limited contexts that are not associated with quantitative traits. In particular, we could connect long noncoding RNAs with quantitative traits. rs78568419 in *LINC00670*, an EAS-specific variant (also present at very low frequency in admixed American populations), was associated with platelet (PLT) count (Supplementary Table 4 and Extended Data Fig. 4). This long noncoding RNA is called cardinal and is expressed mainly in arterial tissues (coronary artery and aorta in the Genotype-Tissue Expression (GTEx) project[12]). In line with these findings, this variant showed a pleiotropic association with coronary artery disease ($P = 0.001$)[7]. Another example is the association of an EAS-specific variant in *LINC01094* (long intergenic non-protein-coding RNA 1094) with total cholesterol (TC) and high-density lipoprotein cholesterol (HDLC) (Supplementary Note 4.1). *LINC01094* has been associated with gastric cancer and renal cell carcinoma[18]. Other examples include associations of zinc-finger protein genes, such as *ZNF365* with HDLC, *ZNF787* with basophil count, *ZNF423* with hematocrit and hemoglobin, *ZNF468* with eGFR and blood urea nitrogen and *ZNF444* with white blood cell (WBC) count (Supplementary Table 4 and Supplementary Note 4.1).

Previously unreported associations of noncoding variants in genes of known function include the association of an upstream variant of *CD118* with low-density lipoprotein cholesterol (LDL) levels (Extended Data Fig. 4); *CD118* encodes a leukemia inhibitory factor receptor; this variant was not present in Europeans. Other examples include the association of hematocrit and hemoglobin with an upstream variant in *HEY1* that is highly specific to Asians, and the association of eosinophils with an *ETV6* variant (Extended Data Fig. 4 and Supplementary Note 4.1). *HEY1* encodes a crucial transcription factor involved in the NOTCH pathway, which was suggested to have critical roles in erythropoiesis[19]. *ETV6* is implicated in myeloid lymphoma; several *ETV6* fusion protein-positive acute myeloid lymphomas have been associated with clonal eosinophilia[20].

We also found new associations of noncoding variants in genes relevant to complex traits. These include an association between glucose levels and an EAS-specific rare variant upstream of *PAX4* (Extended Data Fig. 4 and Supplementary Note 4.1). As PAX4 is a master regulator of β-cells in the pancreas, this association suggests that this variant affects the development or function of β-cells via altered *PAX4* expression or activity, resulting in increased glucose levels even in individuals without diabetes. Other examples include an association of an intronic variant in *AQP1*, more frequent in EAS than other populations, with eGFR and serum creatinine levels. *AQP1* is a widely expressed water channel, especially in the kidney. Other associations include *ATM* with hemoglobin and hematocrit, *SIRT1* with hemoglobin, *RRAS2* with PLTs and *CD163* with aspartate aminotransferase (AST) levels (Supplementary Table 4 and Supplementary Note 4.1).

Characterization of putatively causal noncoding variants

We found reasonable enrichment of noncoding causal variants for DNase I hypersensitivity sites (DHS) and consensus footprints (CFPs) (Fig. 2a). To further assess the functionality of these noncoding regulatory elements quantitatively, we applied a deep-learning-based method to predict the pathogenicity (disease impact score) of noncoding variants. The disease impact score showed a strong positive association with PPI ($P < 2.9 \times 10^{-30}$; Fig. 2b, Extended Data Fig. 5 and Methods). A typical example was rs146018792 (MAF = 0.48%), a rare Japanese-specific noncoding variant in an intron of *CCND3* significantly associated with red blood cell-related traits (mean corpuscular volume (MCV) and mean corpuscular hemoglobin (MCH); $P_{MCV} = 7.8 \times 10^{-14}$, $PPI_{MCV} = 1.00$; $P_{MCH} = 8.5 \times 10^{-11}$, $PPI_{MCH} = 0.87$). rs146018792 is in a CFP within a myeloid-specific DHS (Fig. 2c,d). This variant had one of the highest disease impact scores (99.98 percentile; Extended Data Fig. 5). Specifically, this variant strongly decreased the affinity of the cFos/JUN transcription factor in the myeloid cell line K562 (Extended Data Fig. 5).

High-impact variants were strongly constrained across protein-truncating, protein-altering and noncoding variants (Fig. 2e,f). Importantly, we found that these putative causal variants (especially those with high PPI) were highly specific to EAS (Fig. 2g). We observed an array of rare, noncoding variants with comparable effect sizes to coding variants. As an example, rs542962114, a Japanese-specific rare noncoding variant located upstream of *LDHB*, was significantly associated with lactate dehydrogenase (LDH) levels with a large effect size (Extended Data Fig. 6; MAF = 1.13%, $\beta_{LDH} = -0.317$, $P_{LDH} = 1.6 \times 10^{-70}$). As most causal variants are noncoding, if we compare the number of variants, more than twice as many noncoding high-impact associations as coding variants were observed (Supplementary Note 5). In line with this finding, among all the putative causal associations (PPI > 0.1), a noncoding variant showed the largest effect size (rs33981098 and MCV; $\beta_{MCV} = -1.67$, $P_{MCV} = 1.2 \times 10^{-89}$, $PPI_{MCV} = 1.00$). rs33981098 is a known noncoding pathogenic variant for β-thalassemia located upstream of *HBB* (hemoglobin B)[21]. These findings underscore the importance of clarifying the mechanisms underlying causal noncoding variants.

Thus, high-quality whole-genome imputation enabled us to assess the impact of these rare noncoding variants on human phenotypes and rare coding variants at the population scale.

New population-specific causal variants in known loci

We also found new EAS or Japanese-specific causal variants (coding and noncoding) in genes previously known for their associations with quantitative traits (variant-level new associations). We found seven signals in the *PCSK9* locus associated with LDL (PPI > 0.1), including four population-specific rare coding variants. A very rare new noncoding variant showed the strongest effect size among the seven variants (Extended Data Fig. 7 and Supplementary Note 4.2).

Among these associations, we found rare Asian-specific variants as probably causal via altered splicing. We observed strong enrichment of predicted cryptic splicing variants by Splice-AI (predicted cryptic splice scores > 0.2) in variants with high PPI (PPI > 0.9, OR = 8.4 (3.8–16.1), $P = 2.1 \times 10^{-6}$; Fig. 3a). rs76080105 is an intronic variant in *FLT3* (ENST00000241453:c.2208-14A>G, MAF = 0.76%), which encodes a tyrosine kinase and whose distinct cryptic splicing variant was recently reported to cause autoimmune thyroid disease in Europeans[22]. rs76080105 was predicted to cause a splice acceptor loss; we found significant associations with several immunological traits supported

https://doi.org/10.1038/s41588-024-01913-5

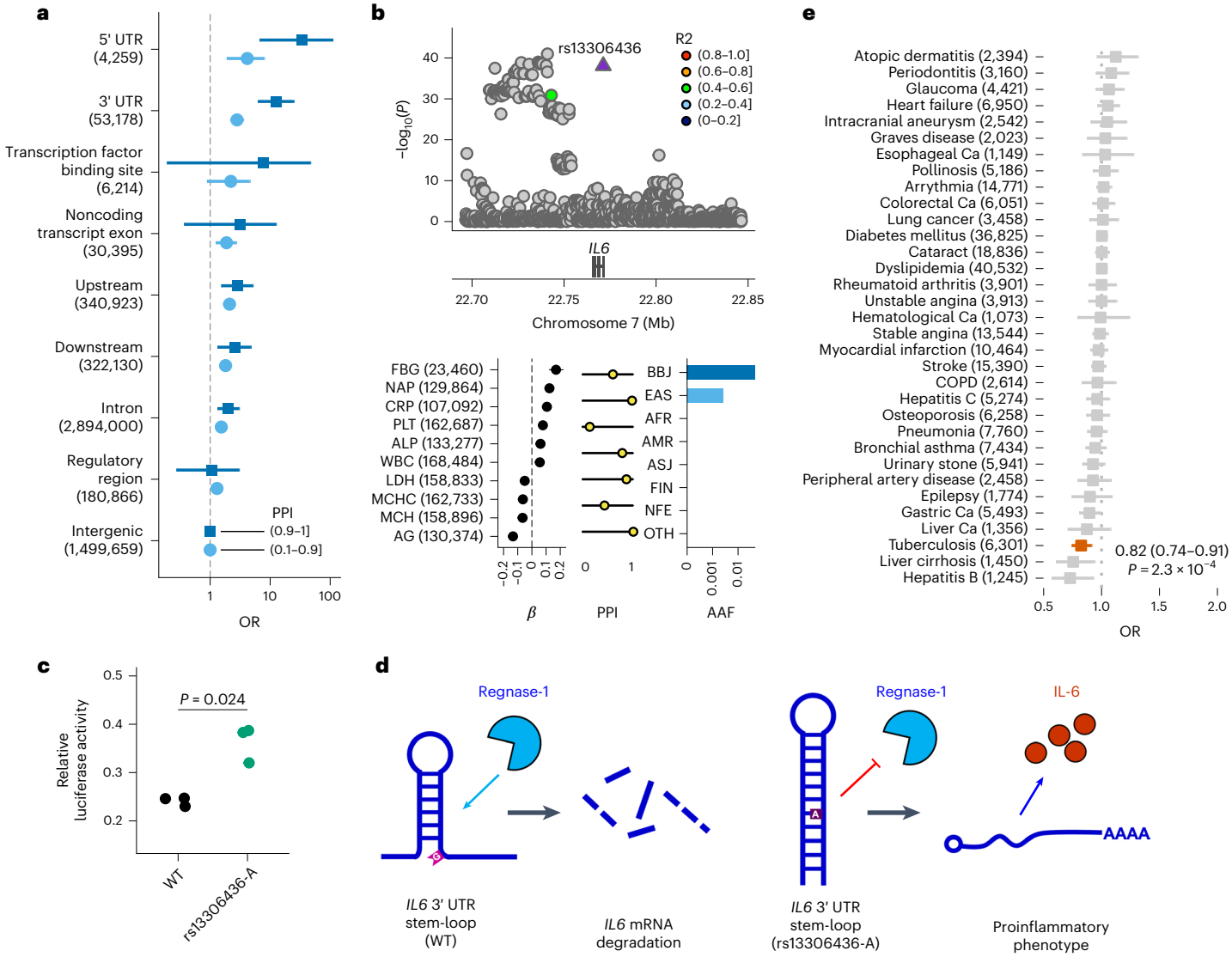


Fig. 4 | Enrichment of putative causal noncoding variants for functional annotations and a new mechanism of causal variants in 3′ UTR. a, Enrichment of causal noncoding variants for functional annotations. Each point and error bar indicates the OR of variants with a high PPI ((0.1, 0.9] or (0.9, 1]) and the 95% CI, respectively. The 95% CIs were estimated using a Fisher's exact test. **b**, Regional association plot and strong associations of the *IL6* locus. The horizontal axis indicates the genomic coordinates and the vertical axis indicates a negative $\log_{10}(P)$. Statistical significance was tested using a linear mixed model. The displayed *P* values are two-sided and were not adjusted for multiple testing. The β estimate, PPI and AAF in the global population of rs13306436 are shown. The error bar for the β estimates indicates the 95% CI. The number of individuals included in the analysis is shown after the trait name. **c**, rs13306436 showed resistance to regnase-1-mediated inhibition of a *IL6* 3′ UTR reporter. Overexpression of regnase-1 (10 ng per well) decreased expression of the reporter harboring the *IL6* 3′ UTR of both the wild-type (WT) (G) and variant (A) alleles of rs13306436, but the variant (A) allele of rs13306436 exhibited less of a decrease. The results are representative of experiments carried out in triplicate. Statistical significance was assessed using two-sided *t*-test. **d**, Working hypothesis of rs13306436 altering the posttranscriptional regulation of *IL6* expression. Regnase-1 recognizes the stem-loop structure within the 3′ UTR of *IL6* and leads to mRNA degradation. rs13306436 is located close to the stem-loop sequence; the variant (A) allele is more structured, which might suppress regnase-1-mediated degradation, thereby making the mRNA more stable (Supplementary Note 6.3). Short transcripts indicate degraded ones. **e**, OR for 29 diseases among carriers of rs13306436 in unrelated BBJ participants. Case counts are shown after the outcomes (n_{Total} = 169,020). The squares indicate the OR; the error bars indicate the 95% CI. Statistical significance was tested using a logistic regression with a two-sided test at $P < 0.05/29$. The displayed *P* values were not adjusted for multiple testing. The AAF was obtained from the gnomAD dataset. The number of individuals included in the association analysis is found in Supplementary Table 1; the abbreviations for the phenotypes are found in Supplementary Table 2.

by high PPI (Supplementary Note 4.2). We also found that rs76080105 was associated with rheumatoid arthritis and systemic lupus erythematosus in the Japanese population[23] (Supplementary Note 4.2). We experimentally validated this cryptic splice alteration (Fig. 3b–d; $P = 1.26 \times 10^{-6}$, Fisher's exact test). Another example is rs141440582, a rare missense variant in *MMP2* (ENST00000219070:c.1453A>T, p.I485F), MAF = 0.63%), which is associated with height ($P = 6.9 \times 10^{-9}$, PPI = 0.15). This missense variant is predicted to introduce a splice donor gain, resulting in a 25-bp frameshift deletion, which we validated experimentally ($P = 2.9 \times 10^{-11}$, Fisher's exact test; Fig. 3e,f and Supplementary Note 4.2). In agreement with our observation, *Mmp2* knockout in mice resulted in short stature and abnormal bone formation[24].

High PPI variants enriched in pathogenic variants

We observed a 15-fold enrichment of pathogenic variants in ClinVar among putative causal variants (PPI > 0.1, $P = 2.2 \times 10^{-10}$; Fig. 3g,h and Supplementary Table 9). All these clinically determined pathogenic variants were connected with quantitative trait-relevant diseases

(Supplementary Note 5.2). For example, rs75326924, a missense variant in *CD36* and known to be causal for *CD36* deficiency[25], showed putative causal associations with multiple quantitative traits, including PLT count, fatty acids, ejection fraction and heart failure (Fig. 3i,j and Supplementary Note 5.2), in line with the biological functions of CD36. rs119480069, a known pathogenic missense variant in *ABCG5*, showed causal associations with TC and LDL (Fig. 3k,l).

Functional annotations for noncoding causal variants

To further elucidate the consequences of putative causal noncoding variants on functional annotations that are in line with previous studies[26], we assessed the overlap of putative causal noncoding variants in DHS and CFP regions and found significant enrichment in a phenotype-relevant tissue-specific manner (Supplementary Table 11 and Extended Data Fig. 8). One of the most significant enrichments was observed in height-associated noncoding variants in musculoskeletal-specific DHS (OR_{Height} = 3.2 (2.4–4.1), $P_{Height} = 7.2 \times 10^{-15}$). Noncoding variants associated with hematological traits and antibody production were significantly enriched in myeloid-specific and lymphoid-specific DHS, respectively. We also found significant enrichment of causal variants for causal expression quantitative trait locus (eQTL) variants in the GTEx (Extended Data Fig. 9 and Supplementary Note 6.1).

We performed an enrichment analysis of functional annotations of noncoding variants neither in DHS nor CFPs to explore potential mechanisms. We found that such variants with high PPI are strongly enriched in the 3′ UTR or 5′ UTR of transcripts, suggesting crucial roles and distinct mechanisms of these regions on quantitative traits and underscoring diverse mechanisms of causal variants (Fig. 4a). These enrichments were also observed in the UK Biobank (UKB) (Extended Data Fig. 9 and Supplementary Note 6.2).

One such example was rs13306436, a rare EAS-specific variant in the 3′ UTR of *IL6*, associated with ten traits with high PPI (PPI > 0.9: fibrinogen (FBG), NAP, C-reactive protein (CRP), PLT, MCH and AG; PPI > 0.1: alkaline phosphatase (ALP), WBC, LDH and mean corpuscular hemoglobin concentration; Fig. 4b), concordant with the multipotency of *IL6*. The direction of the effects of this variant suggested an increase in immunogenicity (increased FBG, CRP, NAP and WBC). This variant is located near the binding site of regnase-1, an RNA-binding protein targeting the 3′ UTR of transcripts to degrade mRNA and control protein levels[27,28]. We experimentally showed that a reporter carrying the *IL6* mRNA 3′ UTR with the rare minor allele rs13306436 was resistant to degradation by regnase-1 (Fig. 4c), suggesting that the mRNA structure is altered by this variant, resulting in stable mRNA, increased interleukin-6 levels and consequently increased immunogenicity (Fig. 4d and Supplementary Note 6.3). We found that this variant decreased the risk of tuberculosis (Fig. 4e), in agreement with a previous report that identified an increased risk of tuberculosis infection in *Il6* knockout mice[29]. As regnase-1 targets several immune-related genes, our results may suggest more potential 3′ UTR variants as targets of regnase-1. In line with these findings, genes in which we found probable causal variants at the 3′ UTR showed enrichment for the target genes of regnase-1 (hypergeometric test, $P = 5.2 \times 10^{-5}$; Supplementary Note 6.3).

Possibility of drug repurposing

Genes with putatively causal coding variants were enriched in known genes causing monogenic disorders (genes with 'pathogenic' variants in the ClinVar database) or drug targets (Extended Data Fig. 10). These genes were more frequently included in protein–protein networks, regardless of genes with coding and noncoding variants (Extended Data Fig. 10). Genes encoding currently available drug targets that showed new associations in the current study included *CACNA1S*, *RYR1*, *PDE10A*, *SIRT1* and *CYP19A1* (Supplementary Tables 13 and 14). These raise the possibility of repurposing drugs currently available to other phenotypes or diseases (or potential side effects of the currently available drugs), taking advantage of direct or indirect connections to drug targets via the molecular network.

Discussion

In the current study, we combined finely imputed genotypes with a high-density Japanese-orientated imputation reference panel, well-powered multi-trait GWAS in a homogeneous population, a sensitive algorithm to determine the likelihood of causality of variants in the associated loci and in silico and experimental functional analyses. Together, these enabled us to detect many new associations especially specific to EAS, characterize putative causal genes and variants, and find new mechanisms for how causal noncoding variants affect complex traits, despite the bias of existing resources for these purposes toward European populations.

The new associations and putative causal associations in this study will be valuable for further functional follow-up studies for several traits with high sensitivity (Supplementary Note 4 and Data availability). Our study also provides several insights into the genetic architecture of causal associations, especially for coding variants. Discovery of many new associations in rare causal variants indicates the presence of many population-specific rare variants (both for coding and noncoding variants, in line with previous studies[30,31]) and the importance of deeply analyzing large-scale single populations. We also found that many new associations were driven by common variants in Japanese or EAS populations that were much more frequent in these populations, not limiting to rare variants (Supplementary Note 4).

We found a possible new mechanism underlying causal variants at 3′ UTRs. Further analyses would expand the yet-to-be-identified mechanisms of noncoding variants. The noncoding putative causal variants included rare variants having strong effect sizes on the phenotypes comparable with damaging-coding variants. As most associations are driven by noncoding causal variants, our observations suggest that we could drastically extend potential intervention targets as therapeutic or preemptive options targeting noncoding causal variants.

Our results coincide with the current effort to transition from whole-exome to whole-genome space[32]. Continuous efforts to expand WGS in single populations by leveraging large sample sizes and extending to the global population would uncover further population-specific associations and variants, from which we can identify causal variants and mechanisms and advance efforts toward personalized medicine.

Online content

Any methods, additional references, Nature Portfolio 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/s41588-024-01913-5.

References

1. Akiyama, M. et al. Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. *Nat. Genet.* **49**, 1458–1467 (2017).
2. 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).
3. Christophersen, I. E. et al. Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation. *Nat. Genet.* **49**, 946–952 (2017).
4. Ghoussaini, M. et al. Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res.* **49**, D1311–D1320 (2021).

5. Ishigaki, K. et al. Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases. *Nat. Genet.* **52**, 669–679 (2020).
6. Kanai, M. et al. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* **50**, 390–400 (2018).
7. Koyama, S. et al. Population-specific and trans-ancestry genome-wide analyses identify distinct and shared genetic risk loci for coronary artery disease. *Nat. Genet.* **52**, 1169–1177 (2020).
8. Ozaki, K. et al. Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nat. Genet.* **32**, 650–654 (2002).
9. Terao, C. et al. GWAS of mosaic loss of chromosome Y highlights genetic effects on blood cell differentiation. *Nat. Commun.* **10**, 4719 (2019).
10. Huang, J. et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6**, 8111 (2015).
11. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
12. Aguet, F. et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
13. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
14. Hormozdiari, F. et al. Widespread allelic heterogeneity in complex traits. *Am. J. Hum. Genet.* **100**, 789–802 (2017).
15. Arvanitis, M., Tayeb, K., Strober, B. J. & Battle, A. Redefining tissue specificity of genetic regulation of gene expression in the presence of allelic heterogeneity. *Am. J. Hum. Genet.* **109**, 223–239 (2022).
16. Abell, N. S. et al. Multiple causal variants underlie genetic associations in humans. *Science* **375**, 1247–1254 (2022).
17. Berger, M., Geng, B., Cameron, D. W., Murphy, L. M. & Schulman, E. S. Primary immune deficiency diseases as unrecognized causes of chronic respiratory disease. *Respir. Med.* **132**, 181–188 (2017).
18. Luo, C. et al. LINC01094 promotes pancreatic cancer progression by sponging miR-577 to regulate LIN28B expression and the PI3K/AKT pathway. *Mol. Ther. Nucleic Acids* **26**, 523–535 (2021).
19. Robert-Moreno, A., Espinosa, L., Sanchez, M. J., de la Pompa, J. L. & Bigas, A. The notch pathway positively regulates programmed cell death during erythroid differentiation. *Leukemia* **21**, 1496–1503 (2007).
20. Montano-Almendras, C. P. et al. ETV6-PDGFRB and FIP1L1-PDGFRA stimulate human hematopoietic progenitor cell proliferation and differentiation into eosinophils: the role of nuclear factor-κB. *Haematologica* **97**, 1064–1072 (2012).
21. Takihara, Y., Nakamura, T., Yamada, H., Takagi, Y. & Fukumaki, Y. A novel mutation in the TATA box in a Japanese patient with β^+-thalassemia. *Blood* **67**, 547–550 (1986).
22. Saevarsdottir, S. et al. *FLT3* stop mutation increases FLT3 ligand level and risk of autoimmune thyroid disease. *Nature* **584**, 619–623 (2020).
23. Yin, X. et al. Meta-analysis of 208370 East Asians identifies 113 susceptibility loci for systemic lupus erythematosus. *Ann. Rheum. Dis.* **80**, 632–640 (2021).
24. Mosig, R. A. et al. Loss of MMP-2 disrupts skeletal and craniofacial development and results in decreased bone mineralization, joint erosion and defects in osteoblast and osteoclast growth. *Hum. Mol. Genet.* **16**, 1113–1123 (2007).
25. Hanawa, H. et al. Identification of cryptic splice site, exon skipping, and novel point mutations in type I CD36 deficiency. *J. Med. Genet.* **39**, 286–291 (2002).
26. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
27. Matsushita, K. et al. Zc3h12a is an RNase essential for controlling immune responses by regulating mRNA decay. *Nature* **458**, 1185–1190 (2009).
28. Mino, T. et al. Regnase-1 and Roquin regulate a common element in inflammatory mRNAs by spatiotemporally distinct mechanisms. *Cell* **161**, 1058–1073 (2015).
29. Ladel, C. H. et al. Lethal tuberculosis in interleukin-6-deficient mutant mice. *Infect. Immun.* **65**, 4843–4849 (1997).
30. Bomba, L., Walter, K. & Soranzo, N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biol.* **18**, 77 (2017).
31. Wainschtein, P. et al. Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nat. Genet.* **54**, 263–273 (2022).
32. Smedley, D. et al. 100,000 Genomes pilot on rare-disease diagnosis in health care—preliminary report. *N. Engl. J. Med.* **385**, 1868–1880 (2021).

[1]Laboratory for Cardiovascular Genomics and Informatics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [2]Medical and Population Genetics and Cardiovascular Disease Initiative, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [3]Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA. [4]Laboratory for Statistical and Translational Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [5]Laboratory for Bone and Joint Diseases, RIKEN Center for Medical Sciences, Tokyo, Japan. [6]Department of Orthopedic Surgery, Hokkaido University Graduate School of Medicine, Sapporo, Japan. [7]Laboratory for Pharmacogenomics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [8]Department of Cancer Biology, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [9]Laboratory of Complex Trait Genomics, Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan. [10]Department of Medical Chemistry, Graduate School of Medicine, Kyoto University, Kyoto, Japan. [11]Department of Orthopedic Surgery, Shimane University Faculty of Medicine, Izumo, Japan. [12]Department of Orthopedic Surgery, Keio University School of Medicine, Tokyo, Japan. [13]Department of Orthopedic Surgery, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan. [14]Department of Ocular Pathology and Imaging Science, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan. [15]Clinical Research Center, Shizuoka General Hospital, Shizuoka, Japan. [16]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. [17]Center for Genomic Medicine, Massachusetts General

Hospital, Boston, MA, USA. [18]Personalized Medicine, Mass General Brigham, Boston, MA, USA. [19]Department of Medicine, Harvard Medical School, Boston, MA, USA. [20]Laboratory for Genomics of Diabetes and Metabolism, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [21]Department of Psychiatry, School of Medicine, Fujita Health University, Toyoake, Japan. [22]Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan. [23]Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [24]Research Institute, National Center for Geriatrics and Gerontology, Obu, Japan. [25]Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [26]Department of Applied Genetics, School of Pharmaceutical Sciences, University of Shizuoka, Shizuoka, Japan. [27]These authors contributed equally: Satoshi Koyama, Xiaoxi Liu, Yoshinao Koike. ✉e-mail: chikashi.terao@riken.jp

Biobank Japan Project

Koichi Matsuda[22]

A list of members and their affiliations appears in the Supplementary Information.

Methods

Ethics oversights

All participants provided written informed consent according to the protocols approved by following institutional ethical committees: the RIKEN Center for Integrative Medical Sciences; the Institute of Medical Sciences; the University of Tokyo; the National Center for Geriatrics and Gerontology (NCGG); the Tohoku University Graduate School of Medicine; and Iwate Medical University.

Study cohorts

First, we included three different datasets constructed from the contemporary Japanese population (BBJ first cohort, BBJ second cohort, NCGG cohort). We subjected these datasets to imputation using our reference panel (described below) to obtain the results for harmonized variants and then meta-analyzed the results. Additionally, to maximize statistical power to identify new signals especially specific to the EAS population, we also analyzed the quantitative trait data of 53,083 individuals from the Tohoku Medical Megabank Organization (ToMMo) community-based cohort study (67K; Extended Data Fig. 1 and Supplementary Note 1).

The BBJ[33,34] is a nationwide hospital-based biobank with 12 collaborating medical institutions. The first cohort targeted 47 diseases and recruited 200,000 people between 2003 and 2013; the second cohort targeted 38 diseases and recruited 67,000 people between 2013 and 2018 (https://biobankjp.org/en/index.html). In this study, 12,098 people with available genotypes were included from the BBJ second cohort. The NCGG Biobank is a hospital-based biobank maintained by the NCGG since 2012. Participants were recruited from the NCGG hospital and nearby medical institutes (https://www.ncgg.go.jp/english/index.html). ToMMo is a population-based cohort in which study participants were recruited from the health checkups conducted in two prefectures of Northeastern Japan: Miyagi (n = 32,459) and Iwate (n = 20,906).

WGS and creation of the imputation reference panel

The procedures for WGS and reference panel construction are described elsewhere[35]. The 3,256 individuals sequenced are from the BBJ cohort. Briefly, 1,502 individuals were sequenced aiming at a 30× coverage (high coverage) and 1,786 at a 15× coverage (medium coverage) with a HiSeq 2500 (Rapid mode or V4, Illumina) or HiSeq X Five platform. Samples with low sequence quality or from closely related individuals were removed. Sequenced reads were aligned to a human reference genome (hg19) using the Burrows–Wheeler Aligner[36]; duplicated reads were removed. Then, we conducted joint genotype calling using HaplotypeCaller and GenotypeGVCFs implemented by the Genome Analysis Toolkit[37] (v.3.5-0, v.3.8-0 for high coverage, v.3.6-0 for medium coverage, v.3.8-0 for joint calling) according to germline short variant discovery best practice workflows. We removed variants with: (1) read depth (DP) < 5 from high coverage samples; DP < 2 from medium coverage samples; (2) genotype quality (GQ) < 20; (3) DP > 60 and GQ < 95 from high coverage; (4) failed in variant quality score recalibration. The procedures for reference panel construction were as follows. From WGS VCF files generated as above, we removed multiallelic or monomorphic sites, singleton variants and variants deviating from Hardy–Weinberg equilibrium ($P < 1 \times 10^{-6}$). The genotypes from the 1000 Genomes Project[13] (phase 3, v.5) were similarly processed. Then, these datasets were merged using IMPUTE2 (ref. 38) v.2.3.2. For the X chromosome, we used BEAGLE[39] v.4.1 to merge the male WGS genotypes, and then combined them with the female genotypes. For the ToMMo dataset, the 3,552 Japanese genomes in ToMMo were sequenced using the HiSeq platform and the sequenced reads were aligned to the human reference genome (GRCh37). Genotypes were called using the Genome Analysis Toolkit best practice pipeline and used as a reference panel (ToMMo 3.5KJPNv2)[40].

Haplotype phasing and imputation

Genotypes were determined using either (1) the Illumina HumanOmniExpressExome BeadChip or (2) a combination of Illumina HumanOmniExpress and HumanExome arrays for the BBJ first cohort. For the BBJ second cohort, genotypes were determined using the HumanOmniExpressExome BeadChip; for the NCGG cohort, genotypes were determined using the Illumina AsianScreeningArray (the NCGG data were obtained from the NCGG Biobank database). Quality control (QC) was performed by removing individuals who withdrew consent, had call rates lower than 98%, gender mismatch or non-East Asian ancestry. Any samples overlapping with those in the reference panels were also removed. QC on variants excluded those with a call rate lower than 99%, fewer than five heterozygotes, extreme deviation from the Hardy–Weinberg equilibrium ($P < 1 \times 10^{-6}$) and palindromic variants. We also compared the array genotype and WGS to exclude variants with a concordance rate lower than 99.5%. After QC, the BBJ first cohort, BBJ second cohort and NCGG cohort were separately phased using SHAPEIT2 (ref. 41) (BBJ first cohort, v.2.837) or EAGLE2 (ref. 42) (BBJ second cohort and NCGG cohort, v.2.39), followed by whole-genome imputation using Minimac4 (ref. 43) (v.1.0.0).

For the ToMMo dataset, the array dataset in PLINK binary format (659,326 SNPs) and the imputed genotype dataset in the Oxford BGEN format (54,041,917 variants) for 53,365 study participants were obtained. The genotyping and imputation procedures have been described elsewhere[40]. All samples were genotyped using the Affymetrix Axiom Japonica array. After QC, autosomal variants were phased using SHAPEIT2 (v2.R837) and subsequently imputed using IMPUTE2 (v.2.3.2). We conducted further QC and excluded samples with (1) an array call rate lower than 97% or (2) non-Japanese ancestry identified using principal component analysis with all samples from the 1000 Genomes Phase III dataset. For variants, we excluded variants with an imputation INFO score lower than 0.3 from the downstream analysis. The final dataset consisted of 37,167,587 variants for 53,083 individuals.

Variant annotation

We used VEP[44] v.87 to annotate the tested variants. To obtain a single annotation for a variant, we used the --pick option to prioritize annotation on the canonical transcript. rsIDs were assigned using VEP; if an rsID was not assigned, we annotated the variant as chromosome:position:reference allele:alternate allele. The summary and definition of variant annotation are summarized in Supplementary Table 15.

Quantitative phenotype curation, QC and normalization

Quantitative phenotypes were extracted from the BBJ participant health records. The NCGG phenotype data were obtained from the NCGG Biobank database. The ToMMo data were obtained from the ToMMo database. Raw phenotype data were filtered using the mean ± four s.d. Then, phenotype-specific corrections were applied as follows. For individuals taking a lipid-lowering agent, TC and LDL were divided by 0.8 and 0.7, respectively. For individuals taking antihypertensive agents, systolic and diastolic blood pressure were added (15 and 10 mmHg, respectively). Phenotype-specific exclusion criteria were also applied as follows. Individuals taking an antiuremic agent were excluded from the uric acid analysis; individuals taking warfarin were excluded from the prothrombin time analysis; and individuals with diabetes were excluded from the HbA1c and blood sugar analyses. The raw phenotypes were regressed and residualized according to age, sex and principal components (PCs) 1–10 used as covariates. Additionally, we introduced 47 target disease statuses for the BBJ first cohort, 38 target disease statuses for BBJ second cohort and prefecture of enrollment for the ToMMo cohort into the model[6]. Then, residuals were inverse-rank-normalized and used as quantitative phenotypes. After normalization, we conducted association analysis using the BOLT-LMM algorithm without covariates. The distributions of phenotypes are summarized in Supplementary Tables 1 and 2.

Quantitative and case-control association analysis and meta-analysis

For the quantitative phenotypes, we applied BOLT-LMM[45] (v.2.3.4) for the single-variant association test in each cohort separately.

When the model was not converged, we applied a linear regression model implemented in PLINK2 (ref. 46) software excluding related individuals (defined as PI-HAT > 0.25). For the X chromosome, males and females were tested separately and meta-analyzed using inverse-variance-weighted fixed-effect meta-analysis implemented in the METAL[47] software. The applied model is summarized in Supplementary Table 1. Then, the results were meta-analyzed with the METAL software using an inverse-variance-weighted fixed-effect meta-analysis. After the meta-analysis, the variant with an overall MAF < 0.1% or *P* value for heterogeneity < 1×10^{-6} were excluded from the results. For the case-control analysis, we conducted the logistic regression analysis implemented in PLINK2 to associate the genetic dosage and case-control status registered in the BBJ first cohort, introducing age, sex and PCs 1–10 as covariates excluding related individuals (PI-HAT > 0.25).

LD Score regression

Lambda GC, LD Score regression intercept and its ratio were determined using the LDSC software[48] (v.1.0.1). We used the LD Score calculated from the 1000 Genomes Project EAS individuals using the LDSC software.

Locus definition

Genome-wide significant loci were determined as follows: (1) extracting variants with $P < 5 \times 10^{-8}$; (2) adding a 5×10^5 base length to each position of these variants bilaterally; (3) merging any overlapping regions. For the variants located in the major histocompatibility complex region (defined as chromosome 6 coordinates from 25000000 to 35000000), 1×10^6 base length was added to the position of variants with genome-wide significance. If the locus did not contain coordinates with previously reported genome-wide significant variants, the locus was annotated as a new.

Statistical fine-mapping

We applied FINEMAP[49] (v.1.4) for each genome-wide significant locus. We used the meta-analysis results of the primary datasets (BBJ first, BBJ second and NCGG; Supplementary Note 1). We uniformly used the genotype dosage of the first cohort of the BBJ to calculate LD matrices using the Ldstore software (v.2.0) as it was the largest cohort in this study. The maximum number of causal variants in the locus was used as ten in the first round. If the number of causal variants was estimated at ten, we reran FINEMAP using 20 as the maximum number of causal variants (Supplementary Note 3). To control fine-mapping quality, we first excluded 48 loci overlapping the major histocompatibility complex region (chromosome 6 25000000–35000000) because of its extensive LD structure[50]. In addition, we removed 16 loci where the causalities of the variants were not supported by the conditional analysis. In total, we completed statistical fine-mapping for 3,309 of the 3,390 genome-wide significant loci (97.6%). The marginal PPI was used for each variant throughout the study. Detailed processes are described in Supplementary Note 3. For the UKB, we downloaded the summary statistics generated previously (http://www.nealelab.is/uk-biobank) for 37 corresponding phenotypes. We used the LD matrix calculated using the dosage data for White British individuals in the UKB using LDstore. Otherwise, we defined the loci, ran FINEMAP and processed the output data as described for the BBJ.

Estimation of enrichment and PPI

For each PPI bin, the ORs of the variants annotated as 'high' or 'moderate' (Supplementary Table 15) by the VEP software to the variants annotated as 'modifier' were calculated in comparison with the lowest PPI bin (0–0.1) and tested using a Fisher's exact test.

ClinVar annotation

We downloaded the VCF file from the ClinVar[51] website (https://www.ncbi.nlm.nih.gov/clinvar, 27 January 2020). For each PPI bin, the ORs of variants with each level of clinical significance to variants with 'benign' annotation were calculated in comparison with the lowest PPI bin (0–0.1) and tested using a Fisher's exact test.

Protein visualization

We used the PyMOL software (https://pymol.org/2) to visualize the three-dimensional (3D) structure of proteins. We obtained the 3D protein structures from the Protein Data Bank website (https://www.rcsb.org). The following accession codes were used for the visualization: 6KN8, 5LGD and 5DO7 for *Troponin-T*, *CD36* and *ABCG5/ABCG8*, respectively.

Drug target

The list of genes encoding drug targets was defined using a previous report[52]. We counted the number of genes with high-PPI coding and noncoding variants overlapping such drug target genes. A Fisher's exact test was used to estimate the OR and 95% CI. The *P* value was calculated by comparing genes with the highest PPI > 0.1 and highest PPI ≤ 0.1.

Protein–protein interactions

Protein–protein interaction data were downloaded from the STRING[53] website (https://string-db.org/). High-confidence protein–protein interactions were determined using a combined score greater than 0.9. We counted the number of edges from each gene within this set of interactions. Then, we computed the mean number of interactions for genes in each PPI bin. A Wilcoxon rank-sum test was used to test the difference in the number of protein–protein interactions between genes with a gene PPI > 0.1 and genes with a gene PPI ≤ 0.1.

DeepSEA and disease model

We applied the DeepSEA-based disease impact score predicting model[54] to all noncoding variants in genome-wide significant loci ($n = 7{,}289{,}211$, https://hb.flatironinstitute.org/asdbrowser/about). The baseline DeepSEA[55] model returned the probability differences for 2,002 epigenetic features. Then, the disease impact score was estimated from these predicted probability differences as a single scalar value for each variant. We estimated the effect sizes of PPI on the disease impact score; we conducted linear regression modeling in the variants with the highest PPI in the loci as follows:

$$\text{Disease impact score} \sim \beta_1 \text{minor allele frequency} + \beta_2 \text{PPI}$$

Splice-AI

We downloaded the precomputed Splice-AI score (https://basespace.illumina.com). The precomputed score file contained all the substitutions around the exon–intron boundary, provided the delta score and predicted the position for the alternative splicing for these substitutions. We annotated all the variants in the genome-wide-significant loci using the score. As the cutoff, we applied a delta score greater than 0.2 (high sensitivity cutoff[56]).

Splicing assay

The precise method for the in vitro alternative splicing assay was described elsewhere[57]. Briefly, we cloned exon–intron–exon structures harboring reference and alternate alleles for the predicted cryptic splice variant on the minigene construct. Each construct was transfected into HEK 293T cells. After 24 h of incubation, RNA was extracted and sequenced using the Illumina MiSeq platform. Sequenced reads were processed using our open-source software (https://github.com/SplicingVariant/SplicingVariants_Beta) to quantify the number of non-splicing, normal splicing and aberrant splicing. We calculated the *P* value using a Fisher's exact test by normalizing the analyzed reads to 100 for each allele. The oligonucleotide sequences used in this study are provided in Supplementary Table 24.

Regulatory element and tissue enrichment analysis

We obtained the definitions of DHS and CFP from the ENCODE3 projects[58]; then, we mapped the positions of these elements to the hg19 coordinates using the liftOver software. Next, we counted the overlap of variants with each regulatory element and calculated the OR of variants in each PPI bin to the lowest PPI bin. For each DHS–phenotype pair, we created a contingency table including: (1) variants with a high PPI (0.1–1.0] and located in the DHS of interest; (2) variants with a high PPI and not located in any DHS; (3) variants with a low PPI [0–0.1] and located in the DHS of interest; (4) variants with a low PPI and not located in any of the DHS. We tested the OR using a Fisher's exact test. For the tissue enrichment analysis, P values were Bonferroni-adjusted; only the associations with an adjusted $P < 0.05$ were considered significant and are displayed in Extended Data Fig. 8.

For noncoding variants not in CFP or DHS, we tested the enrichment of functional annotations. The ORs of variants with a high PPI ((0.1–0.9] and (0.9–1.0]) in reference to the intergenic variants with a low PPI [0–0.1] were tested using a Fisher's exact test.

Population-specific alleles defined using gnomAD

We download the site VCF files, including the allele frequency information, from the gnomAD[59] website (v.2.1.1, https://gnomad.broadinstitute.org/downloads). We extracted the variants of interest; if a variant was found only in a single population, we defined it as a population-specific variant. We excluded variants found only in the Japanese WGS (current dataset) from the analysis.

Plasmids

To construct the luciferase reporter vector, the human *IL6* 3′ UTR sequence (1–428) was amplified using genomic DNA derived from HeLa cells as a template and was inserted into the pGL3-Promoter vector (Promega Corporation) using the In-Fusion HD Cloning Kit (Takara Bio). The rs13306436 point mutation was introduced using the QuikChange Lightning Site-Directed Mutagenesis Kit (Agilent Technologies). The regnase-1 expression vector was constructed by inserting the coding sequence of regnase-1 into the pcDNA3.1(+) vector (Invitrogen).

Luciferase assay

Both *IL6* WT and *IL6* rs13306436 mutant reporter plasmids were cotransfected with *Renilla* luciferase plasmid into HeLa cells using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. A pGL3-Promoter vector without *IL6* 3′ UTR (empty) was used as the control. After 24-h incubation, cells were lysed and the luciferase activity was determined using the Dual-Luciferase Reporter Assay system (Promega Corporation). We further examined the luciferase activity under regnase-1 overexpression. The fold of repression due to regnase-1 was calculated by normalizing the luciferase level of regnase-1-overexpressing cells with that of empty vector transfected cells.

Statistical analysis of luciferase assay and enrichment for target genes of regnase-1

Data are presented as the mean ± s.d. Statistical significance was calculated with a Student's t-test. The significance level at $P < 0.05$ (*) is shown. We analyzed the enrichment of genes where causal variants were in 3′ UTR for the target genes of regnase-1, which were experimentally validated[28]. We used a hypergeometric test for this enrichment (Supplementary Note 6.4).

Reporting summary

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

Data availability

The GWAS summary statistics and the results of statistical fine-mapping are available without any restriction at the Japanese ENcyclopedia of GEnetic associations by Riken website (http://jenger.riken.jp/en) and the National Bioscience Database Center (https://biosciencedbc.jp/en) under research ID hum0014. The imputation reference panel containing the 3,256 high-depth Japanese individuals will be made available to researchers at the National Bioscience Database Center under research ID hum0014 after approval by the Human Data Review Board. The protein 3D structure data were obtained from the Protein Data Bank (https://www.rcsb.org/). The human tissue expression data were obtained from the GTEx Portal (https://www.gtexportal.org/home/). The DNase1 hypersensitivity site and transcription factor footprints were obtained from public repositories (https://zenodo.org/records/3838751 and https://zenodo.org/records/3905306, respectively[60,61]). The chromatin immunoprecipitation data were obtained from the ENCODE website (https://www.encodeproject.org/). The allele frequency information for the diverse human populations was obtained from the gnomAD project website (https://gnomad.broadinstitute.org/). The list of clinically curated pathogenic variants was obtained from the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/).

Code availability

Standalone software to create the LD matrix (LD store) and conduct the statistical fine-mapping (FINEMAP) is available at http://www.christianbenner.com/. We deposited the custom analysis codes for the association analysis and fine-mapping at https://doi.org/10.5281/zenodo.10934238 (ref. 62). Further detailed scripts are available upon reasonable request to the corresponding author.

References

33. Nagai, A. et al. Overview of the BioBank Japan Project: study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).
34. Hirata, M. et al. Cross-sectional analysis of BioBank Japan clinical data: a large cohort of 200,000 patients with 47 common diseases. *J. Epidemiol.* **27**, S9–S21 (2017).
35. Akiyama, M. et al. Characterizing rare and low-frequency height-associated variants in the Japanese population. *Nat. Commun.* **10**, 4393 (2019).
36. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
37. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
38. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
39. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
40. Fuse, N. et al. Genome-wide association study of axial length in population-based cohorts in Japan: the Tohoku Medical Megabank Organization Eye Study. *Ophthalmology Sci.* **2**, 100113 (2022).
41. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2011).
42. 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).
43. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
44. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
45. Loh, P.-R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).

46. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
47. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
48. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
49. Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
50. Weissbrod, O. et al. Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020).
51. Landrum, M. J. et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–D985 (2014).
52. Finan, C. et al. The druggable genome and support for target identification and validation in drug development. *Sci. Transl. Med.* **9**, eaag1166 (2017).
53. Szklarczyk, D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
54. Zhou, J. et al. Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nat. Genet.* **51**, 973–980 (2019).
55. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
56. Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548 (2019).
57. Ito, K. et al. Identification of pathogenic gene mutations in *LMNA* and *MYBPC3* that alter RNA splicing. *Proc. Natl Acad. Sci. USA* **114**, 7689–7694 (2017).
58. Moore, J. E. et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
59. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
60. Meuleman, W. & Stamatoyannopoulos, J. A. Index and biological spectrum of accessible DNA elements in the human genome. *Zenodo* https://zenodo.org/records/3838751 (2019).
61. Viestra, J. & Stamatoyannopoulos, J. A. Global consensus map of human transcription factor footprints. *Zenodo* https://zenodo.org/records/3905306 (2020).
62. Satoshi, K. Population-specific putative causal variants shape quantitative traits. *Zenodo* https://zenodo.org/records/10934238 (2024).

Acknowledgements

We thank the study participants and the research and medical staff at the study sites and hospitals. We thank the study participants of ToMMo and the staff at the Center for Genome Platform Projects of ToMMo (no. 2019-0075) and its computational resources supported by the Japan Agency for Medical Research and Development under grant no. JP21tm0424601. The list of participating ToMMo members is available at https://www.megabank.tohoku.ac.jp/english/a210901/.

We thank N. Parrish at the RIKEN Center for Integrative Medical Sciences for editing the manuscript. We thank Y. Onouchi and K. Yamazaki at Chiba University, the Department of Public Health and the RIKEN Center for Integrative Medical Sciences for providing materials and support. S.K. is supported by the Japan Society for the Promotion of Science Overseas Research Fellowship. This study was supported by the Japan Agency for Medical Research and Development under grant nos. JP21ek0109555, JP21tm0424220, JP21ck0106642, 23ek0410114, 23tm0424225 and JP21ae0121030, the Japan Society for the Promotion of Science KAKENHI grant nos. JP20H00462, 22H03207 and 18H05278, the Medical Research Support Project of the Shizuoka Prefectural Hospital Organization and the BBJ project, which was supported by the Ministry of Education, Culture, Sports, Sciences and Technology of the Japanese Government and the Japan Agency for Medical Research and Development under grant nos. 17km0305002 and 18km0605001.

Author contributions

S.K., K.I. and C.T. conceived the study design. M.A., C.B., P.N., P.T.E., T.M., M.H., S. Ikegawa, O.T., K.I. and C.T. supervised the project. S.K., Y.K., K.H. and M.K. conducted the phenotyping. X.L., K.T., M.A., M.I., N.I., Y.M. and C.T. conducted the WGS and imputation. S.K., X.L., Y.K., K.H., M.K., S. Ito, N.O., H.S., S.Y., K.S., Y.I., C.B., M.I., N.I. and C.T. conducted the statistical analysis. W.L., K.A., O.T. and K.I. conducted the functional analyses. K.M., Y.M., K.I. and C.T. generated the BBJ data. S.N. and K.O. generated the NCGG data. S.K., K.I. and C.T. wrote the manuscript; all authors reviewed the manuscript and provided valuable edits.

Competing interests

P.N. reports research grants from Allelica, Amgen, Apple, Boston Scientific, Genentech/Roche and Novartis; personal fees from Allelica, Apple, AstraZeneca, Blackstone Life Sciences, Creative Education Concepts, CRISPR Therapeutics, Eli Lilly & Co, Foresite Labs, Genentech/Roche, GV, HeartFlow, Magnet Biomedicine, Merck and Novartis; and scientific advisory board membership of Esperion Therapeutics, Preciseli and TenSixteen Bio. He is scientific cofounder of TenSixteen Bio; holds equity in MyOme, Preciseli and TenSixteen Bio; and reports spousal employment at Vertex Pharmaceuticals, all unrelated to the present work. The other authors declare no competing interests.

Additional information

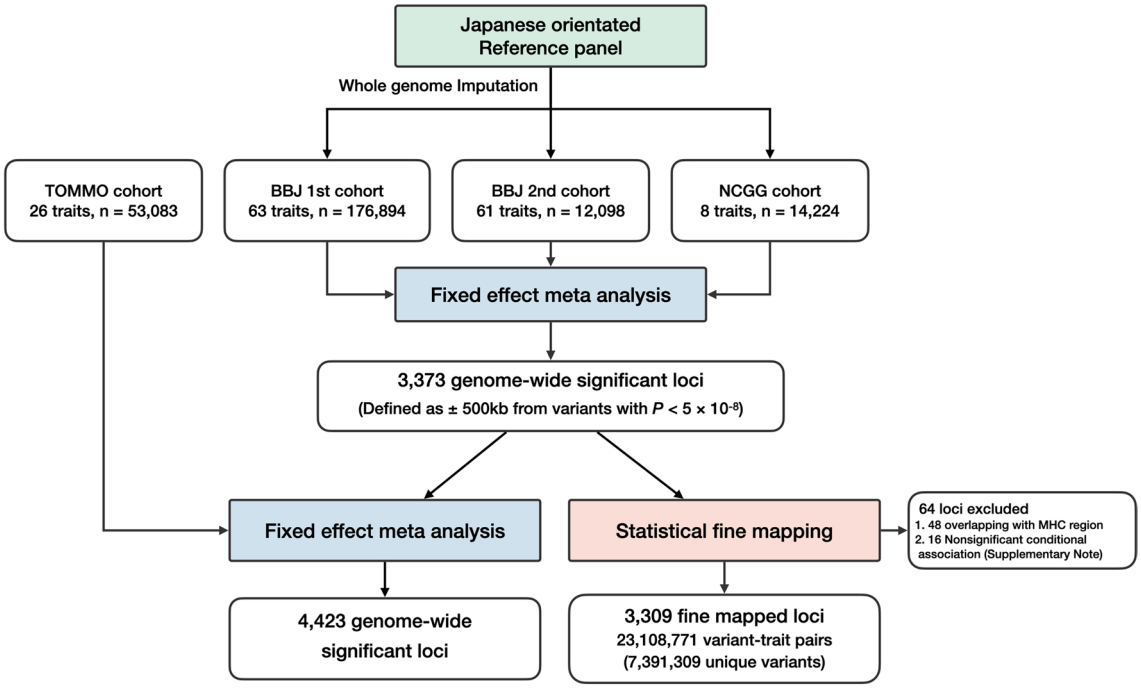
Extended data is available for this paper at https://doi.org/10.1038/s41588-024-01913-5.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-024-01913-5.

Correspondence and requests for materials should be addressed to Chikashi Terao.

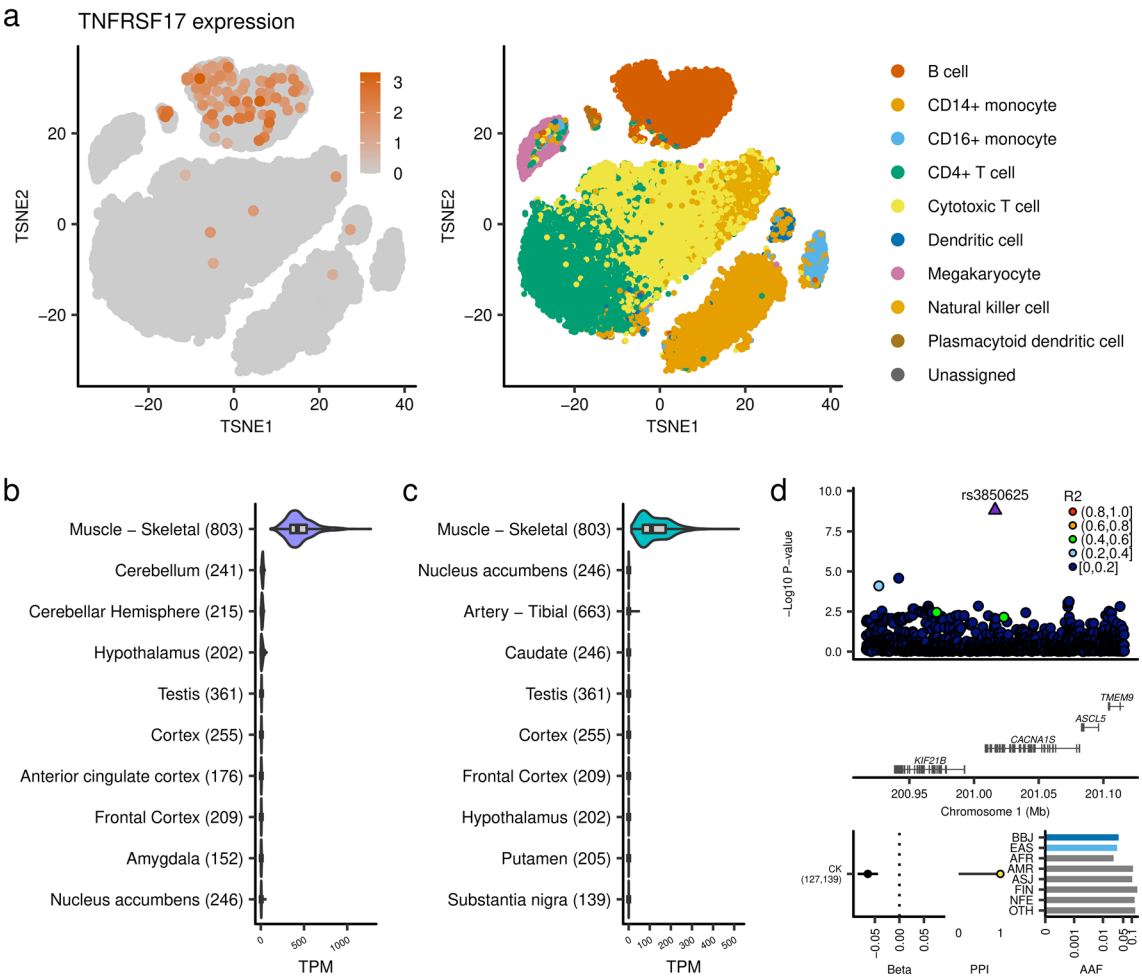
Peer review information *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work.

https://doi.org/10.1038/s41588-024-01913-5



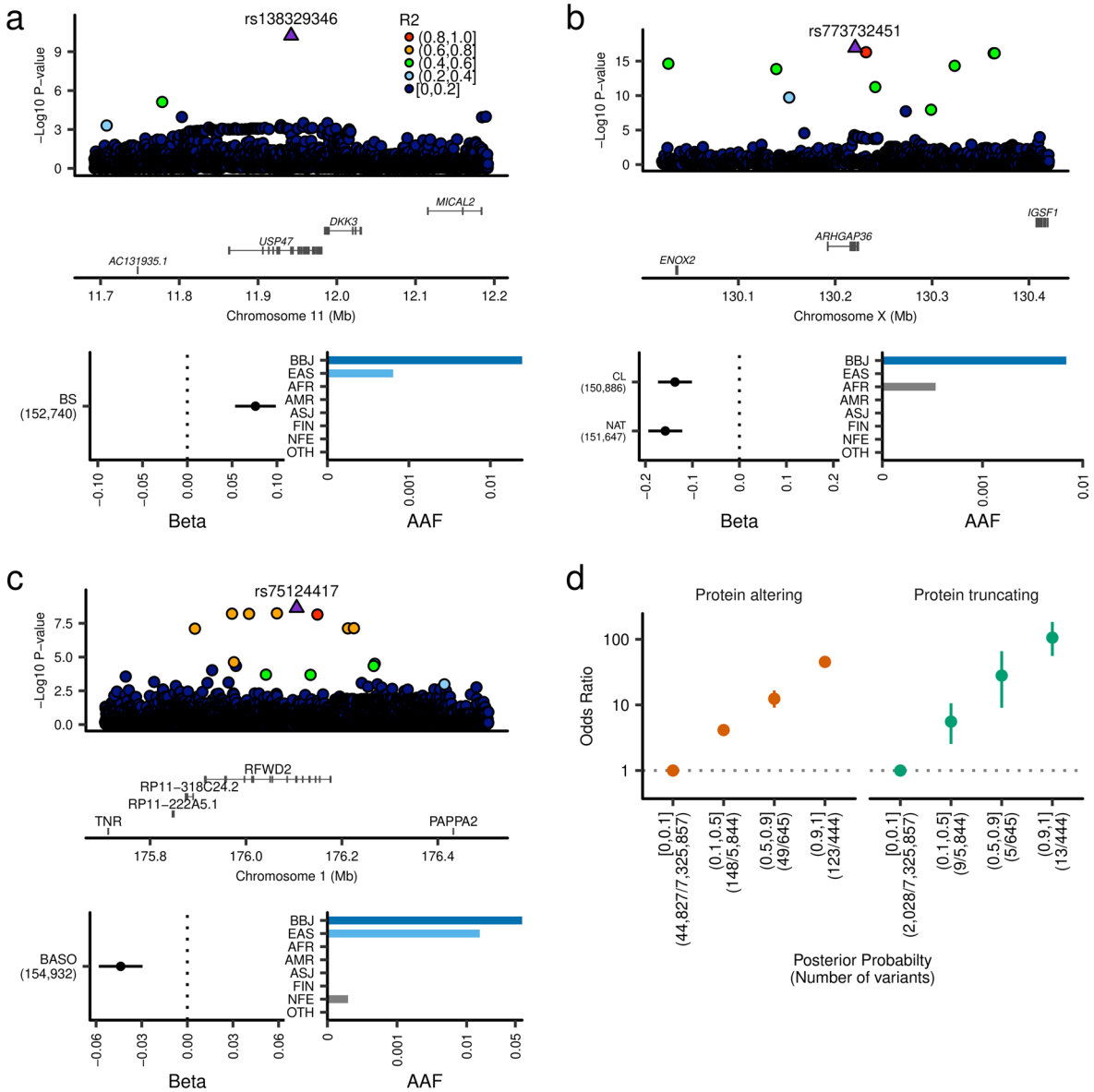
Extended Data Fig. 1 | Schematic of the study design. BBJ, Biobank Japan; NCGG, National Center for Geriatrics and Gerontology; TOMMO, Tohoku Medical Megabank Organization; MHC, Major histocompatibility complex. The numbers of study participants (*n*) are those after quality control.

https://doi.org/10.1038/s41588-024-01913-5



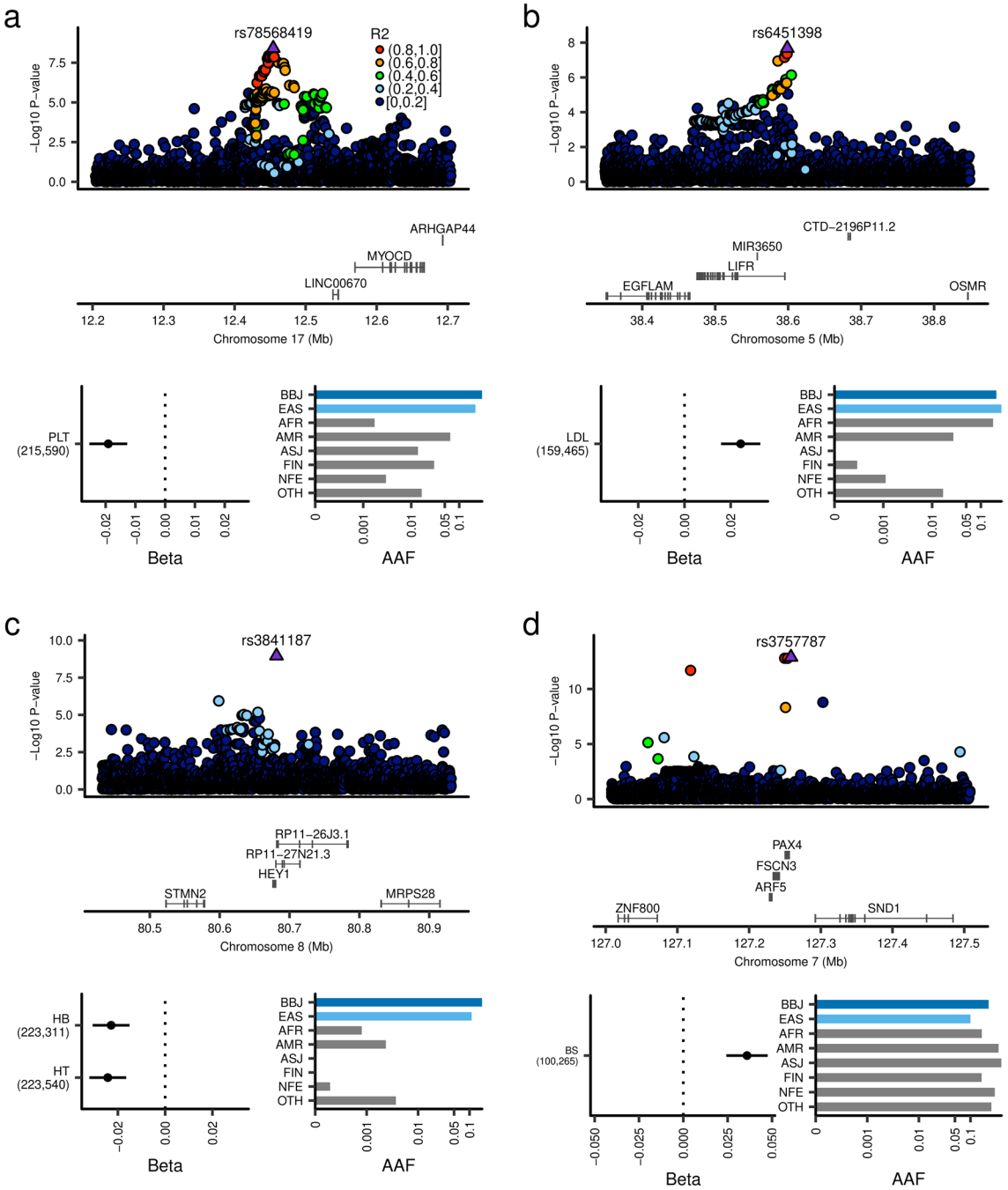
Extended Data Fig. 2 | B cell-specific expression of *TNFRSF17* and muscle-specific expression of *RYR1* and *CACNA1S*. **a**, Single-cell expression status of *TNFRSF17* in 31,021 human peripheral blood mononuclear cells. In the right panel, *TNFRSF17*-expressing cells are highlighted. Color intensity indicates *TNFRSF17* expression level. The left panel shows the cell population. Data were obtained from Single Cell Portal (Single Cell Comparison: PBMC data). **b**,**c**, Muscle-specific expression of *RYR1* (**b**) and *CACNA1S* (**c**). Numbers of samples are shown after the organ name. Violin plots show distribution of gene expression in TPM. Boxplot shows the median value as the centerline; box boundaries show the first and third quartiles and whiskers extending 1.5 times the interquartile range. **d**, Strong association of *CACNA1S* with creatine kinase (CK) levels. Regional plot, beta estimate, PPI, and AAF of rs3850625 are indicated. The numbers of individuals included in the analysis are shown after the trait names. PPI, posterior probability of inclusion; AAF, alternate allele frequency; BBJ, Biobank Japan; EAS, East Asian; AFR, African; AMR, Admixed American; ASJ, Ashkenazi Jewish; FIN, Finnish; NFE, non-Finnish European; OTH, others. AAF was obtained from the gnomAD dataset. The numbers of individuals included in the association analysis are found in Supplementary Table 1.

https://doi.org/10.1038/s41588-024-01913-5



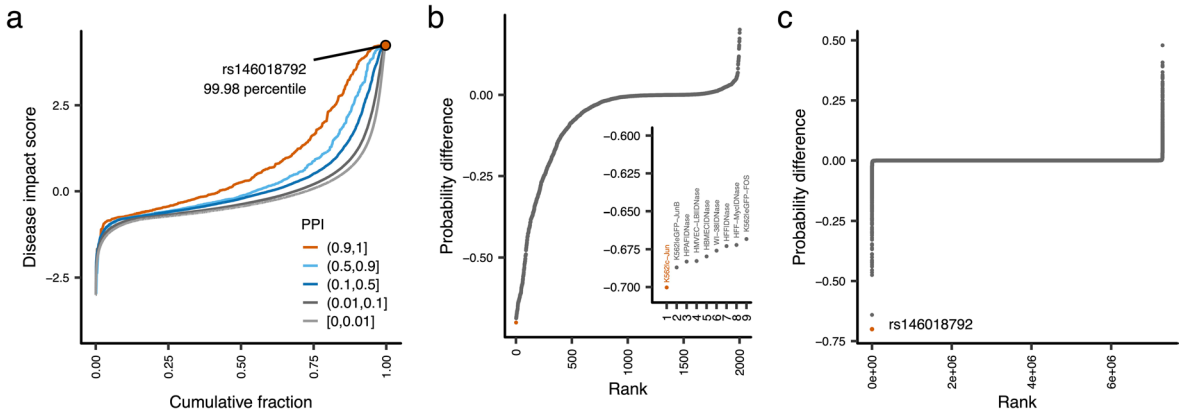
Extended Data Fig. 3 | Rare population-specific coding variants in novel gene-phenotype pairs. a, The EAS-specific rare missense variant in *USP47*, rs138329346, is strongly associated with blood glucose levels. **b**, The Japanese-specific rare missense variant in *ARHGAP36*, rs773732451, is strongly associated with blood sodium and chloride levels. **c**, The EAS-specific rare missense variants in *RFWD2*, rs75124417, is associated with basophil counts. Beta estimates, PPI, and AAF of the associated variants are also indicated in each panel. The error bar for beta estimates indicates 95% confidence interval. The numbers of individuals included in the analysis are shown after the trait names. **d**, Enrichment of coding deleterious variants in variants with high PPI. The numbers of variants included in the analysis are shown after the PPI bins. PPI, posterior probability of inclusion; AAF, alternate allele frequency; BBJ, Biobank Japan; EAS, East Asian; AFR, African; AMR, Admixed American; ASJ, Ashkenazi Jewish; FIN, Finnish; NFE, non-Finnish European; OTH, others. AAF was obtained from the gnomAD dataset. The numbers of individuals included in the association analysis are found in Supplementary Table 1, and abbreviations for phenotypes are found in Supplementary Table 2.

https://doi.org/10.1038/s41588-024-01913-5



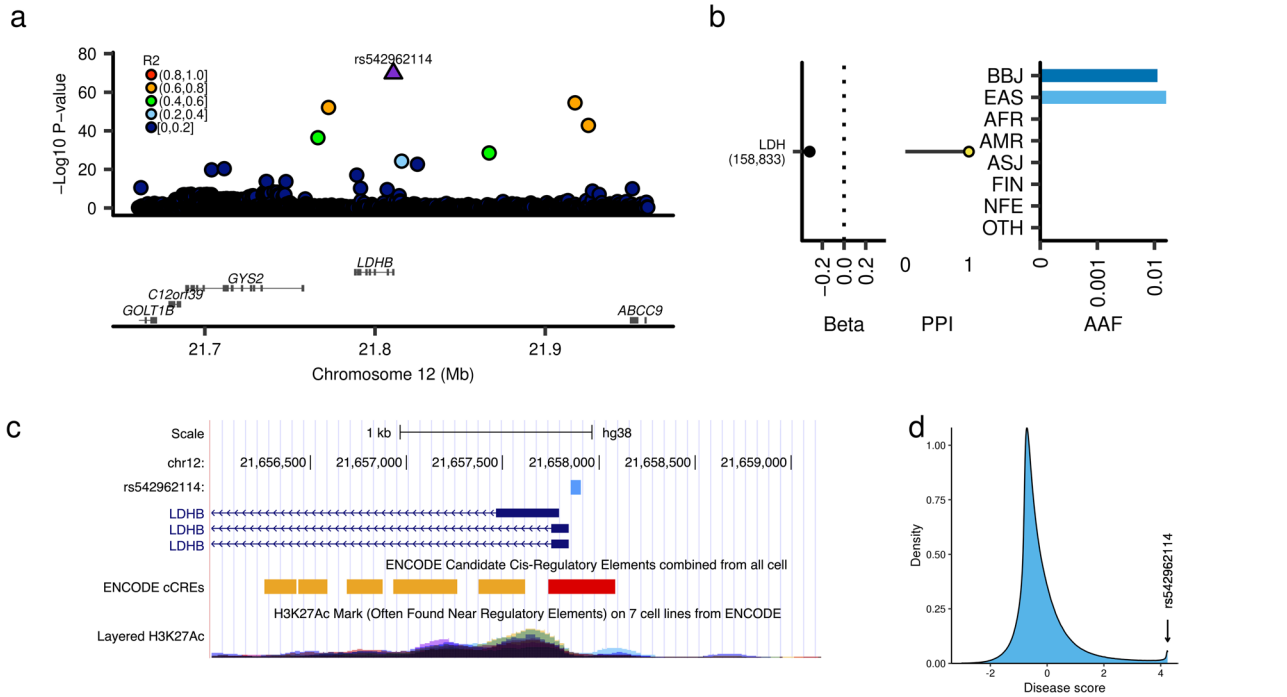
Extended Data Fig. 4 | Non-coding variants much more frequent in East Asians than Europeans in novel gene-phenotype pairs. a, The non-coding variant in the *LINC00670* region, rs78568419, which is much more frequent in EAS than EUR, is associated with platelet counts. **b**, The non-coding variant in the *LIFR* region, rs6451398, quite rare in Europeans, is associated with LDL levels. **c**, The non-coding variant in the *HEY1* region, rs3841187, which is much more frequent in EAS than the other populations (almost absent in Europeans), showed an association with hemoglobin and hematocrit. **d**, The non-coding variant in the *PAX4* region is associated with blood glucose levels. While this variant is similarly frequent between EAS and EUR, this association was not previously reported. Beta estimates, PPI, and AAF of the associated variants are also indicated in each panel. The error bar for beta estimates indicates 95% confidence interval. The numbers of individuals included in the analysis are shown after the trait names. PPI, posterior probability of inclusion; AAF, alternate allele frequency; BBJ, Biobank Japan; EAS, East Asian; AFR, African; AMR, Admixed American; ASJ, Ashkenazi Jewish; FIN, Finnish; NFE, non-Finnish European; OTH, others. AAF was obtained from the gnomAD dataset. The numbers of individuals included in the association analysis are found in Supplementary Table 1, and abbreviations for phenotypes are found in Supplementary Table 2.

https://doi.org/10.1038/s41588-024-01913-5



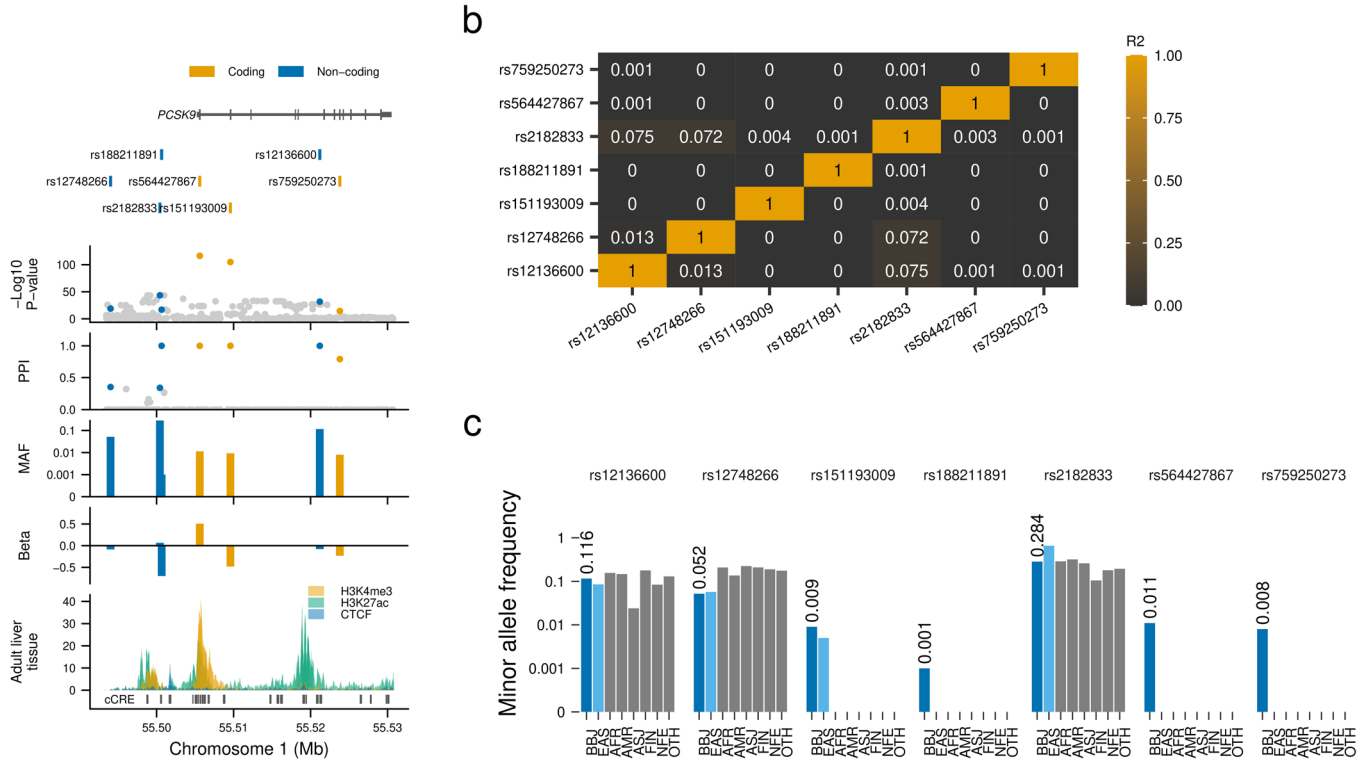
Extended Data Fig. 5 | *In silico* functional assessment of rs146018792, a blood-trait associated non-coding rare variant. Functional prediction of a very rare putative causal variant rs146018792 by DeepSEA. **a**, Distribution of disease impact score of 7,289,211 non-coding variants in the 3,309 fine-mapped loci. **b**, Distribution of probability differences caused by rs146018792 in 2,002 regulatory features implemented in the DeepSEA model. The inset is a zoomed plot of the top 9 features negatively dysregulated by rs146018792. **c**, Distribution of probability differences of K562|c-Jun caused by 7,289,211 variants. PPI, posterior probability of inclusion.

https://doi.org/10.1038/s41588-024-01913-5



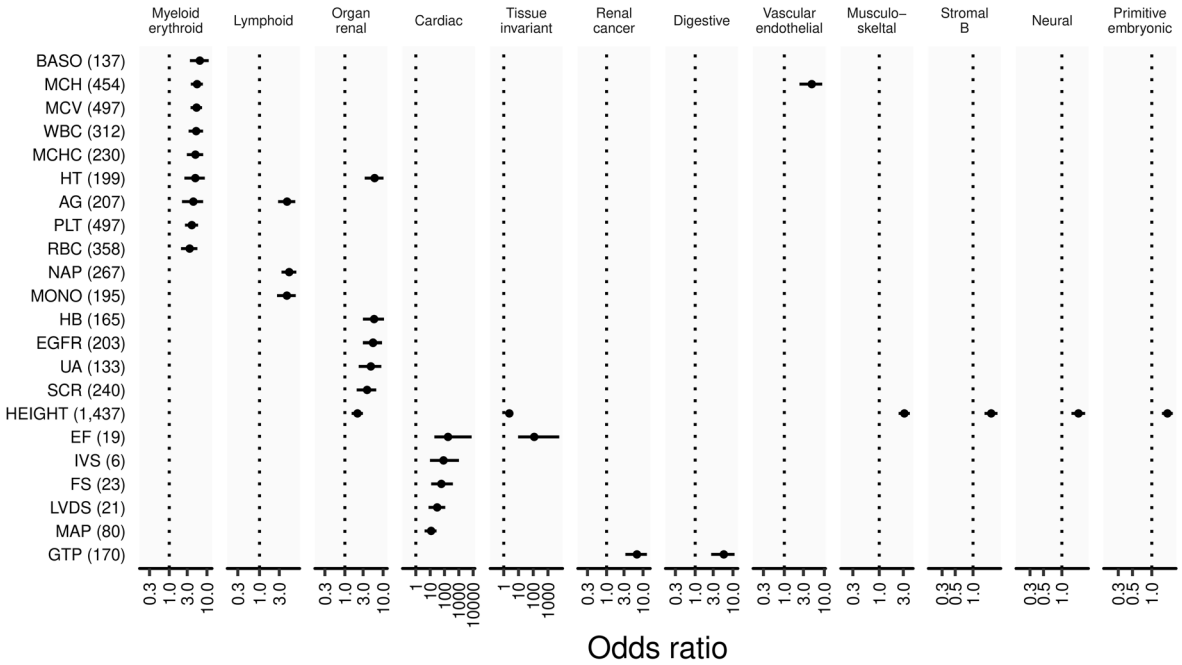
Extended Data Fig. 6 | High impact non-coding variant in the *LDHB* locus. **a**, Regional association plot for the *LDHB* locus. The horizontal axis indicates genomic coordinates, and the vertical axis shows the negative $\log_{10}$ *P*-value. **b**, Beta estimate, PPI, and allele frequency in the global population of rs542962114. **c**, Schematic representation of *LDHB* locus where rs542962114 is located. The horizontal axis shows the genomic coordinate. **d**, Machine learning derived feature for rs542962114 (Methods). PPI, posterior probability of inclusion; AAF, alternate allele frequency; BBJ, Biobank Japan; EAS, East Asian; AFR, African; AMR, Admixed American; ASJ, Ashkenazi Jewish; FIN, Finnish; NFE, non-Finnish European; OTH, others. AAF was obtained from the gnomAD dataset. The numbers of individuals included in the association analysis are found in Supplementary Table 1, and abbreviations for phenotypes are found in Supplementary Table 2.

https://doi.org/10.1038/s41588-024-01913-5



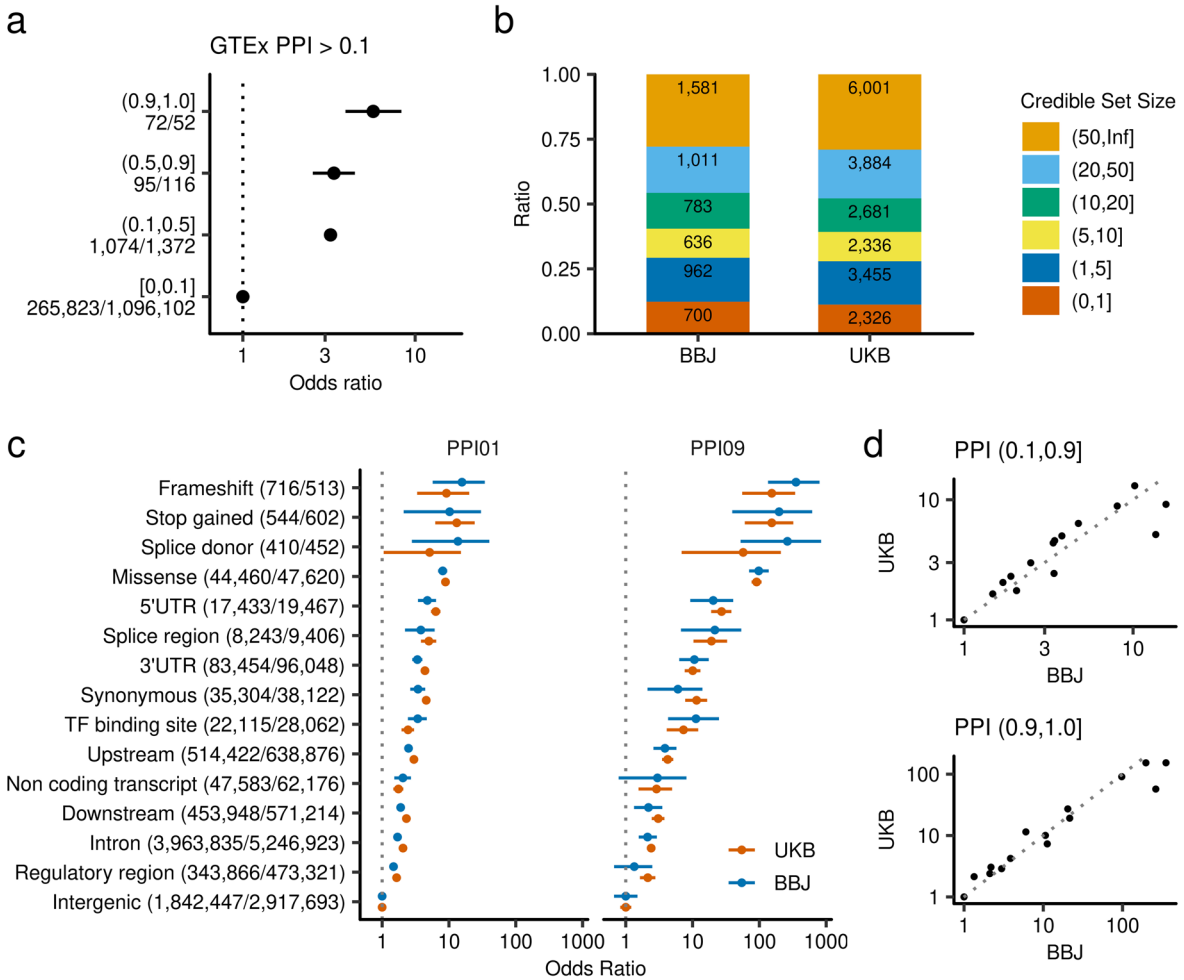
Extended Data Fig. 7 | A novel rare non-coding variant in the *PCSK9* locus confers very strong association with LDLC levels. a, Estimated causal variant configuration at the *PCSK9* locus for serum LDLC. The horizontal axes indicate genomic coordinates. Beta and *P*-value were determined by LDLC GWAS ($n = 111{,}048$). PPIs were determined by FINEMAP (Methods). The very rare non-coding variant rs188211891 showed a very strong association with the LDLC levels with high PPI. **b**, Pairwise linkage disequilibrium matrix of 7 putative causal variants in *PCSK9* locus for LDLC association. Numeric values inside the rectangles indicate r^2. **c**, Population frequencies of seven putative causal variants in this locus. Population frequencies were obtained from the gnomAD database. Chromatin immune-precipitation data were obtained from ENCODE portal. LDLC, low-density lipoprotein cholesterol; MAF, minor allele frequency; PPI, posterior probability of inclusion; AAF, alternate allele frequency; BBJ, Biobank Japan; EAS, East Asian; AFR, African; AMR, Admixed American; ASJ, Ashkenazi Jewish; FIN, Finnish; NFE, non-Finnish European; OTH, others. AAF was obtained from the gnomAD dataset. The numbers of individuals included in the association analysis are found in Supplementary Table 1, and abbreviations for phenotypes are found in Supplementary Table 2.

https://doi.org/10.1038/s41588-024-01913-5



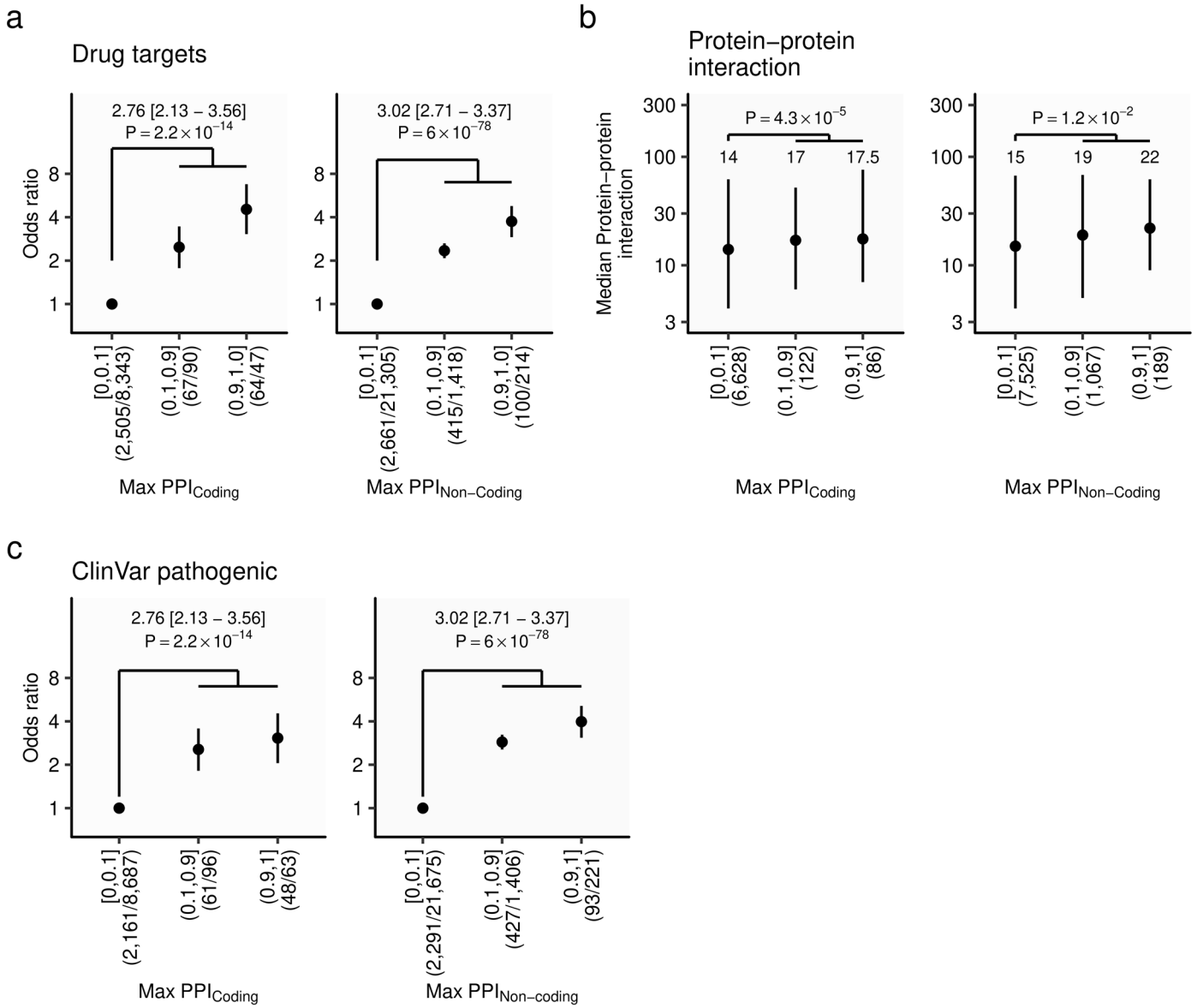
Extended Data Fig. 8 | Tissue-specific enrichment of putative causal variants in regulatory elements. The horizontal axes indicate the odds ratio of high PPI (0.1–1.0] variants within tissue-specific DHS to low PPI [0.0–0.1] variants. We display only DHS-vocabulary and trait pairs which showed significant associations after multiple-testing adjustment. Each point and error bar shows the odds ratios and 95% confidence intervals. The odds ratio and its 95% confidence interval were estimated by Fisher's exact test. The numbers of variants included in the analysis are shown after the trait names. The numbers of individuals included in the association analysis are found in Supplementary Table 1, and abbreviations for phenotypes are found in Supplementary Table 2.

https://doi.org/10.1038/s41588-024-01913-5



Extended Data Fig. 9 | Enrichment of high PPI variants for causal eQTL variants and comparable enrichment of causal variants for functional annotations between UK and Japan. a, Enrichment of causal eQTL variants in GTEx for variants with high PPI in the current study. The fine-mapped eQTL variants are obtained from results using DAP-G as a representative. Each point and error bar shows the enrichment odds ratios and 95% confidence interval. The odds ratio and its 95% confidence interval were estimated by Fisher's exact test. The numbers of variants included in the analysis are shown after the PPI bins. **b**, Comparable distribution of credible set sizes between UKB and BBJ. **c**, Enhanced enrichment of causal variants in functional annotations in high PPI variants and comparable enrichment for functional annotations between Japanese data and UKB data. Each point and error bar shows the enrichment odds ratios and 95% confidence interval. The odds ratio and its 95% confidence interval were estimated by Fisher's exact test. The numbers of variants included in the analysis are shown after the variant annotations (BBJ/UKB). **d**, Correlations of functional enrichment between UK and Japan in both sets of variants with different PPI. PPI, posterior probability of inclusion; BBJ, Biobank Japan; UKB UK Biobank; UTR, untranslated region; TF, transcription factor.

https://doi.org/10.1038/s41588-024-01913-5



Extended Data Fig. 10 | Enrichment of coding and non-coding causal variants in druggable genes. a, Enrichment of drug-target genes for fine-mapped genes with variants with high PPI for coding and non-coding variants. **b**, Enrichment of genes in protein-protein networks for genes with variants with high PPI for coding and non-coding variants. **c**, Enrichment of genes containing pathogenic variants in the ClinVar for fine-mapped genes with variants with high PPI for coding and non-coding variants. Error bars indicate the first and third quartiles. Annotated *P*-value was estimated comparing genes with the highest PPI > 10% to the highest PPI ≤ 10% by two-sided Fisher's exact test (**a**,**c**) and Wilcoxon rank-sum test (**b**). The numbers of variants included in the analysis are shown after the PPI bins. PPI, posterior probability of inclusion.

nature portfolio

Corresponding author(s): Chikashi Terao

Last updated by author(s): May 6, 2024

Reporting Summary

Nature Portfolio 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 Portfolio 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
Data analysis	GATK(3.5.0, 3.6.0, 3.8.0), IMPUTE2 (2.3.2), BEAGLE(4.1), SHAPEIT2(2.837), EAGLE2(2.39), Minimac4(1.0.0), VEP(87), BOLT(2.3.4), PLINK2(2.0), LDSC(1.0.1), LDstore(2.0), FINEMAP(1.4), PAINTOR (3.0), susieR (version 0.12.19), PyMOL (2.5.2). Standalone software to create LD-matrix (LD-store) and conduct statistical finemapping (FINEMAP) are available at http://www.christianbenner.com/. We deposited custom analysis codes for association analysis and finemapping at https://doi.org/10.5281/zenodo.10934238. Further detailed scripts are available upon reasonable request to the corresponding author.

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 Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The GWAS summary statistics and the results of statistical finemapping is available at JENGER website (http://jenger.riken.jp/en) and the National Bioscience Database Center (https://biosciencedbc.jp/en) under research ID hum0014 without any restriction. The imputation reference panel containing the 3,256 high-depth Japanese subjects will be available at the National Bioscience Database Center (https://biosciencedbc.jp/en) under research ID hum0014 and available to the

researchers after approval by the Human Data Review Board.
The protein 3D structure data was obtained from the Protein Data Bank (https://www.rcsb.org/). Human tissue expression data was obtained from Genotype-Tissue Expression (GTEx) Portal (https://www.gtexportal.org/home/). DNase1 hypersensitivity site and transcription factor footprints were obtained from public repositories (https://zenodo.org/records/3838751 and https://zenodo.org/records/3905306, respectively). Chromatin immunoprecipitation data was obtained from ENCODE website (https://www.encodeproject.org/). Allele frequency information for diverse human populations was obtained from gnomAD project website (https://gnomad.broadinstitute.org/). The list of clinically curated pathogenic variants was obtained from ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/).

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	BBJ first cohort n = 176,894, BBJ second cohort n = 12,098, NCGG cohort n = 14,224, ToMMo cohort n = 53,365. Specific sample sizes for each GWAS are described in Supplementary table 1. We excluded traits with total sample size < 4,000 to attain 80% power for 1 S.D. effect size at allele frequency 0.5%.
Data exclusions	For BBJ subjects, we excluded outliers from East Asian clusters in the first and second genetic principal component space in which we projected subjects in combination with 1000 Genomes Project samples. We also excluded samples with call rates less than 0.98, and samples whose reported sex information was not supported by genotypes in the X-chromosome. For ToMMo subjects, we excluded samples with call rate less than 0.97 and non-Japanese identified by principal component analysis (PCA) analyzed with combination of 1000 Genomes Project samples.
Replication	We perform replication analysis for 26 phenotypes which were available from both BBJ and ToMMo. We observed 85.4% (1,304/1,528) of lead signals were replicated in the ToMMo dataset with P < 0.05. Further information will be found in the supplementary information.
Randomization	Randomization is not applicable since genetic variants are inherently randomly distributed in the population and the study focuses on the association between the genetic variants and human phenotypic variation.
Blinding	Blinding is not applicable since the genetic and phenotype data were independently collected in advance.

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

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HEK293T cells and HeLa cells (#CCL-2) were purchased from the American Type Culture Collection (ATCC).
Authentication	HEK293T cells and HeLa cells (#CCL-2) were authenticated by the supplier (ATCC) using STR profiling.
Mycoplasma contamination	No contaminated cell lines were used
Commonly misidentified lines (See ICLAC register)	No misidentified cell lines were used.

Human research participants

Policy information about studies involving human research participants

Population characteristics	The BBJ is composed of ~200k participants who had one of the 47 diseases predefined. This cohort has trend of old age (mean age at enrollment 63.5 years and s.d. 14.0 years) and relatively high fraction of male subjects (54.0%). Detailed study design of BBJ is described in the following literatures. The NCGG is a hospital based cohort with elder participants (mean age at enrollment 70.7 years and s.d. 13.5 years) and relatively low fraction of male subjects (44.3%). ToMMo was established to contribute to the reconstruction of the Tohoku area which suffered from the Great East Japan Earthquake and contains large-scale adult general population cohort (mean age at enrollment 60.4 years and s.d. 11.22 years, 38.4% male). Nagai, A. et al. Overview of the BioBank Japan Project: Study design and profile. J Epidemiol 27, S2–S8 (2017). Hirata, M. et al. Cross-sectional analysis of BioBank Japan clinical data: A large cohort of 200,000 patients with 47 common diseases. J Epidemiol 27, S9–S21 (2017). Yasuda, Jun, et al. Genome analyses for the Tohoku Medical Megabank Project towards establishment of personalized healthcare. The journal of biochemistry 165.2 (2019): 139-158.
Recruitment	In this study, we included three different datasets constructed from the contemporary Japanese population [Biobank Japan (BBJ) 1st cohort, BBJ 2nd cohort, and National Center for Geriatrics and Gerontology (NCGG) cohort] and meta-analyzed the results. BBJ is a nationwide hospital-based biobank with 12 collaborating medical institutions. The first cohort targeted 47 diseases and recruited 200,000 people between 2003 and 2013, and the second cohort targeted 38 diseases and recruited 67,000 people between 2013 and 2018 (https://biobankjp.org/en/index.html). In this study, 12,098 people with available genotypes were included from BBJ 2nd cohort. The NCGG Biobank is a hospital-based biobank maintained by NCGG since 2012. The participants were recruited from NCGG hospital, Obu City, Aichi prefecture, and nearby medical institutes (https://www.ncgg.go.jp/english). The subjects in ToMMo were recruited from the health checkups conducted in two prefectures of Northeastern Japan: Miyagi and Iwate (https://www.megabank.tohoku.ac.jp/english/).
Ethics oversight	All participants provided written informed consent following the protocols approved by following institutional ethical committees, the ethics committees of RIKEN Center for Integrative Medical Sciences, the Institute of Medical Sciences, the University of Tokyo, and National Center for Geriatrics and Gerontology.

Note that full information on the approval of the study protocol must also be provided in the manuscript.