Bayesian estimation of shared polygenicity identifies drug targets and repurposable medicines for human complex diseases

Noah Lorincz-Comi[1,2], Feixiong Cheng[1,2,3,*]

[1]Cleveland Clinic Genome Center, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44195, USA

[2]Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44195, USA

[3]Department of Molecular Medicine, Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland, OH 44195, USA

*Correspondence to Feixiong Cheng, Ph.D.
Lerner Research Institute, Cleveland Clinic
Tel: +1-216-444-7654; Fax: +1-216-636-0009
Email: chengf@ccf.org

Keywords: Polygenicity, pleiotropy, druggable genes, complex disease, genome-wide association study, gene association testing

Abstract

Background: Complex diseases may share portions of their polygenic architectures which can be leveraged to identify drug targets with low off-target potential or repurposable candidates. However, the literature lacks methods which can make these inferences at scale using publicly available data.

***Methods*:** We introduce a Bayesian model to estimate the polygenic structure of a trait using only gene-based association test statistics from GWAS summary data and returns gene-level posterior risk probabilities (PRPs). PRPs were used to infer shared polygenicity between 496 trait pairs and we introduce measures that can prioritize drug targets with low off-target effects or drug repurposing potential.

***Results*:** Across 32 traits, we estimated that 69.5 to 97.5% of disease-associated genes are shared between multiple traits, and the estimated number of druggable genes that were only associated with a single disease ranged from 1 (multiple sclerosis) to 59 (schizophrenia). Estimating the shared genetic architecture of ALS with all other traits identified the *KIT* gene as a potentially harmful drug target because of its deleterious association with triglycerides, but also identified *TBK1* and *SCN11B* as putatively safer because of their non-association with any of the other 31 traits. We additionally found 21 genes which are candidate repourposable targets for Alzheimer's disease (AD) (e.g., *PLEKHA1*, *PPIB*) and 5 for ALS (e.g., *GAK*, *DGKQ*).

Conclusions: The sets of candidate drug targets which have limited off-target potential are generally smaller compared to the sets of pleiotropic and putatively repurposable drug targets, but both represent promising directions for future experimental studies.

Keywords: polygenic, pleiotropy, complex disease, druggable genes, drug repurposing

Introduction

Understanding the extent to which the heritability of a complex trait is conferred by few or many genes, i.e., its 'polygenicity', is a first step in understanding its genetic etiology (Visscher et al., 2021). Similarly, quantifying the degree to which two distinct traits have heritability which is contributed by shared and non-shared genes helps researchers better contextualize phenotypic similarities between them (Jiang et al., 2019; Ballard & O'Connor, 2022). This information can even be used to guide drug targeting decisions (Lounkine et al., 2012), namely how the targeting of one gene with a drug may reduce the risk and/or symptoms of a target disease while simultaneously inducing side effects because the gene is also associated with other traits (Bedi et al., 2016; Nguyen et al., 2019; Woodward et al., 2024). On the other hand, genes associated with known disease risk factors can also be candidate drug targets (Mao et al., 2024), a notable example of which is the targeting of low-density lipoprotein by inhibiting *PCSK9* expression to reduce coronary artery disease risk (Horton et al., 2007; Cohen et al., 2006). The first step in inferring shared gene associations between multiple traits is generally to separately test each gene for association with each trait and to make a joint inference for the pair.

Inference of a shared gene association with multiple traits can be accomplished using standard hypothesis testing with gene-based test statistics from genome-wide association studies (GWAS) (Lorincz-Comi et al., 2024a; Liu et al., 2010; de Leeuw et al., 2015) and a joint inference using the intersection-union test at the SNP (Solovieff et al., 2013) or gene level (Sivakumaran et al., 2011), but this is highly sensitive to GWAS

sample size, inherently limits the available interpretation only to nonzero disease association, and cannot be used to reliably characterize polygenicity genome-wide. Researchers cannot make inferences about non-association at the gene level using hypothesis testing, which is required for identifying genes whose drugs have limited off-target potential. Hypothesis testing also relies almost entirely on the statistical power available to detect disease-associated SNPs in GWAS, which is primarily driven by the polygenic structure itself and GWAS sample size (Bulik-Sullivan et al, 2015; Wu et al., 2022). For example, counts of rejected independent gene-based null hypotheses estimate the number of associated genes that exist for a single trait, but this estimate is biased in finite GWAS samples. Similarly, inference of no association between a gene and trait can only be validly made from a non-rejected null hypothesis when GWAS sample size is infinitely large.

These limitations make polygenic inference using hypothesis testing alone potentially unreliable and unflexible (Schaid et al., 2016). Instead, directly modelling the polygenic structure of traits and leveraging statistical methods which can supply posterior risk probabilities for each trait in a pair under the presumed model may provide more consistent estimates of shared polygenicity and make a wider range of inferences available to researchers. However, currently available methods do not completely account for linkage disequilibrium correlation between their measured units (i.e., SNPs, genes) and can only provide polygenic inferences at the genome-wide or SNP levels (Frei et al., 2019; Parker et al., 2024; Akdeniz et al., 2024; Frei et al., 2024), which in many applications are of at least secondary interest to the gene level.

We present a general statistical model which computes the posterior probability that each gene contributes to additive heritability using gene-based test statistics and an empirically derived model of polygenicity. This approach is leveraged to estimate the total number of genes which are putatively associated with each of 32 complex traits and the proportions of genes contributing shared and non-shared additive heritability across them. Our results suggest large variability across traits in the estimated number of associated genes, from less than 20 to over 900, and that the vast majority of disease associated genes are associated with multiple traits. These results highlight the role of prioritizing trait-exclusive risk genes as putatively drug target candidates with less off-target effects, and we show as an example that targeting the *KIT* gene for ALS prevention may reduce ALS risk while simultaneously increasing triglyceride levels, which could harm overall health. On the other hand, genes with evidence of pleiotropy across multiple traits can be leveraged if their associations are in the direction of risk, and we provide two demonstrative examples for AD with *PLEKHA1* and ALS with *GAK*.

Methods

Overview of methods

Our method requires only gene-based association test statistics, their null and alternative distributions, and pairwise correlations between gene-based test statistics under their null hypotheses. Because of its generality and widespread availability and use in the literature (Lu et al., 2010; Vsevolozhskaya et al., 2020; Lorincz-Comi et al., 2024a) we assume gene-based test statistics are calculated as the sums of SNP-level

chi-square statistics used to test the null hypothesis of no association between the SNP and phenotype (Lu et al., 2010; Vsevolozhskaya et al., 2020; Lorincz-Comi et al., 2024a), denoted here as T_k for the kth gene. SNPs used in these tests are generally in linkage disequilibrium (LD) with each other and proximal to the transcription start site of the gene (Vsevolozhskaya et al., 2020). Gene-based test statistics can be calculated from publicly available GWAS summary statistics and an accompanying population-matched LD reference panel, and a repository of the necessary data required for our method is available for over 50 traits from Lorincz-Comi. et al. (2024a). We assume each gene-based test statistic in the genome-wide set $\{T_k\}$ is drawn from a mixture of null and non-null distributions parameterized by the SNP heritability of the trait, GWAS sample size, linkage disequilibrium scores (LD; Bulik-Sullivan et al., 2015) and SNP minor allele frequencies, and SNP LD structure. We introduce a general method which can be used to estimate the posterior probability that a gene contributes to the additive heritability of a trait not explained by dominance, epistasis, or gene-environment interactions.

SNP-level model

Our gene-level model is built from a SNP-level random effect model for a single trait. Let β_j^k represent the marginal association between the jth SNP of the kth gene and a trait, $r_{jj'}$ the LD correlation between SNPs j and j' corresponding to the gene, ℓ_j the LD score of the jth SNP (i.e., sum of squared LD correlations with surrounding SNPs in a window of fixed size; Bulik-Sullivan et al., 2015), a_j the corresponding minor allele frequency, N the GWAS sample size, and ϕ_h^2 the average SNP heritability explained by

each of $\ddot{M}$ causal SNPs genome-wide. Let G_{ij} represent the dosage of risk alleles of the jth SNP from the ith person in the GWAS and $\sigma_j^2 \coloneqq \mathrm{E}\left(G_{ij}^2\right) = 2\alpha_j\left(1+\alpha_j\right)$. We use the following SNP-level models:

$$\beta_j^k \sim N\left(0, \phi_h^2 \ell_j\right), \qquad \mathrm{Cov}\left(\beta_j^k, \beta_{j'}^k\right) = 0$$

$$u_j \coloneqq \left(\hat{\beta}_j^k - \beta_j^k\right) \xrightarrow{D} N\left(0, \frac{\phi_Y^2}{N\sigma_j^2}\right), \qquad \mathrm{Corr}\left(u_j, u_{j'}\right) = r_{jj'}$$

$$\hat{\beta}_j^{\mathrm{k}} = \beta_j^k + u_j,$$

$$\sqrt{N}\sigma_j \phi_Y^{-1} \hat{\beta}_j^k \xrightarrow{D} N\left(0, \tau \ell_j N \sigma_j^2 + 1\right), \qquad \tau \coloneqq \phi_h^2 \phi_Y^{-2}$$

where ϕ_Y^2 is the scaled conditional variance of the phenotype given the genotype for the jth SNP. For binary phenotypes, ϕ_Y^2 acts also as a scaling factor which is proportional to the trait prevalence that in effect indicates a transformation from the observed binary scale to the underlying liability scale (Wu & Sham, 2021). It follows from the definition of τ that $\tau \coloneqq h^2 \ddot{M}^{-1} \phi_Y^{-2}$ is likely to be very small in practice when $\ddot{M}$ is large and/or ϕ_Y^{-2} is large and SNP heritability h^2 is small. Let $\widehat{\boldsymbol{\beta}}_k = \left(\hat{\beta}_j^k\right)$ be the vector of estimated SNP effect sizes for all SNPs in the set corresponding to the kth gene. The correlation matrix of $\widehat{\boldsymbol{\beta}}_k$ is approximately the LD matrix $\mathbf{R}_k$ under this model (Zhu et al., 2017).

Gene-level model

The SNP-level model is used to construct a gene-level model for the distribution of gene-based test statistics $\{T_k\}$. Let the kth of M genes across the genome be tested using the statistic T_k calculated from m_k SNPs in the set $\mathcal{S}_k$, i.e.

$$T_k = \sum_{j \in \mathcal{S}_k} N \widehat{\sigma_j}^2 \hat{\phi}_Y^{-2} \left(\hat{\beta}_j^k\right)^2 \qquad (1)$$

Lorincz-Comi et al. (2024a) showed that under $H_{0k}\colon \mathrm{E}(T_k) = m_k$, $T_k \sim \Gamma(\alpha_{0k}, \xi_{0k})$ approximately and under $H_{1k}\colon \mathrm{E}(T_k) > m_k$, $T_k \sim \Gamma(\alpha_{1k}, \xi_{1k})$ approximately, using the shape-rate parameterizations of each Gamma distribution. Let $\mathbf{z}_k = \left(\sqrt{N}\hat{\sigma}_j\hat{\phi}_Y^{-1}\hat{\beta}_j^k\right)$, $\mathbf{L}_k = \mathrm{diag}\left(\ell_j\right)$ and $\mathbf{D}_k = \mathrm{diag}\left(\sigma_j^2\right)$ for $j \in \mathcal{S}_k$, and $\mathbf{H}_k \coloneqq \mathbf{D}_k^{1/2}\mathbf{L}_k\mathbf{D}_k^{1/2}$ such that $\mathrm{Cov}(\mathbf{z}_k|H_{1k}) = N\tau\mathbf{H}_k + \mathbf{R}_k$ and $\mathrm{Cov}(\mathbf{z}_k|H_{0k}) = \mathbf{R}_k$. It is shown in the **Supplement** that

$$\xi_{1k} = \frac{m_k + N\tau\sum_{j=1}^{m_k}\sigma_j^2\ell_j}{2\mathrm{tr}(\mathbf{R}_k\mathbf{R}_k) + 2(N\tau)^2\sum_{j=1}^{m_k}\sigma_j^4\ell_j^2 + 4N\tau\sum_{j=1}^{m_k}\sigma_j^2\ell_j} \tag{2}$$

and

$$\alpha_{1k} = \left(m_k + N\tau\sum_{j=1}^{m_k}\sigma_j^2\ell_j\right)\xi_{1k}. \tag{3}$$

It follows immediately that $\xi_{0k} = m_k\mathrm{tr}(2\mathbf{R}_k\mathbf{R}_k)^{-1}$ and $\alpha_{0k} = m_k\xi_{0k}$. We use these quantities to model the data-generating process of the statistic T_k as

$$T_k \sim \delta\Gamma(\alpha_{0k}, \xi_{0k}) + (1-\delta)\Gamma(\alpha_{1k}, \xi_{1k}) \tag{4}$$

where $1-\delta$ is the marginal probability that gene k is causally associated with the phenotype under this model. This mixture can alternatively be expressed as

$$T_k = (1-I_k)\Gamma(\alpha_{0k}, \xi_{0k}) + I_k\Gamma(\alpha_{1k}, \xi_{1k}) \tag{5}$$

where $\mathrm{P}(I_k = 1) = 1-\delta$, $\mathrm{P}(I_k = 0) = \delta$, and I_k is a latent indicator of causality for the kth gene.

Estimating the number of disease associated genes and their shared counts

We present a Bayesian method to infer the polygenic architectures of complex traits (BPACT) and the sharing of architectures between pairs of them. In this model, we

calculate the posterior probability that the kth gene is associated with risk of the tth trait (PRP_{kt}) as

$$\mathrm{PRP}_{kt} = \int_{\delta_t \in \Delta_t} \int_{\tau_t \in \mathrm{T}_t} P(I_k^t | T_k^t, \delta^t, \tau^t) p_{\delta_t}(\delta_t | T_k^t) p_{\tau_t}(\tau_t | T_k^t) d\delta_t d\tau_t \quad (6)$$

$$= E(I_k^t | T_k^t)$$

where $p_{\delta_0^t}(\delta^t | T_k^t)$ and $p_{\tau_0^t}(\tau^t | T_k^t)$ are posterior distributions of δ^t and τ^t, respectively, which are assumed conditionally independent given T_k^t, and

$$P(I_k^t | T_k^t, \delta^t, \tau^t) = \frac{f_1^t(T_k^t; \tau^t)(1-\delta^t)}{f_1^t(T_k^t; \tau^t)(1-\delta^t) + f_0^t(T_k^t; 0)\delta^t} \ , \quad (7)$$

where $f_1^t(\cdot\,;\, v_1^t)$ and $f_0^t(\cdot\,;\, v_0^t)$ respectively are the non-null and null distributions of T_k^t parameterized by $\tau^t \leftarrow v_1^t$ and $\tau^t \leftarrow v_0^t$. The quantity PRP_{kt} is interpreted as the posterior probability that gene k causes trait t given the statistic T_k^t, δ^t, τ^t, and their distributions $p_{\delta_0^t}$ and $p_{\tau_0^t}$ in the spaces Δ_t and T_t, which can be interpreted as adjustments for the estimation error of $\left(\hat{\delta}_k^t, \hat{\tau}_k^t\right)$. We estimate the posterior distributions $p_{\delta_0^t}(\delta^t | T_k^t)$ and $p_{\tau_0^t}(\tau^t | T_k^t)$ by sampling from their distributions computationally using the Metropolis-Hastings algorithm described in the **Supplement** and thereafter model them with Beta and Trapezoidal distributions, respectively.

We infer the shared polygenic architecture for two traits t and t' using $\{\mathrm{PRP}_{kt}\}$ and $\{\mathrm{PRP}_{kt'}\}$ while correcting for participant overlap between the GWAS from which $\{T_k^t\}$ and $\left\{T_k^{t'}\right\}$ were calculated (see the next subsection). Inferences of genes with potentially low off-target effects are those for which

$$\mathrm{PRP}_{kt} \times \prod_{\{s:s\neq t\}} (1 - \mathrm{PRP}_{ks}) \tag{8}$$

is high, and inferences of potentially repurposable genes for the tth trait are made for those for which

$$\mathrm{PRP}_{kt} \times \sum_{\{s:s\neq t\}} \mathrm{PRP}_{ks} r_g(t, s; k) d_s \tag{9}$$

is high, where $r_g(t, s; k)d_s$ is the re-signed genetic correlation between traits t and s in the local region of the kth gene, i.e., $d_s \in \{1, -1\}$. We use the risk-directed genetic correlation $r_g(t, s; k)d_s$ to infer that the therapeutic targeting of the kth gene for the tth trait may be accomplished by targeting the sth trait. Although PRP_{kt} is a conditional probability under a model representing statistical causality, we describe our practical epidemiological inferences in specific disease contexts just to disease risk association and not causality, which could be better assessed using experimental approaches.

We estimate the total number of disease associated genes for trait t using S randomly selected, mutually weakly correlated, and chromosome-specific sets of genes, denoted as $\mathcal{C}_r(s)$ for chromosome $r = 1, \ldots, R$ at the sth iteration, and pooling results across the S replicates and R chromosomes to produce the estimate:

$$c(t) = \frac{1}{S} \sum_{s=1}^{S} \sum_{r=1}^{R} \sum_{k \in \mathcal{C}_r(s)} \mathrm{PRP}_{kt} \,. \tag{10}$$

We can estimate the number of shared associated genes between traits t and t' as

$$c(t, t') = \frac{1}{S} \sum_{s=1}^{S} \sum_{r=1}^{R} \sum_{k \in \mathcal{C}_r(s)} \mathrm{PRP}_{kt} \mathrm{PRP}_{kt'} \,, \tag{11}$$

which assumes independence between T_k^t and $T_k^{t'}$ for all k. Independent sets of genes from each chromosome at the sth iteration are constructed such that the correlations between all pairs of gene-based test statistics in the set $\mathcal{C}_r(s)$ are below a fixed level, which in practice we set as $\sqrt{0.5}$. Since many of such sets may exist, we first define approximately independent blocks of genes using the $\sqrt{0.5}$ threshold for $\mathrm{Corr}(T_k, T_{k'})$ and the correlation block-sorting method of Prive (2022) in the `bigsnpr` R package. For each block and chromosome at each iteration, we randomly select one gene and add it to the set $\mathcal{C}_r(s)$. Across the 16,324 genes which were tested for association with all 32 traits in our real data analysis, 8,148 independent blocks of gene-based test statistics were present. The inferences of total, $c(t)$, and shared, $c(t, t')$, association are therefore based on composite posterior densities fitted multiple times for S iterations, over which the results are pooled to produce the final estimates. We calculate standard errors for estimated quantities using the imputation principles of Rubin (1996) across the S iterations. We estimated the number of genes which are associated with any trait by first subtracting from 1 the product of all trait-specific posterior risk probabilities and summing them across all independent genes. When then divided this quantity by the estimated number of independent gene blocks, 8,148, and multiplied it by 20K.

Effect of sample overlap on estimated shared gene counts

In the previous subsection, we introduced $c(t, t')$ which assumed T_k^t was uncorrelated with $T_k^{t'}$ for the kth gene and tth, t'th phenotypes. We show in this subsection that overlapping subjects between the GWAS in which T_k^t and $T_k^{t'}$ were calculated can induce nonzero spurious correlation between T_k^t and $T_k^{t'}$, and how we can correct for it.

For generality and to illustrate the motivation for considering independent blocks of genes in our polygenicity estimation, we show the correlation between gene-based test statistics in the case of different genes and for two traits from different GWAS cohorts which may contain overlapping subjects. Let $\mathbf{R}_{kk'}$ represent the matrix of LD correlations between SNPs in the gene-specific sets for genes k and k', $\boldsymbol{\Upsilon}_{tt'} = (\upsilon_{tt'})$, and $\upsilon_{tt'} \approx N_{tt'}(N_t N_{t'})^{-1/2}\mathrm{Corr}(t,t')$ (LeBlanc et al., 2018), which represents the approximate correlation between T_h^t and $T_h^{t'}$ for any gene index h using GWAS of continuous traits t and t' of sizes N_t and $N_{t'}$ containing $N_{tt'}$ overlapping subjects. For binary traits, $\upsilon_{tt'}$ will have a slightly different expression but the general principle is the same (LeBlanc et al., 2018). It follows that

$$\tilde{\mathbf{z}}(k,k';t,t') \coloneqq \begin{pmatrix} \mathbf{z}_k^t \\ \mathbf{z}_{k'}^{t'} \end{pmatrix} \sim N\left(\mathbf{0}, \mathbf{G}_{kk'}^{tt'} \coloneqq \begin{bmatrix} \mathbf{R}_{kk} & \upsilon_{tt'}\mathbf{R}_{kk'}^{\top} \\ \upsilon_{tt'}\mathbf{R}_{kk'} & \mathbf{R}_{k'k'} \end{bmatrix}\right) \tag{12}$$

under H_{0k}^t and $H_{0k}^{t'}$ and that $\|\tilde{\mathbf{z}}(k,k';t,t')\|_2^2 = T_k^t + T_{k'}^{t'}$. It remains to find $\mathrm{Var}\left(T_k^t + T_{k'}^{t'} | H_{0k}^t, H_{0k}^{t'}\right)$, which is equal to $\mathrm{Var}(T_k^t | H_{0k}^t) + \mathrm{Var}\left(T_k^{t'} | H_{0k'}^{t'}\right) + 2\mathrm{Cov}\left(T_k^t, T_k^{t'} | H_{0k}^t, H_{0k}^{t'}\right)$ and where

$$\begin{aligned} \mathrm{Var}\left(T_k^t + T_{k'}^{t'} | H_{0k}^t, H_{0k}^{t'}\right) &= 2\mathrm{tr}\left(\mathbf{G}_{kk'}^{tt'}\mathbf{G}_{kk'}^{tt'}\right) \\ &= 2\mathrm{tr}(\mathbf{R}_k\mathbf{R}_k) + 2\mathrm{tr}(\mathbf{R}_{k'}\mathbf{R}_{k'}) + 4\upsilon_{tt'}^2\mathrm{tr}\left(\mathbf{R}_{kk'}^{\top}\mathbf{R}_{kk'}\right), \end{aligned} \tag{13}$$

since $\mathrm{Var}(T_k^t | H_{0k}^t) = 2\mathrm{tr}(\mathbf{R}_k\mathbf{R}_k)$ and $\mathrm{Var}\left(T_{k'}^{t'} \middle| H_{0k'}^{t'}\right) = 2\mathrm{tr}(\mathbf{R}_{k'}\mathbf{R}_{k'})$, implying that

$$4\upsilon_{tt'}^2\mathrm{tr}\left(\mathbf{R}_{kk'}^{\top}\mathbf{R}_{kk'}\right) = 2\sqrt{2\mathrm{tr}(\mathbf{R}_{kk}\mathbf{R}_{kk})2\mathrm{tr}(\mathbf{R}_{k'k'}\mathbf{R}_{k'k'})}\mathrm{Corr}(T_{kt}, T_{k't'}) \tag{14}$$

$$\Rightarrow \mathrm{Corr}(T_{kt}, T_{k't'}) = \upsilon_{tt'}^2 \frac{\mathrm{tr}\left(\mathbf{R}_{kk'}^{\top}\mathbf{R}_{kk'}\right)}{\sqrt{\mathrm{tr}(\mathbf{R}_{kk}\mathbf{R}_{kk})\mathrm{tr}(\mathbf{R}_{k'k'}\mathbf{R}_{k'k'})}}.$$

This quantity is strictly positive and states that the correlation between gene-based test statistics from different genes and for different traits from separate GWAS is proportional to the product of shared LD between gene-specific SNP sets, the proportion of shared GWAS subjects, and the phenotypic correlation between the traits. The quantity $v_{tt'}^2$ has as its maximum the square of phenotypic correlation $\mathrm{Corr}(t, t')$ in a single GWAS cohort. This shows that any pair of gene-based tests statistics, and by extension PRP_{kt} and $\mathrm{PRP}_{kt'}$, are not generally independent if there are overlapping GWAS subjects and SNPs in LD between the SNP sets. This motivates correction of $c(t, t')$ for nonzero correlation between PRP_{kt} and $\mathrm{PRP}_{kt'}$.

To perform this correction in practice, researchers can either (i) de-correlate SNP-level Z-statistics for each trait pair using the method of LeBlanc et al (2018) before applying gene-based association testing and subsequent polygenicity evaluations to each trait, or (ii) apply our post-hoc correction to the estimated number of shared disease associated genes using the principles of simulation extrapolation (SIMEX; Stefanski & Cook, 1995). We intend to measure the effect of GWAS sample overlap on estimates of shared disease associated gene counts between pairs of traits using simulation because an analytic expression of it using the definition of $c(t, t')$ is challenging to derive. In this procedure, we begin by estimating posterior risk probabilities PRP_{kt} and $\mathrm{PRP}_{kt'}$ for traits t and t' and all genes $k = 1, \ldots, |\mathcal{S}(t, t')|$ in the set of genes $\mathcal{S}(t, t')$ tested for association with both traits. We require the estimated number of disease associated genes $c(t)$ and $c(t')$ for each trait, their estimated SNP heritability or their τ^t and $\tau^{t'}$ values, and the estimated sample overlap correlation parameter $v_{tt'}$ for the trait pair,

which is estimable from GWAS summary statistics as the empirical correlation between non-significant SNP-level Z-statistics (Lorincz-Comi et al., 2024b).

We then specify a grid of $v^*_{tt'}$ values in the interval $v_{tt'} < v^*_{tt'} \leq 1$, generate simulated SNP-level summary statistics under the above model which includes GWAS sample overlap, perform gene-based association testing using the sum of SNP chi-squares, estimate gene-level posterior risk probabilities for each trait directly from the likelihoods of the latent indicators (i.e., without integrating over the prior distributions of δ and τ), and estimate the number of shared disease associated genes using the $c(t,t')$ estimator with $|\mathcal{S}(t,t')| = 8{,}148$ independent genes. We use the simulation-averaged estimates of shared counts at each $v^*_{tt'}$ value to extrapolate from the observed shared count $c(t,t')$ at $v_{tt'}$ back to the estimated shared count when there is no GWAS sample overlap, denoted as $c\left(t,t;v^*_{tt'}=0\right)$. We then multiply the original $c(t,t')$ estimate by $c\left(t,t;v^*_{tt'}=0\right)/c\left(t,t;v^*_{tt'}=v_{tt'}\right)$ to adjust it for sample overlap. We provide four examples of the performance of this procedure in **Figure S7** in the **Supplement**, which shows that increasing overlap proportions generally increase the estimated numbers of shared disease associated genes quadratically for pairs of phenotypically correlated traits, but that our proposed procedure removes bias from this source. We also show in the **Supplement** that our original shared disease associated gene estimates from the $c(t,t')$ estimator only differ from SIMEX-adjusted estimates for 5.2% of pairs, and that the average difference between original and adjusted disease associated gene counts for these 5.2% of pairs was only 1.

Real data application

We estimated shared and non-shared disease associated gene counts using gene-based test statistics (T_k) for 32 complex traits and statements of their null and non-null distributions downloaded from a public database (Lorincz-Comi et al., 2024a). A full list of the repositories from which these GWAS data were accessed is available in the **Appendix**, where the phenotype label abbreviations and GWAS sample sizes are also present. All GWAS were performed in populations of predominantly or exclusively European ancestry. Genetic correlation estimates using LDSC are presented for all 496 trait pairs in **Supplementary Figure S11**. Gene-based test statistics were calculated as the sum of SNP-level association chi-square statistics from GWAS for all SNPs within $\pm$50 Kb of the gene start and end base pair positions defined using Ensembl (Harrison et al., 2024), and which could be matched to the 1000 Genome Phase 3 European (1KGv3-EUR) LD reference panel (Siva, 2008). We defined Bonferroni significant trait-genes as those with a P-value less than $0.05/d$, where d is the number of independent gene-based association test statistics estimated using the method of Jiang et al. (2022) applied to each chromosome separately and summed across them. Let $\mathbf{\Sigma}_c$ be the $R_c \times R_c$ matrix of correlations between gene-based test statistics for a single trait calculated using the method described previously for the cth chromosome and where $\lambda_{1c}, \ldots, \lambda_{R_c c}$ are its eigenvalues. We calculated d_c as

$$d_c = \sum_{i=1}^{R_c} I(\lambda_{ci} \geq 1) + \lambda_{ci} I(\lambda_{ci} < 1) \qquad (15)$$

where $I(a)$ is 1 if the argument a is true and 0 otherwise. Where any $\mathbf{\Sigma}_c$ matrix was not positive definite, we used the nearest positive definite matrix via the transform of Choi et

al. (2019). The number of independent gene-based association tests performed across the genome was approximated as $d = \sum_c d_c$ which we found to be 12,272 using the 1KGv3-EUR reference. LD scores and minor allele frequencies which parameterized the SNP-level and subsequent gene-level models were calculated using the 1KGv3-EUR reference panel. LD scores were calculated using SNPs in windows 1 centimorgan wide.

To count the numbers of independent Bonferroni- and false discovery rate (FDR)-significant genes in gene-based association testing as we show in **Figure 1a**, we first formed the set $\{T_\ell\}$ such that $T_\ell > T_s \; \forall \; s \in \left\{s: s \neq \ell, \mathrm{Corr}(T_\ell, T_s) > \sqrt{0.5}\right\}$. The number of independent significant genes was then the size of the set $\{T_\ell\}$ using either Bonferroni or FDR correction. To identify drug target candidates, we joined lists of drug-target interactions from ChEMBL (Gaulton et al., 2017), BindingDB (Liu et al., 2024), and GtoPdb (Harding et al., 2022) to their molecules indicated in DrugBank (Wishart et al., 2018), of which there were 3,369. We also estimated the risk-directed genetic correlation between traits (*cf.* Equation 9), and all traits (see **Appendix**) were inferred to already be coded in the direction of risk in their respective GWAS except HDL, intelligence, and education.

Results

Estimating polygenicity among 32 human complex diseases

We estimated the number of disease-associated genes (DAGs) for 32 complex traits and display the counts in **Figure 1a**. These results suggest that traits such as body

353 mass index (BMI), high-/low-density lipoprotein (HDL/LDL), schizophrenia (SCZ),
354 intelligence (INT), and diastolic, systolic, and pulse pressure (DBP, SBP, PP) may have
355 more than 500 DAGs, while other traits such as Alzheimer's disease (AD) or
356 amyotrophic lateral sclerosis (ALS) may only have only 100-300. These counts are also
357 highly correlated with the number of independent Bonferroni- and FDR-significant genes
358 scaled by the square root of GWAS sample size, denoted respectively as sBonf and
359 sFDR, which in the **Supplement** are shown to be approximately proportional to the true
360 number of disease associated genes. Hence, linear correlation of our estimated
361 numbers of DAGs with sBonf (Pearson r=0.92) and sFDR (r=0.98) implies at least linear
362 correlation between our estimates and the true numbers of true disease-associated
363 genes. Independence between genes that were significant in hypothesis testing with
364 gene-based association test statistics was inferred using the clumping procedure in
365 Lorincz-Comi et al. (2025a) based on shared LD between gene-specific SNP sets. Our
366 estimated counts of disease associated genes are also not correlated with GWAS
367 sample size (Pearson r=0.02, P-value=0.900) or is square root (Pearson r=0.07, P-
368 value=0.691) across the 32 traits, suggesting that they are not heavily influenced by
369 statistical power in GWAS, the primary source of which is sample size.
370
371 **Figure 1b** shows an example of the relationship between gene-level posterior risk
372 probabilities (PRPs) using our method and gene-based association test P-values from
373 the GenT method (Lorincz-Comi et al., 2025a) for AD on chromosome 1. These results
374 show that genes with the smallest P-values are most likely to be assigned a large PRP,
375 such as *CR1* and *B4GALT3*, but that genes which fail to meet the level of genome-wide

significance in gene-based testing can still be assigned relatively large PRPs under our model, such as *SH2D2A* which had PRP of 0.97 for AD but gene-based test P-value of 1.1E-5, above the Bonferroni significance level of 3.9E-6 (*cf.* **Methods**) and therefore not detected using hypothesis testing. *SH2D2A* is associated with neuronal signaling via synapse formation (Sachse et al., 2019) and has previously been shown to be overexpressed in AD cases vs controls in immune cells (Chen et al., 2024), suggesting it may indeed be associated with AD risk. Generally, there is close concordance between PRPs and the level of significance in gene-based hypothesis testing, but this example demonstrates that PRPs can be used as an additional inferential tool to discover gene-disease associations missed by hypothesis tests.

Figure 1c displays the gene- and trait-specific posterior risk probabilities (PRPs) summed within each chromosome. These results suggest that the genes associated with the 32 traits are not uniformly distributed across the 22 chromosomes, but that chromosome 17 may contain the largest number of DAGs, despite it only containing the fifth largest number of tested genes (see **Supplementary Figure S9**). These counts are contributed from associated genes from all traits, and no small subset of traits dominates the chromosome-specific counts for any chromosome. We show in **Figure 1d** an example of how gene-level posterior risk probabilities can be used to provide an inference of shared association between coronary artery disease (CAD) and LDL, a known CAD risk factor (Ballantyne, 1998). These results highlight two lead loci on chromosome 7 which are likely shared between CAD and LDL, indexed by *DDX56*

(7p13) and *DUS4L* (7q22.3), due to their posterior shared risk probabilities (S-PRPs) near 1.

Estimating shared polygenicity among complex diseases

On average, an estimated 2.4 to 29.7% (mean=16.8%) of trait-associated genes are not shared with any other traits (**Figure 2a**). For example, of the estimated 120 genes which are inferred to associate with risk of major depression, only 14 (SE=0.89; 11.7%) are not associated with at least one other trait. Similarly, of the 58 genes inferred to associate with chronic kidney disease (CKD), only 13 (SE=0.81; 22.4%) are estimated to be specific only to CKD. We also estimate the total number of genes across the genome which are not associated with any of the 32 traits as 8,312 (SE=40.25), implying that approximately 50.9% (SE=2.5E-3) of the 16,324 genes tested for association with each of the 32 traits we studied may contribute to the SNP heritability of at least one of them (*cf.* Methods). **Figure 2b** shows the matrix of cosine similarity values (see Xie et al., 2021) between all trait pairs, which is the ratio of shared disease associated gene counts to the geometric means of total disease associated gene counts for each trait pair such that larger values indicate greater sharing of associated genes and smaller values indicate less sharing. The row/column ordering of traits is determined by hierarchical clustering, and these results suggest that genetic similarity may be used to approximately group phenotypically similar traits into distinct clusters including a psychiatric/behavioral cluster, a metabolic/cardiovascular cluster, and an age-related/ autoimmune cluster. These clusters show greater evidence of associated gene sharing within them than between, and some traits such as stroke and chronic kidney disease

show little evidence of gene sharing with other traits, potentially explained by their low SNP heritability (**Supplementary Figure S4**). Together, these results suggest that cosine similarity indices applied to our estimates of shared association at the gene-level can be used to identify phenotypically similar traits, supporting our use of shared PRPs to identify shared genetic architecture between traits.

A case study of polygenic sharing between traits is presented in **Figure 2c** for Lewy body dementia (LBD) using both the cosine and Jaccard indices (see Frech & Chen, 2010). These results show that the average LBD-associated gene is most likely to be shared with highly polygenic traits such as TG, SCZ, and BMI, but that, after considering the estimated sizes of disease associated gene sets of LBD and the other traits, LBD is most genetically similar to Parkinson's disease (PD), Alzheimer's disease (AD), and a multivariate index of healthy aging (mvAGING; Rosoff et al., 2023), which is supported by their shared phenotype of age-related neurodegeneration. For example, there is an estimated 0.19 probability that a randomly selected LBD-associated gene is associated with PD risk, and an estimated 0.32 probability that a randomly selected LBD risk gene is associated with AD risk. We show in the **Supplement** that cosine index values for AD with the other 31 traits are mildly linearly correlated with their estimated genetic correlations (Pearson r=0.31) using LD score regression (Bulik-Sullivan et al., 2015), but that for some traits such as MDD, LDL, and EDU, there is evidence of a nonzero proportion of shared associated genes but no evidence of a nonzero global genetic correlation between them. Across all 496 trait pairs, the linear correlation between absolute genetic correlations and cosine index values was 0.22 (P=8.0E-7).

Two case studies of gene sharing between AD and LBD are presented in **Figure 2d** for two loci which are respectively indexed by *KCTD13* (16p11.2) and *SLC8B1* (12q24.13). *KCTD13* had a posterior probability of being shared by AD and LBD of 1.00 and is associated with brain morphology (Qiang et al., 2023) and known to affect short-term memory by regulating synaptic activity in the hippocampus (Arbogast et al., 2019). The inhibition of *SLC8B1* has also been shown to reduce cognitive performance and memory in mice (Jadiya et al., 2019) and its expression may protect against neuronal cell death (Jadiya et al., 2017). Together, this supporting evidence suggests associations of *KCTD13* and *SLC8B1* with impaired memory, a hallmark symptom of both AD and LBD.

Discovery of non-pleiotropic drug target candidates in ALS

We next provide a motivating example of how shared PRPs can be used to identify candidate drug targets for ALS with limited potential for off-target effects on other traits (*cf.* Equation 8). **Figure 3a** displays PRPs for ALS and for ALS but not any of the other 31 traits using the set of 3,369 genes with drug targets (i.e., 'druggable genes'; *cf.* **Methods**). These results suggest that *TBK1* and *SCNN1B* have the largest PRPs (0.87 and 0.89, respectively) of associating with ALS but none of the other 31 phenotypes, suggesting they may have lower off-target potential compared to other drug targets with respect to effects on the other 31 traits. As a counterexample, the *KIT* gene is associated with ALS, TG, HDL, BIP, and BMI. **Figure 3b** shows the SNP-level associations between ALS, TG, and gene expression in the thyroid from GTEx v8

(Lonsdale et al., 2013) in the *KIT* locus, first suggesting evidence of nonzero association between SNPs in this locus and ALS, TG, and *KIT* expression. There is also evidence of a negative local genetic correlation between ALS and TG and between ALS and thyroid gene expression, and a positive local genetic correlation between TG and thyroid gene expression. Mendelian Randomization using all FDR-significant thyroid eQTLs as instruments and observed eQTL and ALS Z-statistics as in Lorincz-Comi et al. (2024b) suggested a negative causal effect of *KIT* expression on ALS risk (Estimate=-0.12; P=1.8E-8). These results suggest that while increased *KIT* expression in the thyroid may reduce ALS risk, this may simultaneously be associated with increased TG levels, potentially harmful to overall health.

The sharing of disease association between *KIT* with ALS and TG, and the corresponding non-sharing of the *SCNN1B* and *TBK1* genes with ALS only, is demonstrated in **Figure 3c**. We note that non-sharing of association between ALS and non-ALS traits is not a requirement for a putatively safer ALS drug target, but that the set of genes with low non-ALS trait associations may contain genes with less complex biological pathways of effect and hence better candidates as therapeutics with limited off-target effects. **Figure 3d** displays estimated counts of shared and non-shared disease associated genes that have drug-target interactions for all traits, made by directly summing PRPs for these genes. These results suggest that for some traits such as multiple sclerosis (MS), PD, and LBD, currently available drug targets are likely to associate with at least one other trait. These results also suggest that approximately 14 druggable genes are candidate ALS targets with limited evidence of conferring off-target

effects by modifying any of the other 31 traits. Of these 14, the *TBK1* and *SCNN1B* genes have the strongest evidence of ALS association and limited non-ALS association. We provide PRPs and S-PRPs for all 3,369 druggable genes for all traits in the **Supplemental Data**.

Discovery of non-pleiotropic drug target candidates in Alzheimer's disease (AD)

We also leveraged estimated shared polygenicity between AD and each of the other 31 traits to identify candidate drug targets which have limited evidence of off-target potential. Some examples of these genes include *EARS2*, *EPHA1*, *ITGA2B*, *MME*, and *MS4A2*, each of which have a PRP with AD greater than 0.9, and no PRP with any other trait greater than 0.9 (**Supplementary Figure S10**). The *MS4A2* gene is a member of the *MS4A* gene cluster (11q12.1) which is known to play a critical role in autoimmune activation (Mattiola et al., 2021; Luo et al., 2024) and even moderate *TREM2* effects on AD risk (Ma et al., 2015; Deming et al., 2019). On the other hand, many well-known druggable AD risk genes have evidence of pleiotropic association with multiple other traits (**Supplementary Figure S10**). These genes include *ACE*, *APOC2*, *APOE*, *BIN1*, and *CLU*. As examples, *CLU* had PRP with AD of 0.99 and with SCZ of 1.00; *APOC2* had PRP with AD of 1.00 and PRPs with TG, LDL, HDL, and CAD each of 1.00. *CLU* may simultaneously confer risk of AD and SCZ because of its role in and response to autoimmunity and inflammation, which is dysregulated in both AD and SCZ cases vs controls (Falgarone & Chiocchia, 2009; Sardi et al., 2011; Bergink et al., 2014). *APOC2* is known to be involved in the metabolism of lipoprotein (Jong et al., 1999), and so has

been implicated in hypertriglyceridemia risk (Gao et al., 2020), suggesting that its conferral of AD risk may be mediated by lipidemia.

Repurposable drug candidates using BPACT

We can also leverage the shared association of a gene with multiple disease trait(s) to identify candidate drug targets that have a therapeutic effect on multiple diseases simultaneously. In this subsection, we refer to these genes as candidate repurposable drug candidates since they may already have some demonstrated evidence of treating one condition but have so far not been evaluated for the treatment of another, though the genetic evidence suggests they may be able to. For these genes, a genetic correlation between the target disease and other traits in the direction of risk (e.g., positive correlation for T2D and LDL cholesterol, negative correlation for T2D and intelligence), suggests that targeting them with a drug may have therapeutic effects on multiple traits simultaneously. We identified these genes as those with large PRPs for multiple traits with which the index trait was locally genetically correlated in the direction of risk with all other traits on average (*cf.* Equation 3) and present these quantities in **Figure 4a**. These results suggest that traits such as SCZ, CAD, and T2D may have the largest number of repurposable candidates, which is expected since these traits have evidence of being highly polygenic (**Figure 1a**), and that traits such as lupus and MS may not have any known associated genes that are not associated with at least one other trait.

As a demonstrative example, we show in **Figure 4b** that *PLEKHA1* is associated with BMI, INT, and T2D, and that the directions of their local genetic correlations with AD suggests that targeting *PLEKHA1* may have therapeutic/preventative effects on all traits including AD. In this locus, six SNPs (rs10788284, rs6585827, rs2421016, rs7097701, rs10510110, rs2280141) are associated with both AD and each of BMI, INT, and T2D at level P<5E-5 in intersection-union tests (IUTs). The rs10510110 SNP of *PLEKHA1* is associated with AD (FDR=3.2E-3) and gene expression (GTEx v8; Lionsdale et al., 2018) in visceral adipose (FDR=3.6E-7), brain cortex (FDR=3.4E-2), and thyroid tissue (FDR=1.8E-9) (**Figure 4c**). Citric acid (CA) is a nutraceutical agent interacting with *PLEKHA1* (DrugBank, DB04272; Knox et al., 2024) that is traditionally recognized for its association with body mass index, lipid cholesterol, and glucose metabolism (Tomar et al., 2019; Yadikar et al., 2022), though there is also some evidence suggesting CA can modulate oxidative stress in the brain by reducing inflammation and lipid peroxidation (Amin et al., 2011), potentially via its role in fatty acid metabolism (Abdel-Salam et al., 2014) and/or its inhibition of the acetylcholinesterase (AChE) enzyme (Suner et al., 2021). AChE inhibitors have demonstrated efficacy in treating AD symptoms (Marucci et al., 2021), and one hypothesized mechanism is via the deformation of amyloid-beta tangles (Cerbai et al., 2007; Furukawa-Hibi et al., 2011).

Similarly to AD and *PLEKHA1*, two SNPs in the *GAK* locus (rs3775121, rs873785) are associated with HDL, PD, and TG (P<5E-5), and their local genetic correlations with ALS are each in the direction of risk for all traits. The rs3775121 SNP is associated with ALS (FDR=5.3E-3) and gene expression (GTEx v8; Lionsdale et al., 2018) in aortic

(FDR=1.4E-3), blood (FDR=2.8E-4), and colon tissue (FDR=2.2E-3) (**Figure 4c**). Fostamatinib is a drug which primarily inhibits spleen-associated kinase (*SYK*), but can also inhibit many additional kinases including cyclin-G-associated kinase (*GAK*) (DrugBank, DB12010; Knox et al., 2024). Fostamatinib was originally developed to treat immune conditions such including autoimmune hemolytic anemia, thrombocytopenia, and immunoglobulin A nephropathy (Markham, 2018), though bioinformatic analyses using Mendelian Randomization and molecular docking support fostamatinib as a potential therapeutic target for ALS and PD (Yergolkar et al., 2020; Duan et al., 2024; Eshak & Arumugam, 2024). Fostamatinib may have evidence as a candidate ALS target because *SYK* and *GAK* can cause inflammation-associated cell death and cognitive impairment (Birkle & Brown, 2023; Zhou et al., 2024; Miyazaki et al., 2021; Pan et al., 2013).

In summary, BPACT identified genes which may be drug repurposing candidates for complex traits including ALS and AD. These genes had evidence of association with multiple traits each in the direction of risk such that targeting them with a drug could lead to therapeutic effects for all associated traits. We demonstrated this phenomenon using the AD-associated *PLEKHA1* gene and ALS-associated *GAK* gene. These genes interact with available compounds, citric acid with *PLEKHA1* and fostamatinib with *GAK*, and the literature provides supporting evidence that these compounds associate with AD and ALS pathology. Future functional and experiment studies may provide greater insight into the degree to which these compounds may simultaneously confer protective effects against AD/ALS and other traits.

Discussion

We present BPACT that can be used to estimate the number of genes which contribute to the SNP heritability of a trait, and the number of genes which contribute SNP heritability to multiple traits. We showed how this method can be used to estimate the number and proportion of genes which are uniquely associated with an index trait but no other trait in a set, and that more than half of the genome has evidence of explaining SNP heritability in at least one of the 32 traits we studied. Results from analyses of shared heritability highlighted druggable genes that may have evidence of reducing disease risk while minimizing the potential for side effects if targeted clinically, and we demonstrated this phenomenon using *KIT* for ALS as an example.

We first showed that there is substantial variability in the estimated number of disease associated genes (DAGs) across the 32 traits. We were also able to identify a highly polygenic group of traits which included BMI, DBP, SBP, PP, LDL, HDL, TG, intelligence, BIP, and SCZ. Estimates of polygenicity correlated almost perfectly with the number of GWAS significant genes adjusted for sample size, which we showed in the **Supplement** to be proportional to the true number of associated genes for a given trait. This concordance provides supporting evidence that our estimator of the number of disease associated genes is at least proportional to the true number of disease associated genes for the trait, and since it is adjusted for GWAS sample size, may be unbiased by GWAS statistical power. We also showed how the method which produces estimates of genome-wide trait associated gene counts can be used to make gene-level inferences

of risk association using AD and chromosome 1 as an example. These results showed that gene-level posterior risk probabilities can be used to infer a nonzero association between a gene and phenotype in the absence of genome-wide significance during hypothesis testing at the SNP or gene levels. We then showed that the estimated number of disease associated genes (DAGs) across all 32 traits is not uniformly distributed across chromosomes after adjusting for the number of genes they contain, and that chromosome 17 had the largest estimated number of genes which were putatively associated with any trait.

We next showed that most genes which explained nonzero heritability in a trait were shared with at least one other trait. For example, we estimated that 24.9% of genes associated with ischemic stroke risk and only 2.5% of genes associated with the index of aging (mvAGING) are not associated with any other trait we tested. We then showed that Jaccard and Cosine indices applied to total and shared DAG counts can be used to measure their shared polygenicity. These quantities were reported in the **Supplement** to moderately positively correlate with estimated global absolute genetic correlations, but that for some trait pairs we find evidence of substantial sharing but estimated genetic correlations very close to 0. These results provide empirical evidence that shared polygenicity is a weaker form of genetic similarity than genetic correlation, and that global null estimates of genetic correlation do not imply genetic independence between traits.

The evidence also suggested that most trait-associated genes with known drug interactions were shared across multiple traits, and we showed an example of the *KIT* gene associated with ALS, BIP, TG, HDL, and BMI. Expression of *KIT* in the thyroid was negatively associated with ALS risk but positively associated with TG levels, suggesting that targeting of *KIT* with an agonist may reduce ALS risk but simultaneously elevate TG levels, a potentially harmful side effect. We presented the *TBK1* and *SCNN1B* genes as examples of alternative druggable candidates, each of which were associated with ALS but had no evidence of association with any of the other 31 traits. *TBK1* can mediate activation of the NF-κB transcription factor which regulates innate and adaptative immunity processes (Liu et al., 2017; Shi et al., 2018; Balka et al., 2020), and targeting of *TBK1* with fostamatinib has been demonstrated to reduce ALS risk in vitro (Duan et al., 2024). *SCNN1B* has been shown to suppress the MAPK signaling pathway (Qian et al., 2023) which is dysregulated in ALS patients (Sahana et al., 2021; Yadav et al., 2021).

We also showed that leveraging gene pleiotropy across multiple traits may nominate repurposable drug targets, and presented examples of *PLEKHA1* with AD and *GAK* with ALS. These targets can be viewed as potentially therapeutic for an index trait because of modification to one or more of its risk factors, or simultaneous therapeutic effects on multiple traits via independent biological pathways. *PLEKHA1* is expressed across many tissues included the brain, heart, and adipose tissue (GTEx v8; Lonsdale et al., 2013), and contains the pleckstrin homology protein folding domain which is associated with the binding of phosphorylated lipids onto inositol rings (Lemmon, 2007). Knockout

of *GAK* has been shown to disrupt the homeostasis of lysosomes during autophagy (Miyazaki et al., 2021), and like *PLEKHA1* it is expressed across many tissues including the brain, heart, and adipose tissue (Lonsdale et al., 2013). The pleiotropic effects of these genes on multiple phenotypically distinct traits may be explained by their broad expression patterns across multiple tissues. We showed that the proportion of genes for which shared associations with other traits can be detected is generally large, emphasizing that the richest source of candidate drug targets for a complex disease may be provided by the set of pleiotropic genes.

Our study is strengthened by the generality of the underlying statistical model and the wide range of inferences which can be made once it is fitted. The model is also advantageous because it accounts for its own parameter misspecification by integrating over the prior parameter space. Gene-level posterior risk probabilities can be used to infer trait-specific association, trait-shared association, and trait non-association in the context of the BPACT statistical model. We also provide expressions for the effect of GWAS sample overlap on gene-based test statistic correlations and the working number of independent gene-based association tests genome-wide, each of which henceforth have been absent in the literature and have respectively precluded direct comparison of gene-based test statistics from multiple GWAS cohorts and well-calibrated control of Type I error rates in genome-wide testing. We proposed a new approach to address the challenge of correlated gene-based test statistics during inference, which is based on the well-studied SIMEX approach (Stefanski & Cook, 1995) that is shown in the **Supplement** to reduce any extant bias in estimated counts of shared associated genes

between trait pairs. We also present a new model-fitting approach for correlated test statistics based on composite posterior densities which are iteratively evaluated over randomly selected and weakly correlated gene sets. This approach uses the principles of imputation to make its inference, and without it our approach could lead to slightly inflated estimates of disease associated gene counts because of shared LD between gene-specific SNP sets. Finally, **Supplementary Table S1** shows that our method is computationally efficient, spending approximately 15 minutes to run on an Intel® Xeon® Gold 6148 CPU 2.40GHz machine.

Our method has the following limitations. Estimates of disease associated gene (DAG) counts and their shared proportions adjust for GWAS sample size, but do assume accurate phenotyping, that genes only have non-interacting effects on the trait, and correct specification of the gene-based test statistic null distributions. Violations of any of these assumptions may bias the estimated DAG counts via mis-specified model priors and likelihood functions. Our model intends to account for misspecification of model parameters, but does assume a correctly specified model structure of additive SNP heritability. If the structure and/or parameters of these models are inappropriate for some traits, or for some loci for some traits, it may cause the model to return a posterior risk probability which is far from the true value of the latent binary indicator of association, which could have downstream consequences on estimated shared and non-shared DAG counts for traits and their pairs. Our model also assumes that within each block of correlated gene-based test statistics, only a single disease associated gene is present. Future extensions of BPACT method may attempt to relax this

assumption. Finally, an inherent limitation of using gene-based test statistics to make the aforementioned inferences is that any nuance at the SNP level may not be completely captured at the gene level.

Future researchers may also use BPACT and its results to evaluate the degree to which increases in GWAS sample sizes are likely to detect additional genes explaining heritability. For example, we estimate that 133 genes may associate with AD, and a recent AD GWAS has detected 82 independent loci (Bellenguez et al., 2021). This suggests that increasing AD GWAS sample size may be likely to produce new meaningful insights into the genetic etiology of AD. However, for traits such as BMI, for which we estimate that approximately 943 associated genes may exist, the currently reported ~1,000 BMI-associated loci (Loos & Yeo, 2022) may cover most of the entire associated gene set and so further investment in continually larger BMI GWAS sample sizes may not return the same value which it requires. We also assert that BPACT may be used to identify drug targets for complex disease which have a putatively low probability of off-target effects on other phenotypes. We estimate that such genes do exist for many of the traits we studied but reiterate that these sets are often of quite small size. Nevertheless, these gene sets may be optimal candidates for developing putatively safer therapeutic targets for complex disease. Finally, the integration of quantitative trait loci (QTL) data into the BPACT model may help it to nominate drug targets with supporting transcriptomic, proteomic, epigenomic, and/or metabolomic evidence that may have greater therapeutic potential than those identified just from GWAS summary statistics.

Appendix

Here we present the names of each trait we studied, its abbreviation displayed throughout the manuscript, the PubMed ID to the study from which the original GWAS data came, and the GWAS sample size using the convention of <**abbreviation**: Full name (PubMed ID; GWAS sample size)>. All GWAS cohorts were of strictly or predominantly European ancestry.

- **AD**: Alzheimer's disease (35379992; 487,511)
- **ADHD**: Attention-deficit/hyperactivity disorder (36702997; 225,534)
- **AFIB**: Atrial fibrillation (36653681; 2,339,188)
- **agingR1**: R1 index of aging-associated brain atrophy (39147830; 49,482)
- **agingR2**: R2 index of aging-associated brain atrophy (39147830; 49,482)
- **agingR3**: R3 index of aging-associated brain atrophy (39147830; 49,482)
- **agingR4**: R4 index of aging-associated brain atrophy (39147830; 49,482)
- **agingR5**: R5 index of aging-associated brain atrophy (39147830; 49,482)
- **ALS**: Amyotrophic lateral sclerosis (34873335; 138,086)
- **BIP**: Bipolar I or II disorder (34002096; 413,466)
- **BMI**: Body mass index (30239722; 694,649)
- **CAD**: Coronary artery disease (36474045; 1,165,690)
- **CKD**: Chronic kidney disease (31152163; 625,219)
- **DBP**: Diastolic blood pressure (30224653; 757,601)
- **EDU**: Educational attainment (35361970; 3,037,499)
- **HDL**: High-density lipoprotein cholesterol (34887591; 1,320,016)

- **INT**: Fluid intelligence (36150907; 216,381)
- **LBD**: Lewy body dementia (33589841; 16,516)
- **LDL**: Low-density lipoprotein cholesterol (34887591; 1,320,016)
- **LUPUS**: Lupus (34278373; 324,698)
- **MDD**: Major depressive disorder (30718901; 807,553)
- **MS**: Multiple sclerosis (34737426; 456,348)
- **mvAGING**: Multivariate index of non-pathologic aging (37550455; 1,958,000)
- **PD**: Parkinson's disease (38155330; 611,485)
- **PP**: Pulse pressure (30224653; 757,601)
- **RA**: Rheumatoid arthritis (34737426; 456,348)
- **SBP**: Systolic blood pressure (30224653; 757,601)
- **SCZ**: Schizophrenia (35396580; 320,404)
- **SLEEP**: Sleep duration (30846698; 446,118)
- **STROKE**: Ischemic stroke (36180795; 1,296,908)
- **TG**: Triglycerides (34887591; 1,320,016)
- **T2D**: Type 2 diabetes (37034649; 1,528,967)

Software availability

We developed an R package to compute gene-level posterior risk probabilities and estimates of shared polygenicity between traits which requires only gene- or SNP-level summary statistics from GWAS and an LD reference panel (1000 Genomes Phase 3 is provided) at https://github.com/noahlorinczcomi/bpact. All R code used to perform the

762 analyses presented in the main text is available at
763 https://github.com/noahlorinczcomi/bpact_analysis.
764

765 **Data Availability**

766 All data used in this study was downloaded from a public database of gene-based test
767 statistics for complex traits: https://nlorinczcomi.shinyapps.io/gent/. Linkage
768 disequilibrium reference panels were from the European population of 1000 Genomes
769 Phase 3 study (Siva, 2008) available at https://www.internationalgenome.org/ or
770 https://github.com/privefl/bigsnpr. We provide posterior risk probabilities for each of the
771 32 traits we studied and up to 17,166 genes and make the results available at
772 https://github.com/noahlorinczcomi/bpact. We also provide estimated total, shared, and
773 non-shared disease associated gene counts in the **Supplemental Data**.
774

775 **Acknowledgements**

776 **Funding**: This work was supported by the National Institute on Aging (NIA) under Award
777 Number R01AG084250, U01AG073323, R01AG066707, R01AG076448,
778 R01AG082118, RF1AG082211, R56AG074001, and R21AG083003, and the National
779 Institute of Neurological Disorders and Stroke (NINDS) under Award Number
780 RF1NS133812 to F.C. This work was partly supported by the Alzheimer's Association
781 award (ALZDISCOVERY-1051936) and the funds from the Alzheimer's Drug Discovery
782 Foundation (ADDF) to F.C.
783

References

Abdel-Salam, O.M.E., Youness, E.R., Mohammed, N.A., Youssef, M.M., Omara, E.A., & Sleem, A.A. (2014). Citric acid effects on brain and liver oxidative stress in lipopolysaccharide-treated mice. *Journal of medicinal food*, *17*(5).

Amin, K. A., Kamel, H. H., & Abd Eltawab, M. A. (2011). The relation of high fat diet, metabolic disturbances and brain oxidative dysfunction: modulation by hydroxy citric acid. Lipids in health and disease, 10, 1-11.

Arbogast, T., Razaz, P., Ellegood, J., McKinstry, S. U., Erdin, S., Currall, B., ... & Katsanis, N. (2019). Kctd13-deficient mice display short-term memory impairment and sex-dependent genetic interactions. Human molecular genetics, 28(9), 1474-1486.

Ballantyne, C. M. (1998). Low-density lipoproteins and risk for coronary artery disease. *The American journal of cardiology*, *82*(8), 3-12.

Ballard, J. L., & O'Connor, L. J. (2022). Shared components of heritability across genetically correlated traits. The American Journal of Human Genetics, 109(6), 989-1006.

Balka, K. R., Louis, C., Saunders, T. L., Smith, A. M., Calleja, D. J., D'Silva, D. B., ... & De Nardo, D. (2020). TBK1 and IKKε act redundantly to mediate STING-induced NF-κB responses in myeloid cells. *Cell reports*, *31*(1).

Bedi, O., Dhawan, V., Sharma, P. L., & Kumar, P. (2016). Pleiotropic effects of statins: new therapeutic targets in drug design. Naunyn-Schmiedeberg's archives of pharmacology, 389, 695-712.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological), 57(1), 289-300.

Bergink, V., Gibney, S. M., & Drexhage, H. A. (2014). Autoimmunity, inflammation, and psychosis: a search for peripheral markers. Biological psychiatry, 75(4), 324-331.

Birkle, T. J., & Brown, G. C. (2023). Syk inhibitors protect against microglia-mediated neuronal loss in culture. *Frontiers in Aging Neuroscience*, *15*, 1120952.

Borchers, H. W., & Borchers, M. H. W. (2019). Package ‘pracma’. Practical numerical math functions, 2(5).

Bulik-Sullivan, B. K., Loh, P. R., Finucane, H. K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, ... & Neale, B. M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, *47*(3), 291-295.

Cerbai, F., Giovannini, M. G., Melani, C., Enz, A., & Pepeu, G. (2007). N1phenethyl-norcymserine, a selective butyrylcholinesterase inhibitor, increases acetylcholine release in rat cerebral cortex: a comparison with donepezil and rivastigmine. European journal of pharmacology, 572(2-3), 142-150.

Chen, X., Hu, F., Chi, Q., & Rao, C. (2024). The study of Alzheimer's disease risk diagnosis based on natural killer cell marker genes in the multi-omics data. Journal of Alzheimer’s Disease, 0(0).

Chib, S., & Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. The American statistician, 49(4), 327-335.

Choi, Y. G., Lim, J., Roy, A., & Park, J. (2019). Fixed support positive-definite modification of covariance matrix estimators via linear shrinkage. Journal of Multivariate Analysis, 171, 234-249.

Cohen, J. C., Boerwinkle, E., Mosley Jr, T. H., & Hobbs, H. H. (2006). Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *New England Journal of Medicine*, *354*(12), 1264-1272.

de Leeuw, C. A., Mooij, J. M., Heskes, T., & Posthuma, D. (2015). MAGMA: generalized gene-set analysis of GWAS data. PLoS computational biology, 11(4), e1004219.

853 Deming, Y., Filipello, F., Cignarella, F., Cantoni, C., Hsu, S., Mikesell, R., ... & Cruchaga,
854 C. (2019). The MS4A gene cluster is a key modulator of soluble TREM2 and
855 Alzheimer's disease risk. Science translational medicine, 11(505), 2291.
856
857 Duan, Q. Q., Wang, H., Su, W. M., Gu, X. J., Shen, X. F., Jiang, Z., ... & Chen, Y. P.
858 (2024). TBK1, a prioritized drug repurposing target for amyotrophic lateral sclerosis:
859 evidence from druggable genome Mendelian randomization and pharmacological
860 verification in vitro. *BMC medicine*, *22*(1), 96.
861
862 Eshak, D., & Arumugam, M. (2024). Unveiling therapeutic biomarkers and druggable
863 targets in ALS: An integrative microarray analysis, molecular docking, and structural
864 dynamic studies. *Computational Biology and Chemistry*, *113*, 108211.
865
866 Falgarone, G., & Chiocchia, G. (2009). Clusterin: A multifacet protein at the crossroad of
867 inflammation and autoimmunity. *Advances in cancer research*, *104*, 139-170.
868
869 Frech, C., & Chen, N. (2010). Genome-wide comparative gene family
870 classification. PLoS one, 5(10), e13409.
871
872 Furukawa-Hibi, Y., Alkam, T., Nitta, A., Matsuyama, A., Mizoguchi, H., Suzuki, K., ... &
873 Yamada, K. (2011). Butyrylcholinesterase inhibitors ameliorate cognitive dysfunction
874 induced by amyloid-β peptide in mice. *Behavioural brain research*, *225*(1), 222-229.
875

876 Gao, M., Yang, C., Wang, X., Guo, M., Yang, L., Gao, S., ... & Xian, X. (2020). ApoC2
877 deficiency elicits severe hypertriglyceridemia and spontaneous atherosclerosis: A rodent
878 model rescued from neonatal death. *Metabolism*, *109*, 154296.
879
880 Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., ... &
881 Overington, J. P. (2012). ChEMBL: a large-scale bioactivity database for drug
882 discovery. *Nucleic acids research*, *40*(D1), D1100-D1107.
883
884 Harding, S. D., Armstrong, J. F., Faccenda, E., Southan, C., Alexander, S. P.,
885 Davenport, A. P., ... & NC-IUPHAR. (2022). The IUPHAR/BPS guide to
886 PHARMACOLOGY in 2022: curating pharmacology for COVID-19, malaria and
887 antibacterials. *Nucleic Acids Research*, *50*(D1), D1282-D1294.
888
889 Harrison, P. W., Amode, M. R., Austine-Orimoloye, O., Azov, A. G., Barba, M., Barnes, I.,
890 ... & Yates, A. D. (2024). Ensembl 2024. Nucleic acids research, 52(D1), D891-D899.
891
892 Horton, J. D., Cohen, J. C., & Hobbs, H. H. (2007). Molecular biology of PCSK9: its role
893 in LDL metabolism. *Trends in biochemical sciences*, *32*(2), 71-77.
894
895 Jadiya, P., Kolmetzky, D. W., Tomar, D., Di Meco, A., Lombardi, A. A., Lambert, J. P., ...
896 & Elrod, J. W. (2019). Impaired mitochondrial calcium efflux contributes to disease
897 progression in models of Alzheimer's disease. Nature communications, 10(1), 3885.
898

899 Jadiya, P., Lombardi, A. A., Lambert, J. P., Luongo, T. S., Chu, J., Praticò, D., & Elrod, J.
900 W. (2017). Genetic Rescue of Mitochondrial Calcium Efflux in Alzheimer's Disease
901 Preserves Mitochondrial Function and Protects against Neuronal Cell
902 Death. Biophysical Journal, 112(3), 445a.
903
904 Jiang, L., Miao, L., Yi, G., Li, X., Xue, C., Li, M. J., ... & Li, M. (2022). Powerful and
905 robust inference of complex phenotypes' causal genes with dependent expression
906 quantitative loci by a median-based Mendelian randomization. The American Journal of
907 Human Genetics, 109(5), 838-856.
908
909 Jiang, X., Finucane, H. K., Schumacher, F. R., Schmit, S. L., Tyrer, J. P., Han, Y., ... &
910 Khusnutdinova, E. (2019). Shared heritability and functional enrichment across six solid
911 cancers. Nature communications, 10(1), 431.
912
913 Jong, M. C., Hofker, M. H., & Havekes, L. M. (1999). Role of ApoCs in lipoprotein
914 metabolism: functional differences between ApoC1, ApoC2, and
915 ApoC3. *Arteriosclerosis, thrombosis, and vascular biology*, *19*(3), 472-484.
916
917 Knox, C., Wilson, M., Klinger, C. M., Franklin, M., Oler, E., Wilson, A., ... & Wishart, D.
918 S. (2024). DrugBank 6.0: the DrugBank knowledgebase for 2024. Nucleic acids
919 research, 52(D1), D1265-D1275.
920

LeBlanc, M., Zuber, V., Thompson, W. K., Andreassen, O. A., Schizophrenia and Bipolar Disorder Working Groups of the Psychiatric Genomics Consortium, Frigessi, A., & Andreassen, B. K. (2018). A correction for sample overlap in genome-wide association studies in a polygenic pleiotropy-informed framework. BMC genomics, 19, 1-15.

Lemmon, M. A. (2007, January). Pleckstrin homology (PH) domains and phosphoinositides. In *Biochemical Society Symposia* (Vol. 74, pp. 81-93). Portland Press Ltd.

Liu, J. Z., Mcrae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., ... & Macgregor, S. (2010). A versatile gene-based test for genome-wide association studies. The American Journal of Human Genetics, 87(1), 139-145.

Liu, T., Zhang, L., Joo, D., & Sun, S. C. (2017). NF-κB signaling in inflammation. *Signal transduction and targeted therapy*, *2*(1), 1-9.

Liu, T., Hwang, L., Burley, S. K., Nitsche, C. I., Southan, C., Walters, W. P., & Gilson, M. K. (2024). BindingDB in 2024: a FAIR knowledgebase of protein-small molecule binding data. *Nucleic Acids Research*, gkae1075.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., ... & Moore, H. F. (2013). The genotype-tissue expression (GTEx) project. *Nature genetics*, *45*(6), 580-585.

Loos, R. J., & Yeo, G. S. (2022). The genetics of obesity: from discovery to biology. Nature Reviews Genetics, 23(2), 120-133.

Lorincz-Comi, N., Song, W., Chen, X., Rivera Paz, I., Hou, Y., …, Cheng, F. Combining xQTL and Genome-Wide Association Studies from Ethnically Diverse Populations Improves Druggable Gene Discovery. Available at http://dx.doi.org/10.2139/ssrn.5080346

Lorincz-Comi, N., Yang, Y., Li, G., & Zhu, X. (2024b). MRBEE: A bias-corrected multivariable Mendelian randomization method. *Human Genetics and Genomics Advances*, *5*(3).

Luo, X., Luo, B., Fei, L., Zhang, Q., Liang, X., Chen, Y., & Zhou, X. (2024). MS4A superfamily molecules in tumors, Alzheimer's and autoimmune diseases. Frontiers in Immunology, 15, 1481494.

Lounkine, E., Keiser, M. J., Whitebread, S., Mikhailov, D., Hamon, J., Jenkins, J. L., ... & Urban, L. (2012). Large-scale prediction and testing of drug activity on side-effect targets. Nature, 486(7403), 361-367.

Ma, J., Yu, J. T., & Tan, L. (2015). MS4A cluster in Alzheimer's disease. *Molecular neurobiology*, *51*, 1240-1248.

Mao, T., Chen, J., Su, T., Xie, L., Qu, X., Feng, R., ... & Lin, Q. (2024). Causal relationships between GLP1 receptor agonists, blood lipids, and heart failure: a drug-target mendelian randomization and mediation analysis. *Diabetology & Metabolic Syndrome*, *16*(1), 208.

Markham, A. (2018). Fostamatinib: first global approval. *Drugs*, *78*, 959-963.

Marucci, G., Buccioni, M., Dal Ben, D., Lambertucci, C., Volpini, R., & Amenta, F. (2021). Efficacy of acetylcholinesterase inhibitors in Alzheimer's disease. *Neuropharmacology*, *190*, 108352.

Mattiola, I., Mantovani, A., & Locati, M. (2021). The tetraspan MS4A family in homeostasis, immunity, and disease. *Trends in Immunology*, *42*(9), 764-781.

Miyazaki, M., Hiramoto, M., Takano, N., Kokuba, H., Takemura, J., Tokuhisa, M., ... & Miyazawa, K. (2021). Targeted disruption of GAK stagnates autophagic flux by disturbing lysosomal dynamics. *International Journal of Molecular Medicine*, *48*(4), 1-18.

Nguyen, P. A., Born, D. A., Deaton, A. M., Nioi, P., & Ward, L. D. (2019). Phenotypes associated with genes encoding drug targets are predictive of clinical trial side effects. Nature communications, 10(1), 1579.

990 Pan, J., Zhang, J., Hill, A., Lapan, P., Berasi, S., Bates, B., ... & Haney, S. (2013). A
991 kinome-wide siRNA screen identifies multiple roles for protein kinases in hypoxic stress
992 adaptation, including roles for IRAK4 and GAK in protection against apoptosis in
993 VHL−/− renal carcinoma cells, despite activation of the NF-κB pathway. Journal of
994 biomolecular screening, 18(7), 782-796.
995
996 Privé, F. (2022). Optimal linkage disequilibrium splitting. Bioinformatics, 38(1), 255-256.
997
998 Qian, Y., Zhou, L., Luk, S. T. Y., Xu, J., Li, W., Gou, H., ... & Wong, C. C. (2023). The
999 sodium channel subunit SCNN1B suppresses colorectal cancer via suppression of
1000 active c-Raf and MAPK signaling cascade. *Oncogene*, *42*(8), 601-612.
1001
1002 Qiang, Y. X., Deng, Y. T., Zhang, Y. R., Wang, H. F., Zhang, W., Dong, Q., ... & Yu, J. T.
1003 (2023). Associations of blood cell indices and anemia with risk of incident dementia: A
1004 prospective cohort study of 313,448 participants. Alzheimer's & Dementia, 19(9), 3965-
1005 3976.
1006
1007 Quarteroni, A., Sacco, R., & Saleri, F. (2010). Numerical mathematics (Vol. 37). Springer
1008 Science & Business Media.
1009
1010 Rosoff, D. B., Mavromatis, L. A., Bell, A. S., Wagner, J., Jung, J., Marioni, R. E., ... &
1011 Lohoff, F. W. (2023). Multivariate genome-wide analysis of aging-related traits identifies
1012 novel loci and new drug targets for healthy aging. Nature aging, 3(8), 1020-1035.

Rubin, D. B. (1996). Multiple imputation after 18+ years. Journal of the American statistical Association, 91(434), 473-489.

Sachse, S. M., Lievens, S., Ribeiro, L. F., Dascenco, D., Masschaele, D., Horré, K., ... & Schmucker, D. (2019). Nuclear import of the DSCAM-cytoplasmic domain drives signaling capable of inhibiting synapse formation. The EMBO journal, 38(6), e99669.

Sahana, T. G., & Zhang, K. (2021). Mitogen-activated protein kinase pathway in amyotrophic lateral sclerosis. *Biomedicines*, *9*(8), 969.

Sardi, F., Fassina, L., Venturini, L., Inguscio, M., Guerriero, F., Rolfo, E., & Ricevuti, G. (2011). Alzheimer's disease, autoimmunity and inflammation. The good, the bad and the ugly. Autoimmunity reviews, 11(2), 149-153.

Schaid, D. J., Tong, X., Larrabee, B., Kennedy, R. B., Poland, G. A., & Sinnwell, J. P. (2016). Statistical methods for testing genetic pleiotropy. Genetics, 204(2), 483-497.

Siva, N. (2008). 1000 Genomes project. Nature biotechnology, 26(3), 256-257.

Sivakumaran, S., Agakov, F., Theodoratou, E., Prendergast, J. G., Zgaga, L., Manolio, T., ... & Campbell, H. (2011). Abundant pleiotropy in human complex diseases and traits. The American Journal of Human Genetics, 89(5), 607-618.

Shi, J. H., Xie, X., & Sun, S. C. (2018). TBK1 as a regulator of autoimmunity and antitumor immunity. *Cellular & molecular immunology*, *15*(8), 743-745.

Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M., & Smoller, J. W. (2013). Pleiotropy in complex traits: challenges and strategies. Nature Reviews Genetics, 14(7), 483-495.

Stefanski, L. A., & Cook, J. R. (1995). Simulation-extrapolation: the measurement error jackknife. *Journal of the American Statistical Association*, *90*(432), 1247-1256.

Suner, S. S., Sahiner, M., Ayyala, R. S., Bhethanabotla, V. R., & Sahiner, N. (2021). Versatile fluorescent carbon dots from citric acid and cysteine with antimicrobial, anti-biofilm, antioxidant, and AChE enzyme inhibition capabilities. Journal of fluorescence, 31(6), 1705-1717.

Tomar, M., Rao, R. P., Dorairaj, P., Koshta, A., Suresh, S., Rafiq, M., ... & Venkatesh, K. V. (2019). A clinical and computational study on anti-obesity effects of hydroxycitric acid. RSC advances, 9(32), 18578-18588.

Visscher, P. M., Yengo, L., Cox, N. J., & Wray, N. R. (2021). Discovery and implications of polygenicity of common diseases. Science, 373(6562), 1468-1473.

Vsevolozhskaya, O. A., Shi, M., Hu, F., & Zaykin, D. V. (2020). DOT: Gene-set analysis by combining decorrelated association statistics. PLOS Computational Biology, 16(4), e1007819.

Werme, J., van der Sluis, S., Posthuma, D., & de Leeuw, C. A. (2022). An integrated framework for local genetic correlation analysis. Nature genetics, 54(3), 274-282.

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., ... & Wilson, M. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research*, *46*(D1), D1074-D1082.

Woodward, D. J., Thorp, J. G., Middeldorp, C. M., Akósílè, W., Derks, E. M., & Gerring, Z. F. (2024). Leveraging pleiotropy for the improved treatment of psychiatric disorders. Molecular Psychiatry, 1-17.

Wu, T., & Sham, P. C. (2021). On the transformation of genetic effect size from logit to liability scale. Behavior Genetics, 51(3), 215-222.

Wu, T., Liu, Z., Mak, T. S. H., & Sham, P. C. (2022). Polygenic power calculator: Statistical power and polygenic prediction accuracy of genome-wide association studies of complex traits. Frontiers in Genetics, 13, 989639.

Xie, J., Wang, M., Xu, S., Huang, Z., & Grant, P. W. (2021). The unsupervised feature selection algorithms based on standard deviation and cosine similarity for genomic data analysis. Frontiers in Genetics, 12, 684100.

Yadav, R. K., Minz, E., & Mehan, S. (2021). Understanding abnormal c-JNK/p38MAPK signaling in amyotrophic lateral sclerosis: potential drug targets and influences on neurological disorders. *CNS & Neurological Disorders-Drug Targets (Formerly Current Drug Targets-CNS & Neurological Disorders)*, *20*(5), 417-429.

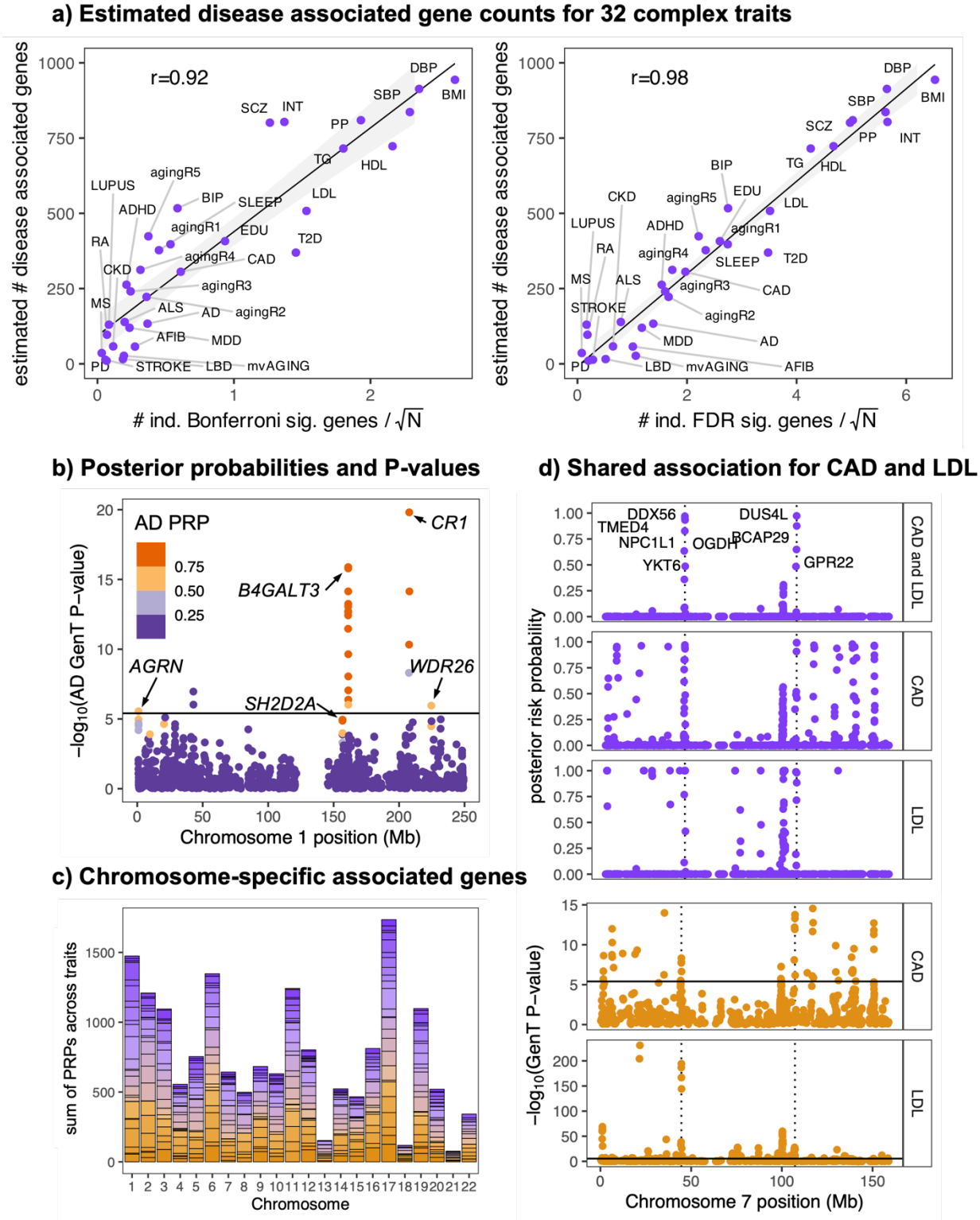
Yadikar, N., Ahmet, A., Zhu, J., Bao, X., Yang, X., Han, H., & Rozi, P. (2022). Exploring the mechanism of citric acid for treating glucose metabolism disorder induced by hyperlipidemia. Journal of Food Biochemistry, 46(12), e14404.

Yergolkar, A. V., Yalamanchili, J., Satish, K., & Saraswathy, G. R. (2020). PND24 target identification and drug repurposing for Parkinson's disease: A novel integrative computational approach. Value in Health Regional Issues, 22, S79.

Zhou, C., Li, J., Wu, X., & Liu, F. (2024). Activation of spleen tyrosine kinase (SYK) contributes to neuronal pyroptosis and cognitive impairment in diabetic mice via the NLRP3/Caspase-1/GSDMD signaling pathway. *Experimental Gerontology*, *198*, 112626.

Zhu, X., & Stephens, M. (2017). Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. The annals of applied statistics, 11(3), 1561.

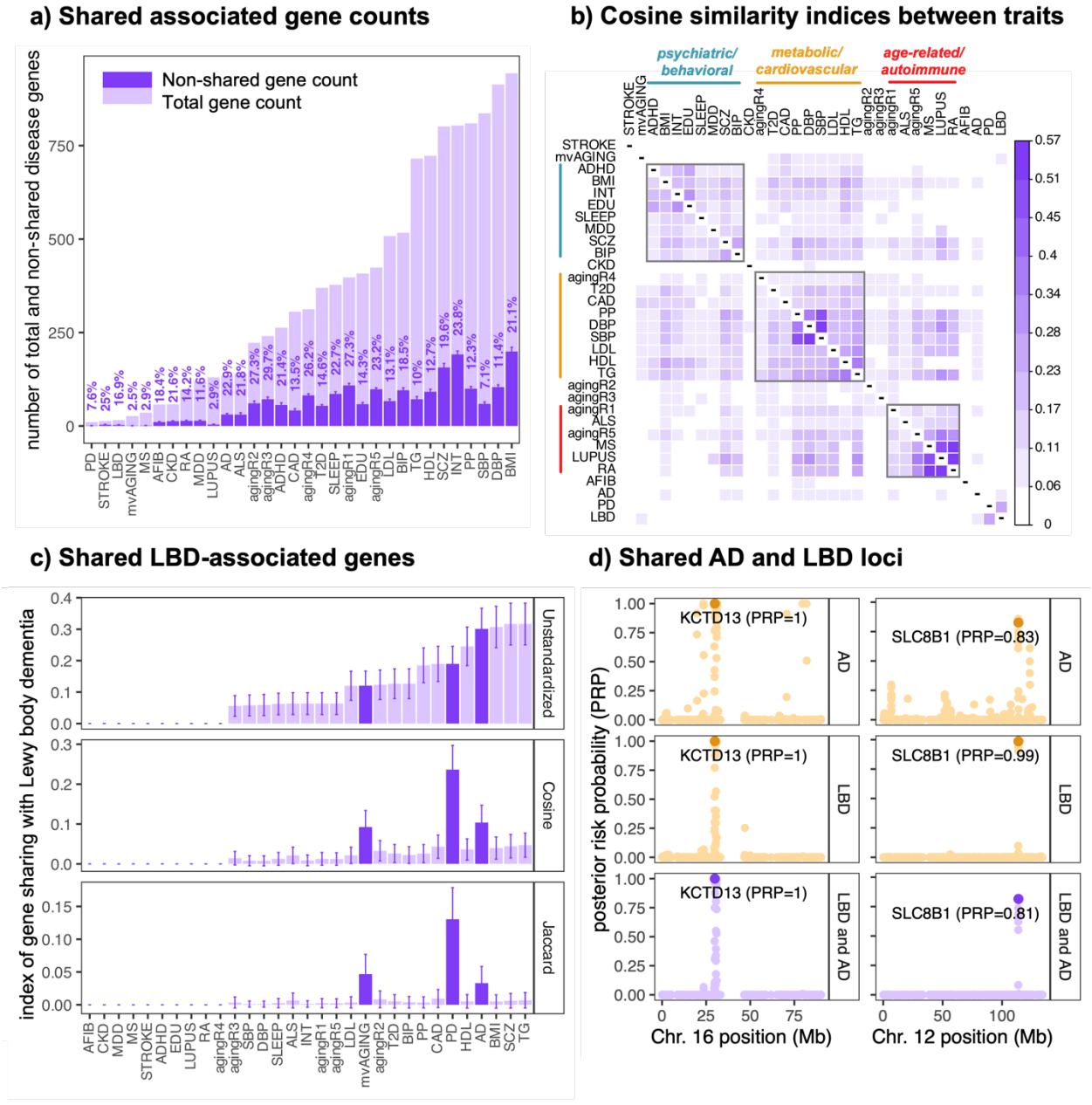
1110 **Figure 1: Estimated disease associated gene counts and example inference**



1111

(a): Estimated counts of disease associated genes for 32 traits and their relationship with the number of Bonferroni (left) and FDR (right) significant genes using gene-based test statistics scaled by GWAS sample size. **(b)** Example of AD posterior risk probabilities (PRPs) for all tested genes on chromosome 1. **(c)** Estimated counts of disease associated genes on each chromosome from each of 32 traits calculated by using the prior distributions of δ (the empirically derived prior proportion of non-disease associated genes) for each trait. Different traits are represented by different colors which are separated by horizontal lines in each vertical bar. **(d)** Example of shared association on chromosome 7 for LDL and CAD. Probabilities in the 'CAD and LDL' panel are the products of LDL- and CAD-specific posterior risk probabilities and represent the posterior probability of shared association for each gene.

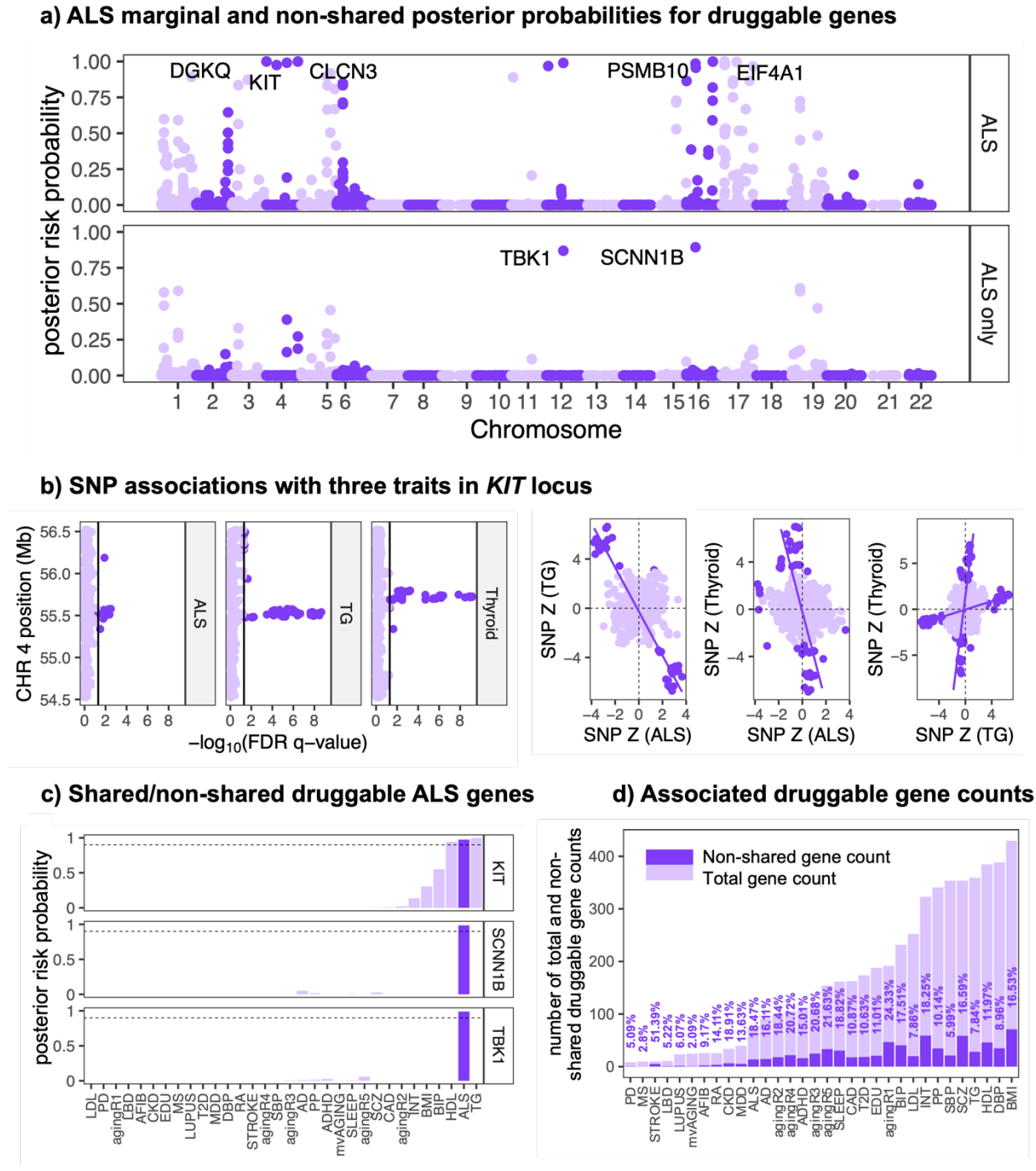
Figure 2: Measuring shared polygenicity across traits



(a): Bars display estimated counts of all disease associated genes (light purple) and non-shared disease associated genes (dark purple). Vertically oriented percentages above each bar indicate the percentage of all disease associated genes that the number of non-shared genes represent. **(b)** Cosine index values between all pairs of traits (*cf.* Methods). Boxes are drawn around select sets of traits heuristically, and their

labels assigned manually. **(c)** Estimated proportions of all LBD-associated genes which are shared with each other trait and their corresponding cosine and Jaccard index values (*cf.* Methods). **(d)** Examples of two loci with evidence of shared association for LBD and AD. 'LBD and AD' represents the posterior probability that each gene is associated with both LBD and AD risk.

1135 **Figure 3: Shared association of ALS druggable genes with other traits**



1136

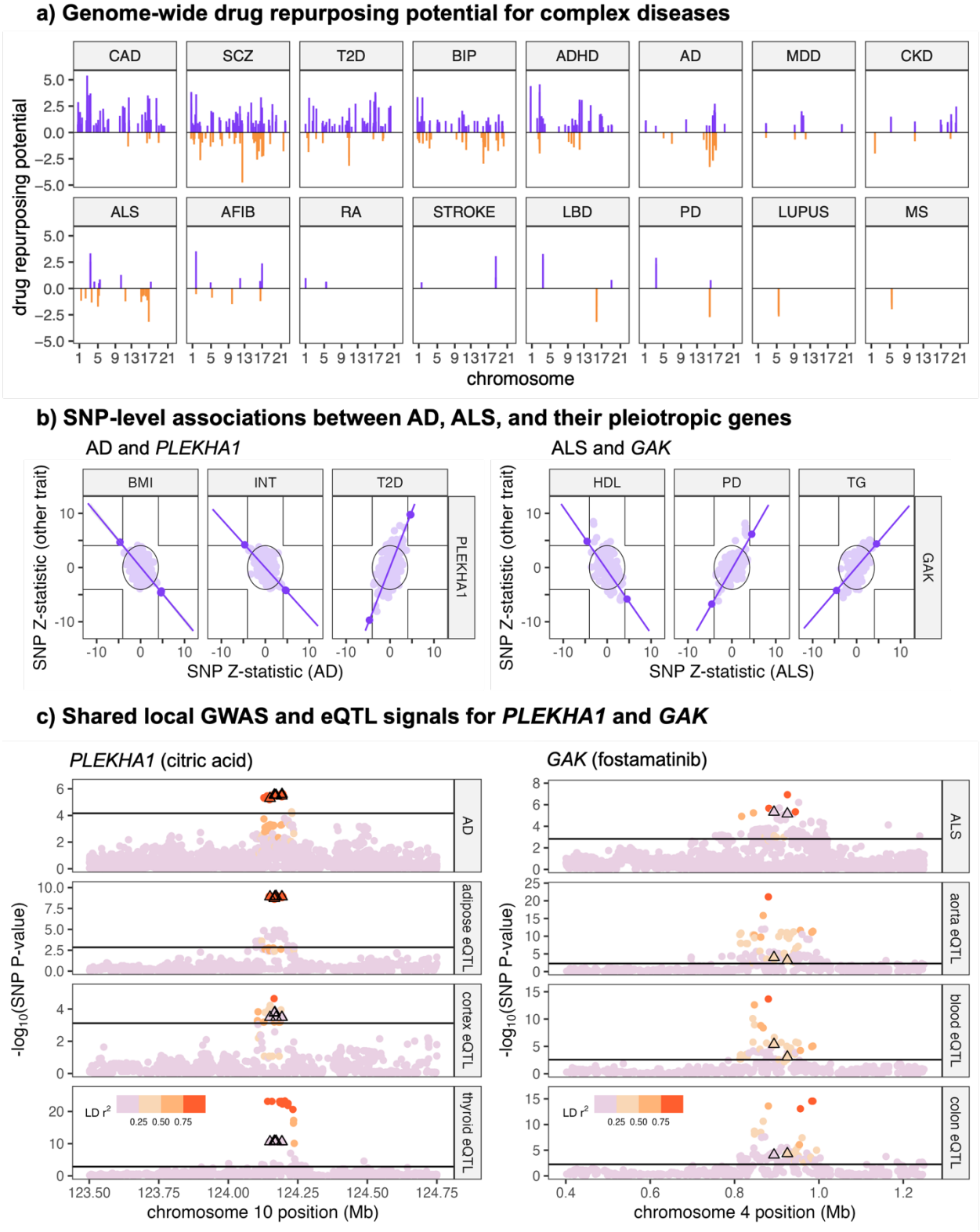
1137 **(a)** Genome-wide plot of posterior risk probabilities (PRPs) for ALS (top) for each gene

1138 and posterior probabilities that each gene is associated with ALS but not with any of the

1139 other 31 studied traits (bottom). **(b)** (left) SNP-level association estimates from GWAS

for ALS, TG, and gene expression in the in the *KIT* locus (Chr4: 54524085-56606881). (right) Bivariate SNP associations between ALS, TG, and thyroid eQTLs in the *KIT* locus. Dark purple points are FDR significant at the 5% level for either trait on the x- or y-axis. Lines are of best fit through sets of FDR-significant SNPs. **(c)** posterior risk probabilities of three select genes for association with each of the 32 studied traits. Vertical bars corresponding to ALS are colored in dark purple; bars not corresponding to ALS are colored in light purple. **(d)** Bars display estimated counts of all druggable disease associated genes (light purple) and non-shared druggable disease associated genes (dark purple). Vertically oriented percentages above each bar indicate the percentage of all druggable disease associated genes that the number of non-shared druggable disease associated genes represent.

Figure 4: Drug repurposing targets for disease traits



(a) Drug repurposing potential values (*cf.* Equation 9) for the 16 complex diseases we studied and each tested druggable gene. Positive values denoted by purple bars are candidate repurposable targets. **(b)** Examples of candidate repurposable drug targets for AD with *PLEKHA1* (left) and ALS with *GAK* (right). Displayed are Z-statistics for association at the SNP level from each respective GWAS for the indicated traits in the $\pm$100Kb window around each gene body in hg19 coordinates. Dark purple points are significant at P<5E-5 in an intersection union test for the pair of traits on the x- and y-axes. Dark purple lines correspond to the best linear fit through these points. The null region of the IUT at the 5E-5 level is indicated by the interior of joined vertical and horizontal lines; the null region of a joint test at the 5E-5 level is indicated by the interior of the circle centered at the origin. **(c)** Local SNP associations between *PLEKHA1*, AD, and select tissue contexts of gene expression (left) and between *GAK*, ALS, and select tissue contexts of gene expression (right). The color of each point represents its squared LD with the lead SNP, which was estimated using the 1KGv3-EUR reference panel (Siva et al., 2008). SNPs represented by triangles correspond to the SNPs which are highlighted in dark purple in panel **(b)**.