ID N. 18806

Università Campus Bio-Medico di Roma

Department of Engineering

Università di Torino

Department of Neurosciences "Rita Levi Montalcini"

Italian National Ph.D. in Artificial Intelligence
Health and Life Sciences
XXXVIII Cycle

Beyond The Genome: Integrative Multi-omic Analysis Of Complex Traits Using Biobank Data

Supervisors

prof. Paolo Provero

prof. Michele Caselle

Candidate

Daniela Fusco

January, 2026

Abstract

Human complex traits arise from the intricate interplay between genetic, molecular, and environmental influences. This thesis leverages large-scale biobank data to move beyond the genome, integrating genomic, transcriptomic, proteomic, and phenotypic layers to dissect the biological mechanisms underlying complex traits and to enhance biomarker discovery.

In the first study, we investigated the molecular basis of the genetic correlation between body mass index (BMI) and brain morphology using the UK Biobank imaging genetics resource. By combining genome-wide association data with brain expression quantitative trait loci (eQTLs), we identified 21 genes whose genetically regulated expression in the brain pleiotropically influences both BMI and regional brain structure. Fine-mapping, colocalization, and epigenetic annotation highlight causal variants in loci such as *TUFM* and *VPS11*, implicating mitochondrial translation and microglial regulation in the neurobiological architecture of obesity.

In the second study, we developed a framework to remove inherited genetic effects from plasma protein levels—genetic adjustment—to refine biomarker prediction. Using proteomic and genomic data from nearly 40,000 UK Biobank participants, we showed that genetically adjusted proteins display stronger associations with 37 diseases and with lifestyle or environmental exposures, corresponding to a 30% median gain in statistical power for biomarker discovery. Multi-protein models

derived from adjusted proteins outperform standard models across multiple conditions, illustrating how subtracting genetic variability enhances detection of biologically relevant, non-genetic signals.

Together, these studies demonstrate that biobank-scale multi-omic integration can reveal the molecular mechanisms linking genotype, intermediate molecular phenotypes, and disease. They advance a generalizable paradigm for moving beyond the genome toward a more complete, mechanistic, and predictive understanding of human health and disease.

Contents

Introduction

Biomedical research has undergone a substantial transformation over the past few decades, driven by advancements in technology and data science. One of the most significant developments has been the creation of large-scale population biobanks, which collect detailed genetic, molecular, and health records from hundreds of thousands and, in some cases, millions of individuals. These biobanks have provided unprecedented access to different datasets, enabling researchers to investigate the complex interactions between genetic variation, environmental exposures, and health outcomes at a population level. By linking these various data types, researchers can gain a deeper understanding of the factors that contribute to health and disease. [1, 2]

Genome-wide association studies (GWAS) within these resources have uncovered multiple loci associated with complex traits and diseases. GWAS have underscored the polygenic nature of many human traits, revealing that most diseases and traits are influenced by the cumulative effects of many genetic variants, each having a small effect on the phenotype [3]. These discoveries provided the foundation for data-driven precision medicine, an approach that adapts medical treatment to individual genetic profiles.
Yet, despite this success, most associations remain statistical rather than mecha-

nistic, and the biological processes connecting genetic variation to phenotype are often unclear. To bridge this gap, the challenge now is to move beyond these associations to understand how genetic variants influence disease through their effects on biological pathways and molecular networks. This has led to a paradigm shift in biomedical research, where the focus is increasingly shifting towards integrating multi-omics data to map how genetic effects propagate through intermediate molecular layers and to provide a more comprehensive biological interpretation of disease.

The integration of GWAS with other molecular layers has enabled a more comprehensive biological interpretation, revealing how variants shape molecular and cellular processes underlying disease. Expression and protein quantitative trait loci (eQTLs and pQTLs) have emerged as powerful tools to connect genetic variation to gene and protein expression. These loci help identify causal pathways and pleiotropic mechanisms, providing insights into how genes and proteins influence complex traits [4, 5].

At the same time, many human traits are also strongly influenced by non-genetic factors, including environmental exposures, lifestyle and socioeconomic factors, that interact with the genome, and their roles in disease are complex and often underexplored [6, 7, 8]. Distinguishing the relative contributions of genetic and non-genetic components is therefore crucial, not only for understanding disease biology but also for improving the predictive power of biomarkers and disease models.

Recent years have seen the emergence of multi-omic approaches to study how genomic, transcriptomic, proteomic, and phenotypic layers contribute to disease. By leveraging the power of large-scale biobank data, multi-omic approaches allow researchers to investigate causality in disease development, uncover shared molecular mechanisms across diseases, and improve the prediction of disease risk. These integrative approaches hold great promise for the discovery of novel biomarkers and therapeutic targets, as they provide a more nuanced understanding of disease at the molecular level, by combining genomics with downstream molecular data to explore both causality and prediction [9].

The work presented in this thesis applies multi-omic integration to biobank

data in order to advance our understanding of complex trait mechanisms and improve disease prediction.
In Chapter 1, the key concepts and methodologies that underpin the work presented in subsequent chapters will be introduced, providing a foundation for understanding the techniques and approaches implemented in the studies discussed.
The first study, presented in Chapter 2, explores the molecular basis of the genetic correlation between body mass index (BMI) and brain morphology, using a combination of genomic, transcriptomic, and epigenomic data. This study identifies shared pleiotropic mechanisms that link BMI with brain structure, offering insights into the genetic underpinnings of both traits.
The second study, presented in Chapter 3, focuses on the development of a novel polygenic-score-based strategy to remove the inherited genetic component from plasma protein levels, thereby enhancing their predictive power for disease. By adjusting for genetic effects, this approach improves the accuracy and utility of plasma proteins as biomarkers for disease risk.
In Chapter 4, additional related work is briefly introduced. Although these studies do not constitute the main focus of this thesis, they are relevant as they apply the methodologies described here to different contexts, thereby highlighting the broader applicability and importance of these approaches across a broad range of research questions.

Together, these studies highlight the power of multi-omic integration at a biobank scale and demonstrate how it can be leveraged to uncover biological mechanisms that drive complex diseases. The findings of this thesis have broad implications for the field of precision medicine, particularly in the identification of new biomarkers for early disease detection and the development of targeted therapeutic strategies. By bridging the gap between genetic, molecular, and phenotypic data, this research contributes to a deeper understanding of disease biology and the advancement of personalized healthcare.

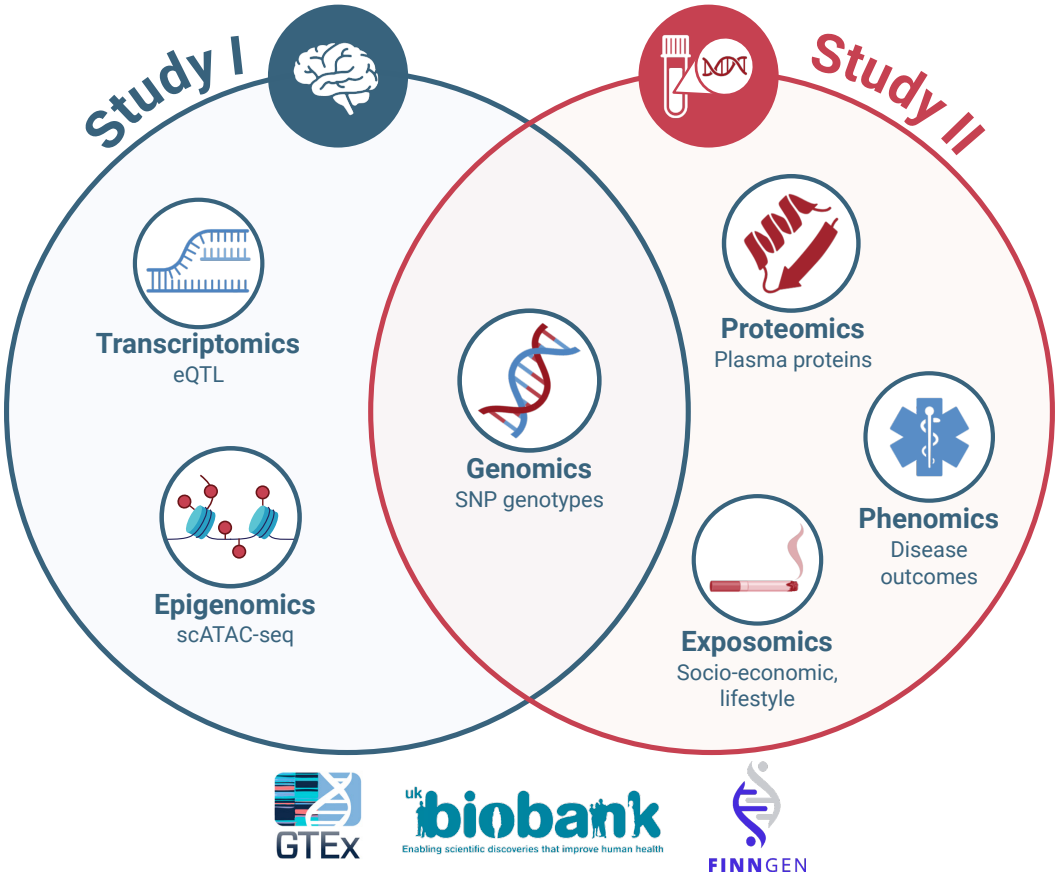


Figure 1: Overview of omic layers used in Study I (Chapter 2) and Study II (Chapter 3).

CHAPTER 1

Literature review

1.1 The GWAS era from the beginning to modern times

The discoveries made through genome-wide association studies (GWAS) have transformed our understanding of complex diseases, revealing the polygenic nature of complex traits. The following sections will explore the fundamental concepts of heritability, polygenicity, and polygenic scores, with a particular focus on their applications in genetic epidemiology.

1.1.1 The rise of GWAS and their main discoveries

The GWAS era led to many scientific discoveries, revealing a powerful approach to the identification of genes involved in common human diseases and the highly polygenic nature of complex traits [3]. The completion of the Human Genome Project (International Human Genome Sequencing Consortium, 2004) and the

development of the HapMap Project (International HapMap Consortium, 2005) enabled systematic cataloguing of common genetic variation and laid the foundation for unbiased, high-throughput association testing across the genome. The first landmark GWAS is the 2005 study by Klein et al., which linked CFH variants to age-related macular degeneration, a landmark demonstration that unbiased genome-wide scans could identify strong associations [10]. Nonetheless, the Wellcome Trust Case Control Consortium (WTCCC) study of 2007 [11] is considered the first large, well-designed GWAS for complex diseases. The work confirmed that complex diseases are highly polygenic, driven by many variants each with small effect size. In the past 20 years a large number of studies identified loci for diseases such as type 2 diabetes [12], Crohn's disease [13], and many other traits and conditions, thanks also to the increase in sample sizes allowed by large consortia and population biobanks.

GWAS rely on the principle that genetic variants located across the genome can be tested for statistical association with a trait in a large group of individuals. Typically, genetic variants are single-nucleotide polymorphisms (SNPs), which are genotyped in many individuals using arrays or derived through genotype imputation [14]. These variants are tested with a linear mixed model for quantitative traits, where for a SNP j the model is:

$$\mathbf{y} = \mu + \mathbf{C}\boldsymbol{\alpha} + \mathbf{g}_j\beta_j + \boldsymbol{\varepsilon},$$

where $\mathbf{y} \in \mathbb{R}^n$ is the phenotype, $\mathbf{C}$ are covariates, $\mathbf{g}_j \in \{0, 1, 2\}^n$ is the mean-centered genotype dosage. The null $H_0 : \beta_j = 0$ is tested with a Wald, score, or likelihood-ratio test; the Wald statistic $t = \hat{\beta}_j/\mathrm{SE}(\hat{\beta}_j)$ yields $t^2 \overset{a}{\sim} \chi^2_1$. For case-control traits, a logistic regression is used:

$$\operatorname{logit} P(y = 1) = \mu + \mathbf{C}\boldsymbol{\alpha} + \mathbf{g}_j\beta_j + \varepsilon.$$

Confounding from population structure or relatedness is mitigated by genetic principal components, mixed models, and post-hoc genomic control. Multiple testing typically uses a genome-wide Bonferroni threshold $P < 5 \times 10^{-8}$. Standard QC in-

cludes variant-level call rates, minor allele frequency filters, Hardy-Weinberg equilibrium, and imputation scores, as well as sample-level relatedness/heterozygosity filters. Effects are reported per allele ($\hat{\beta}_j$) or log-odds; standardized effects use $z_j = \hat{\beta}_j / \mathrm{SE}_j$.

Recent developments in GWAS research have focused on the transferability of GWAS findings across populations, particularly the challenge of applying findings from European ancestry cohorts to non-European populations [15]. This issue arises mainly from the over-representation of European participants in most GWAS [16]. However, large-scale biobanks in East Asia (e.g., Biobank Japan [17], China Kadoorie Biobank [18]) are contributing to a shift toward including more diverse populations in GWAS. As a result, novel statistical methods are being developed to improve the transferability of findings across different populations [19].
The issue of transferability across populations is due to several factors, including differences in haplotype frequencies (e.g., due to differences in linkage disequilibrium) and effect sizes, which may be influenced by gene-by-gene or gene-by-environment interactions. The significance of these factors varies across traits. Additionally, there is evidence that environmental and cultural differences between populations may affect GWAS signals, especially for traits related to mental health [3].

With the collection of more genotype data across diverse populations, the investigation of cross-population differences in the genetic architecture of complex traits can be addressed. Recent advances have also revealed the importance of rare variants through the emergence of whole genome sequencing (WGS) data, which are now increasing [3]. Overall, GWAS has significantly enhanced our understanding of the genetic causes and consequences of complex traits and diseases, and is expected to continue to advance this knowledge in the future.

1.1.2 Heritability, polygenicity, and common variant effects

Heritability, a key concept in quantitative genetics, measures the proportion of variance of the phenotype that is attributable to genetic differences between individuals. It is not a static measure: it is population- and context-specific, and varies across methods of measurement, environmental change, and the effects of migration, selection and inbreeding [20].
Given a phenotype P, we can build a statistical model representing the contribution of the unobserved genotype G and unobserved environmental factors E:

$$\text{Phenotype}(P) = \text{Genotype}(G) + \text{Environment}(E)$$

The variance of the observable phenotypes (σ_P^2) can be expressed as a sum of unobserved underlying variances (σ_G^2 and σ_E^2):

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2$$

The broad sense heritability is defined as a ratio of variances:

$$\text{Heritability (broad sense)} = H^2 = \frac{\sigma_G^2}{\sigma_P^2}$$

We consider narrow-sense heritability, h^2, which captures the proportion of phenotypic variance explained by additive genetic effects. In this definition, dominance effects at a single locus (that is, the extent to which the heterozygote mean deviates from the average of the two homozygote means) and non-additive interaction effects between loci (epistasis) are not included. Broad-sense heritability, H^2, instead encompasses all genetic sources of variance, but is typically more challenging to estimate compared to h^2 [20].
Classically, heritability has been estimated using resemblance between relatives, such as parent–offspring regression, twin and sibling correlations [20]. These approaches use known pedigree relationships to partition phenotypic variance into genetic and environmental components, under assumptions about shared environments, assortative mating and the absence (or specific structure) of geno-

type–environment covariance. In the "genomics era", whole-genome sequence data have enabled a parallel approach based on realized genetic relatedness among nominally unrelated individuals. Methods such as genomic REML (GREML), implemented in tools like GCTA [18], estimate the fraction of phenotypic variance explained by the additive effects of all measured common variants, often referred to as SNP-heritability. Conceptually, these models treat genome-wide SNP genotypes as random effects and estimate how much phenotypic similarity increases with genetic similarity, without requiring explicit family structures. This shift has allowed heritability estimation for traits where classical designs are difficult, and for large biobank cohorts where genealogical information is incomplete or absent [21]. Early SNP-heritability estimates suggested that common genotyped variants explain a substantial, but incomplete, fraction of the heritability inferred from family studies, leading to the notion of "missing heritability" [22]. Subsequent methodological work has shown that part of this gap reflects modelling assumptions about allele frequency, linkage disequilibrium (LD) structure and genotype quality. For example, models that allow SNP effect sizes to depend on minor allele frequency and LD, and that account more carefully for genotype uncertainty, tend to yield higher and more consistent SNP-heritability estimates across traits. This implies that a considerable portion of the "missing" component may arise from imperfect tagging of causal variants, incomplete variant coverage (especially of rare variants and structural variation), and over-simplified assumptions about genetic architecture. At the same time, analyses of functional annotations and regulatory features indicate that heritable variation is widely distributed across the genome: coding regions, introns and diverse regulatory elements all contribute, and thousands of loci of modest effect typically underlie complex traits. This strongly supports a highly polygenic model, where heritability is spread across many variants rather than concentrated in a small number of large-effect loci [23].

A complementary development has been the use of GWAS summary statistics themselves to infer heritability and to distinguish true polygenic signal from confounding. LD score regression [24], for instance, regresses GWAS test statistics on LD scores (a measure of how much variation each SNP tags) and uses the intercept to quantify inflation due to population stratification or cryptic relatedness, while attributing the remaining inflation to polygenic effects. This framework provides

SNP-heritability estimates that are robust to certain forms of confounding and can be applied to very large meta-analyses where individual-level genotypes are unavailable. Together, these methodological advances—from pedigree-based designs to genome-wide relatedness and LD-aware summary-statistics methods—have refined our understanding of heritability for complex traits.

In the context of this thesis, h^2 estimates provide a descriptive summary of how strongly genetic factors shape trait variation and a quantitative benchmark for evaluating downstream analyses such as genetic correlation, polygenic scores, Mendelian randomization and fine-mapping.

1.1.3 Polygenic score

A polygenic score (PGS) or polygenic risk score (PRS) quantifies an individual's genetic predisposition to a specific disease or trait by summing the weighted effects of many genetic variants and can be used as a genetic predictor for that phenotype. This method has proven to be of broad utility in the study of complex diseases.
A PGS is constructed from the estimated effect sizes derived from GWAS. The results from a GWAS estimate the strength of the association at each SNP as well as a p-value for statistical significance. A typical score is then calculated with a weighted sum of allele counts, where the weights reflect the relative magnitude of association between variant alleles and the trait or disease [25].

$$\mathrm{PGS}_i = \sum_{j=1}^{m} w_j \, G_{ij},$$

where G_{ij} is (often standardized) allele dosage for individual i at SNP j, and w_j are weights from GWAS. The associations identified via GWAS are combined to quantify genetic predisposition to a heritable trait, generating a prediction of prognostic outcomes and response to therapy, or a stratification of disease risk [26, 27].

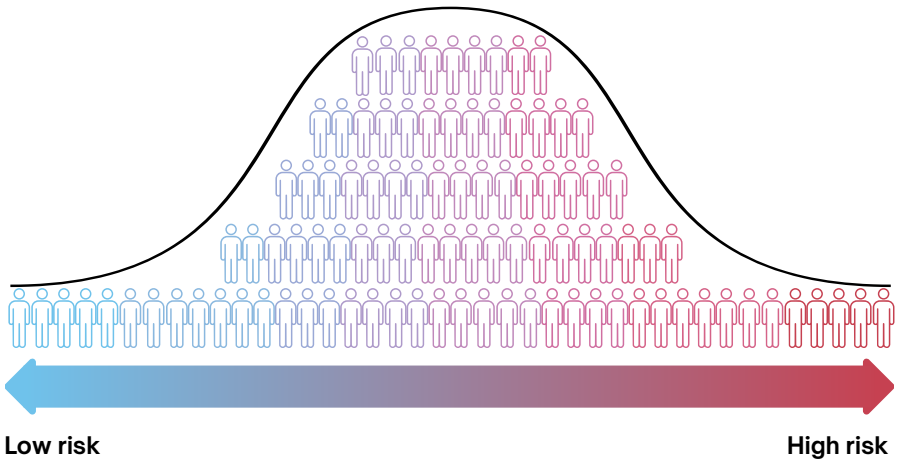


Figure 1.1: Polygenic risk score distribution.

The number of variants included and the method to weight them depend on the algorithm used, and these factors can influence the predictive performance [28].
A simple and widely used approach is "clumping and thresholding", based on the fact that nearby SNPs are not independent of each other due to LD. In other words, nearby variants are inherited together with high probability and therefore do not provide independent predictive power. In practice, SNPs in high LD are pruned so that only the most strongly associated variant is retained, and a p-value threshold is applied to exclude variants that do not reach a chosen level of statistical significance [29].
More advanced Bayesian methods, however, use external LD reference panels to infer SNP effect sizes while accounting for the genome-wide LD structure. For example, PRS-CS employs continuous shrinkage priors that conservatively shrink noisy or small effects towards zero while allowing larger effects to remain relatively unpenalized [30].

Another Bayesian approach, LDpred, uses GWAS summary statistics together with an LD matrix to jointly model SNPs in LD and to estimate posterior mean effect sizes under an assumed fraction of causal variants [31].
PRS is highly important in the clinical setting for identifying a subset of individuals whose absolute risk of disease is significantly higher than that of the average individual in the general population, and who can benefit from a health management strategy that may include intensive screening or more invasive interventions [27]. A typical application is for cardiovascular disease risk, where it has been shown that adding PRS for coronary heart disease to the Framingham risk score and the ACC/AHA pooled risk equations resulted in increased predictive power [32]; or for breast cancer risk combined with the effect of *BRCA* mutations [33].

1.2 Genetic architecture and causal inference in complex traits

While GWAS datasets have generated a wealth of replicable results, interpreting these findings and understanding the underlying molecular mechanisms are still challenging. Post-GWAS analyses have emerged as a crucial further step, focusing on the integration of intermediate phenotypes and molecular-level data [34].
The following sections outline the most commonly used methods for understanding the architecture of complex traits and their interaction.

1.2.1 Foundations of complex trait inheritance

Complex traits (or diseases) do not follow simple Mendelian inheritance, but arise from the interplay between genetic contributions, environmental exposures and lifestyle factors [35]. Typically, complex traits are influenced by a large number of genetic variants, whose single effects on the phenotype are small and mostly additive [35]. When a large number of loci each contribute small effects to a trait, the random assortment of alleles across these loci results in a continuous, normally distributed phenotype in the population [35]. GWASs and exome-sequencing studies have refined our understanding of the genetic bases of complex traits, showing that even the most important loci in the genome have small effect sizes and that,

together, the significant hits only explain a modest fraction of the predicted genetic variance [36]. This pattern is consistent with an approximately infinitesimal model, in which genetic risk is diffusely distributed across the genome and includes contributions from both common and rare variants [37]. Moreover, complex traits are mainly driven by non-coding variants that presumably affect gene regulation [38, 39, 40].

Regulatory variants can act in a tissue- and cell type–specific manner and may influence transcription, chromatin state and RNA processing, often through long-range gene regulatory networks. This has led to an expanded "omnigenic" view of complex traits, in which a limited set of core genes in disease-relevant pathways is modulated by a much larger set of peripheral genes connected through regulatory networks [37].

In such a framework, complex traits emerge as the aggregate consequence of numerous small perturbations distributed across the genome, many of which can also influence multiple phenotypes, setting the stage for the widespread pleiotropy and shared genetic architectures, which will be described in the next sections.

1.2.2 Genetic correlation

The genetic relationship between two traits is described by a quantitative parameter, namely genetic correlation (ρ_g), which reflects the pleiotropic action of genes or the correlation between causal loci. Pleiotropy occurs when a genetic locus affects multiple traits, reflecting different modes of action. Two phenotypes can be part of a causal pathway (vertical or mediated pleiotropy) or not (horizontal pleiotropy). Genetic correlation describes the average effect of pleiotropy across all causal loci, but the underlying architecture of correlations at individual loci can vary [41].

In a quantitative model, two traits, x and y, are defined as the sum of genetic value (g) and a residual value (e, difference between the trait and g):

$$x = g_x + e_x$$

$$y = g_y + e_y,$$

the genetic correlation (ρ_g) of the traits is:

$$\rho_g = \frac{\sigma_{g_x,g_y}}{\sqrt{\sigma^2_{g_x}\sigma^2_{g_y}}}$$

where σ_{g_x,g_y} is the covariance of the genetic values, and σ_{g_x} and σ_{g_y} are the genetic variances of the two traits in the population. ρ_g varies from -1 to 1.
When traits are standardized and only the additive genetic effects are considered, then the genetic variances are narrow-sense heritabilities and the genetic covariance is the coheritability (h_{xy}), so the equation becomes:

$$\rho_g = \frac{h_{xy}}{\sqrt{h^2_x h^2_y}}$$

Depending on data availability, several methodological frameworks have been developed. The earliest approaches are family-based, which use pedigree data to estimate the variance-covariance structure of the genetic relationships. In a bivariate linear mixed model (LMM), phenotypes are assumed to be bivariate normal distributions [41]. Best estimates can be obtained using restricted maximum likelihood (REML) [42]. For disease traits, an easier approach is to use a liability threshold model, which uses normal distribution theory [43].
With the advent of genome-wide SNP data, individual-level genomic methods have been widely used for unrelated individuals. The univariate genome-based REML (GREML) uses a genomic relationship matrix (GRM) from observed genome-wide SNP data to describe the variance-covariance structure of genetic values [44]. In this case, relatives are excluded from the analysis to ensure that haplotypes are tracked by the shared genetic relationships between pairs of individuals. This method assumes additive polygenic architecture, is robust to many model violations, and provides precise estimates but requires access to raw genotype data and large sample sizes.
The most widely used modern approaches rely on GWAS summary statistics. The first method to propose this was linkage disequilibrium score regression (LDSC) [24]. LDSC infers genetic covariance from the relationship between GWAS effect sizes and LD patterns, enabling unbiased estimation even in the presence of sample overlap and without individual-level data. Its efficiency has enabled the creation of

large-scale cross-trait correlation maps across hundreds of traits. Bivariate LDSC uses a weighted regression to estimate the h_{xy} for SNP j of both traits (x_j and z_j) and the SNP LD scores (l_j). The regression also depends on the sample size for the two traits (N_x, N_y) and the total number of SNPs (M). The intercept represents the proportion of shared individuals (sample overlap), where N_s is the sample overlap, and their phenotypic correlation (ρ_p):

$$\rho_p = \sqrt{h_x^2 h_y^2 \rho_g} + \sqrt{(1 - h_x^2)(1 - h_y^2)\rho_e}$$

Extensions include stratified LDSC (sLDSC) [45], which decomposes genetic correlation across functional annotations, and local genetic correlation methods such as ρ-HESS, which estimate correlations within genomic regions to uncover architecture masked by genome-wide averages [46]. Additional regression-based approaches, such as PCGC (Phenotype–Correlation Genotype–Correlation) provide robust estimation in ascertained case–control samples where standard GREML may be biased. Finally, for populations of differing ancestry, methods like Popcorn model LD and allele frequency differences to distinguish genetic-effect versus genetic-impact correlations [47]. Individual-level data enable the direct estimation of genetic variances and covariances, utilising the full information in the genotype matrix. Standard errors of GREML genetic correlations are typically ∼2x smaller than LDSC for the same sample size, making it more powerful for detecting correlations that are significantly different from zero or one. On the other hand, summary statistics-based approaches require less computational effort and resources, allowing for the achievement of larger sample sizes [41].

1.2.3 Mendelian Randomization

Genetic correlation provides evidence of shared genetic determinants between two traits; nonetheless, it does not imply the presence, direction or magnitude of any causal relationship. Genome-wide genetic correlation can arise from horizontal pleiotropy, LD, phenotypic misclassification, assortative mating or shared biological pathways. Therefore, additional methods specifically designed to explore causation are required. Randomized control trials (RCTs) remain the gold standard for causal inference; however, they are expensive and infeasible in certain settings

[48]. In such context, Mendelian randomization (MR) represents a cost-effective genetically informed alternative which leverages genetic variants as instrumental variables to strengthen causal inference. The primary rationale behind MR is that the genetic sequence is fixed from conception, thereby eliminating the possibility of reverse causation [49].
Over the past decade, MR has been widely applied to investigate a large number of causal questions, from molecular biomarkers to behavioural traits, disease risk factors and clinical endpoints [48]. The basic MR framework relies on the fact that a genetic instrument, typically a SNP, associated with an exposure X, can be used to test the causal effect of X on an outcome Y. To derive reliable causal estimates, genetic instruments (Z) must satisfy the core assumptions:

1. Relevance: Z has to be associated with X.
2. Exchangeability: Z should be independent from all confounders.
3. Exclusion restriction: there should not be a causal path from Z to Y other than through X [50].

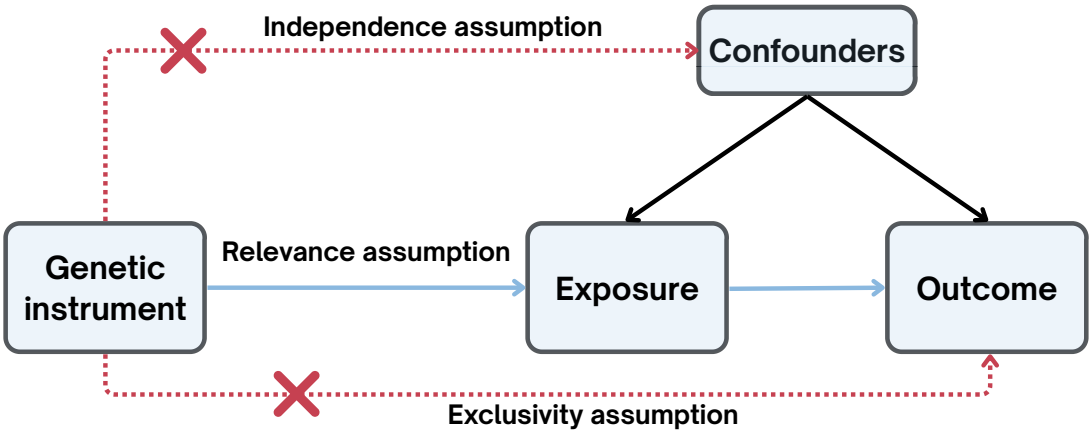


Figure 1.2: Mendelian Randomization assumptions.

Some of these assumptions are not fully testable (e.g., the presence of unmeasurable confounders), and thus they can inflate the validity of the inference. Another problem is the highly polygenic influences on most phenotypes, which

cause very small individual SNP effects, making them weak instruments. Also, polygenicity implies that pleiotropy is widespread [51].
Modern MR approaches investigate possible reciprocal causal relationships between two phenotypes. In bidirectional MR, two MR analyses are conducted on a pair of phenotypes by reversing the exposure and the outcome to establish the direction of the effect; however, results should be interpreted with caution. If there is evidence of the effect of both directions, this can indicate a true bidirectional relationship, or a bias due to horizontal pleiotropy, or confounders resulting from population stratification [52].
Another modern approach is the multivariable MR, which extends the classical one by including multiple exposures (and, respectively, multiple instruments), but each exposure must be robustly predicted by the instruments, conditional on the other exposures included in the estimation.
Moreover, there must be no confounders of the outcome and any of the instruments and none of the instruments can have an effect on the outcome that does not act through at least one of the exposures [52].
In the MR analysis, the effect of exposure on outcome can also be estimated using summary-level data through a method known as summary data-based Mendelian randomization (SMR). SMR represents a valid alternative to traditional MR when individual-level data is not accessible. It works by performing MR on summary statistics derived from independent GWASs of exposure and outcome to infer the causal pathway from genotype to transcription and then to phenotype. This framework is based on the use of expression quantitative trait loci (eQTLs) as instruments, to identify associations between gene expression (the exposure) and complex traits [53].

1.2.4 Colocalization and fine-mapping

With the identification of hundreds of thousands of genetic associations to date, a key concern is that two traits may be causally affected by separate variants that are correlated with each other through LD. This phenomenon leads to the violation of the Mendelian randomization assumption of exchangeability by providing a pathway between a genetic variant and the outcome that bypasses the exposure.

Therefore, colocalization methods have been developed especially to understand whether disease endpoints and potential mediators might share one or more causal variants [54, 55].

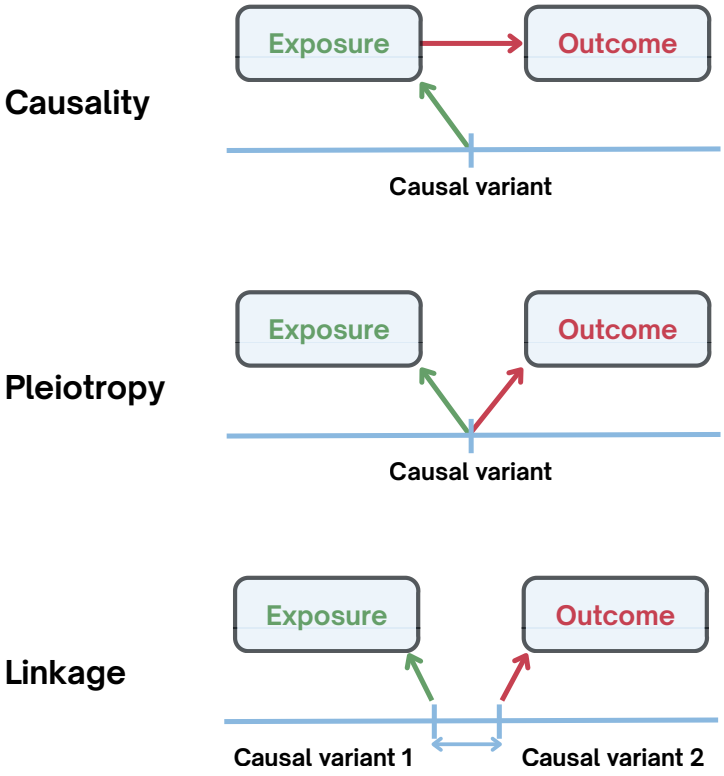


Figure 1.3: Possible relationships between exposure and outcome.

Colocalization tries to distinguish between two distinct causal variants, possibly in LD, or a single shared causal variant for two given phenotypes. This is useful for MR analysis to assess the validity of certain variants used as instrumental variables for the exposure. To define colocalization, a comparison is made between different hypotheses within a Bayesian framework, as implemented in the coloc method [55].
Given two traits and assuming that there is at least one causal variant per trait, a null hypothesis of proportionality of regression coefficients for two traits across any set of SNPs is tested. The hypotheses are:

1. H0: no association with either trait.

2. H1: association with trait 1 but not trait 2.
3. H2: association with trait 2 but not trait 1.
4. H3: association with both traits, but separate causal variants.
5. H4: association with both traits and shared causal variant.

The H4 corresponds to the situation of colocalization. Each hypothesis is calculated from a prior probability, and summary statistics from GWAS data are used to compute the posterior probabilities (PP0, PP1, PP2, PP3, PP4). Interpreting the posterior probabilities requires caution. For example, a low PP4 may not indicate evidence against colocalisation in situations where PP3 is also low. It may simply be the result of limited power, as evidenced by high values of PP0, PP1, and/or PP2. Moreover, a high PP4 is a measure of correlation, not causality [55].

Colocalization and MR share some similarities: the same statistical method can be implemented. However, MR focuses on estimating the slope, assuming a line exists, while colocalization focuses on the variability of errors in the model by determining if a line can be drawn. Moreover, genetic variants included in a MR analysis are selected based on their association with the exposure, and they are included even if not directly linked to the outcome. Instead, colocalization analysis focuses on a region of the genome where both traits exhibit signals of association, aiming to determine whether these signals are caused by the same genetic variants [56].

Fine-mapping enhances both MR and colocalization by refining the genomic regions of interest from a GWAS to a credible set of associated variants that are more likely to include the causal variance. The performance is influenced by several factors, such as the number of causal SNPs in a locus and their effect size on the phenotype, the local LD structure, sample size and SNP density [57].
Fine-mapping is particularly useful when multiple genetic variants in high LD are associated with a phenotype, making it difficult to discern which variant is truly causal. By incorporating information about the local LD structure, fine-mapping methods help to disambiguate these variants, providing more accurate estimates

[57].
Several methods have been implemented to perform fine-mapping, producing, as in colocalization, a posterior probability that a variant in that locus is causal. These methods include Bayesian approaches, such as FINEMAP [58], which estimates the likely set of causal variants given summary statistics from GWAS. CAVIAR [59] and SuSiE [60] are other examples of fine-mapping tools that use Bayesian frameworks to assign posterior probabilities to individual variants based on their potential to explain the association signals observed in GWAS data. An adaptation of coloc, known as coloc-SuSiE, enables multiple labelled comparisons within a genomic region, providing an approach to colocalization in cases of multiple causal variants. While traditional coloc assumes a single causal variant responsible for the observed associations, coloc-SuSiE leverages the SuSiE model to allow for multiple causal variants in the same region. This is particularly useful in genomic regions where multiple variants in high LD are likely to contribute to the observed signals, making it difficult to resolve a single causal variant [61].

1.3 Multi-omics approaches in biological research

Multi-omics approaches combine data from multiple molecular layers—such as genomics, transcriptomics, metabolomics, proteomics, and epigenomics, to offer a more realistic view of biological systems. By leveraging advancements in high-throughput technologies and computational tools, multi-omics has revolutionized biomedical research, offering a promise to fill the gaps in the understanding of human health and disease.
The following sections will describe the key concepts and methodologies underlying multi-omics approaches, with a focus on their application in biological research.

1.3.1 The advantages of multi-omics

The term "multi-omics" refers to a biological analysis approach that encompasses multiple sources of data representing different molecular layers, including genomics, transcriptomics, metabolomics, proteomics, and epigenomics, but also exposomics. By combining these data types, researchers can address complex questions and

uncover novel associations. This approach has revolutionized precision medicine research, enabling the discovery of new biomarkers, unveiling previously unknown disease mechanisms, and broadening our general understanding of biological systems [62].
The omics field has been driven largely by recent advances in technology, which have made high-throughput analysis possible and cost-effective. At the same time, progress in computational biology and machine learning has led to the development of sophisticated integration frameworks capable of managing large, heterogeneous datasets [63].

Compared to single omics, multi-omics can provide a more comprehensive flow of information and relative interactions. Indeed, multi-omics data integration has emerged as a powerful approach to grasp the intricate interplay between different biomolecular layers that contribute to disease states, and thus to uncover more effective strategies for diagnosis, treatment, and prevention. For example, the integration of genomic and transcriptomic data enables the identification of genetic variants that influence changes in gene expression associated with disease. Furthermore, combining epigenomic data (e.g., DNA methylation, ChIP-seq, and ATAC-seq) with transcriptomics within regulatory network frameworks enhances our understanding of how environmental exposures influence disease phenotypes [62].
In some Mendelian diseases, typically driven by a single causal variant, such variant may be difficult to detect. Integrating additional information, such as RNA sequencing (RNA-seq) or network analyses, can be useful for detecting molecular events that prioritize among likely causal variants or provide evidence of pathogenicity for a candidate mutation [64].

Multi-omics approaches are generally based on statistical inference from large datasets. Therefore, the power to detect associations depends on effect size, heterogeneity of the background noise and sample size [9].
Different approaches exist for integrating various omics. A first distinction is between vertical and horizontal integration. Vertical integration is the most common approach, consisting of combining different omics data for the same set of biological

samples to build an individual-based molecular profile. The horizontal integration, on the other hand, combines data from different studies or cohorts to increase statistical power and validate results on a larger scale [62, 65].

Another way to classify integration approaches is to distinguish between sequential and parallel integration. In the first one, datasets are analyzed separately and only combined in a later stage, whereas in the parallel integration, all datasets are analyzed together within a single computational framework. In general, methods that integrate data sequentially are more common in addressing questions related to disease classification, while integrating data simultaneously is often used for biomarker prediction tasks [62].

Recently, network-based methods have become a natural framework for integrating multi-omics data, given their modelling of molecules and interactions as graphs. Networks facilitate the capture of the interplay between omics layers by anchoring heterogeneous measurements to shared interaction structures, enabling the transition from lists of differentially expressed molecules to systems-level interpretations of complex disease mechanisms [66].

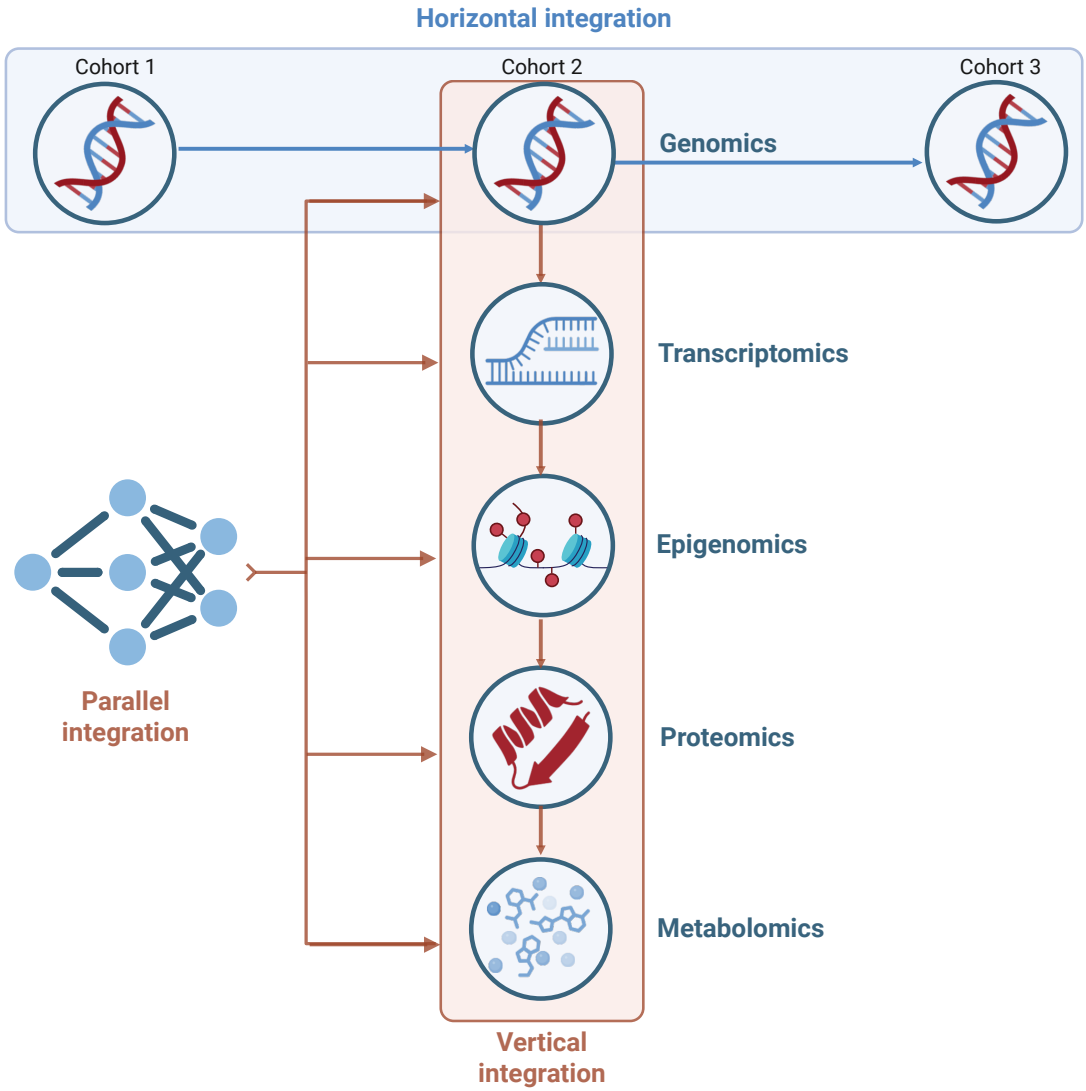


Figure 1.4: Multi-omic data integration.

Despite their potential, the integration of multi-omics data still faces challenges. One main issue is the heterogeneity of integrating datasets from different studies, where different sources of variation, which do not depend on biological variation, namely "batch effects", can compromise the signals and reproducibility of results and obscure true biology. Second, human healthy and disease groups are intrinsically heterogeneous (in terms of population structure, environment, and cell-type composition); therefore, very large, well-phenotyped reference populations are

needed to separate disease signals from confounding background variation. Third, analytical hurdles include the lack of robust, reproducible statistical frameworks that can jointly model multiple disparate data types, distinguish causal from reactive changes, and scale to thousands of samples while remaining interpretable and clinically usable. Fourth, computational demands become substantial as multi-omics, single-cell, and spatial datasets grow, and are compounded by privacy and data-sharing constraints that motivate the development of distributed integration strategies. Finally, even when computational integration succeeds, a major translational gap remains: results must be rigorously validated in independent cohorts and experiments, integrated with electronic health records, and demonstrated to be cost-effective and generalizable across populations before they can be adopted in clinical practice [64, 62, 9].

1.3.2 Quantitative trait loci

A natural bridge between genetic variation and different molecular layers is represented by quantitative trait loci (QTLs). A QTL is a locus that correlates with variation of a quantitative trait, therefore linking two sources of information: genomic and phenotypic data [67].
Most expression traits are shaped by a complex genetic architecture involving multiple QTLs, each of which typically accounts for only a small fraction of the overall trait variation [68].
Among molecular QTLs, expression QTLs (eQTLs) were the first to be extensively characterized, which refers to traits measured by gene expression. Each transcript corresponds to an encoding gene whose genomic position is known. Thus, once mapping studies identify the locations of QTLs, an eQTL can be classified as local if it is near the gene that produces the transcript or distant if it is elsewhere in the genome. Local eQTLs often reflect variants affecting cis-regulatory elements, whereas distant eQTLs tend to arise from trans-acting factors such as transcription regulators or signaling pathways [68].
The combination of GWAS and the measurement of global gene expression allows the systematic identification of eQTLs.
Early eQTL studies collected gene expression data using microarrays, where gene

expression levels need to be normalized to remove batch effects, and the normalized data are analyzed to identify eQTLs. A common way to model the relationship between a SNP and a gene's expression level is a simple linear regression, where expression is explained by the SNP's genotype. The error term in this model is typically assumed to follow a normal distribution. Covariates such as age, sex, or technical factors can also be added to the model to account for non-genetic influences on expression [69].

A typical eQTL study considers more than 20,000 genes and up to millions of SNPs. Because of the large number of SNPs to be tested, researchers often focus on cis-eQTLs for a given gene, given also to the statistical power differences in detecting cis- versus trans- eQTLs. In fact, trans-eQTLs often have smaller effect sizes than cis-eQTLs, and thus require larger sample sizes to be detected [69].

With gene expression data collected through RNA sequencing, such as those from the GTEx project (ref GTEx), the measured gene expression level is the total number of sequence reads mapped to a specific gene, which needs to be adjusted for total sequencing depth and other factors. In this case, other distributions, like a negative binomial, can be more appropriate for modelling the associations. A generalized linear regression model may have better statistical power [69].

Beyond bulk data, studies have been conducted using purified cells of different cell types to infer cell-type-specific eQTLs (ct-eQTLs). The same statistical methods for bulk samples can be applied to infer ct-eQTLs from these data. However, the sample size tends to be smaller, and the measurement noises may be higher. In addition, the purified cell types may be contaminated with other cell types [70, 71, 69, 72, 73].

The identification of eQTLs has been pivotal for interpreting the genetic basis of complex traits, as these loci can provide insights into regulatory mechanisms that drive phenotypic variation. Furthermore, eQTL maps are increasingly used in conjunction with GWAS to identify genes that are causally implicated in disease via techniques such as colocalization and transcriptome-wide association studies (TWAS), helping to bridge the gap between genetic variation and clinical outcomes [74].

Despite the association between genetic variation and complex traits having provided key insights into the heritable component of inter-individual variation, mRNA

levels represent only one step along the path to phenotype, and a substantial proportion of variation is mediated through other intermediate molecular phenotypes, including splicing (sQTLs), methylation of CpG sites (meQTLs), chromatin accessibility (caQTLs), protein expression (pQTLs), and others. For molecular QTLs too, generally, associations within a 1 Megabase (Mb) window surrounding each locus are defined as cis, whereas associations outside this window are considered as trans [75].
In the following section, the focus will be on pQTL studies of the plasma proteome and how they have been leveraged to map disease pathways and identify potential drug targets.

1.3.3 The plasma proteome as a window into disease biology

Protein quantitative trait loci (pQTLs) provide a complementary layer of molecular QTLs by quantifying how variants influence circulating protein concentrations while including the effects of post-transcriptional mechanisms. Plasma proteins play key roles in many biological processes and are often dysregulated in disease mechanisms, making them important biomarkers [5].
Precision medicine has so far been driven largely by the identification of genomic determinants of human disease, with early evidence of clinical utility [76, 77].
However, the complex and context-dependent regulation that links genotype to phenotype through transcription and translation often complicates the identification of causal genes, limiting mechanistic insight and the translation of genome-to-phenome associations into drug targets [78, 79].
Proteins are the primary effectors of genetic and environmental influences on disease, capturing ongoing biological processes and pathophysiological changes in the body. Consequently, defining protein–disease relationships can help delineate molecular signatures of health and disease, supporting more actionable and scalable precision medicine approaches [77].
In the last decade, many studies focused on the characterization of a genetic atlas of the human plasma proteome [80, 5].
Several consortia have invested in large population biobanks to advance genetics-

guided drug discovery, complementing massive phenotype–genotype resources, such as the UK Biobank (UKB) [81, 82], with deep multi-omics profiling of biological samples [83, 84]. Large biobank initiatives, such as the UKB Pharma Proteomics Project (UKB-PPP), enable the construction of extensive plasma pQTL resources to support biomarker discovery and genetics-guided drug development. The UK-PPP consortium, comprising 13 biopharmaceutical companies, conducted plasma proteomic profiling on 54,219 UKB participants using the antibody-based Olink Explore 3072 PEA, identifying a large number of cis- and trans-pQTL [85].
The Olink technology uses Proximity Extension Assay [86], where a matched pair of antibodies labelled with unique complementary oligonucleotides (proximity probes) binds to the respective target protein in a sample. Alongside aptamer-based assays, such as SomaScan [87], Olink is among the most widely used high-throughput platforms for population-scale plasma proteomics. However, the two technologies differ in molecular recognition and thus in the interpretation of pQTLs. Aptamer-based methods quantify proteins through binding of sequence-defined oligonucleotide aptamers, enabling broad proteome coverage and sensitive measurement from small plasma volumes.
Colocalization with eQTLs in individual tissues has demonstrated that pQTLs in plasma reflect the transcriptome in different tissues [88].

Many plasma pQTLs have been found to overlap established disease-susceptibility loci, and Mendelian randomization analyses supported putative causal roles for several protein pathways; integration with disease and drug databases further highlighted candidate therapeutic targets [5, 85].
Recently, a comprehensive atlas of proteome-phenome associations has been published, which, by integrating pQTL data, systematically maps 2,920 plasma proteins to the presence and onset of hundreds of diseases and health-related traits [77].

A key challenge in proteomic analyses is that they frequently uncover associations between circulating proteins and disease; however, these relationships alone cannot distinguish between causal mediators and downstream or reactive changes, thereby limiting their utility for the discovery of therapeutic targets. One strat-

egy to strengthen causal inference is to integrate pQTLs with disease GWAS data and apply proteome-wide MR, which uses inherited genetic variation as an instrumental proxy for long-term differences in protein levels. In a proteome-by-phenome MR framework, genetic instruments (typically prioritising cis-pQTLs to reduce pleiotropy) are leveraged to systematically evaluate the predicted effects of modifying thousands of circulating proteins across a wide spectrum of disease outcomes. Notably, MR-based prioritization of protein–disease associations partially recapitulates established target–indication relationships documented in drug development databases, supporting the broader observation that targets with human genetic support have a higher likelihood of clinical success. At the same time, these analyses nominate additional protein–disease relationships that are not represented among approved indications, highlighting potential opportunities for drug repurposing or novel target development. Nevertheless, such findings require careful interpretation, including sensitivity analyses to assess pleiotropy, replication in independent cohorts, and experimental follow-up to establish underlying biological mechanisms [77].

1.4 Machine Learning approaches in complex traits and disease prediction

As multi-omics data becomes increasingly central to understanding complex traits and diseases, the integration of diverse data types has opened new avenues for predicting disease risk. However, the high dimensionality of these datasets presents a significant challenge, particularly when combining features that capture regulatory activity, genetic variation, and clinical phenotypes. The following sections provide examples of how to apply machine learning techniques to address these challenges and effectively leverage multi-omics data for complex traits.

1.4.1 Predicting regulatory activity and chromatin accessibility in complex traits

A support vector machine (SVM) is a supervised machine learning model most commonly used for binary classification (e.g. regulatory vs non-regulatory se-

quences). Conceptually, an SVM tries to find a decision boundary (a hyperplane) that separates positive from negative examples while maximizing the margin between the two classes. In practice, SVMs become very powerful when paired with a feature representation (e.g., how DNA sequence is encoded) and/or a kernel (a similarity function that allows flexible decision boundaries) [89].

One example application is gapped k-mer SVM (gkm-SVM), in which SVM is used to distinguish cell type–specific regulatory sequences from matched negative genomic sequences, using sequence-derived features, and the learned model can then be used as a quantitative scoring function for regulatory activity.

gkm-SVM (gapped k-mer SVM) [90] is a sequence-based SVM where DNA is represented by counts/presence of gapped k-mers (short sequence "words" that allow gaps), which helps capture transcription factor–like sequence patterns more flexibly than contiguous k-mers. The model is trained on a positive set of regulatory regions (for example, DNase I hypersensitive sites in a specific cell type) and a negative set of matched control sequences (matched for length/GC/repeats). Once trained, the gkm-SVM provides a scoring function that reflects how strongly a sequence matches the learned "regulatory vocabulary." This enables the identification of regulatory elements that are specific to particular cell types or conditions. Then, a deltaSVM is evaluated, which quantitatively predicts the functional effect of a variant by computing the change in model score between alleles. Operationally, deltaSVM is computed by summing the differences in learned weights across the set of overlapping sequence words (in the authors' implementation, 10-mers spanning the variant), so that variants disrupting (or creating) high-weight regulatory words produce large negative (or positive) changes in score. The workflow is: (1) train a cell type–specific gkm-SVM on regulatory vs control sequences, (2) derive the weighted sequence vocabulary, and (3) compute deltaSVM for each candidate variant as the predicted impact on regulatory activity. Empirically, the authors validate that deltaSVM tracks measured regulatory effects, including strong correlation with DNase I–sensitivity QTL (dsQTL) effect sizes and agreement with results from targeted and massively parallel enhancer mutagenesis assays [90] .
Chromatin accessibility and transcription factor binding sites influence gene ex-

pression, which in turn can have a profound impact on phenotypes. The regulatory activity of specific genes, as predicted by models such as deltaSVM, can help elucidate the underlying molecular mechanisms of complex traits and their interactions.

In this context, the integration of regulatory activity in complex traits through expression quantitative trait loci (eQTL) will be explored in Chapter 2, in the work titled "Exploring the molecular basis of the genetic correlation between body mass index and brain morphological traits".

1.4.2 Feature selection and regularization for predicting disease risk

Disease prediction models, in particular for complex diseases, often face the challenge of dealing with high-dimensional, sparse data. The sheer number of potential predictors, which can span across various data types such as diagnosis codes, medications, laboratory test results, polygenic risk scores, and multi-omics data (proteomics, genomics, metabolomics), can overwhelm the ability of traditional statistical methods to produce meaningful and generalizable insights. At the same time, the relatively small number of disease events for many outcomes, particularly in rare diseases, creates further difficulty for these models, as it exacerbates the issue of overfitting [91].

Feature selection is a crucial step in mitigating this issue. By narrowing down the predictors to those that are most informative and consistently associated with the outcome of interest, feature selection can improve both model efficiency and performance. The goal is to identify a set of features that not only correlate with the outcome but also capture reproducible signals that are not specific to the training dataset or cohort, helping the model generalize better across different populations or healthcare settings. Moreover, feature selection can contribute to clinical interpretability. For example, by selecting a small, clinically relevant subset of features (such as age, sex, lab values, or known disease biomarkers), clinicians can gain insights into which risk factors are most predictive, facilitating easier integration into clinical practice and decision-making [91].

In the context of multi-omics data, where features are often highly correlated, effective feature selection can help mitigate the multicollinearity that arises between omic layers, which is a common challenge when integrating data from genomics, transcriptomics, proteomics, and metabolomics. For instance, genes within the same pathway might show similar patterns of expression or interaction, and without proper feature selection, a model might focus disproportionately on a small subset of features, leading to overfitting [91].
Embedded methods for feature selection perform the process during model training, making them particularly effective in predictive tasks. Unlike filter methods, which evaluate features independently of the model, embedded methods integrate feature selection directly into the fitting process. This means that the model automatically determines the most relevant predictors based on its internal optimization criterion, typically a loss function that is designed to minimize prediction error or some other performance metric. Some examples of embedded methods include decision tree-based algorithms (e.g., decision tree, random forest, gradient boosting), and feature selection using regularization models (e.g., LASSO or Elastic Net) [91].
Regularization constrains model complexity by adding a penalty to the loss function, trading a small increase in bias for a large reduction in variance, often improving calibration and transportability. L1 regularization (LASSO) induces sparsity by setting many coefficients to zero, effectively removing less informative features from the model. While this can produce simpler, more interpretable models, LASSO is sensitive to correlated predictors. In cases where features are highly correlated, LASSO may arbitrarily select one feature over others, which could result in a less stable model across cohorts [92].

In Chapter 3, the application of LASSO for feature selection and regularization will be discussed in the work titled "Removing genetic effects on plasma proteins enhances their utility as predictive disease biomarkers", where it has been used to predict diseases using a selected set of proteins.

1.5 Population Biobanks to study complex traits

Large-scale biobanks are pivotal to advancing biomedical research and public health by providing researchers with comprehensive, high-quality data on genetic, environmental, and health-related factors.
The following sections will examine the substantial impact of large-scale biobanks on biomedical research and public health, highlighting how these initiatives address health inequities, support targeted clinical trials, and accelerate drug development. The legal and ethical challenges will be discussed, particularly in areas such as ownership, informed consent, privacy, and data security.

1.5.1 The emergence of population-scale biobanks

Biobanks are large, systematically curated biorepositories that collect biological samples together with extensive phenotypic information, such as electronic health records, disease registries, laboratory measurements, imaging, lifestyle questionnaires, and genome-wide genotyping or sequencing, for use in research. Their data are the product of standardized and ethically governed collection. Participants are typically recruited through voluntary enrollment, provide informed consent for broad research use, and are followed longitudinally through periodic questionnaires, measurements, or linkage to national health registries. This ensures both data consistency and long-term reusability across research domains.
Modern biobanks also integrate multi-omics (proteomics, metabolomics, transcriptomics) and repeated measurements, providing rich training data for machine learning models that aim to predict health trajectories or infer variant function by connecting molecular profiles to clinical outcomes. Their scale and longitudinal follow-up make it possible to study the genetic architecture of complex traits, discover biomarkers, and quantify disease risk across the life course, often in real-world clinical settings.

The UK Biobank (https://www.ukbiobank.ac.uk/) is one of the world's largest population-based biobanks, comprising over 500,000 participants across the United Kingdom. It has become a cornerstone for investigating the genetic, environmen-

tal, and lifestyle factors that influence health outcomes. Participants provided detailed health data, including lifestyle questionnaires, physical measures, and biological samples. Coupled with genome-wide genetic data, this resource has been instrumental in studying a wide range of diseases, identifying genetic risk factors, and offering insights into disease prevention and personalized medicine. The UK Biobank's large and diverse sample size allows for comprehensive genetic analyses, while its integration with national health registries enables long-term tracking of health outcomes, making it a pivotal tool for population health research.

FinnGen [2] is a nationwide research initiative that is particularly notable for its integration of genomic data with longitudinal health data from Finland's extensive registries. With over 520,000 participants and growing, FinnGen represents a unique opportunity to investigate the genetic basis of disease in a population with relatively homogeneous genetic backgrounds. Sample collection started in 2017 and data releases and analyses will continue through 2027. The Finnish healthcare system and the availability of detailed, long-term health records allow for in-depth studies on rare diseases, complex traits, and drug responses. FinnGen is a key resource for studying the genetic underpinnings of diseases like cancer, cardiovascular diseases, and neurological disorders, providing insights that are crucial for improving healthcare outcomes and personalizing treatments.

The Genotype-Tissue Expression (GTEx) [93] project focuses on understanding how genetic variation influences gene expression across different tissues in the human body. GTEx was launched in 2010 and it has collected and analyzed genetic and transcriptomic data from up to 1000 postmortem donors and 50 tissues, spanning a wide array of organs.
The version 8 (v8), with 838 donors and 15,201 samples, released a catalog of genetic regulatory variants affecting gene expression and splicing in cis and trans across 49 tissues [94].

1.5.2 Advantages and limitations

Large-scale biobank databases have revolutionized biomedical research given their size, overcoming traditional sample size constraints, and offering valuable insights into health indicators that are often underrepresented or absent in national datasets. They enable estimates for smaller population groups that are frequently excluded from many studies, whether due to limited sample size, specific geographic areas, or time-based limitations in research. With their comprehensive and diverse data, they have the potential to drive public health solutions tailored to the needs of various populations [95].

Addressing health inequities by developing effective policies and programs for marginalized populations is a crucial goal, yet there is limited evidence on what interventions work for groups facing health disparities. The "All of Us" initiative, for example, targets populations underrepresented in biomedical research, considering factors like ethnicity, sexual orientation and gender identity, geographic location, access to healthcare, socioeconomic status, and disability. With over 80% of its participants identifying with one or more of these groups, the data from All of Us is invaluable for understanding and addressing health disparities within and across these populations [96]. Furthermore, the data's volume and geographic diversity offer opportunities for public health practitioners to tailor evidence-based interventions to specific regions or demographic groups [97].

These initiatives enable cross-sector collaboration and enhance the responsiveness of public health systems [95].

By providing well-characterized patient biosamples, biobanks accelerate the drug development process, reducing both the time and costs associated with clinical trials. Furthermore, biobanks help identify patient subpopulations that may respond differently to specific drugs, enabling more targeted and effective clinical trials [98].

However, situated between research and clinical units, biobanks face many legal and ethical challenges throughout their development that must be addressed to ensure both scientific progress and care for participant rights'. First of all, determining the ownership of biological samples has been a longstanding concern. While samples may be used for research, drug discovery, or treatment develop-

ment, questions around the commercial benefits and the rights of donors persist. Legal clarity on ownership, especially regarding sample transfers and posthumous use, is essential for biobanks. https://doi.org/10.1038/ejhg.2014.45 Informed consent is another key issue, since it is crucial but it presents challenges due to the evolving nature of research. Traditional consent models, which limit sample use to a specific project, are insufficient for biobanks, which require broader, more flexible consent forms [99, 100].
Ensuring the privacy and security of donor data is vital, particularly when storing biological samples over long periods. Simple anonymization is not enough; pseudonymization and secure coding systems are commonly used. However, in cases of accidental findings, researchers must be able to contact donors while maintaining privacy [101].
The storage of whole genome sequences raises concerns about data security and the potential misuse of sensitive genetic information. While open access to genomic data can enhance research, it must be balanced with the protection of donors' privacy and rights [102].

The General Data Protection Regulation (GDPR), which came into effect in May 2018, has had a significant impact on biobanks, particularly in the European Union. GDPR is designed to protect the privacy and personal data of individuals and harmonize data protection laws across Europe. For biobanks, GDPR provides a comprehensive framework for handling personal and sensitive data, which includes genetic and health data collected from participants. The implementation of GDPR in biobanks has significantly strengthened the protection of personal data while promoting transparency, accountability, and ethical research practices.

CHAPTER 2

Exploring the molecular basis of the genetic correlation between body mass index and brain morphological traits

Obesity is linked to many chronic diseases and its prevalence worldwide is increasing. Susceptibility to obesity is known to be due, to some degree, to genetic factors, and such genetic contribution is mediated in part by the brain through the control of food intake. In this work, published in 2025 at *PLOS Genetics*, we investigated the genetic variants that affect both obesity and brain morphology, identifying 21 genes whose expression affects both body mass index and morphological measures obtained from brain magnetic resonance imaging data. These results provide new insight into the complex relationships between the brain and obesity, also in view of the recent development of revolutionary anti-obesity drugs, whose mechanisms of action are believed to include effects on the reward-related areas of the brain.

2.1 Abstract

Several studies have demonstrated significant phenotypic and genetic correlations between body mass index (BMI) and brain morphological traits derived from structural magnetic resonance imaging (sMRI). We use the sMRI, BMI, and genetic data collected by the UK Biobank to systematically compute the genetic correlations between area, volume, and thickness measurements of hundreds of brain structures on one hand, and BMI on the other. In agreement with previous literature, we find many such measurements to have negative genetic correlation with BMI. We then dissect the molecular mechanisms underlying such correlations using brain eQTL data and summary-based Mendelian randomization, thus producing an atlas of genes whose genetically regulated expression in brain tissues is pleiotropic with brain morphology and BMI. Fine-mapping followed by colocalization analysis allows, in several cases, the identification of credible sets of variants likely to be causal for both the macroscopic phenotypes and for gene expression. In particular, epigenetic fine mapping identifies variant rs7187776 in the $5'$ UTR of the *TUFM* gene as likely to be causal of increased BMI and decreased caudate volume, possibly through the creation, by the alternate allele, of an ETS binding site leading to increased chromatin accessibility, specifically in microglial cells.

2.2 Introduction

In 2022 one out of eight people were diagnosed with obesity worldwide, amounting to over 890 million adults and 160 million children and adolescents living with this chronic complex disease. In the majority of cases, obesity is a combination of environmental, psycho-social, and genetic factors. If the current increasing trend continues, 60% of the entire human population is estimated to be overweight or obese by 2030 [103].

The central nervous system has a role in susceptibility to obesity through the control of food intake [104]. This notion is strengthened by the fact that the heritability of body mass index (BMI), a parameter typically used by the World

Health Organization to distinguish normal-weight ($18.5 \leq \text{BMI} \leq 25$) from obese people ($\text{BMI} \geq 30$), was found to be enriched in genes expressed in the brain and central nervous system [45, 105]. In particular, a recent enrichment analysis of the genes located near 97 BMI-associated single-nucleotide polymorphisms (SNPs), aimed at identifying tissues and cell types with high expression of such genes, revealed that 27 out of 31 significantly enriched tissues were part of the central nervous system. However, these results are not enough to identify the specific brain regions involved [105].

Beyond the aforementioned results, several neuroimaging studies investigated the correlation between obesity and structural alterations in brain regions, especially differences in gray matter volumes. While the results of these studies have been somewhat conflicting, with both positive and negative correlations reported between gray matter volumes and obesity, most neuroimaging studies reported reduced gray matter volume in many brain regions, including several involved in executive control, to be associated with higher BMI [106]. In particular, reduced volume of cerebellum [107, 108], basal ganglia [109], putamen [108], prefrontal cortex, temporal lobes, and subcortical structures [110], as well as reduced cortical thickness, were found to be associated with higher BMI [111, 112, 113, 114].

On the other hand, a positive correlation between gray matter volume and BMI was found in nucleus accumbens and hypothalamus [115], while a twin study identified positive correlation with ventromedial prefrontal cortex and the right cerebellum [110].

The relationship between BMI and white matter integrity is more difficult to characterize since alterations show a more complex pattern [116]. Studies involving diffusion tensor imaging (DTI) investigated the influence of obesity on white matter integrity by comparing BMI with DTI parameters in adults, revealing a negative correlation between several microstructure architectures of the white matter, such as in the corpus callosum, and BMI [116].

To date, most studies have focused on phenotypic correlation. Regarding genetic correlation, a study explored the effects of a collection of obesity-related SNPs on 164 regional brain volume traits from the UK Biobank, finding that 17 such SNPs were associated with 51 regional brain volumes (both positively and negatively) [117]. A bivariate linkage and quantitative analysis on Mexican

American individuals looked for genetic factors associated with increased BMI and reduced cortical surface area and subcortical volume, localizing two genome-wide significant QTLs at 17p13.1 and 3q22.1. The former pleiotropically affects ventral diencephalon volume and BMI, suggesting the involvement of such region in obesity through leptin-induced signaling in the hypothalamus. The latter involves the surface area of the supramarginal gyrus and BMI and might be relevant to the food-related reward mechanism [118].

Finally, a study integrating single-cell-RNA-sequencing and genome-wide association studies (GWAS) has shown that susceptibility to obesity is enriched for some hypothalamic cell types such as VMH Sf1-expressing neurons, most of which are involved in the integration of sensory stimuli, learning and memory. Despite neuroanatomical differences, these brain cell types share transcriptional patterns related to obesity [119].

Here, we systematically investigated the genetic correlation between BMI and brain morphology using GWAS summary data from the UK Biobank. To investigate the molecular mechanisms behind these correlations, we used summary-based Mendelian randomization (SMR) to identify the genes whose genetically regulated expression (GReX) in brain tissues is pleiotropic with BMI and brain morphology, suggesting a mediating effect. Finally, we used fine-mapping and colocalization to identify cases in which a single variant is causal of gene expression, BMI, and brain morphology, and epigenetic fine-mapping based on single-cell chromatin accessibility data to formulate mechanistic hypotheses on the relevant regulatory mechanisms and cell types.

2.3 Results

2.3.1 Genetic correlation

In order to assess the shared genetic basis between brain morphology and BMI, we relied on variant-trait associations for 435 cortical and subcortical measurements estimated on more than 30 thousand donors from the UK Biobank [120]. These measurements included volume, thickness, areas from white and pial surfaces, gray-white matter contrast for 54 cortical regions [121]; volumes and mean intensities

for subcortical segments [122] (see Methods for details on the selection of traits).

Adopting cross-trait LD Score regression [123], the state-of-the-art summary-based approach, we found 108 nominally significant ($P < 0.05$) genetic correlations (r_g) across all categories. Most of the significant correlations turned out to be negative (Fig. 2.1A), including in particular those between BMI and cortical areas (white and pial surfaces), cortical and subcortical volumes. The most significant correlations ($P < 0.001$ in at least one hemisphere) were all negative (Fig. 2.1B), suggesting that, overall, genetic variants associated with increased BMI are also related to reduced size of brain structures. This is evident from the global subcortical volume and the globus pallidus in particular, and the total areas of pial and white surfaces, with especially negative r_g values in cortical regions adjacent to the central sulcus, and in the temporal and cingulate lobes (Fig. 2.1B). As expected, results for the two hemispheres are largely consistent with each other (Fig. 2.1C).

Subcortical intensity traits, as provided by the UK Biobank (UKB), were analyzed, revealing a significant negative genetic correlation for brainstem intensity ($P<0.001$). However, interpreting this association is not straightforward. Although intensity values partially reflect what voxel-based morphometry (VBM) measures, VBM leverages the optimized intensity contrast among gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) tissues from T1-weighted images to classify voxels and estimate tissue volumes. Voxel intensity itself, however, can be influenced by MRI scanner artifacts and other sources of noise, introducing biases which are difficult to control. To address this, studies employing this method typically perform statistical analyses after normalizing tissue intensities, thereby estimating the contrasts between different tissues [124].

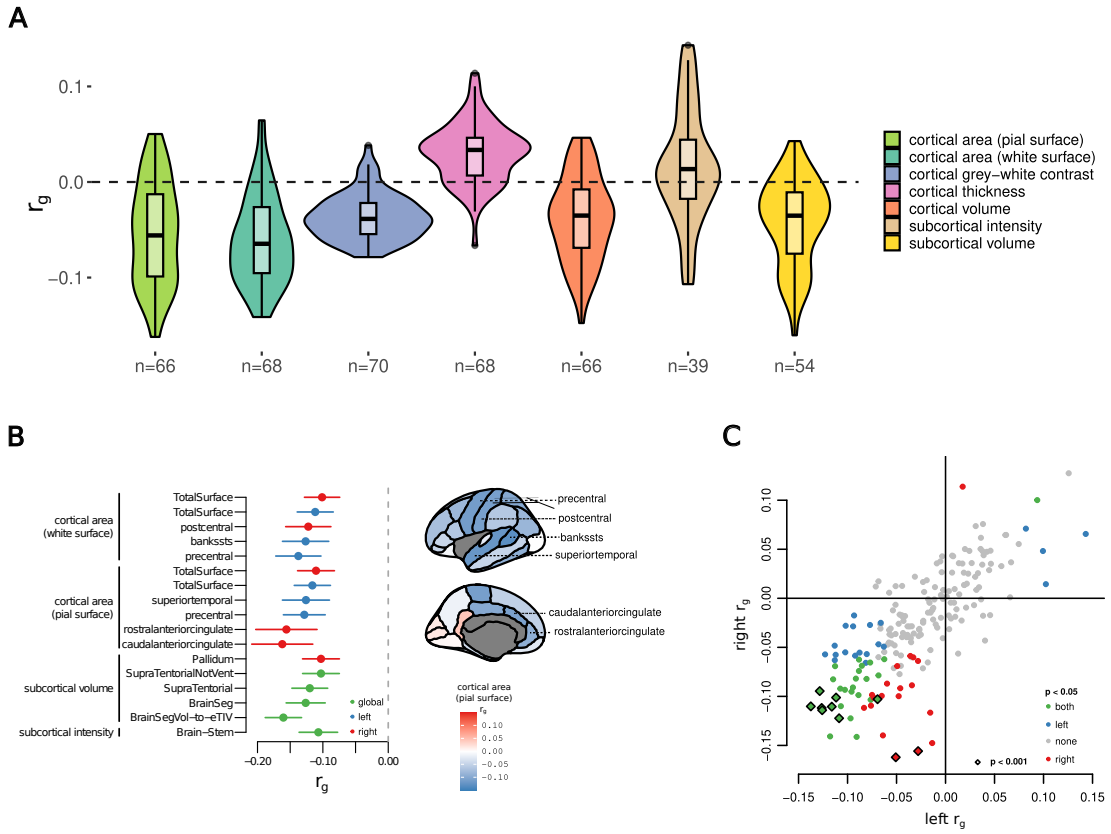


Figure 2.1: Genetic correlations between sMRI traits and BMI. (A) Genetic correlations (y axis) of all 435 sMRI traits with BMI separated by category. n, number of traits for each category. (B) Traits with the most significant ($P < 0.001$) genetic correlation (x axis) with BMI, grouped by category. A few subcortical volume traits were removed to avoid redundancy. Measurements refer to the left (blue) or right (red) emisphere, or are global (green). Significant cortical regions are indicated on the brain surfaces on the right, which are colored according to their genetic correlation between pial surface area and BMI (left-right averages). (C) Comparison between genetic correlations in left (x axis) and right (y axis) hemispheres. Dots are colored based on their significance level. Diamonds represent correlations with $P < 0.001$ in either or both hemispheres.

2.3.2 Summary-based Mendelian randomization

We hypothesized the existence of molecular mechanisms at the level of gene expression underlying these genetic correlations. Mendelian randomization [52] can help elucidate the causal effect of an exposure on an outcome. Thanks to large

databases of expression quantitative trait loci (eQTLs), gene expression can be used as an exposure, allowing the identification of putative genes whose regulation is pleiotropically related to a complex trait, suggesting a mediation effect of their expression. We used summary-based Mendelian randomization (SMR) [53] with gene expression in each of 13 brain tissues included in the GTEx database [4] as the exposure and sMRI traits and BMI as outcomes, reasoning that genes associated with both sMRI and BMI in such analysis could provide a basis for the investigation of the molecular mechanisms driving the genetic correlation between the two types of traits.

SMR is unable, by itself, to distinguish between a causal relationship and a pleiotropic one, that is when gene expression and complex trait are influenced by the same genetic variants but not causally related. In addition, it can identify relationships which are actually mediated by different causal variants for exposure and outcome, when they are in linkage disequilibrium (LD) with each other. For this reason, SMR is often coupled with a test for heterogeneity in dependent instruments (HEIDI) [53], which can rule out a LD-mediated relationship.

A SMR + HEIDI analysis using the 435 sMRI traits and 13 GTEx brain tissues [4] found several genes potentially mediating the shared genetic components of BMI and sMRI traits through their expression (Fig. 2.2A, respectively blue and green dots). A total of 21 genes (Fig. 2.2A, red dots) were associated with both BMI and one or more sMRI traits.

We confirmed the gene/trait associations found by SMR through an individual-level study of the association between GReX and traits, as in transcriptome-wide association studies (TWAS) [125]: the genetic component of gene expression for the 21 genes was predicted in the relevant brain tissues for the UKB subjects for which the relevant phenotype was available (BMI or sMRI), and its correlation with the trait was computed. For BMI the TWAS association was nominally significant in 23 out of 34 cases, and the sign of the correlation was concordant with the one found by SMR for all of them. For sMRI traits, nominal significance and sign concordance was achieved in all of the 687 tested associations.

We asked whether the overlap between BMI- and sMRI-associated genes was greater than expected by chance. Since the significant genes for all traits are clearly clustered into genomic loci (see e.g. the loci in chromosomes 8 and 16

in Fig. 2.2A, the latter including the well-known obesity associated *FTO* gene [126], we devised a circular permutation procedure preserving the gene order (see Methods) which revealed that the overlap between BMI- and sMRI-associated genes is indeed greater than expected by chance (empirical $P < 10^{-4}$).

Furthermore, the cis- regions around the pleiotropic genes were enriched in genetic covariance between BMI and sMRI traits, although only nominally and for just 10 sMRI traits out of 97 tested (Monte-Carlo P-value from partial genetic covariance estimates). Nevertheless, as shown in Fig. 2.2B, when considering the genetic regions surrounding all genes mediating BMI, i.e. without restricting to the 21 found to also mediate sMRI traits, we could not observe any significant enrichment: their estimated log-fold changes were markedly smaller and never distinguishable from 0. This suggests that the portion of the genome surrounding the 21 identified pleiotropic genes, albeit small, indeed carries a larger-than-expected proportion of the shared genetic determinants of brain morphology and BMI.
As shown in Supplementary Fig. A.1 , there is a dense network of genetic correlations across the measurements from different brain regions for which we found pleiotropic genes with BMI. This is in line with a common genetic basis across many sMRI traits: it is therefore non-trivial to quantitatively pinpoint which morphological traits are the most affected by these genes. However, all categories show at least one pleiotropic gene (Fig. 2.2C), and even though cortical gray-white contrast measurements seem very prominent, this can be readily explained by an especially strong genetic correlation among them (Supplementary Fig. A.1). Similarly, all GTEx tissues are involved, although with a suggestive prominence of subcortical tissues.

SMR also provides an estimate of the direction and size of the effect of gene expression on each trait. We observed that in most cases the effects of gene expression on BMI and sMRI traits have opposite signs (Supplementary Fig. A.2), in agreement with the prevalence of negative genetic correlations. Taken together, these results suggest that the genetic correlations between BMI and morphological brain measures are indeed driven in part by regulatory variants acting on brain gene expression, and provide us with a list of candidate genes for the more detailed variant-level investigations described in the following.

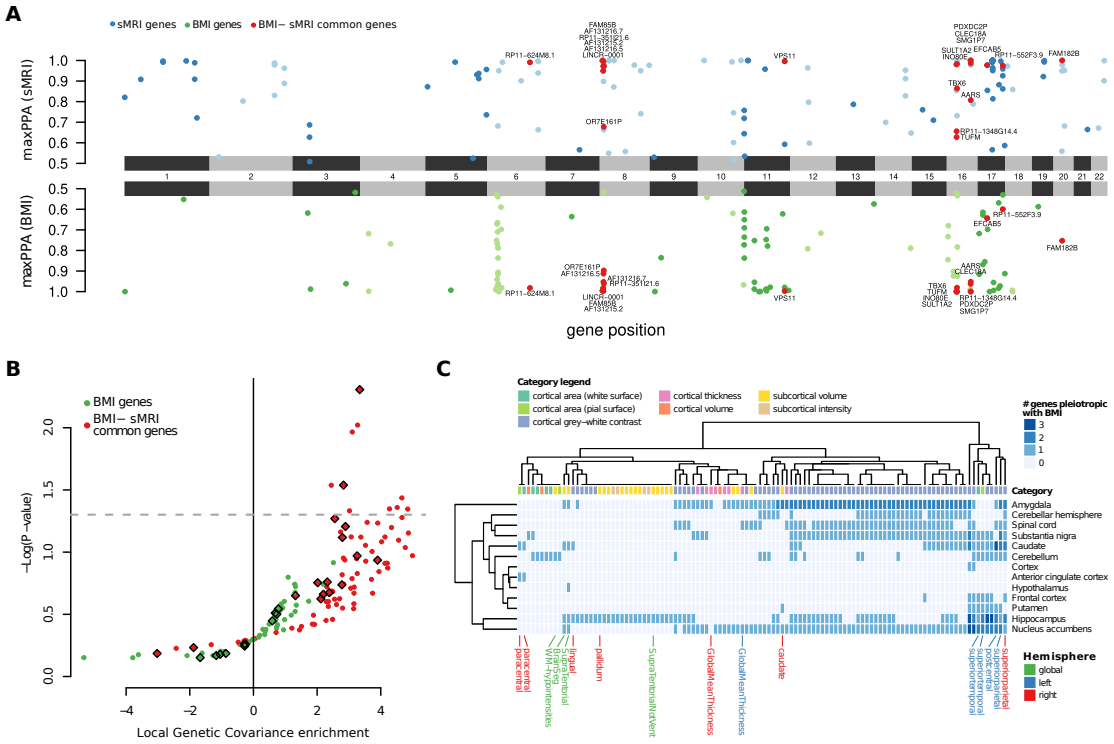


Figure 2.2: SMR reveals common molecular patterns. (A) Chicago plot showing the posterior probability of association (PPA, y axis) of gene expression on sMRI traits (top) and BMI (bottom) according to SMR. The x axis represents the genomic position. For each gene-sMRI trait, the reported PPA is the highest among all tissue/trait combinations. Red dots represent genes whose expression is associated with BMI and at least one sMRI trait in at least one tissue. (B) Volcano plot showing the local genetic covariance enrichment of each sMRI trait and its respective p-value. The enrichment is referred to the genomic regions surrounding the genes found by SMR to be in pleiotropy with BMI (green dots) or with both BMI and sMRI traits (red dots). Diamonds represent sMRI traits with significant ($P < 0.001$) genetic correlation with BMI. (C) Heatmap representation of the number of genes whose genetically determined tissue expression pleiotropically affects BMI and each sMRI trait. sMRI traits refer to the left (blue) or right (red) emisphere, or are global (green)

2.3.3 Fine-mapping and colocalization analysis

SMR reveals genes associated to both brain morphology and BMI through their expression in brain tissues, but does not provide direct information about the genetic

variants involved. Although the heterogeneity test should rule out non-colocalizing associations between gene expression variation and complex traits (but not necessarily between BMI and sMRI), we used colocalization analysis to determine in which cases the variants likely to be causal of gene expression variation, BMI, and brain morphology coincided. Colocalization was investigated for all the BMI/sMRI trait/gene expression trios found through SMR analysis. Specifically, given such a trio, we used the Sum of Single Effects (SuSiE) regression framework combined with the COLOC algorithm (coloc-SuSie) [61], which avoids the assumption of a single causal variant for each trait in each locus, to test colocalization for the three pairs of traits (gene expression - sMRI, gene expression - BMI, and BMI - sMRI). The results are shown in Fig. 2.3, as a network whose nodes are traits, and edges indicate colocalized causal variants.

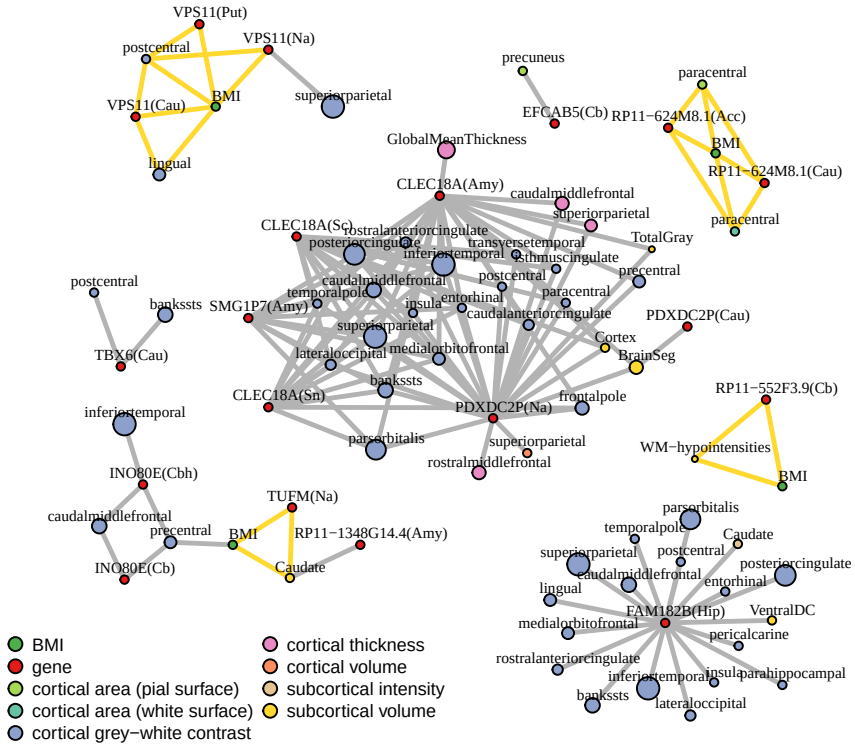


Figure 2.3: Network of colocalized causal variants. Each edge represents an instance of colocalization, that is a pair of causal variants (or rather genetic components tagged by such variants) for which the hypothesis of a common causal variant (H4) is the most likely according to COLOC; edges highlighted in yellow form a colocalization clique between eQTL, BMI and sMRI causal variants. Each node represents a trait: red nodes stand for gene expression in a GTEx tissue, green nodes for BMI, while sMRI nodes are colored according to their trait category. Traits can be duplicated if multiple independent causal variants are found for them. sMRI traits belonging to the same category with genetic correlation higher than 0.9 were pruned to leave only a representative node, sized proportionally to the trait count that it represents. For all left-right hemisphere pairs this pruning left a single representative node.

Of particular interest are cliques involving BMI, an sMRI trait, and the expression of a gene. In these cases a single variant is likely to be causal for the three traits. In particular, this happens for the protein-coding genes *VPS11* and *TUFM*, and the antisense transcripts *RP11-624M8.1* and *RP11-552F3.9*. For other genes, colocalization is limited to gene expression and sMRI traits. Note that the HEIDI test we used after SMR is also a test of colocalization between gene expression on one hand and BMI or sMRI traits on the other, which was passed by all the trait pairs shown in the figure. The fact that not all of them pass the coloc-SuSiE test indicates that the results of HEIDI and coloc-SuSiE agree only partially, a fact previously noted in [127]. Thus we conservatively consider the cliques in the network as having strong evidence of three-way colocalization, and the other phenotypic trios as having weak evidence of colocalization.

The Manhattan plots (Fig. 2.4) show the SuSie 95% credible sets for the three traits for the TUFM and VPS11 loci (the loci associated to the two antisense transcripts are shown in Supplementary Fig. A.3). In both cases, as expected, the three-way intersection of the SuSie credible sets contains variants in high LD with each other ($R^2 > 0.5$). Notably, such intersections overlap the TUFM and VPS11 gene loci. Coloc analysis of each pair of traits showed strong evidence of colocalization (posterior probability of colocalization (PPH4) > 0.8).

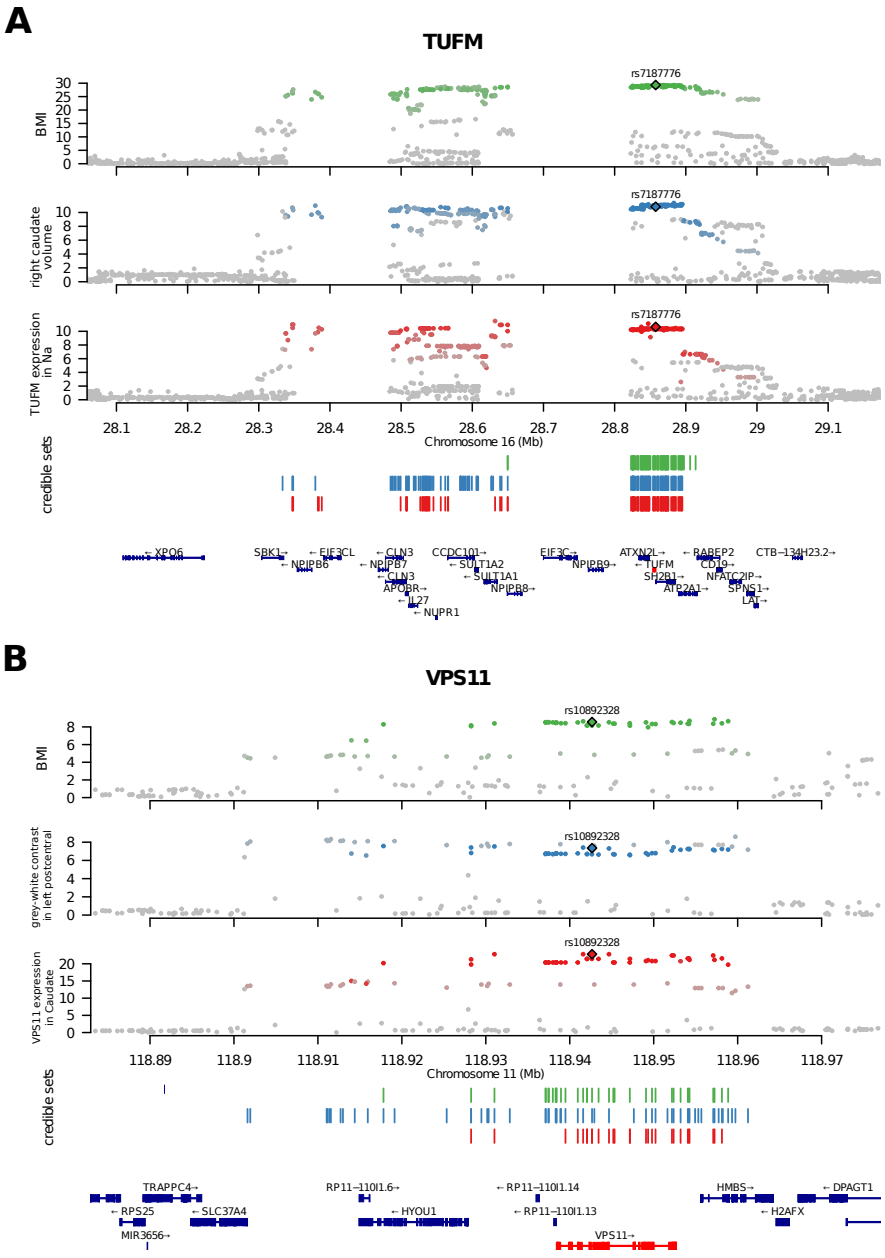


Figure 2.4: Credible sets and evidence of colocalization. Manhattan plots showing colocalization in the TUFM (A) and VPS11 (B) loci among the trait trios. Variants are plotted with their respective GWAS $-log_{10}(P)$ (see y axis). BMI (green), sMRI (blue), and gene expression (red) credible sets are shown by vertical bars. For each gene, the three-way intersection of the SuSie credible sets is tagged by a genome-wide significant variant represented by a diamond. Other variants in each credible set are represented by dots whose color represents the level of LD with the tagging variant (gray: $R^2 < 0.5$; colored: $R^2 > 0.5$, with color intensity proportional to R^2). Na= Nucleus accumbens.

2.3.4 Epigenetic fine-mapping

While colocalization analysis strongly suggests that the same variant is causal of gene expression, BMI, and brain morphology for the four loci described above, it cannot identify the individual causal variant: Indeed (Fig. 2.4, Supplementary Fig. A.3), in each of these loci, the intersection of the three credible sets still contains a sizable number of variants.

In order to further restrict the set of candidate causal variants for each locus, we investigated the effect of each of the variants included in the intersection of the credible sets of the three traits (gene expression, BMI, and sMRI) on chromatin accessibility in brain cells. This analysis was limited to the four loci with strong evidence of three-way colocalization, for which the intersections of the credible sets included a total of 128 variants, henceforth referred to as "candidate pleiotropic variants".

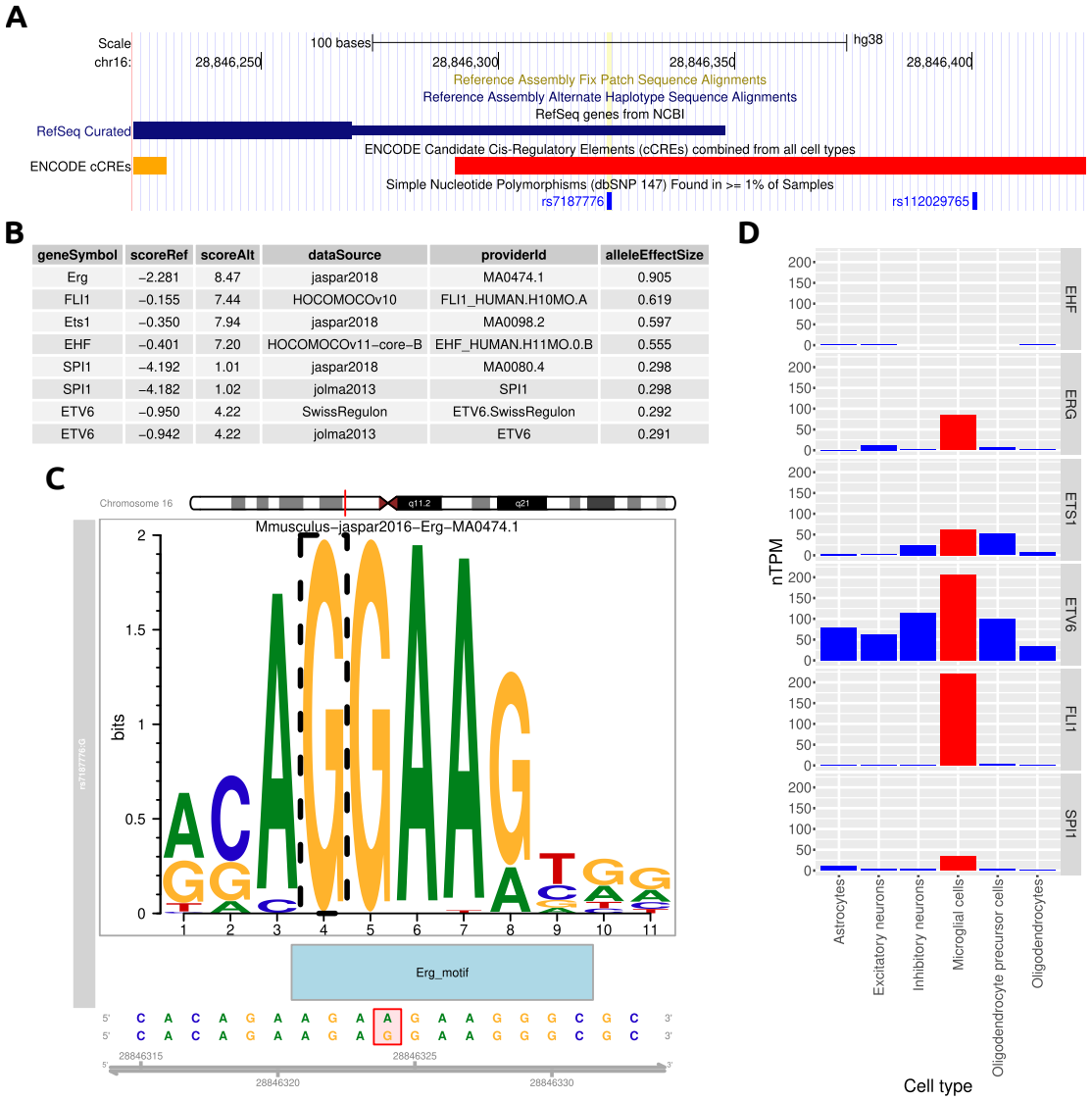
To this purpose, we first trained a gapped k-mer support vector machine (gkm-SVM [90]) on single-nuclei ATAC-seq data of 107 brain cell types produced in [128]. For 105 out of 107 cell types we had sufficient data to train a SVM, which we then used to predict the cell type-specific effect of each candidate pleiotropic variant on chromatin accessibility. Nine of the 128 candidate pleiotropic variants were predicted ($|\text{deltaSVM}| > 2$) to affect chromatin state in at least one brain cell type.

The strongest epigenetic fine mapping prediction concerned variant rs7187776, located in the 5' UTR of the *TUFM* gene (Fig. 2.5A), and previously associated by GWAS to various complex traits including hip circumference adjusted for BMI [129]. The alternate allele was predicted by gkm-SVM to be associated with increased chromatin accessibility specifically in microglial cells (deltaSVM = 3.17). Indeed the same variant was found to be positively associated ($P = 1.1 \cdot 10^{-4}$) with chromatin accessibility in glia, but not neurons, in a recent study of the genetic determinants of chromatin accessibility in the human brain [130].

We thus decided to investigate in greater depth the possible mechanisms by which the alternate allele of rs7187776 could promote the opening of the chromatin in microglia. We used motifbreakR [131] to identify transcription factors (TFs) whose motif was altered by rs7187776. The analysis revealed that the alternate

allele introduces a strong binding site for TFs of the ETS family, by creating the core ETS motif "GGAA" in lieu of "AGAA" found in the reference sequence. Specifically, motifbreakR identified several TFs whose affinity is predicted to be increased by the alternate allele, all belonging to the ETS family (Fig. 2.5B,C).

Thus we hypothesized that rs7187776 opens the chromatin near the *TUFM* transcription start site (TSS) by creating a binding site for an ETS TF. To gain insight into the precise identity of such TF we used single-cell RNA-seq data curated by the Human Protein Atlas [132] to identify which of these TFs are expressed in microglia (Fig. 2.5D). ETV6 and FLI1 showed the most robust expression in microglial cells, with a remarkable level of cell type-specificity especially for the latter.



geneSymbol	scoreRef	scoreAlt	dataSource	providerId	alleleEffectSize
Erg	−2.281	8.47	jaspar2018	MA0474.1	0.905
FLI1	−0.155	7.44	HOCOMOCOv10	FLI1_HUMAN.H10MO.A	0.619
Ets1	−0.350	7.94	jaspar2018	MA0098.2	0.597
EHF	−0.401	7.20	HOCOMOCOv11−core−B	EHF_HUMAN.H11MO.0.B	0.555
SPI1	−4.192	1.01	jaspar2018	MA0080.4	0.298
SPI1	−4.182	1.02	jolma2013	SPI1	0.298
ETV6	−0.950	4.22	SwissRegulon	ETV6.SwissRegulon	0.292
ETV6	−0.942	4.22	jolma2013	ETV6	0.291

Figure 2.5: Epigenetic fine mapping and motif analysis for the TUFM locus. (A) Variant rs7187776 in the 5' UTR of TUFM is predicted by gkm-SVM to affect chromatin state specifically in microglial cells, with the alternate allele G associated with higher accessibility. (B) Motif analysis identifies eight motifs, corresponding to six transcription factors belonging to the ETS family, whose binding is strongly favored by the alternate allele. (C) The alternate allele creates the core ETS motif GGAA, here shown in the context of the ERG binding site. (D) The expression levels of the six transcription factors in brain cell types from the Human Protein Atlas: several of them are expressed in microglial cells, FLI1 and ETV6 showing especially robust and cell type-specific expression.

2.4 Discussion

We have devised a strategy to dissect the molecular underpinnings of the genetic correlation between a macroscopic complex trait (BMI) and a set of endophenotypes (brain morphology parameters assessed by sMRI). Since the effects of genetic variants on complex traits is thought to be mediated mostly by gene regulation, we first used SMR to generate a catalog of 21 genes whose genetically regulated expression pleiotropically affects both the endophenotypes and BMI. The enrichment of both the number of genes found and of the contribution of the respective genomic loci to the genetic correlation suggests that this approach can indeed explain at the molecular level a sizable portion of the observed correlations. In four of the identified loci, colocalization analysis confirmed the three-way pleiotropic effect of the fine-mapped genetic variants on gene expression, BMI, and brain morphology. For these loci we used epigenetic fine mapping to identify putative causal variants and the related mechanisms, highlighting in particular variant rs7187776, located in the 5' UTR of the *TUFM* gene and predicted to alter chromatin accessibility in microglia, possibly by creating a strong ETS binding site.

SMR identified genetically regulated *TUFM* expression to be positively associated with BMI and negatively associated with the volume of caudate in the right hemisphere. *TUFM* encodes a ubiquitously expressed nuclear-encoded mitochondrial protein, involved in mitochondrial translation, that has been associated to several biological processes, including autophagy, and human phenotypes, including obesity (see [133] for a recent review). A study on subcortical volumes across lifespan has shown that *TUFM* expression in several brain regions in older and young adults is associated with caudate nucleus volume. In particular, increased *TUFM* expression was associated with smaller caudate nucleus volume, with evidence for colocalization in several tissues [134]. TUFM was also found associated with caudate volume and other subcortical morphological features in a recent GWAS meta-analysis [135].

Epigenetic fine mapping and motif analysis suggest a mechanism by which variant rs7187776 in the 5' UTR of *TUFM* creates a strong ETS binding site resulting in microglia-specific chromatin opening. Of note, chronic microglia activation has been associated with decreased hippocampus and parahippocampus volume in

neurodegenerative diseases [136]. According to dbSNP, the variant is common in all populations ascertained, with minor allele frequency (MAF) between 0.11 and 0.48. Since the magnitude and direction of causal eQTL effects are known to be consistent across populations [137], it is reasonable to expect that the findings related to this SNP should be transferable to other populations, although we did not formally check whether this is the case.

Other three-way colocalization signals involved BMI, *VPS11* expression in various subcortical structures, and gray-white contrast phenotypes from sMRI. *VPS11*, a subunit of the CORVET complex involved in the endosome/lysosome pathway, has an essential role in brain white matter development and neuron survival in zebrafish, while its reduced expression leads to an impaired autophagic activity in human cells [138]. To the best of our knowledge it has not been investigated in the context of BMI and obesity, although GWASs found *VPS11*-associated variants associated to BMI [139] and HDL cholesterol measurements [140].

The remaining high-confidence colocalization signals involved two non-coding RNAs, namely *RP11-552F3.9* and *RP11-624M8.1* (also known as *HEY2-AS1*), antisense transcripts of *TRIM47* and *HEY2*, respectively. Variants mapped to these genes have been associated by GWASs to both brain morphology traits [120, 141] and BMI [142, 143]. Indeed *TRIM47* was among the genes found in [117] to harbor variants associated with both BMI and regional brain volumes.

2.4.1 Limitations

Some limitations of this work should be noted. It is possible that some of the pleiotropy signals were lost by summarizing the genetic basis of our complex traits of interest into effect-sizes and applying summary-based analyses. While the high number of sMRI traits analyzed initially warranted this approach, individual-based techniques capable of dealing with large sample sizes and multiple traits simultaneously are now being developed, opening the possibility of an individual-based analysis. Notably, the fact that the estimated GReX was predictive of BMI and sMRI traits for the vast majority of the genes identified by SMR, is instead reassuring against the risk of false positives.

Another limitation is that even if we set out to evaluate genetic correlation and

pleiotropy for fine-grained brain morphology measurements, the strong genetic correlation across all the sMRI traits analyzed and the lack of an explicitly differential strategy prevented us from pinpointing specific brain regions consistently involved throughout all steps of our analysis.

Epigenetic fine-mapping was performed with the SVM-based methods of [90], while methods to predict open chromatin from sequence based on convolutional neural networks, such as Basset [144] or Sei [145], might in principle achieve higher predictive power. However, the method we used has the advantage of a much shorter training time, a key issue as it had to be applied to each of the more than 100 cell types identified in [128].

Finally, the use of BMI as an obesity-related quantitative trait should be complemented by other obesity-related measures, such as waist-to-hip ratio, body fat percentage, fat-free mass, etc. The analysis and comparison of the genetic correlations of these traits with sMRI measurements is likely to allow a more nuanced view of the relationship between obesity and brain morphology.

2.5 Conclusion

We have shown how the integration of data from GWASs, population-based transcriptomic studies, and single-cell chromatin accessibility assays allows the molecular dissection of the genetic correlation between complex traits and the formulation of data-driven hypotheses on the relevant regulatory mechanisms. Specifically, we have confirmed and expanded previous observations on the genetic correlation between brain morphological features and BMI, and generated an atlas of 21 genes whose genetically regulated expression mediates a significant part of such correlation.

2.6 Methods

2.6.1 Summary statistics and individual data

We used the GWAS summary statistics of brain imaging-derived phenotypes (IDPs) processed with a sample size of about 33 thousands individuals (sample sizes

vary across traits) from the UKB release 2020 [120]. In particular, we selected all the IDPs defined according to the Desikan-Killiany cortical atlas [121] and subcortical volumetric segmentation (ASEG) [122]. After excluding the traits "aseg_lh_number_HolesBeforeFixing" and "aseg_rh_number_HolesBeforeFixing," which correspond to technical artifacts produced by segmentation rather than morphological measurements, we were left with 435 IDPs. GWAS summary statistics for BMI were derived from the UK Biobank GWAS version 3, and made available by the Neale Lab (http://www.nealelab.is/uk-biobank). All summary statistics for both IDPs and BMI were filtered for imputation INFO score ≥ 0.95 and minor allele frequency (MAF) ≥ 0.01.

For methods that required individual-level information (LD matrices, stratified LD scores, GReX), we used 38,406 samples of European ancestry from the UKB. Subjects were selected among the donors with brain MRI-derived phenotypes as of October 2023 (ASEG whole brain volume, ID 26514, used as representative trait for filtering) identified as British with West-European ancestry (field 22006, code 1), and not considered outliers for missingness and heterozygosity (field 22027). Finally, we pruned relatives with kinship coefficient ≥ 0.04419 (third or lower degree relatives), reaching a subset with the sample size mentioned above. A total of 9,252,961 variants with INFO score > 0.8 and minor allele frequency (MAF) > 0.01 were included in this dataset.

2.6.2 Genetic correlation

We estimated the genetic correlation among each of the 435 sMRI-derived traits and BMI using GWAS summary statistics with the cross-trait LD Score regression (LDSC) software obtained from https://github.com/bulik/ldsc. We used genome-wide LD scores computed by the Pan-UK consortium on UKB samples of European descent (https://pan.ukbb.broadinstitute.org/): scores were available for 1,094,844 SNPs which met their exporting criteria, most notably after filtering for HapMap3 SNPs, MAF > 0.01 and removing the human major histocompatibility complex (MHC) region.

2.6.3 Summary-based Mendelian randomization

We conducted summary-based Mendelian randomization with the Omics Pleiotropic Association (OPERA) software tool [146] to identify pleiotropic associations. Summary-level expression quantitative trait loci (eQTL) data of European individuals (n = 698) from the GTEx [4] dataset, including 13 brain tissues, were used for the analysis as the exposure. We estimated the cis-eQTLs with the R package Matrix eQTL [147], testing all variants with MAF ≥ 0.01 within 2 megabases (Mbs) of each gene's TSS. All the covariates included by GTEx for each tissue, including sex, genomic PCs, and PEER factors, were used in the regression models.

BMI and sMRI-derived traits were considered one by one as outcomes. We used 503 European samples from the 1000 Genomes Project [148] as the LD reference data required for the HEIDI test. We set the significance threshold of posterior probability of association (PPA) to 0.5, then performed the HEIDI test for the genes passing this threshold and discarded those for which the null hypothesis of non heterogeneity could be rejected with $P < 0.05$.

The circular permutation procedure to test for BMI/sMRI gene overlap was devised as follows: for 10,000 times, a list of all genes tested, ordered for TSS position and tagged for association status with BMI, was split into K genomic chunks and a random spin was applied to each chunk; chunks were then reassembled in a random order. This resulted in 10,000 random sets with the same positional clustering of the genes identified for BMI, which were used to compute an empirical P-value for the size of the intersection with genes associated with at least one sMRI trait. The procedure was repeated with K= 1,5,10, obtaining the same result.

2.6.4 Individual-level GREx-trait association

An in-house implementation of prediXcan [125] was used to train an elastic net regression for the 21 putatively pleiotropic genes on the 13 GTEx brain tissues mentioned above. We used for training 80% of the European samples used for eQTL estimation, using 1 Mb as cis-distance from each gene's TSS and including all encompassed variants in the training. Gene expression was first transformed into residuals using the approach described in [127], then regressed on the genotype

matrices. We filtered out models achieving Pearson's correlation $r < 0.1$ between predicted and measured expression, or with zero variance, in the 20% holdout testing set. Filtering on the signed correlation coefficient, rather than on its square, allows excluding negative correlations between predicted and actual expression, which clearly do not signal good predictive power [149]. The resulting models were applied on individual UKB genetic data described above, for which the GReX was imputed and used as predictor for the BMI and sMRI phenotypes identified by SMR+HEIDI analysis, together with sex and age as covariates.

2.6.5 Partitioned genetic covariance

We aimed to test enrichment in genetic covariance in the genomic loci around two sets of genes: a) those found associated with both BMI and at least one sMRI trait by SMR analysis, and b) all genes associated with BMI by SMR. These regions were identified by considering a 4 Mb window surrounding the TSS of each gene in each set.

A cross-trait stratified LD Score Regression (sLDSC) [45] was conducted to estimate the partitioned genetic covariance between BMI and sMRI phenotypes at each of the two aforementioned annotations, using GWAS summary statistics as described previously. We limited this test to the 97 sMRI phenotypes previously found to be genetically correlated with BMI at nominal significance.

We used the individual-level UKB genetic data described above to compute partitioned LD scores, setting a 2 Mb LD window and keeping SNPs in HapMap3 and with MAF > 0.01 for the regression. Enrichments and the corresponding P-values were computed starting from the genetic covariance estimates and relative standard errors reported in the LDSC logs, by generating 50,000 Monte Carlo randomizations and testing how often the genetic covariance fraction was less or equal than the SNP fraction, thus obtaining one-sided P-values. The enrichment estimates were extracted from the medians of such distributions and then log2-transformed.

2.6.6 Fine-mapping and colocalization

In order to assess three-way colocalization assuming multiple causal variants per locus, we applied coloc to the decomposed signals found by SuSiE (coloc-SuSiE [61]) for each locus whose expression in at least one brain tissue was found by SMR+HEIDI analysis to be in association with BMI and a sMRI phenotype. The analysis was performed on each pair of phenotypes resulting from the combination of BMI, sMRI traits and gene expression, then the colocalization results were compared across the three variables. Only SNPs with effect sizes available for all three phenotypes were considered. The *coloc_susie* function from the coloc 5.2.3 R package was run with default parameters. We computed the LD matrix for effects on gene expression with the GTEx European samples [4] from each tissue. The LD matrix for the effects on the complex trait from UKB was computed on the individual UKB genetic data described above. We considered that colocalization occurred when the colocalization posterior probability (PPH4) was greater than the probability of association with different causal variants (PPH3). Manhattan plots were produced with locuszoom [150].

2.6.7 Epigenetic fine mapping and motif analysis

The open chromatin regions of the brain cell types identified in [128] were used to train gkm-SVM [90] models which were then used to predict the effect of each variant on the chromatin accessibility of each cell type. The analysis was performed on the 105 cell types (out of a total of 107 identified in [128]) for which at least 5,000 open chromatin regions were available. Each model was trained on 5,000 open chromatin regions drawn at random from the complete set. This analysis was applied to the 128 variants included in the three-way intersection of credible sets (BMI, sMRI, and gene expression) for the four loci for which we had strong evidence of pleiotropy (supported by both HEIDI and COLOC).
Candidate transcription factor binding sites altered by the variants of interest were identified with motifbreakR [131], using the whole MotifDb collection of positional frequency matrices and log probability scores. We considered as significantly motif-altering the variants whose effect was evaluated as "strong" by motifbreakR and such that the logarithmic score was positive for one allele and negative for the

other one.

CHAPTER 3

Removing genetic effects on plasma proteins enhances their utility as predictive disease biomarkers

This work, conducted at the Institute for Molecular Medicine Finland (FIMM) and currently under review at *Nature Genetics*, highlights the need to focus also on non-genetic factors when working with plasma proteomics data. Plasma protein levels, in fact, are shaped by both heritable and environmental factors. In the context of disease relationships, it is necessary to capture these effects when they are correlated but not causally linked to the disease. We showed that this approach can enhance statistical power in detecting associations and it can be a strategy for biomarker discovery and clinical trial design, consistent with the largely non-causal role of most plasma proteins in disease risk.

Abstract

Plasma protein levels are influenced by both genetic and non-genetic factors and can serve as early disease biomarkers. When a protein is correlated with, but not causally linked to, disease, its genetic determinants can add unwanted variability to protein–disease associations. In such cases, removing the genetic component may improve their predictive performances. Here, we tested this hypothesis by genetically adjusting 94 highly heritable proteins spanning diverse biological pathways and evaluating their associations with the onset of 37 diseases in 39,871 UK Biobank participants. Genetically adjusted proteins showed stronger associations in 88% of 1,312 significant protein–disease pairs, equivalent to a 30% median reduction in required sample size for comparable power. Of 96 protein–disease pairs with significant differences, all but one showed larger effects for adjusted proteins. Most proteins also showed consistently stronger associations with environmental and lifestyle factors once genetic effects were removed. Finally, we constructed multi-protein scores from genetically adjusted proteins and demonstrated that they significantly improve prediction for 7 diseases compared to unadjusted proteins. These findings demonstrate that removing genetic effects from plasma proteins is an effective strategy to increase power for biomarker discovery and clinical trial design, consistent with the largely non-causal role of most plasma proteins in disease risk.

3.1 Introduction

Antibody-based, broad-capture proteomics enables the measurement of thousands of plasma proteins across large cohorts[85]. Applied to biobank studies, these scalable approaches have led to the discovery of hundreds of disease biomarkers[77] and shown that multi-protein scores can strongly predict disease onset across a wide range of disorders, from neurological to cardiometabolic[151, 152].

Plasma protein levels are partially shaped by inherited genetic variation, with the average SNP-based heritability of proteins measured on the Olink Explore platform estimated at ∼0.16[85]. At the same time, proteins capture dynamic,

non-genetic risk factors such as smoking, diet, and lifestyle[8, 153, 154], which enhance their value as disease predictors.

Traditionally, genetic effects on protein expression (protein quantitative trait loci, pQTLs) have been leveraged to test causal links between proteins and diseases using Mendelian randomization[155, 5]. Here, we take a related but orthogonal approach: removing genetic effects from protein expression to aid biomarker discovery by removing noise due to random genetic variation. This strategy, sometimes referred to as "de-Mendelization"[156], has not yet been applied at scale, but a notable example is prostate-specific antigen (PSA), where adjusting for genetic determinants of constitutive, non-cancer-related PSA variation improves screening utility[157].

In this work, we first describe the conditions under which removing genetic effects from proteins can improve the discovery of disease associations. We then demonstrate that, in practice, this strategy consistently increases power to detect protein–disease associations.

3.2 Results

3.2.1 Conceptual framework

If protein expression causally influences the risk of developing a certain disease, then genetic factors affecting protein expression are also, through the protein, likely to be associated with the disease. This is the fundamental principle underlying Mendelian randomization[158]. In this scenario, removing genetic signals that influence protein expression is expected to weaken the observed association between proteins and disease. This observation extends to the removal of any genetic factor that is causal to the disease (Figure 3.1A, upper DAGs).

However, a protein may not causally influence disease risk. This can occur when protein levels are altered by disease pathophysiology or treatment, or, more commonly, when external non-genetic factors influence both protein levels and

disease risk (Figure 3.1A, lower DAGs). Such factors act as confounders of the protein–disease relationship. In this case, genetic influences on proteins represent unwanted variation that adds noise to the associations between proteins and disease. In other words, genetics can make proteins less predictive of disease risk by introducing variability unrelated to the disease of interest.

In the latter scenario, removing the genetic effect on a protein can increase the power to detect the association with a disease. In the Supplementary Methods, we provide the theoretical derivation showing that the gain in power is a function of both the proportion of protein variance explained by genetics (R^2) and the strength of the protein-disease association.

3.2.2 Protein selection and genetic adjustment

We compared protein-disease associations using both directly measured and genetically adjusted protein levels. We initially considered 1,108 proteins measured with Olink or SomaScan in the INTERVAL study[159], for which OmicsPred provides polygenic score (PGS) weights[160]. These PGSs were calculated in 39,880 participants of European ancestry from the UK Biobank (UKB) (Figure 3.1B). From this set, we retained 94 proteins that were well predicted by their PGS ($R^2 > 0.2$) and pruned for correlation ($|r| < 0.5$) to capture independent biological effects and increase generalizability. Of the 94 proteins, 37 PGSs were derived from genome-wide association studies (GWAS) of Olink-measured proteins and 57 from GWAS of SomaScan-measured proteins. We then regressed the PGS out of each protein, using the residuals as genetically adjusted protein levels (see Methods). As a replication analysis, we further considered seven proteins not highly correlated ($|r| < 0.5$) with the initial 94 and with PGS weights derived from FinnGen[2]

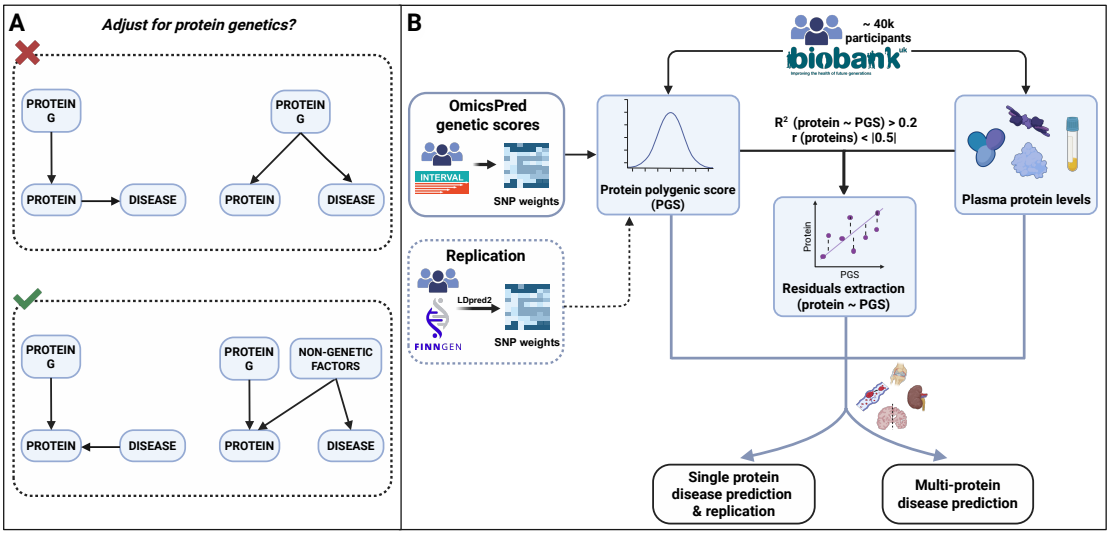


Figure 3.1: **Graphical abstract A** Directed acyclic graphs (DAGs) illustrating how adjusting proteins for their genetically predicted component (G) influences disease prediction. In the top panels (red cross), adjusting for G removes the genetic signal that is causal for disease, thereby reducing the protein's predictive performance. In contrast (green check), when G is not causal, adjustment may improve prediction. **B** Protein polygenic scores (PGSs) were derived using OmicsPred genetic scores and calculated in UK Biobank participants of European ancestry. Proteins with high genetic predictability and low correlation were retained. Genetically adjusted proteins were obtained by regressing out the PGS component to extract residuals, which were then used as inputs for both single- and multi-protein disease prediction models. For replication, SNP weights were computed with LDpred2 using FinnGen data. Created with BioRender.com.

3.2.3 Genetically adjusted proteins show stronger associations with disease risk

Among the 94 selected proteins, PGS explained a substantial proportion of variability, with R^2 values ranging from 20% to 81% (Supplementary Figure B.1). These proteins span across all the major pathways represented in the overall Olink Explore 3072 panel, including inflammation, neurology, cardiometabolism, and oncogenesis.

We tested associations between the 94 proteins (both unadjusted and genet-

ically adjusted) and the onset of 37 high-burden diseases based on phenotype definitions previously described by Jermy et al.[161], adjusting for age and sex. This yielded 3,478 protein–disease pairs, of which 1,312 were statistically significant (false discovery rate, FDR < 0.05) in at least one model and were considered in subsequent analyses. All diseases except appendicitis had at least one significant protein association, underscoring the value of plasma proteins as biomarkers across organ systems.

While effect sizes were generally consistent across the two models (Figure 3.2A), genetically adjusted proteins showed an overall trend toward stronger disease associations: in 88% of the 1,312 protein–disease pairs, adjusted proteins had larger absolute z-scores, with an average 57.6% increase in absolute z-scores compared to unadjusted proteins. One way to interpret the improvements in z-scores is to consider the potential gain in statistical power for biomarker discovery, i.e., identifying new protein-disease associations. Among the 1,312 significant protein-disease pairs, the same level of statistical power achieved with unadjusted proteins could have been obtained with a median sample size 30% smaller when using genetically adjusted proteins.

We identified 96 protein-disease pairs that showed nominally significantly different effect sizes ($p < 0.05$) between the genetically adjusted and unadjusted model. Of these, 95 pairs (spanning 43 unique proteins) had stronger effects in the genetically adjusted model, while only one pair showed a stronger effect in the unadjusted model. Type 2 diabetes, chronic kidney disease, and heart failure showed the greatest enrichment (see Methods) of associations among the 95 pairs (Supplementary Figure B.2).

3.2.4 Stronger disease associations are explained by removal of genetic noise from proteins

To assess whether the observed improvements in the genetically adjusted models can be explained by PGS being strong predictors of protein levels but not directly linked to disease risk, we evaluated associations between protein PGS and disease

outcomes with logistic regression (Figure 3.2B). Among all the 3,478 tested pairs, only four reached significance (FDR < 0.05).

As expected, the only protein–disease pair showing a larger effect in the unadjusted model (ALPI and venous thromboembolism) also showed a significant association between ALPI-PGS and venous thromboembolism ($P = 3.04 \times 10^{-8}$). By contrast, among the 95 protein–disease pairs where genetically adjusted proteins showed stronger effects, most PGS showed weak or no association with disease. Only two pairs reached significance (FDR < 0.05): MAN2B2-PGS with Alzheimer's disease and SELE-PGS with venous thromboembolism. In these two cases, genetic adjustment does not simply remove genetic noise but might induce a spurious association with the disease; we discuss these two cases in more detail in the following sections.

Overall, these results support our hypothesis that the stronger disease associations observed for genetically adjusted proteins are primarily explained by the removal of genetic effects that influence protein variability but do not affect disease risk, thereby reducing noise in protein-disease associations.

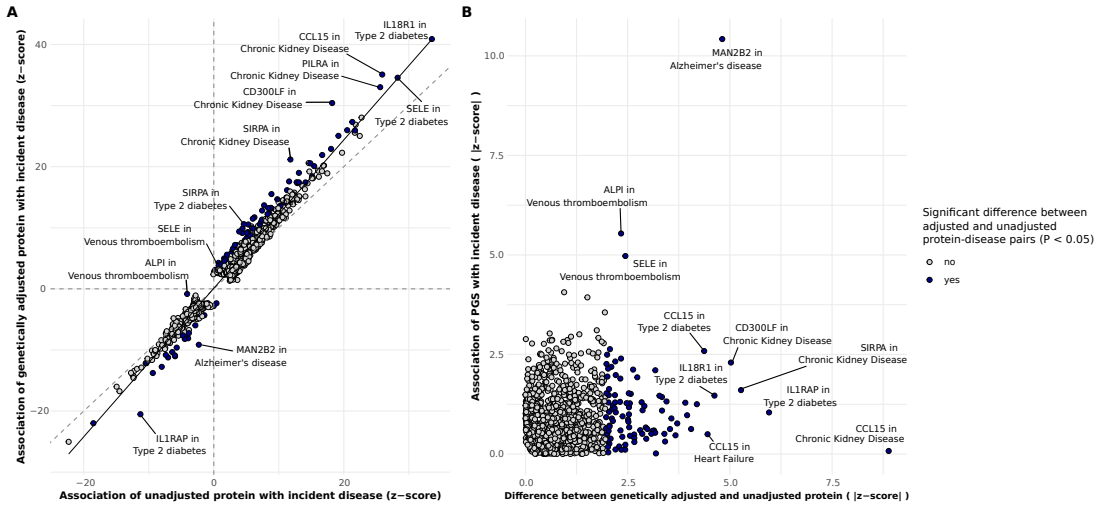


Figure 3.2: **Significant associations between 94 proteins and 37 diseases in UK Biobank. A** Logistic regression z-scores for associations with incident diseases for unadjusted proteins (x-axis) and genetically adjusted proteins (y-axis). **B** Absolute differences in z-score from genetically adjusted and unadjusted protein models (x-axis) and absolute z-score for the association between protein PGS and diseases (y-axis). Blue colour indicates a significant difference in effect sizes between adjusted and unadjusted protein associations ($p < 0.05$). We report only significant (FDR < 0.05) protein-disease associations in at least one model (unadjusted or adjusted protein); a plot including all pairs is provided as Supplementary Figure B.3.

3.2.5 Both cis and trans genetic effects on proteins contribute to stronger disease associations, but cis effects are more robust

The PGS weights calculated by OmicsPred incorporate both cis and trans genetic signals. Among the 94 protein PGS, 55 were constructed using both cis and trans variants, 36 included only cis variants, and 3 contained only trans variants.

For the 55 proteins with both cis and trans scores available, we computed separate cis- and trans-based PGS. Overall, cis-PGS explained more variation in protein expression than trans-PGS. However, for a subset of proteins (e.g., LRPAP1, TNFRSF10C, CD209, FAM3D), trans-PGS showed higher R^2 (Supplemen-

tary Figure B.4), suggesting that distal genetic regulation also plays a role in the expression of specific proteins.

We then repeated the disease association analyses, adjusting proteins separately for cis- or trans-PGS. In the cis-adjusted analysis, among 3,367 protein–disease pairs tested, 1,250 were significantly associated (FDR < 0.05) in either the cis-adjusted or unadjusted models. Seventy-five pairs showed significantly different effect sizes ($p < 0.05$), all with stronger effects for cis-adjusted proteins (Supplementary Figure B.5).

A similar but less pronounced pattern was observed for trans-adjusted proteins: among 2,146 pairs tested, 629 were significant at FDR < 0.05, with significantly different effect sizes in 9 pairs—8 stronger for trans-adjusted proteins and only one stronger for unadjusted proteins.

For both cis- and trans-analyses, the improvements in association strength (relative to unadjusted proteins) scaled with PGS prediction accuracy, with the largest gains observed for proteins with higher PGS–protein R^2 values (Supplementary Figure B.6).

An important difference between cis and trans analyses was the magnitude of direct associations between protein PGSs and disease. Cis-PGSs, despite being stronger predictors of protein levels, were not associated with any disease (FDR < 0.05). In contrast, trans-PGSs were more likely to associate with diseases, including MAN2B2-trans-PGS with Alzheimer's disease and SELE-trans-PGS with venous thromboembolism (Supplementary Figure B.7).

These findings indicate that both cis and trans genetic effects can be leveraged to reduce genetic noise in proteins and enhance disease prediction, proportionally to how well they genetically predict the protein. However, trans-PGSs, due to the inclusion of a larger number of variants, are more likely to capture pleiotropic variants that are themselves causal for disease. Removing these variants may, in turn, induce spurious correlations between proteins and diseases.

3.2.6 Illustrative examples of genetically adjusted proteins

We present two examples: one illustrating how genetic adjustment of proteins can improve protein-disease associations, and another showing instead how it may induce spurious associations.

The first example reflects most scenarios where protein-disease associations improved after genetic adjustment. Higher CCL15 protein levels were associated with increased risk of chronic kidney disease (CKD) (OR = 1.68, $p = 4.71 \times 10^{-148}$; Figure 3.3A), consistent with previous reports showing elevated CCL15 in plasma of patients with chronic renal failure[162]. The PGS for CCL15 was strongly predictive of protein levels ($R^2 = 0.47$) but showed no association with CKD (OR = 1.00, $p = 0.93$). Consequently, genetically adjusted CCL15 levels displayed an even stronger positive association with CKD (OR = 2.20, $p = 1.59 \times 10^{-269}$). Another way to interpret this result is that the same magnitude of association with CKD observed for unadjusted CCL15 could have been detected with 45% fewer individuals when using genetically adjusted CCL15.

A similar pattern was observed for other CKD-associated proteins, including CD300LF, PILRA, SIRPA, and TDGF1 (Supplementary Figure B.8), consistent with genetic contributors to these proteins adding noise to the protein–disease association.

A second example highlights a contrasting scenario. Higher MAN2B2 protein levels were weakly associated with decreased risk of Alzheimer's disease (AD) (OR = 0.91, $p = 0.02$), whereas higher MAN2B2-PGS was strongly associated with increased AD risk (OR = 1.25, $p = 1.06 \times 10^{-25}$; Figure 3.3B), showing opposite directions of effect. After genetic adjustment, MAN2B2 levels exhibited an even stronger inverse association with AD (OR = 0.86, $p = 5.56 \times 10^{-5}$).

To investigate this further, we decomposed the PGS into cis and trans components. The association signal was primarily driven by the trans-PGS (OR = 2.02, $p = 8.72 \times 10^{-110}$), whereas the cis-PGS showed no significant effect (OR = 0.93,

$p = 0.072$; Supplementary Figure B.9A). Notably, one of the trans-acting variants included in the PGS mapped to the *APOE* locus, a well-established genetic risk factor for AD[163].

These findings suggest that *APOE* influences both MAN2B2 expression and AD risk through independent mechanisms, supporting a model of horizontal pleiotropy (Figure 3.1A, upper right DAG) rather than a direct role of MAN2B2 in AD pathogenesis. Consistent with this, there are no prior reports directly implicating MAN2B2 in AD, and protein–protein interaction databases reveal no functional connections between MAN2B2 and *APOE*.

A similar scenario may underlie the association between genetically adjusted SELE and venous thromboembolism, where the SELE PGS showed a strong association with venous thromboembolism ($p = 6.59 \times 10^{-7}$), driven only by trans effects. Genetic adjustment in this case may induce a spurious association between SELE and venous thromboembolism that is not observed with the unadjusted protein (Supplementary Figure B.9B).

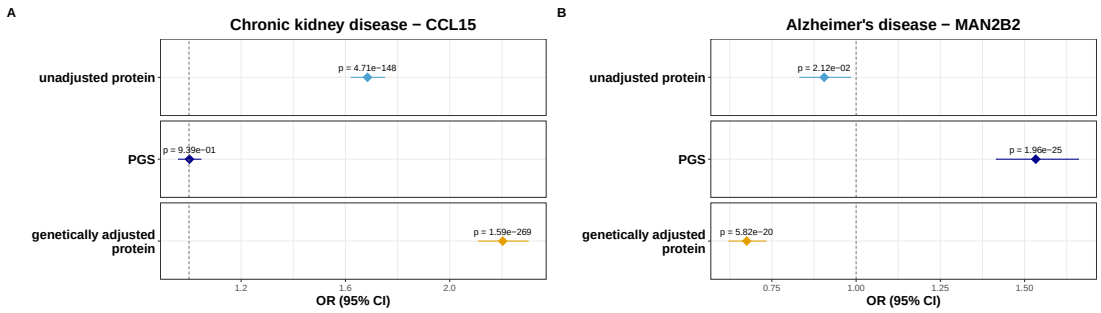


Figure 3.3: **Associations of CCL15 and MAN2B2 with disease risk.** Logistic regression Odds Ratios (ORs) with 95% confidence intervals are shown for the association of CCL15 with chronic kidney disease (**A**), and MAN2B2 with Alzheimer's disease (**B**), using unadjusted protein, PGS, and genetically adjusted protein.

3.2.7 Genetically adjusted proteins show stronger correlation with the exposome

By removing genetic effects from proteins, we expect genetically adjusted proteins to exhibit stronger associations not only with diseases but also with non-genetic factors, which we collectively refer to as the exposome. To better characterize which environmental and lifestyle influences are captured by genetically adjusted proteins, we analyzed associations with 86 variables spanning environmental, lifestyle, and sociodemographic domains, along with three laboratory measurements (body mass index (BMI), platelet count, and creatinine). Logistic regressions (or linear regressions for continuous traits) were used to compare associations between the 94 proteins and the exposome, both before and after genetic adjustment (Figure 3.4A). In total, we identified 10,894 significant associations (FDR < 0.05) out of 28,388 protein–exposome pairs tested. Adjusted proteins showed an average 48.7% increase in absolute z-scores compared to unadjusted proteins. Moreover, 921 of 10,894 protein–exposome pairs showed significantly different effect sizes ($p < 0.05$), and in 96% of these cases, genetically adjusted proteins displayed stronger associations.

Most association improvements (see Methods) were observed for BMI, creatinine, and smoking, consistent with their roles as major factors influencing plasma protein levels. The most enriched proteins included IL1RAP, SIRPA, CD300LF, ENPP5, and CCL15.

Returning to the example of CCL15 and CKD, we found that genetically adjusted CCL15 was more strongly associated with disability benefits, major dietary changes due to illness, and smoking (Figure 3.4B). For instance, the association with current tobacco smoking, a likely confounder of the CCL15-CKD relationship, increased from OR = 1.23 [1.19–1.27] to OR = 1.36 [1.32–1.41] after genetic adjustment (Figure 3.4C).

Overall, these results suggest that stronger associations between adjusted proteins and the exposome may underlie their improved associations with diseases.

In other words, genetically adjusted proteins may serve as better proxies for environmental and lifestyle factors that influence both protein levels and disease risk.

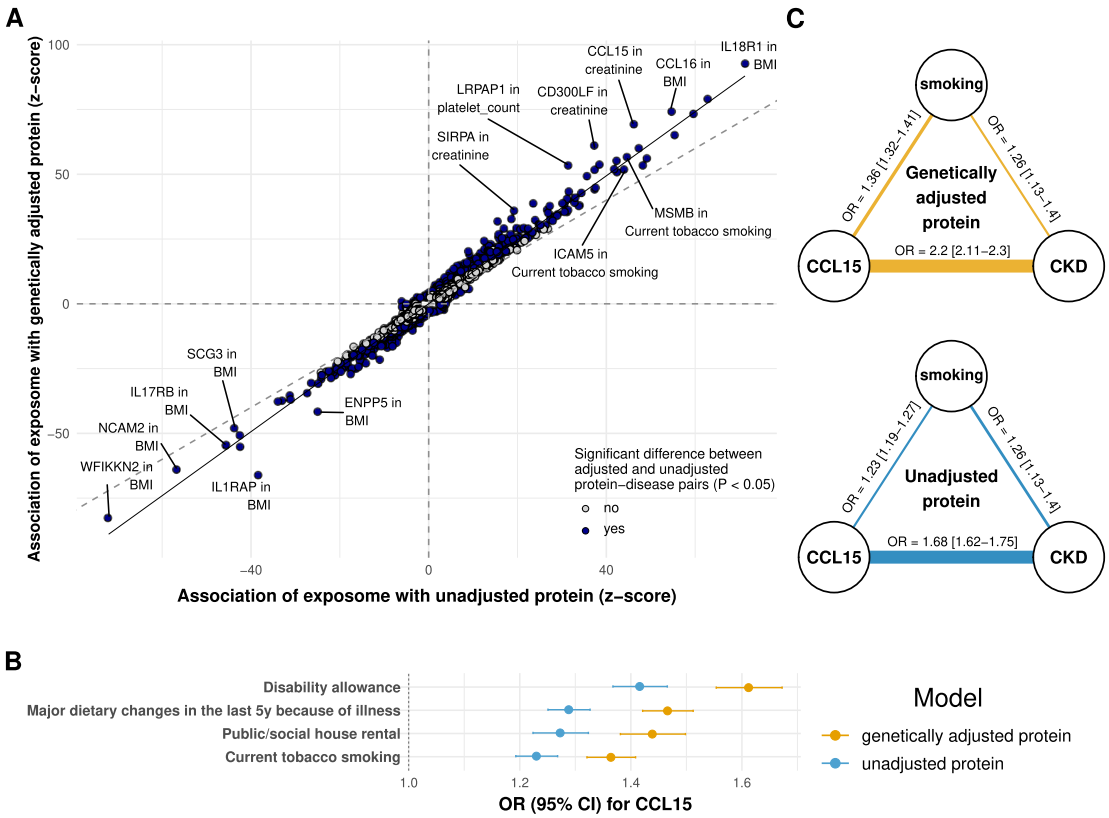


Figure 3.4: **Associations with the exposome. A** Associations with exposome features for unadjusted protein (x-axis) versus genetically adjusted protein (y-axis), expressed as z-scores from logistic regressions. Blue indicates a significant difference in effect sizes between adjusted and unadjusted protein associations ($p < 0.05$). Only associations significant at FDR < 0.05 in at least one model (unadjusted or adjusted protein) are shown. **B** Top associated exposures with CCL15, showing OR (95% CI) for the adjusted model (orange) and the unadjusted model (blue). **C** Associations between CKD and CCL15 (unadjusted protein, blue; adjusted protein, orange) and "current tobacco smoking" (binarized). 5y = 5 years.

3.2.8 Multi-protein scores using adjusted proteins are stronger predictors of disease

Previous work has shown that combining multiple proteins can yield powerful predictors of disease risk [151, 152]. We compared the standard approach–using protein levels to build multi-protein score–with our approach based on genetically adjusted proteins. Using LASSO regression, we constructed two multi-protein models (unadjusted and genetically adjusted) for the same 37 diseases analyzed previously. Overall, the genetically adjusted models achieved higher AUCs across all diseases, with 7 showing significantly improved performance (one-tailed $p < 0.05$ for ΔAUC; Figure 3.5).

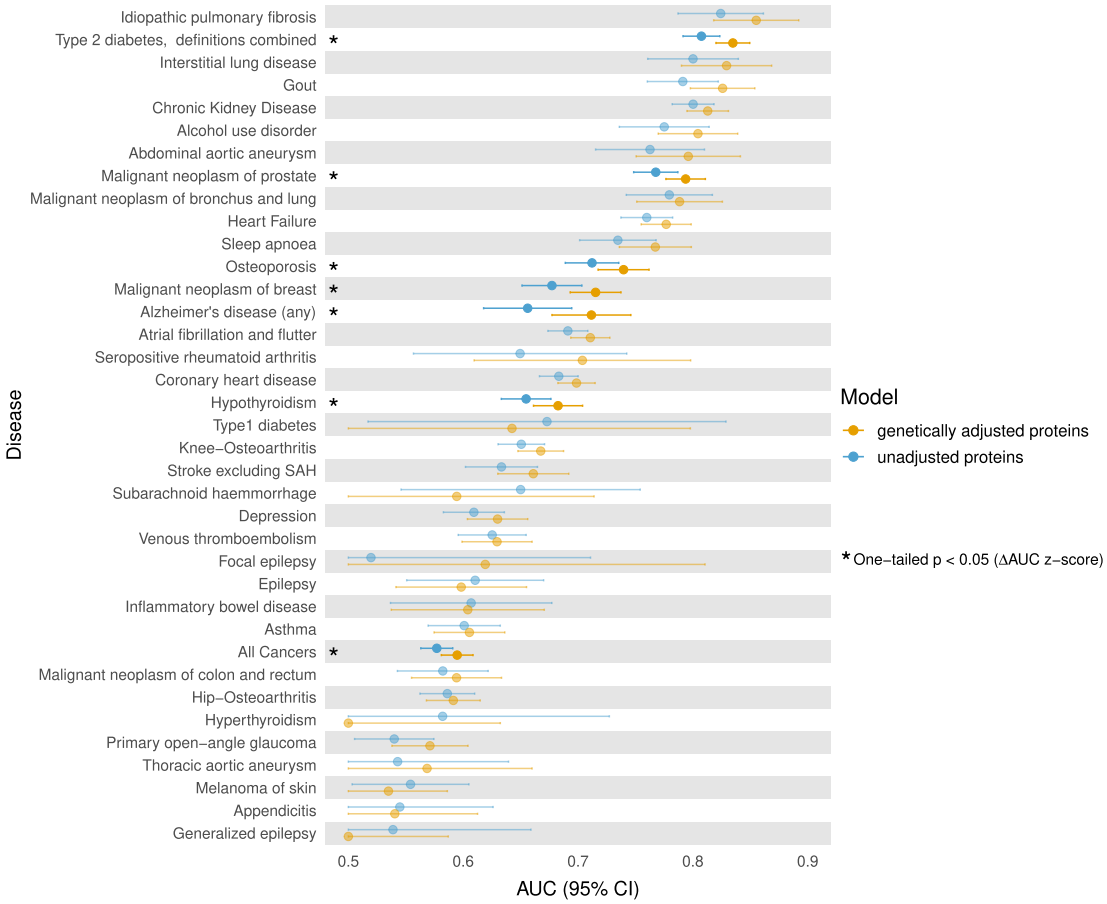


Figure 3.5: **Multi-protein disease prediction.** LASSO-derived AUC for disease prediction using unadjusted proteins (blue) or genetically adjusted proteins (orange). Stars and solid-coloured points indicate nominal significance (one-tailed $p < 0.05$) for ΔAUC ($\text{AUC}_{\text{adjusted}} - \text{AUC}_{\text{unadjusted}}$) z-score.

3.2.9 Replication using FinnGen-derived polygenic scores and 7 additional proteins

So far, genetic adjustment of proteins was performed using PGS derived by OmicsPred based on GWAS from the INTERVAL study using the SomaScan assay (v.3) and four different panel-specific Olink assays. To further validate our results, we considered protein GWAS from 1,829 individuals in FinnGen measured with the same Olink platform used in the UKB, and we derived PGS using an alternative approach (LDpred2[164]). We identified seven additional proteins not included

in the main analysis that met our criteria of low correlation with the existing proteins and strong predictive performance of the PGS. Results were consistent with the main analysis: among 59 significant protein–disease pairs (FDR < 0.05), two showed significantly stronger associations after genetic adjustment, while none favored the unadjusted proteins (Figure 3.6). On average, z-scores increased by 32.1% when using genetically adjusted proteins.

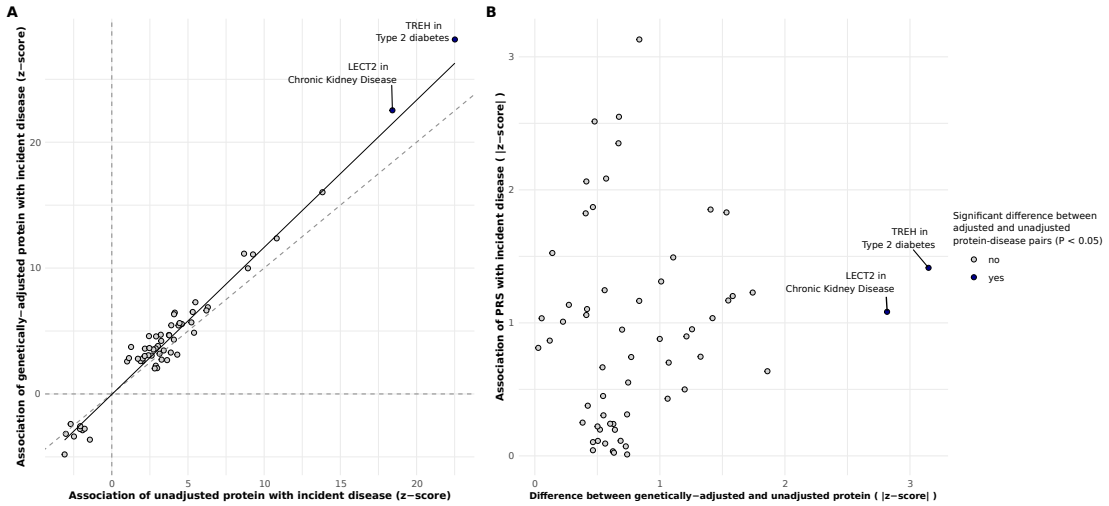


Figure 3.6: **Significant associations between seven additional proteins and 37 diseases in UK Biobank, using PGS weights from FinnGen. A** Logistic regression z-scores for associations with incident diseases for unadjusted proteins (x-axis) and genetically adjusted proteins (y-axis). **B** Absolute differences in z-scores between genetically adjusted and unadjusted protein models (x-axis) and absolute z-scores for the association between protein PGS and diseases (y-axis). Blue indicates a significant difference in effect sizes between adjusted and unadjusted protein associations ($p < 0.05$). Only associations significant at FDR < 0.05 in at least one model (unadjusted or adjusted protein) are shown; a plot including all pairs is provided as Supplementary Figure B.10.

3.3 Discussion

In this study, we show that, in general, removing genetic effects from highly heritable proteins makes them more powerful predictive biomarkers of common disease. This benefit increases with how well PGSs capture protein variation (see Supplementary Methods for a technical explanation). Therefore, the value of genetic

adjustment is expected to increase as larger protein GWASs are performed and more powerful PGSs are derived.

For most proteins, improvements arise because genetic effects, while being strong predictors of protein levels, are not associated with disease. Removing inter-individual differences in protein levels attributable to inherited genetic variation therefore strengthens associations with diseases or environmental factors linked to disease.

Interestingly, genetic adjustment also uncovered associations that were not significant in the unadjusted model. For example, SIRPA showed significant associations with knee osteoarthritis, epilepsy, and three cardiovascular endpoints only after genetic adjustment. SIRPA is best known for its role in the CD47-SIRPA checkpoint[165], and increased SIRPA expression has been reported in inflamed synovial tissue from rheumatoid arthritis[166]. However, the role of SIRPA as a predictive biomarker for these conditions remains poorly understood and may have been obscured by the strong genetic influence on SIRPA levels ($R^2 = 0.78$).

Although it may be tempting to extend the principle of removing *unwanted variation* beyond genetics, it is important to recognize that most non-genetic factors are likely to act as confounders or mediators of the protein–disease relationship, rather than as causal factors for protein levels unrelated to disease. An exception is technical variation (e.g., batch effects), which fits this definition, and adjusting for it has been shown to improve protein-disease associations[167, 168].

A corollary of our findings is that smaller sample sizes are required to detect significant protein–disease associations when using genetically adjusted proteins. We show theoretically (Supplementary Methods) that the percentage reduction in required sample size is linearly related to the proportion of protein variance explained by genetics (R^2). We estimate that the significant protein–disease pairs in our study could have been identified with similar statistical power using, on average, $\sim$30% fewer individuals. This has implications for the design of biomarker discovery studies and clinical trials where non-causal protein biomarkers are used

as outcomes.

To capture the genetic contribution to protein expression, we used polygenic scores (PGSs). While this approach maximizes the proportion of genetic variance explained, it can also introduce biases. In particular, PGS may include variants that influence both protein levels and disease (horizontal pleiotropy), which ideally should not be removed. These pleiotropic effects are more likely to arise from trans variants, simply because they are included in the PGS in larger numbers than cis variants. One illustrative example is MAN2B2 and Alzheimer's disease: an *APOE* variant has a small trans effect on MAN2B2 but a large causal effect on Alzheimer's risk. A PGS that includes *APOE* variants can therefore induce a strong, possibly spurious, association between genetically adjusted MAN2B2 and Alzheimer's disease. On a technical note, even trans variants with small effects on protein expression can have their impact magnified when combined with strong-effect cis variants in the PGS used for genetic adjustment. In the MAN2B2 example, neither cis-PGS nor trans-PGS adjustment alone induced a spurious association with Alzheimer's disease, but the combined PGS did (Supplementary Figure B.9A). We provide a technical explanation of this phenomenon in the Supplementary Methods.

Based on these observations, and also considering the well-known challenge of PGS transferability across ancestry groups, we suggest using cis-pQTLs rather than genome-wide PGSs when the primary goal is to minimize false positives. Nonetheless, we also note that, overall, trans-adjusted proteins still showed improved disease associations compared to unadjusted proteins.

A limitation of this study is that we considered only 94 proteins in the main analyses and seven in the replication analyses. Whether these findings generalize beyond this set remains to be determined. Of the 1,108 proteins initially evaluated, only 99 had a PGS that satisfied our inclusion criteria ($R^2 > 0.2$), and five were excluded due to high correlation with other proteins. The relatively small number of highly genetically predicted proteins can be partly explained by some genetic scores being based on SomaScan and applied to predict Olink proteins in

the UK Biobank, where cross-platform correlations between Olink and SomaScan proteins are modest (≈ 0.3)[167].

There is no reason to believe that the benefits of genetic adjustment should not extend to proteins with lower PGS $R^2 < 0.2$. For our results not to generalize, one would have to assume that highly heritable proteins behave fundamentally differently from less heritable proteins and are less likely to be causal for disease, an assumption for which we find little support. It is instead more likely that most proteins are not causal and instead capture risk factors shared with disease[169].

Another conceptual limitation is that we did not use genetic information directly for disease prediction. One might argue that, once genetic information is available, it could be added on top of unadjusted proteins to improve disease prediction, rather than being used to remove genetic effects from proteins. While this approach would likely enhance predictive performance, our aim was to demonstrate that genetic adjustment is, in itself, a valuable strategy-particularly when the focus is on protein-disease discovery or when proteins are used as outcomes in intervention studies.

In conclusion, highly heritable proteins are strong predictors of common diseases, but their genetic component often reduces their utility as predictive biomarkers. Accounting for genetic effects offers a promising strategy to enhance biomarker discovery and enable the development of more powerful multi-protein-based prediction models.

3.4 Methods

3.4.1 Study population

The UK Biobank (UKB) is a prospective cohort of ∼500,000 participants aged 40–69 years at recruitment. We restricted the analyses to individuals of European ancestry (EUR) with available plasma proteomics data, measured using the antibody-based Olink Explore 3072 proximity extension assay[85, 170] (N = 49,005). We

further restricted analyses to individuals with at least 2,500 proteins measured at baseline (i.e., the time of recruitment) and whose genetic sex (UKB field ID: 22001) was concordant with their self-reported sex (UKB field ID: 31). After all filters, the final cohort consisted of 39,880 individuals.

3.4.2 Proteomics imputation

Within the UKB, we imputed missing values using a random forest-based approach. We first excluded proteins with more than 20% missingness, resulting in a set of 2,919 proteins. We then randomly selected a sample of 10,000 individuals with any genetic ancestry and trained a random forest imputation model using the MissForest[171] library in Python. We imputed a single dataset using a maximum of five iterations during training, with the other parameters left at default values.

3.4.3 Selection of OmicsPred genetic scores

We used protein genetic scores from the OmicsPred database[160], trained on the INTERVAL cohort[159]. We extracted weights derived from both Olink (302 proteins) and SomaLogic (989 proteins) assays, filtering for those also measured in the UKB by the Olink platform. We chose to also include those coming from the SomaLogic assay, given the limited availability of Olink-based summary statistics within the database. Among these proteins, there was an intersection of 183 common proteins, resulting in 1,108 unique proteins.

3.4.4 Polygenic score-based adjustment of proteins

We estimated polygenic scores (PGS) for each protein in EUR UKB individuals using PLINK 2.0.0[172], based on SNP weights obtained from OmicsPred.

Proteins were filtered to retain those for which the R^2 between their observed levels and the corresponding PGS exceeded 0.2, using results from the platform with the higher R^2 in cases where the same protein was included in both platforms. The R^2 values were calculated using protein measurements at baseline, removing missing values (the sample size varied across proteins). To reduce redundancy, we further excluded highly correlated proteins (absolute Pearson correlation > 0.5)

using the `findCorrelation` function in `caret` in R [173], resulting in a final set of 94 proteins.

Then, using the previously described imputed values for 39,880 individuals, we isolated the non-genetic component of each protein by regressing out the variance explained by the corresponding PGS using linear regression and extracting residuals, which were then standardized across individuals. These standardized residuals represent the genetically adjusted proteins used in downstream analyses.

3.4.5 Disease definitions

We selected 37 diseases based on phenotype definitions previously described by Jermy et al.[161], extending the panel by applying the same criteria—specifically, harmonization using FinnGen ICD code lists. We defined end of follow-up as the age of the first record of a disease diagnosis (for individuals with the disease), age at death for a cause other than the disease, age at last record available in the registries or electronic health records, or age 80, whichever occurred first. We excluded prevalent cases, which had their first occurrence before or up to the date of the first assessment visit, and incident cases recorded within the first 6 months of follow-up. A total of 39,871 individuals were included in the analysis. The sample size varied across diseases.

3.4.6 Disease association analysis

We evaluated associations between proteins and disease endpoints using logistic regression models across all 3,478 protein-disease pairs (94 proteins $\times$ 37 diseases). For each pair, three models were fitted independently, each using a different predictor, using the `glmnet` R package[174] and adjusting for age and sex. We used these predictors: 1) unadjusted protein; 2) genetically adjusted protein; 3) PGS. All predictors were standardized to facilitate the comparison of effect sizes across models (mean = 0, standard deviation = 1). Significant associations were defined as those with a false discovery rate (FDR)-adjusted p-value < 0.05 in either the unadjusted or genetically adjusted protein model. In order to test for a significant improvement in the associations, we computed the z-score of the difference in effect

sizes between the genetically adjusted and the unadjusted protein models, and a two-tailed p-value assessing the significance of the difference.

We performed an enrichment analysis restricted to significant protein–disease pairs. Within this set, we defined a subset of significantly different pairs, in which the absolute z-score was greater for the genetically adjusted model compared to the unadjusted one. For each disease, we counted its occurrences in both the subset and the full set, and tested for over-representation using Fisher's exact test. Then we ranked diseases by Fisher's p-value.

To quantify the effective change in statistical power when using unadjusted vs. genetically adjusted proteins, we compared the z-scores from the disease–protein association models. The relative sample size R was defined as the squared ratio between the unadjusted and genetically adjusted z-scores. R values greater than 1 indicate that fewer individuals are required to reach the same power when using residualized proteins, and R values smaller than 1 indicate a relative loss of power.

3.4.7 Cis- trans- analysis

To isolate cis- and trans-acting genetic effects, we defined for each protein the cis-locus as a 1 megabase (Mb) window surrounding the start and end positions of the corresponding gene. To retrieve the genomic positions, we used the GRCh37 Ensembl archive from the `biomaRt` R package[175]. Using BEDTools[176], we filtered OmicsPred genetic scores by performing genomic intersection with the defined cis- and trans-loci. The resulting filtered weights were then used to compute PGS following the same methodology described previously. Among the proteins analyzed, three had only trans-pQTLs, 36 had only cis-pQTLs, and 55 had both cis- and trans-pQTLs. We then extracted the corresponding genetically adjusted proteins from each cis- and trans-PGS, where available. Three proteins had only trans-pQTLs, while 36 proteins had only cis-pQTLs, so 55 proteins had both. Logistic regression analyses were conducted using the cis- and trans-genetically adjusted proteins as predictors for the same disease outcomes, adjusting for age and sex.

3.4.8 Exposome analysis

To better interpret the information captured by the genetically adjusted proteins, we assessed the association with a set of environmental exposures. We selected a list of exposures with fewer than 1,000 missing values from our dataset of 39,880 individuals, resulting in a final set of 86 traits. In addition, we included three laboratory measurements: body mass index (UKB field ID: 21001), creatinine (UKB field ID: 30700), and platelet count (UKB field ID: 30080). We performed separate logistic regression analysis for each categorical exposure, and linear regression for each continuous exposure, using the 94 genetically adjusted proteins as predictors and adjusting for age and sex. We further ran a parallel model replacing genetically adjusted proteins with the standardized unadjusted proteins, allowing for a direct comparison of effect sizes. Continuous exposures were standardized prior to analysis. Categorical exposures were one-hot encoded, with categories represented as separate binary variables; categories with fewer than 100 observations were excluded to ensure model stability. Sample size varied across traits. Significant associations were defined as those with an FDR-adjusted p-value < 0.05 in either the unadjusted or genetically adjusted protein model. In order to test for a significant improvement in the associations, we computed the z-score of the difference in effect sizes between the genetically adjusted and the unadjusted protein models, and a two-tailed p-value assessing the significance of the difference. We performed an enrichment analysis restricted to protein–exposure pairs with FDR < 0.05 in at least one model. Within this set, we defined a subset of pairs with significantly different effect sizes, in which the absolute z-score was greater for the genetically adjusted model. For each exposure, we counted its occurrences in both the subset and the full set of significant pairs, and tested for over-representation using Fisher's exact test. Then we ranked exposures by Fisher's p-value. The same procedure was applied to identify the most enriched proteins.

3.4.9 Multi-protein analysis

To investigate the predictive performance of multiple proteins, we implemented a regularized logistic regression framework using the Least Absolute Shrinkage and Selection Operator (LASSO). As in previous analyses, prevalent cases and inci-

dent cases recorded within the first 6 months of follow-up were excluded. For each disease endpoint, we constructed two models using two different sets of predictors: genetically adjusted proteins and unadjusted proteins. For each model, we randomly partitioned the data into training (70%) and test (30%) sets using the `caret::createDataPartition()` function in R [173], matching the proportion of cases and controls in the train and test sets to avoid class imbalance. LASSO logistic regression was fitted to the training data using the `glmnet` 4.1-9 R package[177]. We performed 10-fold cross-validation to select the optimal lambda parameter (λ) that minimized binomial deviance (λ_{min}). Features with non-zero coefficients at the optimal λ were retained as predictive variables. Using the selected features and their coefficients, we computed weighted linear combinations of scores for individuals in both the training and test sets, which were standardized within each set. A secondary logistic regression model was then fitted in the training set using the LASSO-derived score as the predictor. This model was applied to the test set to quantify predictive performance by calculating the area under the curve (AUC) using the `pROC` R package[178]. To quantify significant differences, we calculated the z-score of the difference between the AUC of adjusted and unadjusted protein models, and the corresponding one-tailed p-value of this difference.

3.4.10 Replication with FinnGen pQTLs

We first selected 2,832 proteins from the FinnGen[2] 3 QCd Olink proteomics dataset of 1,829 independent samples that were also measured in the UKB. Protein heritability (h^2) estimation in FinnGen was performed using the Genome-based Restricted Maximum Likelihood (GREML) method, as implemented in GCTA[179]. We first constructed a genetic relationship matrix (GRM) based on 190,180 high-quality, common, and independent SNPs, filtered with the following criteria: INFO score > 0.9, minor allele frequency (MAF) > 0.01, genotyping missingness $< 3\%$, and linkage disequilibrium pruning using a window size of 1,500 kb and pairwise $R^2 < 0.2$. GREML analysis was then conducted using the resulting GRM and protein level measurements from 1,829 eligible individuals in the FinnGen cohort.

We then filtered for proteins with a significant h^2 estimate ($p < 0.05$) and

$h^2 > 0.3$, resulting in a set of 237 proteins. To reduce redundancy, we retained proteins with pairwise absolute intercorrelation (Pearson r) below 0.5. Next, we removed proteins that were either present in or highly correlated (absolute Pearson correlation > 0.5) with our original set of proteins, yielding a final subset of 135 proteins. Using FinnGen GWAS data for 135 selected proteins, we applied the LDpred2-auto model implemented in the R package `bigsnpr` to estimate SNP weights for the PGS calculation[164]. We used the HapMap3+ linkage disequilibrium (LD) matrix provided with the package[180], which offers improved genome coverage. After excluding SNPs with minor allele frequencies (MAF) below 1%, a total of 1,301,162 SNPs with available LD information were used. The model failed to converge for two proteins (APOBR and LRRC37A2), which were subsequently excluded. For the remaining 133 proteins, we estimated the PGS for UKB individuals with PLINK 2.0.0[172], based on the SNP weights derived from LDpred2. Finally, we retained only those proteins ($N = 7$) for which the variance explained by the PGS (R^2) in measured protein levels was greater than 0.2, resulting in a high-confidence set of genetically predictable proteins for downstream analysis.

CHAPTER 4

Other related work

In this chapter, additional work is briefly introduced to complement the core studies presented in this thesis. Although these analyses do not constitute the primary focus of the thesis, they are nonetheless relevant, as they extend and apply the methodological frameworks described and developed here to different contexts, such as evolutionary and population genetics. By doing so, they illustrate the broader applicability, flexibility, and robustness of these approaches beyond the specific use cases explored in the main chapters, and across a diverse range of research questions.

4.1 Sequence composition and conservation predict the phenotypic relevance of transposable elements

4.1.1 Abstract

Transposable elements (TEs) are powerful drivers of genome evolution, in part through their ability to rapidly rewire the host regulatory network by creating transcription factor binding sites that can potentially be turned to the host's advantage in a process called exaptation. However, the effects on the host phenotype vary widely among different TEs. Here, we classify TEs based on their contribution to the human host phenotypes at both molecular and macroscopic scales.
TE contributions to chromatin accessibility, gene expression, and the heritability of complex traits are strongly correlated to each other, confirming that the main mechanism through which TEs affect the host phenotype is through the rewiring of the regulatory network. TE sequence and evolutionary features are able to explain a large fraction of the variance in phenotypic relevance, and in particular more than 50% of the variance of their contribution to the heritability of complex traits. A conspicuous exception to this pattern is represented by TEs of the ERV1 family, whose phenotypic impact cannot be explained by our model: in particular, this family includes a set of relatively young TEs whose phenotypic relevance is much larger than would be expected based on their sequence and evolutionary parameters. These TEs are involved in fast-evolving biological processes related to the interaction of the organism with its environment.
In conclusion, our results confirm quantitatively that TE insertions affect the host phenotype mostly through the rewiring of its regulatory network; identify a signature of phenotypic relevance based on sequence and conservation properties; and highlight several TEs as promising candidates for functional studies.

4.1.2 My contribution

My contribution was related to assessing the macroscopic phenotypic effects of TEs that are driven by regulatory intermediate molecular phenotypes. In particular, for each TE, its effect on gene expression was considered, evaluated as the number

of genes with at least one expression quantitative trait locus (eQTL) residing inside a copy of the TE in at least one GTEx tissue, divided by the number of common variants found in the TE. This quantity estimates the effect of TE sequence on gene expression by evaluating the effect of the corresponding sequence variants. We also assessed the contribution of each TE with complex trait heritability and chromatin accessibility. As shown in Fig. 4.1A, all pairwise correlations between the three measures are positive and highly significant (all $P < 2.2 \times 10^{-16}$). As expected, the two molecular phenotypes show high concordance (Spearman's $\rho = 0.59$), confirming that the effect of TE exaptation on gene expression is partly driven by the creation of regulatory elements characterized by accessible chromatin.

Perhaps less obvious is the high correlation between effects at the molecular and macroscopic levels ($\rho = 0.57$ and 0.63 of complex trait heritability with chromatin accessibility and gene expression, respectively). These results generally apply to all TE families (Fig. 4.1B–E), and suggest that the contribution of TEs to complex trait variability is largely driven by their effects on the regulatory network. Macroscopic traits are the targets of selection, and complex traits in particular are known to be subjected to stabilizing selection.

The full manuscript is currently under review at *Genome Biology* and is available as a preprint on bioRxiv.

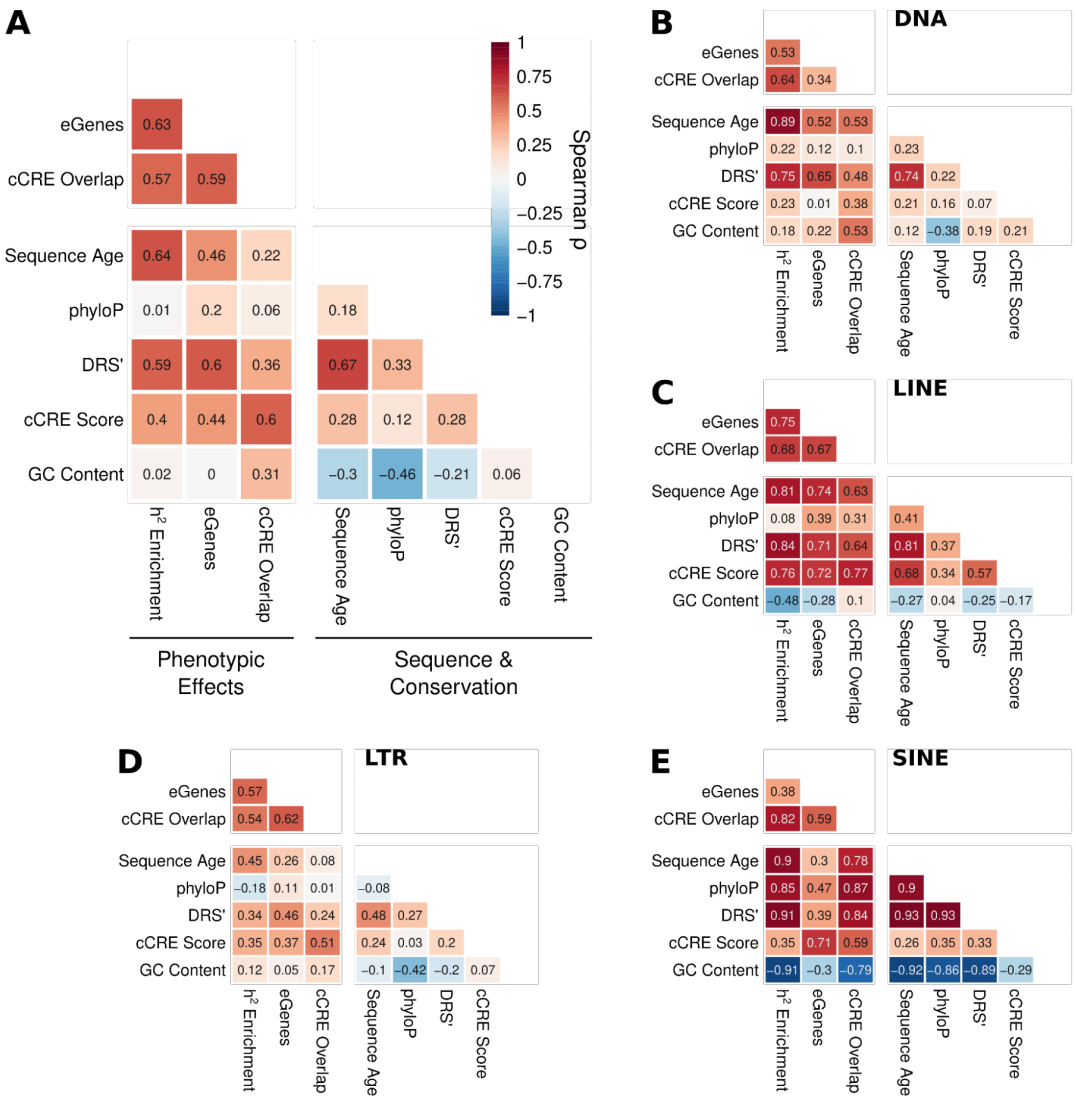


Figure 4.1: A: Spearman correlation between measures of phenotypic impact and sequence/conservation features. B-E: Same by TE class.

4.2 Ancestral genetic components are consistently associated with the complex trait landscape in European biobanks

4.2.1 Abstract

The genetic structure in Europe was mostly shaped by admixture between the Western Hunter-Gatherers, Early European Farmers and Steppe Bronze Age ancestral components. Such structure is regarded as a confounder in GWAS and follow-up studies, and gold-standard methods exist to correct for it. However, it is still poorly understood to which extent these ancestral components contribute to complex trait variation in present-day Europe. In this work we harness the UK Biobank to address this question. By extensive demographic simulations, exploiting data on siblings and incorporating previous results we obtained from the Estonian Biobank, we carefully evaluate the significance and scope of our findings. Heart rate, platelet count, bone mineral density and many other traits show stratification similar to height and pigmentation traits, likely targets of selection and divergence across ancestral groups. We show that the reported ancestry-trait associations are not driven by environmental confounders by confirming our results when using between-sibling differences in ancestry. The consistency of our results across biobanks further supports this and indicates that these genetic predispositions that derive from post-Neolithic admixture events act as a source of variability and as potential confounders in Europe as a whole.

4.2.2 My contribution

My specific task was to evaluate the heritability enrichment of trait-associated genomic regions (TAGRs) defined in this work. In particular, TAGRs were defined for each trait based on genome-wide significant associations reported in the GWAS Catalog, extending ±20 kb around each lead variant and merging overlapping intervals.

To evaluate if TAGRs capture a meaningful polygenic signal, we quantified their enrichment for the heritability of SNP using stratified LD score regression (sLDSC).

These regions were treated as custom annotations in the sLDSC framework.
We applied sLDSC to 50 complex traits using UK Biobank GWAS summary statistics obtained from Neale Lab (http://www.nealelab.is/UK Biobank/). For each trait, we estimated the proportion of SNP heritability attributable to TAGRs relative to their genomic coverage, yielding an enrichment statistic and the corresponding significance estimate.
As a negative control, we constructed "negative TAGRs" by excluding all genomic regions within ±500 kb of any GWAS Catalog hit, thereby defining regions depleted of known trait associations. These negative TAGRs were analyzed in parallel using the same sLDSC framework.
Across almost all traits, TAGRs showed consistent enrichment for trait heritability (nominal $P < 0.05$), whereas non-TAGRs generally showed depletion or no enrichment in trait heritability (Figure 4.2), although likely harboring variants contributing to the polygenic traits analyzed.
These results demonstrate that the TAGR definitions capture regions that are disproportionately enriched for trait-relevant genetic signal, supporting their use in downstream covariance and integration analyses.
The full manuscript has been published in 2024 in the *European Journal of Human Genetics*.

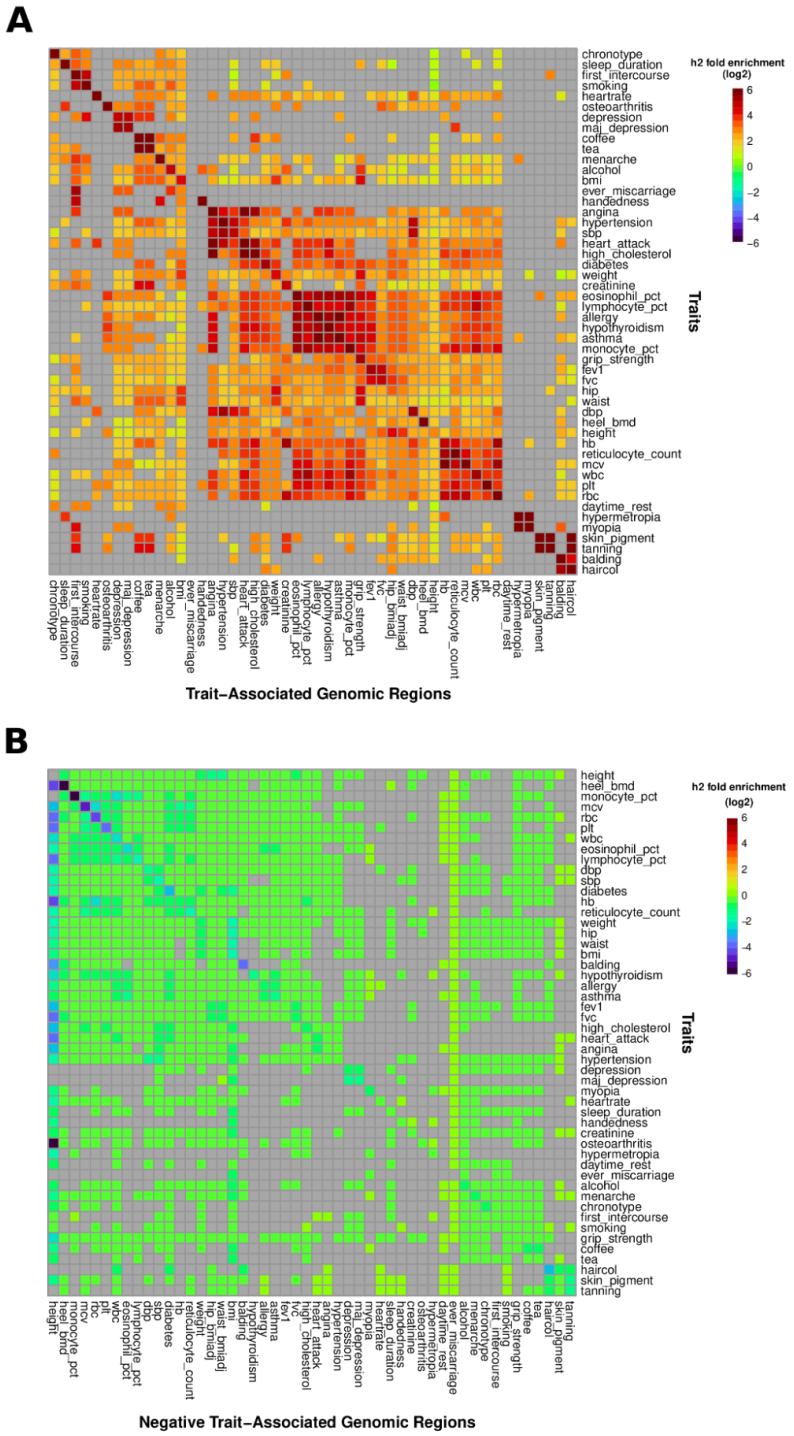


Figure 4.2: Heritability enrichment estimates for (A) TAGRs (50 annotations), and (B) negative TAGRs, defined excluding GWAS hits. Enrichment of traits expressed as (Proportion of heritability)/(Proportion of SNPs). All colored cells indicate nominal significance at P¡0.05, while gray cells are not significant. Annotations are ordered by clustering. Overall, all traits show their highest enrichment in correspondent genomic regions.

Conclusion

In this thesis, I have presented the integration of multiple molecular layers, including genomics, transcriptomics, proteomics, and environmental data, to explore the complex relationships between genetic variation, molecular mechanisms, and disease outcomes.
By leveraging large-scale biobank data, we have demonstrated how multi-omic approaches can uncover new associations that were previously difficult to detect, offering deeper insights into the biological mechanisms underlying complex traits and diseases.
This approach has not only enhanced our understanding of the common genetic basis between phenotypes but has also provided more accurate predictive biomarkers by removing genetic noise from molecular data, particularly in the context of plasma proteins.

In the first study, we demonstrated how integrating data from GWASs, population-based transcriptomic studies, and single-cell chromatin accessibility assays enables the molecular dissection of the genetic correlation between complex traits and the formulation of data-driven hypotheses on the relevant regulatory mechanisms.

In the second study, we showed that, although highly heritable proteins are

strong predictors of common diseases, their genetic component often reduces their utility as predictive biomarkers. Accounting for genetic effects offers a promising strategy to enhance biomarker discovery and enable the development of more powerful multi-protein-based prediction models.

However, despite the advances made in these works, there are notable limitations that warrant further attention. One key limitation is the focus on European populations in both studies. While these populations are well-characterized, the transferability of the findings to other populations remains a significant gap.
This focus was primarily driven by data availability (large, deeply phenotyped cohorts are still mainly European) and by methodological considerations that make ancestry stratification essential rather than optional. In particular, many of the analyses used here rely implicitly on population-specific genetic architecture: LD structure differs across ancestries and influences fine-mapping resolution, polygenic score construction, and the interpretation of the signals of the associations; allele frequencies differences affect power, instrument strength, and the stability of effect estimates; and imputation quality as well as reference panel match can vary substantially, propagating bias into downstream results. These issues are especially relevant when deriving or applying polygenic scores for molecular traits (e.g., proteins), where prediction accuracy can drop significantly, and where trans-acting components may capture pleiotropic signals differently across populations. Similarly, QTL-based integration (eQTL/pQTL) can be sensitive to ancestry-dependent regulatory architectures and differences in cohort composition and recruitment strategies, meaning that causal inference frameworks or prioritization of candidate genes and pathways may not generalize. All these genetic, as well as environmental and lifestyle factors, which vary across populations, can influence the generalizability of the results.
Addressing these issues requires more inclusive and diverse population-based studies to ensure that the insights gained from this research are applicable to a broader range of individuals worldwide. These factors, together, motivate and highlight the need for future work extending these approaches to multi-ancestry datasets, developing ancestry-aware modeling strategies, and validating whether the discovered associations and predictive improvements replicate across diverse genetic and

environmental contexts.
Furthermore, as the field of computational biology rapidly evolves, new machine learning and deep learning techniques are emerging almost daily, offering sophisticated tools for analyzing increasingly large and complex datasets. The current statistical methods, while being gold standards in many areas, may no longer be sufficient to fully capture the intricate relationships between molecular layers, especially as the data becomes high-dimensional. Therefore, there is a need to apply these novel computational techniques to improve the precision of our analyses and to overcome the limitations of traditional methods. Incorporating these advanced approaches will not only enhance the statistical power of existing models but also open up new avenues for biomarker discovery and disease prediction, ultimately advancing the field of personalized medicine.

In conclusion, this thesis demonstrates how the integration of multiple molecular layers can enhance our understanding of complex trait biology. By leveraging large-scale biobank resources, the work highlights the potential of multi-omic approaches to move beyond purely statistical associations toward biologically interpretable and clinically relevant insights. Looking ahead, the next major step for the field lies in integrating data across multiple European biobanks. Pan-European initiatives and infrastructures, along with the emerging European Health Data Space (EHDS), offer a unique opportunity to combine genetic, molecular, and electronic health record data. Realizing this potential will require extensive harmonization efforts, such as genetic data processing, ancestry-aware analytical frameworks, and the standardization of clinical phenotypes, disease definitions, and longitudinal health records across heterogeneous healthcare systems. Such harmonization is not only a technical challenge but a prerequisite for ensuring comparability, reproducibility, and equitable transferability of findings across populations.
As computational methodologies continue to advance, integrated, harmonized pan-European biobank data will provide the foundation for next-generation predictive and mechanistic models. Ultimately, these efforts will enable more robust, generalizable, and clinically actionable insights, positioning European biobank integration as a cornerstone of future precision medicine and population health research.

List of publications

Peer-reviewed articles

- **Fusco D**, Marinelli C, André M, Troiani L, Noè M, Pizzagalli F, Marnetto D, Provero P. (2025). Exploring the molecular basis of the genetic correlation between body mass index and brain morphological traits. *PLOS Genetics* 21(4): e1011658. https://doi.org/10.1371/journal.pgen.1011658

- Pankratov M, Mezzavilla M, Aneli S, Kuznetsov I, **Fusco D**, Wilson J, Metspalu M, Provero P, Pagani L, Marnetto D. (2024). Ancestral genetic components are consistently associated with the complex trait landscape in European biobanks. *European Journal of Human Genetics*. https://doi.org/10.1038/s41431-024-01678-9

Preprints and manuscripts under review

- **Fusco D**, Yang Z, Viippola E, Cajuso T, Corbetta A, Caime C, German J, Fu M, Argentieri M.A., FinnGen, Kachuri L, Provero P, Nakanishi T, Yang Z, Ganna A. (2025). *medRxiv* 2025.10.14.25337894. https://doi.org/10.1101/2025.10.14.25337894 (under review at *Nature Genetics*)

- Cerruti Y, **Fusco D**, Jahanbin M, Marnetto D, Provero P. (2025). Se-

quence composition and conservation predict the phenotypic relevance of transposable elements. *bioRxiv* 2024.03.11.584453. https://doi.org/10.1101/2024.03.11.584453 (under review at *Genome Biology*)

- Viippola E*, Hartonen T*, Zhu M, Yang Z, Edahiro R, Nakanishi T, Lan A, Kanai M, Cajuso T, **Fusco D**, ... Ganna A. (2025). Sex-related signature in the plasma proteome not attributable to XX and XY chromosomes and their impact on health and disease. (under review at *Science*)

- Rossi F, Sportelli L, Kikidis G.C., Grassi G, Di Camillo F, Bertolino A, Blasi G, **Fusco D**, ... Pergola G. (2026). Co-expression-based models improve eQTL predictions and highlight many novel transcriptome-wide genes associated with schizophrenia. (under review at *Nature Genetics*)

Acknowledgements

Bibliography

[1] O. G. Bahcall. "UK Biobank — a new era in genomic medicine". en. In: *Nature Reviews Genetics* 19.12 (Dec. 2018), pp. 737–737.

[2] M. I. Kurki et al. "FinnGen provides genetic insights from a well-phenotyped isolated population". en. In: *Nature* 613.7944 (Jan. 2023), pp. 508–518.

[3] A. Abdellaoui et al. "15 years of GWAS discovery: Realizing the promise". en. In: *The American Journal of Human Genetics* 110.2 (Feb. 2023), pp. 179–194.

[4] GTEx Consortium. "The GTEx Consortium atlas of genetic regulatory effects across human tissues". eng. In: *Science (New York, N.Y.)* 369.6509 (Sept. 2020), pp. 1318–1330.

[5] B. B. Sun et al. "Genomic atlas of the human plasma proteome". en. In: *Nature* 558.7708 (June 2018), pp. 73–79.

[6] V. Laville et al. "Gene-lifestyle interactions in the genomics of human complex traits". en. In: *European Journal of Human Genetics* 30.6 (June 2022), pp. 730–739.

[7] E. Herrera-Luis et al. "Gene–environment interactions in human health". en. In: *Nature Reviews Genetics* 25.11 (Nov. 2024), pp. 768–784.

[8] J. Carrasco-Zanini et al. “Mapping biological influences on the human plasma proteome beyond the genome”. eng. In: *Nature Metabolism* 6.10 (Oct. 2024), pp. 2010–2023.

[9] Y. Hasin, M. Seldin, and A. Lusis. “Multi-omics approaches to disease”. en. In: *Genome Biology* 18.1 (Dec. 2017), p. 83.

[10] R. J. Klein et al. “Complement Factor H Polymorphism in Age-Related Macular Degeneration”. en. In: *Science* 308.5720 (Apr. 2005), pp. 385–389.

[11] T. W. T. C. C. Consortium. “Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls”. In: *Nature* 447 (2007), pp. 661–678.

[12] R. Sladek et al. “A genome-wide association study identifies novel risk loci for type 2 diabetes”. en. In: *Nature* 445.7130 (Feb. 2007), pp. 881–885.

[13] the NIDDK IBD Genetics Consortium et al. “Genome-wide association defines more than 30 distinct susceptibility loci for Crohn’s disease”. en. In: *Nature Genetics* 40.8 (Aug. 2008), pp. 955–962.

[14] J. Marchini and B. Howie. “Genotype imputation for genome-wide association studies”. en. In: *Nature Reviews Genetics* 11.7 (July 2010), pp. 499–511.

[15] A. B. Popejoy and S. M. Fullerton. “Genomics is failing on diversity”. en. In: *Nature* 538.7624 (Oct. 2016), pp. 161–164.

[16] M. C. Mills and C. Rahal. “The GWAS Diversity Monitor tracks diversity by disease in real time”. en. In: *Nature Genetics* 52.3 (Mar. 2020), pp. 242–243.

[17] A. Nagai et al. “Overview of the BioBank Japan Project: Study design and profile”. en. In: *Journal of Epidemiology* 27.3 (Mar. 2017), S2–S8.

[18] R. G. Walters et al. “Genotyping and population characteristics of the China Kadoorie Biobank”. en. In: *Cell Genomics* 3.8 (Aug. 2023), p. 100361.

[19] Y. Ruan et al. “Improving polygenic prediction in ancestrally diverse populations”. en. In: *Nature Genetics* 54.5 (May 2022), pp. 573–580.

[20] P. M. Visscher, W. G. Hill, and N. R. Wray. "Heritability in the genomics era — concepts and misconceptions". en. In: *Nature Reviews Genetics* 9.4 (Apr. 2008), pp. 255–266.

[21] J. Yang et al. "Concepts, estimation and interpretation of SNP-based heritability". en. In: *Nature Genetics* 49.9 (Sept. 2017), pp. 1304–1310.

[22] O. Zuk et al. "The mystery of missing heritability: Genetic interactions create phantom heritability". en. In: *Proceedings of the National Academy of Sciences* 109.4 (Jan. 2012), pp. 1193–1198.

[23] P. Wainschtein et al. "Estimation and mapping of the missing heritability of human phenotypes". en. In: *Nature* (Nov. 2025).

[24] Schizophrenia Working Group of the Psychiatric Genomics Consortium et al. "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies". en. In: *Nature Genetics* 47.3 (Mar. 2015), pp. 291–295.

[25] H. Wand et al. "Improving reporting standards for polygenic scores in risk prediction studies". en. In: *Nature* 591.7849 (Mar. 2021), pp. 211–219.

[26] S. A. Lambert, G. Abraham, and M. Inouye. "Towards clinical utility of polygenic risk scores". eng. In: *Human Molecular Genetics* 28.R2 (Nov. 2019), R133–R142.

[27] A. Torkamani, N. E. Wineinger, and E. J. Topol. "The personal and clinical utility of polygenic risk scores". en. In: *Nature Reviews Genetics* 19.9 (Sept. 2018), pp. 581–590.

[28] N. Chatterjee, J. Shi, and M. García-Closas. "Developing and evaluating polygenic risk prediction models for stratified disease prevention". en. In: *Nature Reviews Genetics* 17.7 (July 2016), pp. 392–406.

[29] C. M. Lewis and E. Vassos. "Polygenic risk scores: from research tools to clinical instruments". eng. In: *Genome Medicine* 12.1 (May 2020), p. 44.

[30] T. Ge et al. "Polygenic prediction via Bayesian regression and continuous shrinkage priors". eng. In: *Nature Communications* 10.1 (Apr. 2019), p. 1776.

[31] B. J. Vilhjálmsson et al. "Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores". eng. In: *American Journal of Human Genetics* 97.4 (Oct. 2015), pp. 576–592.

[32] G. Abraham et al. "Genomic prediction of coronary heart disease". eng. In: *European Heart Journal* 37.43 (Nov. 2016), pp. 3267–3278.

[33] A. Lee et al. "BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors". en. In: *Genetics in Medicine* 21.8 (Aug. 2019), pp. 1708–1718.

[34] W. Cookson et al. "Mapping complex disease traits with global gene expression". eng. In: *Nature Reviews. Genetics* 10.3 (Mar. 2009), pp. 184–194.

[35] R. A. Fisher. "XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance." en. In: *Transactions of the Royal Society of Edinburgh* 52.2 (1919), pp. 399–433.

[36] T. A. Manolio et al. "Finding the missing heritability of complex diseases". eng. In: *Nature* 461.7265 (Oct. 2009), pp. 747–753.

[37] E. A. Boyle, Y. I. Li, and J. K. Pritchard. "An Expanded View of Complex Traits: From Polygenic to Omnigenic". en. In: *Cell* 169.7 (June 2017), pp. 1177–1186.

[38] Y. I. Li et al. "RNA splicing is a primary link between genetic variation and disease". eng. In: *Science (New York, N.Y.)* 352.6285 (Apr. 2016), pp. 600–604.

[39] J. K. Pickrell. "Joint analysis of functional genomic data and genome-wide association studies of 18 human traits". eng. In: *American Journal of Human Genetics* 94.4 (Apr. 2014), pp. 559–573.

[40] D. Welter et al. "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations". eng. In: *Nucleic Acids Research* 42.Database issue (Jan. 2014), pp. D1001–1006.

[41] W. Van Rheenen et al. "Genetic correlations of polygenic disease traits: from theory to practice". en. In: *Nature Reviews Genetics* 20.10 (Oct. 2019), pp. 567–581.

[42] A. Tenesa and C. S. Haley. "The heritability of human disease: estimation, uses and abuses". en. In: *Nature Reviews Genetics* 14.2 (Feb. 2013), pp. 139–149.

[43] D. S. Falconer. "The inheritance of liability to certain diseases, estimated from the incidence among relatives". en. In: *Annals of Human Genetics* 29.1 (Aug. 1965), pp. 51–76.

[44] S. Lee et al. "Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood". en. In: *Bioinformatics* 28.19 (Oct. 2012), pp. 2540–2542.

[45] H. K. Finucane et al. "Partitioning heritability by functional annotation using genome-wide association summary statistics". eng. In: *Nature Genetics* 47.11 (Nov. 2015), pp. 1228–1235.

[46] H. Shi et al. "Local Genetic Correlation Gives Insights into the Shared Genetic Architecture of Complex Traits". en. In: *The American Journal of Human Genetics* 101.5 (Nov. 2017), pp. 737–751.

[47] B. C. Brown et al. "Transethnic Genetic-Correlation Estimates from Summary Statistics". en. In: *The American Journal of Human Genetics* 99.1 (July 2016), pp. 76–88.

[48] J.-B. Pingault et al. "Using genetic data to strengthen causal inference in observational research". en. In: *Nature Reviews Genetics* 19.9 (Sept. 2018), pp. 566–580.

[49] G. Davey Smith and S. Ebrahim. "What can mendelian randomisation tell us about modifiable behavioural and environmental exposures?" eng. In: *BMJ (Clinical research ed.)* 330.7499 (May 2005), pp. 1076–1079.

[50] *Burgess, S. & Thompson, S. G. Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation. (CRC Press, Boca Raton, 2015).*

[51] N. A. Sheehan and V. Didelez. "Commentary: Can 'many weak' instruments ever be 'strong'?" eng. In: *International Journal of Epidemiology* 40.3 (June 2011), pp. 752–754.

[52] E. Sanderson et al. "Mendelian randomization". en. In: *Nature Reviews Methods Primers* 2.1 (Feb. 2022), p. 6.

[53] Z. Zhu et al. "Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets". en. In: *Nature Genetics* 48.5 (May 2016), pp. 481–487.

[54] C. Wallace. "Statistical Testing of Shared Genetic Control for Potentially Related Traits". en. In: *Genetic Epidemiology* 37.8 (Dec. 2013), pp. 802–813.

[55] C. Giambartolomei et al. "Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics". en. In: *PLoS Genetics* 10.5 (May 2014). Ed. by S. M. Williams, e1004383.

[56] V. Zuber et al. "Combining evidence from Mendelian randomization and colocalization: Review and comparison of approaches". en. In: *The American Journal of Human Genetics* 109.5 (May 2022), pp. 767–782.

[57] D. J. Schaid, W. Chen, and N. B. Larson. "From genome-wide associations to candidate causal variants by statistical fine-mapping". eng. In: *Nature Reviews. Genetics* 19.8 (Aug. 2018), pp. 491–504.

[58] C. Benner et al. "FINEMAP: efficient variable selection using summary data from genome-wide association studies". en. In: *Bioinformatics* 32.10 (May 2016), pp. 1493–1501.

[59] F. Hormozdiari et al. "Identifying causal variants at loci with multiple signals of association". en. In: *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. Newport Beach California: ACM, Sept. 2014, pp. 610–611.

[60] G. Wang et al. "A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping". en. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82.5 (Dec. 2020), pp. 1273–1300.

[61] C. Wallace. “A more accurate method for colocalisation analysis allowing for multiple causal variants”. en. In: *PLOS Genetics* 17.9 (Sept. 2021). Ed. by H. J. Cordell, e1009440.

[62] P. Sibilio et al. “Integrating multi-omics data: Methods and applications in human complex diseases”. en. In: *Biotechnology Reports* 48 (Dec. 2025), e00938.

[63] C. N. Hayes et al. “From Omics to Multi-Omics: A Review of Advantages and Tradeoffs”. en. In: *Genes* 15.12 (Nov. 2024), p. 1551.

[64] K. J. Karczewski and M. P. Snyder. “Integrative omics for health and disease”. en. In: *Nature Reviews Genetics* 19.5 (May 2018), pp. 299–310.

[65] T. Das et al. “Integration of Online Omics-Data Resources for Cancer Research”. In: *Frontiers in Genetics* 11 (Oct. 2020), p. 578345.

[66] J. Yan et al. “Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data”. en. In: *Briefings in Bioinformatics* (June 2017).

[67] M. C. M. Wayne Marta. “Quantitative trait locus (QTL) analysis”. In: *Nature Education* (2008).

[68] M. V. Rockman and L. Kruglyak. “Genetics of global gene expression”. en. In: *Nature Reviews Genetics* 7.11 (Nov. 2006), pp. 862–872.

[69] J. Zhang and H. Zhao. “eQTL studies: from bulk tissues to single cells”. en. In: *Journal of Genetics and Genomics* 50.12 (Dec. 2023), pp. 925–933.

[70] H.-J. Westra et al. “Cell Specific eQTL Analysis without Sorting Cells”. en. In: *PLOS Genetics* 11.5 (May 2015). Ed. by T. Pastinen, e1005223.

[71] D. V. Zhernakova et al. “Identification of context-dependent expression quantitative trait loci in whole blood”. en. In: *Nature Genetics* 49.1 (Jan. 2017), pp. 139–145.

[72] F. Avila Cobos et al. “Benchmarking of cell type deconvolution pipelines for transcriptomics data”. en. In: *Nature Communications* 11.1 (Nov. 2020), p. 5650.

[73] S. Kim-Hellmuth et al. “Cell type–specific genetic regulation of gene expression across human tissues”. en. In: *Science* 369.6509 (Sept. 2020), eaaz8528.

[74] A. Gusev et al. “Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights”. en. In: *Nature Genetics* 50.4 (Apr. 2018), pp. 538–548.

[75] F. Aguet et al. “Molecular quantitative trait loci”. en. In: *Nature Reviews Methods Primers* 3.1 (Jan. 2023), p. 4.

[76] J. C. Denny and F. S. Collins. “Precision medicine in 2030-seven ways to transform healthcare”. eng. In: *Cell* 184.6 (Mar. 2021), pp. 1415–1419.

[77] Y.-T. Deng et al. “Atlas of the plasma proteome in health and disease in 53,026 adults”. en. In: *Cell* 188.1 (Jan. 2025), 253–271.e7.

[78] N. S. Abell et al. “Multiple causal variants underlie genetic associations in humans”. en. In: *Science* 375.6586 (Mar. 2022), pp. 1247–1254.

[79] M. Koprulu et al. “Proteogenomic links to human metabolic diseases”. en. In: *Nature Metabolism* 5.3 (Feb. 2023), pp. 516–528.

[80] L. Wu et al. “Variation and genetic control of protein abundance in humans”. en. In: *Nature* 499.7456 (July 2013), pp. 79–82.

[81] O. Canela-Xandri, K. Rawlik, and A. Tenesa. “An atlas of genetic associations in UK Biobank”. en. In: *Nature Genetics* 50.11 (Nov. 2018), pp. 1593–1599.

[82] C. Bycroft et al. “The UK Biobank resource with deep phenotyping and genomic data”. en. In: *Nature* 562.7726 (Oct. 2018), pp. 203–209.

[83] T. J. Littlejohns et al. “The UK Biobank imaging enhancement of 100,000 participants:rationale, data collection, management and future directions”. eng. In: *Nature Communications* 11.1 (May 2020), p. 2624.

[84] J. D. Szustakowski et al. “Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank”. eng. In: *Nature Genetics* 53.7 (July 2021), pp. 942–948.

[85] B. B. Sun et al. “Plasma proteomic associations with genetics and health in the UK Biobank”. en. In: *Nature* 622.7982 (Oct. 2023), pp. 329–338.

[86] M. Lundberg et al. "Homogeneous antibody-based proximity extension assays provide sensitive and specific detection of low-abundant proteins in human blood". en. In: *Nucleic Acids Research* 39.15 (Aug. 2011), e102–e102.

[87] J. C. Rohloff et al. "Nucleic Acid Ligands With Protein-like Side Chains: Modified Aptamers and Their Use as Diagnostic and Therapeutic Agents". eng. In: *Molecular Therapy. Nucleic Acids* 3.10 (Oct. 2014), e201.

[88] E. Ferkingstad et al. "Large-scale integration of the plasma proteome with genetics and disease". en. In: *Nature Genetics* 53.12 (Dec. 2021), pp. 1712–1721.

[89] C. Cortes and V. Vapnik. "Support-vector networks". en. In: *Machine Learning* 20.3 (Sept. 1995), pp. 273–297.

[90] D. Lee et al. "A method to predict the impact of regulatory variants from DNA sequence". en. In: *Nature Genetics* 47.8 (Aug. 2015), pp. 955–961.

[91] N. Pudjihartono et al. "A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction". In: *Frontiers in Bioinformatics* 2 (June 2022), p. 927312.

[92] R. Tibshirani. "Regression Shrinkage and Selection Via the Lasso". en. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (Jan. 1996), pp. 267–288.

[93] J. Lonsdale et al. "The Genotype-Tissue Expression (GTEx) project". en. In: *Nature Genetics* 45.6 (June 2013), pp. 580–585.

[94] The GTEx Consortium et al. "The GTEx Consortium atlas of genetic regulatory effects across human tissues". en. In: *Science* 369.6509 (Sept. 2020), pp. 1318–1330.

[95] J. A. Pagán, V. H.-C. Wang, and H. Sur. "Time to use large-scale biobank databases in health policy and public health research". en. In: *Health Affairs Scholar* 3.8 (Aug. 2025), qxaf158.

[96] The All of Us Research Program Investigators. "The "All of Us" Research Program". en. In: *New England Journal of Medicine* 381.7 (Aug. 2019), pp. 668–676.

[97] J. Alvidrez et al. "Building the Evidence Base to Inform Planned Intervention Adaptations by Practitioners Serving Health Disparity Populations". en. In: *American Journal of Public Health* 109.S1 (Jan. 2019), S94–S101.

[98] D. Batheja et al. "Understanding the value of biobank attributes to researchers using a conjoint experiment". en. In: *Scientific Reports* 13.1 (Dec. 2023), p. 22728.

[99] K. S. Steinsbekk, B. Kåre Myskja, and B. Solberg. "Broad consent versus dynamic consent in biobank research: Is passive participation an ethical problem?" en. In: *European Journal of Human Genetics* 21.9 (Sept. 2013), pp. 897–902.

[100] F. Colledge et al. "Sample and data sharing barriers in biobanking: consent, committees, and compromises". en. In: *Annals of Diagnostic Pathology* 18.2 (Apr. 2014), pp. 78–81.

[101] M. Sariyar et al. "Sharing and Reuse of Sensitive Data and Samples: Supporting Researchers in Identifying Ethical and Legal Requirements". en. In: *Biopreservation and Biobanking* 13.4 (Aug. 2015), pp. 263–270.

[102] W. Paskal et al. "Aspects of Modern Biobank Activity – Comprehensive Review". en. In: *Pathology & Oncology Research* 24.4 (Oct. 2018), pp. 771–785.

[103] T. Kelly et al. "Global burden of obesity in 2005 and projections to 2030". en. In: *International Journal of Obesity* 32.9 (Sept. 2008), pp. 1431–1437.

[104] M. W. Schwartz et al. "Central nervous system control of food intake". eng. In: *Nature* 404.6778 (Apr. 2000), pp. 661–671.

[105] A. E. Locke et al. "Genetic studies of body mass index yield new insights for obesity biology". en. In: *Nature* 518.7538 (Feb. 2015), pp. 197–206.

[106] I. García-García et al. “Neuroanatomical differences in obesity: meta-analytic findings and their validation in an independent dataset”. en. In: *International Journal of Obesity* 43.5 (May 2019), pp. 943–951.

[107] Y. Taki et al. “Relationship Between Body Mass Index and Gray Matter Volume in 1,428 Healthy Individuals”. en. In: *Obesity* 16.1 (Jan. 2008), pp. 119–124.

[108] N. Pannacciulli et al. “Brain abnormalities in human obesity: A voxel-based morphometric study”. In: *NeuroImage* 31.4 (July 2006), pp. 1419–1425.

[109] C. A. Raji et al. “Brain structure and obesity”. en. In: *Human Brain Mapping* 31.3 (Mar. 2010), pp. 353–364.

[110] C. M. Weise et al. “The obese brain as a heritable phenotype: a combined morphometry and twin study”. en. In: *International Journal of Obesity* 41.3 (Mar. 2017), pp. 458–466.

[111] M. R. Caunca et al. “Measures of obesity are associated with MRI markers of brain aging: The Northern Manhattan Study”. en. In: *Neurology* 93.8 (Aug. 2019).

[112] N. Opel et al. “Brain structural abnormalities in obesity: relation to age, genetic risk, and common psychiatric disorders: Evidence through univariate and multivariate mega-analysis including 6420 participants from the ENIGMA MDD working group”. en. In: *Molecular Psychiatry* 26.9 (Sept. 2021), pp. 4839–4852.

[113] M. E. Shaw et al. “Body mass index is associated with cortical thinning with different patterns in mid- and late-life”. en. In: *International Journal of Obesity* 42.3 (Mar. 2018), pp. 455–461.

[114] L. Chen et al. “Genetic Insights into Obesity and Brain: Combine Mendelian Randomization Study and Gene Expression Analysis”. en. In: *Brain Sciences* 13.6 (May 2023), p. 892.

[115] A. Horstmann et al. “Obesity-Related Differences between Women and Men in Brain Structure and Goal-Directed Behavior”. In: *Frontiers in Human Neuroscience* 5 (2011).

[116] S. Kullmann et al. "Compromised white matter integrity in obesity". eng. In: *Obesity Reviews: An Official Journal of the International Association for the Study of Obesity* 16.4 (Apr. 2015), pp. 273–281.

[117] X. Pan et al. "Exploring the genetic correlation between obesity-related traits and regional brain volumes: Evidence from UK Biobank cohort". en. In: *NeuroImage: Clinical* 33 (2022), p. 102870.

[118] J. E. Curran et al. "Identification of Pleiotropic Genetic Effects on Obesity and Brain Anatomy". In: *Human Heredity* 75.2-4 (Sept. 2013), pp. 136–143.

[119] P. N. Timshel, J. J. Thompson, and T. H. Pers. "Genetic mapping of etiologic brain cell types for obesity". en. In: *eLife* 9 (Sept. 2020), e55851.

[120] S. M. Smith et al. "An expanded set of genome-wide association studies of brain imaging phenotypes in UK Biobank". en. In: *Nature Neuroscience* 24.5 (May 2021), pp. 737–745.

[121] R. S. Desikan et al. "An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest". en. In: *NeuroImage* 31.3 (July 2006), pp. 968–980.

[122] B. Fischl et al. "Whole Brain Segmentation". en. In: *Neuron* 33.3 (Jan. 2002), pp. 341–355.

[123] ReproGen Consortium et al. "An atlas of genetic correlations across human diseases and traits". en. In: *Nature Genetics* 47.11 (Nov. 2015), pp. 1236–1241.

[124] for the Alzheimer's Disease Neuroimaging Initiative et al. "Gray & white matter tissue contrast differentiates Mild Cognitive Impairment converters from non-converters". en. In: *Brain Imaging and Behavior* 9.2 (June 2015), pp. 141–148.

[125] E. R. Gamazon et al. "A gene-based association method for mapping traits using reference transcriptome data". en. In: *Nature Genetics* 47.9 (Sept. 2015), pp. 1091–1098.

[126] K. A. Fawcett and I. Barroso. "The genetics of obesity: FTO leads the way". en. In: *Trends in Genetics* 26.6 (June 2010), pp. 266–274.

[127] A. N. Barbeira et al. "Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics". en. In: *Nature Communications* 9.1 (May 2018), p. 1825.

[128] Y. E. Li et al. "A comparative atlas of single-cell chromatin accessibility in the human brain". en. In: *Science* 382.6667 (Oct. 2023), eadf7044.

[129] The ADIPOGen Consortium et al. "New genetic loci link adipose and insulin biology to body fat distribution". en. In: *Nature* 518.7538 (Feb. 2015), pp. 187–196.

[130] B. Zeng et al. "Genetic regulation of cell type–specific chromatin accessibility shapes brain disease etiology". en. In: *Science* 384.6698 (May 2024), eadh4265.

[131] S. G. Coetzee, G. A. Coetzee, and D. J. Hazelett. "*motifbreakR* : an R/Bioconductor package for predicting variant effects at transcription factor binding sites". en. In: *Bioinformatics* 31.23 (Dec. 2015), pp. 3847–3849.

[132] M. Karlsson et al. "A single–cell type transcriptomics map of human tissues". en. In: *Science Advances* 7.31 (July 2021), eabh2169.

[133] N. Liu et al. "TUFM in health and disease: exploring its multifaceted roles". In: *Frontiers in Immunology* 15 (May 2024), p. 1424385.

[134] Q. Le Grand et al. "Genomic Studies Across the Lifespan Point to Early Mechanisms Determining Subcortical Volumes". eng. In: *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging* 7.6 (June 2022), pp. 616–628.

[135] L. M. García-Marín et al. "Genomic analysis of intracranial and subcortical brain volumes yields polygenic scores accounting for variation across ancestries". eng. In: *Nature Genetics* 56.11 (Nov. 2024), pp. 2333–2344.

[136] G. D. Femminella et al. "Does Microglial Activation Influence Hippocampal Volume and Neuronal Function in Alzheimer's Disease and Parkinson's Disease Dementia?" In: *Journal of Alzheimer's Disease* 51.4 (Apr. 2016), pp. 1275–1289.

[137] D. J. Taylor et al. "Sources of gene expression variation in a globally diverse human cohort". eng. In: *Nature* 632.8023 (Aug. 2024), pp. 122–130.

[138] J. Zhang et al. "A Founder Mutation in VPS11 Causes an Autosomal Recessive Leukoencephalopathy Linked to Autophagic Defects". en. In: *PLOS Genetics* 12.4 (Apr. 2016), e1005848.

[139] J. Huang et al. "Genomics and phenomics of body mass index reveals a complex disease network". en. In: *Nature Communications* 13.1 (Dec. 2022), p. 7973.

[140] T. G. Richardson et al. "Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis". en. In: *PLOS Medicine* 17.3 (Mar. 2020). Ed. by D. J. Rader, e1003062.

[141] D. Van Der Meer et al. "Understanding the genetic determinants of the brain with MOSTest". en. In: *Nature Communications* 11.1 (July 2020), p. 3512.

[142] F. Koskeridis et al. "Pleiotropic genetic architecture and novel loci for C-reactive protein levels". en. In: *Nature Communications* 13.1 (Nov. 2022), p. 6939.

[143] Z. Zhu et al. "Shared genetic and experimental links between obesity-related traits and asthma subtypes in UK Biobank". en. In: *Journal of Allergy and Clinical Immunology* 145.2 (Feb. 2020), pp. 537–549.

[144] D. R. Kelley, J. Snoek, and J. L. Rinn. "Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks". eng. In: *Genome Research* 26.7 (July 2016), pp. 990–999.

[145] K. M. Chen et al. *A sequence-based global map of regulatory activity for deciphering human genetics*. en. July 2021.

[146] Y. Wu et al. "Joint analysis of GWAS and multi-omics QTL summary statistics reveals a large fraction of GWAS signals shared with molecular phenotypes". en. In: *Cell Genomics* 3.8 (Aug. 2023), p. 100344.

[147] A. A. Shabalin. "Matrix eQTL: ultra fast eQTL analysis via large matrix operations". en. In: *Bioinformatics* 28.10 (May 2012), pp. 1353–1358.

[148] 1000 Genomes Project Consortium et al. "A global reference for human genetic variation". eng. In: *Nature* 526.7571 (Oct. 2015), pp. 68–74.

[149] D. Zhou et al. "A unified framework for joint-tissue transcriptome-wide association and Mendelian randomization analysis". eng. In: *Nature Genetics* 52.11 (Nov. 2020), pp. 1239–1246.

[150] R. J. Pruim et al. "LocusZoom: regional visualization of genome-wide association scan results". en. In: *Bioinformatics* 26.18 (Sept. 2010), pp. 2336–2337.

[151] J. Carrasco-Zanini et al. "Proteomic signatures improve risk prediction for common and rare diseases". eng. In: *Nature Medicine* 30.9 (Sept. 2024), pp. 2489–2498.

[152] J. Woerner et al. *Large-scale evaluation of proteomic and polygenic risk scores reveals complementary contributions to incident disease prediction*. en. July 2025.

[153] S. Isaac et al. *Human Plasma Proteomics Links Modifiable Lifestyle Exposome to Disease Risk*. en. May 2025.

[154] K. Tsuo et al. *Proteomic prediction of disease largely reflects environmental risk exposure*. en. Aug. 2025.

[155] A. D. Bretherick et al. "Linking protein to phenotype with Mendelian Randomization detects 38 proteins with causal roles in human diseases and traits". en. In: *PLOS Genetics* 16.7 (July 2020). Ed. by G. Davey Smith, e1008785.

[156] A. D. Kjaergaard et al. *Biomarker de-Mendelization: principles, potentials and limitations of a strategy to improve biomarker prediction by reducing the component of variance explained by genotype*. en. Sept. 2018.

[157] L. Kachuri et al. "Genetically adjusted PSA levels for prostate cancer screening". eng. In: *Nature Medicine* 29.6 (June 2023), pp. 1412–1423.

[158] G. D. Smith and S. Ebrahim. "'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?" eng. In: *International Journal of Epidemiology* 32.1 (Feb. 2003), pp. 1–22.

[159] C. Moore et al. "The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial". eng. In: *Trials* 15 (Sept. 2014), p. 363.

[160] Y. Xu et al. "An atlas of genetic scores to predict multi-omic traits". eng. In: *Nature* 616.7955 (Apr. 2023), pp. 123–131.

[161] B. Jermy et al. "A unified framework for estimating country-specific cumulative incidence for 18 diseases stratified by polygenic risk". en. In: *Nature Communications* 15.1 (June 2024), p. 5007.

[162] R. Richter et al. "Increase of expression and activation of chemokine CCL15 in chronic renal failure". eng. In: *Biochemical and Biophysical Research Communications* 345.4 (July 2006), pp. 1504–1512.

[163] C. Bellenguez et al. "New insights into the genetic etiology of Alzheimer's disease and related dementias". en. In: *Nature Genetics* 54.4 (Apr. 2022), pp. 412–436.

[164] F. Privé, J. Arbel, and B. J. Vilhjálmsson. "LDpred2: better, faster, stronger". eng. In: *Bioinformatics (Oxford, England)* 36.22-23 (Apr. 2021), pp. 5424–5431.

[165] M. E. W. Logtenberg, F. A. Scheeren, and T. N. Schumacher. "The CD47-SIRPα Immune Checkpoint". eng. In: *Immunity* 52.5 (May 2020), pp. 742–752.

[166] M. M. Xie et al. "An agonistic anti-signal regulatory protein α antibody for chronic inflammatory diseases". eng. In: *Cell Reports. Medicine* 4.8 (Aug. 2023), p. 101130.

[167] G. H. Eldjarn et al. "Large-scale plasma proteomics comparisons through genetics and disease associations". eng. In: *Nature* 622.7982 (Oct. 2023), pp. 348–358.

[168] D. A. Gadd et al. "Blood protein assessment of leading incident diseases and mortality in the UK Biobank". en. In: *Nature Aging* 4.7 (July 2024), pp. 939–948.

[169] M. Pietzner et al. *Machine learning-guided deconvolution of plasma protein levels*. en. Jan. 2025.

[170] L. Wik et al. "Proximity Extension Assay in Combination with Next-Generation Sequencing for High-throughput Proteome-wide Analysis". eng. In: *Molecular & cellular proteomics: MCP* 20 (2021), p. 100168.

[171] Hindy, M. V. Vallelado, and Sep905. *yuenshingyan/MissForest: MissForest in Python - Arguably the best missing values imputation method*. Aug. 2024.

[172] C. C. Chang et al. "Second-generation PLINK: rising to the challenge of larger and richer datasets". en. In: *Gigascience* 4.1 (Dec. 2015), s13742–015–0047–8.

[173] M. Kuhn. "Building Predictive Models in *R* Using the **caret** Package". en. In: *Journal of Statistical Software* 28.5 (2008).

[174] J. Friedman, T. Hastie, and R. Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". en. In: *Journal of Statistical Software* 33.1 (2010).

[175] S. Durinck et al. "BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis". en. In: *Bioinformatics* 21.16 (Aug. 2005), pp. 3439–3440.

[176] A. R. Quinlan and I. M. Hall. "BEDTools: a flexible suite of utilities for comparing genomic features". en. In: *Bioinformatics* 26.6 (Mar. 2010), pp. 841–842.

[177] J. K. Tay, B. Narasimhan, and T. Hastie. "Elastic Net Regularization Paths for All Generalized Linear Models". en. In: *Journal of Statistical Software* 106.1 (2023).

[178] X. Robin et al. "pROC: an open-source package for R and S+ to analyze and compare ROC curves". en. In: *BMC Bioinformatics* 12.1 (Dec. 2011), p. 77.

[179] J. Yang et al. “Common SNPs explain a large proportion of the heritability for human height”. en. In: *Nature Genetics* 42.7 (July 2010), pp. 565–569.

[180] F. Privé et al. “Inferring disease architecture and predictive ability with LDpred2-auto”. eng. In: *American Journal of Human Genetics* 110.12 (Dec. 2023), pp. 2042–2055.

APPENDIX $\mathcal{A}$

Supplementary material to Chapter 2

A.1 Supplementary Figures

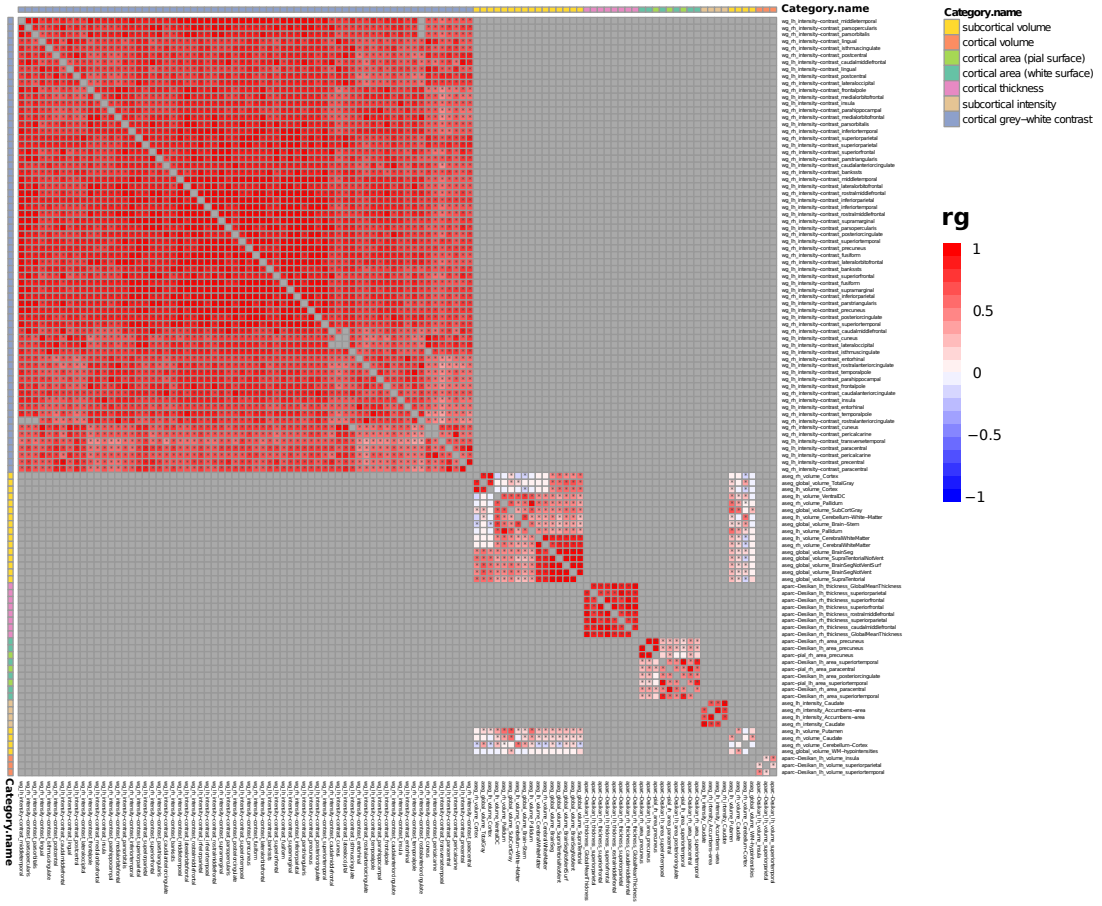


Figure A.1: **Genetic correlation between sMRI traits.** Only the correlations between traits in the same category were computed, plus those between cortical area (pial surface) and cortical area (white surface). *, nominally significant ($P < 0.05$).

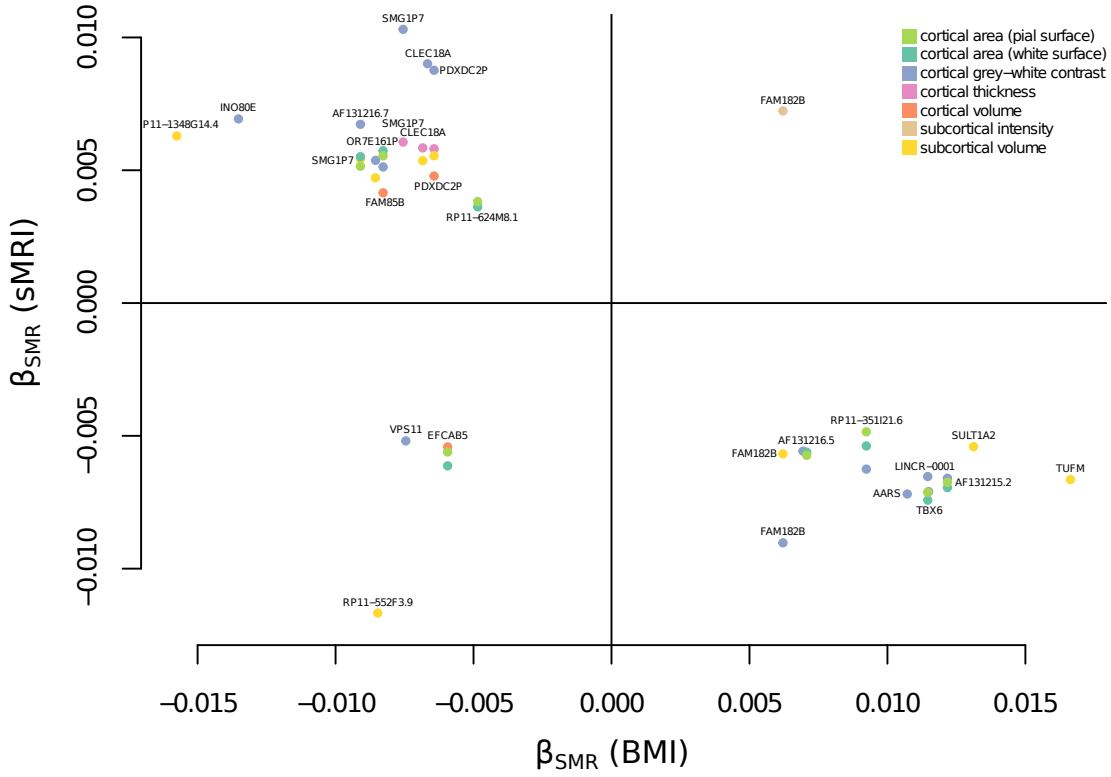


Figure A.2: **Effect size of GReX on BMI and sMRI traits for the 21 common genes.**

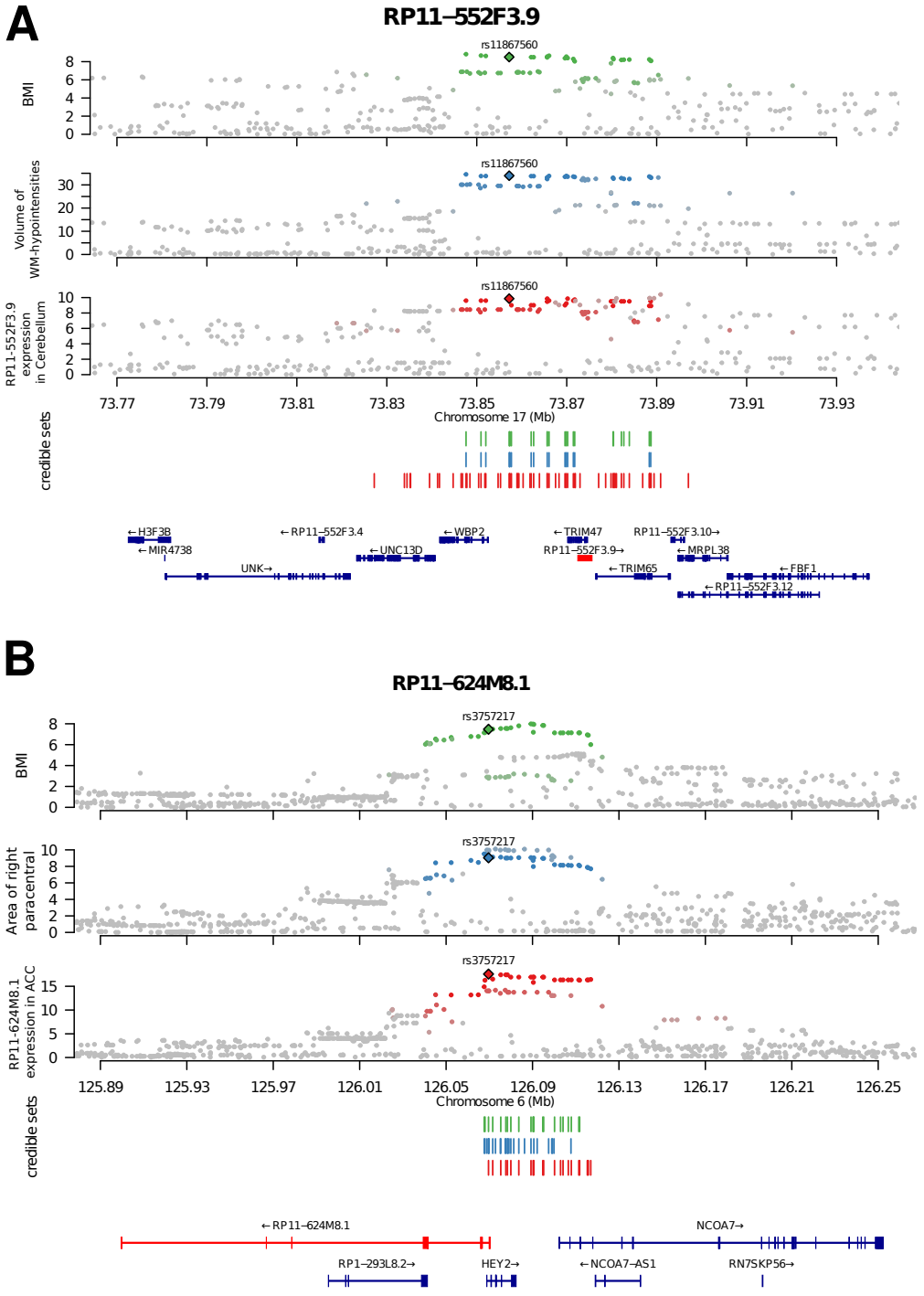


Figure A.3: **Manhattan plots as in 2.4 for the non-coding RNAs RP11-552F3.9 and RP11-624M8.1, respectively antisense transcripts of TRIM47 and HEY2.**

APPENDIX $\mathcal{B}$

Supplementary material to Chapter 3

B.1 Supplementary Figures

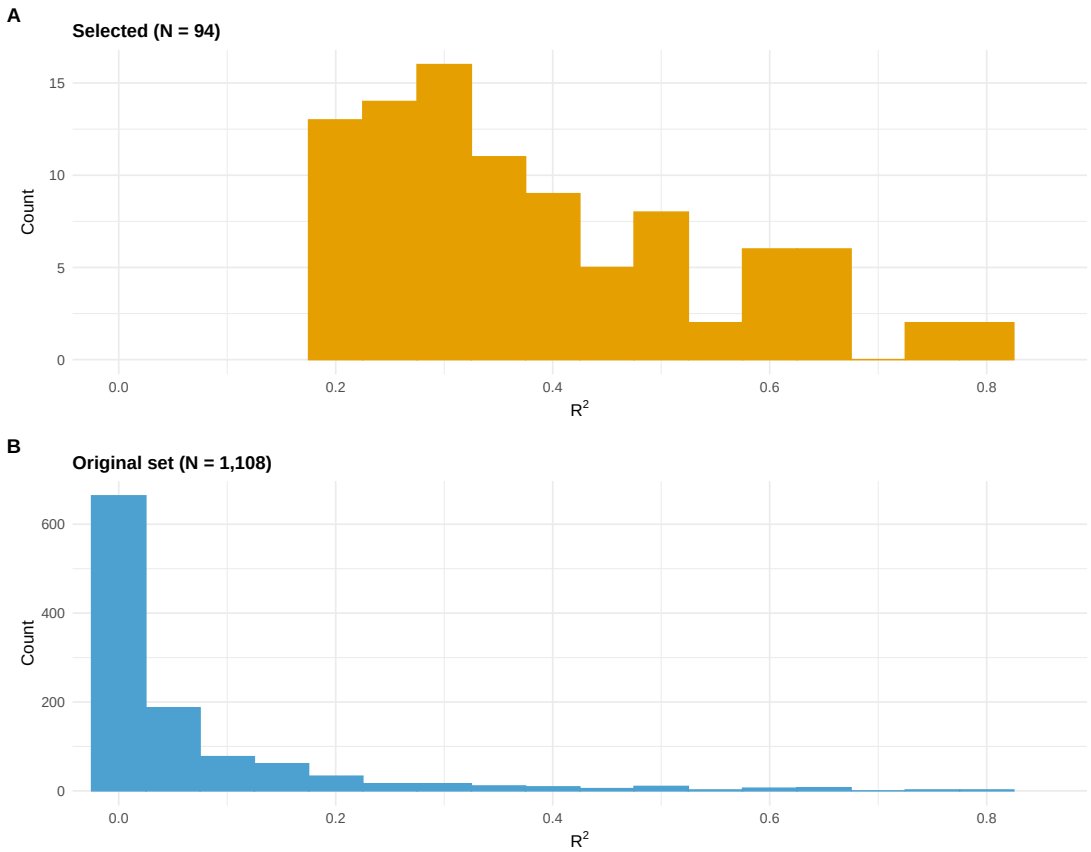


Figure B.1: **Protein variance explained by PGS (R^2).** Distributions of the R^2 between protein levels and PGS across the 94 selected proteins (**A**) and the original set of proteins before filtering (**B**).

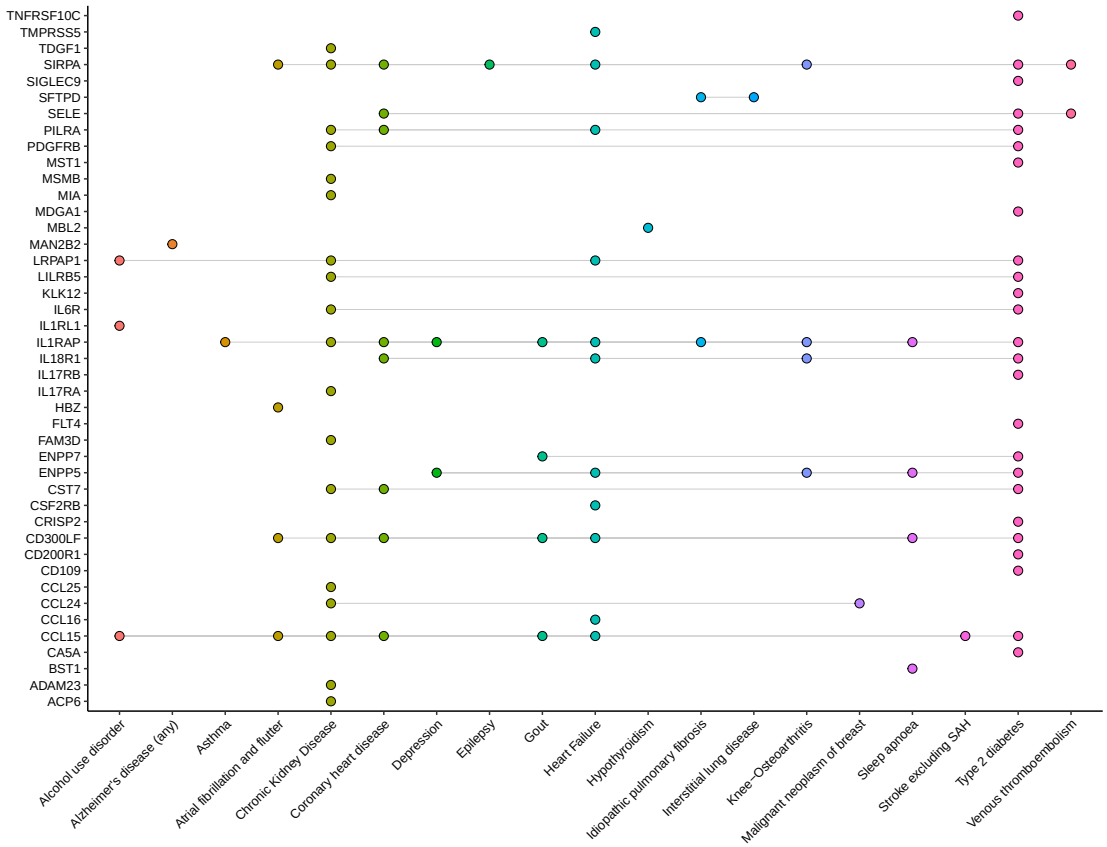


Figure B.2: **Upset plot of the top 95 protein–disease pairs showing stronger effects in the genetically adjusted versus unadjusted analysis.** Dots indicate significant associations (FDR < 0.05 in at least one model and p for the difference < 0.05) between a disease (x-axis) and a protein (y-axis). Gray lines denote proteins associated with multiple diseases, and dot colours represent diseases.

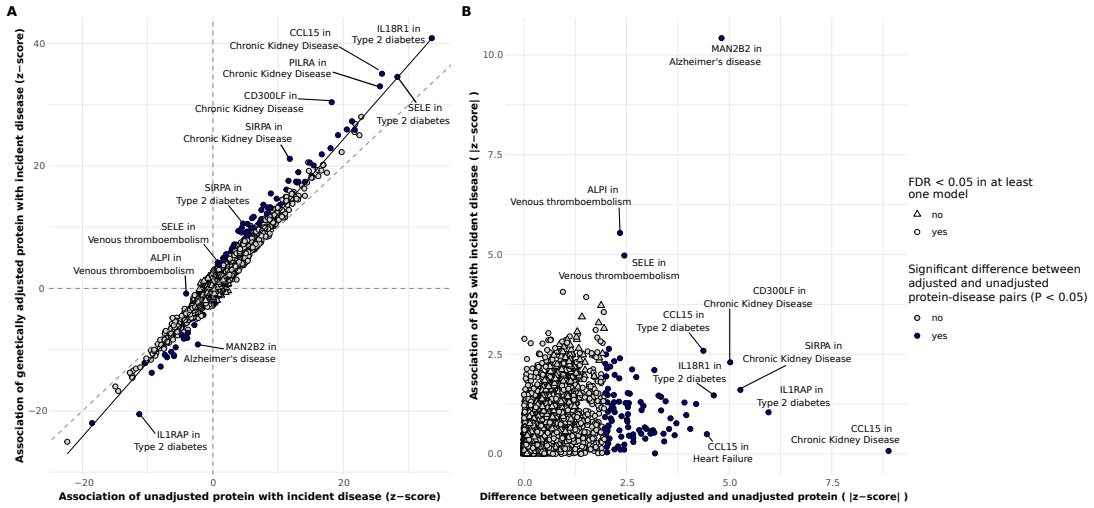


Figure B.3: **Associations between 94 proteins and 37 diseases in UK Biobank. A** Logistic regression z-scores for associations with incident diseases for unadjusted proteins (x-axis) and genetically adjusted proteins (y-axis). **B** Absolute differences in z-scores between genetically adjusted and unadjusted protein models (x-axis) and absolute z-scores for the association between protein PGS and diseases (y-axis). Blue indicates a significant difference in effect sizes between adjusted and unadjusted protein associations ($p < 0.05$); shape indicates FDR significance in at least one model (unadjusted or adjusted protein).

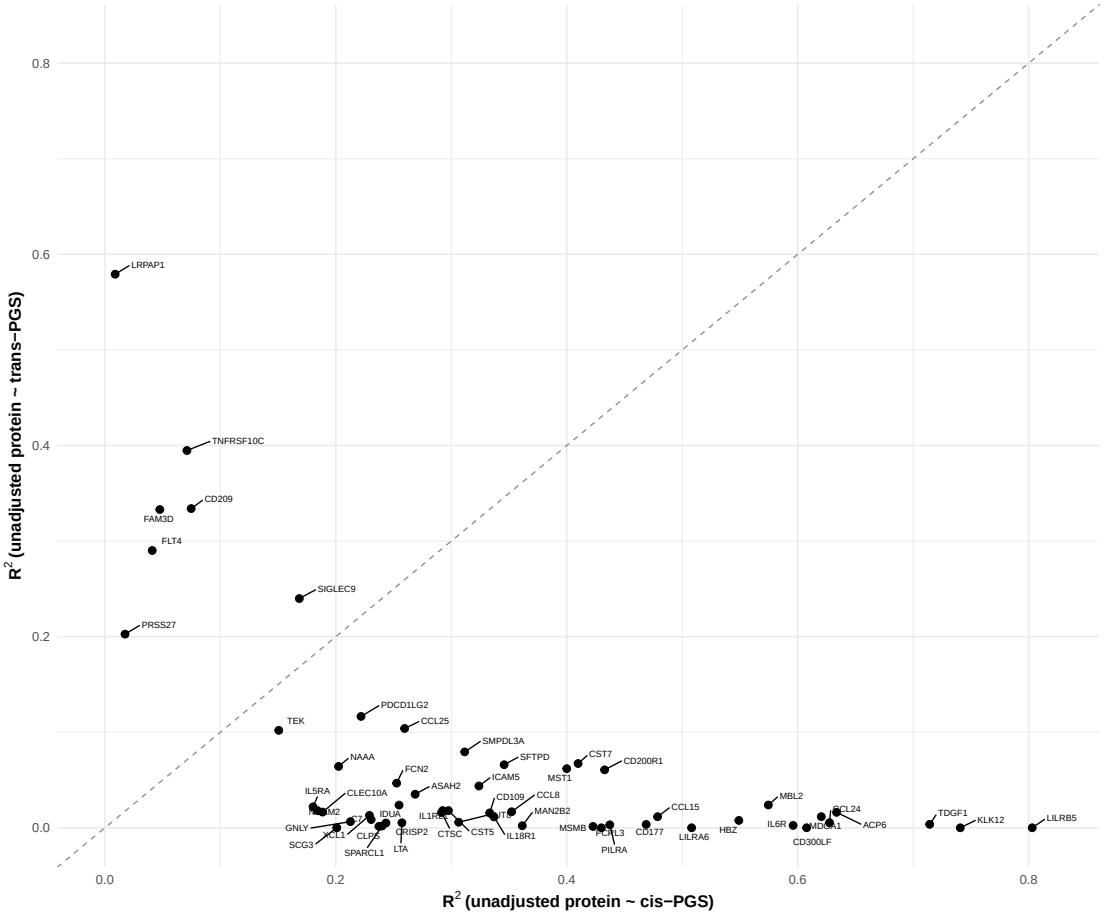


Figure B.4: **Comparison of protein variance explained by cis- and trans-PGS.** Protein variance explained (R^2) by cis-PGS (x-axis) and trans-PGS (y-axis). Each point represents a protein with both cis and trans variants derived from OmicsPred. Proteins with only cis- or only trans-variants are not shown.

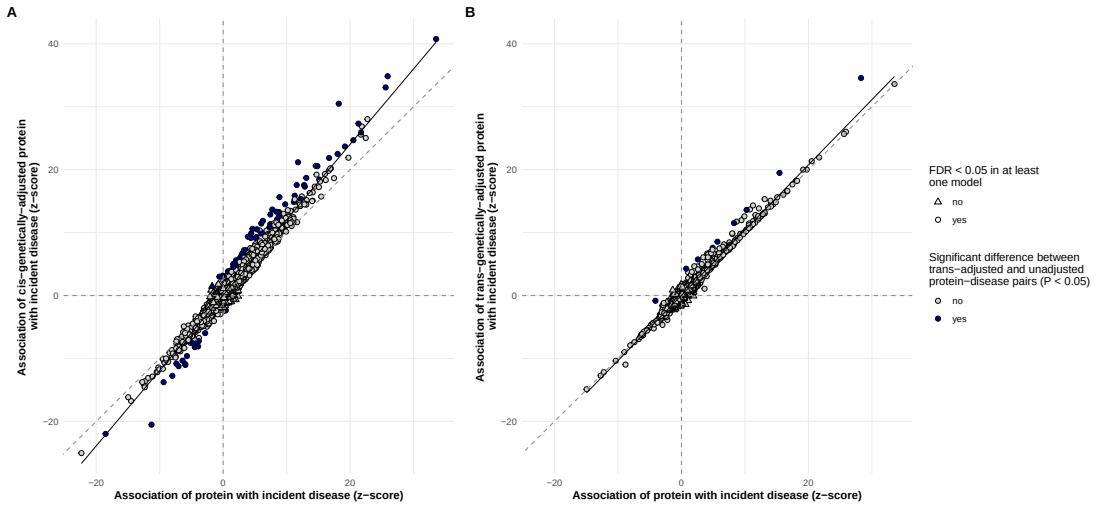


Figure B.5: **Results from cis- and trans-adjusted logistic regressions.** Logistic regression z-scores for associations of unadjusted proteins (x-axes) and cis-adjusted proteins (**A**, y-axis) or trans-adjusted proteins (**B**, y-axis) with incident diseases. Blue indicates a significant difference in effect sizes between adjusted and unadjusted protein associations ($p < 0.05$), and shape indicates FDR significance in at least one model (unadjusted or adjusted protein).

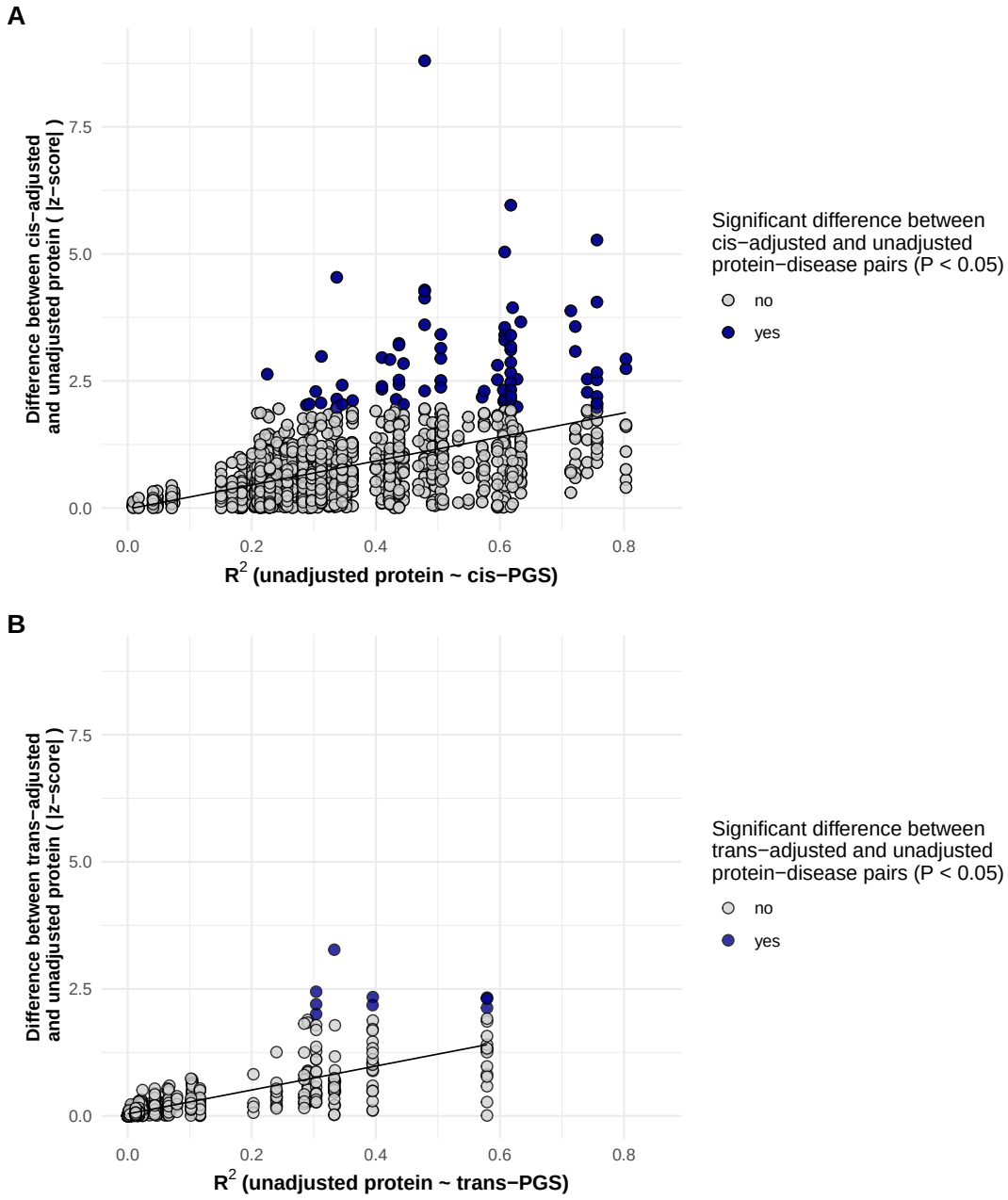


Figure B.6: **Association between PGS-explained protein variance and cis- or trans-adjusted disease associations.** Relationship between the protein variance explained (R^2) by the corresponding PGS and the absolute z-score difference between unadjusted protein associations and cis- (**A**) or trans- (**B**) adjusted protein associations with diseases. Only associations significant at FDR < 0.05 in at least one model (unadjusted or corresponding adjusted protein) are shown. Colour indicates a significant difference in effect sizes between adjusted and unadjusted protein associations ($p < 0.05$).

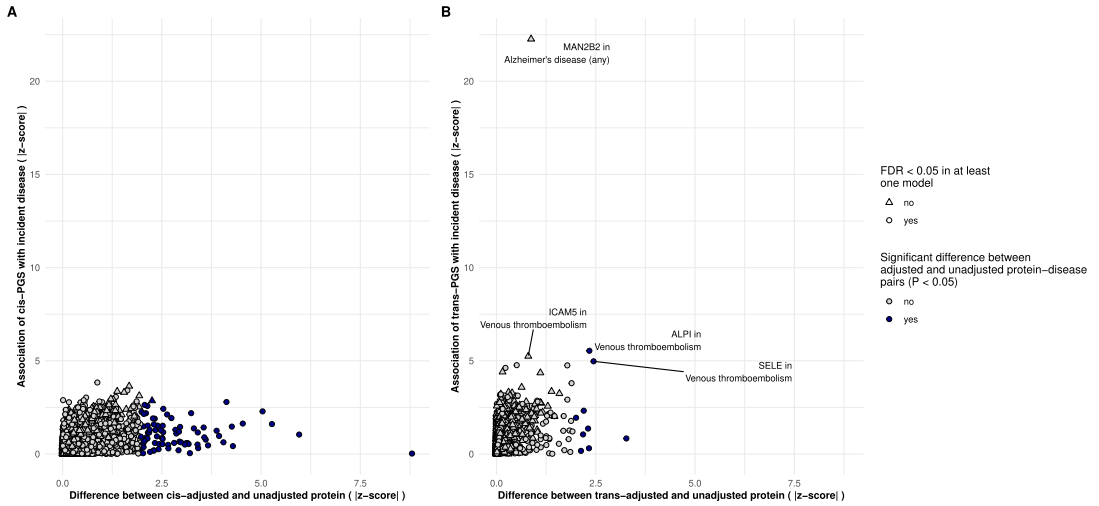


Figure B.7: **Associations between cis- and trans-PGS with diseases compared to the difference between respective adjusted and unadjusted models.** Absolute differences in z-scores between cis- (**A**) or trans- (**B**) genetically adjusted and unadjusted protein models (x-axes) and absolute z-scores for the association between protein PGS and diseases (y-axes). Blue indicates a significant difference in effect sizes between adjusted and unadjusted protein associations ($p < 0.05$); shape indicates FDR significance in at least one model (unadjusted or adjusted protein). Labels are shown only for PGS–disease associations with FDR < 0.05.

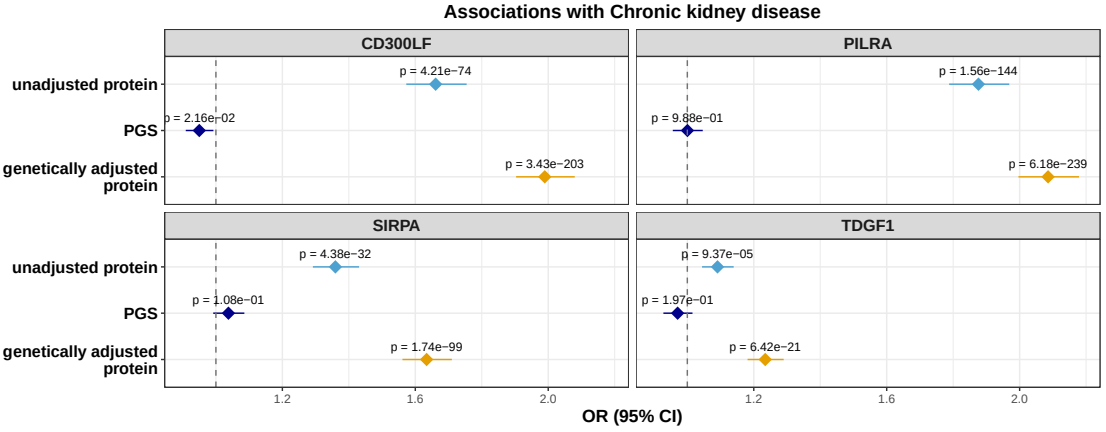


Figure B.8: **Top associations with chronic kidney disease.** Logistic regression odds ratios (ORs) with 95% confidence intervals are shown for associations of four proteins-CD300LF, PILRA, SIRPA, and TDGF1-with chronic kidney disease, using unadjusted protein levels, PGS, and genetically adjusted protein levels.

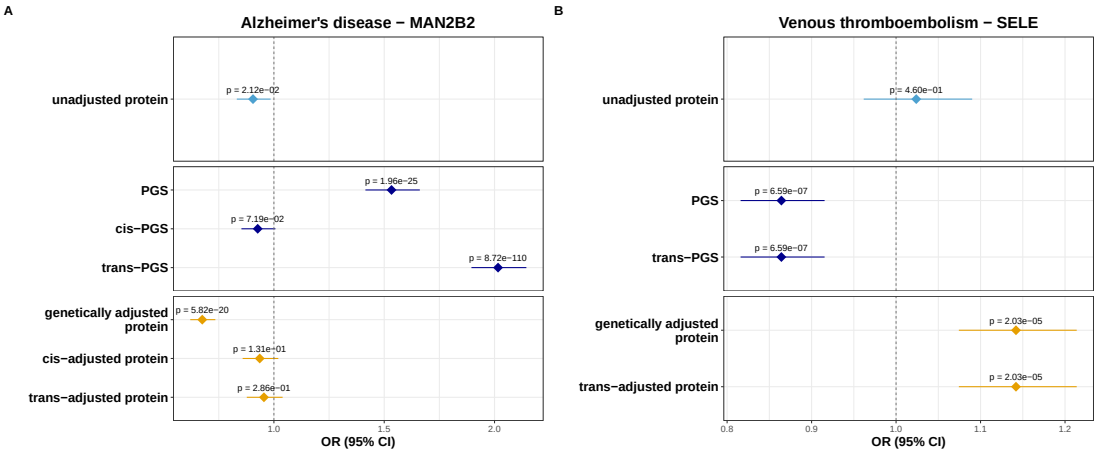


Figure B.9: **Associations of MAN2B2 and SELE with disease risk.** Logistic regression odds ratios (ORs) with 95% confidence intervals are shown for the association of MAN2B2 with Alzheimer's disease (**A**) and SELE with venous thromboembolism (**B**), using unadjusted protein, PGS, cis-PGS, trans-PGS, genetically adjusted protein, and cis- and trans-adjusted protein. SELE PGS was computed using only trans-variants; therefore, SELE cis-PGS is not available.

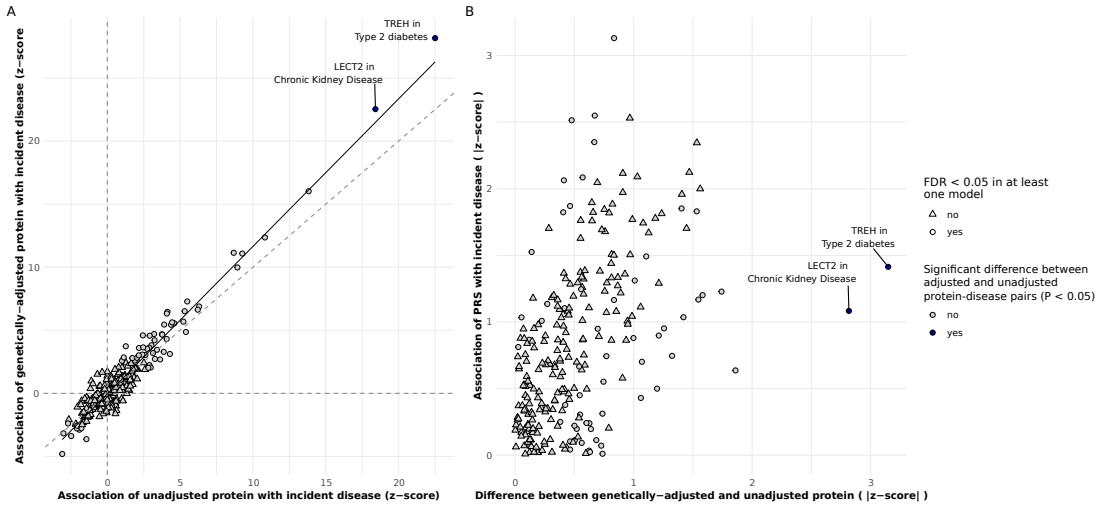


Figure B.10: **Associations between seven additional proteins and 37 diseases in UK Biobank, using PGS weights from FinnGen. A** Logistic regression z-scores for associations with incident diseases for unadjusted proteins (x-axis) and genetically adjusted proteins (y-axis). **B** Absolute differences in z-scores between genetically adjusted and unadjusted protein models (x-axis) and absolute z-scores for the association between protein PGS and diseases (y-axis). Blue indicates a significant difference in effect sizes between adjusted and unadjusted protein associations ($p < 0.05$), and shape indicates FDR significance in at least one model (unadjusted or adjusted protein).

B.2 Supplementary Methods

Supplementary methods 1: Technical note on how cis- and trans-protein adjustment affects associations

Here, we provide a technical explanation for why even a trans-PRS that is a weak predictor of the protein—but a strong predictor of the disease—can, when considered together with a cis-PRS that is not associated with the disease, lead to a genetically adjusted protein showing a strong and spurious association with the disease. This is what was observed in the case of the MAN2B2–Alzheimer's disease association.

Let a linear predictor of interest P (the protein) with polygenic determinant G

(the PGS) and non-genetic determinants ε_p (the residuals of the model):

$$P \;=\; \beta_{pG} G \;+\; \varepsilon_p \tag{B.1}$$

Where G and P are standardized ($\text{var}(G) = \text{var}(P) = 1$), ε_p and G are assumed independent ($\varepsilon_p \perp G$)

Suppose the polygenic determinant G can be divided into two independent genetic components g_c and g_T

$$G \;=\; \beta_{g_c} g_c \;+\; \beta_{g_T} g_T \tag{B.2}$$

Let g_A and g_B be homoscedastic ($\text{var}(g_c) = \text{var}(g_T) = 1$), independent between them ($g_c \perp g_T \rightarrow \beta_{g_c}^2 + \beta_{g_T}^2 = 1$), and independent to the non-genetic determinants ε_p ($g_c \perp \varepsilon_p$ and $g_T \perp \varepsilon_p$).

COMMENT: a protein's PGS can be decomposed into two parts: a pleiotropic (trans) component that also influences the disease and a protein-specific (cis) component that does not. By construction, these components should be approximately independent.

Let g_T be a causal genetic component (or in LD with the causal genetics) for disease D whereas g_c is not:

$$D \;=\; \beta_{D_T} g_T \;+\; \varepsilon_D \tag{B.3}$$

Where $\text{var}(D) = 1$, ε_D denotes any other genetic or non-genetic determinants for disease D. $\varepsilon_D \perp g_T$, $g_c \perp D$ (therefore also $\varepsilon_D \perp g_c$). In this case, a linear unbiased estimate for the association between D and g_T would be β_{D_T}, and between D and g_c would be 0.

Suppose now we have $\beta_{g_c} \gg \beta_{g_T} > 0$.

COMMENT: this corresponds to the association between disease and partial protein PGS separated by pleiotropy. For the MAN2B2 case, β_{D_T} seems to have large magnitude and is greater than 0.

An unbiased estimate for the association between D and G is:

$$\beta_{DG} \;=\; \frac{\text{cov}(D, G)}{\text{var}(G)} \;=\; \text{Cov}(\beta_{D_T} g_T + \varepsilon_D,\; \beta_{g_c} g_c + \beta_{g_T} g_T) \;=\; \beta_{D_T} \beta_{g_T} \tag{B.4}$$

COMMENT: this corresponds to the association between disease and full protein PGS.
In the MAN2B2 case, this also seems to be rather large, because of large β_{D_T}.

An unbiased estimate for the association between D and P is:

$$\begin{aligned}\beta_{DP} &= \frac{\text{cov}(D,P)}{\text{var}(P)} = \text{cov}(\beta_{D_T} g_T + \varepsilon_D,\ \beta_{pG}(\beta_{g_c} g_c + \beta_{g_T} g_T) + \varepsilon_p) \\ &= \beta_{D_T}\beta_{pG}\beta_{g_T} + \text{cov}(\varepsilon_D, \varepsilon_p) \end{aligned} \tag{B.5}$$

COMMENT: this corresponds to the association between disease and the unadjusted protein. In the MAN2B2 case, this seems to be very small, because in this case $\text{cov}(\varepsilon_D, \varepsilon_p) < 0$ *(see below), and* $\beta_{D_T}\beta_{pG}\beta_{g_T} > 0$.

If we remove genetic part (G) from P (genetic-adjusted protein denoted by P_{G_-}), an unbiased estimate for the association between D and P_{G_-} is:

$$\begin{aligned}\beta_{DP_{G_-}} &= \frac{\text{Cov}(D, P_{G_-})}{\text{Var}(P_{G_-})} = \frac{\text{cov}(\beta_{D_T} g_T + \varepsilon_D, \varepsilon_p)}{\text{var}(\varepsilon_p)} \\ &= \frac{\text{cov}(\varepsilon_D, \varepsilon_p)}{1 - \beta_{pG}^2} \end{aligned} \tag{B.6}$$

COMMENT: this corresponds to the association between disease and the total PGS-adjusted protein. In the MAN2B2 case, this also seems to be negative, indicating $\text{cov}(\varepsilon_D, \varepsilon_p) < 0$.

More specifically, if we remove only g_c from G (cis-adjusted protein denoted by P_{C_-}), an unbiased estimate for the association between D and P_{C_-} is:

$$\begin{aligned}\beta_{DP_{C_-}} &= \frac{\text{Cov}(D, P_{C_-})}{\text{Var}(P_{C_-})} = \frac{\text{cov}(\beta_{D_T} g_T + \varepsilon_D,\ \beta_{pG}\beta_{g_T} g_T + \varepsilon_p)}{\text{var}(\beta_{pG}\beta_{g_T} g_T + \varepsilon_p)} \\ &= \frac{\beta_{D_T}\beta_{pG}\beta_{g_T} + \text{cov}(\varepsilon_D, \varepsilon_p)}{1 - (\beta_{pG}\beta_{g_c})^2} \end{aligned} \tag{B.7}$$

COMMENT: this corresponds to the association between disease and the "cis"-adjusted protein. In the MAN2B2 case, this also seems to be very small. Note that the numerator in this form is the same as the disease association with the

unadjusted protein, which is almost 0 as observed above.

if we remove only g_B from P (cis-adjusted protein denoted by $P_{B\_}$), an unbiased estimate for the association between D and $P_{B\_}$ is:

$$\begin{aligned}\beta_{DP_{T\_}} &= \frac{\mathrm{Cov}(D, P_{T\_})}{\mathrm{Var}(P_{T\_})} \\ &= \frac{\mathrm{cov}(\beta_{D_T} g_T + \varepsilon_D,\ \beta_{pG}\beta_{g_c} g_c + \varepsilon_p)}{\mathrm{var}(\beta_{pG}\beta_{g_c} g_c + \varepsilon_p)} \\ &= \frac{\mathrm{cov}(\varepsilon_D, \varepsilon_p)}{\mathrm{var}(\beta_{pG}\beta_{g_c} g_c + \varepsilon_p)} = \frac{\mathrm{cov}(\varepsilon_D, \varepsilon_p)}{1 - (\beta_{pG}\beta_{g_T})^2}\end{aligned} \tag{B.8}$$

COMMENT: this corresponds to the association between disease and the "trans"-adjusted protein. In the MAN2B2 case, this also seems to be very small.

Note the numerator of $\beta_{DP_{T\_}}$ (disease association with the trans-adjusted protein) in (8) is the same as $\beta_{DP_{G\_}}$ (disease association with the PGS-adjusted protein) in (6) when the hypothesis in (3) holds. However, we see the association $\beta_{DP_{T\_}}$ smaller than $\beta_{DP_{G\_}}$ because their denominators $1 - (\beta_{pG}\beta_{g_T})^2 \gg 1 - \beta_{pG}^2$ (due to small β_{g_T}). Therefore:

$$\left|\frac{\mathrm{cov}(\varepsilon_D, \varepsilon_p)}{1 - \beta_{pG}^2}\right| > \left|\frac{\mathrm{cov}(\varepsilon_D, \varepsilon_p)}{1 - (\beta_{pG}\beta_{g_T})^2}\right| \ \rightarrow\ \left|\beta_{DP_{G\_}}\right| > \left|\beta_{DP_{T\_}}\right| \tag{B.9}$$

CONCLUSION Small-effect trans variants can be amplified during genetic adjustment when combined with strong cis effects in the PGS.

Supplementary Methods 2: Relationship between power gain and proportion of protein variance explained by genetics

Here, we provide a technical explanation of why, when protein genetics is not associated with the disease of interest, if genetics explains much of a protein's variability, then removing that genetic part from the protein boosts the statistical power to detect its association with disease.

Let a linear predictor of interest P (the protein) with polygenic determinant G (the PGS) and non-genetic determinants ε_p (the residuals of the model):

$$P \;=\; \beta_{pG}G \;+\; \varepsilon_p \tag{B.10}$$

Where G and P are standardized ($\text{var}(G) = \text{var}(P) = 1$), ε_p and G are assumed independent ($\varepsilon_p \perp G$). Since P depends on G with coefficient β_{pG}, the proportion of variance of P explained by G is β_{pG}^2 and the residual variance of P is then $\text{var}(\varepsilon_p) = 1 - \beta_{pG}^2$.

Let a target disease D with $\text{var}(D) = 1$, suppose the genetics of P is independent of D ($D \perp G$).

The unbiased estimate for the linear regression coefficient of D on P is:

$$\hat{\beta}_{DP} \;=\; \frac{\text{cov}(D,P)}{\text{var}(P)} \;=\; \frac{\text{cov}(D,\beta_{pG}G+\varepsilon_p)}{1} \;=\; \beta_{pG}\,\text{cov}(D,G)+\text{cov}(\varepsilon_p,D) \;=\; \text{cov}(\varepsilon_p,D) \tag{B.11}$$

By the definition of the variance of an ordinary least squares (OLS) estimate, it can be written as:

$$\begin{aligned}\text{var}\Big(\hat{\beta}_{DP}\Big) &= \frac{\text{var}\Big(D-\hat{\beta}_{DP}P\Big)}{(n-1)\text{var}(P)} = \frac{\text{var}\Big(D-\hat{\beta}_{DP}P\Big)}{n-1} \\ &= \frac{\text{var}(D)+\hat{\beta}_{DP}^2\,\text{var}(P)-2\hat{\beta}_{DP}\,\text{cov}(\varepsilon_p,D)}{n-1} = \frac{1+\hat{\beta}_{DP}^2-2\hat{\beta}_{DP}^2}{n-1} = \frac{1-\hat{\beta}_{DP}^2}{n-1}\end{aligned} \tag{B.12}$$

Therefore

$$z_{DP} = \frac{\hat{\beta}_{DP}}{SE_{\hat{\beta}_{DP}}} = \frac{\hat{\beta}_{DP}}{\sqrt{\frac{1-\hat{\beta}_{DP}^2}{n-1}}} \tag{B.13}$$

Now, let us remove genetics G from P hence evaluating only ε_p (in our case the residuals of the linear model), the unbiased estimate for the linear regression coefficient of P_{G_-} on D becomes:

$$\hat{\beta}_{DP_{G_-}} = \frac{\text{cov}(\varepsilon_p,D)}{\text{var}(\varepsilon_p)} = \frac{\text{cov}(\varepsilon_p,D)}{1-\rho_{PG}^2} = \frac{\hat{\beta}_{DP}}{1-\rho_{PG}^2} \tag{B.14}$$

With a variance of

$$\begin{aligned}\text{var}\Big(\hat{\beta}_{DP_{G\_}}\Big) &= \frac{\text{var}\Big(D - \hat{\beta}_{DP_{G\_}}\varepsilon_p\Big)}{(n-1)\text{var}(\varepsilon_p)} = \frac{\text{var}\Big(D - \frac{\hat{\beta}_{DP}}{1-\rho^2_{PG}}\varepsilon_p\Big)}{(n-1)(1-\rho^2_{PG})} \\ &= \frac{1-\hat{\beta}^2_{DP}-\rho^2_{PG}}{(n-1)(1-\rho^2_{PG})^2}\end{aligned} \tag{B.15}$$

and a z-score of:

$$z_{DP_{G\_}} = \frac{\hat{\beta}_{DP_{G\_}}}{SE_{\hat{\beta}_{DP_{G\_}}}} = \frac{\hat{\beta}_{DP}}{1-\rho^2_{PG}} \Big/ \sqrt{\frac{1-\hat{\beta}^2_{DP}-\rho^2_{PG}}{(n-1)(1-\rho^2_{PG})^2}} = \frac{\hat{\beta}_{DP}}{\sqrt{\frac{1-\hat{\beta}^2_{DP}-\rho^2_{PG}}{n-1}}} \tag{B.16}$$

Let us compare the z-score obtained with their ratio:

$$\frac{z_{DP_{G\_}}}{z_{DP}} = \frac{\frac{\hat{\beta}_{DP}}{\sqrt{\frac{1-\hat{\beta}^2_{DP}-\rho^2_{PG}}{n-1}}}}{\frac{\hat{\beta}_{DP}}{\sqrt{\frac{1-\hat{\beta}^2_{DP}}{n-1}}}} = \frac{\sqrt{1-\hat{\beta}^2_{DP}}}{\sqrt{1-\hat{\beta}^2_{DP}-\rho^2_{PG}}} \tag{B.17}$$

This means that the more variance the polygenic score explains in the protein ρ^2_{PG} i.e. the R^2, the larger the ratio of z-scores becomes when we regress the disease on the non-genetic component of protein (residualised on genetics) versus the unadjusted protein. In other words, when genetics of P is not associated with target disease D, the stronger the genetic contribution to P, the more power we gain by removing that genetic component when testing the association with disease.

Note that since by the definition of Z-score:

$$Z \;=\; \frac{\bar{X}-\mu}{\sigma} \tag{B.18}$$

Where $\bar{X}$ and μ are sample and population mean respectively and σ is the population standard deviation, a ratio of squared z-scores can also be interpreted as the reciprocal of required sample size (n) ratio for detecting a same effect. Therefore, based on #(8), to detect a same association effect, comparing to P, using $P_{G\_}$

may reduce the sample size requirement by:

$$1 - \left(\frac{z_{DP}}{z_{DP_{G\_}}} \right)^2 = 1 - \frac{1 - \hat{\beta}^2_{DP} - \rho^2_{PG}}{1 - \hat{\beta}^2_{DP}} = \frac{\rho^2_{PG}}{1 - \hat{\beta}^2_{DP}} \tag{B.19}$$

CONCLUSION The percentage reduction in required sample size is linearly related to the proportion of protein variance explained by genetics (R^2 or ρ^2_{PG})