medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Co-expression-wide association studies implicate protein–protein interactions in complex disease risk

Mykhaylo M. Malakhov[1] and Wei Pan[1*]

[1]Division of Biostatistics and Health Data Science, School of Public Health, University of Minnesota, Minneapolis, MN, USA.

*Corresponding author(s). E-mail(s): panxx014@umn.edu;
Contributing authors: malak039@umn.edu;

Abstract

Transcriptome-wide association studies (TWAS) have proven successful in prioritizing genes and proteins whose genetically regulated expression modulates disease risk, but they ignore potential co-expression and interaction effects. Here we introduce the co-expression-wide association study (COWAS) method to identify pairs of co-expressed genes or proteins that are associated with complex traits. COWAS first trains models to predict co-expression conditional on genetic variation, and then tests for association between imputed co-expression and the trait while also accounting for direct effects from each exposure. We applied our method to plasma proteomic concentrations from the UK Biobank, identifying dozens of interacting protein pairs associated with cholesterol levels, Alzheimer's disease, and Parkinson's disease. Notably, our results demonstrate that co-expression between proteins may affect complex traits even if neither protein is detected to influence the trait when considered on its own.

Introduction

Translating genetic associations into knowledge of causal genes and proteins is a central problem in genetic epidemiology. Although genome-wide association studies (GWAS) can rapidly identify the single nucleotide polymorphisms (SNPs) and genetic loci associated with any measurable phenotype, most of the significant GWAS hits for complex traits fall outside of protein-coding regions and are thought to affect the phenome through regulatory pathways [1–6]. A popular approach for aggregating these

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

regulatory effects into interpretable gene-level functional units is the transcriptome-wide association study (TWAS) method [7, 8]. TWAS is a two-stage framework that first trains a model to predict gene expression levels from genetic variation, thereby estimating the genetically regulated component of expression, and then tests for association between imputed expression and the trait of interest. Although most commonly applied to gene expression data, TWAS can be used with any heritable molecular phenotype. For example, proteome-wide association studies (PWAS) identify disease-relevant proteins by applying the two-stage TWAS framework to proteomic concentrations [9–11].

Many innovative methodological extensions to TWAS and PWAS have been developed since their initial introductions [12–19], with applications spanning hundreds of outcome traits [20–26]. All existing TWAS/PWAS methods, however, have a major limitation: they fail to account for correlations or interactions among the functional units being studied. In standard TWAS approaches, each gene or protein is considered independently of the rest. This marginal assumption is mathematically simple and provides for a straightforward implementation of the method, but it is biologically implausible. Moreover, discounting interaction effects in TWAS may lead to a loss of statistical power and missed biological insights when considering molecular drivers that primarily affect complex traits through synergistic pathways.

Recent methods have partially addressed the marginal limitation in TWAS by fine-mapping candidate TWAS genes to separate the effects of multiple correlated exposures [27–29]. These methods can tease out the likely causal genes within a larger set of co-expressed genes by conditioning each gene on the others. However, they do not model the genetic regulation of co-expression and cannot be used to infer the impact of gene–gene or protein–protein interactions on the outcome trait. In a separate line of research, protein–protein interaction (PPI) networks have been used to aid in the interpretation of PWAS findings [30]. Such use of PPI networks, however, still relies on the results of testing each protein individually for association with disease, and only utilizes evidence of interactions to cluster those marginal associations. Thus, no existing approaches are able to elucidate the extent to which co-expression and interactions among molecular phenotypes mediate genetic effects on complex traits.

The importance of epistasis, co-expression, and PPIs in complex disease pathogenesis has been well established and is the subject of extensive research despite the challenges of ascertaining interaction effects from genomic data [31–34]. An increasing burden of evidence also highlights the role of genetic variation in regulating gene–gene and protein–protein interactions. For example, single-cell RNA sequencing data has enabled the detection of genetic variants that significantly alter co-expression relationships [35]. More recently, a pan-cancer study demonstrated that point mutations correlate with altered, tumor-specific PPIs and can rewire interaction networks [36]. Other work used gene co-expression networks to link cancer driver genes to cancer GWAS genes, showing that common genetic variants are involved in the regulation of co-expression networks [37]. More generally, large-scale sequencing studies have established that both germline and somatic mutations are responsible for widespread perturbations in PPI networks in human diseases [38]. Such evidence suggests that it

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

should be possible to predict the effects of genetic variation on gene or protein co-expression, and to consequently assess the association between genetically regulated co-expression and disease.

In this paper we introduce the co-expression-wide association study (COWAS) method to identify co-expressed genes or proteins that are associated with complex traits. COWAS analyzes pairs of co-expressed molecular exposures, first imputing their genetically regulated expression and co-expression, and then jointly testing for both direct effects and interaction effects on the outcome trait. We also extend COWAS to a summary statistics setting, making it easy to apply our method to any trait of interest for which GWAS summary-level data are available.

We applied COWAS to plasma proteomic concentrations from the UK Biobank (UKB) [39, 40] and large GWAS datasets for three complex traits [41–43]. We first trained imputation models for pairs of proteins with known PPIs, and then tested each well-imputed pair for association with low-density lipoprotein (LDL) cholesterol, Alzheimer's disease (AD), and Parkinson's disease (PD). Our results demonstrate that COWAS can successfully identify protein pairs whose co-expression impacts complex traits while at the same time disentangling their direct and interaction effects. Our approach also increases power relative to standard PWAS analyses, leading to the discovery of proteins that were missed by PWAS. Notably, we show that co-expression between proteins may affect disease risk even if neither protein influences the disease when considered on its own. Overall, our contribution provides a novel framework for studying the effects of genetically regulated co-expression on complex traits, facilitating interrogation of the phenotypic consequences of gene–gene and protein–protein interactions using GWAS summary statistics.

Results

Overview of COWAS

The co-expression-wide association study (COWAS) method prioritizes pairs of interacting genes or proteins whose genetically regulated expression or co-expression is significantly associated with a complex trait. Note that COWAS can be applied to either gene expression or protein expression data, but since our application concerns the proteome, we will primarily refer to protein expression throughout the rest of the paper.

The key motivation behind our approach is the observation that genetic variation modulates not only protein expression, but also protein co-expression (Fig. 1a). We refer to genetic variants associated with co-expression as co-expression quantitative trait loci (coQTLs) [35], analogously to how variants associated with gene expression are termed expression quantitative trait loci (eQTLs) and variants associated with protein expression are termed protein quantitative trait loci (pQTLs). A variant can belong to one or more of these xQTL classes, but we assume that a coQTL is most likely also an eQTL or a pQTL. Furthermore, we consider co-expression to be a proxy for interaction effects. Although gene–gene and protein–protein interactions are not directly measured in large biobank studies such as the UKB, co-variation of protein abundance is an accurate proxy for PPIs because interacting protein pairs are known

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

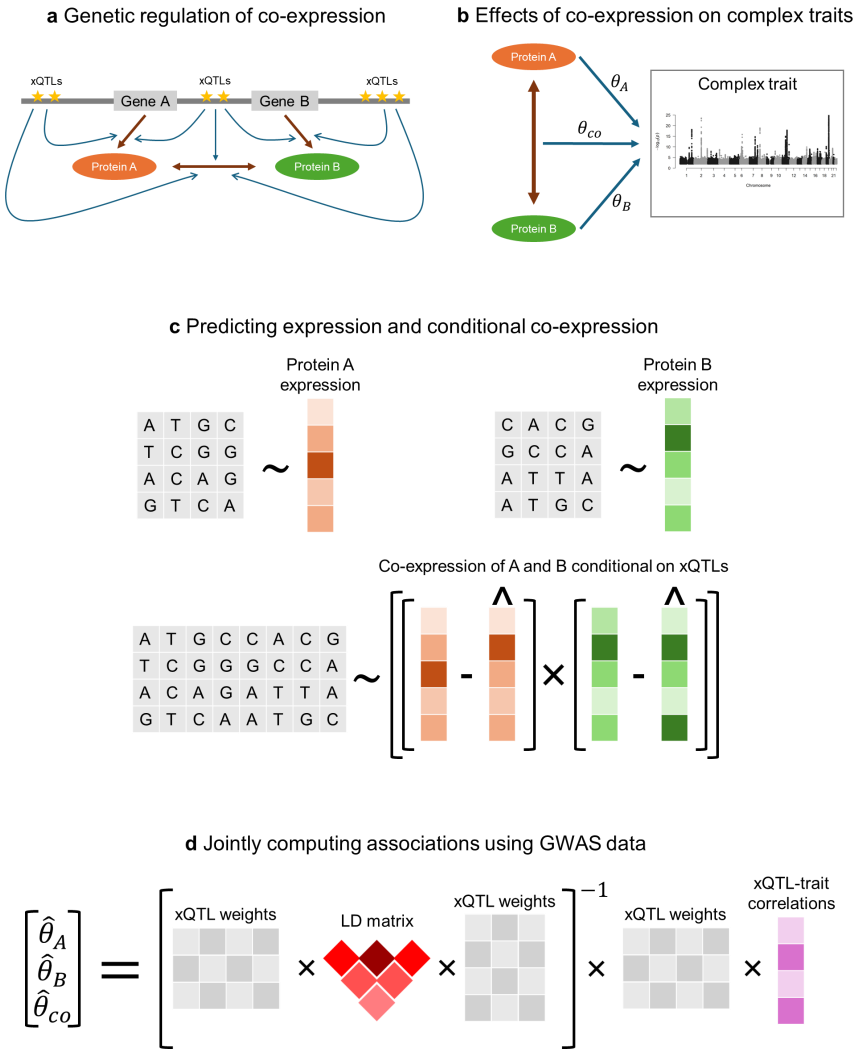

Fig. 1 Overview of the COWAS framework.
a, Genes A and B code for proteins A and B, which interact with each other. The transcription, translation, and interaction processes are regulated by eQTLs, pQTLs, and coQTLs, respectively, which may overlap and are collectively denoted as xQTLs. **b**, Proteins A and B may have direct effects on a complex trait (θ_A and θ_B, respectively), but they may also impact the trait through their interactions with each other (θ_{co}). **c**, The training stage of COWAS involves first building models to impute the expression levels of each protein from pQTL genotypes, then using the imputed expression levels to estimate the conditional co-expression between the two proteins, and finally building a third model to impute their estimated conditional co-expression. **d**, The testing stage of COWAS involves jointly estimating direct and interaction effects on a complex trait of interest using the fitted model weights from the training stage, an LD reference panel, and GWAS summary statistics for the outcome trait.

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

to be highly co-expressed [36, 44]. COWAS leverages pQTL data to learn the patterns of genetic regulation underlying protein expression and co-expression, and ultimately estimates the direct and interaction effects of genetically regulated expression on a complex trait of interest (Fig. 1b).

The COWAS framework is comprised of a training stage (Fig. 1c) and a testing stage (Fig. 1d). The training stage must be performed on individual-level genotype and expression data. First, models are trained to predict the expression levels of each protein from its pQTLs. Next, the measured and imputed expression levels are used to estimate the co-expression of the two proteins conditional on genetic information, which we define in terms of conditional correlation. COWAS exploits the properties of conditional covariance to remove the components of co-expression that are explained by genetic effects on mean expression levels or by factors unrelated to genetics, allowing us to focus on how genetic variation modulates the amount of correlation between the two exposures. Finally, a third model is trained to predict estimated conditional co-expression from the union of all considered pQTLs. Explicitly modeling the conditional correlation of expression is the primary innovation of COWAS, because it enables our approach to incorporate the genetic component of gene or protein co-expression into an association testing framework.

The testing stage of COWAS is typically performed using fitted model weights from the training stage, a linkage disequilibrium (LD) reference panel, and summary-level GWAS data for the outcome trait of interest (Fig. 1d). Here three effect sizes are jointly estimated: the direct effect of the first protein's genetically regulated expression on the trait (θ_A), the direct effect of the second protein's genetically regulated expression on the trait (θ_B), and the effect of their genetically regulated co-expression on the trait (θ_{co}). Note that θ_A and θ_B are distinct from the marginal effects obtained through standard TWAS or PWAS, since here the three effect sizes are estimated together in a joint model. As a result, each effect size is conditional on the other two.

Several hypothesis tests can be performed with these estimated effect sizes and their standard errors. The COWAS global test determines if the protein pair has an overall effect on the outcome trait, potentially boosting power relative to marginal TWAS/PWAS analyses of each exposure. Alternatively, we can test the effect size estimates individually in order to disentangle the impact of each protein's genetically regulated expression from the impact of their genetically regulated co-expression. In particular, the COWAS interaction test determines if co-expression has an effect on the outcome trait while accounting for direct effects from each exposure. This flexibility and increased statistical power enable COWAS to identify novel disease-relevant genes or proteins and aid in the interpretation of GWAS findings.

Accurately imputing genetically regulated co-expression

We trained COWAS models to predict protein expression and co-expression using genotypes and proteomic concentrations from the UKB Pharma Proteomics Project [40]. After quality control, we retained 2,833 proteins coded by autosomal genes. Since training imputation models for each of the $\binom{2,833}{2} = 4,011,528$ possible protein pairs would have been computationally infeasible, we restricted our analysis to pairs with

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

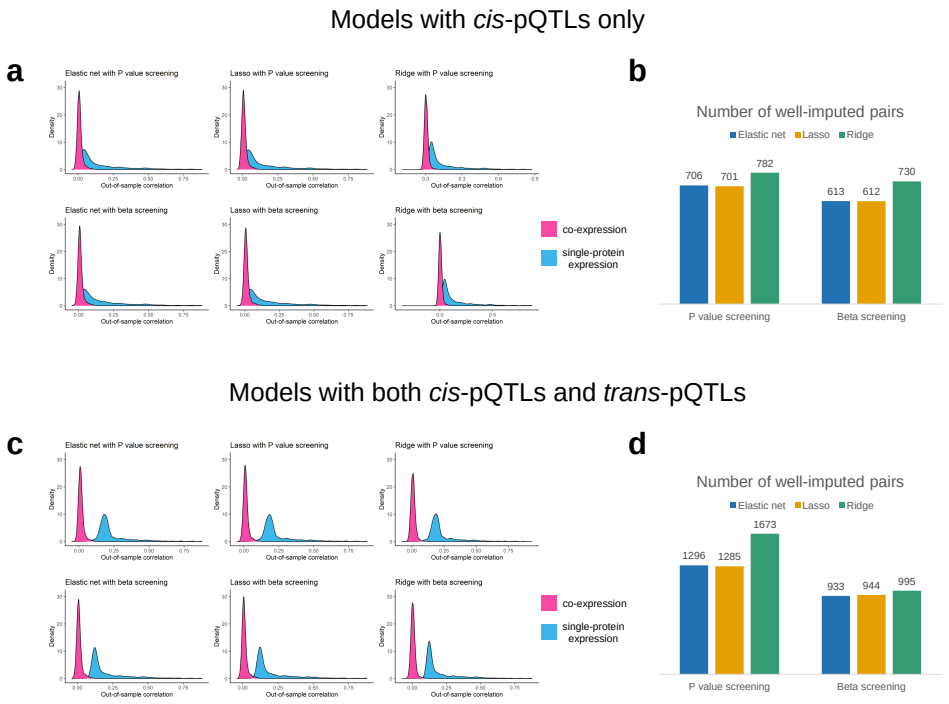

Fig. 2 Performance metrics for COWAS models trained on UKB data.
a,**c**, Density plots of the correlation between measured and imputed expression, as well as estimated and imputed co-expression, on a held-out test set. **b**,**d**, Counts of the numbers of protein pairs in which all three prediction models had an out-of-sample correlation greater than 0.03. Models in **a** and **b** were trained with only *cis*-pQTLs as predictors. Models in **c** and **d** were trained with both *cis*-pQTLs and *trans*-pQTLs as predictors.

some prior evidence of PPIs, as listed in the Human Integrated Protein–Protein Interaction rEference (HIPPIE) database [45]. In total, we trained COWAS models using UKB genotypes and normalized protein abundance residuals for 26,433 protein pairs.

To ensure that COWAS can accurately predict genetically regulated co-expression, we explored the out-of-sample imputation performance of several regression methods (Fig. 2). We considered penalized linear regression models with either an elastic net penalty, a lasso penalty, or a ridge penalty. For each of these three model types, we pre-screened genetic variants using either the P values or the effect sizes of their association with each protein's expression. Additionally, we also considered the extent to which including both local pQTLs (*cis*-pQTLs) and distant pQTLs (*trans*-pQTLs) improved model imputation performance relative to only including *cis*-pQTLs.

Our results show that accurate imputation is more challenging for protein co-expression than for the expression of individual proteins. Across all of the model types we considered, the median out-of-sample correlation between estimated and imputed co-expression was always lower than between measured and imputed single-protein

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

expression (Figs. 2a,c and Supplementary Data 1). This was expected, since interaction effects are known to be more difficult to detect than main effects, with considerably larger sample sizes being needed for the same level of power or prediction quality. Interestingly, including *trans*-pQTLs in addition to *cis*-pQTLs significantly increased the imputation quality for single-protein models, but it did not have a pronounced effect on the performance of co-expression models (Figs. 2a,c). This suggests that *trans*-pQTLs only weakly regulate PPIs, with the bulk of heritability in co-expression stemming from local genetic variation. However, it is also possible that *trans*-coQTLs may not overlap with *trans*-pQTLs. Since we pre-screened genetic variants based on the strength of their association with the individual proteins in each pair, the inclusion of distant variants primarily increases the number of strong pQTLs present in each model and may not necessarily increase the number of strong coQTLs.

Next, we filtered the protein pairs to those in which all three imputation models yielded an out-of-sample correlation greater than 0.03 (Fig. 2b,d). Among these well-imputed pairs, lasso regression with *cis*-pQTLs pre-screened by their effect sizes achieved the highest mean out-of-sample R^2 for predicting co-expression (mean $R^2 = 0.0038$, Supplementary Data 1). On the other hand, ridge regression with both *cis*-pQTLs and *trans*-pQTLs pre-screened by their P values yielded the greatest number of well-imputed protein pairs (Fig. 2d and Supplementary Data 1). We decided to use the former approach in our main analyses in order to maximize the imputation quality of conditional co-expression. Model performance metrics for every combination of protein pair and model type are provided in Supplementary Data 2-13.

COWAS identifies co-expressed proteins associated with complex traits

Having shown that COWAS is able to accurately impute both single-protein expression and protein co-expression, we applied it to three complex trait outcomes: low-density lipoprotein (LDL) cholesterol, Alzheimer's disease (AD), and Parkinson's disease (PD). For each trait, we downloaded summary-level data from the largest publicly available GWAS study [41–43]. To ensure complete overlap between the genetic variants included in the imputation models and the GWAS data, we re-trained COWAS models for each trait using only the intersection of variants found in both the UKB genotype data and the trait's GWAS. We also re-assessed the out-of-sample predictive performance of each model separately for each trait and only kept pairs with sufficiently high imputation accuracy, thus guaranteeing that differences between the GWAS datasets do not negatively impact the validity of association testing. As a result, the numbers of considered protein pairs somewhat differed among the three traits. For LDL cholesterol 613 pairs were accurately imputed, for AD there were only 564 well-imputed pairs, and for PD we retained 592 well-imputed pairs (Supplementary Data 14-16).

To compare our new approach with currently available methods, we also performed a standard PWAS analysis for each protein included in the COWAS analyses. The same training samples, model types, and variant screening strategies were applied for both COWAS and PWAS. Namely, we selected the top 100 pQTLs for each protein by their association effect sizes and used them as features in linear regression models with

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

a lasso penalty. We also used the same LD reference panel derived from UKB data when computing effect sizes in both COWAS and PWAS. Full imputation performance metrics for all analyzed proteins and outcome traits are provided in Supplementary Data 14-16.

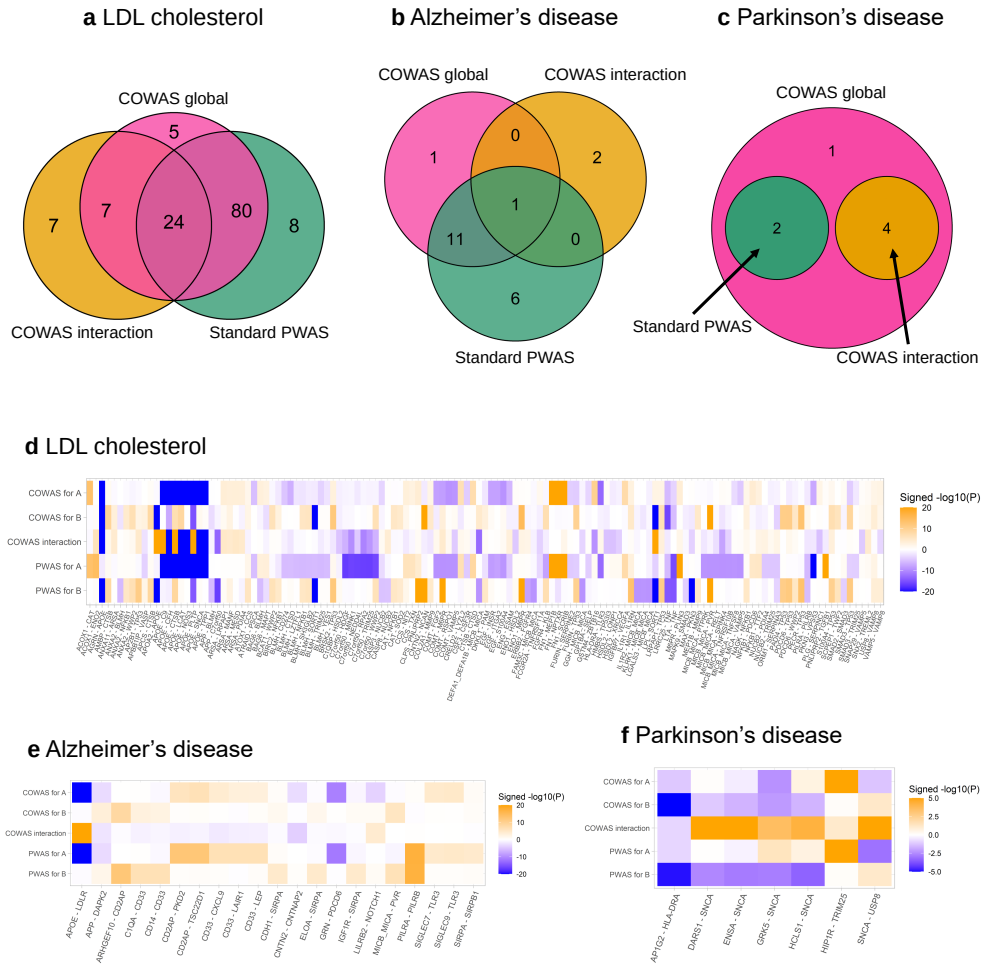

Fig. 3 COWAS and PWAS results for three complex traits.
a-c, Venn diagrams displaying the numbers of protein pairs identified by the COWAS global test, the COWAS interaction test, and a standard PWAS analysis. Here "standard PWAS" refers to pairs in which at least one of the proteins is identified by PWAS. Statistical significance was assessed at the 95% confidence level, with Bonferroni multiple testing corrections for the number of protein pairs (in COWAS) or the number of unique proteins (in PWAS). **d-f**, Heatmaps displaying signed $-\log_{10}(P)$ values from COWAS single-protein and interaction tests as well as from standard PWAS analyses for all pairs included in the Venn diagrams. To facilitate visualization, the $-\log_{10}(P)$ values were capped at 20 for LDL cholesterol and AD, and at 5 for PD. A and B refer to the first and second proteins listed in each pair, respectively.

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Our results demonstrate that COWAS is able to detect PPIs with a significant genetically regulated effect on complex traits (Fig. 3). We identified 38 protein pairs whose co-expression has a significant effect on LDL cholesterol levels after accounting for the direct effects of each protein and adjusting for multiple testing (Fig. 3a). Of these protein pairs, 24 had at least one protein that was also identified by a standard PWAS analysis, while the rest were uniquely identified by our method. We also performed a global test on each pair to assess whether it has an overall effect on LDL cholesterol, which yielded 116 significant pairs after adjusting for multiple testing. As expected, nearly all of those pairs contained at least one protein that was also detected by PWAS. However, the COWAS global test did identify 12 pairs with a significant effect on LDL cholesterol in which neither protein was significant when considered on its own, and 5 of those pairs did not even have a significant interaction term (Fig. 3a). This suggests that explicitly modeling co-expression can boost power relative to standard marginal tests, even when there is no statistically significant effect of co-expression on the outcome trait.

Interestingly, the effect of co-expression on a complex trait can have an opposite direction relative to the effects of the interacting proteins themselves. For example, we found that APOE and PLTP both decrease LDL cholesterol levels, while their co-expression was associated with increased LDL cholesterol levels (3d). In other cases, the direct and interaction effects may all be in the same direction, such as observed for the effects of APOE and AGRN on LDL cholesterol. This illustrates the potential of COWAS to help disentangle the effects of interacting proteins on complex traits, thereby providing a richer picture of the functional consequences of molecular phenotypes. Full results for LDL cholesterol, including the estimated effect sizes and standard errors within each pair, are provided in Supplementary Data 14.

COWAS boosts power and corroborates known PPIs driving Alzheimer's disease risk

We identified fewer significant protein pairs for AD compared to LDL cholesterol, but this was expected due to the lower power of the corresponding GWAS study. Yet here again, our approach was able to detect significant protein pairs missed by standard PWAS (Fig. 3b and Supplementary Data 15). Notably, the COWAS global test identified the pair comprised of amyloid-beta precursor protein (APP) and death-associated protein kinase 2 (DAPK2) as significant for AD (P = 1.25e-05), while a standard PWAS analysis failed to identify either of these proteins (P = 7.72e-04 for APP and P = 1.31e-02 for DAPK2, Fig. 3e and Supplementary Data 15). APP is concentrated in the synapses of neurons and is the precursor molecule for the generation of amyloid beta (Aβ), which contributes to the formation of amyloid plaques—a hallmark pathology in AD [46–48]. Yet despite the central role of APP in Alzheimer's pathogenesis, standard PWAS lacked the power to identify it in our dataset. On the other hand, COWAS was able to boost power and attain statistical significance by jointly considering APP and a member of the DAPK family, which has also been previously implicated in late-onset AD [49].

Furthermore, COWAS discovered a highly significant effect of the interaction between APOE and LDLR on AD risk (P < 1e-50). Although APOE was also highly

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

significant according to a standard PWAS analysis (P < 1e-50), LDLR was not (P = 0.48). This result is notable, because LDLR is known to be a receptor for APOE that preferentially binds lipidated APOE particles and plays an important role in Aβ clearance [50]. Our results are consistent with this mechanistic explanation, since we found APOE and its interaction with LDLR to have opposite effects on AD (Fig. 3e). Thus, COWAS provides strong support to the hypothesis that APOE and LDLR have a synergistic effect in Alzheimer's pathogenesis, even after accounting for the direct effect of APOE on AD risk.

The other two significant interactions implicated by COWAS for AD are also likely true positives, further confirming the sensitivity and power of our approach. We identified a significant effect of the interaction between LILRB2 and NOTCH1 on AD (P = 5.73e-05), whereas standard PWAS failed to identify either protein (P = 0.80 for LILRB2 and P = 0.13 for NOTCH1). LILRB2 is a neuronal cell surface receptor that interacts with Aβ and is being studied as a promising therapeutic target for AD [51, 52], while NOTCH1 has been found to be differentially expressed in Alzheimer's patients [53] and is potentially involved in neurodegeneration-related cell signaling disruptions [54]. Finally, the COWAS interaction test also discovered a significant effect of co-expression between CNTN2 and CNTNAP2 on AD (P = 8.42e-05), whereas standard PWAS again failed to detect either protein as significant (P = 0.94 and P = 0.38, respectively). The mechanisms by which these proteins are involved in Alzheimer's pathology have not yet been thoroughly studied, but earlier genetic and functional genomic evidence indicates that they do play a role [55].

Co-expression analysis identifies SNCA interactions in Parkinson's disease pathogenesis

For PD the COWAS global test identified all of the protein pairs that were also discovered by the COWAS interaction test or by a standard PWAS analysis (Fig. 3c). In addition to those pairs, the COWAS global test also uniquely identified an effect of GRK5 and SNCA on the risk of Parkinson's (P = 4.81e-06). Note that both of these proteins have been previously implicated in PD pathogenesis. Alpha-synuclein (SNCA) is a protein that regulates the release of neurotransmitters from the axon terminals of presynaptic neurons, and insoluble forms of SNCA accumulate in the form of Lewy bodies, leading to nerve cell death and the development of PD symptoms [56–58]. As for GRK5, some evidence suggests that it plays a role in the pathogenesis of sporadic forms of Parkinson's [59]. These results further highlight the ability of COWAS to boost power relative to marginal approaches such as PWAS.

Interestingly, all four of the significant co-expression effects on PD that were identified by COWAS are comprised of SNCA interacting with some other protein (Fig. 3f and Supplementary Data 16). In particular, the COWAS interaction test identified significant effects on Parkinson's from genetically regulated co-expression between SNCA and DARS1 (P = 1.21e-13), SNCA and ENSA (P = 1.03e-08), SNCA and HCLS1 (P = 5.50e-05), and SNCA and USP8 (P = 3.64e-07). Note that co-expression between SNCA and each of these four proteins has a positive effect on PD even though the effect of SNCA itself is negative (Fig. 3f). This suggests that a genetically regulated

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

escalation of co-expression between SNCA and each of these proteins elevates PD risk, illustrating potential avenues for therapeutic intervention.

None of the four proteins whose interaction with SNCA had an effect on PD were significant according to a standard PWAS analysis, with marginal PWAS P values ranging from $P = 0.88$ to $P = 0.07$ (Supplementary Data 16). However, the COWAS discoveries are reasonable in light of previous research. For example, USP8 is a deubiquitinase that has also been found in Lewy bodies and plays a role in determining SNCA levels [60, 61]. ENSA has been shown to interfere with SNCA self-assembly and thereby alleviate its neurotoxicity [62], and variants in HCLS1 binding protein 3 were found to be associated with the related condition of essential tremor (but not PD itself) [63]. We are not aware of any existing evidence for the role of DARS1 in Parkinson's, but its identification by COWAS points to a potential avenue for further research.

Discussion

In this paper we introduced the co-expression-wide association study (COWAS) method, the first statistical framework for identifying gene or protein pairs whose genetically regulated interactions are associated with complex traits. COWAS extends the two-stage least squares approach underlying TWAS/PWAS by explicitly estimating and imputing the conditional correlation between pairs of exposures, which we interpret as a proxy for genetically regulated gene–gene or protein–protein interactions. This enables COWAS to jointly test for direct and interaction effects of genetically regulated expression on a complex trait of interest, thereby boosting power relative to existing methods and helping to disentangle the functional mechanisms by which molecular exposures influence the outcome trait. We also extended COWAS to a summary statistics setting, making it easy to apply our method to any trait for which GWAS summary data are available.

In our application of COWAS to the UKB Pharma Proteomics Project dataset, we first explored the performance of different regression models for imputing genetically regulated co-expression and then applied our method to identify protein pairs associated with three complex traits. Our method was able to discover biologically relevant co-expressed proteins for all three traits, highlighting the importance of interaction effects in driving complex disease risk. Notably, COWAS identified a number of protein pairs with a significant interaction term in which neither protein had a significant effect when analyzed independently via standard PWAS. These results underscore the importance of considering interaction effects in future research, since the marginal TWAS/PWAS approaches currently used to analyze molecular phenotypes may be missing important sources of signal. Moreover, our results demonstrate that the COWAS global test is able to identify more disease-relevant protein pairs than methods that consider one protein at a time, even in the absence of significant interaction effects and despite the better imputation quality of single-protein models.

Notwithstanding the many advantages of COWAS, our approach has several limitations. First of all, COWAS only considers one pair of molecular units at a time.

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Although this assumption is more realistic than the single-exposure setting of existing methods, it does not reflect the full range of possibilities. Proteins may interact in larger, multi-protein interaction networks with complex topological structures [64–66]. Extending COWAS to allow for interactions among more than two exposures at a time could illuminate even more disease-relevant genes and proteins, but it is not obvious how to do so in a computationally efficient way. Furthermore, we found that the predictive capacity of protein co-expression imputation models is lower than that of expression imputation models for individual proteins. This was expected given the difficulty of ascertaining interaction effects in general, yet even so we were able to obtain sufficiently good imputation quality for over a thousand protein pairs. However, more work could be done to explore different machine learning algorithms for training co-expression imputation models. Finally, we only considered individuals of a single genetic ancestry in this study. Since transcriptome and proteome imputation models are not portable across ancestry groups [67–69], we subset the UKB data to the largest genetically-inferred ancestry subgroup, which roughly corresponds to White British individuals, and correspondingly used GWAS studies conducted on European individuals for our three outcome traits. An extension of COWAS to handle multiple genetic ancestries and admixed individuals would expand the diversity and relevance of its applications.

The field of human genetics has historically focused on studying linear, marginal effects. This is exemplified by the popularity of GWAS and TWAS/PWAS analyses, which only consider one genetic variant or one functional molecular unit at a time. By providing a simple yet powerful approach for analyzing genetically regulated gene or protein co-expression using existing biobank data, our work joins the growing body of evidence emphasizing the limitations of this historical paradigm. The COWAS method exhibits high statistical power, provides flexibility in modeling direct and interaction effects, and is easy to use. We envision that COWAS, along with its future improvements and extensions, will enhance the interpretation of genomic findings and lead to the discovery of new biological insights and therapeutic targets.

Methods

Modeling genetically regulated co-expression

The co-expression-wide association study (COWAS) method is applied to one outcome trait and two molecular exposures at a time. Let $A, B \in \mathbb{R}^n$ denote the expression or abundance levels of the two exposures, as measured in n individuals. Further, let $\boldsymbol{Z_A} \in \mathbb{R}^{n \times p_A}$ be the genotype matrix of p_A xQTLs for exposure A, which are genotyped in the same set of individuals. Similarly, let $\boldsymbol{Z_B} \in \mathbb{R}^{n \times p_B}$ be the genotype matrix of p_B xQTLs for exposure B, and let $\boldsymbol{Z} \in \mathbb{R}^{n \times p}$ be the joint matrix of all p xQTLs, where p is the number of unique variants in the union of xQTLs for the two exposures. (If there is no overlap among the xQTLs for the two exposures, then $p = p_A + p_B$.) Finally, let $Y \in \mathbb{R}^n$ be the outcome trait of interest. All of these vectors and each column of these matrices are assumed to be centered around 0 and scaled to have a variance of 1.

Just like in standard TWAS or PWAS, we assume that the mean genetically regulated expression of each molecular exposure can be modeled as a linear combination

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

of its xQTL genotypes. That is,

$$A = \gamma_A + \boldsymbol{Z_A}\beta_A + \varepsilon_A, \tag{1}$$
$$B = \gamma_B + \boldsymbol{Z_B}\beta_B + \varepsilon_B. \tag{2}$$

Here $\beta_A \in \mathbb{R}^{p_A}$ and $\beta_B \in \mathbb{R}^{p_B}$ are unknown xQTL weights, while $\gamma_A \in \mathbb{R}$ and $\gamma_B \in \mathbb{R}$ are unknown intercepts. The error terms ε_A and ε_B are assumed to be normally distributed.

What sets COWAS apart from previous methods, however, is that we also model the genetically regulated co-expression of the two functional units instead of analyzing them independently of each other. The most popular metric for co-expression is the Pearson correlation between measured expression levels [70, 71]. Therefore, genetically regulated co-expression should be defined as the Pearson correlation conditional on genetic information. Formally, we define the genetically regulated co-expression of A and B as

$$Corr(A, B \mid \boldsymbol{Z_A}, \boldsymbol{Z_B}) = \frac{Cov(A, B \mid \boldsymbol{Z_A}, \boldsymbol{Z_B})}{\sqrt{Var(A \mid \boldsymbol{Z_A})Var(B \mid \boldsymbol{Z_B})}}, \tag{3}$$

where the conditional covariance between A and B is

$$Cov(A, B \mid \boldsymbol{Z_A}, \boldsymbol{Z_B}) = E\big((A - E(A \mid \boldsymbol{Z_A}))(B - E(B \mid \boldsymbol{Z_B})) \mid \boldsymbol{Z_A}, \boldsymbol{Z_B}\big). \tag{4}$$

To simplify estimation of this quantity, we make the assumption that $Var(A \mid \boldsymbol{Z_A})$ and $Var(B \mid \boldsymbol{Z_B})$ are both constant. In other words, we assume that genetic variants only regulate the mean of each molecular phenotype and their covariance, but not their individual variances. Although not exactly true from a biological perspective, this simplifying assumption is reasonable because any effects of genetic variation on the variance of protein concentrations are likely to be much smaller than the effects of genetic variation on the quantities we are considering.

With the above in mind, we can approximate the conditional correlation of A and B as

$$Corr(A, B \mid \boldsymbol{Z_A}, \boldsymbol{Z_B}) \propto Cov(A, B \mid \boldsymbol{Z_A}, \boldsymbol{Z_B}) \tag{5}$$
$$= E\big((A - E(A \mid \boldsymbol{Z_A}))(B - E(B \mid \boldsymbol{Z_B})) \mid \boldsymbol{Z_A}, \boldsymbol{Z_B}\big) \tag{6}$$
$$\approx E\big((A - \widehat{A})(B - \widehat{B}) \mid \boldsymbol{Z_A}, \boldsymbol{Z_B}\big), \tag{7}$$

where $\widehat{A} = \boldsymbol{Z_A}\widehat{\beta}_A$ and $\widehat{B} = \boldsymbol{Z_B}\widehat{\beta}_B$ are the genetically imputed expression levels of the two exposures. Thus, the appropriate way to model genetically regulated co-expression is

$$(A - \widehat{A})(B - \widehat{B}) = \gamma_{co} + \boldsymbol{Z}\beta_{co} + \varepsilon_{co}, \tag{8}$$

where $\beta_{co} \in \mathbb{R}^p$ is the unknown vector of coQTL weights and $\gamma_{co} \in \mathbb{R}$ is an unknown intercept. The random error term ε_{co} is again assumed to be normally distributed. For conciseness, let $C = (A - \widehat{A})(B - \widehat{B})$ and $\widehat{C} = \boldsymbol{Z}\widehat{\beta}_{co}$.

We can interpret $\widehat{\beta}_{co}$ as representing the effects of genetic information on the correlation between A and B. In general, correlation between A and B may be due to some combination of the following three factors:

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

1. Genetic effects on the mean expression levels of the two proteins. In other words, co-expression can be induced through correlation between the genetically regulated expression levels $E(A \mid \boldsymbol{Z_A})$ and $E(B \mid \boldsymbol{Z_B})$.
2. Genetic effects that modulate the correlation between A and B. That is, genetic variation can influence the level of correlation between ε_A and ε_B.
3. Factors unrelated to genetics. For example, a shared tissue environment or various other environmental effects may cause A and B to be correlated.

Our formulation of conditional covariance in Equation 7 effectively removes the first factor, so what remains may be some combination of the second and third factors. However, the effects of any factors unrelated to genetics should be constant with respect to genetic variation, and so they will be captured by the intercept term γ_{co}. Notice that we estimate this intercept term and then discard it before imputing $\widehat{C}$, thereby removing all environmental effects on co-expression. In the end, this procedure for estimating $\widehat{C}$ isolates the genetic component of co-expression.

Next, we estimate the effect of genetically regulated co-expression on the outcome trait (Y) while accounting for direct effects of A and B on Y. We assume that the outcome trait depends on a linear combination of the genetically regulated expression levels of both molecular exposures and their genetically regulated co-expression. Formally, our model for the outcome trait is

$$Y = \widehat{A}\theta_A + \widehat{B}\theta_B + \widehat{C}\theta_{co} + \varepsilon_Y, \tag{9}$$

where $\theta_A, \theta_B, \theta_{co} \in \mathbb{R}$ are unknown scalars and ε_Y is a normally distributed, independent error term. The ultimate goal of COWAS is to estimate θ_A, θ_B, and θ_{co} and to test each of them for statistical significance. We will also derive a global test to determine if the model in Equation 9 is significantly better than a null model.

Two-sample model estimation and hypothesis testing

In practice, the COWAS models are estimated in a two-sample setting akin to standard TWAS and PWAS. Suppose we have two nonoverlapping, individual-level datasets with sample sizes n_1 and n_2. Genotypes for all p xQTLs are available in both datasets, but the molecular exposures A, B are only measured on the n_1 samples in the first dataset, while the outcome trait Y is only measured on the n_2 samples in the second dataset.

In the model training stage, COWAS first trains models to estimate β_A and β_B using data from the first dataset. Then it imputes expression for each exposure on that same dataset. That is, we compute

$$\widehat{A} = \boldsymbol{Z_A}\widehat{\beta}_A, \tag{10}$$

$$\widehat{B} = \boldsymbol{Z_B}\widehat{\beta}_B, \tag{11}$$

using the same n_1 individuals used for model training. Next, COWAS computes the conditional correlation $C = (A - \widehat{A})(B - \widehat{B})$. These conditional correlation values are

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

then used as the outcome for training the model in Equation 8, yielding the fitted weights $\widehat{\beta}_{co}$.

In the testing stage, the fitted weights $\widehat{\beta}_A$, $\widehat{\beta}_B$, and $\widehat{\beta}_{co}$ are used to impute expression and co-expression for the n_2 samples in the second dataset. That is, we compute

$$\widehat{A}^* = \boldsymbol{Z}^*_{\boldsymbol{A}}\widehat{\beta}_A, \tag{12}$$

$$\widehat{B}^* = \boldsymbol{Z}^*_{\boldsymbol{B}}\widehat{\beta}_B, \tag{13}$$

$$\widehat{C}^* = \boldsymbol{Z}^*\widehat{\beta}_{co}, \tag{14}$$

where the $*$ symbol is used to distinguish quantities measured or imputed in the outcome dataset from those in the expression dataset. Finally, we fit the outcome trait model

$$Y = \widehat{A}^*\theta_A + \widehat{B}^*\theta_B + \widehat{C}^*\theta_{co} + \varepsilon_Y \tag{15}$$

to estimate each of its coefficients and their standard errors. Any linear model hypothesis tests can be performed on the estimated coefficients $\widehat{\theta}_A$, $\widehat{\theta}_B$, and $\widehat{\theta}_{co}$. In this study, we primarily consider an interaction test and a global test.

Interaction test: To determine if co-expression has an effect on the outcome trait, we test the hypothesis $H_0 : \theta_{co} = 0$ against its two-sided alternative using a Wald test. Namely, the test statistic is $w = \widehat{\theta}^2_{co}/Var(\widehat{\theta}_{co})$, which asymptotically follows a χ^2 distribution with 1 degree of freedom under H_0.

Global test: To determine if the two exposures have an overall effect on the outcome trait, we test whether the model in Equation 15 fits the data better than an intercept-only model using an F test. Namely, the test statistic is $f = \frac{RSS_{\text{null}}-RSS}{(n_2-1)-(n_2-4)} / \frac{RSS}{n_2-4} = \frac{n_2-4}{3RSS}(n_2-1-RSS)$, where RSS is the residual sum of squares from the COWAS model in Equation 15 and $RSS_{\text{null}} = n_2 - 1$ is the residual sum of squares from an intercept-only model. f follows an F distribution with $(3, n_2 - 4)$ degrees of freedom under the null hypothesis.

The interaction test and the global test can help to disentangle the effects of co-expression from the direct effects of the individual exposures. If the global test rejects its null hypothesis but the interaction test does not, we can conclude that the molecular exposures directly influence the outcome trait. Our implementation of COWAS in R also provides P values for Wald tests on θ_A and θ_B, enabling users to test whether each exposure has a significant effect while accounting for the other exposure and the co-expression term.

Extension of COWAS for use with GWAS summary data

Here we extend the association testing stage of COWAS for use with summary-level GWAS data. The formulas we derive here only require fitted weights for expression and co-expression imputation models, Z scores from a GWAS for the outcome trait, and an LD reference panel. Note that we used this summary-level version of COWAS to obtain all of the results reported in this paper.

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Let $\widehat{\beta}_A, \widehat{\beta}_B, \widehat{\beta}_{co} \in \mathbb{R}^p$ be the trained model weights for molecular phenotypes A, B, and their co-expression, respectively. We will denote the joint matrix of all model weights by $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_A, \widehat{\beta}_B, \widehat{\beta}_{co}) \in \mathbb{R}^{p \times 3}$.

Furthermore, let $z_1, \ldots, z_p \in \mathbb{R}$ be Z scores from a GWAS study for the outcome trait of interest (Y) for the same set of p genetic variants. We assume that the GWAS was conducted in a population of the same genetic ancestry as the population used to train COWAS model weights. Importantly, reference and effect alleles must be consistent between the GWAS summary data and the COWAS weights. Our implementation of COWAS automatically checks for allele consistency, flips GWAS Z scores when necessary, and removes variants which cannot be harmonized. Next, COWAS converts the GWAS Z scores to pseudocorrelation estimates. This is done by relying on the monotonic relationship between Z scores and correlations [72], leading to the following approximate formula for the pseudocorrelation between Y and variant i:

$$\widehat{c}_i = \frac{z_i}{\sqrt{n' - 1 + z_i^2}}. \tag{16}$$

Here z_i is the Z score for the effect of variant i on the outcome trait and n' is the sample size of the GWAS cohort for the outcome trait.

Finally, let $\boldsymbol{G} \in \mathbb{R}^{m \times p}$ be a genotype matrix for m individuals and the same set of p variants included in the COWAS models. We assume that these m individuals are of the same genetic ancestry as those used to train the COWAS model weights and conduct the outcome trait GWAS. Moreover, we assume that each column of $\boldsymbol{G}$ has been centered around 0 and scaled to a variance of 1. An LD reference panel represents correlations among genetic variants, and we compute it from $\boldsymbol{G}$ as

$$\widehat{\boldsymbol{D}} = \frac{1}{m} \boldsymbol{G}^{\mathsf{T}} \boldsymbol{G}. \tag{17}$$

Now we will derive an estimator for $\theta = (\theta_A, \theta_B, \theta_{co})^{\mathsf{T}}$ in terms of the trained COWAS model weights $\widehat{\boldsymbol{\beta}}$, the variant-outcome pseudocorrelation vector $\widehat{c} = (\widehat{c}_1, \ldots, \widehat{c}_p)^{\mathsf{T}}$, and the LD reference panel $\widehat{\boldsymbol{D}}$. Suppose that $\widehat{\boldsymbol{X}}^* = (\widehat{A}^*, \widehat{B}^*, \widehat{C}^*) \in \mathbb{R}^{n_2 \times 3}$ is the matrix of imputed expression and co-expression in an individual-level dataset for the outcome trait, so that Equation 15 can be rewritten as $Y = \widehat{\boldsymbol{X}}^* \theta + \varepsilon_Y$. This is a multiple linear regression model, and the ordinary least squares estimator of θ is

$$\widehat{\theta} = \left((\widehat{\boldsymbol{X}}^*)^{\mathsf{T}} \widehat{\boldsymbol{X}}^* \right)^{-1} (\widehat{\boldsymbol{X}}^*)^{\mathsf{T}} Y \tag{18}$$

$$= \left((\boldsymbol{Z}^* \widehat{\boldsymbol{\beta}})^{\mathsf{T}} \boldsymbol{Z}^* \widehat{\boldsymbol{\beta}} \right)^{-1} (\boldsymbol{Z}^* \widehat{\boldsymbol{\beta}})^{\mathsf{T}} Y \tag{19}$$

$$= \left(\widehat{\boldsymbol{\beta}}^{\mathsf{T}} \frac{\boldsymbol{Z}^{*\mathsf{T}} \boldsymbol{Z}^*}{n_2} \widehat{\boldsymbol{\beta}} \right)^{-1} \widehat{\boldsymbol{\beta}}^{\mathsf{T}} \frac{\boldsymbol{Z}^{*\mathsf{T}} Y}{n_2}. \tag{20}$$

Observe that $\frac{\boldsymbol{Z}^{*\mathsf{T}} \boldsymbol{Z}^*}{n_2}$ is a matrix of correlations among the xQTLs in $\boldsymbol{Z}^*$, so we can estimate it with $\widehat{\boldsymbol{D}}$. Moreover, $\frac{\boldsymbol{Z}^{*\mathsf{T}} Y}{n_2}$ is a vector of correlations between each variant and

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Y, so we can estimate it with $\widehat{c}$. Therefore, the effects of expression and co-expression on the outcome trait are jointly estimated by

$$\widehat{\theta} = \left(\widehat{\boldsymbol{\beta}}^{\mathsf{T}}\widehat{\boldsymbol{D}}\widehat{\boldsymbol{\beta}}\right)^{-1}\widehat{\boldsymbol{\beta}}^{\mathsf{T}}\widehat{c}. \tag{21}$$

Similarly, the variance of $\widehat{\theta}$ can be estimated in terms of $\widehat{\theta}$, $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{D}}$, and $\widehat{c}$. The residual sum of squares for $Y = \widehat{\boldsymbol{X}}^*\theta + \varepsilon_Y$ is

$$\begin{aligned}
RSS &= \|Y - \widehat{\boldsymbol{X}}^*\widehat{\theta}\|^2 && (22)\\
&= Y^{\mathsf{T}}Y - 2Y^{\mathsf{T}}\widehat{\boldsymbol{X}}^*\widehat{\theta} + \widehat{\theta}^{\mathsf{T}}(\widehat{\boldsymbol{X}}^*)^{\mathsf{T}}\widehat{\boldsymbol{X}}^*\widehat{\theta} && (23)\\
&= (n_2 - 1)\frac{Y^{\mathsf{T}}Y}{n_2 - 1} - 2Y^{\mathsf{T}}\boldsymbol{Z}^*\widehat{\boldsymbol{\beta}}\widehat{\theta} + \widehat{\theta}^{\mathsf{T}}\widehat{\boldsymbol{\beta}}^{\mathsf{T}}\boldsymbol{Z}^{*\mathsf{T}}\boldsymbol{Z}^*\widehat{\boldsymbol{\beta}}\widehat{\theta} && (24)\\
&= (n_2 - 1)\frac{Y^{\mathsf{T}}Y}{n_2 - 1} - 2n_2\left(\frac{\boldsymbol{Z}^{*\mathsf{T}}Y}{n_2}\right)^{\mathsf{T}}\widehat{\boldsymbol{\beta}}\widehat{\theta} + n_2\widehat{\theta}^{\mathsf{T}}\widehat{\boldsymbol{\beta}}^{\mathsf{T}}\frac{\boldsymbol{Z}^{*\mathsf{T}}\boldsymbol{Z}^*}{n_2}\widehat{\boldsymbol{\beta}}\widehat{\theta}. && (25)
\end{aligned}$$

Observe that $\frac{Y^{\mathsf{T}}Y}{n_2-1} = 1$ because we assumed that Y was scaled to have a variance of 1, $\frac{\boldsymbol{Z}^{*\mathsf{T}}Y}{n_2}$ can be estimated by $\widehat{c}$, $\frac{\boldsymbol{Z}^{*\mathsf{T}}\boldsymbol{Z}^*}{n_2}$ can be estimated by $\widehat{\boldsymbol{D}}$, and the leftover n_2 can be replaced by n'. Therefore, we estimate the RSS by

$$RSS \approx n'\left(1 - 2\widehat{c}^{\mathsf{T}}\widehat{\boldsymbol{\beta}}\widehat{\theta} + \widehat{\theta}^{\mathsf{T}}\widehat{\boldsymbol{\beta}}^{\mathsf{T}}\widehat{\boldsymbol{D}}\widehat{\boldsymbol{\beta}}\widehat{\theta}\right) - 1. \tag{26}$$

Finally, we estimate the variance of $\widehat{\theta}$ by

$$\begin{aligned}
\widehat{Var}(\widehat{\theta}) &= \left((\widehat{\boldsymbol{X}}^*)^{\mathsf{T}}\widehat{\boldsymbol{X}}^*\right)^{-1}\frac{RSS}{n_2 - 4} && (27)\\
&\approx \left(\widehat{\boldsymbol{\beta}}^{\mathsf{T}}\widehat{\boldsymbol{D}}\widehat{\boldsymbol{\beta}}\right)^{-1}\frac{RSS}{n'(n' - 4)}. && (28)
\end{aligned}$$

A Wald test for the effect of A, B, or C on the outcome trait can be performed as described above. An F test of overall significance can also be performed using the formula derived for individual-level data above, except that the sample size of the individual-level outcome cohort (n_2) should be replaced with the sample size of the GWAS cohort (n').

Expression and co-expression imputation models

Any statistical or machine learning algorithms can be used to build expression and co-expression imputation models for COWAS when individual-level data is available for the outcome trait. In order to perform the testing stage of COWAS using summary-level GWAS data, however, the models must be built using an algorithm that can provide a vector of xQTL weights. In this paper, we evaluated penalized linear regression models with three different penalties: the elastic net penalty, the lasso penalty, and the ridge penalty. All models were trained using the glmnet package in R. The α

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

hyperparameter was set to $\alpha = 0.5$ for elastic net regression, $\alpha = 1$ for lasso regression, and $\alpha = 0$ for ridge regression. The λ hyperparameter, which controls the strength of the penalty, was chosen through 10-fold cross validation. Our implementation of COWAS also provides the option for linear regression with stepwise variable selection, but we did not use that method in this study due to its much longer runtime.

To further increase the computational performance of COWAS, we pre-screened genetic variants before including them as features in the penalized regression models. First, we conducted a proteome-wide pQTL mapping study to compute the association between each variant and the standardized residuals of each protein. This enabled us to consider two approaches for pre-screening predictive variants to ensure that only strong pQTLs are considered by each model. For P value screening, we ranked variants by their pQTL P values and kept the top 100. For effect size screening, we instead ranked variants by the absolute values of their pQTL effect sizes and kept the top 100. In both cases, the feature set for the co-expression model was taken to be the union of the top-ranked pQTLs for the two proteins. Note that for the models trained with *cis*-pQTLs only, we restricted the rankings to variants located close to the gene that codes for the given protein (see details below). Performing feature screening before training imputation models greatly decreases the runtime of COWAS, making it computationally feasible to apply to large-scale biobank studies.

It is important to ensure that only well-imputed protein pairs are considered in the testing stage. COWAS assesses the predictive performance of each model by calculating the correlation between imputed and measured expression on a held-out test set. In particular, we randomly selected 80% of the available samples for each protein pair to train imputation models and the remaining 20% to test their predictive performance. For the single-protein models, we calculated the correlation between imputed and measured expression on the test set. Recall that the outcome for the co-expression model, on the other hand, is a quantity estimated using measured expression levels as well as predictions from single-protein models. Thus, to evaluate the performance of the co-expression model, we correlated the conditional co-expression imputed by a model trained on the 80% training set with the conditional co-expression estimated using single-protein models trained on the entire data. After assessing predictive performance, all three models were re-trained on the full dataset to obtain the final xQTL weight vectors. Only pairs in which all three models had out-of-sample correlations greater than 0.03 were used for hypothesis testing.

Standard PWAS analysis

We compared our proposed method with the marginal, single-exposure PWAS approach commonly used today. Note that PWAS is independently performed on one protein at a time, so we considered a pair to be significant according to a PWAS analysis if at least one of its proteins was identified by PWAS. Without loss of generality, we will use the notation for protein A to explain the PWAS association test.

Let $\widehat{\beta}_A \in \mathbb{R}^{p_A}$ be a vector of fitted pQTL weights for imputing the expression level of protein A, as defined above. These weights could be obtained from any regression model, such as the penalized linear regression models we considered in this study. In

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

a setting with individual-level data available for the outcome trait, we first compute

$$\widehat{A}^* = \boldsymbol{Z}^*_{\boldsymbol{A}} \widehat{\beta}_A, \tag{29}$$

where $\boldsymbol{Z}^*_{\boldsymbol{A}}$ is a matrix of pQTL genotypes for individuals in the outcome trait dataset, as defined above. Then we fit the model

$$Y = \widehat{A}^* \theta_{A,M} + \varepsilon_{Y,A}, \tag{30}$$

where $\theta_{A,M} \in \mathbb{R}$ is the coefficient of interest and $\varepsilon_{Y,A}$ is an independent, normally distributed error term. To determine if the genetically regulated component of A has a significant effect on Y, we test the hypothesis $H_0 : \theta_{A,M} = 0$ against its two-sided alternative using a Wald test. The test statistic is $\widehat{\theta}^2_{A,M}/Var(\widehat{\theta}_{A,M})$, which asymptotically follows a χ^2 distribution with 1 degree of freedom under H_0.

Importantly, note that the PWAS effect size $\theta_{A,M}$ is distinct from the COWAS effect size θ_A. Whereas $\theta_{A,M}$ is the marginal effect of $\widehat{A}$ on Y, the COWAS coefficient θ_A is the effect of $\widehat{A}$ on Y after accounting for the effects of $\widehat{B}$ and $\widehat{C}$.

In practice, we performed PWAS using summary-level GWAS data for the outcome trait and an LD reference panel. Using similar notations to those defined above, let $\widehat{c}_A = (\widehat{c}_1, \ldots, \widehat{c}_{p_A})^\intercal$ be a vector of correlations between each of the p_A variants and the outcome trait Y. Also let $\widehat{\boldsymbol{D}}_{\boldsymbol{A}} \in \mathbb{R}^{p_A \times p_A}$ be an LD reference panel for those same p_A pQTLs. Then we can estimate $\theta_{A,M}$ by

$$\widehat{\theta}_{A,M} = \left(\widehat{\beta}_A^\intercal \widehat{\boldsymbol{D}}_{\boldsymbol{A}} \widehat{\beta}_A\right)^{-1} \widehat{\beta}_A^\intercal \widehat{c}_A. \tag{31}$$

The residual sum of squares for $Y = \widehat{A}^* \theta_{A,M} + \varepsilon_{Y,A}$ can be estimated by

$$RSS_A \approx n'\left(1 - 2\widehat{c}_A^\intercal \widehat{\beta}_A \widehat{\theta}_{A,M} + \widehat{\theta}^\intercal_{A,M} \widehat{\beta}_A^\intercal \widehat{\boldsymbol{D}}_{\boldsymbol{A}} \widehat{\beta}_A \widehat{\theta}_{A,M}\right) - 1, \tag{32}$$

and finally we estimate the variance of $\widehat{\theta}_{A,M}$ by

$$\widehat{Var}(\widehat{\theta}_{A,M}) = \left(\widehat{\beta}_A^\intercal \widehat{\boldsymbol{D}}_{\boldsymbol{A}} \widehat{\beta}_A\right)^{-1} \frac{RSS_A}{n'(n'-2)}, \tag{33}$$

where n' is the sample size of the GWAS for Y, as before. Notice that these formulas are analogous to the ones given for COWAS, except with different degrees of freedom and dimensions for each quantity.

Data processing and quality control

We trained protein expression and co-expression imputation models on individual-level data from the UKB. First we downloaded genotype data for 92,457,702 autosomal markers and 487,363 samples. Imputation, phasing, and extensive quality control checks had already been performed on the provided data as detailed previously [39, 73].

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

We further removed any individuals that had a missingness rate above 1% across markers, and then subset the data to only keep high-quality samples of genetically-inferred White British ancestry with no relatives of third degree or closer, using indicators provided by UKB. Lastly, we subset the data to individuals who have proteomic data available at the baseline visit. After these steps, 36,171 samples remained.

We also performed additional variant-level quality control on the UKB genotype data. In particular, we removed all variants with a missingness rate greater than 10% across the remaining individuals, those with a minor allele count (MAC) less than 100, those with a minor allele frequency (MAF) less than 1%, and those that failed a Hardy–Weinberg equilibrium test with $P < 10^{-15}$. To facilitate matching up variants between UKB data and outcome trait GWAS data, we also removed all variants lacking an rsID and those that are palindromic. Finally, we pruned the variants to $r^2 < 0.8$ with a 1,000 base pair (bp) window and a step size of 100 bp. After these steps, 1,689,714 variants remained. All sample-level and variant-level quality control was done in PLINK 2.00.

Next, we computed genetic principal components (PCs) from the quality-controlled genotype data. Before computing PCs, we applied several data processing steps in addition to those described above. In particular, we additionally removed all variants in regions of long-range LD [73] and then pruned the remaining ones to a strict threshold of $r^2 < 0.1$ with a 1,000 bp window and a step size of 100 bp. The computation of genetic PCs was also done in PLINK 2.00.

Proteomic profiling in blood plasma was performed by the UKB Pharma Proteomics Project using the antibody-based Olink Explore 3072 proximity extension assay, which measured 2,941 protein analytes across eight panels and captured 2,923 unique proteins [40]. Various quality control checks and normalization had already been performed as described previously [40]. We downloaded Normalized Protein eXpression (NPX) values for 2,923 proteins in 53,073 samples. After subsetting the data to individuals with high-quality genotypes and protein abundance measurements at the baseline visit, as described above, a total of 36,171 samples remained. The sample sizes for individual proteins ranged from 106 to 35,581 individuals, with a median of 33,643 individuals. Rather than imputing missing values, we used the intersection of samples with non-missing data within each protein pair.

Then we normalized the NPX levels using a rank-based inverse normal transformation. In particular, we utilized the commonly-used Blom transform with an offset of 3/8 and ties broken by averaging. Following the transformation, we regressed out the following standardized covariates: age, age^2, sex, age $*$ sex, age^2 $*$ sex, UKB assessment center, genotyping array, and the first 20 genetic PCs. These protein expression residuals, after normalizing and adjusting for covariates, were used in all downstream analyses.

Protein annotations

We trained models that include pQTLs screened from across the genome, as well as models that only include *cis*-pQTLs. To identify the *cis*-SNPs for each protein, we obtained start and end positions for the genes coding each assayed protein from annotations provided by the UKB Pharma Proteomics Project (Synapse: syn52364558),

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

and then lifted them over to the hg19 genome build used by UKB. For proteins coded by several genes, we only considered the first gene listed in the UKB annotation file. We defined the *cis* region for each gene as beginning 500,000 bp upstream of its transcription start site and ending 500,000 bp downstream of its transcription end site. Thus, the *cis*-pQTLs for a given protein are the top-ranked variants that fall within this genomic window.

In this study we only trained models for pairs of proteins found in the HIPPIE database of PPIs [45]. To identify those pairs, we downloaded version 2.3 of the HIPPIE database and mapped each protein in the database to its gene name using the UniProt ID Mapping web tool (https://www.uniprot.org/id-mapping). The gene names were then matched with protein annotations from UKB, and protein pairs present in the HIPPIE database were retained. Note that we considered all protein pairs listed in the HIPPIE database, regardless of their interaction confidence score.

GWAS data for outcome traits

We considered three complex traits as outcomes in our application of COWAS: low-density lipoprotein (LDL) cholesterol, Alzheimer's disease (AD), and Parkinson's disease (PD). The testing stage of COWAS was performed using summary-level data from the largest available GWAS study for each trait. Since none of the GWAS studies provided Z scores in their summary data, we computed them by dividing each variant's effect size by its standard error.

For LDL cholesterol levels, we downloaded GWAS summary statistics data from the Global Lipids Genetics Consortium (GLGC) [41]. The GLGC aggregated GWAS results from 1,320,016 individuals of European ancestry across 146 cohorts. Their meta-analysis provided summary statistics for 47,006,483 genetic variants and five lipid traits, including LDL cholesterol. Although the authors also conducted a multi-ancestry meta-analysis, we used results that were meta-analyzed solely in the European cohorts to ensure consistency with the genetic ancestry of the majority of UKB participants. A total of 1,624,628 genetic variants remained after harmonization with our quality-controlled UKB genotype data.

For AD status, we downloaded GWAS summary statistics data from the European Alzheimer & Dementia Biobank (EADB) consortium [42]. Namely, we used their stage 1 GWAS of AD and related dementias in individuals of European ancestry. The stage 1 GWAS was a meta-analysis based on 39,106 clinically diagnosed cases, 46,828 proxy cases (with disease status inferred from parental history), and 401,577 controls. Summary statistics for 21,101,114 genetic variants were provided, of which 1,435,986 remained after harmonization with our quality-controlled UKB genotype data.

For PD status, we downloaded GWAS summary statistics data from the International Parkinson's Disease Genomics Consortium (IPDGC) [43]. The IPDGC GWAS is also a meta-analysis, aggregating associations across 17 cohorts with individuals of European ancestry. Their main analysis included 37,688 clinically diagnosed cases, 18,618 proxy cases (with disease status inferred from first-degree relatives), and 1,417,791 controls. However, the publicly available summary statistics exclude three studies with individuals from 23andMe due to data sharing restrictions. We used the publicly available GWAS data in our analysis, which was based on 15,056

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

clinically diagnosed cases, 18,618 proxy cases, and 449,056 controls. Summary-level data were provided for 17,443,094 genetic variants, of which 1,393,959 remained after harmonization with our quality-controlled UKB genotype data.

Data availability. Genotype, covariate, and protein expression data from the UK Biobank are available through the UK Biobank data access process (https://www.ukbiobank.ac.uk/enable-your-research). Access to the UK Biobank data was approved through UK Biobank Application #35107. Annotations for proteins assayed by the UK Biobank Pharma Proteomics Project are publicly available on Synapse (https://www.synapse.org/Synapse:syn51364943). Protein pairs with known interactions are publicly accessible in the HIPPIE web tool (https://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie). Publicly available GWAS summary statistics for cholesterol levels were downloaded from the Global Lipids Genetics Consortium website (https://csg.sph.umich.edu/willer/public/glgc-lipids2021). Publicly available GWAS summary statistics for Alzheimer's disease and Parkinson's disease were obtained from the NHGRI-EBI GWAS Catalog (https://www.ebi.ac.uk/gwas) under accession numbers GCST90027158 and GCST009325, respectively.

Code availability. Our software for COWAS is implemented in R and made available on GitHub under a GPL-3.0 open source license at https://github.com/mykmal/cowas. This GitHub repository also contains the scripts used for data quality control and batch processing. Fitted model weights for all protein expression and co-expression imputation models trained in this study are provided on Synapse at https://synapse.org/cowas.

Acknowledgements. This work was supported by the National Institutes of Health (NIH) under grants R01 AG065636 and RF1 AG067924. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The authors also acknowledge the Minnesota Supercomputing Institute (MSI) at the University of Minnesota for providing high-performance computing resources that contributed to the research results reported within this paper.

Author contributions. W.P. conceived and supervised the project. M.M. developed the method, implemented the software, and performed the analyses. M.M. drafted the manuscript and W.P. proofread it.

Competing interests. The authors declare no competing interests.

Supplementary information

Supplementary Data 1: Summarized out-of-sample performance metrics for each imputation model type.

Supplementary Data 2: Out-of-sample performance metrics for each protein pair imputed by an elastic net penalized regression model with P value screening and only *cis*-pQTLs as predictors.

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Supplementary Data 3: Out-of-sample performance metrics for each protein pair imputed by a lasso penalized regression model with P value screening and only *cis*-pQTLs as predictors.

Supplementary Data 4: Out-of-sample performance metrics for each protein pair imputed by a ridge penalized regression model with P value screening and only *cis*-pQTLs as predictors.

Supplementary Data 5: Out-of-sample performance metrics for each protein pair imputed by an elastic net penalized regression model with effect size screening and only *cis*-pQTLs as predictors.

Supplementary Data 6: Out-of-sample performance metrics for each protein pair imputed by a lasso penalized regression model with effect size screening and only *cis*-pQTLs as predictors.

Supplementary Data 7: Out-of-sample performance metrics for each protein pair imputed by a ridge penalized regression model with effect size screening and only *cis*-pQTLs as predictors.

Supplementary Data 8: Out-of-sample performance metrics for each protein pair imputed by an elastic net penalized regression model with P value screening and both *cis*-pQTLs and *trans*-pQTLs as predictors.

Supplementary Data 9: Out-of-sample performance metrics for each protein pair imputed by a lasso penalized regression model with P value screening and both *cis*-pQTLs and *trans*-pQTLs as predictors.

Supplementary Data 10: Out-of-sample performance metrics for each protein pair imputed by a ridge penalized regression model with P value screening and both *cis*-pQTLs and *trans*-pQTLs as predictors.

Supplementary Data 11: Out-of-sample performance metrics for each protein pair imputed by an elastic net penalized regression model with effect size screening and both *cis*-pQTLs and *trans*-pQTLs as predictors.

Supplementary Data 12: Out-of-sample performance metrics for each protein pair imputed by a lasso penalized regression model with effect size screening and both *cis*-pQTLs and *trans*-pQTLs as predictors.

Supplementary Data 13: Out-of-sample performance metrics for each protein pair imputed by a ridge penalized regression model with effect size screening and both *cis*-pQTLs and *trans*-pQTLs as predictors.

Supplementary Data 14: COWAS (joint) and PWAS (marginal) results for association between genetically regulated (co-)expression and low-density lipoprotein

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

(LDL) cholesterol. Only well-imputed protein pairs are included, and out-of-sample performance metrics for the imputation models are also provided.

Supplementary Data 15: COWAS (joint) and PWAS (marginal) results for association between genetically regulated (co-)expression and Alzheimer's disease. Only well-imputed protein pairs are included, and out-of-sample performance metrics for the imputation models are also provided.

Supplementary Data 16: COWAS (joint) and PWAS (marginal) results for association between genetically regulated (co-)expression and Parkinson's disease. Only well-imputed protein pairs are included, and out-of-sample performance metrics for the imputation models are also provided.

References

[1] Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS. *PLOS Genetics* **6**, e1000888 (2010). URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000888.

[2] Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012). URL https://www.science.org/doi/10.1126/science.1222794.

[3] Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *The American Journal of Human Genetics* **95**, 535–552 (2014). URL https://www.cell.com/ajhg/fulltext/S0002-9297(14)00426-1.

[4] Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* **47**, 1228–1235 (2015). URL https://www.nature.com/articles/ng.3404.

[5] Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015). URL https://www.nature.com/articles/nature14248.

[6] Meuleman, W. *et al.* Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020). URL https://www.nature.com/articles/s41586-020-2559-3.

[7] Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics* **47**, 1091–1098 (2015). URL https://www.nature.com/articles/ng.3367.

[8] Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics* **48**, 245–252 (2016). URL https://www.nature.com/articles/ng.3506.

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

[9] Zheng, J. *et al.* Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nature Genetics* **52**, 1122–1131 (2020). URL https://www.nature.com/articles/s41588-020-0682-6.

[10] Brandes, N., Linial, N. & Linial, M. PWAS: proteome-wide association study—linking genes and phenotypes by functional variation in proteins. *Genome Biology* **21**, 173 (2020). URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02089-x.

[11] Zhang, J. *et al.* Plasma proteome analyses in individuals of European and African ancestry identify *cis*-pQTLs and models for proteome-wide association studies. *Nature Genetics* **54**, 593–602 (2022). URL https://www.nature.com/articles/s41588-022-01051-w.

[12] Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature Communications* **9**, 1825 (2018). URL https://www.nature.com/articles/s41467-018-03621-1.

[13] Hu, Y. *et al.* A statistical framework for cross-tissue transcriptome-wide association analysis. *Nature Genetics* **51**, 568–576 (2019). URL https://www.nature.com/articles/s41588-019-0345-7.

[14] Bhattacharya, A., Li, Y. & Love, M. I. MOSTWAS: Multi-omic strategies for transcriptome-wide association studies. *PLOS Genetics* **17**, e1009398 (2021). URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1009398.

[15] Lin, Z., Xue, H., Malakhov, M. M., Knutson, K. A. & Pan, W. Accounting for nonlinear effects of gene expression identifies additional associated genes in transcriptome-wide association studies. *Human Molecular Genetics* **31**, 2462–2470 (2022). URL https://academic.oup.com/hmg/article/31/14/2462/6511389.

[16] Li, Z. *et al.* METRO: Multi-ancestry transcriptome-wide association studies for powerful gene-trait association detection. *The American Journal of Human Genetics* **109**, 783–801 (2022). URL https://www.cell.com/ajhg/fulltext/S0002-9297(22)00099-4.

[17] Bhattacharya, A. *et al.* Isoform-level transcriptome-wide association uncovers genetic risk mechanisms for neuropsychiatric disorders in the human brain. *Nature Genetics* **55**, 2117–2128 (2023). URL https://www.nature.com/articles/s41588-023-01560-2.

[18] Knutson, K. A. & Pan, W. MATS: a novel multi-ancestry transcriptome-wide association study to account for heterogeneity in the effects of cis-regulated gene expression on complex traits. *Human Molecular Genetics* **32**, 1237–1251 (2023). URL https://academic.oup.com/hmg/article/32/8/1237/6731969.

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

[19] He, J. *et al.* A statistical method for image-mediated association studies discovers genes and pathways associated with four brain disorders. *The American Journal of Human Genetics* **111**, 48–69 (2024). URL https://www.cell.com/ajhg/fulltext/S0002-9297(23)00401-9.

[20] Gusev, A. *et al.* Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nature Genetics* **50**, 538–548 (2018). URL https://www.nature.com/articles/s41588-018-0092-1.

[21] Zhong, J. *et al.* A transcriptome-wide association study identifies novel candidate susceptibility genes for pancreatic cancer. *JNCI: Journal of the National Cancer Institute* **112**, 1003–1012 (2020). URL https://academic.oup.com/jnci/article/112/10/1003/5698709.

[22] Zhao, B. *et al.* Transcriptome-wide association analysis of brain structures yields insights into pleiotropy with complex neuropsychiatric traits. *Nature Communications* **12**, 2878 (2021). URL https://www.nature.com/articles/s41467-021-23130-y.

[23] Gao, G. *et al.* A joint transcriptome-wide association study across multiple tissues identifies candidate breast cancer susceptibility genes. *The American Journal of Human Genetics* **110**, 950–962 (2023). URL https://www.cell.com/ajhg/fulltext/S0002-9297(23)00129-5.

[24] Chen, D. M. *et al.* Transcriptome-wide association analysis identifies candidate susceptibility genes for prostate-specific antigen levels in men without prostate cancer. *Human Genetics and Genomics Advances* **5**, 100315 (2024). URL https://www.cell.com/hgg-advances/fulltext/S2666-2477(24)00054-X.

[25] Gao, G. *et al.* A multi-tissue, splicing-based joint transcriptome-wide association study identifies susceptibility genes for breast cancer. *The American Journal of Human Genetics* **111**, 1100–1113 (2024). URL https://www.cell.com/ajhg/fulltext/S0002-9297(24)00125-3.

[26] Zhu, J. *et al.* Associations between genetically predicted plasma protein levels and Alzheimer's disease risk: a study using genetic prediction models. *Alzheimer's Research & Therapy* **16**, 8 (2024). URL https://alzres.biomedcentral.com/articles/10.1186/s13195-023-01378-4.

[27] Liu, L. *et al.* Conditional transcriptome-wide association study for fine-mapping candidate causal genes. *Nature Genetics* **56**, 348–356 (2024). URL https://www.nature.com/articles/s41588-023-01645-y.

[28] Zhao, S. *et al.* Adjusting for genetic confounders in transcriptome-wide association studies improves discovery of risk genes of complex traits. *Nature Genetics* **56**, 336–347 (2024). URL https://www.nature.com/articles/s41588-023-01648-9.

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

[29] Chan, L. S., Malakhov, M. M. & Pan, W. A novel multivariable Mendelian randomization framework to disentangle highly correlated exposures with application to metabolomics. *The American Journal of Human Genetics* **111**, 1834–1847 (2024). URL https://www.cell.com/ajhg/fulltext/S0002-9297(24)00251-9.

[30] Hu, T. *et al.* Omnibus proteome-wide association study identifies 43 risk genes for Alzheimer disease dementia. *The American Journal of Human Genetics* **111**, 1848–1863 (2024). URL https://www.cell.com/ajhg/fulltext/S0002-9297(24)00225-8.

[31] Cordell, H. J. Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics* **10**, 392–404 (2009). URL https://www.nature.com/articles/nrg2579.

[32] Van Steen, K. Travelling the world of gene–gene interactions. *Briefings in Bioinformatics* **13**, 1–19 (2012). URL https://academic.oup.com/bib/article/13/1/1/218763.

[33] Gonzalez, M. W. & Kann, M. G. Chapter 4: Protein interactions and disease. *PLOS Computational Biology* **8**, e1002819 (2012). URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002819.

[34] Lu, H. *et al.* Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials. *Signal Transduction and Targeted Therapy* **5**, 213 (2020). URL https://www.nature.com/articles/s41392-020-00315-3.

[35] van der Wijst, M. G. P. *et al.* Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nature Genetics* **50**, 493–497 (2018). URL https://www.nature.com/articles/s41588-018-0089-9.

[36] Li, Y. *et al.* Pan-cancer proteogenomics connects oncogenic drivers to functional states. *Cell* **186**, 3921–3944.e25 (2023). URL https://www.cell.com/cell/fulltext/S0092-8674(23)00780-8.

[37] Urzúa-Traslaviña, C. G. *et al.* Co-expression in tissue-specific gene networks links genes in cancer-susceptibility loci to known somatic driver genes. *BMC Medical Genomics* **17**, 186 (2024). URL https://bmcmedgenomics.biomedcentral.com/articles/10.1186/s12920-024-01941-4.

[38] Cheng, F. *et al.* Comprehensive characterization of protein–protein interactions perturbed by disease mutations. *Nature Genetics* **53**, 342–353 (2021). URL https://www.nature.com/articles/s41588-020-00774-y.

[39] Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018). URL https://www.nature.com/articles/s41586-018-0579-z.

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

[40] Sun, B. B. *et al.* Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329–338 (2023). URL https://www.nature.com/articles/s41586-023-06592-6.

[41] Graham, S. E. *et al.* The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600**, 675–679 (2021). URL https://www.nature.com/articles/s41586-021-04064-3.

[42] Bellenguez, C. *et al.* New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nature Genetics* **54**, 412–436 (2022). URL https://www.nature.com/articles/s41588-022-01024-z.

[43] Nalls, M. A. *et al.* Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *The Lancet Neurology* **18**, 1091–1102 (2019). URL https://www.thelancet.com/journals/laneur/article/PIIS1474-4422(19)30320-5.

[44] Romanov, N. *et al.* Disentangling genetic and environmental effects on the proteotypes of individuals. *Cell* **177**, 1308–1318.e10 (2019). URL https://www.cell.com/cell/fulltext/S0092-8674(19)30277-6.

[45] Alanis-Lobato, G., Andrade-Navarro, M. A. & Schaefer, M. H. HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Research* **45**, D408–D414 (2017). URL https://academic.oup.com/nar/article/45/D1/D408/2290937.

[46] Murrell, J., Farlow, M., Ghetti, B. & Benson, M. D. A mutation in the amyloid precursor protein associated with hereditary Alzheimer's disease. *Science* **254**, 97–99 (1991). URL https://www.science.org/doi/10.1126/science.1925564.

[47] O'Brien, R. J. & Wong, P. C. Amyloid precursor protein processing and Alzheimer's disease. *Annual Review of Neuroscience* **34**, 185–204 (2011). URL https://www.annualreviews.org/content/journals/10.1146/annurev-neuro-061010-113613.

[48] Delport, A. & Hewer, R. The amyloid precursor protein: a converging point in Alzheimer's disease. *Molecular Neurobiology* **59**, 4501–4516 (2022). URL https://link.springer.com/article/10.1007/s12035-022-02863-x.

[49] Wu, P.-R. *et al.* DAPK activates MARK1/2 to regulate microtubule assembly, neuronal differentiation, and tau toxicity. *Cell Death & Differentiation* **18**, 1507–1520 (2011). URL https://www.nature.com/articles/cdd20112.

[50] Zhao, N., Liu, C.-C., Qiao, W. & Bu, G. Apolipoprotein E, receptors, and modulation of Alzheimer's disease. *Biological Psychiatry* **83**, 347–357 (2018). URL https://www.sciencedirect.com/science/article/pii/S0006322317313586.

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

[51] Cao, Q. *et al.* Inhibiting amyloid-β cytotoxicity through its interaction with the cell surface receptor LilrB2 by structure-based design. *Nature Chemistry* **10**, 1213–1221 (2018). URL https://www.nature.com/articles/s41557-018-0147-z.

[52] Lao, K. *et al.* Identification of novel Aβ-LilrB2 inhibitors as potential therapeutic agents for Alzheimer's disease. *Molecular and Cellular Neuroscience* **114**, 103630 (2021). URL https://www.sciencedirect.com/science/article/pii/S1044743121000439.

[53] Cho, S.-J. *et al.* Altered expression of Notch1 in Alzheimer's disease. *PLOS ONE* **14**, e0224941 (2019). URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0224941.

[54] Kapoor, A. & Nation, D. A. Role of Notch signaling in neurovascular aging and Alzheimer's disease. *Seminars in Cell & Developmental Biology* **116**, 90–97 (2021). URL https://www.sciencedirect.com/science/article/pii/S1084952120302056.

[55] Bamford, R. A. *et al.* The interaction between contactin and amyloid precursor protein and its role in Alzheimer's disease. *Neuroscience* **424**, 184–202 (2020). URL https://www.sciencedirect.com/science/article/pii/S030645221930692X.

[56] Spillantini, M. G. *et al.* α-synuclein in Lewy bodies. *Nature* **388**, 839–840 (1997). URL https://www.nature.com/articles/42166.

[57] Ozansoy, M. & Başak, A. N. The central theme of Parkinson's disease: α-synuclein. *Molecular Neurobiology* **47**, 460–465 (2013). URL https://link.springer.com/article/10.1007/s12035-012-8369-3.

[58] Bloem, B. R., Okun, M. S. & Klein, C. Parkinson's disease. *The Lancet* **397**, 2284–2303 (2021). URL https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)00218-X.

[59] Arawaka, S. *et al.* The role of G-protein-coupled receptor kinase 5 in pathogenesis of sporadic Parkinson's disease. *The Journal of Neuroscience* **26**, 9227–9238 (2006). URL https://www.jneurosci.org/content/26/36/9227.

[60] Alexopoulou, Z. *et al.* Deubiquitinase Usp8 regulates α-synuclein clearance and modifies its toxicity in Lewy body disease. *Proceedings of the National Academy of Sciences* **113** (2016). URL https://www.pnas.org/doi/full/10.1073/pnas.1523597113.

[61] Amer-Sarsour, F., Kordonsky, A., Berdichevsky, Y., Prag, G. & Ashkenazi, A. Deubiquitylating enzymes in neuronal health and disease. *Cell Death & Disease* **12**, 120 (2021). URL https://www.nature.com/articles/s41419-020-03361-5.

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

[62] Ysselstein, D. *et al.* Endosulfine-alpha inhibits membrane-induced α-synuclein aggregation and protects against α-synuclein neurotoxicity. *Acta Neuropathologica Communications* **5**, 3 (2017). URL https://actaneurocomms.biomedcentral.com/articles/10.1186/s40478-016-0403-7.

[63] Keeling, B. H. *et al.* DRD3 Ser9Gly and HS1BP3 Ala265Gly are not associated with Parkinson disease. *Neuroscience Letters* **461**, 74–75 (2009). URL https://www.sciencedirect.com/science/article/pii/S0304394009007642.

[64] Ruffner, H., Bauer, A. & Bouwmeester, T. Human protein–protein interaction networks and the value for drug discovery. *Drug Discovery Today* **12**, 709–716 (2007). URL https://www.sciencedirect.com/science/article/pii/S1359644607002784.

[65] Chautard, E., Thierry-Mieg, N. & Ricard-Blum, S. Interaction networks: From protein functions to drug discovery. a review. *Pathologie Biologie* **57**, 324–333 (2009). URL https://www.sciencedirect.com/science/article/abs/pii/S0369811408002538.

[66] Wodak, S. J., Vlasblom, J., Turinsky, A. L. & Pu, S. Protein–protein interaction networks: the puzzling riches. *Current Opinion in Structural Biology* **23**, 941–953 (2013). URL https://www.sciencedirect.com/science/article/pii/S0959440X13001541.

[67] Keys, K. L. *et al.* On the cross-population generalizability of gene expression prediction models. *PLOS Genetics* **16**, e1008927 (2020). URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1008927.

[68] Patel, R. A. *et al.* Genetic interactions drive heterogeneity in causal variant effect sizes for gene expression and complex traits. *The American Journal of Human Genetics* **109**, 1286–1297 (2022). URL https://www.cell.com/ajhg/fulltext/S0002-9297(22)00252-X.

[69] Bhattacharya, A. *et al.* Best practices for multi-ancestry, meta-analytic transcriptome-wide association studies: Lessons from the global biobank meta-analysis initiative. *Cell Genomics* **2**, 100180 (2022). URL https://www.cell.com/cell-genomics/fulltext/S2666-979X(22)00125-2.

[70] Song, L., Langfelder, P. & Horvath, S. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics* **13**, 328 (2012). URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-328.

[71] van Dam, S., Võsa, U., van der Graaf, A., Franke, L. & de Magalhães, J. P. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics* **19**, 575–592 (2018). URL https://academic.oup.com/bib/article/19/4/575/2888441.

medRxiv preprint doi: https://doi.org/10.1101/2024.10.02.24314813; this version posted October 4, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

[72] Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X. & Sham, P. C. Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology* **41**, 469–480 (2017). URL https://onlinelibrary.wiley.com/doi/10.1002/gepi.22050.

[73] Bycroft, C. *et al.* Genome-wide genetic data on 500,000 UK Biobank participants (2017). URL https://www.biorxiv.org/content/10.1101/166298. BioRxiv preprint.