Effect heterogeneity reveals complex pleiotropic effects of rare coding variants

Wenhan Lu[1,2,3], Siwei Chen[1,2], Danielle Posthuma[4,5], Benjamin M. Neale[1,2,3,6], Luke J. O'Connor[1,7], Konrad J. Karczewski[1,2,3,6,*]

[1]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
[2]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA
[3]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
[4]Department of Complex Trait Genetics, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[5]Department of Child and Adolescent Psychiatry and Pediatric Psychology, Section Complex Trait Genetics, Amsterdam Neuroscience, Vrije Universiteit Medical Center, Amsterdam, The Netherlands
[6]Novo Nordisk Foundation Center for Genomic Mechanisms of Disease, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
[7]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02446
[*]Corresponding author

Abstract

Genome-wide association studies (GWAS) and rare-variant association studies (RVAS) have identified thousands of genes and variants that affect multiple phenotypes. Here, using rare variant association results from the UK Biobank (UKB) data, we identify pervasive gene-level pleiotropy across diverse phenotypic domains and highlight genes with apparent allelic series that provide additional association support through disease pathways. We develop a statistical test, ALLSPICE, to explore allelic heterogeneity within pleiotropic genes and identify evidence for a heterogeneous effect of a subset of missense variants in the *ALB* gene on albumin and calcium levels, which are further validated at protein structure level. In this way, we demonstrate potential shared and trait-specific biological pathways, providing insights into the genetic basis of pleiotropy in complex human traits.

Introduction

Pleiotropy, or the association between a variant with multiple phenotypes, is pervasive across the human genome. For uncorrelated phenotypes, pleiotropy can help decipher the genetic mechanisms of complex human diseases by indicating the sharing or independence of molecular pathways. Better interpretations of the pleiotropic effects of genetic risk factors can not only inform the development of therapeutic approaches and diagnosis strategies but also help boost the statistical power of gene discovery by providing evidence to pool samples with phenotypes sharing a genetic basis[1].

Pleiotropy can manifest through a diverse range of mechanisms, including the gene's involvement in multiple biological pathways, its influence on multiple cellular functions, or the expression of different gene isoforms in different cell types or tissues. Additionally, pleiotropic effects may result from a gene's role in regulating other genes or interacting with multiple molecular targets, which may also be cell type-specific, leading to diverse phenotypic outcomes[2]. One way that pleiotropy arises is when a gene affects multiple traits through a single biological pathway, leading to vertical pleiotropy, which is mainly addressed by causal inference approaches such as Mendelian Randomization (MR). However, most pleiotropic relationships are not causal[3], and instead, they reflect horizontal pleiotropy, where the same genetic components independently influence multiple traits through distinct, unrelated pathways. Even causally related traits may also exhibit horizontal pleiotropy[4]. Methods[3,5,6,7] have been developed to decompose the structure of phenotypes with shared genetic risk factors and indirectly account for the underlying pleiotropic relationships, with a focus on common variants.

Well-powered GWAS identify widespread pleiotropic effects among common variants[8,9]. Such studies allow the exploration of the molecular basis of pleiotropy within certain phenotypic groups[10], which implicate etiologies within major domains of diseases and disorders, such as psychiatric disorders[11], type II diabetes[12], coronary artery diseases[13]. Previous exome-wide association studies have presented rare pleiotropic variants associated with specific

phenotypes, such as urine and blood biomarkers[14,15] as well as ASD and ADHD[1], which have implicated shared biological mechanisms, including similar biological perturbations across different cell types or gene regulatory circuits (e.g. in metabolism), and empowered further gene discoveries. However, these studies focused on restricted groups of phenotypes lacking a full picture of rare variant pleiotropy across traits. Recently, the rapid growth of large-scale sequencing data from biobanks has enabled broad gene-based association tests across thousands of phenotypes[16,17], which have been used to decipher the polygenic architecture of rare variants[18] and allow further exploration of pleiotropy across diverse phenotypic domains. Cross-domain associations can arise from heterogeneous effects of genetic variants, such as missense variants altering protein structure and function or pLoF variants leading to the loss of gene activity. Additionally, differences in cellular environments can influence how these variants interact with biological pathways and other molecules, further contributing to the diversity of their impacts. Understanding these effects can provide insight into gene-level pleiotropy and lays the groundwork for a thorough analysis of the underlying mechanisms behind these associations.

Here, we leverage gene-based association statistics to enhance our understanding of pleiotropy among rare variants on a broad range of phenotypes. We highlight gene-level pleiotropy by examining allelic series across rare variants within genes and quantifying the extent of pleiotropy across gene sets that were previously defined to have functional relevance in diseases. Conditioning on the shared associations identified at the aggregated gene level, we examine the effect of rare variants within genes and determine the potential shared or heterogeneous effect on the same pair of phenotypes. We propose a novel statistical method, ALLSPICE, to test for a heterogenous effect of different alleles within each pleiotropic gene. We then apply the statistical tests to the pleiotropic gene-annotation pairs and compare the different directions of effects between functional groups. By summarizing results from Genebass and investigating the underlying mechanisms of pleiotropic genes with statistical inferences, we target genes with functional impact on diverse phenotypes and evaluate the magnitudes and

directions of the effects of the variants underlying the target gene, which helps differentiate shared associations due to common biological pathways from those demonstrating heterogeneous variant effects.

Results

Pervasive gene-level pleiotropy across human exome

To characterize the extent of pleiotropy across human exome, we considered 599 biological measurements and health consequences that have at least one significant association ($p < 2.5 \times 10^{-6}$) among burden tests of rare coding variants (MAF < 1%; Methods) and considered a gene as pleiotropic if it had a significant association with two or more of these phenotypes. From these, we curated a maximal subset of 239 independent phenotypes ($r^2 < 0.1$; Methods) and identified 313 pleiotropic gene-annotation pairs (with burden $p < 2.5 \times 10^{-6}$). Among those pairs with at least one association, the observed fractions of genes demonstrating pleiotropic effect are 28%, 23%, and 11%, respectively through pLoF, missense, and synonymous variants (Figure 1A), which are overdispersed compared to the expected Poisson distribution and the pvalue-permuted null data across all four functional groups, suggesting potential non-random variation in the underlying biology to be further investigated (Table S1 and Figure S1). The presence of significant synonymous associations, particularly pleiotropic ones, represents a baseline that is likely due to tagging of GWAS signals with intermediate frequency; however, the enrichment of pLoF and missense variants – which are themselves rarer – suggests an effect of functional impact.

The functional importance of genes often reflects their central role in key biological processes and their influence across multiple traits. Here, we explore pleiotropy across 37 gene categories (e.g. essential or haploinsufficient genes; https://github.com/macarthur-lab/gene_lists), where genes were previously identified to have a specific etiological or functional relevance within each category. For pLoF and missense

variants, pleiotropic genes are more enriched within disease-related groups, such as ACMG genes[19] and the BROCA cancer risk panel. Genes with vital biological functions are already depleted for any associations, resulting in their likewise depletion for pleiotropic associations with a similar extent as the background gene list of all protein-coding genes. Autosomal dominant and ClinGen haploinsufficient genes[20] tend to be more pleiotropic among pLoF associations, indicating that pLoF variants in these genes can have broad impacts across multiple systems, contributing to a wider range of diseases and disorders, when a single functional copy of the gene is not sufficient to maintain normal function (Figure 1C, Figure S3 & SuppTable1). Pleiotropic genes also tend to have more protein-protein interactions than trait-specific or non-association genes, indicating that variations of these types are likely to impact phenotypes by interrupting the interactions of protein complexes (Figure S4).

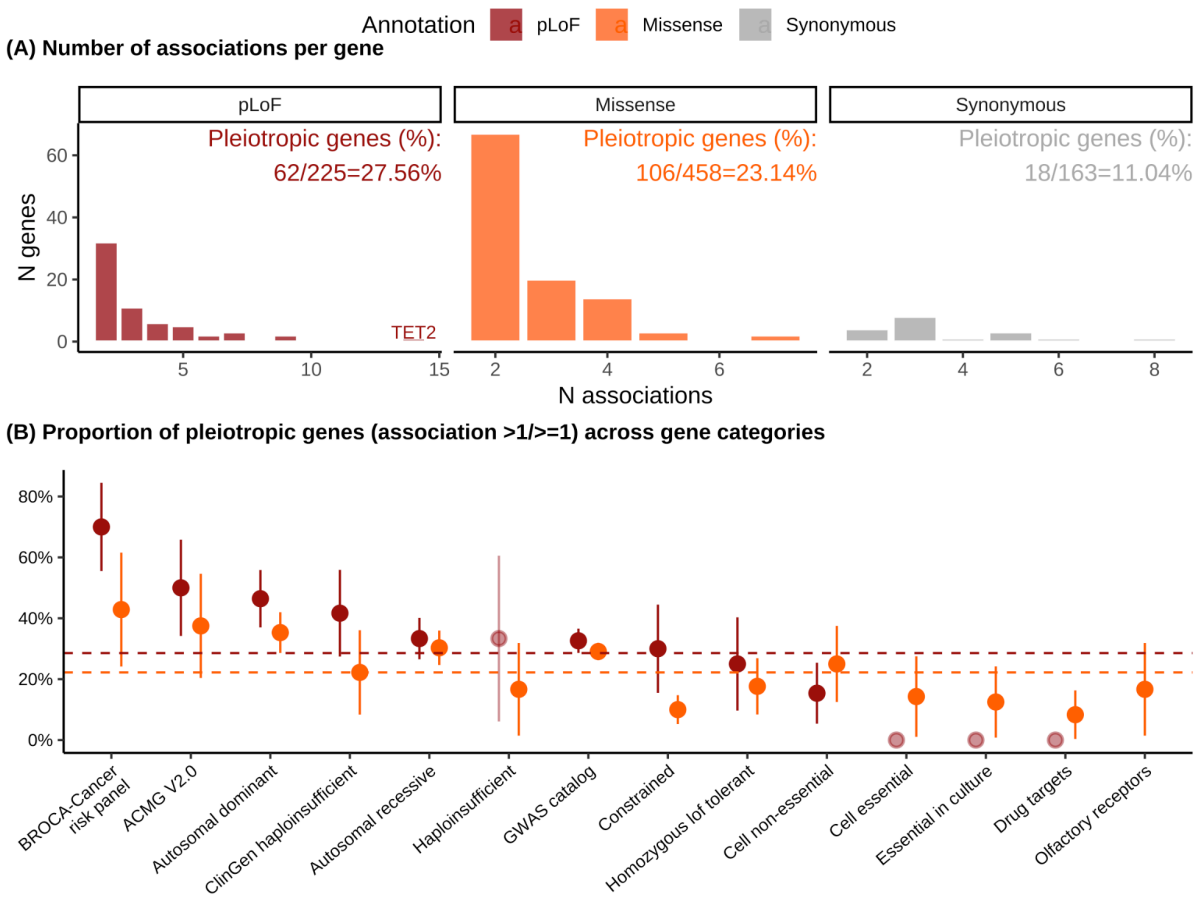


Figure 1 | Pervasive gene-level pleiotropy among rare variants across 239 high-quality independent phenotypes. **A**, distributions of genes with the corresponding number of associations (x-axis) across three variant functional groups (panels and colors). Labels on top of each panel indicate the proportion of genes with more than one association (as shown by the bar plots) among genes with at least one association for the corresponding functional annotation group; **B,** Pervasive gene-level pleiotropy across 14 selected gene categories. The proportion of pleiotropic genes is computed as the number of genes with two or more associations divided by the number of genes with at least one association for each gene category per annotation group (pLoF: red; missense: orange) across the 239 high-quality independent phenotypes. The transparent dots are those groups with less than 5 genes tested in total. The dashed line is the proportion of pleiotropic genes among all genes. Error bars indicate standard errors. Colors are consistent in **A-B**.

Pleiotropy manifests in diverse forms: allelic series and domain-level associations

One form of pleiotropy is an allelic series, where different mutations within the same gene produce a spectrum of related phenotypic outcomes. Additionally, a single gene might be

associated with diverse phenotypic domains, which also underscore the gene's multifaceted role in biological processes. We investigated allelic series present in gene-based test results by identifying those gene-phenotype pairs that are significantly associated with either pLoF and missense variants, but not both, implicating potential connections between the biological pathways of different functional annotation groups.

Across the 599 selected phenotypes, we identified 151 gene-annotation pairs exhibiting pleiotropy (i.e. two or more associations) for non-synonymous variants (i.e., pLoF or missense), among which 31 genes are associated with different sets of phenotypes for their pLoF and missense groups. For instance, the pLoF variants in well-documented tumor suppressor gene *ATM* are associated with 7 cancer diagnoses, such as breast and pancreas cancer, while its missense variants are associated with 3 blood biomarkers. In addition, *PKD1*, *SLC34A3*, and *TMPRSS6* also have pLoF variants associated with diseases, but missense variants associated with blood biomarkers. On the other hand, *FAM234A*, *JAK2*, and *MICA* have missense variants associated with diseases, but pLoF variants associated with blood biomarkers (Table 1 & SuppTable2). These differential associations may arise when different variant types have different functional effects on the given gene. However, there remains a challenge in distinguishing bona fide functional effects from power differences, especially when comparing phenotypic power (i.e. quantitative vs binary traits), and allele frequency (i.e. pLoF variants have lower frequencies than missense variants).

Across six phenotypic domains formed from the 599 selected phenotypes (Extended Data Figure 1A), we identify 114 gene-annotation pairs presenting multi-domain associations (Extended Data Figure 1C). Among pairs with at least one association ($p < 2.5 \times 10^{-6}$), more deleterious functional groups, such as pLoF, show higher proportions of genes with multi-domain associations (12.2%) and fewer with trait-specific associations (50.14%), while in contrast, synonymous variants present opposite trends for the two association levels (3.41% for

multi-domain and 69.66% for trait-specific) (Table S3 and Extended Data Figure 1B). The same relationship of association level with functional deleteriousness is also observed among the 17 disease-level domains (Table S4 and Figure S5). To evaluate the relationship between constraint level and domain-level association status, we performed logistic regression of gene intolerance to loss of function mutations (Upper Bound of the Observed/Expected Ratio for Loss of Function Variants (LOEUF)) (Karczewski et al. 2020) and protein coding sequence (CDS) length on whether a gene is more likely to have multi-domain, domain-specific, and trait-specific associations compared to having no association. For pLoF variants, genes with higher LOEUF value (more tolerant to loss of function mutations) are less likely to have multi-domain ($p_{oe_lof_upper} = 5.67 \times 10^{-3}$, OR = 0.34), domain-specific ($p_{oe_lof_upper} = 1.81 \times 10^{-2}$, OR = 0.60) or trait-specific ($p_{oe_lof_upper} = 7.13 \times 10^{-3}$, OR = 0.60) associations than no association. CDS length does not show strong evidence of contributing to the likelihood of having any type of association pattern (Table S6).

As an example of domain-specific gene, the missense group of *OCA2*, a gene that provides instruction for producing the P protein that affects skin color, is associated with both skin change due to chronic exposure to nonionizing radiation and basal cell carcinoma or skin cancer. Some disease domains are frequently associated with the same genes. For example, five genes are associated with both ankylosing spondylitis (AS) and iridocyclitis, including the missense variants of *NELFE*, *PSMB9*, and *SLC44A4*, and the synonymous variants in *AL645922*, which are genes that play significant roles in immune function and inflammatory processes relevant to various diseases. *PSMB9* has the most direct potential interaction through its role in antigen processing and immunoproteasome, which can influence the presentation of peptides by *HLA-B27*, a well-known genetic risk factor for AS[24]. Additionally, *NELFE* regulates transcription elongation, *SLC44A4* impacts cell membrane integrity and signaling and thus influencing the function of immune cells such as T cells and macrophages, and *AL645922*, a long non-coding RNA, regulates the expression of immune-related genes by modifying

chromatin structure and transcriptional activity, which all potentially affect genes involved in immune responses and inflammation. Together, these genes contribute to the broader landscape of immune regulation and inflammatory responses, shedding light on the mechanisms underlying autoimmune diseases.

Table 1 | Genes with pLoF variants associated with disease(s) and missense variants associated with biomarkers and no diseases or vice versa. (Diseases are in bold)

Gene	pLoF-only associations	Missense-only associations	Both pLoF and missense associations
ATM	•**Cancer diagnosed by doctor** •**C25 Malignant neoplasm of pancreas** •**C77 Secondary and unspecified malignant neoplasm of lymph nodes** •**C78 Secondary malignant neoplasm of respiratory and digestive organs** •**C50 Malignant neoplasm of breast** •**D70 agranulocytosis** •**C79 Secondary malignant neoplasm of other sites**	•Mean reticulocyte volume •Mean corpuscular volume •Mean corpuscular haemoglobin	
PKD1	•**Hypertension** •Urate •Microalbumin in urine •Haemoglobin concentration •Urea •Cystatin C	•Alkaline phosphatase •Monocyte count	•Creatinine
SLC34A3	•**N20 calculus of kidney and ureter**	•Urea	•Cystatin C •Creatinine •Phosphate
TMPRSS6	•**D50 iron deficiency anaemia** •Mean corpuscular haemoglobin concentration •Mean sphered cell volume	•Glycated haemoglobin (HbA1c)	•Haemoglobin concentration •Red blood cell (erythrocyte) distribution width •Haematocrit percentage •Mean corpuscular volume •Mean corpuscular haemoglobin
FAM234A	•Glucose	•**Diabetes diagnosed by doctor**	
JAK2	•Eosinophill count •Eosinophill percentage	•**D75 other diseases of blood and blood-forming organs** •Mean platelet (thrombocyte) volume •Red blood cell (erythrocyte) distribution width	•IGF-1 •Platelet count •Platelet crit
MICA	•Lymphocyte count	•**Hypothyroidism/myxoedema** •**E03 other hypothyroidism**	

Statistical test to model heterogeneous pleiotropic effects

The gene-level results above are influenced by detection ability due to power differences. In order to better understand heterogeneity of effects within genes, we developed a

new statistical method, ALLSPICE to distinguish functional differences among variants. In particular, we consider functionally-influenced directions of effects: pLoF variants cause predictable phenotypic effects by triggering nonsense-mediated decay, disrupting gene function and thus biological pathways, typically leading to uniform impacts on phenotypes, while missense variants result in varied effects on protein function, leading to a broader spectrum of impacts.

Motivated by the observation that many pleiotropic genes have distinct associations for their pLoF vs. their missense variants, we developed a test, ALLSPICE (ALLelic Spectrum of Pleiotropy Informed Correlated Effects; Methods). For each triplet consisting of a gene and two phenotypes that are associated with the gene through burden tests, ALLSPICE identifies variants with heterogeneous effects through a likelihood ratio test, incorporating phenotypic correlation between the two traits, allele frequencies of rare variants within the gene and their GWAS effect sizes on the two phenotypes. The null hypothesis assumes a perfect correlation between the rare variant effect sizes on the two phenotypes indicated by a hypothesized slope c, which is formulated through maximum likelihood estimation (MLE). By finding the best possible c that fits the data in a perfect linear model, ALLSPICE constructs a likelihood ratio and returns the test statistics. A non-null p-value indicates either that variants within the gene are differentially associated with the two phenotypes or, more generally, that their pleiotropic effects are imperfectly correlated, in which case, no c from real number space can possibly fit the data in a perfect linear model, while when p-value is non-significant, c suggests a relationship between the functional effects of the variants on the two phenotypes, conditioned on their phenotypic correlation. While it does not test this explicitly, ALLSPICE distinguishes between vertical pleiotropy and some cases of horizontal pleiotropy: vertical pleiotropy should produce a shared allelic series, whereas horizontal pleiotropy may result in heterogeneity, when different missense variants affect different functions and thus different traits[25]. In simulations, the test

produces uniformly distributed p values under the null hypothesis, and it has sufficient statistical power under the alternative hypothesis (Methods: simulation results).

From the 359 quantitative traits out of the 599 curated phenotypes, we applied the ALLSPICE to 11,810 pleiotropic gene-trait triplets (4,708 pLoF, 5,911 missense, and 1,191 synonymous), each of which consists of a gene (with one functional annotation of the variants specified) and two quantitative phenotypes that are associated with this gene at exome-wide significance, where at least two variants are included in that gene (SuppTable 3). We prefiltered rare variants to those with minor allele frequencies (AF) below 0.01% to control for the effect of linkage disequilibrium. 1,073 of the triplets (292 pLoF, 681 missense, and 100 synonymous) demonstrate significant test results at a nominal level ($p < 0.05$). 124 of the triplets (53 pLoF, 70 missense, and 1 synonymous) demonstrate significant results with multiple testing correction ($p < 4.23 \times 10^{-6}$; 0.05/11,810 tests; Figure S13A). These triplets with significant p-values indicate that, given the gene-level association with both phenotypes, the variant-level effects are not perfectly correlated, exhibiting potentially different underlying genetic mechanisms. Overall, a higher proportion of heterogeneous missense variants (1.18%) is observed across the entire spectrum of phenotypic correlation than the other two functional annotation groups (pLoF 1.13%, and synonymous 0.08%), reflecting the intrinsically diverse effects of missense variants. We also applied ALLSPICE on the same set of triplets across ultra-rare variants with Allele Counts (AC) < 5, which presented similar results compared to those from rare variants (AF < 0.0001) with a slightly refined set of output signals with slightly lower statistical power (Figure S16; SuppTable 4)

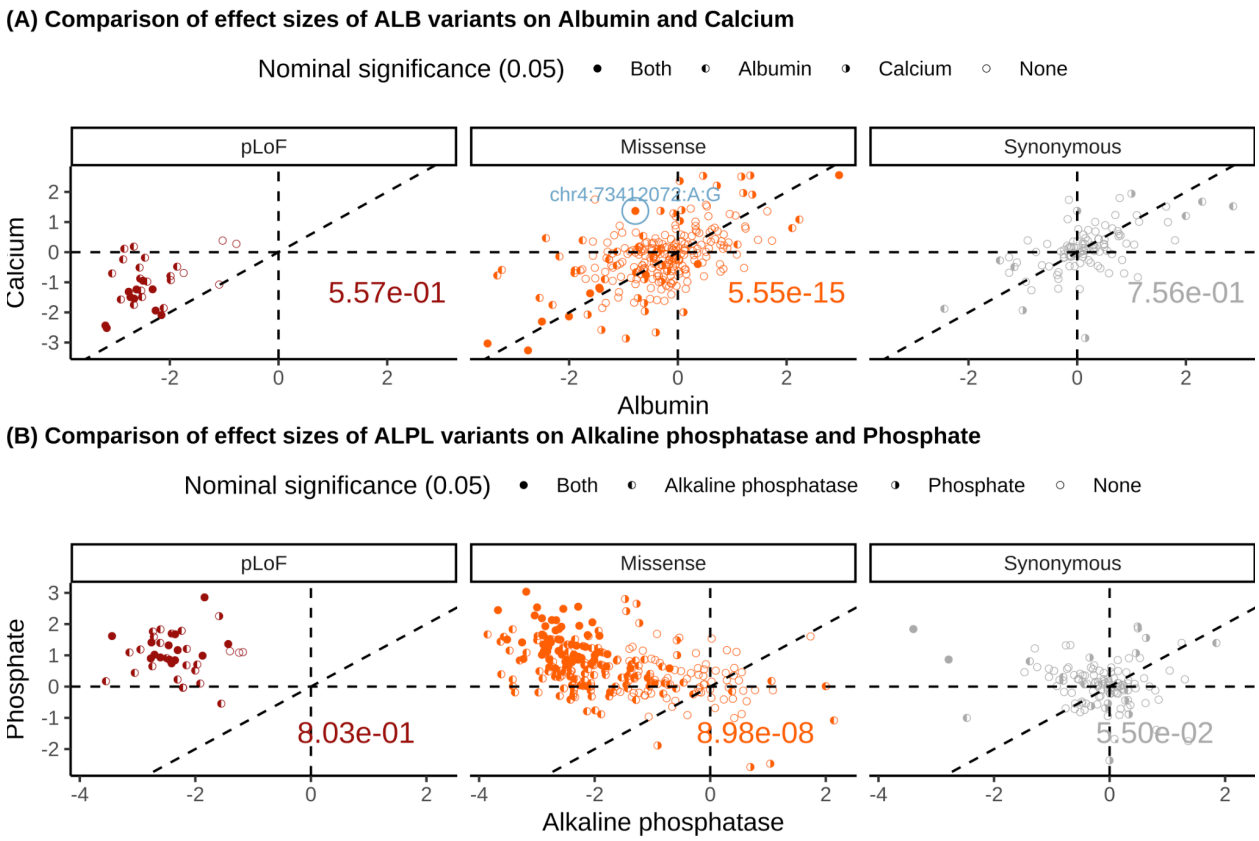


Figure 2 | Examples of significant ALLSPICE triplets. **A,** Comparison of variant effect sizes on Albumin to Calcium within gene *ALB* across three functional annotation groups (colors and columns). The size of the points indicates the negative log of the mean of the pvalues of the two associations of each variant. The values on the x- and y-axis are the effect sizes of variants on albumin and calcium levels respectively. The numbers on the left top of each panel represent the pvalue from ALLSPICE test results of the corresponding annotation group. **B,** Comparison of *ALPL* variant effect sizes on alkaline phosphatase level to phosphate level across three functional annotation groups (colors and columns). The values on the x- and y-axis are the effect sizes of variants on alkaline phosphatase and phosphate levels respectively. The nominal significance ($p < 0.05$) of GWAS association tests for each variant is indicated by the shape of the points. The numbers on the bottom right of each panel represent the pvalue from ALLSPICE test results of the corresponding annotation group.

An example is the *ALB* gene, which is the protein-coding gene for albumin and is associated with calcium and albumin levels in blood samples through both its pLoF and missense variants according to Genebass. It is well-known that biologically circulating calcium is albumin-bound, suggesting that calcium level is mediated by the level of albumin in human

blood, which is supported by a phenotypic correlation of r^2=0.51 in the Genebass data. ALLSPICE test results on triplet *ALB*-albumin-calcium indicate significantly more heterogeneous effects of rare missense variants ($p = 5.5\times10^{-15}$) than the pLoF variants ($p = 0.56$) that are predicted to have consistent directional effects. Particularly, a missense mutation at locus chr4:73412072:A:G has a positive impact on the calcium level but a negative impact on the albumin level (Figure 2A), suggesting that the effects of missense variants on calcium are not mediated by albumin level itself. To determine the *ALB* missense variants with major contributions to the statistical test result, we performed a jackknife cross-validation rerunning ALLSPICE repeatedly each time leaving one variant out and comparing the changes in c_{MLE} and P-value. In the *ALB*-albumin-calcium example, this analysis indicates that no individual variant has an impact significant enough to reverse the test result, where the removals of 159 variants result in an increase in test significance, which are mostly variants with null effect on the two phenotypes (transparent points in Figure S14A), while removing any of the remaining 235 variants attenuates the test significance, indicating stronger heterogeneous effects. The missense variant chr4:73412072:A:G is leading to the most substantial change in c_{MLE} (Figure S14B; SuppTable 5) and the 9th largest change in P-value (Figure S14A; SuppTable 5).

We further examined whether *ALB* missense variants associated with calcium- versus albumin- levels ($p < 0.05$) are distributed differently on the *ALB* protein. We found that missense variants associated with calcium level are located significantly closer to the calcium-binding sites on the 3D protein structure of *ALB*, compared to those exclusively associated with albumin level ($p = 2\times10^{-10}$; Figure 3A and 3B). This pattern suggests that *ALB* missense variants are likely to affect calcium levels by directly disturbing the ability of *ALB* to bind calcium. Missense variants affecting albumin- and calcium-level in an opposite direction also tend to be located closer to the calcium-binding sites at 3D protein level ($p = 6.2\times10^{-4}$; Figure 3C and 3D), which indicates sites around calcium-binding sites are likely to have different functional impacts on the downstream

biological processing mechanisms. The *ALB*-albumin-calcium triplet also has significant ALLSPICE result for ultra-rare missense variants with AC < 5 ($p < 3.5 \times 10^{-7}$; Figure S16; SuppTable4), for which the 3D protein distribution pattern still persists (Figure S17).

Another top gene associated with calcium level is *ALPL*, the protein-coding gene that encodes a member of the alkaline phosphatase family of proteins, which is associated with calcium and C-reactive protein through missense variants (Figure S15A), and both alkaline phosphatase and phosphate through both pLoF and missense variants. Biologically, alkaline phosphatase removes phosphate leading to a negative relationship between the levels of the two biomarkers. All pLoF variants tend to undermine the function of *ALPL*, thus reducing the level of alkaline phosphatase and increasing the level of phosphate, which is reflected by the non-significant test results ($p = 0.8$), along with their negative impacts on alkaline phosphatase and positive impacts on phosphate. However, missense variants in *ALPL* demonstrate a more heterogeneous effect spectrum, as suggested by the significant test result ($p = 8.98 \times 10^{-8}$, Figure 2B). Although most of the variants demonstrate opposite effect sizes, some reduce the alkaline phosphatase level while some increase the level. Moreover, some variants even alter alkaline phosphatase and phosphate at the same time, indicating other potential factors that influence this process. However, no evidence of protein level functional clustering is found in the 3D protein structure of *ALPL* (Figure S15B and Figure S15C).

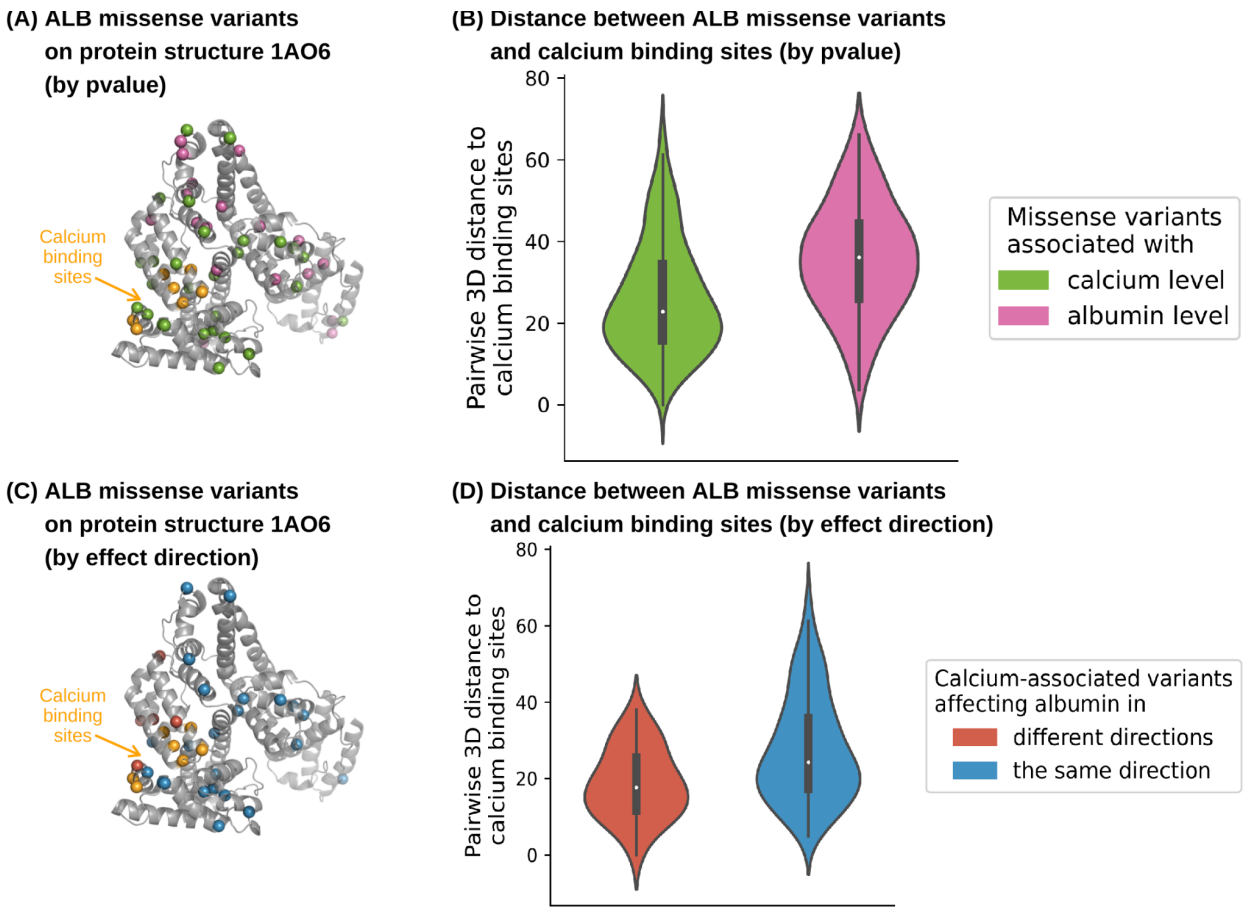


Figure 3 | Rare *ALB* variants (AF < 0.0001) on *ALB* protein structure **A,** Distribution of *ALB* missense variants associated with calcium level but not albumin level (pink), associated with albumin but not calcium level (green), and calcium binding sites (orange) on the 3D structure of *ALB* protein (PDB ID: 1AO6). **B,** Comparison of pairwise distances between missense variants only associated with calcium- (pink) and albumin- level (green) to calcium binding sites on *ALB* protein structure in the 3D space. The distance between two residues was calculated as the Euclidean distance between the alpha carbons of the residues based on their atomic coordinates on the protein structure. *P* value was computed using a two-tailed Mann-Whitney U test. **C,** Distribution of *ALB* missense variants associated with either albumin or calcium level that affect the two biomarker levels in the same (blue) and different (red) direction(s), and calcium binding sites (orange) on the 3D structure of *ALB* protein (PDB ID: 1AO6). **D,** Comparison of pairwise distances between missense variants to calcium binding sites on *ALB* protein structure in the 3D space. The distance between two residues and P value was calculated the same as described in panel **B**. Colors are defined the same as in panel **C.**

Discussion

Here, we conducted systematic analyses of rare variant associations across 599 high-quality phenotypes from Genebass to quantify gene-level pleiotropy. Our survey reveals widespread pleiotropy in human exomes, identifying highly pleiotropic genes with allelic series and heterogeneous effects across individual rare coding variants. To further explore these effects, we developed ALLSPICE, a statistical test that identifies directional heterogeneity of rare variants within pleiotropic genes and assesses their impact on protein functions, adding to the interpretation of potential genetic impacts across phenotypes.

Our analysis has limitations that need to be addressed in future studies. First, as phenotypic correlations induce apparent pleiotropy through shared associations, and pleiotropy is often interpreted in relation to cross-phenotype associations, it is challenging to benchmark a uniform metric for pleiotropy and compare pleiotropic levels across studies. Additionally, given the inherent complexities of rare variants and RVAS design, larger-scale RVAS is needed to enhance the detection of better-powered rare variant associations. The power for burden tests differs between pLoF and missense variants, with pLoF variants having stronger effects but lower allele frequencies, reducing power. Missense variants, though more frequent, have more heterogeneous functional impacts, complicating interpretation. These differences may explain why associations may be detected with one variant type but not the other. To address this, increasing sample size, employing more sophisticated testing methods that account for variant heterogeneity, and improving functional annotations would significantly enhance the ability of burden tests to uncover these complex genetic relationships. Moreover, the choice of maximum minor allele frequency (MAF) is also crucial not only for increasing the power of rare variant associations but also for accurately capturing true pleiotropic effects, as it raises key considerations about the appropriate rarity threshold, influenced by the genetic architecture of phenotypes.

During our exploration of pleiotropy across the frequency spectrum, we observed pleiotropic effects in synonymous variants, which are typically not expected to exhibit such associations. This observation suggests that common variant association signals may be impacting burden tests, particularly at the 1% MAF threshold in Genebass. Further investigation is needed to identify potential confounders and determine the causal effects of shared associations between phenotypes. Our current methods aim to provide a comprehensive view of pleiotropy among rare variants across human exomes, but they do not fully identify causal pathways. More methods will be needed to distinguish between vertical and horizontal pleiotropy and refine their biological interpretations. ALLSPICE aims at highlighting variants with heterogeneous effects on phenotypes presenting horizontal pleiotropy, meaning that variant effects on the two phenotypes should represent divergent biological pathways. Additionally, our ALLSPICE test is currently limited to continuous traits, and future work will focus on detecting horizontal pleiotropy involving binary traits and combinations of binary and quantitative traits.

Pleiotropy helps in understanding the genetic mechanisms of complex human diseases by revealing shared associations across multiple phenotypes. This project systematically investigates the pleiotropic effects of rare coding variants on complex human traits through rigorous statistical inference. By analyzing genes affecting seemingly unrelated traits through rare variants, the study enhances understanding of disease mechanisms and could lead to more targeted therapies.

Methods

Data description

We use gene-based burden test results from the Genebass resource[17], which encompasses 8,074,878 single-variant tests and 75,767 gene-based burden tests across 4,529 phenotypes from the whole exome sequencing data of 394,841 samples in the UK Biobank[26].

Each of the 19,407 unique protein-coding genes is annotated with up to four functional groups: predicted LoF (pLoF), missense (including low-confidence pLoF variants, and in-frame indels), synonymous, and non-synonymous (combination of pLoF and missense variants). Variants grouped in the gene-based burden tests have been pre-filtered to those with minor allele frequency (MAF) less than 0.01. We then filter to the 42,052 recommended high-quality gene-annotation pairs (burden tests) from three main annotation groups (pLoF, missense, and synonymous) with number of variants ≥ 2, mean coverage ≥ 20, expected allele count (AC) ≥ 50, and lambda GC of the synonymous group of the corresponding gene > 0.75, where expected AC is a gene-trait metric defined as combined allele frequency (CAF) times the number of cases for the corresponding trait. CAF of a gene-annotation group is defined here as the sum of MAF of variants with the same functional annotation group within the gene.

Phenotype curation

Focusing more on biological measurements and health consequences, we exclude 1,704 phenotypes related to hospital operations, 435 treatment/medications,177 customized phenotypes, and 64 non-endpoint phenotypes, including injuries, age-related factors, and others. From the 2,149 phenotypes remaining from the previous step of curation, we remove 1,550 phenotypes with no significant associations at the exome-wide significance 2.5×10^{-6} (0.05/20,000). We end up with 599 well-defined phenotypes associated with at least one gene, segregating into six distinct phenotypic domains, including 61 biomarkers, 234 diseases, 216 brain imaging measures, 61 physical measures, 22 mental and cognitive traits, and 5 lifestyle and diet traits. For the purpose of understanding pleiotropy between uncorrelated phenotypes, we also obtain a subset of 239 non-redundant phenotypes with phenotypic correlation below 0.1.

Common variant associations can have pleiotropic effects either on closely related phenotypes within a phenotypic "domain", or across multiple domains. Previous studies have interpreted pleiotropy at different levels: multi-domain (associated with traits from multiple domains), domain-specific (associated with various traits from a single domain), and trait-specific (associated with a single trait), and identified more than 60% of genes with multi-domain associations through common variants across 27 domains[8]. Similarly, we classify the 599 high-quality phenotypes into six broad phenotypic domains (Extended Data Figure 1A), as well as categorize 234 disease phenotypes into 17 ICD-10-based disease domains (Figure S5A).

Statistical test for allelic heterogeneity between quantitative traits

When a gene affects two or more phenotypes, it raises the question of whether the underlying causal variants are the same. We propose a likelihood ratio test, ALLSPICE, to identify allelic heterogeneity within a pleiotropic gene, defined as an imperfect correlation between the effects of variants within that gene on the two traits. Our test uses summary statistics from single-variant association tests without using the original genotype and phenotype data for efficiency and accounts for the phenotypic correlation.

We assume that n samples are sequenced in the gene with m variants. Given a pair of phenotypes, let $y_i = (y_{i1},..., y_{in})^T$ be the i_{th} $(i = 1, 2)$ phenotype for the n samples and $\beta_i = (\beta_{i1},..., \beta_{im})^T$ be the effect sizes of the m variants on the i_{th} phenotype. For our single-variant association test on continuous phenotypes, we consider the linear regression model $Y = \beta X + \epsilon$, where $Y = (y_1, y_2)^T$, $\beta = (\beta_1, \beta_2)^T$, X is an $m \times n$ matrix representing the genotypes for n samples at m variation sites, and $\epsilon_{2\times n} \sim N(0, \Delta^2)$ as the environmental noise

term. We then define $Y \sim N(\beta X,\ R)$, where $R = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$ with r representing the phenotypic correlation between the pair of phenotypes, which is empirically computed as the correlation between effect sizes of all synonymous variants across the exome on the two phenotypes. We restrict to only the rare variants with minor allele frequency < 0.01% within the gene and therefore limit the effect of linkage disequilibrium (LD), and assume that the variants are independent and no individuals have two or more variants out of the m variants:

$$A = \frac{1}{n} diag(XX^T) \approx \frac{1}{n} XX^T. \quad (1)$$

We test the null hypothesis:

$$H_0: \beta_1 = c\beta_2, \quad (2)$$

for some constant c, which is the null hypothesis that the effect sizes of individual variants within a gene for the pair of phenotypes are perfectly correlated. The test statistic is defined as

$$\Lambda = 2log \frac{max_\beta P(Y|\ H_1)}{max_\beta P(Y|H_0)}, \quad (3)$$

following a chi-square distribution with $m - 1$ degrees of freedom.

First, we consider the simplest version, where $m = 1$ and $n = 1$. For the k_{th}individual $(k = 1)$ and the j_{th}variant $(j = 1)$, we have

$$P_{kj}(Y_{.k}|\ \beta_{.j}) \propto e^{-\frac{1}{2}(Y_{.k}-\beta_{.j}X_{jk})^T R^{-1}(Y_{.k}-\beta_{.j}X_{jk})}. \quad (4)$$

Denoting $\hat{\alpha} = \frac{1}{n} YX^T = \hat{\beta}A$, then we have

$$P_{kj}(Y_{.k}|\ \beta_{.j}) \propto e^{\alpha_{.j}^T R^{-1}\hat{\beta}_{.j} - \frac{1}{2}(\beta_{.j}X_{jk})^T R^{-1}(\beta_{.j}X_{jk})}. \quad (5)$$

Therefore, for n unrelated samples and one variant, we have

$$P(Y|\beta_{.j}) = \prod_{k=1}^{n} P_{kj}(Y_{.k}|\beta_{.j}) \propto e^{nA_{jj}(\hat{\beta}_{.j}^{T} - \frac{1}{2}\beta_{.j}^{T})R^{-1}\beta_{.j}}. \quad (6)$$

We then adapt the likelihood to the null and alternative hypothesis and compute the maximum likelihood estimator (MLE) of true effect sizes β_1 and β_2 under both conditions respectively. The MLEs of effect size under the null are

$$\beta_{1j,H_0} = c\beta_{2j,H_0} \text{ and } \beta_{2j,H_0} = \frac{c\hat{\beta}_{ij} - cr\hat{\beta}_{2j} - r\hat{\beta}_{1j} + \hat{\beta}_{2j}}{c^2 - 2cr + 1}. \quad (7)$$

And those for the alternative hypothesis are simply the summary statistics

$$\beta_{1j,H_1} = \hat{\beta}_{1j} \text{ and } \beta_{2j,H_1} = \hat{\beta}_{2j}. \quad (8)$$

Therefore, by plugging the MLEs back into the likelihood ratio, we have

$$\Lambda_j = 2log\frac{maxP(Y|H_1)}{maxP(Y|H_0)} = \frac{nA_{jj}(c\hat{\beta}_{2j} - \hat{\beta}_{1j})^2}{c^2 - 2cr + 1}. \quad (9)$$

Therefore, we end up with our final formula for our test statistic

$$\Lambda = \frac{n}{c^2 - 2cr + 1}\sum_{j=1}^{m} A_{jj}(c\hat{\beta}_{2j} - \hat{\beta}_{1j})^2. \quad (10)$$

To formulate the unknown parameter c, we take the derivative of $\Lambda(c)$ and compute the two extrema of Λ, which are

$$\hat{c}_1, \hat{c}_2 = \frac{-v + \sqrt{v^2 - 4uw}}{2u}, \frac{-v - \sqrt{v^2 - 4uw}}{2u},$$

where $u = \hat{\beta}_2 A(\hat{\beta}_1 - r\hat{\beta}_2)^T$, $v = (\hat{\beta}_2 - \hat{\beta}_1)A(\hat{\beta}_2 + \hat{\beta}_1)^T$, $w = \hat{\beta}_1 A(r\hat{\beta}_1 - \hat{\beta}_2)^T$. (11)

To obtain the maximum likelihood of the null hypothesis, we select the extremum that minimizes the value of Λ, which is $\hat{c} = max(\hat{c}_1, \hat{c}_2)$ when $u > 0$ and $\hat{c} = min(\hat{c}_1, \hat{c}_2)$ when $u < 0$. Finally, we have the equation for the likelihood ratio test statistic

$$\Lambda = \frac{n}{\hat{c}^2 - 2\hat{c}r + 1}(\hat{c}\hat{\beta}_2 - \hat{\beta}_1)A(\hat{c}\hat{\beta}_2 - \hat{\beta}_1)^T. \quad (12)$$

Note that with phenotypic correlation r close to one, the effect sizes of variants on the two phenotypes will be largely identical, resulting in the value of $\hat{c}$ being close to one, which causes instability in the model and thus inflation in the test statistics (Equation 10; Figure S13B). Therefore, we recommend avoiding running ALLSPICE on triplets where the two phenotypes are highly positively correlated, which also aligns with the motivation of ALLSPICE - highlighting variants with heterogeneous effects on phenotypes presenting horizontal pleiotropy.

Simulation design

To evaluate the performance of the test, we conducted extensive simulations by varying the key parameters to construct a wide range of scenarios. We simulated data from 378 combinations of parameter configurations under the null hypothesis and 32 scenarios under the alternative hypothesis to analyze its statistical power. To investigate the false positive rate of the test, we implemented simulations under the null hypothesis and computed the maximum likelihood estimation of c, which maximized the probability of the null hypothesis and minimized the test statistic along the way.

Simulation data generation

We generated minor allele counts of m rare variants from a $Uniform(1,\ 100)$ distribution. For a pair of quantitative phenotypes, the vector representing the true effect sizes of one phenotype (β_2) is simulated from a mixture distribution with a π chance of following a normal distribution $N(0,\ \sigma^2)$ and a $1 - \pi$ chance of being equal to 0. To simulate the condition under the null hypothesis, the vector representing the true effect sizes of the other phenotype (β_1) is defined by $\beta_1 = c\beta_2$, while under the alternative hypothesis, β_1 is generated separately using the same schema as we did for β_2. We then simulated the genotype matrix $X_{m\times n}$ as a $(0, 1)$ matrix with nA_{jj} $1s$ randomly distributed on the j_{th} row. With all that, we generated the phenotype

values of the two traits from a multivariate normal distribution $N(\beta X,\ R)$. We randomized $A,\ \beta\ and\ X$ for all simulations across each scenario, since controlling for any of the three terms does not make substantial changes to the test statistics. And finally, we estimated the summary statistic using $\hat{\beta} = \frac{1}{n} Y X^T A^{-1}$, which approximates the estimator of effect size under the assumption of a linear regression model.

Simulation results

To simulate data under the null hypothesis, we fix the number of samples n at 1000. We then vary parameters within the ranges below, and conduct 100 simulations per combination of parameter configurations:

- Number of variants $m\ \in \{5,\ 10,\ 20, 100\}$,
- True value of slope $c \in \{0,\ 0.2,\ 0.4,\ 0.6,\ 0.8,\ 1\}$
- Phenotypic correlation

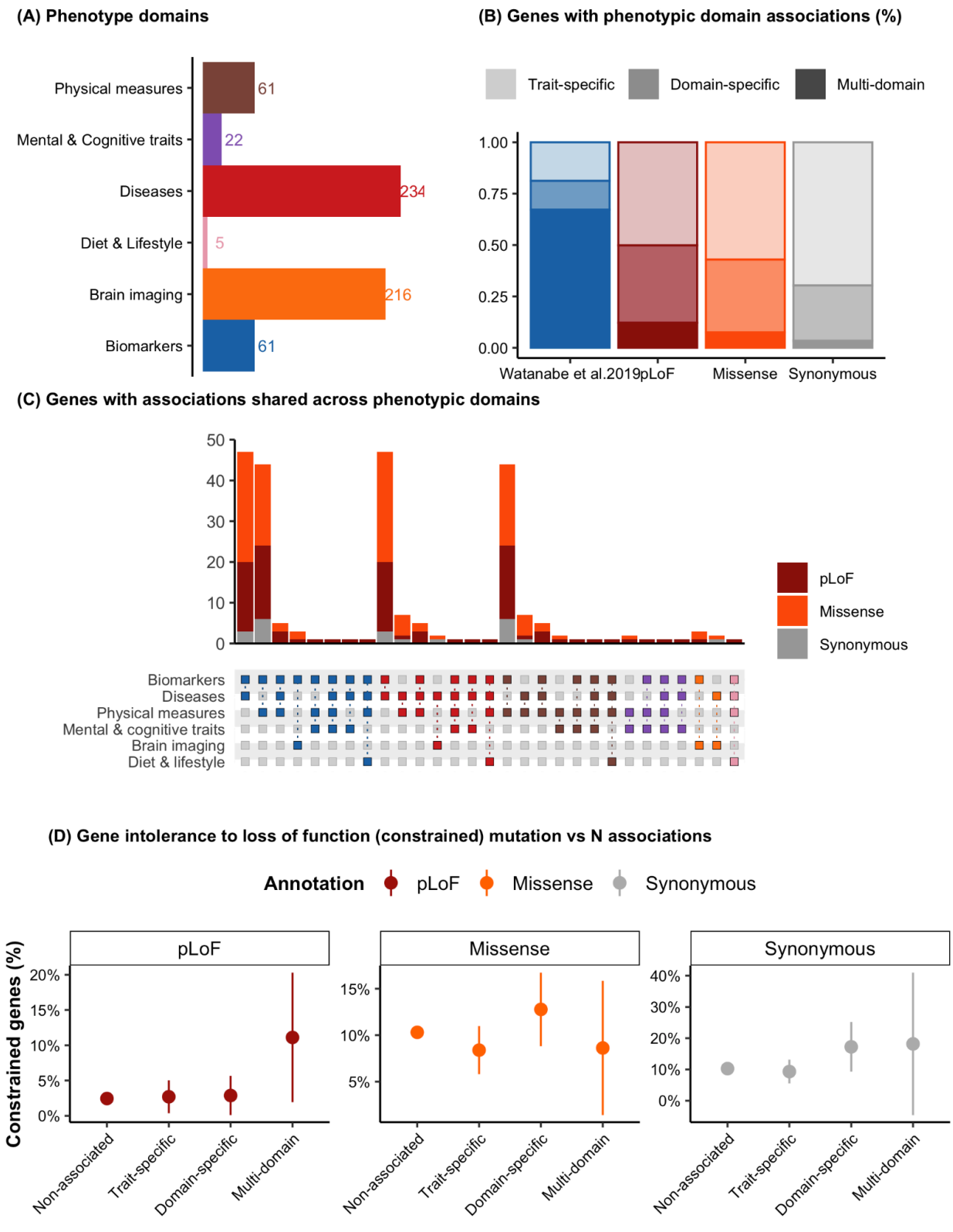
  $r \in \{\ -1,\ -0.8,\ -0.5,\ -0.2,\ -0.1,\ 0,\ 0.1,\ 0.2,\ 0.5,\ 0.8,\ 1\}$
- Variance of the component normal distribution of effect sizes $\sigma^2 \in \{0.01,\ 0.1,\ 1\}$
- Weight for mixture distribution of variant effects $\pi \in \{0.5,\ 0.8\}$

We then ran ALLSPICE on simulated data from all combinations of parameters (Extend Data Figure 2-3; Figure S6-S12; SuppTable 7-12). We observed a well-calibrated distribution of p-value under all scenarios when $\sigma^2 = 1$ and $\pi = 0.5$ (Extended Data Figure 2A), and we summarized the empirical type I error rate by computing the proportion of tests reaching the rejection region of the null hypothesis at 0.05 level of significance under the null. The type I errors centered around 0.05 across most scenarios. This measure inflates slightly as the number of variants increases, but it is still well controlled at a level below 0.1 (Extended Data Figure 2B). Test results for $\sigma^2 = 0.01\ \&\ 0.1$ present a slightly deflated pattern compared to

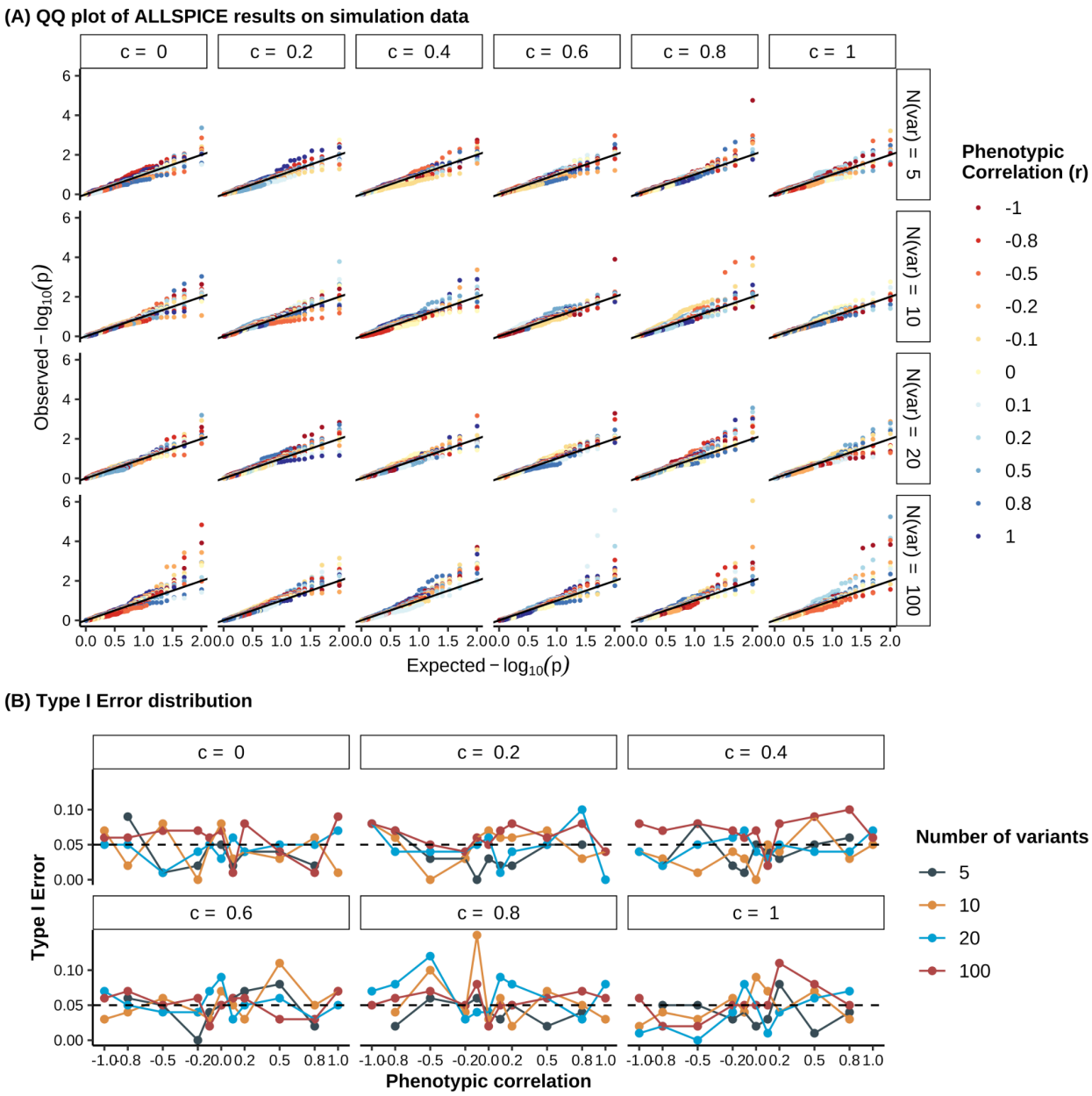
$\sigma^2 = 1$, due to the relatively small scale of effect sizes that have been simulated as σ^2 decreases (Figure S6, S7). Test results for $\pi = 0.8$ are as well calibrated as that for $\pi = 0.5$ (Figure S8, S9, S10).

The empirical statistical power is computed as the proportion of the tests that reach the rejection region of the null hypothesis at a 0.05 level of significance under the alternative. To generate simulated data under the alternative hypothesis, we constructed the vectors of effect sizes of the two phenotypes β_1 and β_2 independently from a mixture of two distributions, with a π chance of following a normal distribution $N(0, \sigma^2)$ and a $1 - \pi$ chance of being equal to 0. Parameter configuration is the same as previously described for simulations under the null scenario. We fixed the number of samples n at 1,000 and σ^2 at 1. For each scenario, we conducted 1,000 simulations to investigate the statistical power of the test. The power of the test increases as the number of variants increases, with power above 0.9 with at least 10 variants (Extended Data Figure 3). When $\sigma^2 = 0.1$ or 0.01, with less variation specified in the simulation, the overall scale of effect sizes (β_1 and β_2) becomes smaller, thus significantly reducing the power of detecting significant triplets (Figure S11 and Figure S12).

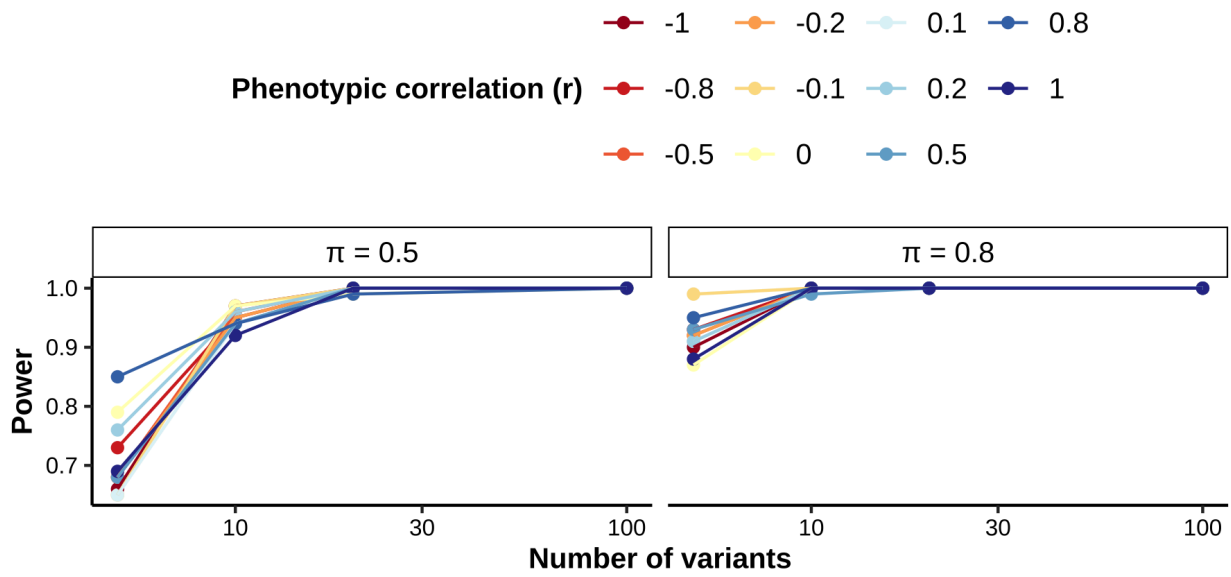
Extended Data Figures


(A) Phenotype domains
Physical measures
Mental & Cognitive traits
Diseases
Diet & Lifestyle
Brain imaging
Biomarkers
61
22
234
5
216
61
(B) Genes with phenotypic domain associations (%)
Trait-specific
Domain-specific
Multi-domain
1.00
0.75
0.50
0.25
0.00
Watanabe et al.2019
pLoF
Missense
Synonymous
(C) Genes with associations shared across phenotypic domains
50
40
30
20
10
0
pLoF
Missense
Synonymous
Biomarkers
Diseases
Physical measures
Mental & cognitive traits
Brain imaging
Diet & lifestyle
(D) Gene intolerance to loss of function (constrained) mutation vs N associations
Annotation
pLoF
Missense
Synonymous
Constrained genes (%)
pLoF
Missense
Synonymous
20%
15%
10%
5%
0%
15%
10%
5%
40%
30%
20%
10%
0%
Non-associated
Trait-specific
Domain-specific
Multi-domain

Extended Data Figure 1 | Pervasive gene-level pleiotropy across six broader phenotypic groups: **A,** Number of phenotypes in each of the six phenotypic domains (color); **B,** Proportion of genes with trait-specific, domain-specific, and multi-domain associations defined at the phenotypic domain level, among gene-annotation pairs with at least one association across all three functional annotation groups. The blue bar is the common variant results from Watanabe et al. 2019,[8]; **C,** Number of genes associated with each combination of phenotypic domains. The bars represent the number of genes with cross-phenotype associations defined by the dots in the matrix below. Colors are consistent in **A** and the bottom matrix in **C**. Each column of the matrix is a possible domain combination that shares any gene-level associations with the colored domain. **D,** Proportion of constrained genes (defined by the lowest LOEUF decile, y-axis) with no association, trait-specific, domain-specific and multi-domain associations (x-axis) in each functional annotation groups (color and column). Error bar indicates standard errors.



Extended Data Figure 2 | ALLSPICE test results on data simulated under the null hypothesis $H_0: \beta_1 = c\beta_2$ across 264 scenarios when $\sigma^2 = 1$ and $\pi = 0.5$. **A,** QQ-plot of test results on the simulated data described above. The solid black line represents line y=x. The rows represent the number of variants included and the columns indicate the true value of c, the slope between two β vectors used in simulating the data. **B,** Distribution of type I error (y-axis) of the test results on simulation data under the null hypothesis $H_0: \beta_1 = c\beta_2$ across combinations of phenotypic correlation (x-axis), number of variants (colors), and true values of slope c (panels). The dashed horizontal line represents the nominal significance level y=0.05.



Extended Data Figure 3 | Distribution of power (y-axis) of ALLSPICE test results for simulation data under the alternative hypothesis $H_1: \beta_1, \beta_2 \in R$, across phenotypic correlation (color), number of variants (x-axis), and weight for mixture components π (panels). The columns represent the probability π of the variant effect being from the non-zero standard normal distribution $N(0, \sigma^2)$, where $\sigma^2 = 1$.

Resources

GitHub repository: https://github.com/atgu/ALLSPICE

References

1. Satterstrom, F. K. *et al.* Autism spectrum disorder and attention deficit hyperactivity disorder have a similar burden of rare protein-truncating variants. *Nat. Neurosci.* **22**, 1961–1965 (2019).
2. Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.* **14**, 483–495 (2013).
3. O'Connor, L. J. & Price, A. L. Distinguishing genetic correlation from causation across 52 diseases and complex traits. *Nat. Genet.* **50**, 1728–1734 (2018).
4. Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy

in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* **50**, 693–698 (2018).

5. Ballard, J. L. & O'Connor, L. J. Shared components of heritability across genetically correlated traits. *Am. J. Hum. Genet.* **109**, 989–1006 (2022).
6. Shi, H., Mancuso, N., Spendlove, S. & Pasaniuc, B. Local Genetic Correlation Gives Insights into the Shared Genetic Architecture of Complex Traits. *Am. J. Hum. Genet.* **101**, 737–751 (2017).
7. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
8. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).
9. Sivakumaran, S. *et al.* Abundant pleiotropy in human complex diseases and traits. *Am. J. Hum. Genet.* **89**, 607–618 (2011).
10. Smith, C. J. *et al.* Integrative analysis of metabolite GWAS illuminates the molecular basis of pleiotropy and genetic correlation. *bioRxiv* 2022.04.02.486791 (2022) doi:10.1101/2022.04.02.486791.
11. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* **381**, 1371–1379 (2013).
12. Udler, M. S. *et al.* Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. *PLOS Medicine* vol. 15 e1002654 Preprint at https://doi.org/10.1371/journal.pmed.1002654 (2018).
13. Do, R. *et al.* Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nat. Genet.* **45**, 1345–1352 (2013).
14. Sinnott-Armstrong, N. *et al.* Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat. Genet.* **53**, 185–194 (2021).
15. Chami, N. *et al.* Exome Genotyping Identifies Pleiotropic Variants Associated with Red

Blood Cell Traits. *Am. J. Hum. Genet.* **99**, 8–21 (2016).

16. Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).
17. Karczewski, K. J. *et al.* Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. *Cell Genom* **2**, 100168 (2022).
18. Weiner, D. J. *et al.* Polygenic architecture of rare coding variation across 400,000 exomes. *bioRxiv* (2022) doi:10.1101/2022.07.06.22277335.
19. Kalia, S. S. *et al.* Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet. Med.* **19**, 249–255 (2017).
20. Rehm, H. L. *et al.* ClinGen--the Clinical Genome Resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015).
21. 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).
22. von Mering, C. *et al.* STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* **33**, D433–7 (2005).
23. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
24. Martinon, F. & Aksentijevich, I. New players driving inflammation in monogenic autoinflammatory diseases. *Nat. Rev. Rheumatol.* **11**, 11–20 (2015).
25. Laddach, A., Ng, J. C. F. & Fraternali, F. Pathogenic missense protein variants affect different functional pathways and proteomic features than healthy population variants. *PLoS Biol.* **19**, e3001207 (2021).
26. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).