Effect heterogeneity reveals complex pleiotropic effects of rare coding variants

Wenhan Lu[1,2,3], Siwei Chen[1,2,†], Chiara Auwerx[1,4,5], Jack Fu[1,4,5], Danielle Posthuma[6,7], Benjamin M. Neale[2,3], Luke J. O'Connor[1,8], Konrad J. Karczewski[1,2,9,*]

[1]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[2]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA
[3]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[4]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA
[5]Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
[6]Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[7]Department of Child and Adolescent Psychiatry and Pediatric Psychology, Section Complex Trait Genetics, Amsterdam Neuroscience, Vrije Universiteit Medical Center, Amsterdam University Medical Center, Amsterdam, The Netherlands
[8]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[9]Novo Nordisk Foundation Center for Genomic Mechanisms of Disease, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[†]Current affiliation: Department of Human Genetics, The University of Chicago, Chicago, IL
[*]Corresponding author

Abstract

Genome-wide association studies (GWAS) and rare-variant association studies (RVAS) have identified thousands of genes and variants associated with diverse human phenotypes, and have revealed pervasive pleiotropy among common variants; however, the extent of pleiotropy among rare variants remains incompletely characterized. Using rare variant association results from the UK Biobank (UKB), we systematically quantify gene-level pleiotropy across hundreds of phenotypes spanning multiple phenotypic domains and identify genes exhibiting allelic series. We develop ALLSPICE (ALLelic Spectrum of Pleiotropy Informed Correlated Effects), a statistical test for assessing variant effect heterogeneity within pleiotropic genes across quantitative traits. In *ALB*, a gene associated with both albumin and calcium levels, ALLSPICE reveals a subset of missense variants with heterogeneous effects across these traits. Protein structure-based analyses further demonstrate that missense variants associated with calcium levels but not albumin, as well as variants with opposing effects on the two traits, are enriched near known calcium-binding sites in the ALB protein. In this way, we present a systematic view of gene-level pleiotropy and effect heterogeneity among rare variants, illustrating how shared and trait-specific genetic effects can arise within the same gene. Our analysis provides a general approach for dissecting pleiotropic architecture and detecting rare variants with heterogeneous effects for downstream biological investigation of complex human traits.

Introduction

Pleiotropy, defined as the association of a single genetic variant with multiple phenotypes, is pervasive across the human genome. Pleiotropy can help characterize the genetic architecture of complex human diseases by indicating shared or independent molecular pathways. Better interpretation of the pleiotropic effects of genetic risk factors can not only enhance the statistical power of gene discovery by pooling samples across phenotypes with a shared genetic basis[1], but it can also facilitate the development of therapeutic approaches and diagnosis strategies by revealing shared genetic mechanisms that inform target prioritization across traits[2,3].

Pleiotropy can manifest through diverse mechanisms, including the involvement of a single gene or its product in multiple biological pathways or cellular functions, as well as the generation of multiple primary products from the same gene through alternative splicing, resulting in distinct transcript isoforms across tissues or cell types[4,5]. In addition, pleiotropic effects may arise from interactions among genes or between genes and their molecular targets, leading to diverse phenotypic outcomes[6]. One way pleiotropy arises is when a gene affects multiple traits through a single biological pathway, leading to vertical pleiotropy, which is primarily addressed by causal inference approaches such as Mendelian Randomization (MR)[7]. However, many pleiotropic relationships do not reflect a causal effect of one phenotype on another[8], and instead, they reflect horizontal pleiotropy, in which the same genetic components, or their downstream products, independently influence multiple traits through distinct or partially separable biological mechanisms. For example, mutations in *MC1R* reduce melanin synthesis, resulting in both fairer skin pigmentation and redder hair color[9], despite neither trait directly driving the other. Notably, even causally related traits can exhibit horizontal pleiotropy[10].

Recently, well-powered GWAS have revealed widespread pleiotropic effects among common variants[11–13]. In parallel, numerous analytical methods[8,14,15,16,17,18] have been developed to dissect the pleiotropic structure underlying shared genetic architecture across phenotypes,

with a focus on common variations. Beyond common variants, exome-wide association studies have identified pleiotropic rare variants associated with specific phenotypes, such as urine and blood biomarkers[19,20], ASD and ADHD[1], and genetic generalized epilepsy and developmental delay[21]. However, these studies have typically focused on specific sets of phenotypes, rather than a comprehensive view of rare-variant pleiotropy across the phenome. Over the past decade, rapid development of large-scale sequencing data from population biobanks has enabled gene-based association testing across thousands of phenotypes[22,23]. These resources have been used to characterize the polygenic architecture of rare protein-coding variants[24], providing new opportunities to examine pleiotropy across diverse phenotypic domains. Such cross-domain associations can arise from heterogeneous effects of genetic variants, including missense variants that alter protein structure or function, and predicted loss-of-function (pLoF) variants that reduce or disrupt gene activity. Moreover, differences in cellular environments can influence how these variants interact with biological pathways and molecular networks, further contributing to phenotypic diversity. Understanding these effects offers critical insight into gene-level pleiotropy and establishes the foundation for systematically dissecting the mechanisms underlying pleiotropic genetic associations.

Here, we leverage gene-based association statistics from Genebass[23], derived from 394,841 UKB participants across 4,529 phenotypes, to characterize pleiotropy among rare protein-coding variants across a broad spectrum of phenotypes. We highlight gene-level pleiotropy by investigating allelic series of rare protein-coding variants within genes and by quantifying the extent of pleiotropy across gene sets with established functional relevance in diseases. Conditioning on the shared associations identified at the aggregated gene level, we examine the effect of rare protein-coding variants within genes, assessing whether their influences on the same phenotype pairs are shared or heterogeneous. We develop ALLSPICE, a statistical test to examine the existence of heterogeneous variant effects within a pleiotropic gene across quantitative traits. We then apply ALLSPICE to the pleiotropic genes identified in

the Genebass RVAS results and compare the directions of effects across variant functional consequence groups. By investigating the underlying mechanisms of pleiotropic genes through statistical inference, we identify genes associated with distinct phenotypes and evaluate the magnitudes and directions of the variant effects underlying the target gene, thereby distinguishing associations shared by genes that demonstrate heterogeneous variant effects from those due to common biological pathways.

Results

Pervasive gene-level pleiotropy across the human exome

To quantify the extent of pleiotropy among rare coding variants, we selected 599 biological measurements and health consequences with at least one significant association ($p_{burden} < 2.5 \times 10^{-6}$) through gene-based burden tests (**Methods**) and considered a gene-annotation pair as pleiotropic if it had significant associations with two or more of these phenotypes (**Figure 1A**). To more stringently quantify pleiotropy across trait-specific associations, we curated a maximal subset of 239 approximately independent phenotypes ($r^2 < 0.1$; **Methods**) and identified 313 pleiotropic gene-annotation pairs (**SuppTable1**), involving 185 unique genes. Among those pairs with at least one association, the observed fractions of genes demonstrating pleiotropic effect through pLoF, missense, pLoF+missense, and synonymous variants are 28%, 23%, 24%, and 11%, respectively (**Figure 1B**). These are overdispersed compared to the expected Poisson distribution and the p-value-permuted null across all four functional groups, indicating variance exceeding Poisson mean and suggesting non-random structure beyond the conditional null expectations, motivating further investigation (**Table S1 and Figure S1**).

The functional importance of genes often reflects their central role in key biological processes and their broad phenotypic influence. We explored pleiotropy across 37 curated gene categories with established functional or etiological relevance

(https://github.com/macarthur-lab/gene_lists). For pLoF and missense variants, pleiotropic genes are more enriched within disease-related groups, such as ACMG genes[25] and the BROCA cancer risk panel, consistent with their known clinical relevance. In contrast, genes with essential biological functions showed fewer associations overall and consequently fewer pleiotropic associations, compared to background rates calculated across all protein-coding genes, reflecting reduced detection of large-effect variants under negative selection[26–28]. Notably, ClinGen haploinsufficient genes[29] exhibited more pleiotropy among pLoF associations, consistent with dosage sensitivity where heterozygous loss of function of a single allele is sufficient to perturb multiple phenotypes (**Figure 1C, Figure S3 & SuppTable3**). pLoF and missense groups of pleiotropic genes also exhibited a greater number of protein-protein interactions than trait-specific or non-association genes, indicating that they may occupy more highly connected positions in protein interaction networks (**Figure S4**).

Under a highly polygenic framework, common variants are more likely to be identified as pleiotropic, in part because they are easier to detect. Consistent with this, we observe frequency-dependent differences in pleiotropy across functional annotation groups (**Figure S2**). However, logistic regression analyses indicate that combined allele frequency (CAF) has a limited impact on pleiotropic status after adjusting for coding sequence (CDS) length, with no consistent effects across annotation groups (**Table S2**). To further evaluate the potential frequency-dependent ascertainment bias, we also applied permutation tests within gene groups stratified by CAF bins and functional annotations. Across most CAF-annotation combinations with sufficient observations, the proportion of pleiotropic genes in the permuted data remains substantially lower than the observed proportion, indicating that the detected pleiotropy is not driven by frequency alone. These results suggest that while frequency and gene features modestly influence detectability, additional multi-trait genetic effects contribute to the observed pleiotropic signals and remain to be uncovered (**Figure S1 and Table S1**).

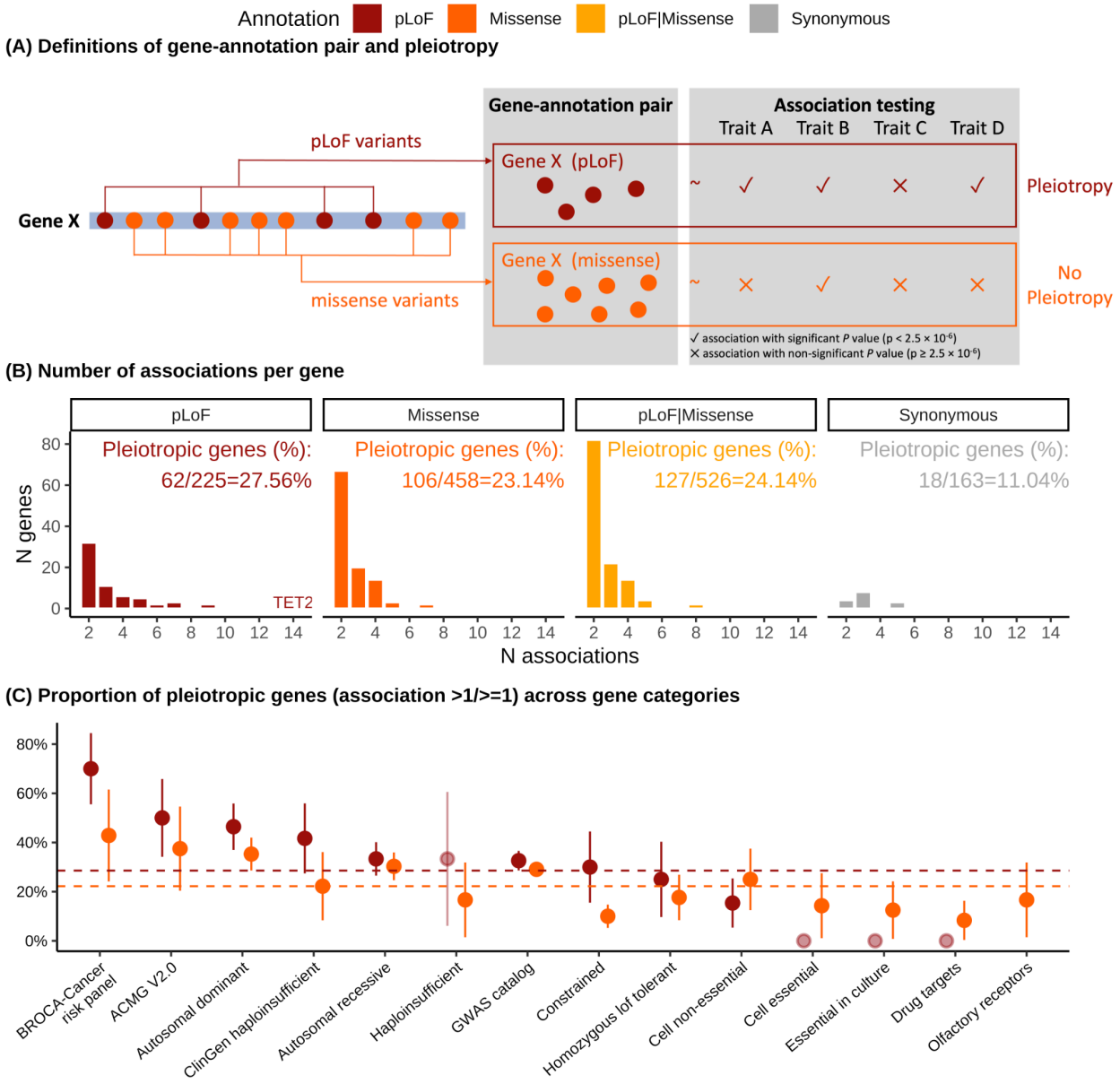

Annotation
pLoF
Missense
pLoF|Missense
Synonymous
(A) Definitions of gene-annotation pair and pleiotropy
Gene-annotation pair
Association testing
Trait A
Trait B
Trait C
Trait D
pLoF variants
Gene X
Gene X (pLoF)
Pleiotropy
Gene X (missense)
No Pleiotropy
missense variants
✓ association with significant P value (p < 2.5 × 10-6)
✕ association with non-significant P value (p ≥ 2.5 × 10-6)
(B) Number of associations per gene
pLoF
Missense
pLoF|Missense
Synonymous
Pleiotropic genes (%): 62/225=27.56%
Pleiotropic genes (%): 106/458=23.14%
Pleiotropic genes (%): 127/526=24.14%
Pleiotropic genes (%): 18/163=11.04%
N genes
TET2
N associations
(C) Proportion of pleiotropic genes (association >1/>=1) across gene categories
80%
60%
40%
20%
0%
BROCA-Cancer risk panel
ACMG V2.0
Autosomal dominant
ClinGen haploinsufficient
Autosomal recessive
Haploinsufficient
GWAS catalog
Constrained
Homozygous lof tolerant
Cell non-essential
Cell essential
Essential in culture
Drug targets
Olfactory receptors

Figure 1 | Pervasive gene-level pleiotropy among rare protein-coding variants across 239 approximately independent phenotypes. **A,** Schematic of gene-annotation pairs, defined as sets of rare variants within a gene grouped by functional annotations (pLoF, missense, synonymous, and pLoF+ missense) for burden testing, and of pleiotropy, defined as a gene-annotation pair showing significant associations ($p_{burden} < 2.5 \times 10^{-6}$) with more than one phenotype. **B**, Distributions of genes by the number of associated phenotypes (x-axis) across four variant functional groups (panels and colors). Labels above the bars in each panel indicate the proportion of genes with more than one association among genes with at least one association in the corresponding functional annotation group. **C,** Gene-level pleiotropy across 14 selected gene categories. The proportion of pleiotropic genes is calculated 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 across the 239 phenotypes. Transparent points indicate categories with fewer than five genes tested. The dashed line denotes the overall proportion of pleiotropic genes, and error bars represent standard errors. Colors are consistent across panels.

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

One form of pleiotropy is an allelic series, in which mutations within the same gene produce a spectrum of related phenotypic outcomes, often reflecting differences in functional severity or molecular mechanisms, such as pLoF versus missense variants[30]. In practice, allelic series are often interpreted within a shared phenotypic pathway, where perturbation of the same gene affects both a proximal quantitative trait and a related clinical endpoint, as exemplified by *PCSK9* variants influencing LDL cholesterol levels and coronary heart disease risk[31]. Here, focusing on biomarkers and disease outcomes, we investigated allelic series in gene-based association results by identifying gene-phenotype associations that are significant for either pLoF and missense variants, but not both, highlighting heterogeneity in variant effects across functional annotation groups.

Across all 599 phenotypes (without restricting by phenotypic correlation), we identified 447 genes associated with more than one phenotype for either pLoF or missense group, among which 31 genes show associations with different sets of phenotypes between their pLoF and missense groups (**SuppTable2**). For example, pLoF variants in *ATM*, a well-documented tumor suppressor gene, are associated with six non-hematologic cancer diagnoses and agranulocytosis, a condition that can arise as a side effect of chemotherapy [32,33], whereas

missense variants in *ATM* are associated with three erythrocyte-related indices, which may capture cancer-related anemia rather than overt malignancy[34]. Similarly, *PKD1*, *SLC34A3*, and *TMPRSS6* show pLoF associations with diseases but missense associations with blood biomarkers. On the other hand, *FAM234A*, *JAK2*, and *MICA* show the opposite pattern: missense associations with disease and pLoF associations with blood biomarkers (**Table 1**, **Figure S5** and **SuppTable4**). These differential associations are consistent with the possibility that different variant classes can have different functional consequences within the same gene.

Motivated by the observation that different variant classes within the same gene can be associated with distinct phenotypic sets, we examined the extent to which gene-level associations extend across broader phenotypic domains. Across six phenotypic domains constructed from the 599 curated phenotypes (**Extended Data Figure 1A**), we identify 191 gene-annotation pairs exhibiting multi-domain associations (**Extended Data Figure 1C**). Among gene-annotation pairs with at least one association, more deleterious functional classes showed a higher proportion of multi-domain associations (12.2% for pLoF, 7.5% for missense, and 3.4% for synonymous) and a lower proportion of trait-specific associations (50.1% for pLoF, 57.1% for missense, and 69.7% for synonymous) (**Table S3 and Extended Data Figure 1B**). A similar relationship between functional deleteriousness and association breadth was observed across 17 disease-level domains (**Table S4 and Figure S6**).

Among genes with multi-domain associations, pLoF variants in genes *GIGYF1* and *KDM5B* are each associated with the most (n = 4) distinct phenotypic domains. For *KDM5B*, these associations span atrial fibrillation and flutter[35], several blood biomarkers, physical measures including hand grip strength[36], spirometry traits, and fluid intelligence score[37]. For *GIGYF1*, associations include hand grip strength[36], current tobacco smoking, biomarkers such as HbA1c and glucose, and type 2 diabetes, consistent with prior reports linking *GIGYF1* to endocrine-related traits[38] (**Table 2**). As for examples of domain-specific pleiotropy, missense variants in *OCA2*, a gene involved in pigmentation biology through its role in melanin

production[39], are associated with both skin changes related to chronic exposure to non-ionizing radiation ($p_{burden} = 6.15 \times 10^{-8}$) and basal cell carcinoma ($p_{burden} = 2.02 \times 10^{-6}$). In addition, ankylosing spondylitis and iridocyclitis share gene-level associations, including missense variants in *NELFE* ($p_{ankylosing\ spondylitis} = 3.44 \times 10^{-13}$; $p_{iridocyclitis} = 9.01 \times 10^{-14}$), *PSMB9* ($p_{ankylosing\ spondylitis} = 1.95 \times 10^{-8}$; $p_{iridocyclitis} = 5.00 \times 10^{-8}$), and *SLC44A4* ($p_{ankylosing\ spondylitis} = 1.45 \times 10^{-7}$; $p_{iridocyclitis} = 3.6 \times 10^{-7}$). Among these, *PSMB9* encodes a subunit of the immunoproteasome, which processes intracellular proteins into peptides for presentation by MHC Class I molecules, including HLA-B27, a genetic risk factor for ankylosing spondylitis[40,41]. *NELFE*, which regulates transcriptional elongation, and *SLC44A4*, a transporter gene, are not disease-specific but may reflect shared regulatory or systemic genetic influences contributing to these inflammatory phenotypes.

Table 1 | Genes for which pLoF variants are associated with disease phenotypes and missense variants with biomarkers and no diseases, or vice versa. Disease phenotypes are shown in bold.

Gene	pLoF-only associations	Missense-only associations	Both pLoF and missense associations
ATM	•**Cancer diagnosed by doctor** •**C25 Malignant neoplasm of pancreas** •**C50 Malignant neoplasm of breast** •**C77 Secondary and unspecified malignant neoplasm of lymph nodes** •**C78 Secondary malignant neoplasm of respiratory and digestive organs** •**C79 Secondary malignant neoplasm of other sites** •**D70 agranulocytosis**	•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
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
JAK2	•Eosinophill count •Eosinophill percentage	•**D75 other diseases of blood and blood-forming organs** •Mean platelet (thrombocyte) volume •Red blood cell (erythrocyte)	•IGF-1 •Platelet count •Platelet crit

		distribution width	
SLC34A3	•**N20 calculus of kidney and ureter**	•Urea	•Cystatin C •Creatinine •Phosphate
FAM234A	•Glucose	•**Diabetes diagnosed by doctor**	
MICA	•Lymphocyte count	•**Hypothyroidism/myxoedema** •**E03 other hypothyroidism**	

Table 2 | Gene burden associations ($p_{burden} < 2.5 \times 10^{-6}$) of pLoF variants in *GIGYF1* and *KDM5B* across phenotypic domains.

Phenotypic domain	***GIGYF1***		***KDM5B***	
	Phenotype	p_{burden}	**Phenotype**	p_{burden}
Diseases	E11 Type 2 Diabetes	1.55×10^{-11}	I48 atrial fibrillation and flutter	2.19×10^{-7}
	J44 Other Chronic obstructive pulmonary disease	1.57×10^{-6}		
	E03 Other hypothyroidism	1.20×10^{-7}		
Biomarkers	Apolipoprotein A	1.27×10^{-6}	IGF-1	2.29×10^{-12}
	Apolipoprotein B	1.18×10^{-12}	Red blood cell (erythrocyte) distribution width	1.15×10^{-6}
	Cholesterol	2.57×10^{-16}	Creatinine	1.34×10^{-6}
	LDL direct	6.45×10^{-13}		
	Total protein	2.08×10^{-6}		
	HDL cholesterol	4.69×10^{-8}		
	Glycated haemoglobin (HbA1c)	2.19×10^{-20}		
	IGF-1	2.03×10^{-6}		
	Cytatin C	6.85×10^{-8}		
	Glucose	1.22×10^{-14}		
Physical measurements	Peak expiratory flow (PEF)	7.40×10^{-9}	Peak expiratory flow (PEF)	3.48×10^{-8}
	Hand grip strength (left)	2.77×10^{-14}	Hand grip strength (left)	2.15×10^{-12}
	Hand grip strength (right)	4.71×10^{-10}	Hand grip strength (right)	7.49×10^{-11}
			Forced expiratory volume in 1 second (FEV1)	3.37×10^{-10}
			Forced vital capacity (FVC)	9.81×10^{-8}
Mental & cognitive measurements			Fluid intelligence score	1.38×10^{-6}
			Mean time to correctly identify	1.33×10^{-12}

			matches
Diet & lifestyle	Current tobacco smoking	1.77×10^{-6}	

Statistical test to model heterogeneous effects within pleiotropic genes

To better characterize the heterogeneity of effects within pleiotropic genes, we developed a statistical method, ALLSPICE (ALLelic Spectrum of Pleiotropy Informed Correlated Effects), to distinguish differences in effect patterns across functional classes of variants. In particular, the method accounts for functionally influenced differences in effect directionality: pLoF variants disrupt gene function through nonsense-mediated decay, and therefore, tend to produce more consistent phenotypic effects across variants within a gene, whereas missense variants can have diverse impacts on protein function, potentially leading to greater heterogeneity in phenotypic effects.

Here, we define a triplet as a combination of gene-phenotype1-phenotype2, in which the gene is associated with both phenotypes through burden tests ($p_{burden} < 2.5 \times 10^{-6}$; **Figure 2A**). For each triplet, ALLSPICE evaluates whether rare protein-coding variants show heterogeneous effects through a likelihood ratio test. The test incorporates phenotypic correlation between the two traits, allele frequencies of rare variants within the same functional annotation class in a gene, and their effect sizes on both phenotypes from single-variant association testing.

Under the null hypothesis, we assume a perfectly correlated linear relationship between the variant effect sizes on the two phenotypes with slope c, which is calculated through maximum likelihood estimation (MLE). ALLSPICE fits the model and constructs a likelihood ratio test statistic by identifying the value of c that maximizes the likelihood. A non-null p-value indicates deviation from perfect correlation, suggesting either heterogeneous variant effects on the two phenotypes or, more generally, imperfectly correlated phenotypic effects within the gene. In contrast, a null p-value indicates that the variant effects are consistent with a shared linear relationship between the two phenotypes, given their phenotypic correlation (**Methods**).

Although ALLSPICE does not explicitly classify pleiotropy types, it is informative for distinguishing vertical pleiotropy from certain forms of horizontal pleiotropy: vertical pleiotropy is expected to yield correlated effects, whereas horizontal pleiotropy may produce heterogeneity in effect sizes when different variants affect distinct molecular functions and traits[42] (**Figure 2A**). In simulation studies, ALLSPICE produces uniformly distributed p-values under the null hypothesis (**Extended Data Figure 2** and **Figures S7-S10**) and demonstrates adequate statistical power under the alternative hypothesis (**Extended Data Figure 3** and **Figures S11-S13**).

We first assessed heterogeneity within a functional annotation class by applying ALLSPICE to 11,810 pleiotropic triplets derived from the 359 quantitative traits (4,708 pLoF, 5,911 missense, and 1,191 synonymous). Each triplet consists of a gene-annotation pair and two quantitative phenotypes, both significantly associated with the pair at exome-wide significance ($p_{burden} < 2.5 \times 10^{-6}$). Analyses were first restricted to gene-annotation pairs with at least two variants contributing to the corresponding burden tests (**SuppTable 3**). To further reduce potential confounding from linkage disequilibrium, we filtered rare protein-coding variants within each gene-annotation pair to those with MAF below 0.01%, and retained only triplets with at least two variants after filtering. Among these 11,810 triplets, 1,073 showed nominal evidence of heterogeneous variant effects ($p_{ALLSPICE} < 0.05$), including 292 pLoF, 681 missense, and 100 synonymous triplets. After correction for multiple testing ($p_{ALLSPICE} < 4.23 \times 10^{-6}$; 0.05/11,810 tests), 124 triplets remained significant (53 pLoF, 70 missense, and 1 synonymous; **Figure 2B**, and **C** and **Figure S14**). Consistent with power considerations, both the number of pLoF and the number of missense variants in a gene's triplet positively correlated with the likelihood for that triplet to be significant (correlation estimate = -0.09, $p_{pearson's}=1.5 \times 10^{-21}$; **Figure S14C**), indicating heterogeneous effects are more readily detected in genes with larger numbers of rare variants, whereas power is reduced for genes with sparse variant counts.

Notably, we identified 53 pLoF triplets with significant ALLSPICE results (**Figure 2B**), despite pLoF variants often being assumed to have largely uniform effect directions. These

significant triplets were concentrated in two relatively long genes with broad biological relevance: *APOB* (CDS length 13,689; 10 significant triplets), and *TET2* (CDS length 6006; 43 significant triplets). *APOB* encodes a structural component of apoB-containing lipoproteins, and loss-of-function mutations in *APOB* cause familial hypobetalipoproteinemia, with reduced low-density lipoprotein cholesterol and apolipoprotein B[43]. *TET2* encodes an epigenetic regulator of hematopoietic stem and progenitor cells, and loss-of-function mutations in *TET2* affect myeloid differentiation and clonal expansion across blood cell lineages[44]. These well-characterized gene-specific effects provide context for the heterogeneous pLoF associations observed across phenotypes. We also applied ALLSPICE to the same set of triplets restricted to ultra-rare protein-coding variants with Allele Counts (AC) < 5, which presented similar results consistent with those from rare protein-coding variants (AF < 0.0001) with a slightly refined set of significant signals (**Figure S17; SuppTable 6**).

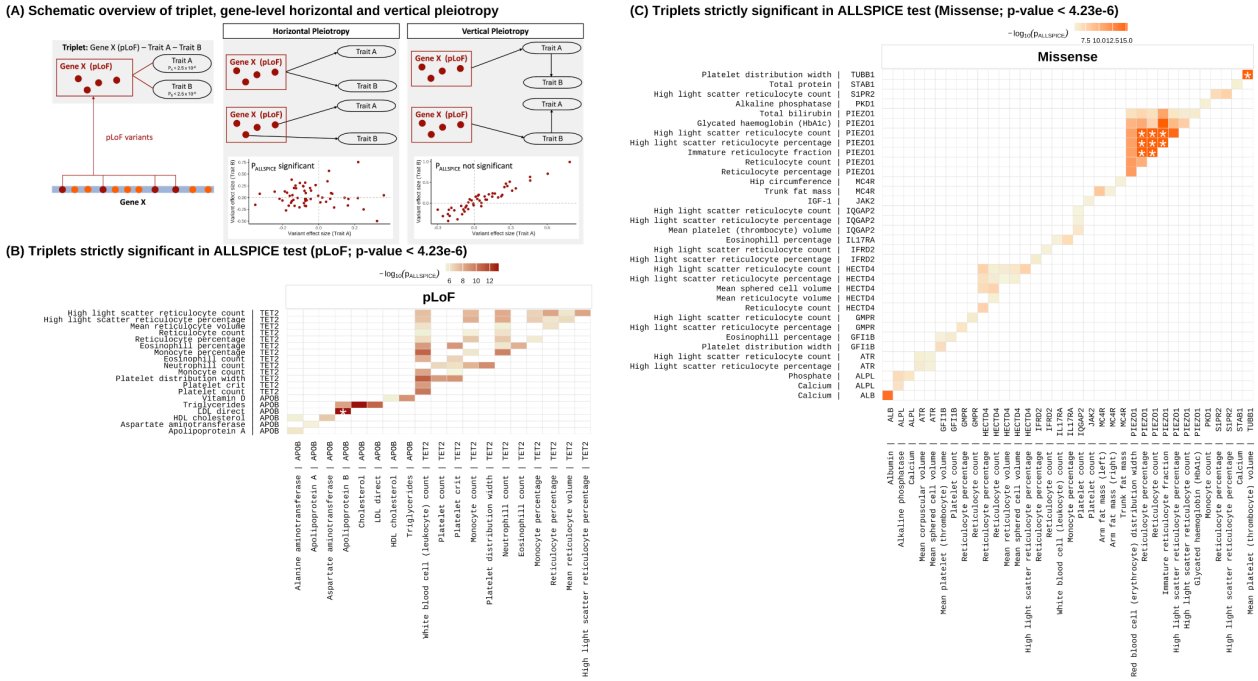


Figure 2 | Conceptual framework and summary of ALLSPICE results. **A**, Definition of a triplet (left), schematic illustrations of gene-level horizontal (middle) and vertical (right) pleiotropy, and examples of the corresponding expected variant-level effect size patterns shown in the scatter plots. **B**, Heatmap of $p_{ALLSPICE}$ for strictly significant pLoF triplets: with phenotype pairs on the x- and y-axes (genes labeled accordingly). Cell color indicates $-\log_{10}(p_{ALLSPICE})$, shown separately by functional annotation group. Cells with asterisk labels are triplets with $p_{ALLSPICE}$=0. **C**, Heatmap of $p_{ALLSPICE}$ for strictly significant missense triplets, with the same display as in **B.**

Examples of variant effect heterogeneity within pleiotropic genes

An illustrative example from ALLSPICE test is *ALB*, which encodes albumin and is associated with blood calcium and albumin levels through both pLoF ($p_{albumin} = 2.40 \times 10^{-256}$; $p_{calcium} = 5.85 \times 10^{-40}$) and missense ($p_{albumin} = 1.73 \times 10^{-62}$; $p_{calcium} = 1.99 \times 10^{-16}$) variants. Circulating calcium is partially albumin-bound; therefore, total serum calcium levels depend on albumin concentration[45,46], as supported by the observed phenotypic correlation between albumin and calcium ($r^2 = 0.51$). ALLSPICE result on triplet *ALB*-albumin-calcium reveals strong evidence of heterogeneous effects among rare missense variants ($p_{ALLSPICE} = 5.5 \times 10^{-15}$), whereas pLoF variants show no evidence of heterogeneity ($p_{ALLSPICE} = 0.557$), consistent with the uniform directional effects both observed and expected (**Figure 3A**). Notably, a missense variant at chr4:73412072:A:G (ENSP00000295897.4:p.Lys264Glu) is associated with increased calcium levels but decreased albumin levels, suggesting that the effect of some missense variants on calcium is not fully explained by their effect on albumin concentration.

To assess which missense variants contribute most strongly to the ALLSPICE signal, we performed a jackknife (leave-one-variant-out) analysis, rerunning ALLSPICE repeatedly while removing one variant at a time and evaluating the changes in c_{MLE} and the $p_{ALLSPICE}$. In the *ALB*-albumin-calcium triplet, removal of 159 variants, primarily variants with near-zero effects on both phenotypes, increased ALLSPICE test significance (transparent points in **Figure S15A**), whereas removal of the remaining 235 variants attenuated the signal. Among these 253 variants, chr4:73412072:A:G produced the largest change in c_{MLE} (**Figure S15B; SuppTable 7**)

and the 9th largest change in $p_{ALLSPICE}$ (**Figure S15A; SuppTable 7**), indicating a disproportionate contribution of this variant to the observed ALLSPICE signal.

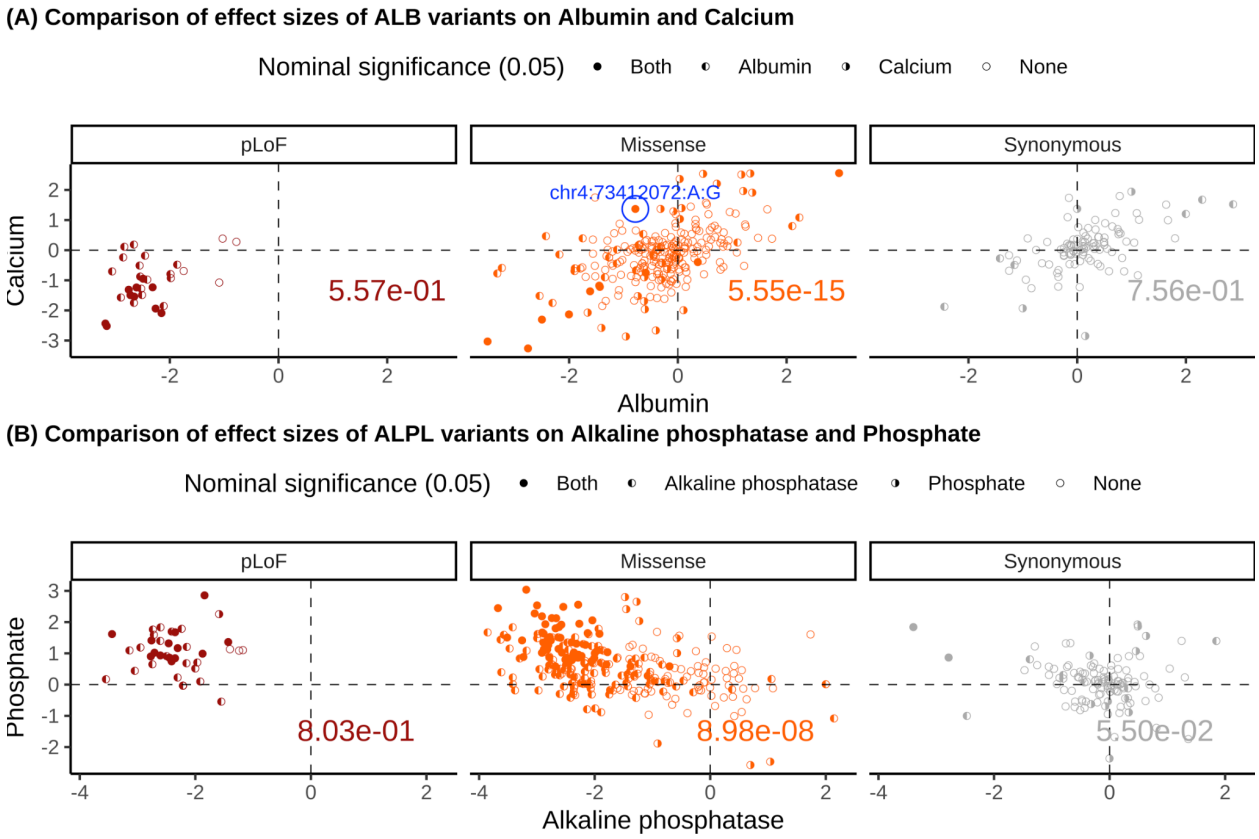


Figure 3 | Examples of significant ALLSPICE triplets. **A,** Comparison of variant effect sizes on albumin (x-axis) and calcium (y-axis) for variants in *ALB*, stratified by functional annotation group (colors and columns). Point shapes indicate nominal significance ($p_{GWAS} < 0.05$) of single-variant association tests for the two phenotypes. The missense variant chr4:73412072:A:G, labeled and highlighted with a blue circle, is significantly associated with both traits but exhibits opposite directions of effect and represents one of the strongest contributors in the LOVO analysis. The value shown in the bottom right of each panel denotes the $p_{ALLSPICE}$ for the corresponding annotation group. **B,** Comparison of *ALPL* variant effect sizes on alkaline phosphatase (x-axis) to phosphate level (y-axis). Panels **A** and **B** share the same legend.

We further examined whether *ALB* missense variants associated with calcium versus albumin levels ($p_{burden} < 0.05$) show different spatial distributions on the *ALB* protein (**Methods**). Missense variants associated with calcium levels are located significantly closer to known calcium-binding sites in the 3D protein structure of *ALB*, compared with variants exclusively

associated with albumin level ($p_{protein} = 1.99 \times 10^{-10}$; **Figure 4A, B**). Missense variants associated with albumin- and calcium- levels in an opposite direction also tend to be located closer to the calcium-binding sites at the 3D protein level ($p_{protein} = 6.2 \times 10^{-4}$; **Figure 4C, D**). These spatial patterns indicate that *ALB* missense variants near calcium-binding sites may affect calcium levels by directly disrupting ALB's calcium-binding capacity, and thereby producing distinct functional effects on albumin-related downstream biological processes. The *ALB*-albumin-calcium triplet also has a significant ALLSPICE result when restricted to ultra-rare missense variants with AC < 5 ($p_{ALLSPICE} = 3.5 \times 10^{-7}$; **Figure S17; SuppTable6**), for which the 3D protein distribution pattern persists (**Figure S18**).

Another ALLSPICE-significant example is *ALPL*, which regulates alkaline phosphatase production and is associated with a wide range of biomarkers. Missense variants in *ALPL* are associated with calcium levels ($p_{burden} = 4.15 \times 10^{-11}$) and C-reactive protein ($p_{burden} = 2.45 \times 10^{-6}$; **Figure S16A**), while both pLoF and missense variants are associated with alkaline phosphatase and phosphate levels (pLoF: $p_{alkaline\ phospate} = 1.99 \times 10^{-110}$, $p_{phospate} = 5.56 \times 10^{-8}$; missense: $p_{alkaline\ phospate} = 0$, $p_{phospate} = 2.17 \times 10^{-175}$). Alkaline phosphatase catalyzes phosphate removal, leading to an inverse relationship between the two biomarkers. Consistent with this biological mechanism, most pLoF variants are associated with decreased alkaline phosphatase levels and increased phosphate levels, and show no evidence of heterogeneous variant effects across the two traits ($p_{ALLSPICE} = 0.803$), consistent with the role of pLoF variants in disrupting *ALPL* activity. In contrast, missense variants exhibit significant heterogeneity ($p_{ALLSPICE} = 8.98 \times 10^{-8}$; **Figure 3B**), displaying a mixture of opposing and concordant effects on alkaline phosphatase and phosphate, concordant with more heterogeneous functional consequences. Unlike *ALB*, missense variants in *ALPL* show no evidence of functional clustering in the 3D protein structure (**Figures S16B, C**).

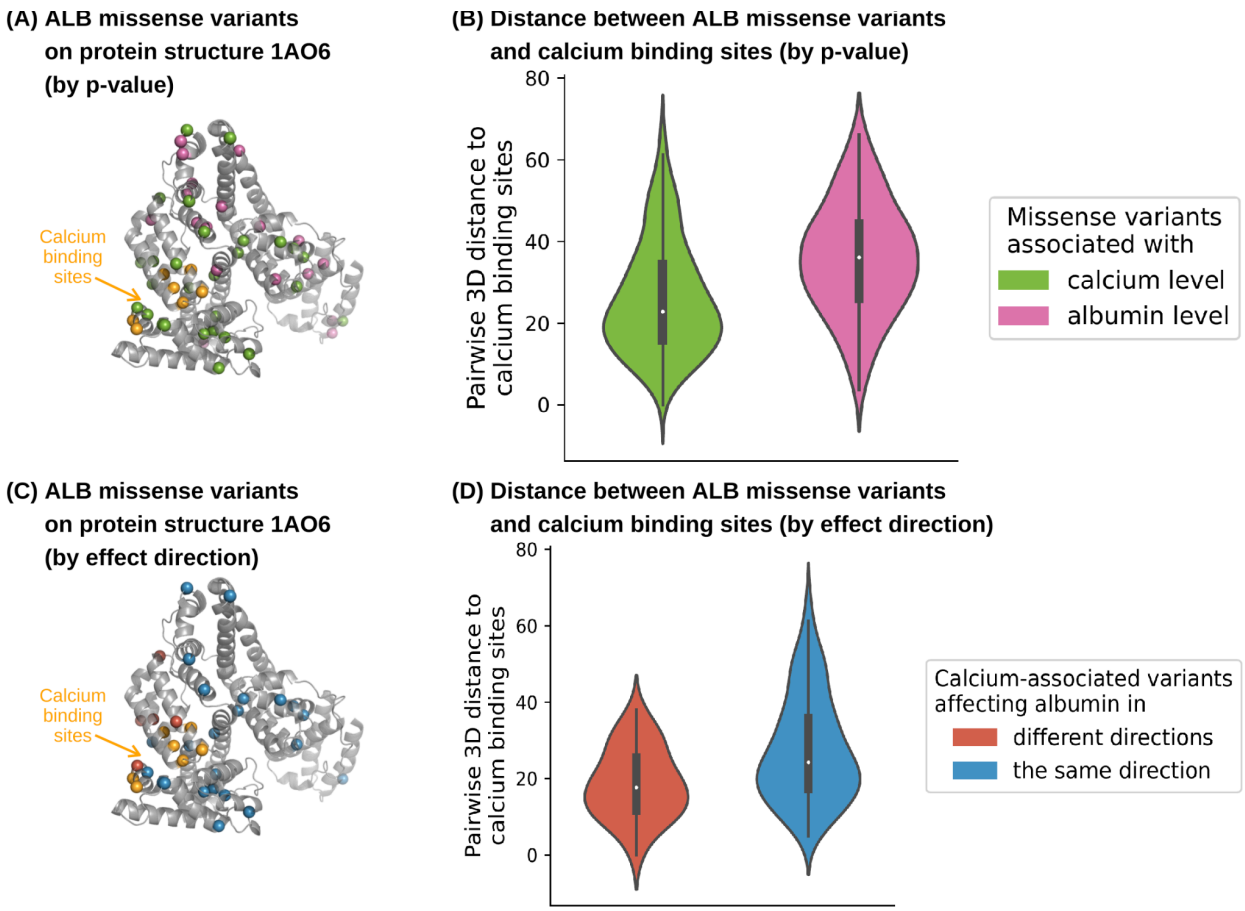


Figure 4 | Rare protein-coding variants in *ALB* (AF < 0.0001) on ALB protein structure. **A,** Distribution of *ALB* missense variants associated with calcium level but not albumin level (pink), variants associated with albumin but not calcium level (green), and known calcium binding sites (orange) on the 3D structure of ALB protein (PDB ID: 1AO6). **B,** Comparison of pairwise distances from missense variants only associated with calcium (pink) or albumin-level (green) to calcium binding sites on ALB protein structure in the 3D space. Colors in panel **B** are defined the same as in panel **A. C,** Distribution of *ALB* missense variants associated with either albumin or calcium level that affect the two biomarker levels in the same (blue) or different (red) direction, 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 the ALB protein structure in the 3D space. Colors in panel **D** are defined the same as in panel **C.**

Discussion

Here, we systematically analyzed rare variant association results across 599 curated phenotypes from Genebass to quantify gene-level pleiotropy from rare protein-coding variants. We observe pervasive pleiotropy across human exomes, highlighting genes with cross-domain

associations and evidence of allelic series. To further examine the structure of these pleiotropic signals, we developed ALLSPICE, a statistical test that detects heterogeneous variant-level effects across traits within pleiotropic genes using RVAS summary statistics. Our study provides a framework for resolving gene-level pleiotropic associations into variant-level effect patterns and demonstrates how heterogeneous effects of rare coding variants within genes can contribute to pleiotropic effects.

Compared to causal inference frameworks commonly used for cross-trait analysis (e.g. MR), ALLSPICE focuses on characterizing heterogeneous variant effects within pleiotropic genes, rather than inferring causal relationships between traits under assumptions about instrument validity and predominantly vertical pleiotropy. By identifying deviations from proportional relationships in variant effects across traits, ALLSPICE can highlight pleiotropic architectures that are likely inconsistent with standard MR assumptions, providing context for interpreting potential horizontal pleiotropy. For example, under a purely vertical pleiotropic model in which trait A causally influences trait B, a gene associated with trait B primarily through its effect on trait A would not be identified as an ALLSPICE-significant triplet. Accordingly, ALLSPICE tests a null hypothesis of a linear relationship between variant effect sizes on the two phenotypes, reflecting the expectation under vertical pleiotropy that effects on the downstream phenotype are approximately proportional to those on the corresponding upstream phenotype. Rejection of the null is therefore suggestive of horizontal pleiotropy, although additional biological and statistical validation is required.

Several conceptual and methodological limitations remain. Pleiotropy is commonly defined as genes associated with more than one trait, a definition that conflates biological pleiotropy with statistical detectability[6]. Under comparable effect size distributions, common variants are inherently more likely to be identified as pleiotropic, introducing frequency-dependent ascertainment bias. This bias is further shaped by negative selection, which constrains high-impact pleiotropic alleles to low frequencies and reduces their power for

detection[26,28,47]. Even after accounting for allele frequency, quantifying pleiotropy remains challenging without knowledge of the joint distribution of effects across traits and their collective impact on fitness. As an exception, genetic correlation provides a genome-wide measure of shared genetic architecture and can often be estimated without bias at finite sample sizes[16]. However, it captures a specific form of pleiotropy that induces consistent positive or negative covariance across the genome, which differs from the focus of this manuscript.

Systematic differences across variant classes of rare variants further exacerbate power limitations. pLoF variants often have larger effects but lower allele frequencies, whereas missense variants are more common but exhibit greater functional heterogeneity, leading to associations being detected for one class but not the other. These differences also influence how within-gene heterogeneity manifests in practice. For pLoF variants, which are often assumed to act in a largely uniform functional direction, heterogeneity detected by ALLSPICE may reflect multiple non-mutually exclusive factors, including variant misannotation, isoform-specific effects, or more complex gene-level architectures that give rise to genuine horizontal pleiotropy. A related consideration emerges from the observation of pleiotropic effects among synonymous variants, which are not generally expected to have strong biological effects. This suggests that residual confounding or tagging of common variant signals may influence burden tests, particularly at the 1% MAF threshold used in Genebass. This observation highlights the challenges of uncovering true biological pleiotropy from power- and frequency-driven artifacts in rare variant analyses, underscoring the need for larger-scale RVAS, better variant annotation methods, and improved modeling frameworks to enable more reliable interpretation of shared genetic associations.

Aside from power, phenotypic correlations can further induce apparent pleiotropy through shared genetic associations across related traits. Grouping phenotypes into broad domains provides a practical framework for systematic analysis but is an imperfect approximation of the underlying biological landscape, sacrificing phenotypic resolution for

statistical robustness and interpretability. Traits within a domain may still differ in etiology, and residual correlations across domains, such as between biomarkers and related diseases, can persist. ALLSPICE accounts for phenotypic correlation between traits within a triplet when evaluating the heterogeneous effects of variants within pleiotropic genes, but correlated phenotypes may still influence the interpretation of pleiotropic signals.

From a methodological perspective, ALLSPICE currently applies only to continuous traits, and therefore, extending the framework to binary traits and mixed trait types remains an important direction for future work. ALLSPICE also relies on simplifying assumptions about RVAS data. In particular, it assumes that there is no LD among ultra-rare variants within genes. While this assumption is approximately valid for ultra-rare variants with extremely low LD, within-gene LD can exist in real data and is known to inflate gene-based burden test statistics in proportion to the number of alleles per gene[48]. Moreover, the relationship between ALLSPICE significance and the number of contributing variants underscores the power limitations in detecting heterogeneous effects for genes with sparse variant counts. Similar considerations apply to gene-based burden tests, which form the input to ALLSPICE and are likewise less powered for genes with few rare variants. These limitations motivate the development of more flexible models that can accommodate more complex genetic architectures, particularly for genes with limited rare variants. Future work integrating complementary genetic evidence, such as eQTLs or pQTLs, may further validate and contextualize pleiotropic signals identified by ALLSPICE.

Pleiotropy provides an important perspective for organizing shared genetic associations across complex traits. By introducing ALLSPICE to identify rare-variant heterogeneity within pleiotropic genes, this work offers a systematic approach for describing pleiotropic association signals from RVAS results. More broadly, it underscores the value of studying pleiotropy and motivates future efforts to integrate additional biological resources and methods to better interpret the underlying mechanisms of shared genetic effect across traits.

Methods

Data description

We use gene-based burden test results from Genebass[23], comprising 8,074,878 single-variant tests and 75,767 gene-based burden tests across 4,529 phenotypes (1,233 quantitative and 3,296 binary), derived from whole exome sequencing data of 394,841 UK Biobank participants[49]. A total of 19,407 protein-coding genes were annotated with up to four functional groups: predicted LoF (pLoF), missense (including low-confidence pLoF variants and in-frame indels), synonymous, and non-synonymous (pLoF + missense). Variants included in burden tests were pre-filtered to those with minor allele frequency (MAF) < 0.01, and gene-based association testing was performed using SAIGE, as described in Genebass. Analyses were restricted to 58,559 high-quality gene-annotation pairs, following Genebass-recommended quality-control criteria, including a minimum of two variants per test, mean sequencing coverage ≥ 20, expected allele count (AC) ≥ 50, and genomic control inflation factor (λ_{GC}) > 0.75 for synonymous tests. Expected AC was defined as the combined allele frequency (CAF) of the gene-annotation group multiplied by the number of cases for the corresponding phenotype.

Phenotype curation

Starting from the 4,529 phenotypes available in Genebass, we focus on biological measurements and health-related outcomes by excluding phenotypes related to hospital operations (1,704), treatment/medications (435), customized phenotypes (177), and non-endpoint phenotypes (64), including injuries and age-related measures. Of the remaining 2,149 phenotypes after the initial curation, we further excluded 1,550 phenotypes with no gene associations reaching exome-wide significance ($p_{burden} < 2.5\times10^{-6}$). This filtering results in a final

set of 599 phenotypes with at least one gene-based association. To facilitate analyses of pleiotropy across largely independent phenotypes, we additionally derived a subset of 239 non-redundant phenotypes with pairwise phenotypic correlation below 0.1.

Following prior frameworks for characterizing pleiotropy among common variants[11,13], we classified the 599 phenotypes into six broad 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 (**Extended Data Figure 1A**). Disease phenotypes were further grouped into 17 ICD-10-based disease sub-domains (**Figure S6A**).

Statistical test for variant effect heterogeneity between quantitative traits

When a gene is associated with multiple phenotypes, an important question is whether the same causal variants underlie these associations. We introduce ALLSPICE, a likelihood ratio test to identify variant effect heterogeneity within pleiotropic genes, defined as imperfect correlation of variant-level effects across two traits. The likelihood ratio test explicitly accounts for uncertainty in effect size estimates arising from sampling variation, which influences the shape of the likelihood function through its dependence on both sample size and allele frequency. Under the null hypothesis of perfectly correlated effects, the likelihood-ratio statistic follows an asymptotic chi-square distribution, with the corresponding critical values reflecting the variation expected from finite-sample noise. ALLSPICE operates efficiently on summary statistics from single-variant association tests, accounts for phenotypic correlation between traits, and does not require individual-level genotype or phenotype data.

We assume that n samples are sequenced for a 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 m variants on the i_{th} phenotype. For single-variant association summary statistics on continuous phenotypes, we consider the linear regression

model $Y = \beta X + \epsilon$, where $Y = (y_1, y_2)^T$, $\beta = (\beta_1, \beta_2)^T$, where X is an $m \times n$ matrix denoting the genotypes of n samples at m variation sites, and $\epsilon_{2\times n} \sim N(0, \Delta^2)$ is 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 protein-coding variants with minor allele frequency < 0.01%, limiting the effect of linkage disequilibrium (LD), and assume that the variants are independent, as no individuals have two or more variants among 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

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

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

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

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

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

And those for the alternative hypothesis are simply the summary statistics

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

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

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

The final test statistic is computed as

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

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

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

When phenotypic correlation r between two traits approaches one, variant effect sizes on the two phenotypes become nearly identical, leading the estimated slope $\hat{c}$ to be close to one and thus the denominator of the test statistic to be close to zero (Equation 10). We therefore recommend avoiding the application of ALLSPICE to triplets involving highly positively correlated phenotypes. This is also consistent with the intended use of ALLSPICE, which is designed to highlight variants with heterogeneous effects across traits, particularly in settings compatible with horizontal pleiotropy.

Simulation design

To evaluate the performance of ALLSPICE, we conducted extensive simulations across a wide range of parameter combinations. We simulated data under 1,584 parameter configurations under the null hypothesis and 792 scenarios under the alternative hypothesis to assess statistical power.

Given the definitions of null and alternative hypotheses, the simulation differs in how effect size vectors are generated. Under the null hypothesis, the true effect sizes for the second phenotype (β_2) were sampled from a mixture distribution $\beta_2 = g_2 z_2$, $z_2 \sim N(0, \sigma^2)$ where $g_2 \sim Bernoulli(\pi)$ indicates whether an individual variant has a nonzero effect and π denotes the probability of those variants with nonzero effects. Effect sizes for the first phenotype (β_1) were then defined as $\beta_1 = c\beta_2$, enforcing perfect correlation between effects across the two phenotypes.

Under the alternative hypothesis, we consider three distinct scenarios in which the perfect linearity is violated.

(1) Independent effects: effect sizes for the first phenotypes β_1 were generated independently using the same mixture model as for the second phenotype β_2, representing completely uncorrelated effects across phenotypes.

   (a) $\beta_1 = g_1 z_1$ with $g_1 \sim Bernoulli(\pi)$ and $z_1 \sim N(0, \sigma^2)$

   (b) $\beta_2 = g_2 z_2$ with $g_2 \sim Bernoulli(\pi)$ and $z_2 \sim N(0, \sigma^2)$.

(2) Correlated but non-proportional effects: β_1 was generated to have a moderate Pearson correlation ($\rho = 0.5$) with β_2, to form a scenario different from the null hypothesis. Specifically, we first sampled two latent variables $z_1, z_2 \sim N(0, \sigma^2)$, and defined $\beta_1 = g_2(\rho z_2 + \sqrt{1 - \rho^2} z_1)$, while keeping $\beta_2 = g_2 z_2$. This construction preserves equal marginal variances while inducing correlated but not proportional effects across the two phenotypes.

(3) Non-linear effects: to model deviations from linearity, effect sizes for the first phenotype β_1 were generated as a nonlinear function of β_2, $\beta_1 = g_1(\beta_2^2 + \beta_2 + \epsilon)$, $\epsilon \sim N(0, \sigma^2)$, where $g_1 \sim Bernoulli(\pi)$ and β_2 is defined the same as previous scenarios. This scenario introduces nonlinear dependence and additional noise, representing complex genetic architectures in which variant effects on one phenotype depend nonlinearly on their effects on another phenotype.

We fixed the number of samples n at 1,000, and then varied parameters over the ranges as specified below, performing 100 simulations for each parameter combination:

- Number of variants within a gene $m \in \{5,\ 10,\ 20, 100\}$,
- Slope between variant-level effect sizes on two phenotypes

  $c \in \{0,\ 0.2,\ 0.4,\ 0.6,\ 0.8,\ 1\}$
- Phenotypic correlation between the two phenotypes

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

We simulated minor allele counts for m rare protein-coding variants from a $Uniform(1,\ 100)$ distribution, and the genotype matrix $X_{m \times n}$ as a binary $(0, 1)$ matrix, where each variant j contains nA_{jj} minor alleles randomly distributed across samples. Phenotype values for the two traits were generated from a multivariate normal distribution $N(\beta X,\ R)$, which encodes the phenotypic correlation between the traits. For each simulation, allele counts A, effect sizes β, and genotype matrix X were independently randomized, since fixing for any of these components did not substantially affect the test statistics. Summary statistics were estimated as $\hat{\beta} = \frac{1}{n} Y X^T A^{-1}$, which approximates the effect sizes under a linear regression model. Finally, for the null hypothesis, we computed the maximum likelihood estimate of c by maximizing the likelihood under the null, thereby minimizing the test statistic.

Simulation results

We applied ALLSPICE to simulated data across all combinations of parameters. Under the null hypothesis with $\sigma^2 = 1$ and $\pi = 0.5$, the distribution of $p_{ALLSPICE}$ was calibrated across all scenarios (**Extended Data Figure 2A; SuppTable 9**). We computed the empirical type I error rate as the proportion of tests rejecting the null hypothesis at the nominal significance level ($p_{ALLSPICE} < 0.05$). Across most scenarios, type I error rates were centered around 0.05. Although error rates increased slightly with the number of variants, they remained well controlled below

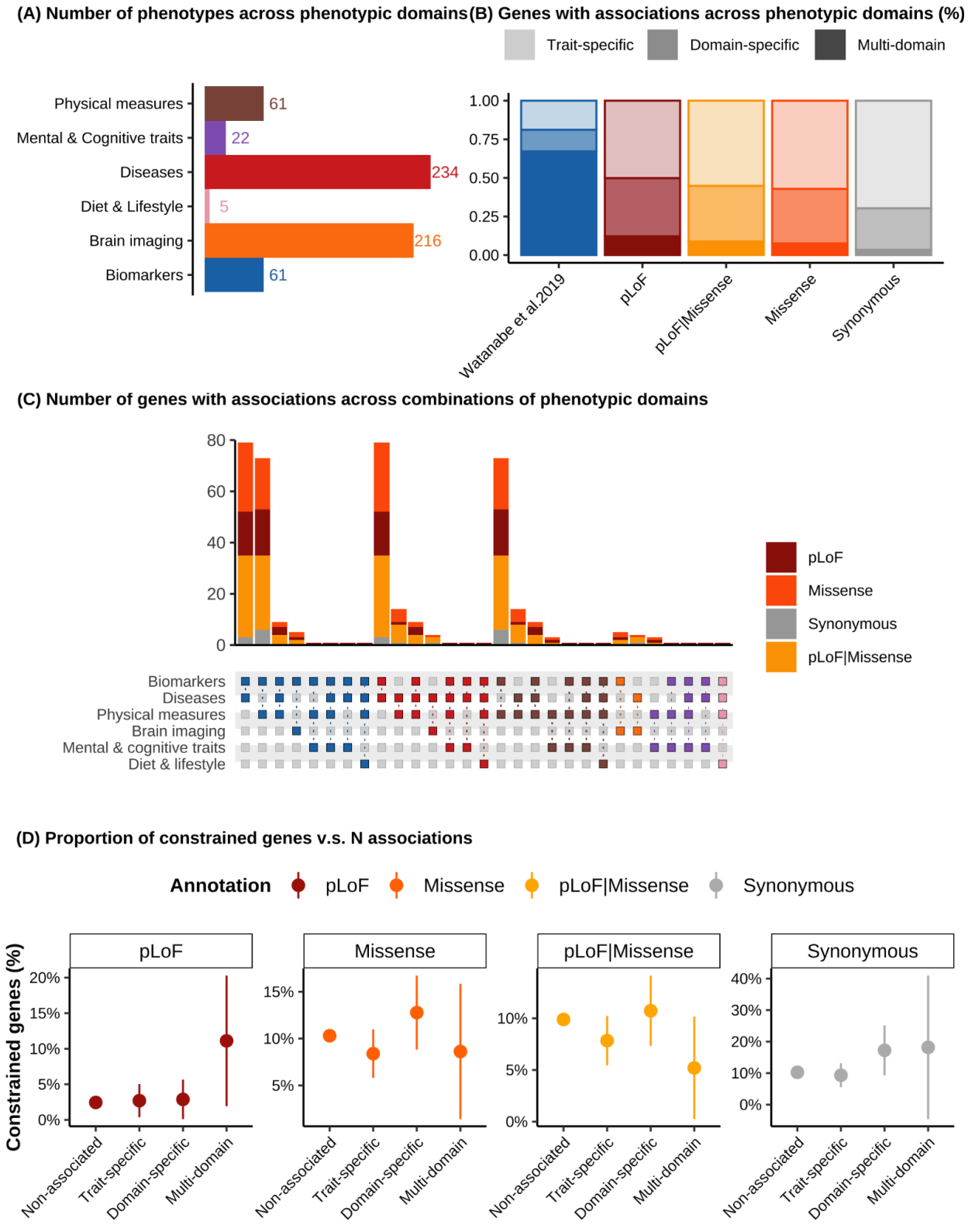
0.1 (**Extended Data Figure 2B**). When σ^2 was reduced to 0.01 or 0.1, the test exhibited mildly conservative patterns compared to $\sigma^2 = 1$, reflecting the smaller effect sizes simulated as σ^2 decreases (**Figures S7, S8; SuppTable 10, 11**). Results for higher mixture proportions ($\pi = 0.8$) were similarly well calibrated compared to $\pi = 0.5$ (**Figures S9-S11; SuppTable 9-11**).

Empirical statistical power was computed as the proportion of the tests rejecting the null hypothesis at a nominal significance level of 0.05 under the alternative. Parameter configuration is the same as previously described for simulations under the null scenario. We fixed the sample size n at 1,000 and σ^2 at 1, and performed 1,000 simulations for each parameter configuration. Statistical power increased with the number of variants included in the gene, exceeding 0.9 when at least 10 variants were present, which is consistent across three alternative scenarios (**Extended Data Figure 3; SuppTable 12**, **15**, and **18**). However, when σ^2 was reduced to 0.01 or 0.1, the overall scale of effect sizes for both $\beta_1$and β_2 decreased, thus leading to a reduction in power to detect significant triplets, particularly for the scenarios with independent effect sizes (**Figure S12 and Figure S13; SuppTable 13**, **14**, **16**, **17**, **19**, and **20**).

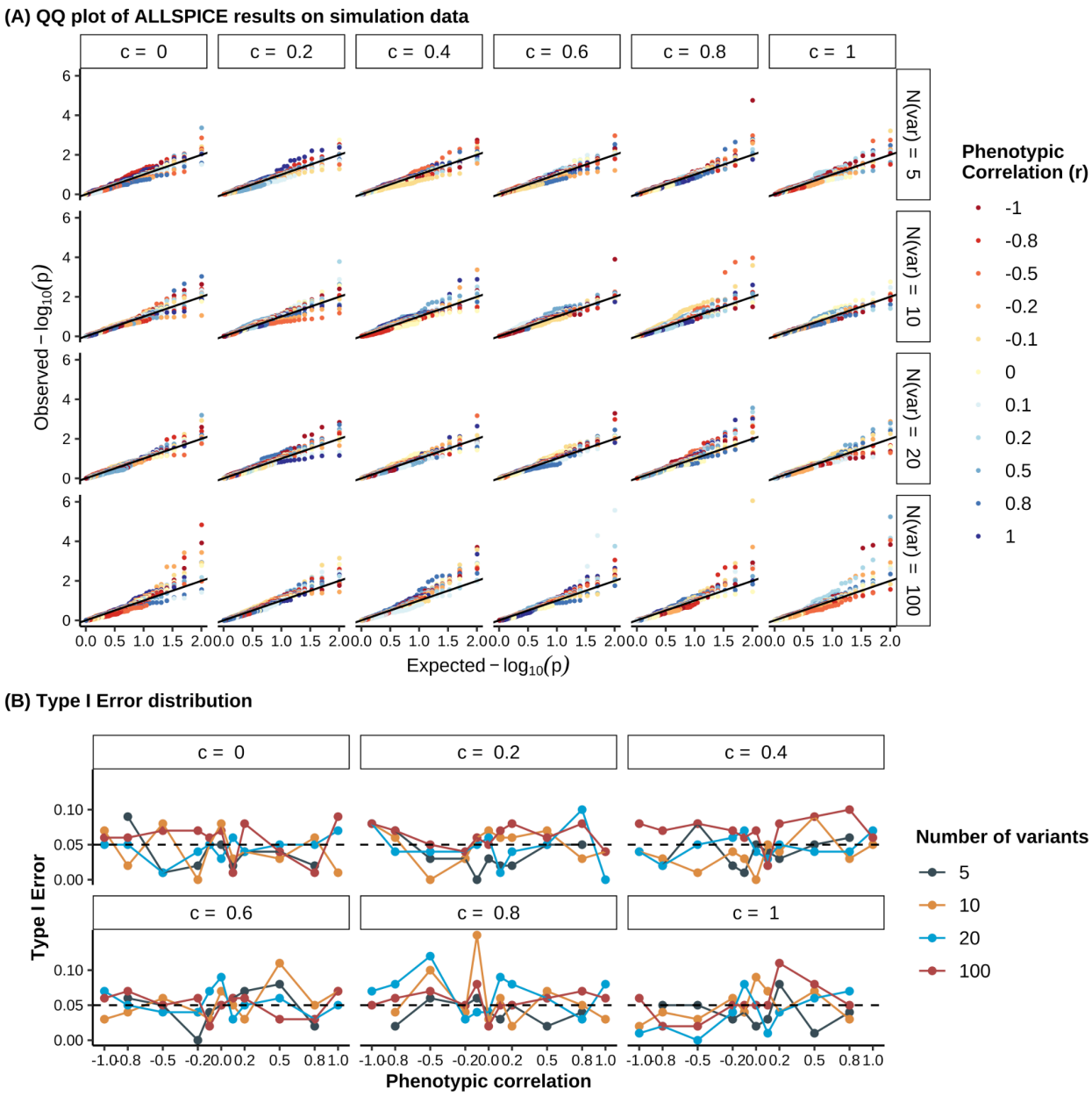
Protein structure-based analysis

Pairwise distances between residuals on the 3D protein structures were calculated as the Euclidean distances between the alpha carbons of the corresponding residues, using their atomic coordinates from the Protein Data Bank (PDB). p-values ($p_{protein}$) indicating the differences in distance distributions between groups were computed through a two-tailed Mann-Whitney U test (**Figure 4**, **Figure S16**, and **Figure S18**).

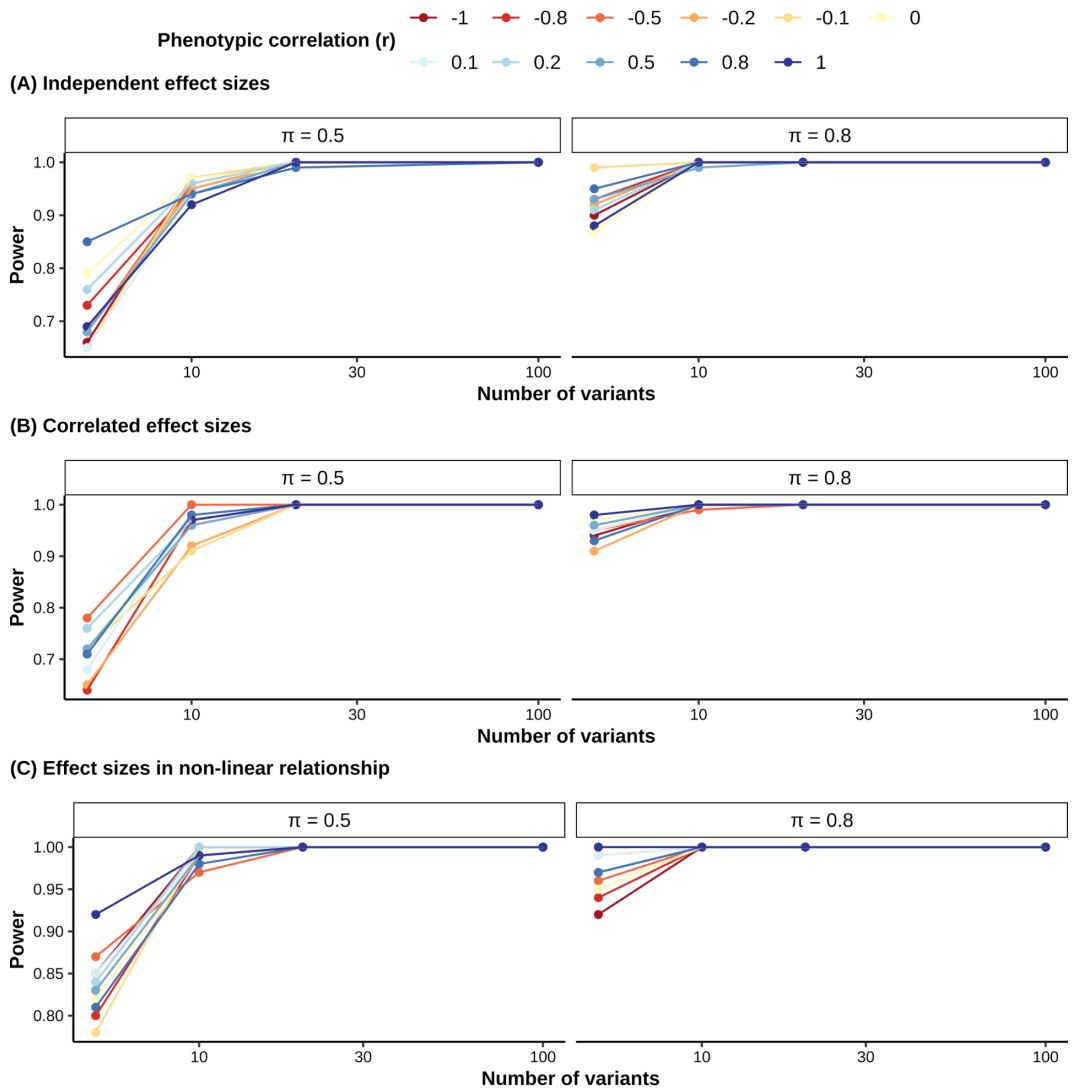
Extended Data Figures



Extended Data Figure 1 | Pervasive gene-level pleiotropy across six broader phenotypic domains: **A,** Number of phenotypes in each of the six phenotypic domains (color-coded); **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 functional annotation groups (color-coded). The blue bar shows the common variant results from Watanabe et al. 2019[11]; **C,** Number of genes associated with each combination of phenotypic domains. The bars indicate the number of genes with cross-domain associations, as shown in the bottom matrix, where each column represents a domain combination that shares at least one gene-level association. Colors are consistent between panel **A** and the bottom matrix in panel **C**. **D,** Proportion of constrained genes, defined as those in the lowest LOEUF decile (y-axis), with no association, trait-specific, domain-specific, or multi-domain associations (x-axis) across functional annotation groups (color and column). Error bars represent 95% confidence intervals.



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 the variance $\sigma^2 = 1$ and the mixture component weight $\pi = 0.5$. **A,** QQ plot of test results from the simulated null data described above. The solid black line denotes the line y=x. Rows correspond to the number of variants included, and the columns indicate the true value of the slope parameter c relating the two effect size vectors β_1, β_2 used in the simulation. **B,** Distribution of type I error rates (y-axis) 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 empirical statistical power of ALLSPICE test under the alternative hypotheses ($\sigma^2 = 1$). Empirical power (y-axis) was evaluated in simulated data across phenotypic correlation (color), number of variants (x-axis), and mixture component weights π (columns), indicating the probability that a variant effect is drawn from the non-zero component, modeled as a standard normal distribution $N(0, \sigma^2)$, where $\sigma^2 = 1$. Results are shown for three alternative scenarios: **A,** independent effect sizes across phenotypes; **B**, correlated but non-proportional effect sizes; **C**, non-linear relationship between effect sizes.

Resources

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

Acknowledgements

Grant from the Swiss National Science Foundation to C.A. (P500-3_235131)
D.P. is supported by the Netherlands Organization for Scientific Research - Gravitation project 'BRAINSCAPES: A Roadmap from Neurogenetics to Neurobiology' (024.004.012)

Declaration of interests

B.M.N is a member of the scientific advisory board at Deep Genomics. K.J.K. is a member of the scientific advisory board of Nurture Genomics. All other authors have no competing interests to declare.

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. Woodward, D. J. *et al.* Leveraging pleiotropy for the improved treatment of psychiatric disorders. *Mol Psychiatry* **30**, 705–721 (2025).
3. Bao, C. *et al.* A cross-disease, pleiotropy-driven approach for therapeutic target prioritization and evaluation. *Cell Rep Methods* **4**, 100757 (2024).
4. Paaby, A. B. & Rockman, M. V. The many faces of pleiotropy. *Trends in genetics : TIG* **29**, 66 (2012).
5. Stearns, F. W. One Hundred Years of Pleiotropy: A Retrospective. *Genetics* **186**, 767 (2010).
6. 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).
7. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet* **23**, R89–98 (2014).
8. O'Connor, L. J. & Price, A. L. Distinguishing genetic correlation from causation across 52 diseases and complex traits. *Nat. Genet.* **50**, 1728–1734 (2018).
9. Valverde, P., Healy, E., Jackson, I., Rees, J. L. & Thody, A. J. Variants of the

melanocyte–stimulating hormone receptor gene are associated with red hair and fair skin in humans. *Nature Genetics* **11**, 328–330 (1995).

10. 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).
11. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).
12. Sivakumaran, S. *et al.* Abundant pleiotropy in human complex diseases and traits. *Am. J. Hum. Genet.* **89**, 607–618 (2011).
13. Levin, M. G. *et al.* Genome-Wide Assessment of Pleiotropy Across >1000 Traits from Global Biobanks. *medRxiv* (2025) doi:10.1101/2025.04.18.25326074.
14. Ballard, J. L. & O'Connor, L. J. Shared components of heritability across genetically correlated traits. *Am. J. Hum. Genet.* **109**, 989–1006 (2022).
15. 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).
16. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
17. Liang, X., Sha, Q., Rho, Y. & Zhang, S. A hierarchical clustering method for dimension reduction in joint analysis of multiple phenotypes. *Genet Epidemiol* **42**, 344–353 (2018).
18. Lazarev, D., Chau, G., Bloemendal, A., Churchhouse, C. & Neale, B. M. GUIDE deconstructs genetic architectures using association studies. *bioRxiv* (2025) doi:10.1101/2024.05.03.592285.
19. Sinnott-Armstrong, N. *et al.* Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat. Genet.* **53**, 185–194 (2021).
20. Chami, N. *et al.* Exome Genotyping Identifies Pleiotropic Variants Associated with Red

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

21. Collaborative, E. Exome sequencing of 20,979 individuals with epilepsy reveals shared and distinct ultra-rare genetic risk across disorder subtypes. *Nature Neuroscience* **27**, 1864–1879 (2024).
22. Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).
23. 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).
24. Weiner, D. J. *et al.* Polygenic architecture of rare coding variation across 400,000 exomes. *bioRxiv* (2022) doi:10.1101/2022.07.06.22277335.
25. 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).
26. Zeng, J. *et al.* Signatures of negative selection in the genetic architecture of human complex traits. *Nat Genet* **50**, 746–753 (2018).
27. Gazal, S. *et al.* Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. *Nature Genetics* **50**, 1600–1607 (2018).
28. O'Connor, L. J. *et al.* Extreme Polygenicity of Complex Traits Is Explained by Negative Selection. *Am J Hum Genet* **105**, 456–476 (2019).
29. Rehm, H. L. *et al.* ClinGen--the Clinical Genome Resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015).
30. McCaw, Z. R. *et al.* A scalable framework for identifying allelic series from summary statistics. *Am J Hum Genet* **112**, 2772–2788 (2025).
31. Cohen, J. C., Boerwinkle, E., Mosley, T. H., Jr & Hobbs, H. H. Sequence Variations in PCSK9, Low LDL, and Protection against Coronary Heart Disease. (2006)

doi:10.1056/NEJMoa054013.

32. Lu, Y. *et al.* Drug-induced agranulocytosis: a disproportionality analysis and umbrella review. *Front Pharmacol* **16**, 1641747 (2025).
33. Mohamed Jiffry, M. Z. *et al.* Agranulocytosis Secondary to Cancer Chemotherapy Associated With Higher In-Hospital Mortality in Patients With Central Line Insertion During a Hospital Stay. *Cureus* **15**, e34717 (2023).
34. Lin, H. *et al.* The mean reticulocyte volume is a valuable index in early diagnosis of cancer-related anemia. *PeerJ* **12**, e17063 (2024).
35. Zöller, B., Manderstedt, E., Lind-Halldén, C. & Halldén, C. Rare-variant collapsing and bioinformatic analyses for different types of cardiac arrhythmias in the UK Biobank reveal novel susceptibility loci and candidate amyloid-forming proteins. *Cardiovasc Digit Health J* **5**, 15–18 (2024).
36. Huang, Y. *et al.* Rare genetic variants impact muscle strength. *Nat. Commun.* **14**, 3449 (2023).
37. Chen, C.-Y. *et al.* The impact of rare protein coding genetic variation on adult cognitive function. *bioRxiv* (2022) doi:10.1101/2022.06.24.22276728.
38. Deaton, A. M. *et al.* Gene-level analysis of rare variants in 379,066 whole exome sequences identifies an association of GIGYF1 loss of function with type 2 diabetes. *Sci. Rep.* **11**, 21565 (2021).
39. Bellono, N. W., Escobar, I. E., Lefkovith, A. J., Marks, M. S. & Oancea, E. An intracellular anion channel critical for pigmentation. *Elife* **3**, e04543 (2014).
40. Martinon, F. & Aksentijevich, I. New players driving inflammation in monogenic autoinflammatory diseases. *Nat. Rev. Rheumatol.* **11**, 11–20 (2015).
41. Martin, T. M. & Rosenbaum, J. T. An update on the genetics of HLA-B27 associated acute anterior uveitis. *Ocular immunology and inflammation* **19**, 108 (2011).
42. 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).

43. Peloso, G. M. *et al.* Rare Protein-Truncating Variants in APOB, Lower Low-Density Lipoprotein Cholesterol, and Protection Against Coronary Heart Disease. *Circ Genom Precis Med* **12**, e002376 (2019).
44. Guarnera, L. & Jha, B. K. TET2 mutation as prototypic clonal hematopoiesis lesion. *Semin Hematol* **61**, 51–60 (2024).
45. Desgagnés, N., King, J. A., Kline, G. A., Seiden-Long, I. & Leung, A. A. Use of Albumin-Adjusted Calcium Measurements in Clinical Practice. *JAMA Netw Open* **8**, e2455251–e2455251 (2025).
46. Besarab, A., DeGuzman, A. & Swanson, J. W. Effect of albumin and free calcium concentrations on calcium binding in vitro. *J Clin Pathol* **34**, 1361–1367 (1981).
47. Novo, I., López-Cortegano, E. & Caballero, A. Highly pleiotropic variants of human traits are enriched in genomic regions with strong background selection. *Hum Genet* **140**, 1343–1351 (2021).
48. Weiner, D. J. *et al.* Polygenic architecture of rare coding variation across 394,783 exomes. *Nature* **614**, 492–499 (2023).
49. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).