Technische Universität München
Fakultät für Medizin

Convergence of genetic mechanisms:

Pathway identification and patient stratification from imputed gene expression in complex diseases

Lucia Trastulla

Vollständiger Abdruck der von der Fakultät für Medizin der Technischen Universität München zur Erlangung einer

Doktorin der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Thomas Korn

Prüfer*innen der Dissertation:

1. apl. Prof. Dr. Bertram Müller-Myhsok
2. Prof. Dr. Julien Gagneur

Die Dissertation wurde am 08.12.2022 bei der Technischen Universität München eingereicht und durch die Fakultät für Medizin am 21.03.2023 angenommen.

Abstract

Genome-wide association studies have so far identified thousands of genetic variants associated with complex diseases, however their interpretation remains challenging. Here, we developed a pipeline called CASTom-iGEx that builds tissue-specific gene expression models, converts genotype to imputed gene expression and pathway-score at the individual level and further stratifies patients driven by disease-related biological mechanisms. Applying CASTom-iGEx to coronary artery disease and schizophrenia, we discern patients' subgroups that exhibit different molecular and phenotypic manifestations.

Zusammenfassung

Genomweite Assoziationsstudien haben tausende von genetischen Varianten als Risikofaktoren für komplexe Erkrankungen identifiziert. Ihre biologische und klinische Interpretation ist jedoch eine große Herausforderung. Hier wird die CASTom-iGEx Pipeline vorgestellt, welche es erlaubt die Genexpression und Pathway-Aktivität basierend auf dem Genotyp vorherzusagen und diese zur Stratifikation zu nutzen. Die Anwendung von CASTom-iGEx auf die koronare Herzkrankheit und Schizophrenie erlaubt es, Untergruppen mit unterschiedlichen molekularen und klinischen Merkmalen zu identifizieren.

Acknowledgement

First and foremost I would like to thank my supervisors Prof. Dr. Bertram Müller-Myhsok and Prof. Dr. Julien Gagneur for their guidance, support and insightful discussions during my doctorate. I would like to express my gratitude to Prof. Dr. Michael Ziller for his mentorship and for allowing me to work on exciting projects under his leadership. In addition, I would like to thank Prof. Dr. Heribert Schunkert for the critical feedback on the project and for providing access to the German cohorts of the CARDIoGRAM consortium. I wish to express my gratitude also to the Psychiatric Genomic Consortium for curating and providing the genomic data on schizophrenia cohorts.

My deepest gratitude goes to the members of the Ziller Lab that made everyday fun and full of cookies. I particularly thank Christine for having taken me under her wing from day one, supporting me with any German-related issues, for the extremely useful (as much as necessary) lessons of molecular biology and all the nice dinners together. I would also like to thank Liesa, Laura and Vanessa for being the best computational biology (dark) side and precious friends. Additionally, I would like to express my deepest gratitude to Dr. Francesco Iorio and my new colleagues at the Iorio Lab for their support and extreme patience in this final process of my doctoral studies. I am very grateful to be working in such a collaborative and fun environment.

This PhD journey would have been impossible without the support and the love of so many passionate, inspiring and caring people I had met during these years. Among those, I would like to express my sincere gratitude to a great friend and mentor from day one, Ezgi. I do not think I would be the person I am proud to be today without meeting you. Thanks for making fun every night and spicy every meal. My appreciation also goes to Fabrizia for being a kind friend sharing with me all the goodies sent from home as well as an inspiring scientist, and Sylvain for being an incredible and patient collaborator, never mad at me for talking about projects while having dinner. I would also like to thank Tibor and Federica for being my rock during the pandemic years, providers of delicious meals and the most entertaining players of “Pandemic”. In particular, I would like to show my deepest gratitude to Tibor for cooking me the best soup ever when I had a bit too much headache.

This journey would have also been hopeless without the love of the persons that were not physically with me in Munich at the time and still manage to be my strength from afar. First and foremost, I would like to thank my parents and my brother for all the love and support provided during these years. I am also hugely thankful to Marica for

her unwavering support and belief in me. All of these would have been intolerable and meaningless without you, thank you for always being on my side, no matter what. Last but not least, the greatest gratitude with all the love I can give goes to Luca. It was a difficult path but you made it possible. Thanks for having always been, every day of every year, the shoulder I would actually cry on, and the person with the toughest job of all, enduring me at my worst, reminding me constantly of my value and that life does not have to be so difficult. Thank you for your patience and for making me feel loved and valued, simply for making this life beautiful. You are truly my sunshine.

Contents

Acronyms

BH Benjamini-Hochberg.

CAD Coronary Artery Disease.

CASTom-iGEx CAses STratification from imputed Gene Expression.

CMC Common Mind Consortium.

CRM Cluster Reliable Measure.

CV Cross-Validation.

DHS DNase I hypersensitive site.

DLPC Dorsolateral Prefrontal Cortex.

FDR False Discovery Rate.

gene-RS Gene risk-score.

GLM Generalized Linear Model.

GO Gene Ontology.

GRE gene regulatory element.

HWE Hardy-Weinberg Equilibrium.

LD Linkage Disequilibrium.

MAF minor allele frequency.

MHC major histocompatibility complex.

MLE Maximum Likelihood Estimation.

MR Mendelian Randomization.

MSE Mean Squared Error.

NMI Normalized Mutual Information.

PALAS PAthway-level association study.

PC principal component.

PGC Psychiatric Genomic Consortium.

PriLer prior learned elastic-net regression.

PRS Polygenic Risk Score.

SCZ Schizophrenia.

SNN Shared Nearest Neighbor.

SNP Single Nucleotide Polymorphism.

TSS Transcription Starting Site.

TWAS transcriptome-wide association study.

UKBB UK Biobank.

UMAP Uniform Manifold Approximation and Projection.

WMW Wilcoxon-Mann-Whitney.

1 Introduction

1.1 Motivation

Complex diseases represent a primary biomedical challenge in today's healthcare. Not confined to a single gene inheritance, the underlying genetic mechanisms and their interactions with environmental and lifestyle factors have not yet been fully elucidated. Genome-wide association studies (GWASs) represented a turning point in identifying genetic components associated with the etiology of complex diseases. From the first published GWAS in 2005 [1], the number of associated genetic variants, i.e. Single Nucleotide Polymorphism (SNP) and indels, has grown exponentially. This was aided by ever-increasing sample sizes that nowadays exceed million of participants [2]. GWASs are likely to remain a pillar in dissecting complex disease mechanisms due to cost-effectiveness of microarrays and the possibility to impute more than 150 million variants leveraging the latest whole-genome sequencing projects [3].
However, it remains a challenge to understand the mechanisms through which disease-associated SNPs lead to disease occurrence. Thus, in the post-GWAS era a wide range of methodologies has been developed to pinpoint functional genes and molecular mechanisms to reveal effective drug targets [4, 5]. Difficulties arise, in part, from the location of associated SNPs, the vast majority residing in non-coding regions of the genome ($\sim 90\%$ [6]). This hampers the identification of the perturbed genes modulated by those variants. Moreover, both large and small effect variants contribute to the heritability of a trait [7], i.e. variability of a phenotype that can be explained by genetic variation in a certain population. Importantly, the small effect class of variants has grown in number concurrently with increase in sample sizes and necessitates a thorough understanding. In addition, complex diseases are highly polygenic [5], with each individual carrying a unique combination of alleles conferring a certain disease risk. This is in accordance with the heterogeneity of complex diseases that are usually defined through multiple criteria and are characterized by co-morbidity that complicates treatment effectiveness [8].

Hence, there is a critical gap between the information that arises from GWAS in terms of associated variants and the implementation of precision medicine strategies guided by the deconvolution of these genetic findings.

1.2 Objectives

Here we hypothesize that genetic variants associated with a complex disease converge into disrupted biological processes. Moreover, we hypothesize that individual genetic liability profiles converge into altered functional mechanisms and give rise to patient-specific phenotypic manifestations. To show the aforementioned points, we focus on the following critical aspects.

1. Identify genes perturbed by genetic changes associated with complex diseases. In particular, this would improve the understanding of functional consequences of associated variants, focusing on their regulatory cis-effects on adjacent genes.
2. Detect biological pathways characteristic of complex diseases. Starting from the hypothesis that associated variants perturb genes that in turn converge into meaningful biological pathways, the aim is to identify biological mechanisms that are disrupted by the aggregated effect of disease-related SNPs.
3. Determine which clinically relevant features are causal or protective for a complex disease mediated via the impairment of genes and pathways.
4. Understand whether the heterogeneous genetic configuration converges into different disease manifestations and symptoms. The aim here is to stratify individuals affected by a certain complex disease based on their genetic background and to detect existing differences in disease severity and treatment response.

1.3 Contribution

To address the previously outlined points, we developed a novel framework called CAses STratification from imputed Gene Expression (CASTom-iGEx) in a unified pipeline that can be applied to any complex disease. In particular, we first developed a novel method to convert variant information into tissue-specific gene expression derived solely from cis-effects, i.e. variants in the proximity of a gene transcription starting site. This approach is an extension of Transcriptome-wide association study (TWAS) methodologies developed so far [9, 10]. TWAS methodologies build gene expression prediction models from cis-effects leveraging data sets with genetic and gene expression measured on the same set of individuals. The newly developed method described here and called prior learned elastic-net regression (PriLer) incorporates prior information on variants, conferring a higher relevance to SNPs located in gene regulatory regions (e.g. open chromatin regions and enhancers). Afterwards, CASTom-iGEx imputes gene expression on large genotype-only data sets and identifies genes whose cis-variants modulated component is associated with the disease of interest. Moreover, we moved a step further and addressed the second objective aggregating the imputed gene expression into individual-level pathway

scores. These pathway scores are now interpretable measures and can be associated with the disease of interest similarly to GWAS or TWAS. We briefly refer to this type of analysis as PAthway-level association study (PALAS). Importantly, this framework offers the unique possibility to investigate whether small effect variants converge onto specific molecular mechanisms, omitting any p-value thresholding strategy. Using summary statistics for associations at the level of genes and pathways, we can then investigate whether endophenotypes related to a disease contribute to or are preventive of the disease etiology (and vice versa) via a Mendelian Randomization approach [11]. In particular, we leverage genes and pathways as instrumental variables and hence point to the candidate mechanisms mediating the causal relationship. With this strategy, it is possible to address the third objective and to effectively empower an endophenotypic deconstruction of complex disease to obtain insights into their specific biological basis. Finally, we used the imputed gene expression and developed a stratification approach for affected individuals which captures different genetic liabilities representing a specific configuration of genes. Once the affected individuals have been partitioned, our framework addresses the fourth objective by detecting any association of the clustering structure with clinical-related features, general endophenotypes, or treatment responses. Since a large genetic data set rarely also includes additional endophenotypic information, we developed a strategy to approximate phenotypes via gene risk-scores and still detect differences in groups of patients. The reliability of the group-specific results depends on the heritability of the considered phenotype as well as the strength of the associations.
Jointly, our newly developed pipeline represents another step towards the understanding of complex disease mechanisms and the implementation of precision medicine strategies.

In this thesis, I discuss the results of our pipeline applied to two complex diseases, coronary artery disease and schizophrenia. CASTom-iGEx recovers known genes and biological mechanisms, points to possible new candidates and molecular pathways that arise from aggregated effects, and identifies the genetic subgroup of patients that are associated with different clinical aspects and disease severity.

The pipeline is freely available at gitlab.mpcdf.mpg.de/luciat/castom-igex. Part of the analyses shown here is also presented in a future publication under review (from now on referred to as "Trastulla et al., in prep."). The usage or rearrangement of any figure from this publication is specified in this thesis at the beginning of the caption's figure.

1.4 Thesis structure

The remaining content of this thesis is organized as follows.
Chapter 2 provides a general overview of the GWAS application for complex diseases, giving insights on the developed strategies for post-GWAS analysis in terms of identification of target genes, pathway-based strategies, and patient stratification approaches. In addition, section 2.6 reports the major findings from the aforementioned strategies for coronary artery disease and schizophrenia.
Chapter 3 explains the developed methodology grouped into a unique pipeline, CASTom-iGEx, and is divided into three major sections. Section 3.1 describes our new method for imputed gene expression *PriLer*, providing details on the mathematical formulation and implementation. Section 3.2 focuses on *TWAS and PALAS* analyses and explains the procedure of testing imputed gene expression and individual-level pathway scores against a trait. Section 3.3 involves *Patient stratification* and informs on clustering strategy applied to stratify affected individuals in an unsupervised manner. In addition, it describes the approach to investigate possible differences in endophenotypes, both in the situation of available endophenotypes and non-available ones which are approximated via risk scores.
Chapter 4 contains the results of the developed method. After an initial explanation of data inclusion and pre-processing in section 4.1, an investigation of the predictive power of PriLer and a comparison with existing methodologies follows in section 4.2. The remaining sections include the results of CASTOm-iGEx applied to coronary artery disease (section 4.3) and schizophrenia (section 4.4). We discuss detected genes and pathways comparing our results to GWAS output, identify putative causal traits mediated by those genes and pathways, and characterize the genetically stratified groups of patients. At the same time, we perform multiple benchmarks via label randomization for disease status and clustering, together with ad-hoc comparison strategies.
Finally, **Chapter 5** discusses major findings and possible further developments.

2 State of the Art

2.1 Genome-wide association studies in complex diseases

Complex diseases are a result of a mixture of genetic, environmental, and lifestyle factors [12]. Different from Mendelian diseases caused by mutations in a single coding gene, complex diseases are associated with a combination of common and rare genetic variants (SNPs and indels). Taken singularly, the effect of any such variant can be marginal. However, in aggregate they can still contribute to the disease manifestation [5, 13]. Cardiovascular, neurodegenerative, psychiatric, and autoimmune diseases are part of this broad class of diseases and they are the leading cause of death and an important economic burden in today's healthcare. Genome-wide association studies (GWASs) represent a breakthrough in elucidating complex disease mechanisms. Indeed, from the first published GWAS in 2005 [1], an ever-increasing number of trait-associated variants have been detected, with over $128,000$ associations grouped in $55,000$ unique loci detected for almost $5,000$ heritable diseases and traits on $4,500$ GWASs [5, 14] (Fig. 2.1). The latest GWASs have even reached sample sizes exceeding a million of participants [2, 15].

At its core, GWAS technology is an experimental framework to discover associations between a trait and genetic variants in individuals from a single or multiple populations, by testing differences in allele frequency (workflow summarized in Fig. 2.2 as described by Uffelmann et al. [16]).
In details, DNA is gathered for a selected group of individuals together with phenotypic information such as disease status and possible confounders e.g. age, sex, and demographic information (Fig. 2.2**a**). Usually, large cohort studies lead by international consortia focus on a single disease. For instance, the Psychiatric Genomic Consortium (PGC) studies 11 psychiatric disorders, among which schizophrenia, bipolar, and major depressive disorders. Instead, an example of a large biobank with deep phenotypic data is the UK Biobank [17] cohort, a prospective genetic study not focusing on a specific disease or trait but systematically collecting a wide range of phenotypic, health-related information and biological measurement. The selected individuals are genotyped via SNP microarrays focusing on common variants, i.e. minor allele frequency (MAF) > 0.01, or next-generation sequencing methods for whole-genome sequencing or whole-exome sequencing that can in principle capture rare variant information (MAF ≤ 0.01) (Fig. 2.2**b**). Afterward,

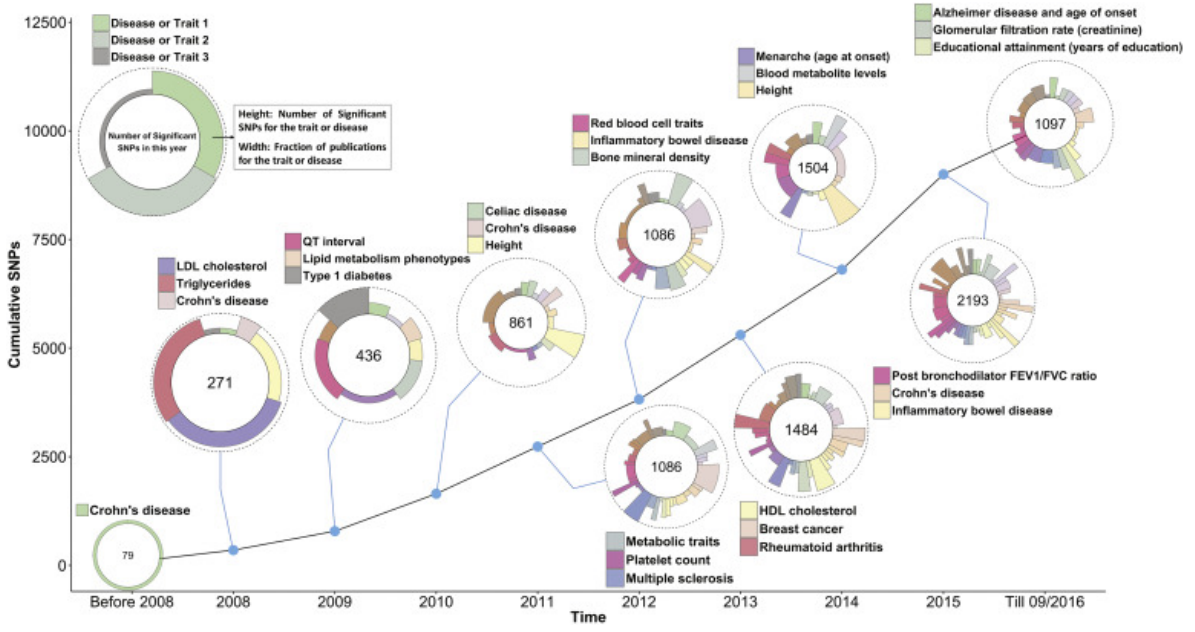

Fig. 2.1.: Figure from [5]. Data used for generating the graph were taken from the GWAS Catalogue[18]. SNPs and traits were selected according to the following filters. SNPs were selected with a p-value $< 5 \cdot 10^{-8}$. For each trait with two or more selected SNPs, SNPs were removed if they had an LD $r^2 > 0.5$ (calculated from 1000 Genomes phase 3 data) with another selected SNPs and their p value was larger. For each year of discovery, only the top three traits and diseases with the largest number of SNPs are labeled in the circle.

quality control is performed, including steps such as genotype calling, DNA switches, removal of not properly called SNPs and poor quality individuals and computation of principal components (PCs) to detect population stratification based on ancestry (Fig.2.2 **c**). Because SNP arrays only tag a subset of common variants (e.g. Affymetrix Human SNP 5.0 GeneChip genotype over $500,000$ human SNPs), genotype data can be phased and not tagged genotypes can be imputed leveraging the information from reference population (matched by ancestry) such as from 1000 Genomes Project repository, hence allowing the coverage of millions of SNPs (Fig. 2.2**d**). Afterward, an association test of each genetic variant with the phenotype of interest is run, correcting for confounders such as population structure captured by PCs, age, and sex (Fig. 2.2**e**). Nowadays, a variety of methods have been proposed for this specific step, from linear models in PLINK/PLINK2 [19] to mixed models such as BOLT-LMM [20] and fastGWA [21]. In general, it is fundamental that the study is properly conceived and that cases and controls are matched by ancestry to avoid confounding. In case multiple cohorts (usually of relatively small size) are available, results are combined to obtain overall summary statistics (variant specific p-value, odds ratio, standard error) via meta-analysis, for instance using GWAMA software [22] (Fig. 2.2**f**). Afterward, the reliability of the observed associations must be explored via internal replication (e.g. one cohort is left aside iteratively) or external replication (independent cohort). In the latter situation, it is crucial that the independent cohort is ancestrally matched to the one considered and there are no shared individuals of family members (Fig. 2.2**g**). Finally, the results from GWAS association are investigated via post-GWAS

techniques (Fig. 2.2**h**) such as silico fine-mapping, SNP enrichment, SNP to gene mapping, gene to function, pathway analysis, genetic correlation analysis, Mendelian randomization and polygenic risk prediction (examples of developed approaches discussed below).

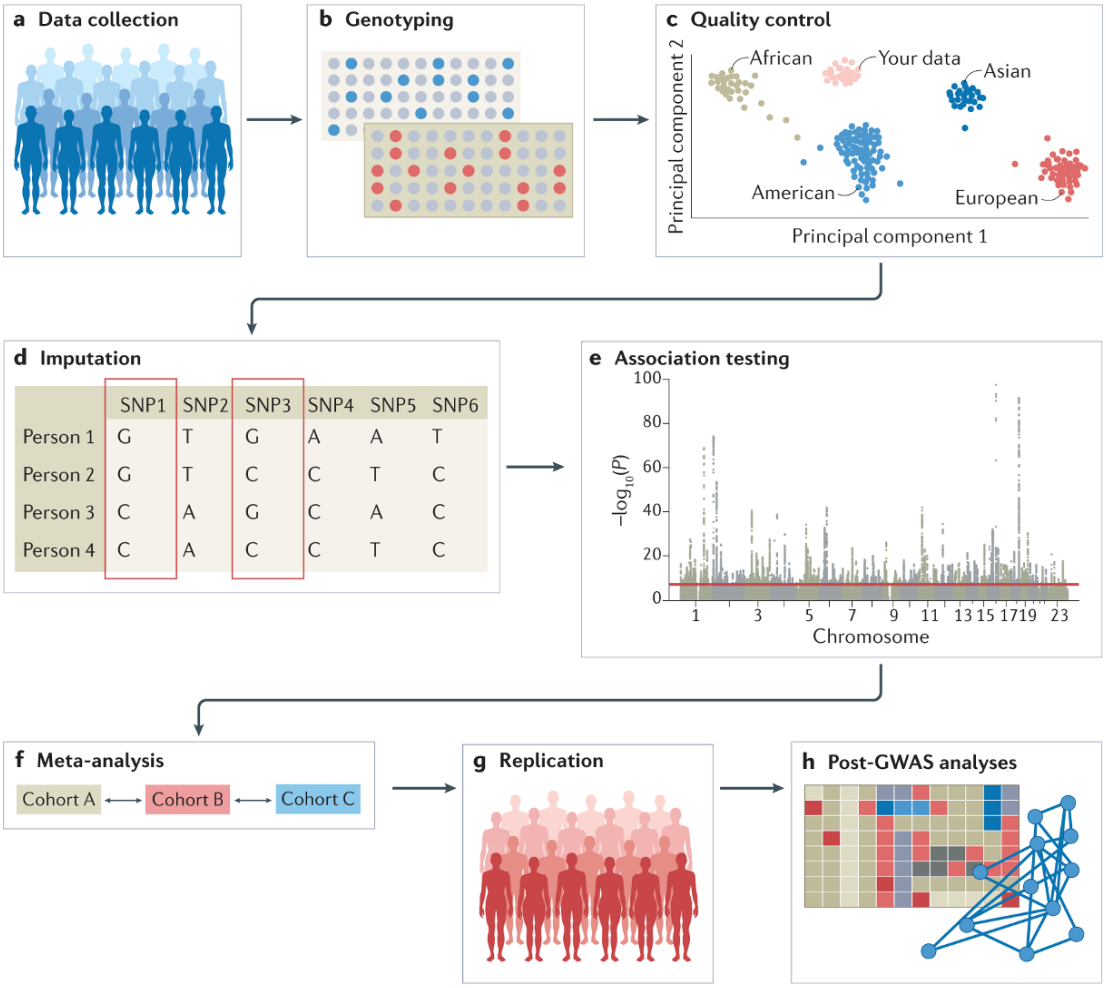

Fig. 2.2.: Figure from [16]. **a** Collection of DNA and phenotypic information from a group of individuals, **b** genotyping of each individual using available GWAS arrays or sequencing strategies, **c** quality control (figure depicts clustering of individuals according to genetic substrata), **d** imputation of untyped variants using haplotype phasing and reference populations, **e** conducting the statistical test for association (typical visual inspection from Manhattan plot in figure), **f** conducting a meta-analysis (optional), **g** seeking an independent replication, **h** interpreting the results via multiple post-GWAS analyses.

Indeed, translating GWAS findings is not a straightforward process [5]. The typical output of a GWAS is a list of p-values and odds ratio or coefficients (binomial and continuous trait respectively), that needs to be properly interpreted in order to understand the most likely causal variants, their consequences in gene products and gene regulation, and the convergence to biological pathways.

First of all, the majority of associated variants are located in non-coding regions, with an over-representation in regulatory elements such as enhancers and promoters [23]. Disease-associated variants disrupt binding sites of transcription factors, alter chromatin states,

and hence perturb regulatory networks [24, 25]. Nevertheless, in which pathological cell types they act is not easily inferred. Due to an incomplete picture of the regulome, a clear dynamic of this gene regulation is still missing, together with the identification of genes that are the direct target of those variants.

In addition, GWAS results are reported in terms of risk loci (i.e. sets of correlated variants with statistically significant associations) rather than single variants. The reason lies within Linkage Disequilibrium (LD) phenomenon leading to neighboring genetic variants being often inherited together because of the co-segregation effect happening at meiotic recombination [26]. LD effect implies a correlation of allelic status for neighboring SNPs. Consequentially, multiple variants genomically close can result in being disease-associated purely due to this phenomenon but not necessarily causal in the disease etiology. It is important to stress that LD between genetic variants, measured as a squared correlation r^2, is also used during statistical imputation of ungenotyped variants to recover lost information via imperfect LD between tagged genotypes and unobserved causal variants. This is predicted through the haplotypes inferred from multiple observed SNPs and those observed from a full sequence reference panel such as 1000 Genome or International HapMap Project [27], hence recovering allele information for millions instead of hundred thousands of variants. The application of GWAS technology to complex traits revealed their association with thousands of variants [14] grouped into hundreds of loci, nevertheless only marginally contributing to the disease risk. This is in accordance with "common disease common variant" hypothesis, stating that common diseases are affected by multiple genetic changes common in the population (Fig. 2.3). On the one hand, common variants cannot have a high effect size in disease association typical of rare disorders. Indeed, if they would, it would result in a complete correlation between allele frequency and population, contrasting with the rareness of this scenario. On the other hand, because common alleles can only have a small effect, it is the combination of multiple ones that influence the disease risk for those disorders having a certain heritability (i.e. the fraction of phenotype variability that can be attributed to genetic variation) [28]. Nevertheless, the heritability of complex diseases has not been yet entirely explained ("missing heritability" problem [13]) and the attention has been pointed towards the contribution from rare variants and their interplay with common associated variants. For instance, it was shown that low-frequency SNPs have moderate-to-large effects in complex traits such as height [29] and that disease-risk genes harbor both common and rare risk variants. Generally, the genetic architecture representing the variability in a disease across individuals of complex traits is regarded as polygenic: each associated variant only gives a small contribution to the overall risk of an individual. Consequently, an individual carries certain alleles that increase and others that decrease the disease risk in a unique combination [5].

To discern functional consequences of variants identified from GWASs, a broad range of analyses and methods have been developed as **post-GWAS** strategies. As previously mentioned, part of the identified variants from GWASs are likely non-causal, with a significant association arising from LD structure and correlating with a causal variant without contributing to the disease etiology [26]. To pinpoint likely causal variants in each

detected loci, fine-mapping is a post-GWAS analysis that prioritizes a subset among the associated variants based on association strength and LD information [30].

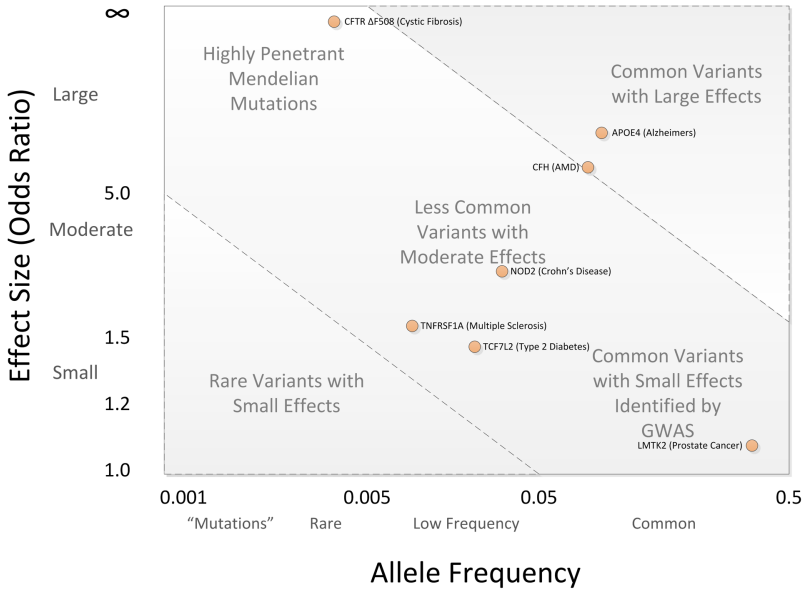

Fig. 2.3.: Figure from [28]. Highly penetrant alleles for Mendelian disorders are extremely rare with large effect sizes (upper left), while most GWAS findings are associations of common SNPs with small effect sizes (lower right). The bulk of discovered genetic associations lie on the diagonal denoted by the dashed lines.

An elementary strategy in this context is to perform a conditional association analysis for each locus, adjusting the local association results via step-wise inclusion of the most significant variant as a covariate in the trait-genotype regression [27]. More sophisticated approaches are based on Bayesian models and optimize the selection of variables for regression model via prior distributions that include imputation accuracy and association strength to estimate the posterior probability of a variant of being casual (e.g. CAVIAR [31]). In addition, fine-mapping strategies based on summary statistics have been developed to overcome the necessity of individual genotypes, which are more difficult to retrieve than the general summary of a GWAS and usually available to the scientific community. Examples of the latter approach are GCTA-COJO [32] built on conditional analysis, FINEMAP [33] based on Bayesian approaches and SuSIE [34] which is a mixed version of the two strategies. Nevertheless, the assumptions of these developed methods hinder the consistency of the results, particularly in the event of multiple independent associations in a locus. In addition, the statistical power to detect a set of putative causal variants is inversely proportional to the number of independent genetic variants [34]. Fine-mapping aids in understanding the most likely causal set of variants but does not point out their functional consequences. Subsequent analyses include the identification of gene(s) that mediate variant effects in a locus on the disease, the relevant tissues or cell types affected by those genetic changes, and the effects on biological pathways and networks that lead to dis-regulation in physiological functions [16]. Examples of developed strategies to tackle these points will be explained in sections 2.2 and 2.3.

In light of the avalanche of segregating genetic variants associated with multiple traits, it is unfeasible that these associations are all uniquely disease-specific in the paradigm of "one gene - one function - one disease". Indeed, widespread pleiotropy is observed for complex traits, where a single variant is associated with multiple phenotypes hindering the understanding of loci associated functional consequences, e.g. multiple auto-immune diseases [35] and psychiatric disorders [36]. This pleiotropy is indeed evident from the genetic correlation between traits, a post-GWAS strategy that aids in understanding common mechanisms between traits but does not inform on causation between two traits. To understand the genetically mediated casual relationship, Mendelian Randomization (MR) is employed instead. Based on GWAS summary statistics, MR is an epidemiological strategy that considers genetic variants (instrumental variables) as an approximation of environmental exposure and is applied as a replacement for not available randomized control trials [37]. Opportunities, limitations, and suggested strategies for genetic correlation and Mendelian Randomization are briefly discussed in section 2.4.

Furthermore, due to the polygenic nature of complex diseases, the predictive information from GWAS studies can be summarized for each individual via Polygenic Risk Score (PRS). The PRS is a score computed for each individual as the sum of the alleles frequency weighted by their SNP effect sizes at independent loci for a complex trait [38]. Thus, PRS represents the risk that an individual carries of developing that complex trait. Although not powerful enough in a clinical setting to predict the actual onset of a disease, it is however useful to detect groups at high and low risk and will be the object of discussion in the context of patient stratification in section 2.5. A similar idea was also extended to molecular endophenotypes such as gene expression [10]. Details on the developed methods for gene expression prediction models from cis-regulatory effects are deepened in section 2.2.3.

The functional characterization of complex diseases eases the development of proper treatment interventions. Drugs with genetically supported targets from GWAS are more likely to pass clinical trials [39]. Notably, small effect sizes of variants detected in a population can still imply a relevant effect in a molecular phenotype and consequentially have an impact on drug-gene targets, for example through disease relevant biological pathways [40]. An understanding of GWAS functional consequences, not simply focusing on the strongest associations, can improve drug development and repurposing for complex diseases. Given this broad overview of post-GWAS methodologies, the first step in untangling the functional implication of GWAS results is rooted in our ability to get insight into their transcriptional consequences and relationship.

2.2 GWAS functional interpretation: from location to target genes

Understanding the target gene from a disease associated variant is challenging. As already mentioned, GWAS results are grouped into loci hence hampering the identification of 1) causal variants, 2) affected genes in the locus and 3) cell types in which those genes are active. The limited amount of fine-mapped GWAS loci in correspondence of the coding region of a gene can be analysed with tools such as ANNOVAR [41] to understand the functional consequence on the mapped gene product. However, these variants corresponds to only $2-3\%$ [5] of the total disease associated ones, with the vast majority ($\sim 90\%$) located in non-coding regions and hence not easily connected to a putative causal gene [24, 42]. In this context, projects like Roadmap Epigenomics [43], ENCODE (Encyclopedia of DNA Elements) [44], and BLUEPRINT [45] gave a unique opportunity in understanding the regulatory consequences by the characterization of epigenetic marks across human tissues and cell types. Indeed, the provided genomic annotations from the aforementioned initiatives have been extensively used in SNP enrichment methods that detect the overlap among GWAS variants and regulatory regions more frequently than what would be expected by chance and prioritize cell types on which those regulatory mechanisms are active (section 2.2.1 for proposed methods).
Another approach for identifying target genes of associated variants is based on the integration of GWAS results with quantitative trait loci (QTL) output for a certain molecular phenotype such as gene expression (eQTL). In particular, eQTL analysis tests the association between gene expression and allele frequency for each variants that are in cis- or trans- positions, i.e. in proximity of or distant from Transcription Starting Site (TSS) respectively, hence the genetic impact on gene expression regulation. In addition, the integration of eQTL results with GWAS output via colocalization methods pinpoints loci sharing a disease effect and a regulatory mechanism for gene expression, thus likely to regulate the molecular mechanisms leading to disease etiology (section 2.2.2).
Nevertheless, the eQTL strategy tests one SNP at a time and for each gene separately, being dependent and confounded by the LD structure. Thus, this framework was extended to transcriptome-wide association study (TWAS) studies that build a gene expression prediction model from the overall cis-variants contribution and imputed the expression of large-scale cohort composed of GWAS genotype-only dataset, directly testing the imputed gene effect on the trait rather than the single SNP (section 2.2.3). Both colocalization and TWAS methods rely on reference panels composed of genotype-gene expression matching data such as The Genotype-Tissue Expression (GTEx) project[46]. In both cases, the accuracy in translating GWAS findings is dependent on the variety of collected cell types and the available sample size. Finally, the fine-mapped GWAS variants in regulatory regions such as enhancers can be linked to the controlled gene via high-throughput chromosome conformation capture (Hi-C), hence removing the spatial constraint of close proximity and

reflecting enhancer-promoter loops [47].

The next subsections elaborate on the previously mentioned strategies.

2.2.1 SNPs enrichment in functional categories

SNP enrichment methods aim at identify relevant functional categories of the genome that overlap with GWAS disease-associated variant more frequently than expected by chance and consequently prioritize cell types influenced by those categories.

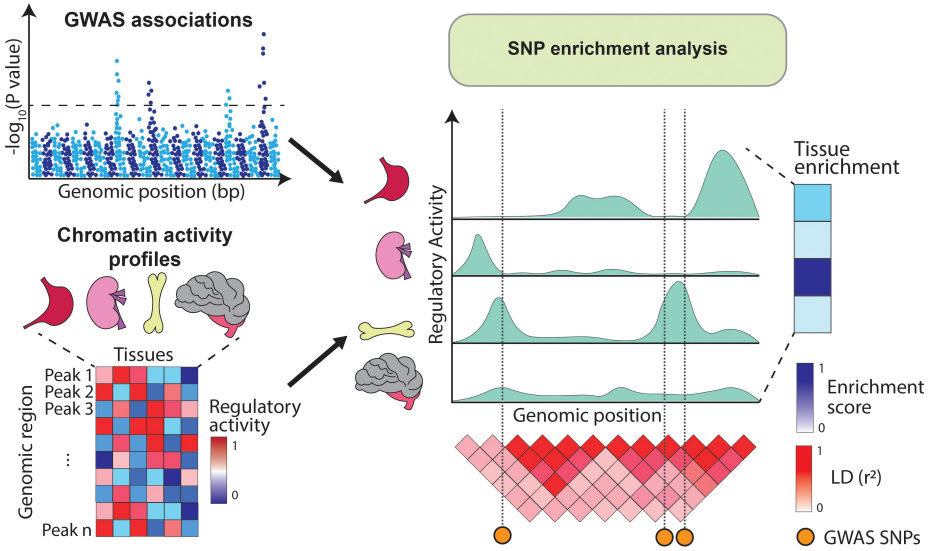

Fig. 2.4.: Figure from [48] "Overview of SNP enrichment analysis using chromatin annotations. SNP enrichment analysis integrates association signals from GWAS (Manhattan plot on the top left) with functional genomics data such as chromatin annotations (heatmap on the bottom left). GWAS SNPs are overlapped with regulatory elements (right panel) and if in a given tissue the overlap occurs more frequently than expected by chance, the tissue is assigned a high enrichment score."

For example, Hu et al.[49] developed a pivotal approach called SNPsea that detected the enrichment of tissue-specific expressed genes with genes overlapping GWAS loci, hence identifying pathogenic cell states. Similarly, GWAS variants can be integrated with tissue-specific epigenomic annotations for open-chromatin states such as DNase-hypersensitivity (DHS), ATAC-seq, for enhancer and silencers regions such as H3K4me1, H3K4me3, H3K27ac or H3K27me3 histone modifications and DNA methylation (Fig. 2.4). For example, via binomial test it was found that GWAS signals were enriched in tissue-specific DHS regions [24], such as variants for heart diseases was enriched in DHS regions observed in fetal cells. With a more sophisticated test that took into account peak properties, epiGWAS [50] method found an enrichment of type 2 diabetes variants in gene promoters (detected from ChIP-seq for histone modification technology) active in liver and pancreatic cells. More recently, Iotchkova et al. developed a method called GARFIELD [51] that identifies SNP enrichment on regulatory regions allowing different threshold levels

of significance and incorporating distance to nearest TSS and LD structure. Strikingly, they found that glycemic indices β-cell activity index resulted enriched in pancreatic islets enhancers not at the standard GWAS significance level of $P \leq 10^{-8}$ but only at lower significance threshold of $P \leq 10^{-5}$, suggesting a role of low effect variants that is not retained via the Bonferroni correction threshold. A widely used method in the context of enrichment analysis is the partition heritability. In particular, SNP heritability of a complex disease is defined as the amount of phenotypic variance that is explained by the additive effects of genotyped and imputed SNPs. Approaches partitioning heritability such as LD-score regression (LDSC) [52] test for an enrichment of phenotype heritability in specific functional categories of the genome, with the idea that if disease associated variants overlap more frequently with a specific functional category, it implies that those variants explain more heritability. Indeed, LDSC methodology revealed a generally higher heritability in conserved regions, and a disease-specific one in cell-type specific enhancers, such as central nervous system for schizophrenia and bipolar disorders and adrenal or pancreas for fasting glucose [52]. This framework was also extended to specifically expressed genes (LDSC-SEG) [53] that tested whether disease heritability is enriched in neighbouring regions for genes that have a tissue-specific expression. In this case, there was not a specific connection to putative causal genes whose disruption is associated with the disease but rather an overall analysis on the gene expression role in the disease etiology. With this methodology the authors revealed that SNP heritability enrichment for schizophrenia was induced by glutamatergic neurons, and bipolar disorder SNP heritability enrichment was instead induced by GABAergic neurons.

2.2.2 Co-localization with quantitative trait loci

The integration of eQTL with GWAS results allows to map disease-associated variants to likely causal genes in a tissue-specific context (Fig. 2.5) and understand the molecular mechanisms that are altered by these variants. A comprehensive catalog of QTL (expression, splicing, open chromatin, etc) is made available to the entire community by initiatives such as the GTEx consortium and Common Mind Consortium (CMC), the former across multiple tissues and cell types and the latter specifically focusing on brain sample collections. Nowadays, GTEx includes matched genotype and gene expression RNA-seq data available in more than 800 postmortem donors and across 52 tissues [46]. Instead, CMC generated multiple data modalities (RNA-seq, genotype, and epigenetic) from individuals affected by schizophrenia and bipolar disorders, and unaffected ones [54]. Of note, eQTL is just one of the molecular traits that can be investigated for the association with regulatory variants and indeed the collection ranges from protein concentrations (pQTLs) [55] to DNA methylation (mQTLs) [56] and chromatin accessibility (caQTLs) [57].

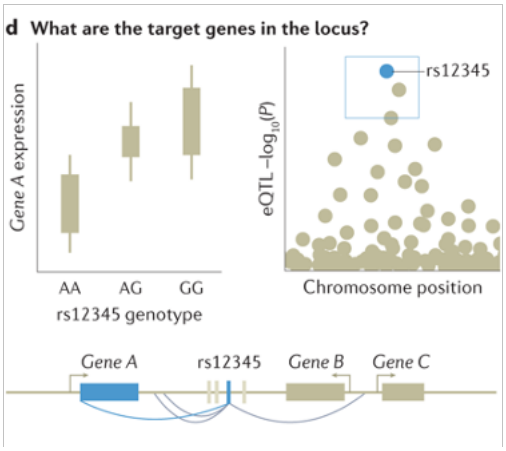

Fig. 2.5.: Figure from [16] "Target gene for a GWAS locus can be prioritized by mapping expression quantitative trait loci (eQTLs) (left) and their co-localization (right) to identify loci where the causal variant from GWAS is also a causal variant affecting gene expression. For GWAS variants in enhancers, high-throughput chromosome conformation capture (Hi-C) data and maps of enhancer target genes can be used together with simple prioritization by distance to identify genes affected by the causal variant (below)."

In order to integrate GWAS and eQTLs results, co-localization approaches have been developed to identify a variant in each locus that is causal for both the regulatory association with a gene and the disease etiology [58–60]. Indeed, LD structure hinders the identification of such putative causal variants. An overlap between eQTL and GWAS signals can arise from 1) independently causal variants (for GWAS and eQTL respectively) in LD with each other, 2) a single causal variant with two independent effects on the disease and gene expression or 3) a single causal variant that leads to the disease by its changes on gene expression. Co-localization methods aim at estimating the probability of overlap, identifying variants such that this probability is more than what is expected by chance, and distinguishing between the aforementioned scenarios. For example, Giambartolomei et al. developed a method called COLOC [59] that for each locus having an eQTL and disease (GWAS) association calculates colocalization odds as probability estimated via a Bayesian approach. The method tests the alternative hypothesis of the observed scenario arising from the effect of a single shared SNP versus the null hypotheses of i) no association with either eQTL or disease, ii) only associated with the disease, iii) only associated with an eQTL, iv) associated with both but due to independent variants. Although being a reference method for co-localization, it is 1) constrained to assuming a single SNP in each locus as causal and 2) it solely tests GWAS - gene expression mechanisms without the inclusion of molecular traits such as epigenetic phenotypes. To solve the second issue, an extension called MOLOC [61] was developed that included a third eQTL mechanism (methylation) that significantly increased the power to connect variants to genes. On the other hand, because one locus can include multiple causal SNPs for both gene expression regulation and disease association, Hormozdiari et al. proposed a method called eCAVIAR [58], from the fine-mapping strategy CAVIAR [31]. eCAVIAR is based on an integration of the fine-mapping arising from GWAS and eQTL results by defining a probability of colocalization as the product of the two fine-mapping posterior probabilities, thus losing

the assumption of a single variant per locus while accounting for LD. Nevertheless, these methodologies are only observational and cannot discern actual consequential causality from pleiotropy mechanisms, in which a single variant influences disease pathogenesis and gene expression or another molecular trait independently. Indeed putative colocalization should be further biological validated for example via gene-editing technologies.
Finally, gene expression is not only mediated via cis-eQTL but also via variants distantly located for their TSS (trans-eQTL). For this reason, He at al. [60] proposed a method called Sherlock that compared a gene QTL signature (both cis- and trans-) and GWAS result genome-wide instead of one locus at the time. Their methodology allowed them to identify four candidate genes mediated by GWAS variants for T2D, among which two could only be identified via the integration of trans-eQTL effects.

2.2.3 Transcriptome-wide association studies

The integration of GWAS and eQTL results to pinpoint possible mechanisms of action via gene expression is based on the usage of genome-wide significant variants. Nevertheless, the identification of variants related to complex diseases has not yet reached a plateau due to the small effect sizes that can only be detected via an adequate sample size [5]. In addition, colocalization methods cannot discern between pleiotropy and causal mediating effects. Thus, a recently proposed alternative to identify target genes are transcriptome-wide association studies (TWAS) [9, 10], that test the association between a trait and the genetically regulated component of gene expression, drastically reducing the number of association tests performed from millions to thousands and consequently multiple testing burden. TWAS leverage reference panels composed of matching genotype and gene expression data such as GTEx and CMC projects to learn the gene expression prediction models from cis-components, expressed in terms of weights, and subsequently impute gene expression into large-scale genotype data for GWAS (possible both at the individual and summary-statistic level) to directly associate it with a trait [62]. The first two methodologies developed (in parallel) were Fusion (initially called simply TWAS) [9] and prediXcan [10]. Despite being based on the same principle, they differ in the approach applied to model gene expression from cis-effects. In particular, a schematic of the common principle is depicted in Fig. 2.6. Briefly, for a certain tissue, let M be the number of samples with matched genotype and gene expression and P the number of cis-variants for a gene g. For each gene g, $\boldsymbol{Y}^g = (Y_1^g, \dots, Y_M^g)$ and $X = [\boldsymbol{X}_1 | \dots | \boldsymbol{X}_P]$ denote the vector of observed expression and the genotype matrix of dosages for imputed variants across M samples, respectively. The gene expression is modeled via additive genetic components as

$$\boldsymbol{Y}^g = \sum_{p=1}^{P} w_p^g \boldsymbol{X}_p + \epsilon \tag{2.1}$$

with ϵ representing additional factors contribution to gene expression independent from genetic components. The problem in (2.1) that estimates optimal weights $\boldsymbol{w}^g$ modeling

gene expression is solved via elastic-net regression [63] in prediXcan [10] and using the best performing model among single most significant cis-eQTL, BLUP (best linear unbiased predictor) [64], BLSMM (Bayesian sparse linear mixed models) [65], LASSO (least absolute shrinkage and selection operator) [66] and elastic-net models in Fusion [9].

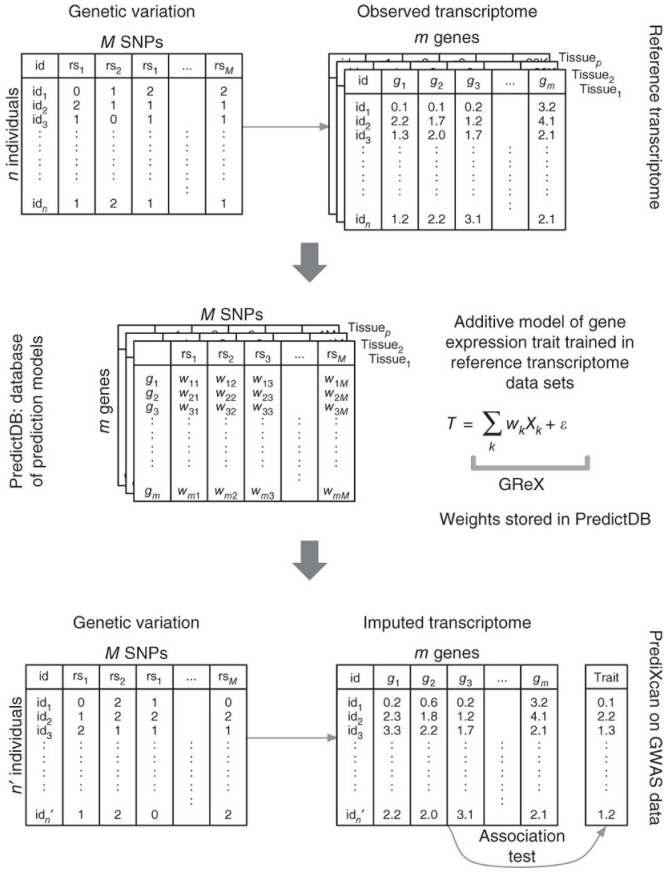

Fig. 2.6.: Transcription-wide association studies workflow, figure from prediXcan method [10]. Top panel: reference transcriptome studies such as GTEx are used to derived gene expression prediction models and fit tissue-specific model coefficients (weights) w for each gene and SNP. Bottom panel: PrediXcan is applied to a GWAS dataset, imputing expression levels on the whole transcriptome based on model weights. These levels are then correlated with trait of interest via linear, logistic or Cox regression.

Once the gene expression prediction model is created and weights $\hat{\boldsymbol{w}}^g$ are estimated, imputed gene expression on large-scale genotype-only datasets $\tilde{X} = [\tilde{\boldsymbol{X}}_1|\ldots|\tilde{\boldsymbol{X}}_P]$ is obtained as

$$\boldsymbol{T}^g = \sum_{p=1}^{P} \hat{w}_p^g \tilde{\boldsymbol{X}}_p. \tag{2.2}$$

Finally, the imputed gene expression is tested for association with the trait in the same way as GWAS, for instance via generalized linear models correcting for additional covariates.

Of note, the prediction of gene expression based on cis-genetic information, i.e. SNPs around the gene TSS, is possible due to the high heritability (explained variance) of gene expression that can be attributed to variants in proximity of genes [67]. On the one hand,

TWAS methodologies allow to test the association of gene expression with a trait in a tissue-specific manner, avoiding the profiling of expression in the large-scale GWAS cohort(s), especially difficult for hard to collect tissues such as brain-related regions. On the other hand, the imputed gene expression solely model the heritable component attributed to cis-genetic effects, hence minimizing the confounding induced by changes in gene expression caused by the presence of the disease. Most importantly, the single variants that would be associated with a trait in a GWAS study are now aggregated in a unique value that represent gene expression, regardless their level of trait-related significance and without applying any p-value based filtering strategy. Finally, the resulting TWAS output (in the same form as GWAS) inform in the directionality of the effect, namely the increase of gene expression is associated to trait presence or vice-versa, that is otherwise not possible in a simple "closest gene" analysis or hard to retrieve from a colocalization approach. This gene directionality can aid at nominating successful drug targets and can be integrated with post-perturbation gene expression screenings [68] to direct towards the development of new therapies and drug repurposing [69].

The described framework presumes the availability of individual-level genotypes $\tilde{X}$, nevertheless both methodologies were extended to the usage of GWAS summary statistics in order to exploit publicly available GWAS results from a large sample collection, achieving comparable results [9, 70]. In addition, Fusion methodology was integrated with epigenetic data in the form of chromatin marks (H3K27ac, H3K4me1 and H3K4me3) to identify genes whose association with the disease was mediated via gene-chromatin interaction in the context of schizophrenia [71]. Moreover, due to the relevance and overlap of regulatory regions both for GWAS and detected eQTLs [50], an extension of prediXcan called EpiXcan [69] was proposed that included epigenetic annotations such as DNA methylation and histone modifications in a Bayesian hierarchical model, weighting SNP regression coefficients by the overlap with gene regulatory elements.

Although being a powerful tool to reveal mediating effect via gene expression, TWAS results are still affected by LD structure similarly to GWAS and additionally confounded by gene co-expression network that can lead to spurious associations and false positive results [62]. This issue was attenuated by an extension of Fusion that included a fine-mapping approach called FOCUS [72] and modeled the correlation among TWAS signals creating a posterior probability of credibility for each gene in a loci via a Bayesian approach.

In general, TWAS methodologies highlighted that the majority of disease-associated genes were not the nearest gene of a significant GWAS result [9]. Moreover, genes-trait significant association were found mostly tissue-specific across a variety of phenotypes and tissues [70], indicating the need to collect and predict gene expression in disease-relevant cell types [62]. Recently, an extension of TWAS was proposed by Zhang et al. [69] that integrated epigenetic information to model gene expression. Their method called EpiXcan increased the number of genes accurately predicted and thus tested for trait associations, being particularly relevant for tissues with restricted sample size [69]. Hence, TWAS approaches are assisting in revealing target tissues and genes affected in complex diseases such as schizophrenia (SCZ) and coronary artery disease (CAD). Application of TWAS

strategies on SCZ and CAD are discussed in section 2.6.

2.3 Pathway-based strategies to decipher variants roles

Many complex diseases such as schizophrenia [73] and coronary artery disease [74] have proven to be highly polygenic, with hundreds of variants associated and thousands of additional SNPs only carrying a small effect. Although the entire mechanistic pattern leading to the etiology of a complex disease is still not clear, polygenic signals from GWAS can be linked to biological processes and molecular functions, highlighting a convergence on specific pathways.

2.3.1 Pathway and functional gene-set databases

The term pathway refers to a collection of interactions between molecules in a cell that brings to a new product or specific modifications in the cell itself. Laboratory studies led to the identification of pathways in human and model systems, nevertheless the complete picture and collection are far from complete. The current knowledge on biological pathways is available to the entire community via databases among which Reactome [75], Gene Ontology (GO) [76] and WikiPathways [77]. Reactome [75] derivation is based on reactions and the corresponding players such as proteins, nucleic acids, small molecules, and complexes. These entities participating in a certain reaction are grouped together to form a pathway i.e. a network of biological interactions. The pathways are supported by literature citations, experimental verification and inferred from non-human experimental details additionally confirmed by expert biologists. In addition, pathways in Reactome are hierarchically organized such that related detailed pathways are combined into larger domains, for instance potassium channels and protein-protein interactions at synapses are categorized under neuronal system class. Gene Ontology collection [76] instead aims at unifying the representation of gene and gene product attributes focusing on their function and covers three domains: cellular component (parts of a cell and the extracellular environment), biological process (molecular events related to cells, tissues, organs, and organisms) and molecular function (activities of a gene product at the molecular level). In addition, GO is structured as a directed acyclic graph with each term having defined relationships to one or more other terms. The ontology is often updated with additions, corrections, and alterations suggested by members of the research and annotation communities and those in the GO project. Finally, WikiPathways [77] incorporates the joint knowledge of the scientific community in biological pathways, with a contribution from any user that is monitored by the database admins. Initially focused on genes and protein products, it recently included an annotation on metabolites and their

interactions, with an ever-increasing in size and accuracy.

2.3.2 Pathway analysis tools

To interpret GWAS results and identify enriched biological pathways towards which GWAS results converged, genetic variants are first aggregated into genes and then to biological pathways in form of gene-sets. The definition of gene-sets and their annotation accuracy is hence critical as randomly selected genes do not carry the information of shared biological meaningful mechanisms. Among the available tools to assess pathway relevance in a trait etiology, MAGENTA [78] was one of the first methods developed. MAGENTA first transforms GWAS significant results to gene scores based on genome location overlap considering only the most significant associated variant, then corrects for gene size and LD structure and finally applies a non-parametric test for Gene Set Enrichment Analysis (GSEA) to discover enrichment of genes in biological pathways based on a predefined gene-score cut-off. Similarly, INRICH [79] finds enriched pathways by overlap of the most significant SNP within a gene region (interval) and tests the number of interval obtained combining all the genes from a pathway compared to the empiric null distribution that matched interval size, gene overlap and LD structure. Other methods instead do not rely on the most significant SNP in a locus. For example, PASCAL [80] creates a gene score and subsequently a pathway score from GWAS summary statistics by the merging of SNPs and genes in the same locus, leveraging LD structure from reference population, thus calculating a pathway significance not subject to a prior defined gene-score threshold. For example, these last two methodologies were used in conjunction in multi-trait analysis to detect enriched pathway from GWAS summary statistics across 25 traits finding shared functional mechanisms for example between the immune and the psychiatric group [81]. Another widely used method called MAGMA [82] first builds a gene score considering mapped SNPs using a linear regression framework that test against the phenotype (or a combination of SNPs summary statistics if the individual phenotype is not available), and then computes gene-set score by including the correlations among genes in a linear regression framework.

The methodologies mentioned so far do not take into consideration the tissue specificity of the mapped gene and hence of the biological pathway. In addition, gene-set annotation is manually created or assessed from molecular evidence but is in general biased towards well-studied genes. Thus, DEPICT [83] address these issues starting from the hypothesis that relevant genes for a disease should share functional annotations and identify functional gene sets enriched in genes within associated loci. From GWAS summary statistics, it first define disease independently related loci from a p-value predefined cut-off and corresponding mapped genes. Then, from co-regulation derived from gene expression it predicts a gene function, defines "reconstituted" gene sets and computes the probability of a gene to belong to that gene sets. Finally, it searches for reconstituted gene sets enriched by genes in the associated loci as well as tissue cell type of interest by considering

genes in associated loci that are over-expressed in a certain tissue. Finally, a recently developed method called PoPS [84] uses the same assumption of "causal genes share same functions" to create a prioritization score for each gene based on a generalized linear model that consider as feature the membership to a biological pathway or expression in tissues disease-relevant. In this way, the authors could inspect the biological pathways with higher relevance in assigning gene priority and for example identified chromatin organization and lipid biosynthesis as the pathways with highest feature score for schizophrenia and LDL traits respectively. Nevertheless, their gene prioritization starts from MAGMA application and hence relies on GWAS summary statistic initial cut-off as well as variants mapping to genes based on genomic location.
The strategies aforementioned rely on genes alignment to disease-relevant loci rather than their tissue-specific functional consequences on gene expression. Other methods instead work in the context of additionally leveraging eQTL results or TWAS in pathway analysis. For instance, Wu et al. [85] integrated eQTL SNP-gene weights, GWAS summary statistic and LD information to identify disease-related pathways in a self-contained manner i.e. having as null hypothesis that no gene in the pathway is associated with the disease. When applied to SCZ summary statistic and with gene weights built from dorsolateral prefrontal cortex in CMC, they identified GABA receptor complex as a novel disease-relevant pathway that does not include any TWAS significant gene. In addition, the same authors also used a competitive gene-set analysis approach, considering TWAS significant results and testing for enrichment via hyper-geometric test (DAVID [86]), as practice in pathway analysis from differentially expressed genes. However, this two-step approach identified only 1 pathway relevant for SCZ ("sequence-specific DNA binding") compared to 15 of their proposed methodology, possibly due to an aggregation of signal from small effect genes that cannot be detected in a competitive gene-set analysis.

Moreover, an implementation of GSEA for TWAS output was developed (TWAS-GSEA) that uses a similar strategy of MAGMA methodology [87]. TWAS-GSEA builds a linear mixed model regression from TWAS Z-score summary statistic on gene set membership, while considering the gene correlation from LD structure. However, this method still relies on an a predefined cut-off for disease-associated genes. When applied to depression, TWAS-GSEA revealed 7 enriched pathways such as macro-molecular complex binding [88].
Finally, the methods mentioned so far are based on aggregation of variants/genes signals from summary statistics results that are usually public available to the community and hence easier to retrieve. However, sample-based pathway analysis can highlight disruption of a certain pathway specific for each affected individual. For example, this has been optimized in the context of transcriptional signatures for cancer [89] where individual enrichment scores are calculated from gene expression for each patient and used to detect pathways potentially functioning as prognostic biomarkers. In connection with GWAS instead, individual pathway levels were proposed in form of pathway-specific polygenic risk scores (see section 2.5 for details on PRS) for Alzheimer's disease [90], focusing solely on SNPs mapping pathway related genes. The authors associated pathway-specific PRS with

cognitive performances and found Aβ Clearance and cholesterol pathways significantly associated with working memory, Aβ deposition and neurodegenerative biomarkers, with the majority of the strength in associations driven by the inclusion of APOE genotype. Similarly, Choi et al. [91] (preprint) argued against the direct usage of PRS as due to information loss compared to an informed individual risk scores based on biological pathways. Hence they developed a software called PRSet that for each individual builds a PRS across multiple biological pathways. This individual-level pathway score was tested for enrichment in a competitive manner accounting for pathway size using permutation but has lower power in a small target sample size setting when compared to MAGMA and LDSC methods. Nevertheless, this method was developed focusing on disease stratification other than pathway identification and will be further discussed in section 2.5. Of note, computing individual-level pathway scores provides a unique opportunity to untangle individual risk mediated by a biological mechanisms, nevertheless the methods mentioned here are based on associated variants from GWAS to which genes are then mapped by location, hence missing their regulatory effect in gene expression.

In conclusion, the state-of-the-art methods for post-GWAS pathway analysis mostly rely on a p-value cut-off decided a priori (either at the SNP or gene level). On the one hand, a loose cut-off increases the noise introducing false positive results, on the other hand a stringent threshold can remove small effect but still relevant information. Indeed, not genome-wide significant results can still explain the majority of complex diseases heritability [7]. A focus on the actual gene regulation can aid to narrow down functional consequences of the associated SNPs, while avoid removing small effect variants. It is important to note that, the polygenic nature of complex diseases hampers the discovery of disrupted biological mechanisms and small effects variants can still have functional consequences when converging on a common molecular mechanism. In addition, an elevated polygenicity implicates that each affected individual has a unique allelic variation and SNP configuration [92] that can converge to different biological mechanisms and possibly change treatment results and intermediate phenotypes relevant for the disease. Hence, individual pathway levels represent a unique opportunity to understand the functional consequences and pathway disruption in each patient.

2.4 Genetic correlation and causality between complex traits

GWAS results highlighted that complex traits are associated with a number of variants going from hundreds to thousands, suggesting that some of these variants can be shared and related to multiple traits [5]. This phenomenon is also known as pleiotropy in which a single variant can affect two different and seemingly unrelated traits. Indeed, it was confirmed that the majority of trait-associated loci overlap for multiple traits [93]. This

relationship between two traits can be inferred from genetic correlation i.e. the degree in which the variants responsible for a trait are also relevant in another trait. Understanding the genetic correlation between complex traits can elucidate shared etiological mechanisms and point to putative causal relationship. Tools such as cross-trait LD score regression, estimate genetic correlation between traits solely using GWAS summary statistics and providing a frame-work that is unbiased with respect to the possible sample overlap [94]. With this method, the authors showed a positive correlation between psychiatric disorders, even in case of one trait being under-powered such as schizophrenia and anorexia, or coronary artery disease and LDL/triglyceride level traits. Nevertheless, the observed genetic correlation does not imply a causality between two traits but only a possibility of existence of causal effects. In particular, genetic correlation can be a consequence of 1) vertical pleiotropy, i.e. causality, in which a trait causes the other, 2) horizontal pleiotropy with the associated variants connected to the two traits independently or horizontal pleiotropy induced by LD where variants are in high LD and still related to the two traits independently, and 3) pleiotropy induced by polygenicity in which the influence on traits arise from a mixture of the aforementioned configurations [95]. To understand whether the genetic correlation observed is a consequence of a causal relationship, Mendelian Randomization (MR) strategy is preferred [37]. MR is an epidemiological technique to obtain unbiased estimates of a phenotype being a risk factor for another trait without the establishment of a traditional randomized control trial and instead considering genetic variants as instrumental variables. Because variants alleles are randomly inherited from the parents they are regarded as similar to the assignment of randomized treatment. In addition, being variants fixed at conception, they are supposedly not related to additional environmental factors that can imply confounding. In the MR approach, the two considered traits are divided in exposure and outcome with exposure representing a risk factor for the outcome, while the outcome being the disease of interested (e.g. blood pressure risk factor for hypertension). MR methodology can thus test the causality of the exposure on the disease via genetic variants considered as instrumental variables (Fig. 2.7).

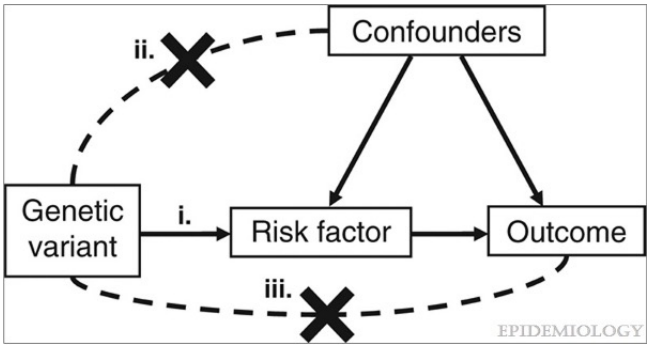

Fig. 2.7.: Figure from [96] "Diagram of instrumental variable assumptions for Mendelian randomization. The three assumptions (i, ii, iii) are illustrated by the presence of an arrow, indicating the effect of one variable on the other (assumption i), or by a dashed line with a cross, indicating that there is no direct effect of one variable on the other (assumptions ii and iii)."

However, the validity of the test depends on the validity of three strong assumptions: i) the genetic variants are associated with the exposure, ii) the same genetic variants considered are not associated with any confounders either of the exposure nor of the outcome and iii) those genetic variants are only related to the outcome via the exposure [96]. Some of the assumptions are not fully testable such as the non association with confounders of risk factor and outcome as it is improbable that all confounders are measured or even known. In the case of the risk factor being a protein biomarkers, variants within the gene region can be used as being the most informative proxies of the risk factor. Nevertheless, using multiple genetic variants encompassing more than one gene region is particularly suited for risk factors that are complex diseases and hence polygenic in being caused my multiple variants. In addition, the usage of genome-wide variants can enhance MR power, despite the drawback of inflating false positive when the assumption of valid instrumental variables are not satisfied even for a single variant. If outcome association is in accordance among all the genetic variants considered, a causal association between exposure and outcome would be reasonable, even for some instrumental variables that are not satisfying the assumptions [96].

Multiple methodologies have been proposed so far, the majority of which do not require individual level data and are built on GWAS summary statistics, known as two-sample Mendelian Randomization. Examples of widely used methods are inverse-weighted variance (IWV) with fixed or random effects [97], weighted median and mode estimator [98] and MR-Egger regression [99], some of which allow for weaker hypothesis such as possibility of horizontal pleiotropic effects (MR-Egger and IWV random effects) or at least half of the genetic variants being valid instrument (median and mode estimator). In addition, two-sample MR methods are based on the assumption that summary-level variant-exposure and variant-outcome associations are estimated from two different populations. Nevertheless, it was shown that some of these methodologies can be reliably applied from summary statistics coming from the same population when the sample size is large enough [100]. Details and formulation of IWV method are provided in section 3.2.3. In general, the comparison of estimates among a range of methods providing a similar results is the preferred strategy to pinpoint towards putative causal effects [96].

Of note, MR strategies have been also developed in the context of gene expression with the aim of estimating the causal effect of genes on the complex trait considered. This strategy is indeed another alternative to TWAS and colocalization studies and exemplar methodology is SMR (summary data-based Mendelian randomization) [101], where genetic variants are regarded as instrumental variables and the effect size of a gene for a disease is approximated via the ratio of GWAS summary statistics and eQTL effect sizes for the top cis-variant, considering only the variant related to the expression of the gene and in a tissue-specific manner. Similarly to TWAS, SMR allows to test for the association of a gene to a trait through genetic mechanisms, nevertheless only considering top eQTL variant (most significant) and hence requiring a large sample size of eQTL catalogs to overcome polygenicity as well as limiting the gene regulation to a unique genetic factor without accounting for the more complex interplay of cis-variants in gene expression

regulation. Recently an extension called transcriptome-wide summary statistics-based Mendelian Randomization approach (TWMR) was developed by Porcu et al. [102] that uses multiple variants as genetic instruments and multiple genes expressions regulated by the same variants as exposures. The multi-variable setting can address horizontal pleiotropy (violation of assumption iii. Fig. 2.7) in case mediators are additionally genes regulated by the same variants and mitigate bias via a joint estimation of exposures effect on the outcome disease.
Established results arisen from MR application in the context of coronary artery disease and schizophrenia will be discussed in sections 2.6.1 and 2.6.2 respectively.

2.5 Patients stratification

Being complex diseases highly polygenic, each affected individual carries a unique combination of alleles in the associated loci, representing a background liability that can be pushed into the emergence of the disease via environmental interactions or rare variant configurations [103]. To define individual risk liability from common additive effects, polygenic risk score (PRS) methodology was soon introduced after the advent of GWAS [38]. Effect sizes of selected independent variants associated with a diseases from GWAS output are considered as weight and multiplied by the genetic dosage (reflecting the allele configuration) of each individual, finally summing across all loci and creating a unique score (Fig. 2.8). Prediction model built on PRS are not particularly informative in accurately separating the space of cases and controls, with a variability dependent on the polygenicity architecture of the disease, its genetic heritability and variants effect sizes [16]. Nevertheless, PRS can be used to identify individuals with highest and lowest risks and suggest specific prevention strategies [104]. For instance, PRS identified a subset of individuals with more than three times the risk of developing coronary artery disease (CAD), atrial fibrillation, type 2 diabetes, inflammatory bowel disease, or breast cancer. In case of CAD, the 8% of the population was 20-fold bigger than the population having rare monogenic mutations of familial hypercholesterolemia, carrying a similar disease risk [105]. Interestingly, it was also shown that polygenic background of an individual for a diseases can modify the risk given by monogenic variants (e.g. CAD and familial hypercholesterolemia), indicating that individuals with low polygenic risk score and carrying a monogenic mutation have a risk similar to the average of non-carrier individuals, and hence highlighting an interplay between monogenic mutation and complex disease polygenic predisposition [106]. Although PRS development encourages a clinical application especially in the context of prevention, challenges remain in term of accuracy decreasing when ancestries of GWAS discovery and target cohorts differ, thus particularly for non European populations [107]. On the other hand, most causal variants are supposed to be in common across populations, especially those that are related to functional consequences and overlap with functional annotations [108], and indeed

transferability can be improved with the inclusion of binding transcription factors and gene regulatory region information into PRS computation [109].

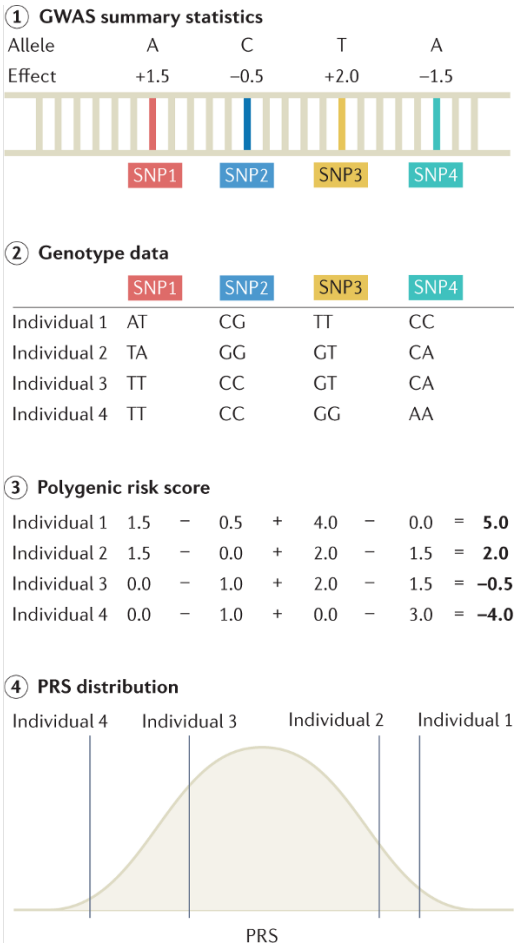

Fig. 2.8.: Figure from [16] "Overview of the steps necessary for calculating PRSs. Step 1: genome- wide association studies (GWAS) summary statistics are obtained, which detail the effect of each single-nucleotide polymorphism (SNP) on the phenotype of interest. Step 2: genotype data for a set of individuals are referenced against GWAS summary statistics. Here, genotype data for four SNPs are shown for four individuals. Step 3: polygenic risk scores (PRSs) can be calculated for each individual by summing up the effect sizes of all risk alleles for each individual. Step 4: linear regression analysis is performed on the calculated PRS to assess the effect of the PRS on the outcome measure"

Aside from defining an overall disease liability, attempts have been made to understand whether individuals affected by complex diseases are formed by subgroup of patients more similar to another disease from a genetic point of view. In this direction, BUHMBOX methodology tested whether the pleiotropy observed in autoimmune and psychiatric diseases reflects a specific subgroup of individuals genetically more similar to another disease/trait [110, 111]. The authors consider the variants associated with a trait A (e.g. major depressive disorder) and obtain the dosages for affected and control individuals on another trait B (e.g. schizophrenia). BUHMBOX calculates the correlation of trait B dosages (corrected for confounders) considering only trait B cases or trait B controls and then computes a delta-correlation z-score matrix representing the differences in correlation

among the cases and controls. Finally, it computes an overall statistic representing the weighted sum of correlations differences that under the null hypothesis of no subgroup heterogeneity follows a normal distribution. Hence, BUHMBOX tests for excessive positive correlation among trait B risk alleles in trait A cases, specifically searching for heterogeneity arising from a subgroup that includes another a priori known trait. However, this methodology did not detect any subgroup heterogeneity among 11 autoimmune diseases [110] nor found a subset of individuals affected by depression more genetically similar to individuals affected by other psychiatric disorders such as schizophrenia and autism spectrum disorder [111].

Nevertheless, the heterogeneity of complex diseases is clearly exhibited from the diversity of symptoms and characteristics that define a certain disease as well as the diversity in terms of endophenotype spectrum i.e. measurable entities genetically determined having a variability concordant with the disease variability [8]. Endophenotype can be seen as an intermediate manifestation of the actual disease for example at the biochemical, anatomical or psychological level, possible but not necessarily causal, that provide an additional diagnostic level which can be leveraged for a proper and more effective treatment strategy. For instance, the general diagnosis of asthma can be divided according the underlying inflammation mechanisms, distinguishing between subtypes with increased levels of neutrophilic and eosinophilic [112]. In term of CAD, it was possible to cluster individuals that experienced heart failure in six subgroups via unsupervised clustering based on 92 cardiovascular biomarkers, with the partition showing clear differences in term of clinical profile, prognosis and therapy response [113]. In a recent study [114], Nguyen et al. investigated the genetic heterogeneity of major depression (MD) performing subtype-specific GWAS of 16 MD subtypes in eight comparison groups defined from the symptomatology (vegetative symptoms, symptom severity, co-morbid anxiety disorder, age at onset, recurrence, suicidality, impairment, and postpartum depression). Clinically demanding subtypes such as recurrent, suicidal and early-onset had a higher genetic correlation with other psychiatric diseases, e.g. schizophrenia and bipolar disorder and estimated SNP-heritability via LDScore [52] was divergent across subtypes, generally higher for more severe manifestations. They also identified 47 genome-wide subtype-specific loci of which only 22 are significant in the most recent MD GWAS (with a 5 to 10-fold higher sample size). Together, these findings suggest a phenotypic characterization of MD that partially depends on a subtype-specific genetic liability. The already mentioned pathway PRS methodology developed by Choi et al. [91] outperformed genome-wide PRS in supervised disease stratification tasks, highlighting the relevance of independent pathways that can discriminate disease subtypes, in contrast to genome-wide PRS led by variants with a similar effect across subtypes. Nevertheless, pathway PRS failed to recognize well-characterized inflammatory bowel disease sub-classification in a unsupervised manner. Another study on complex disease genetic heterogeneity was recently performed in the context of Alzheimer disease (AD) [115]. The authors analyzed four forms of genetic risk, namely APOE-$\epsilon 4$ and APOE-$\epsilon 2$ alleles, polygenic risk computed via PRS and familial risk related to parental history of AD, associating them with 273 phenotypes in UK

Biobank cohort (e.g. blood biochemistry, psychological health and cognitive functions). Different forms of AD risk were associated with different traits, such as APOE alleles and lipid metabolism, APOE-$\epsilon 4$ and C-reactive protein, familial risk with psychological health and AD PRS with 16 traits among blood biochemistry, blood cell traits, metabolic health and general health classes.

In summary, complex diseases are characterized by an heterogeneity in endophenotypes and pathophysiology, however their connection to genetic profiles has been an object of study only of recent years and represent a promising goal towards personalized treatments.

2.6 Case studies: coronary artery disease and schizophrenia

We describe here the major findings from a genetic point of view obtained in a GWAS framework for two complex diseases, Coronary Artery Disease (CAD) and Schizophrenia (SCZ), as they will be the focus of the application Chapter 4.

2.6.1 Coronary artery disease

Coronary artery disease (CAD) is caused by the narrowing of coronary arteries due to an increase of fatty material and plaque formation inside the arteries, with a consequential reduction of the bloodstream to the heart muscle. Despite the progress in identifying reductable risk factors such as smoking, sedentary life style, and obesity, CAD is still the leading cause of deaths worldwide [116]. Much of the disease etiology is rooted in the genetic component, with CAD heritability estimated between 40 to 60% from family and twin studies [117] and cumulative explained heritability from GWAS studies around 40% [74, 118]. Since the first GWAS focusing on CAD in 2007, an increasing in sample-sizes and an improvement in sequencing power together with detailed phenotyping have led to the discovery of up to 321 CAD risk loci [119]. Post-GWAS analyses made possible gene prioritization and the connection to molecular pathway at the causal loci [74, 118, 120] (Fig. 2.9) with the latest characterization obtained from the largest GWAS in term of sample size in an unpublished study under review [74] that included more than 1 million participants and $180,000$ affected individuals. In GWAS context, the tendency moved from a Bonferroni genome-wide correction with a significant threshold of 5×10^{-8} being very stringent and increasing false negatives results (owing to the genome LD structure) to a False Discovery Rate (FDR) thresholding strategy [74, 120]. An FDR value of 0.05 correspond to an expected percentage of 5% of false positives in rejected null hypothesis (i.e. associated variants) and is less conservative than a family-wise error rate correction

such as Bonferroni (see section 3.2.2). Interestingly, the majority of variants identified as significant at an FDR level of 0.05 [120] became then significant at the genome-wide level of $P \leq 5 \times 10^{-8}$ in a recent study with increased sample size [74], highlighting that many FDR significant results are indeed true association from small effect variants.

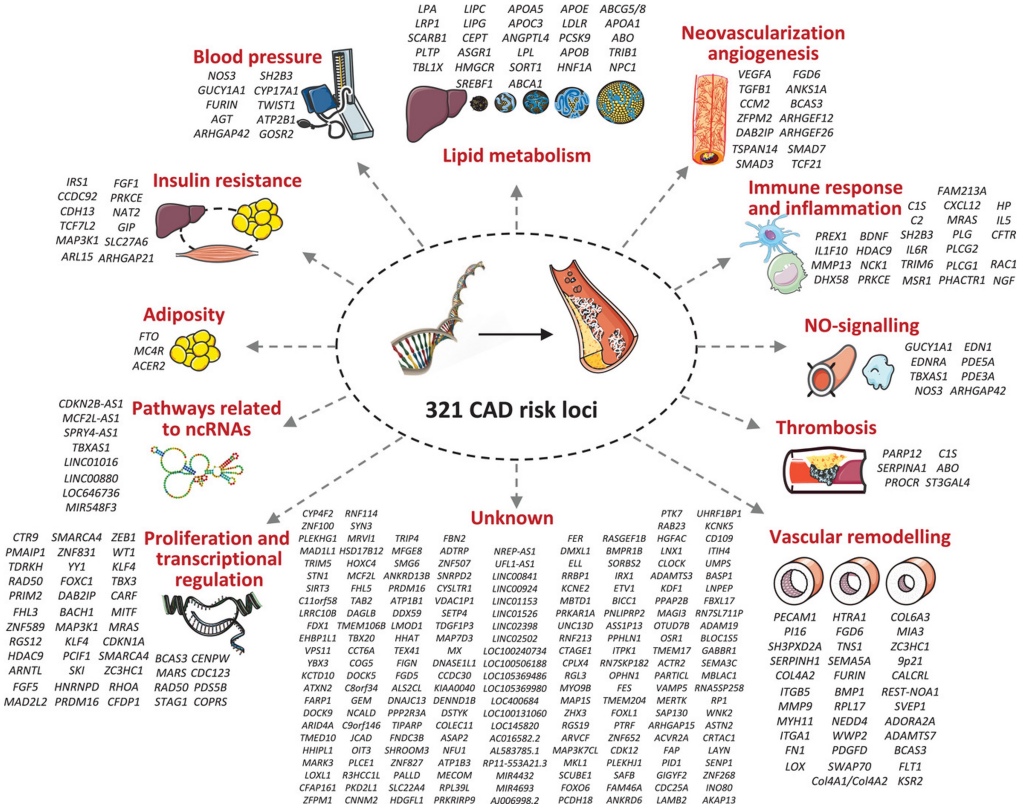

Fig. 2.9.: Figure from [119] "Genes mapped to 321 coronary artery disease (CAD) risk loci and related pathophysiological pathways of atherosclerosis"

Similar to other complex diseases, the genetic architecture of CAD is characterized by a large heritability explained by common variants with small effect sizes, the majority of which are located in non-coding regulatory regions [118]. The strongest association to date was already identified in the first GWAS in 2007 in 9p21 region, connected to the alteration of non-coding RNA ANRIL expression and nearby cyclin dependent kinase inhibitors CDKN2A and CDKN2B [121, 122] involved in cell cycle mechanisms and vascular remodeling. With the naive strategy of connecting relevant loci to genes based on genomic location, it was still possible to identify genes related to core CAD mechanisms such as metabolism of LDL, triglycerides or lipoprotein, blood pressure [123]. Among putative casual genes from GWAS, one striking and mechanistically understood example is represented by the association at 1p13 locus increasing CAD risk, connected to SORT1 decreased expression and consequential LDL increase via experimental validation. In particular, this locus contains a variant (rs12740374) associated both with CAD occurrence and LDL cholesterol levels [124] that was shown in in-vivo experiments to create a novel transcription factor binding site increasing the affinity to a enhancer-binding protein that

boosted the expression of SORT1 [25]. An increase of SORT1 expression regulates in turn LDL catabolism and APOB secretion and hence causes a decrease in LDL levels in plasma [125]. The relevance of lipoprotein blood circulating values emerged also from other CAD-loci associated positions. Indeed, 20% of associated loci are located in the proximity of genes involved in LDL, triglyceride or lipoprotein(a) metabolism [123]. Other than SORT1, GWAS pointed to genes such as LPA, LDLR, APOB, PCSK9, LPL, ANGPTL4, APOA5 and APOC3 some of which represent a promising therapeutic target, with on-going clinical trials [126] and additionally supported by family-based exome-sequencing studies. For example, mutation in coding genes PCSK9, LDLR and APOB are additionally identified as causes of familial hypercholesterolaemia [127], a genetic inherited disorder characterized by high LDL. Interestingly, a recent study investigated the interplay between monogenic risk variants in the aforementioned genes for familial hypercholesterolaemia and polygenic risk in CAD patients and found that the likelihood of CAD by age 75 years varied from 17% to 78% for carriers and from 5% to 30% for non-carriers with increasing range of PRS [106]. This interplay is consistent with the theory that the penetrance of a pathogenic variant in causing a certain disease and its manifestation is influenced by the accumulated genetic risk of the individual, i.e. polygenic background. Finally, LDL cholesterol was long identified and continuously validated by Mendelian Randomization (MR) studies as a functional component through which genotypes increased CAD risk [128], and hence represented a risk factor that can be modulated to reduce CAD risk. Moreover, MR studies together with epidemiological studies identified over the years many causal factor for CAD such as blood pressure, obesity, type 2 diabetes, alcohol and smoking [119]. Nevertheless, MR applications and epidemiological results were not always concordant. For instance, C-reactive protein (CRP), that is found related to the subsequent CAD risk in a log linear manner from observational epidemiological studies ([129]), was not causally related to CAD from MR studied [130]. This scenario was similar for HDL, for which a well-established reduction of HDL in CAD patient did not translate into protective role from a causal point of view estimated via MR studies [131], using as genetic instruments 14 common variants exclusively associated with HDL. Other than pathomechanisms related to lipid metabolisms, genes nearby CAD associated loci detected so far are also connected to blood pressure, immune response and inflammation, and vascular remodeling, however the functions and consequences of the majority of genes located near a significant loci remains unknown [119] (Fig. 2.9). An example of genes with functionally validated role is ADAMTS7, with knock-out mice showing a lower cellular proliferation and increased vascular remodelling and cell repair after vascular injury [132] as well as reduced aortic lesion formation indicating lower atherosclerosis burden [133].

Focusing on gene expression regulation from associated variants, Zhang et al. [69] applied a TWAS extension called EpiXcan to 58 traits among which CAD. In particular, the tissue specific expression allowed to identify pathophysiologically relevant tissue for CAD such as mammary artery and artery aorta together with related phenotypes such as liver for LDL and total cholesterol. Gene-set enrichment analysis of the corresponding CAD associated genes (not tissue-specific) identified known mechanisms, e.g. apolipoprotein binding and

cell proliferation Gene Ontology pathways. A recent TWAS applying EpiXcan on a larger cohort of CAD affected individuals and controls [134], identified 114 transcriptome-wide significant genes (via Bonferroni correction) of which 18 were outside previously reported significant loci. Among the novel putative genes, two genes detected in liver with possible effects on lipid metabolism (RGS19 and KPTN) were additionally functional validated via CRISPR-Cas9 knockdown in human liver cells, with a consequent reduction in cholesterol levels and leading to dysregulated expression of genes enriched in lipid metabolism mechanisms. Moreover, gene-set enrichment analysis from the overall TWAS hits detected cholesterol metabolism and regulation of lipoprotein levels as putative mechanisms as well as regulation of blood pressure and vascular remodeling. Finally, in a recent (unpublished) study, Aragam et al. [74] conducted a GWAS with the so far largest sample size that surpassed one million of individuals including more than $180,000$ affected by CAD and further implemented a systematic approach that prioritized variants as well as genes from multiple line of evidence. Combining 8 locus-based and similarity-based criteria including nearest gene to the most significant variant, causing a monogenic disorder, relevant phenotype in knock-out mouse, being an eQTL in relevant tissue and being prioritized by PoPS methodology [84], 94 genes with at least 3 concordant predictors were established among the 241 genome-wide significant and independent variants, with PCSK9 and NOS3 at the top of the list having 7 concordant line of evidence. In addition, PoPS methodology applied to find putative genes allowed also to identify related biological mechanisms as most informative features and detected as most relevant mechanisms lipoprotein homeostasis, endothelial cell proliferation, collagen matrix formation as well as less established signaling pathways involving cell cycle.

In the context of precision medicine for CAD, polygenic risk scores represents the disease risk that is predicted at the individual level aggregating the individual's genetic variation weighted by the GWAS strength of association. PRS can differentiate and inform of low and elevated risk, for example identifying 8.0% of the population with higher than three-fold increase of CAD risk [105]. As previously mentioned, rare monogenic conditions such as familial hypercholesterolaemia interplay with the common genetic background increasing the overall CAD risk compared to non-carriers [106], nevertheless common variants still play the major part in accumulating CAD risk [123]. The putative causal genes identified from GWAS pointed to different mechanisms of actions and plausible consequential differences in endophenotypes for the disease. Indeed, CAD is an heterogeneous disease in clinical manifestations [135]. For instance, patients affected by heart failure were clustered in a unsupervised manner via cardiovascular biomarkers, leading to different clinical profiles, prognosis and response therapy [113]. Genkel et al. [8] proposed a classification of endophenotypes for CAD and the corresponding genetic connection from GWAS including LDL, hypertension, inflammation and diabetes mellitus. Although these clinical manifestations are not mutually exclusive, understanding they role in CAD etiology from a genetic point of view is essential to provide the proper therapy in a personalized manner.

In summary, recent advancement in genotyping technology and cost reduction has led to an increasing number of associations and hypotheses generation of genetic targets and disease etiology mechanisms that can help, together with the reduction of environmental risk factors, to treat and prevent such widespread and mortal disease. Nevertheless, there is a need in understanding novel pathways on which genetic effect converge, even of small sizes, regardless the well established lipid metabolism as well as the phenotypic consequences of individual genetic configuration with the final goal of assigning the proper treatment at the individual level.

2.6.2 Schizophrenia

Schizophrenia (SCZ) is a devastating disease with an average lifetime prevalence of 1% that varies for different geographic regions up to 5% [136] and charactherized by an hetereogneity of symptoms divided in positive, negative and cognitive. In particular, positive symptoms include recurrent psychosis such as hallucinations, delusions and disorganized behaviour; negative symptoms are typical of amotivational syndrome and include anhedonia, social withdrawal and reduced energy; cognitive symptoms are observed as a widespread range of congitive dysfunctions. The illness onset is usually in the adolescence phase with a typical decline of cognitive and social functions and psychotic episodes emerging later on. Individuals affected by SCZ have a life expectancy drastically reduced by 20 years compared to the overall population, with suicide being the largest contributor connected to impairments in everyday life like maintaining social relationship, employment and personal independence [137]. Anti-psychotic medications treat positive symptoms and are the first line of action, increasing the quality of life and personal independence. Nevertheless, functional behaviours are related to negative and cognitive symptoms, becoming nowadays target outcomes for drug therapy [138].
From a etiology point of view, SCZ is a complex disease arising from the interplay of both environmental (e.g. prenatal maternal conditions, paternal age, urban environment, drug addiction and childhood adversities [139]) and genetic risk factors, with an heritability estimation from twin studies around 80% [140]. Indeed, in the past years GWAS have led to great advancement in revealing the genetic architecture of this complex disease [73, 141, 142], with an increasing sample size that reached more than $300,000$ individuals in the latest study and identified 342 independent common variants associated with SCZ in 287 loci [73]. In particular, this study confirmed the strong heritability of the disease from common variants (MAF $\leq 1\%$) estimated at 24%, nevertheless still impractical at the clinical level. Indeed, PRS assessment showed a median area under the receiver operating characteristic curve of solely 0.72. On the other hand, PRS still informed on the extreme scenarios of disease risk score, with the highest percentile having an odds ratio for SCZ of 39 compared to the lowest percentile. In addition, Trubetskoy et al. leveraged gene expression data from multiple brain regions and identified an enrichment in genes with increased expression for excitatory glutamatergic neurons in cerebral cortex and

hippocampus and for cortical inhibitory interneurons. These significant associations were computed via LDSC method [53] and MAGMA [82] and became observable only in the latest GWAS as compared to the previous ones thanks to the increased sample size and hence power. Pathway enrichment analysis identified mechanisms related to neuronal functions such as synaptic transmission and cellular components such as ion channels, synapses, axon and dendritic annotations. This was also confirmed from previous less powered GWAS studies that additionally pointed to immune mechanisms and histone methylation processes being enriched in associated risk loci [143]. In the latest GWAS [73], the authors also applied a gene prioritization strategy to understand the effects of genetic associations to changes in gene expressions via SMR [101], FINEMAP [33] and chromatin conformation analysis (Hi-C) data, obtaining a putative causal set of 120 genes. A concurrently study [144], focusing on ultra-rare damaging mutation having large effect on SCZ risk and likely to disrupt a protein function, identified 32 genes comparing whole-exome sequencing of $> 110,000$ individuals among people with schizophrenia and healthy controls. Interestingly, this gene-set is enriched for common variant associations as well as genes with variants in the FINEMAP credible set, indicating a convergence between rare and common variants that are likely to disrupt shared mechanisms in SCZ [73]. These two studies hint at 4 prioritized genes in both contexts, STAG1, FAM120A, GRIN2A and SP4. The first two genes are related to regulatory mechanisms of expression and post-transcription, whereas the second two are connected to N-methyl-D-aspartate (NMDA) receptor biology. The latter represent supporting evidence for one leading hypothesis of SCZ, N-methyl-d-aspartate receptor (NMDAR) hypofunction, namely being related to the hypofunctional glutamatergic neurotransmission via NMDA receptor and connected to dysfunction of parvalbumin-positive interneurons in the cerebral cortex and hippocampus. The NMDAR hypofunction hypothesis was formulated from observations of drugs classified as NMDA receptor antagonists (e.g. ketamine) that induce SCZ-like positive symptoms and cognitive deficits [145]. The interplay between common and rare variants was also already hypothesized leveraging previous GWAS and exome-sequencing studies [146, 147]. For instance Chang et al. [147] discovered that genes having common SNPs significant for SCZ were more inclined to be connected to genes with de novo mutations via protein-protein interaction network and identified NMDA receptor interactome being connected to multiple types of genetic risk factors.

Ever since the first GWAS on SCZ was conducted [148], the strongest association reside in the major histocompatibility complex (MHC) [141]. This strong effect as well as additional ones connected to immune related genes residing outside MHC locus highlighted the relevance of inflammatory processes, in accordance with the occurrence of aberrant inflammatory mechanisms from epidemiological studies [149]. Despite the challenge of gene prioritization in MHC locus due to high LD effect, complement factor 4 genes C4A and C4B were identified as relevant in SCZ etiology due to their role in synapse elimination validated by differential expression for brain related genes [150] and in in-vivo experiments showing a reduced cortical synapse density and altered mouse behaviour [150, 151]. Nevertheless, MHC locus contribution cannot simply be reduced to complement

factor 4 genes, with plausible candidates further identified in NOTCH4 gene encoding a transmembrane protein involved in neurodevelopmental processes and connected to cognitive traits in SCZ or TRIM26 encoding for a protein of unknown function but further supported by differential gene expression in case-control setting [152].

More recently, TWAS allowed to integrate GWAS signal effect to tissue-specific expression modulation to pin-point directly affected genes [71, 153, 154]. Focusing on brain, blood and adipose tissues, Gusev et al. [71] identified 157 significant genes, 35 of which not overlapping with at the time GWAS significant loci. The authors excluded genes in MHC locus due to its complexity but specifically validated C4A gene, confirming a strong association. This approach was additionally extended to epigenetic data via the integration of imputed gene expression with chromatin marks observed in lymphoblastoid cell lines and identified 42 genes out of the 157 also associated with epigenetic changes. The showcase MAPK3 gene was validated in in-vivo zebra fish model, highlighting its involvement in neuro-proliferation. In a extended effort including 11 brain regions plus thyroid from GTEx and Dorsolateral Prefrontal Cortex (DLPC) from CommonMind Consortium reference panel, Huckins et al. [153] imputed gene expression applying prediXcan method [10] and retrieved 413 significant associations across 256 genes, of which 67 outside MHC locus were prioritized as independent associations via a stepwise forward conditional analysis and 19 representing novel target with respect to previous GWAS results. The associated genes highlighted implicated 33 molecular mechanisms identified via MAGMA [82] such as porphyrin metabolism whose dsyfunction might lead to psychiatric symptoms, hexosaminidase activity with deficiency resulting in mental problems, and Ras and Rab signaling as well as GTPase activity related to neuronal cell differentiation and migration. Owing to the developmental nature of SCZ, the 67 independent genes were further investigated for developmental expression pattern, obtaining 4 clusters of genes being expressed in four different spatiotemporal regions including early pre-natal and post-natal expressions. This supports the hypothesis that SCZ is originated in early life from the specific genetic architecture and emerge later on in adolescence possibly due to changes in cortical biology and alterations in cortical synaptic arrangement [139]. Similarly, Hall et al. [154] applied TWAS methodology developed in [9] (Fusion) in DLPC tissue from CommonMind Consortium, identifyng via MAGMA two significant pathways after FDR correction: *antigen processing and presentation of peptide antigen via MHC class I* and *Abnormal CNS synaptic transmission*, the latter including six TWAS significant genes. Interestingly, they investigated the increase in identified molecular pathways using summary statistics from GWAS instead of TWAS, reasoning that the discrepancy arise from a lower power with less genes included in TWAS (genes must be significant and explained by cis-genetic effect in TWAS compared to all genes overlapping an identified significant loci in GWAS) and a lower signal due to smaller background set considered.

The genetic association from GWASs permits also the identification of other complex traits or endophenotypes correlated with SCZ from a genetic point of view. For instance, cross-trait LD Score regression [94] was developed to compute genetic correlation solely via GWAS summary statistics and confirmed a genetic correlation of SCZ with bipolar

disorder, long hypothesized due to the shared genetic background [148], and highlighted novel hypotheses to be further investigated such as the association with anorexia nervosa. A recent study by Reay et al. [155] tested for both correlation and casual effect between 10 psychiatric diseases and the blood-based biomarker collected in UK Biobank from a genetic point of view. The genetic correlation estimated via LD Score regression detected a significant positive correlations between SCZ and multiple blood measurements such as lymphocyte count, sex hormone binding globulin and mean sphered cell volume and a negative correlation with C-reactive protein (CRP). Applying Mendelian randomization methods, the authors found a protective causal role of CRP towards SCZ, even after conditioning for IL-6 signaling and body mass index, indicating a discrepancy with observational study that require further investigation. Finally, SCZ was also negatively correlated with cognitive function [156] and performance intelligent quotient [157] from a genetic perspective, consistent with cognitive disturbances as being one of its core symptoms and regarded as independent clinical phenotype, underlying the need to develop treatment strategies that target and improve cognition impairments.
Taking into consideration the heterogeneity of SCZ in pathophysiological manifestation, treatment approaches relying on anti-psychotic medication are a reductive one-fits-all strategy and might only be appropriate for specific genetic sub-groups (however not yet identified). Thus, there is a need for retrieving genetically homogeneous cohorts with specific endophenotypic manifestations on which optimal treatment strategies can be tailored. In that direction, Ruderfer and et al. [158] considered 28 sub-phenotypes collected among SCZ and bipolar disorder (BD) patients and investigated the relationship with disease specific PRS, finding a positive correlation between BD PRS and manic symptoms in SCZ patients, BD PRS and psychotic features in BD patients as well SCZ PRS and SCZ patients with increased negative symptoms.

In summary, GWAS technology has revolutionized genetic understanding of SCZ, pointing at impaired mechanisms such as synaptic biology. Despite having identified multiple target genes, we still lack a comprehension of their roles as well as the consequences of small variant effect and their possible convergence onto other biological pathways. In addition, precision strategies targeting a genetically homogeneous sub-group that address different disease manifestations are still lacking.

3

Methods

To disentangle genetic mechanisms that lead to disrupted biological processes and reveal patient stratification with clinically relevant phenotypic manifestations, we develop a novel comprehensive pipeline for complex diseases called CASTom-iGEx available at gitlab.mpcdf.mpg.de/luciat/castom-igex. The three modules of the pipeline (Fig. 3.1) are explained in detail in this chapter divided per section.

Section 3.1: First, tissue-specific gene expression prediction models are estimated from reference panel data sets such as GTEx and CMC that are composed of matched genotype arrays and RNA-seq data. Instead of using state-of-the-art methods such as Fusion [9] and prediXcan [10], we developed a new strategy called *PriLer* that integrates variants biological annotation to improve the selection of regulatory variants.

Section 3.2: Second, gene expression is imputed on genotype-only cohorts and combined into individual level pathway-scores leveraging gene-sets and biological pathway databases such as Gene Ontology and Reactome. Afterward, TWAS (transcriptome-wide association study) and PALAS (pathway level association study) are performed to test the associations of computed genes and pathways with the disease or trait of interest. We additionally apply Mendelian Randomization (MR) methodology and use these associations to study the causal role of multiple traits and endophenotypes (e.g. LDL levels) that can contribute to the disease of interest (e.g. CAD) and identify the corresponding genes and molecular pathways mediating this contribution. The code for systematic MR analysis can be found at gitlab.mpcdf.mpg.de/luciat/castom-igex_mr.

Section 3.3: Third, patients are stratified based on imputed gene expression. The differences in clinical features and endophenotypes are investigated together with cluster-specific differences in biological pathways. When additional clinical data is not available for a certain disease (such as SCZ data from PGC cohorts), we derive a risk-score mimicking the actual endophenotype based on imputed gene expression (gene-RS) to find plausible different trajectories in comorbidities and disease characteristics.

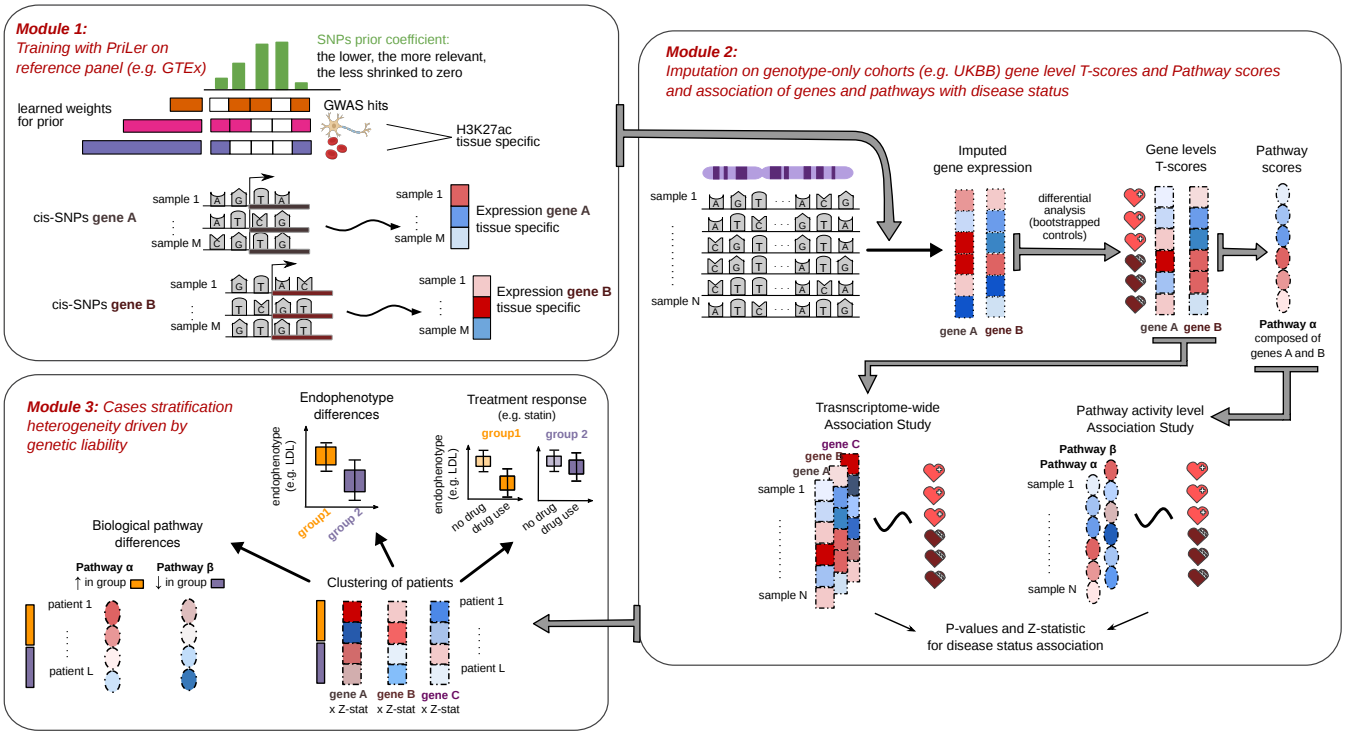

Fig. 3.1.: (From Trastulla et al., in prep.) CASTom-iGEx modules. **Module 1**: creation of gene expression prediction models via PriLer. The top colored array represent different prior information on variants, e.g. from GWAS or open chromatin tissue-specific regions. **Module 2**: inference of gene expression on genotype-only cohorts, conversion to gene T-scores, computation of pathway scores and association with trait of interest (TWAS and PALAS). **Module 3**: clustering of patients based on imputed genes followed by characterization of pathway, endophenotypic differences and differential treatment responses.

3.1 PriLer: prior learned elastic-net regression

In order to create tissue-specific gene expression prediction models from genotype data, we developed a new method called PriLer (Prior learned elastic-net regression). Similarly to previously developed methods such as Fusion [9] and prediXcan [10], PriLer estimates gene expression based on genetic cis-effects i.e. SNPs and indels surrounding a gene's transcription starting site (TSS), in a reference panel cohort composed of matching genotype and gene expression data. Our new method further integrates biological evidence of single genetic variants defined as prior, e.g. cell type specific open chromatin states or GWAS association. The relevance of these prior features in assisting the selection of tissue-specific regulatory variants is unknown a priori and it is automatically learned by the algorithm in a iterative procedure via nested cross-validation. Our new method is inspired by Lirnet algorithm [159], also learning regulatory potential of variants to gene expression (see section 3.1.5 for further discussion).

3.1.1 Problem formulation and solution

For a certain tissue, we denote with M the number of individual with matched genotype and gene expression, N the number of genes expressed for that tissue, P the number of variants (SNPs and indels) across all genome and K the number of prior features considered for that tissue. We shortly indicate with SNP_p the p^{th} variant ordered according genomic location. For n in $1, \dots, N$, let $\boldsymbol{Y}^n = (Y_1^n, \dots, Y_M^n)$ be the vector of observed expression for gene n and $X = [\boldsymbol{X}_1 | \dots | \boldsymbol{X}_P]$ the $M \times P$ genotype matrix of dosages for imputed variants with values between 0 and 2, where 0 indicates homozygous reference (REF/REF), 1 heterozygous (REF/ALT) and 2 homozygous alternative call (ALT/ALT). Generally, an entry of the matrix X referring to sample m and variant p is indicated with the notion $X_{m,p}$. Since we only model cis-effects on gene expression, we denote with $X^n = [\boldsymbol{X}_{n_1} | \dots | \boldsymbol{X}_{n_P}]$ the $M \times P_n$ genotype matrix specific for gene n with $n_1, \dots, n_P$ referring to the indexes in correspondence of gene n cis-variants.
We define as prior information $A = [\boldsymbol{A}_1 | \dots | \boldsymbol{A}_K]$ the $P \times K$ binary matrix where each column represents the intersection between a prior feature and the variants included. We derive prior features using cell-type and tissue-specific open chromatin regions from ChIP-seq H3k27ac or ATAC-seq data, from here on denoted as gene regulatory elements (GREs), and GWAS results converted to a binary format (p-value lower than a certain threshold). For instance, if prior feature k is the collection of GREs found in certain cell-type, then $A_{p,k} = 1$ indicates that the genomic position of SNP_p is located in at least one of those GREs. Full details of prior information construction are in section 4.1.1.
Finally, let $\|\cdot\|_2 : \mathbb{R}^n \longrightarrow \mathbb{R}$ be the Euclidean norm of a vector defined as $\|\boldsymbol{a}\|_2 = \sqrt{\sum_{i=1}^n a_i^2}$ and $\|\cdot\|_1 : \mathbb{R}^n \longrightarrow \mathbb{R}$ the Manhattan norm equivalent to $\|\boldsymbol{a}\|_1 = \sum_{i=1}^n |a_i|$.
PriLer is an extension of elastic-net regression that incorporates weights for variants in order to give a lower penalty to SNPs and indels that are more likely to be casual from external evidence. For each gene n, elastic-net without prior information (enet) aims at find the optimal vector of regression coefficients $(\beta_0^n, \boldsymbol{\beta}^n)$ solving the following minimization problem:

$$\min_{(\beta_0^n, \boldsymbol{\beta}^n) \in \mathbb{R}^{P_n+1}} \left[\frac{1}{2M} \|\boldsymbol{Y}^n - \beta_0^n - X^n \boldsymbol{\beta}^n\|_2^2 + \lambda_n \left(\frac{1-\alpha_n}{2} \|\boldsymbol{\beta}^n\|_2^2 + \alpha_n \|\boldsymbol{\beta}^n\|_1 \right) \right] \tag{3.1}$$

The first term minimizes the distance between the actual gene expression and the predicted one. The second term instead represent the elastic-net penalization of the regression coefficient that is built as a combination of Ridge penalty ($\|\cdot\|_2$) and LASSO penalty ($\|\cdot\|_1$). The Ridge penalty encourages highly correlated variants to be averaged exhibiting the grouping effect, while the LASSO penalty leads to a sparse solution in the coefficients of these averaged variants [63]. The elastic-net penalization term in (3.1) can be briefly expressed as

$$\sum_{p=1}^{P_n} \lambda_n \left(\frac{1-\alpha_n}{2} \beta_p^{n2} + \alpha_n |\beta_p^n| \right) =: \sum_{p=1}^{P_n} \mathcal{L}(\beta_p^n, \lambda_n, \alpha_n), \tag{3.2}$$

with $\mathcal{L}$ indicating the elastic-net penalty function for each variable. The hyper-parameters couple $\lambda_n \geq 0$ and $\alpha_n \in [0, 1]$ control the amount of shrinkage towards zero and the Ridge/LASSO contribution respectively, with the optimal configuration estimated via a 5-fold cross-validation strategy (see section 3.1.2).

Our PriLer extension of the baseline elastic-net model described in (3.1) is built on the evidence that eQTLs are enriched within promoter, open chromatin and enhancer regions of their associated genes [160]. Hence, we assume that variants carrying biological prior information such as GREs location are more likely to be involved in gene expression regulation. To facilitate the selection of these variants as regulatory, i.e. $\beta_p^n \neq 0$ for at least one gene n (denoted as reg-SNPs), in PriLer the penalty term of each variant p ($\mathcal{L}(\beta_p^n, \lambda_n, \alpha_n)$) is multiplied by a prior coefficient v_p derived as a non linear combination of prior information A:

$$v_p = 2\left(1 - \frac{1}{1 + e^{-\sum_{k=1}^{K} \gamma_k A_{p,k}}}\right) \tag{3.3}$$

with $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K) \in \mathbb{R}_+^K$ a vector prior weights associated to each prior feature. The prior coefficient v_p relieves the individual variant penalty term $\mathcal{L}(\beta_p^n, \lambda_n, \alpha_n)$ in (3.1). As shown in Fig. 3.2, v_p reaches is maximum value of 1 when there are no prior information that support SNP$_p$ as regulating any gene and hence the shrinkage should only be determined by the actual association with considered gene expression. Instead, v_p will be close to zero when the linear combination of prior features $\sum_{k=1}^{K} \gamma_k A_{p,k}$ is considerably > 0, meaning there are additional evidences that SNP$_p$ is regulatory for certain genes. In that case, the penalty term applied will be lower compared to variants without any prior information, encouraging but not forcing the selection of SNP$_p$ as regulating a gene expression, thus will be more likely an estimation of $\beta_p^n \neq 0$.

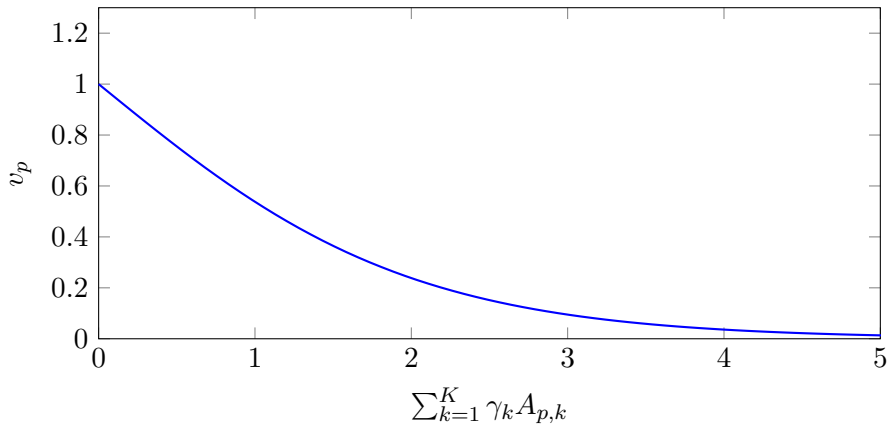

Fig. 3.2.: Prior coefficient function depending on the linear combination of prior information

Prior coefficient formulation in (3.3) depends on a sigmoid function which introduces a saturation effect so that the penalty term will smoothly and boundedly decrease to zero.

The prior weight γ_k represents the contribution of prior feature k to the overall prior coefficient, namely high γ_k indicates that the presence of feature k increases the chance of having reg-SNPs. The constrain of $\boldsymbol{\gamma}$ being non-negative reflects the assumption that prior features can only increase the overall relevance of a variant.
Prior information matrix A is provided from external references but the relevance of each feature γ_k is automatically learned by the algorithm together with the variants effect on gene expression $\boldsymbol{\beta}^n$ solving the PriLer problem:

$$\min_{\substack{(\beta_0^n,\boldsymbol{\beta}^n)\in\mathbb{R}^{P_n+1}\,\forall n;\\ \boldsymbol{\gamma}\in\mathbb{R}_+^K}} \left\{ \sum_{n=1}^{N} \left[\frac{1}{2M} \|\boldsymbol{Y}^n - \beta_0^n - X^n\boldsymbol{\beta}^n\|_2^2 + \sum_{p=1}^{P} v_p \mathcal{L}(\beta_p^n, \lambda_n, \alpha_n) \right] + E\|\gamma\|_2^2 \right\} \tag{3.4}$$

where the last term $E\|\gamma\|_2^2$ is a Ridge regularization term for prior weights to avoid unbounded increase and is controlled by hyper-parameter $E \in \mathbb{R}_+$. Note that $\boldsymbol{\gamma}$ estimation is not gene specific but consider the regulatory effect of cis-variants across all genes. Hence, the optimization problem (3.4) is an extension of (3.1) that includes individual weights for each variant, considers N genes simultaneously and iterates over all P variants, formulated by setting $\beta_p^n = 0$ for SNP_p not in TSS cis-window of gene n.

Solution of (3.4)

Suppose hyper-parameters λ_n, α_n and E are fixed. The optimization problem (3.4) is solved in a 3-step iterative procedure based on coordinate descent algorithm as shown in Fig. 3.3. Coordinate descent methods are a typical approach for solving convex problems and they work optimizing the objective function over each parameter with the others fixed and repeating cycles until convergence. In particular, for PriLer

Step **(0)** individual variant prior coefficients are computed based on (3.3), with an initial prior weights status set to zero $(\gamma_1, \ldots, \gamma_K) = (0, \ldots, 0)$;

Step **(1)** problem (3.4) is then solved with respect to $(\beta_0^n, \boldsymbol{\beta}^n)$ for each gene n separately having prior coefficients v_p with $p = 1, \ldots, P$ fixed;

Step **(2)** the same objective function is minimized with respect to γ_k for each prior feature k keeping regression coefficients $(\beta_0^n, \boldsymbol{\beta}^n)$ fixed.

These 3 steps are repeated until convergence is reached, namely the improvement in objective function's decreasing is not relevant anymore and lower than a certain threshold. In details, prior coefficients v_p for all variants P are computed based on (3.3) in Step **(0)**. Afterward, PriLer optimization problem is solved with respect to regression coefficients $(\beta_n^0, \boldsymbol{\beta}_n)$ in Step **(1)**. In particular, the optimization problem described in (3.4) reduces to N independent elastic-net problems with variant-specific penalties (prior coefficients):

$$\min_{(\beta_0^n,\boldsymbol{\beta}^n)\in\mathbb{R}^{P_n+1}} \left[\frac{1}{2M} \|\boldsymbol{Y}^n - \beta_0^n - X^n\boldsymbol{\beta}^n\|_2^2 + \sum_{p=1}^{P} v_p \mathcal{L}(\beta_p^n, \lambda_n, \alpha_n) \right] \quad \text{for } n = 1, \ldots, N \tag{3.5}$$

The N optimization problems are solved in parallel using `glmnet` R package that uses again a cyclical coordinate descent algorithm [161]. Omitting gene n notation, the optimal value of β_p is determined subsequently the estimation of $\tilde{\beta}_0$ and $\tilde{\beta}_j$ with $j \neq p$ as

$$\tilde{\beta}_p := \frac{S\left(\frac{1}{M}\sum_{m=1}^{M} X_{m,p}\left(Y_m - \tilde{Y}_m^{(-p)}\right), \lambda\alpha v_p\right)}{\lambda(1-\alpha)v_p + \frac{1}{M}\sum_{m=1}^{M} X_{m,p}^2} \tag{3.6}$$

with $S(z, w)$ the soft-thresholding operator

$$S(z,w) = \begin{cases} z - w & \text{if } z > 0 \text{ and } |z| > w \\ z + w & \text{if } z < 0 \text{ and } |z| > w \\ 0 & \text{if } |z| \leq w \end{cases}$$

and $\tilde{Y}_m^{(-p)} := \tilde{\beta}_0 + \sum_{\substack{j=1; \\ j\neq p}}^{P} X_{m,j}\tilde{\beta}_j$ the current fitted value for gene expression in sample m having excluded the contribution of $X_{m,p}$.

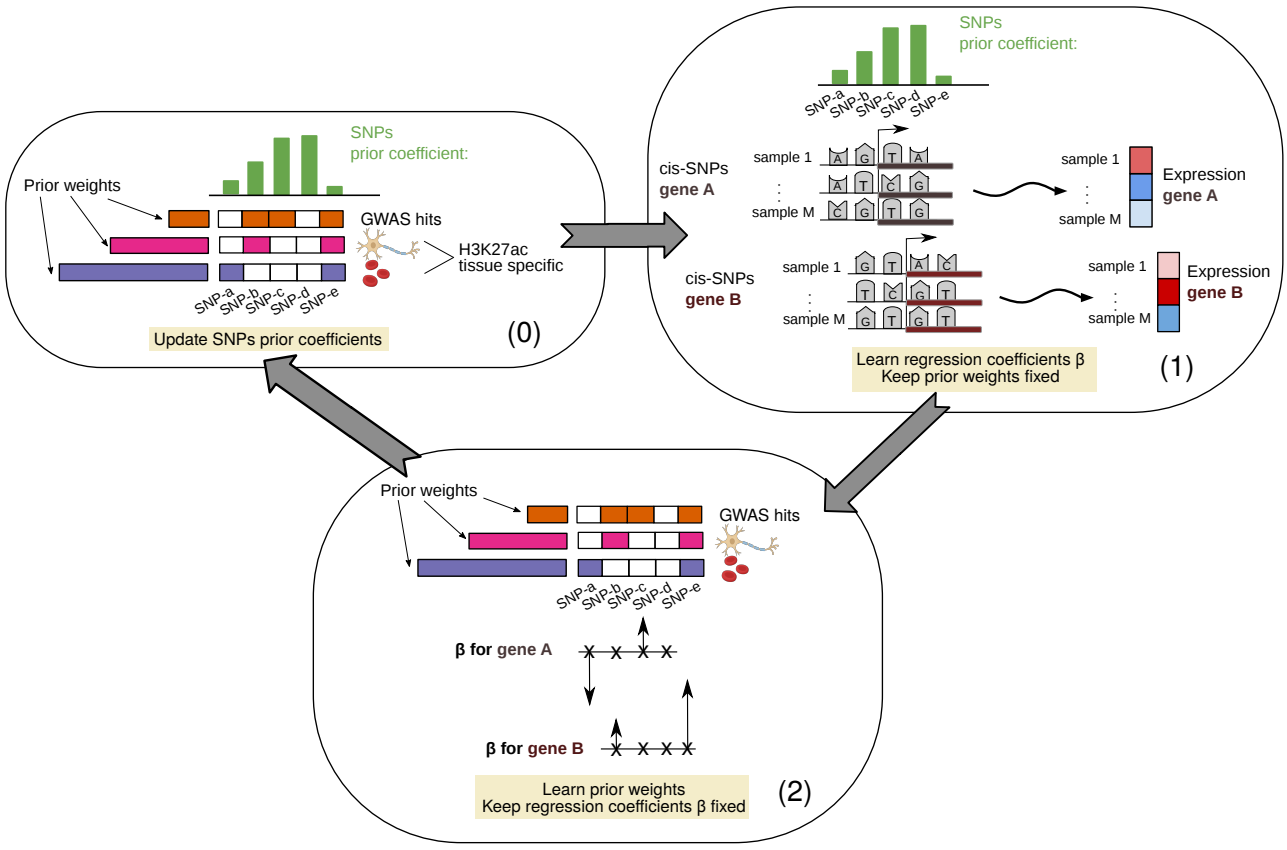

Fig. 3.3.: **Step (0)** Variants prior coefficients are initially computed as a combination of prior weights $\boldsymbol{w}$ and prior matrix A with the initial weight state set as $\mathbf{0}$. **Step (1)** Then, regression coefficients β are estimated for each gene n with variant penalty given by prior coefficient: the lower the more likely is the variant to be selected in a regression model. **Step (2)** Afterward, prior weights are updated based on the previously found regression coefficients so that weights are higher for prior features intersecting more reg-SNPs. These 3 steps are repeated until convergence is reached.

Once regression coefficients $(\beta_0^n, \boldsymbol{\beta}^n)$ are estimated, PriLer objective function is minimized with respect to prior weight vector $\boldsymbol{\gamma}$ in Step **(2)**. Hence (3.4) reduces to

$$\min_{\boldsymbol{\gamma}\in\mathbb{R}_+^K} \left\{ \sum_{n=1}^{N} \left[\sum_{p=1}^{P} v_p \mathcal{L}(\beta_p^n, \lambda_n, \alpha_n) \right] + E\|\boldsymbol{\gamma}\|_2^2 \right\} \tag{3.7}$$

This problem is solved via globally-convergent method-of-moving-asymptotes (MMA) [162] implemented in `nloptr` R package [163] with the option `"algorithm" = "NLOPT_LD_MMA"`. This type of algorithm is guaranteed to converge to some local minimum from any feasible starting point and can solve inequality-constrained nonlinear programming problems based on conservative convex separable approximations [162]. The general formulation is

$$\begin{aligned} &\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) \\ \text{subject to } &g_i(\boldsymbol{x}) \leq 0 \quad i = 1, \ldots, m \\ &x_l^{min} \leq x_l \leq x_l^{max} \quad l = 1, \ldots, n \end{aligned}$$

where $\boldsymbol{x} = (x_1, ..., x_n)$ is a real value vector of variables, x_l^{min} and x_l^{max} are given real numbers, and f, $g_1, ..., g_m$ are given real-valued functions twice continuously differentiable. In our case, we do not have additional inequality constraints (g_i functions) and $\boldsymbol{\gamma}$ vector is only upper bounded by 0 ($\gamma_k \geq 0 \, \forall k$). In order to apply this method, 1) we need to prove that the objective function in (3.7) is continuous and that the first and second derivatives exist and are continuous as well; 2) the explicit formulation for the first derivative must be provided in `nloptr` to be used as `eval_grad_f` argument. As a proof of the two previous points, let $f : \mathbb{R}^K \longrightarrow \mathbb{R}$ be the objective function in (3.7), f is continuous in $\boldsymbol{\gamma}$ being a linear combination of continuous functions, namely the squared linear norm operator and the sigmoid function in v_p (3.3). The gradient of f is a real-valued function $\nabla f : \mathbb{R}^K \longrightarrow \mathbb{R}^K$ where each component is the partial derivative $\frac{\partial f}{\partial \gamma_k}$ for $k = 1, \ldots, K$, and has the following form (see Appendix A.1 for computation)

$$\frac{\partial f}{\partial \gamma_k} = 2E\gamma_k + \sum_{n=1}^{N} \left[\sum_{p=1}^{P} A_{p,k} v_p \left(\frac{v_p}{2} - 1 \right) \mathcal{L}(\beta_p^n, \lambda_n, \alpha_n) \right]. \tag{3.8}$$

Thus, the gradient ∇f is again continuous in all its components as a product and linear combination of continuous functions. The same conclusion can be proven for the Hessian $K \times K$ matrix of the second derivative $\nabla^2 f$ in each entry $\frac{\partial^2 f}{\partial \gamma_h \partial \gamma_k}$ (see Appendix A.1).
These steps allow for an interactive change of prior coefficients based on genes regression coefficients and vice-versa. A small prior coefficient v_p implies that regression coefficients for SNP$_p$ across all N genes β_p^n will be less shrink to zero. Consequently, SNP$_p$ has a higher relevance in the overall gene expression regulation, despite not forcing its selection that will depend on specif genes. On the other hand, the weights for prior features γ_k will increase with the fraction of reg-SNPs intersecting that feature and making even more likely the selection of those SNPs as regulatory. However, if the decrease given by the mean

squared error for predicted gene expression to approximate $\boldsymbol{Y}^n$ is not convenient enough, that SNP_p will not be selected and the associated prior features weight will not increase.

Next section gives details on the implementation to solve problem (3.4) and the nested-cross validation procedure applied to find the optimal hyper-parameter configuration.

3.1.2 Implementation and hyper-parameters search

PriLer problem (3.4) is solved in $K + \sum_{n=1}^{N}(P_n + 1)$ variables keeping a total of $2N + 1$ hyper-parameters fixed: λ_n gene-specific sparsity, α_n gene-specific Ridge/LASSO contribution for $n = 1, \ldots, N$ and E limiting the variance of prior weights. In order to find an optimal space of hyper-parameters that would reduce model overfitting and best perform on external data, we apply a nested 5-fold Cross-Validation (CV) strategy [164] to find a configuration leading to minimal average Mean Squared Error (MSE) on the test-folds. As clarified below, optimal solutions for both elastic-net and PriLer are computed simultaneously, which consequentially allows to assess the improvement in integrating prior information in gene expression modelling.
In the scenario of a single-CV, M samples are split into $L = 5$ equal sized parts (folds) such that $L - 1$ folds are combined in a set called *train set* on which gene expression model is fitted, whereas the prediction error is calculated fitting this model on the remaining l^{th} fold called *test set*.

We denote with M_l and ID_l the number of samples in l^{th} fold and the corresponding sample indices, respectively. Let $\boldsymbol{X}(m, \cdot) = (X_{m,1}, \ldots, X_{m,P})$ the vector of sample m dosages referring to cis-SNPs for a certain gene and $\hat{\boldsymbol{\beta}}^{-l}$ the solution of elastic-net (3.1) or PriLer (3.4) problems having excluded the l^{th} fold. The CV estimate of prediction error for a certain gene is defined as the average across folds MSE between true and predicted gene expression, with the latter derived from the model fitted on all sample but l^{th} fold:

$$CV_{err} = \frac{1}{L}\sum_{l=1}^{L}\left[\frac{1}{M_l}\sum_{m \in ID_l}\left(Y_m - \boldsymbol{X}(m, \cdot)^T \hat{\boldsymbol{\beta}}^{-l}\right)^2\right]. \tag{3.9}$$

Since PriLer approach consider all genes simultaneously, we will refer to the overall CV error as the sum across all genes included $CV_{err}^{tot} := \sum_{n=1}^{N} CV_{err}^{n}$.
In implementing PriLer, we both use nested-CV and single-CV strategies to evaluate prediction performance and create final gene expression prediction model, a summary of the entire procedure is depicted in Fig. 3.5. Different from single-CV described above, nested-CV adds a second layer of partitioning (inner folds) to the external division (outer folds) as outlined in Fig. 3.4. At each step of the outer loop, nested-CV randomly divide the train set of $M - M_l$ samples in $J = 5$ equally sized parts creating inner folds such that a model is built on $J - 1$ inner folds and prediction is performed for the left out j^{th} part.

Usually, the inner-CV division is used to find optimal hyper-parameter so that CV_{err} (3.9) is minimum. That parameter is then used to build the model on the corresponding outer training $L-1$ folds and evaluated on the outer test fold. This procedure is repeated for each l^{th} fold in the outer loop and in order to derive CV_{err} averaging on the external folds. Generally, nested-CV is preferred to single-CV scenario in which both hyper-parameter selection and model performances are estimated since it avoids optimistically biased estimates [165]. However, PriLer is built on a nested-CV due the nested configuration of hyper-parameters (as explained below) as opposed to a grid-search that would have been unfeasible in terms of time and resources. Hence, in the nested external loop PriLer both finds the optimal E hyper-parameter and evaluate the final model. An more unbiased strategy would have required a 3-level nested-CV and a consequential increase in the minimum number of samples in the reference panel and computational time. Thus, we used a 2-level nested-CV, while remaining aware of possible biases in the final performance evaluation.

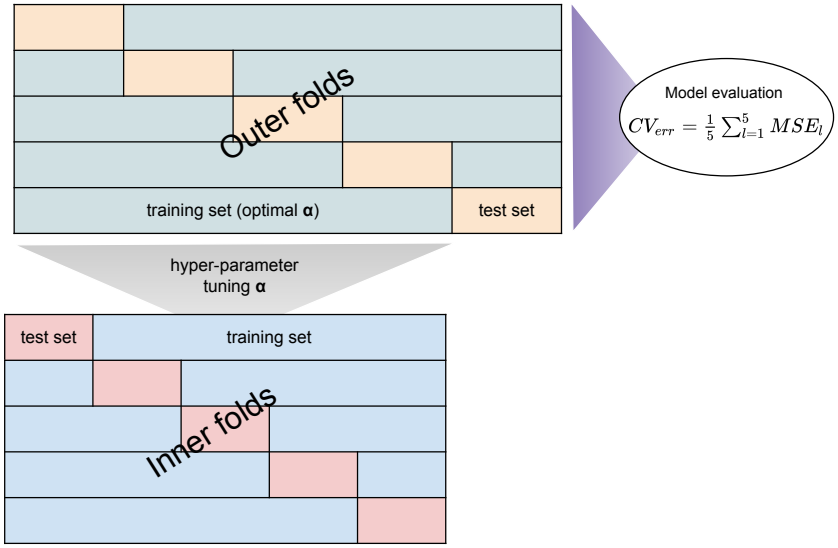

Fig. 3.4.: Nested-CV comprehends two loops, an internal one that searches for optimal hyper-parameters and an external one that based on the corresponding optimal model evaluates the result on the remaining samples (test set).

Before describing the actual strategy to search for an optimal Priler solution, we first highlight the following points.

(a) Prior weights γ_k are derived from reg-SNPs selection across all genes. It is hence critical to obtain prior weights and corresponding prior coefficients based solely on genes whose expression depends on genetic effects. We define this set of genes as cis-heritable genes. Following Fusion convention [9], cis-heritable genes are those whose proportion of gene expression variance is explained by cis-genetic variation. These gene are selected based on a likelihood ratio test implemented in GCTA software [32] that tests the genes cis-genetic heritability h^2_{cis} being equal to 0

(see section 4.2 for derivation). Nevertheless, a gene expression prediction model will be estimated for each gene (also not cis-heritable genes), leveraging the prior coefficients v_p assessed from the cis-heritable ones (see below).

(b) In `glmnet` package, the solution for each gene n of elastic-net problem (3.1) is computed for a decreasing sequence of λ_n that starts from the smallest value for λ_n such that the entire vector β^n is zero [161]:

$$\lambda_n^{max} = \max_{l=1,\dots,P} \frac{|\langle \boldsymbol{X_{n_l}}, \boldsymbol{Y}^n \rangle|}{M\alpha_n}. \tag{3.10}$$

The smallest tested value is $\lambda_n^{min} = \epsilon\lambda_n^{max}$ with ϵ equal to 0.01 or 0.0001 if $M > P_n$ or $M \leq P_n$ respectively. The search space for λ_n is then constructed as the decreasing sequence of 100 values from λ_n^{max} to λ_n^{min} in log scale:

$$\Lambda_n := \left\{ c = 99, \dots, 0 \mid \lambda_n^{min} \cdot s,\ s = e^{c\frac{log(\lambda_n^{max}) - log(\lambda_n^{min})}{99}} \right\}.$$

(c) In case of variable-specific penalties such as prior coefficients, the denominator of (3.10) is also multiplied by v_l. However, due to the interactive update of prior coefficients, the search space Λ_n would also change at each step of PriLer iteration, leading to a different hyper-parameter set that should be instead fixed for the entire iterative procedure, as an initial fixed set of hyper-parameters. To overcome this, we initially compute a fixed Λ_n space from elastic-net without any variant penalty.

(d) As shown in (3.10), Λ_n in `glmnet` depends on the choice of α_n. Thus, regularization hyper-parameters are derived sequentially, with α_n assuming 9 possible values in $(0, 1)$ interval (i.e. $0.1, 0.2, \dots, 0.9$) and $\Lambda_n(\alpha_n)$ path consequentially derived. We excluded α_n extreme values 0 and 1 as the first corresponds to LASSO penalization that would only selects a variant among correlated ones and the latter corresponds to Ridge penalization that does not directly set any variable to zero.

(e) In elastic-net, $\alpha_n, \Lambda_n(\alpha_n)$ search space are independent for each gene n, contrary to PriLer in which genes are considered simultaneously. Suppose the number of $E \geq 0$ parameters tested is C, to perform a hyper-parameter search that consider all the potential combinations, we would need to evaluate $C \cdot (9 \cdot 100)^N$ possible configurations, which is unfeasible for computational time and resources even only including around 3000 cis-heritable genes. Instead, we find the optimal combination of $\hat{\alpha}_n$, $\hat{\lambda}_n(\hat{\alpha}_n)$ for elastic-net (3.1) separately for each gene n from a total of $9 \cdot 100 \cdot N$ tests that can be parallelized over N. These optimal values are then directly used in PriLer that now need only to be optimized over the possible E parameter values.

We will now explain the pipeline in detail as depicted in Fig. 3.5. This is divided in 4 steps that correspond to R implementation in castom-igex/Software/model_training. We indicate with $\hat{\cdot}$ notation the optimal hyper-paramter choice and N_h the number of cis-heritable genes.

STEP 1 Elastic-net regression without prior is built in a nested-CV setting for each cis-heritable gene, with both inner and outer loop having a 5-fold partition. This step is used for two purposes. The first is to find optimal $\hat{\alpha}_n$, $\hat{\lambda}_n$ configuration for each gene n in the inner loop. The results are thus specific for each outer fold ($2 \times N_h \times 5$ tensor) and will be applied in PriLer step as mentioned in point (c). The second is to evaluate elastic-net regression based on average coefficient of determination R^2_{CV} on outer test folds (R^2 derivation section). As previously explained in point (b) and (d), $\hat{\lambda}_n$ is chosen among Λ_n sequence that depends on α_n initial choice. Hence, for each value of $\alpha_n \in \{0.1, \dots, 0.9\}$, we first find $\hat{\lambda}_n(\alpha_n)$ such that $CV^n_{err}(\hat{\lambda}_n(\alpha_n)) = CV^n_{err}(\hat{\lambda}_n(0.1)), \dots CV^n_{err}(\hat{\lambda}_n(0.9))$ is minimum in each entry and $\hat{\alpha}_n$ as the one for which $CV^n_{err}(\hat{\lambda}_n(\hat{\alpha}_n))$ reaches the lowest value. Practically, this is achieved using `cv.glmnet` function for each possible value of α.
From an implementation point of view, step 1 can be parallelized on two levels: chromosome-wise and gene-wise with the actual time depending on the number of cores provided and the number of genes/samples.

STEP 2 PriLer is applied combining all cis-heritable genes in the same outer loop division elastic-net was built. Optimal hyper-parameters $\hat{\alpha}_n$, $\hat{\lambda}_n$ derived from previous step are used in (3.4), leaving only E parameter to be customized. This is reached solving PriLer problem for each chosen value of $E = \{e_1, \dots, e_C\} \in \mathbb{R}^C_+$ in all the 5 outer training sets and computing the average MSE on the test folds. The optimal $\hat{E}$ parameter is the one minimizing CV^{tot}_{err}. This step is additionally performed to evaluate PriLer performance on the same sample division of elastic-net based on average test R^2_{CV} and MSE, as well as obtain for each outer fold the model prior coefficients v_p (3.3) to be directly used for not cis-heritable genes. In this case, we use a single-CV configuration (that has the same sample division of the outer fold for the previous nested-CV) to both find the optimal hyper-parameter E and evaluate PriLer performances with a possible increase of biases in the actual error estimation [165]. As previously mentioned, the creation of an additional layer into the nested structure, although possible, would be unfeasible due to a decrease in the sample size for the model training and an increase in computational time. From an implementation point of view, this step can be parallelized with respect to E possible values.

STEP 3 A final model for elastic-net and PriLer for all cis-heritable genes is built using all M samples. Since in step 1 the optimal α_n, λ_n where specific for nested-CV strategy, we now need to find optimal hyper-parameters for this enlarged set of samples via single-CV. Similarly to step 1, $\hat{\alpha}_n$, $\hat{\lambda}_n$ are then used for PriLer together with the optimal E parameter found in step 2. In addition, an overall R^2 is computed based on all samples for both elastic-net and PriLer that underlies the general performance. Finally, prior coefficients v_p are stored to be directly used as prior for not cis-heritable genes. This step can be parallelized over N_h genes.

STEP 4 Lastly, the entire procedure is repeated for $N - N_h$ not cis-heritable genes, with

α_n, λ_n parameters still customized for each gene but prior coefficients v_p fixed and already computed from step 2 and 3. First, a total gene expression prediction model is built for both elastic-net and PriLer as explained in step 3 (single-CV) and performance is overall evaluated based on R^2. Then, nested-CV is repeated both for elastic-net and PriLer (with fixed prior coefficients) to evaluate prediction on external folds based on R^2_{CV}.

In summary, with our implementation we solve PriLer problem in (3.4) for all genes N having cis-variants in a predefined window (default is 200kb window as maximum distance from TSS), creating gene expression prediction models that are optimized with respect to hyper-parameters. Furthermore, an equivalent elastic-net regression model for each gene is also built to investigate the differences in performances and model selection when introducing prior information on variants.

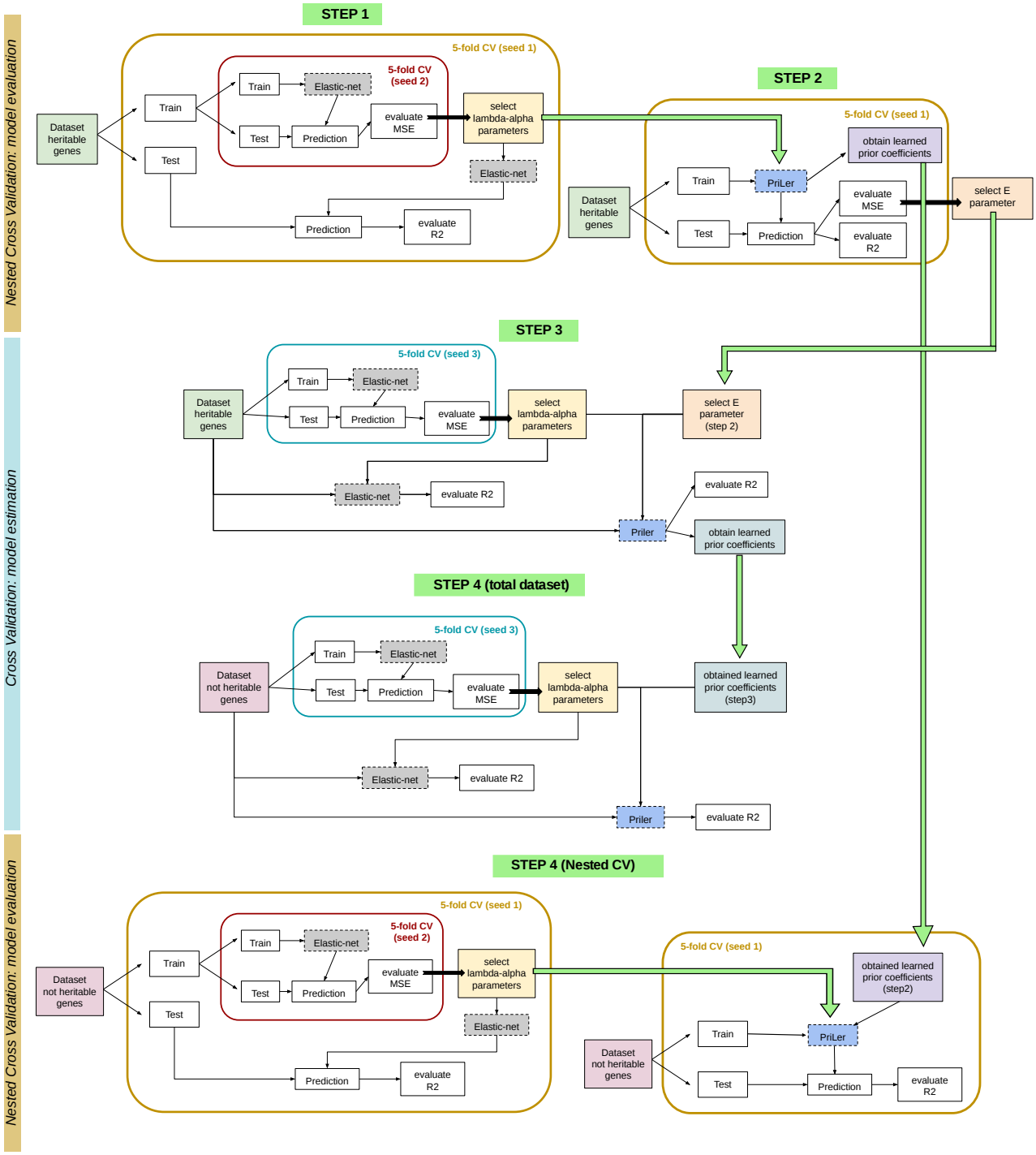

Fig. 3.5.: (From Trastulla et al., in prep.) PriLer implementation is divided in 4 different parts. In step1 elastic-net for only cis-heritable genes is performed in a nested-CV setting. The optimal α_n,λ_n gene-specific combinations found in step 1 are used in step 2 to build PriLer model, with the same outer fold from nested-CV configuration considered to find the optimal E parameter. In step 3, a single-CV is applied for elastic-net to find the optimal α_n,λ_n selection for the entire set of samples and the same pairs together with the optimal E parameter are used to build PriLer on the entire set. Finally, step4 repeats step 1 to 3 for not cis-heritable genes, with fixed prior coefficients previously derived.

3.1.3 Additive confounder effects

Until now, we took into consideration only genetic effects modeling gene expression. As in a eQTL analysis, when modeling this interaction is important to correct for possible confounders such as hidden batch effects, ancestry differences and sex information [166]. Namely, being Y the response variable of gene expression, X genotype matrix and Z covariate matrix, accounting for confounders effects on genotype-gene expression interactions means to model

$$\boldsymbol{Y} = \beta_0 + \boldsymbol{\beta} X + \boldsymbol{\mu} Z + \boldsymbol{\epsilon} \tag{3.11}$$

with $\boldsymbol{\epsilon}$ error vector. One option that is sometimes considered is to regress out these confounders from the response variable and use the residuals as a new response. However, this version is equivalent to solving (3.11) only in the situation of genetic and covariate effects being orthogonal which is highly unlikely when ancestry derived from genotype data are among the covariates. For this reason, we implement in Priler the possibility to model confounder and genetic effects to gene expression together, assuming linear interaction. Let Z be the $M \times D$ matrix of D covariates for each individual and unique to all genes. We model gene expression as a linear combination of both genetic and confounder effects, transforming Priler problem in (3.4) as

$$\min_{\substack{(\beta_0^n,\boldsymbol{\beta}^n)\in\mathbb{R}^{P_n+1}\,\forall n;\\ \boldsymbol{\mu}^n\in\mathbb{R}^D\,\forall n;\\ \gamma\in\mathbb{R}_+^K}} \left\{ \sum_{n=1}^{N} \left[\frac{1}{2M} \|\boldsymbol{Y}^n - \beta_0^n - X^n\boldsymbol{\beta}^n - Z\boldsymbol{\mu}^n\|_2^2 + \sum_{p=1}^{P} v_p \mathcal{L}(\beta_p^n, \lambda_n, \alpha_n) \right] + E\|\gamma\|_2^2 \right\} \tag{3.12}$$

Note that we do not add additional penalty term for confounders in (3.12), forcing the regression model to approximate the corresponding regression coefficients different than zero and only shrinking to null the effects referring to genotype. This is practically achieved with the `penalty.factor` term in `glmnet` `R` package and setting to zero those in correspondence of confounders. Hence, the only difference with respect to previous formulation in (3.4) is the gene expression approximation error term.

3.1.4 Performance estimation

For model performance estimation, we utilize the coefficient of determination R^2 equivalent to `dev.ratio` from `glmnet` output, indicating the fraction of deviance explained by the model. For a certain gene, R^2 can be expressed as

$$R^2 = 1 - \frac{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|_2^2}{\|\boldsymbol{Y} - \overline{Y}\|_2^2} = 1 - \frac{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|_2^2}{M\sigma_{\boldsymbol{Y}}^2} \tag{3.13}$$

where we dropped gene notation, indicated with $\overline{Y} := \frac{1}{M}\sum_{m=1}^{M} Y_m$ the mean of gene expression, with $\hat{\boldsymbol{Y}} := \hat{\beta}_0 + X\hat{\boldsymbol{\beta}} + Z\hat{\boldsymbol{\mu}}$ the imputed gene expression by the model and with $\sigma^2_{\boldsymbol{Y}} = \frac{1}{M}\sum_{i=1}^{M}(Y_i - \overline{Y})^2$ the population variance of $\boldsymbol{Y}$. Note that, when predicting on the same samples the model was estimated (in-sample prediction) as in step 3 and 4 (Fig. 3.5), R^2 is bounded between 0 and 1 with 1 reached when the real and predicted gene expression is almost identical and 0 achieved in case the best estimate is actually the average expression value and there is no advantage in introducing genotype nor confounders to model genes.
Since we are mostly interested by the effect of genotype on gene expression, we decompose R^2 in (3.13) in three parts, in order to differentiate between solely genotype contribution, confounders contribution or interaction of these two components to explained variance. We define with $\widehat{\boldsymbol{W}} := X\hat{\boldsymbol{\beta}}$ the predicted gene expression based solely on genotype effects, with $\boldsymbol{W} := \boldsymbol{Y} - \hat{\beta}_0 - Z\hat{\boldsymbol{\mu}}$ the gene expression removed of the confounder effects and supposedly carrying only the genotype contribution and with $\widehat{V} := Z\hat{\boldsymbol{\mu}}$ the imputed gene expression due to confounders. We also indicate with $\overline{\boldsymbol{W}}$ and $\overline{\widehat{\boldsymbol{V}}}$ the sample mean of vectors $\boldsymbol{W}$ and $\widehat{\boldsymbol{V}}$ respectively. It is possible to decompose R^2 in (3.13) as

$$
\begin{aligned}
& R^2 = R^2_g + R^2_c + R^2_{g,c} \\
\text{with} \quad & R^2_g := \frac{\|\widehat{\boldsymbol{W}} - \overline{W}\|_2^2 + 2\langle \boldsymbol{W} - \widehat{\boldsymbol{W}}, \widehat{\boldsymbol{W}} - \overline{W}\rangle}{M\sigma^2_{\boldsymbol{Y}}} \\
& R^2_c := \frac{\|\widehat{\boldsymbol{V}} - \overline{\widehat{V}}\|_2^2}{M\sigma^2_{\boldsymbol{Y}}} \\
& R^2_{g,c} := \frac{2\langle \boldsymbol{W} - \overline{W}, \widehat{\boldsymbol{V}} - \overline{\widehat{V}}\,\rangle}{M\sigma^2_{\boldsymbol{Y}}}
\end{aligned}
\tag{3.14}
$$

where $\langle\cdot,\cdot\rangle : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$ is the Euclidean inner product defined as $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \sum_{i=1}^{n} x_i y_i$. The explicit derivation of (3.14) is in Appendix A.2. Thus, we deconstruct R^2 in three parts representing the variance in gene expression that is due to genetic contribution (R^2_g), the one related to confounders (R^2_c) and the variance that is due to the joint effect of confounders and genetic plus any other contributor of $\boldsymbol{Y}$ not acknowledged ($R^2_{g,c}$). Note that none of the component is bound to be lower than 1 and only R^2_c is higher or equal than 0, however the sum is restricted in the interval $[0, 1]$ for in-sample estimation.

In PriLer pipeline, we use R^2 both to measure the in-sample and out-of-sample model performances, the latter referring to a prediction on external data that was not considered to build gene expression regression model. This particularly applies for cross-validation strategies, indicated with R^2_{test}. The average estimation on L test folds is constructed as (see also (3.9))

$$
R^2_{test} = \frac{1}{L}\sum_{l=1}^{L}\left[1 - \frac{\sum_{m\in ID_l}\left(Y_m - \hat{\beta}_0^{-l} - \boldsymbol{X}(m,\cdot)^T\hat{\boldsymbol{\beta}}^{-l} - \boldsymbol{Z}(m,\cdot)^T\hat{\boldsymbol{\mu}}^{-l}\right)^2}{\sum_{m\in ID_l}\left(Y_m - \overline{Y}^l\right)}\right] \tag{3.15}
$$

with $\overline{Y}^l = \frac{1}{|ID_l|}\sum_{m\in ID_l} Y_m$ the sample mean for test fold l and $|ID_l|$ the cardinality of corresponding fold l. In this case, R^2_{test} is not bounded to $[0,1]$ and can be also negative when the estimates from the fold average value is better (e.g. lower Euclidean distance) than the model prediction built on all the other sample but that fold. Similarly to in-sample performance, it is possible to estimate R^2_g, R^2_c and $R^2_{g,c}$ as average across fold performance following (3.14), with the regression model built on all the samples but test fold and prediction computed on that test fold.
Finally, we also use Pearson correlation other than R^2 to evaluate model performances, in particular when comparing PriLer with previously developed methods (see section 4.2.4). Focusing on the genetic contribution to gene expression, we define the correlation between predicted and adjusted gene expression as

$$corr := \frac{\left\langle \boldsymbol{W} - \overline{W}, \widehat{\boldsymbol{W}} - \overline{\widehat{W}} \right\rangle}{\left\| \boldsymbol{W} - \overline{W} \right\|_2 \left\| \widehat{\boldsymbol{W}} - \overline{\widehat{W}} \right\|_2} \tag{3.16}$$

which can also be estimated in a CV setting (indicated as $corr_{test}$) as average for each model fold similarly to R^2_{test}, or concatenating all predictions on test folds together to cover all M samples (referred as $corr^c_{test}$).

3.1.5 Discussion

The first module of CASTom-iGEx pipeline is a novel method to predict gene expression from genotype data that integrates biological prior information on variants. The relevance of these biological priors is automatically learned and takes into account additive effects from hidden and explicit confounders, formulated in a comprehensive pipeline that includes hyper-parameter optimization to avoid overfitting.
It is worth noting that we assume prior information matrix A to be binary, nevertheless prior can be continuous (but non-negative), representing different degrees of importance for a certain variant specific information. In this case, an initial scaling for prior is necessary to uniform the starting point of the algorithm with respect to prior features. Furthermore, a prior feature that overlaps with a considerable number of variants will be assigned a high prior weight simply because thosr variants will intersects more reg-SNPs by coincidence. Nonetheless, when a prior feature intersects many variants but is not actually relevant for the considered tissue-regression model, the corresponding weight remains stable during the iterative procedure (Fig. 3.3) as will be shown in details in section 4.2.
The approach of adding individual feature penalties (in our case variant prior coefficient) in a shrinkage regression setting has been long investigated. For instance, an extension of LASSO with adaptive weights was proposed in [167] and penalizes coefficients in the L1 penalty (Manhattan Norm). The new method called adaptive LASSO chooses weights in a data-dependent way as the reciprocal of ordinary least square solutions and it enjoys the oracle property i.e. performing as good as if the true underlying model was previously

provided.
Our new methodology PriLer is inspired by the Lirnet algorithm described in [159] that learns the regulatory potential of an individual SNP together with the regulatory network from gene expression and matched genetic data. Different from Lirnet which was applied to maximum 112 samples, our method is adapted to large reference data of combined gene expression and genotype. The maximum number of samples we tested is around 600 but potentially PriLer usage is still feasible for an higher number, depending on the available computational resources. Compared to Lirnet, we also simplify the prior coefficient formula using only individual prior for each variant without summing up the contribution of a certain region and not based on the gene regulation information. In addition, the problem (3.4) compared to Lirnet considers an individual penalty for both Ridge and LASSO terms instead of only for LASSO part and provides gene-specific elastic-net hyper-parameters λ_n, α_n optimization rather than forcing the same sparsity across all genes.
Finally, a recent method that also integrates prior information and works well in a large cohort scenario is EpiXcan [69]. EpiXcan also uses tissue-specific epigenetic information from ROADMAP to derive variant relevance and applies variant weights in elastic-net regression. The prior coefficient for each variant in a tissue is computed using the qtlBHM method [168], a Bayesian hierarchical model that measures SNP causality and incorporates eQTL and variant annotation. Afterward, prior are converted to penalty factors based on best model performance improvement with respect to prediXcan and a subset of genes that is representative of a certain variance explained level. Instead, in PriLer priors are initially defined from epigenetic (or other relevant) information and subsequently adapted in the iterative procedure based on elastic-net results, instead of external eQTL as in EpiXcan. Hence, prior feature relevance and consequently prior coefficients are learned iteratively by the algorithm itself, while estimating joint genotype causality to gene expression.

3.2 Transcriptome-wide association studies and Pathway activity level studies

Once tissue-specific gene expression prediction models are built via Priler on matched reference panels, the next step of CASTom-iGEx is to impute gene expression on cohorts having only genotyping data available, subsequently agglomerate this information to pathway level and test for association with a certain disease or trait of interest (Fig. 3.1 Module 2). Gene expression can be inferred on large cohorts (e.g. UK Biobank [17]) or smaller multiple cohorts (e.g. PGC cohorts) followed by a meta-analysis to summarize the overall associations. Hence, the TWAS analysis performed after gene imputation becomes similar to GWAS in term of test applied to discover associated genetic features (in case of TWAS genes), nevertheless the number of tests go from million when testing variants to only thousands when testing genes, drastically reducing the multiple testing burden. Furthermore, we move forward the gene expression imputation and associations and

aim at integrating the single SNP information to meaningful entities such as biological pathways, from an individual level point of view. The individual level information on pathways give the possibility to perform pathway-wide associations (PALAS), similarly to TWAS and GWAS, with the advantage of investigating the aggregation of small effects variants, impossible to detect via a p-value threshold filtering approach.
Because genes and pathways can be tested across multiple phenotypes, we finally use TWAS and PALAS summary statistics to deconstruct endophenotypes contribution to complex disease etiology through shared genes and molecular pathways via Mendelian Randomization approach.

3.2.1 Conversion of imputed gene expression into gene T-scores and pathway-scores

First of all, it is necessary to identify the set of genes whose variance can be explained by the corresponding cis-variants in PriLer. Let $\mathcal{G} = \{g_1, \dots, g_N\}$ be the set of genes expressed in a tissue for which a PriLer model is built, we consider the decomposition of R^2 in (3.14) and the corresponding definition in CV setting (3.15) and define the set of reliable genes $\mathcal{G}_{rel}$ as those satisfying

$$\mathcal{G}_{rel} = \left\{ g \in \mathcal{G} \,|\, R^2_g \geq 0.01 \text{ and } R^2_g(CV) > 0 \right\} \tag{3.17}$$

i.e. genes whose variance explained by genetic component is detectable in the final model (built on all samples available) and is not zero in the external CV validation. The final number of reliable genes per tissue depends mostly on the number of available samples in the reference panel to build PriLer model but can be influenced also by the number of included prior features (see section 4.2).
Suppose $\tilde{X}$ is a $\tilde{M} \times P$ matrix of variant dosages for $\tilde{M}$ new individuals in the genotype-only data set and P SNPs matching the ones available in reference panel. The common set of SNPs that have the same REF/ALT annotation and similar allele frequency is identified during pre-processing for PriLer model (see section 4.1), such that gene expression models are built on the same set of variants that will be used for imputation. For a reliable gene g in a tissue, let $\hat{\boldsymbol{\beta}}^g$ be the PriLer coefficients estimated solving (3.12) and associated with P_g cis-variants, then the **imputed gene expression** on $\tilde{M}$ new individuals is computed as

$$\widetilde{\boldsymbol{W}}^g := \tilde{X}^g \hat{\boldsymbol{\beta}}^g \quad \text{with } g \in \mathcal{G}_{rel} \tag{3.18}$$

Note that we reduce at minimum confounders effects since the contribution of cis-variants on gene expression is performed accounting for covariates information hence adjusting for those. The computed $\widetilde{\boldsymbol{W}}^g$ will be different from a null vector because we only consider reliable genes (3.17) and the genotype-only data includes the same set of variants used to estimate the model. Nevertheless, the variability of the imputed expression can deviate greatly among genes depending on R^2 estimates. Hence, to test for association of a gene

with a certain trait as well as to subsequently compute pathway-scores, we converted imputed gene expression to **gene T-scores** for each individual (briefly T-scores). T-scores are based on the differences in distribution between a bootstrapped reference set (usually a subset controls for the trait of interest) and all the other samples, thus enhancing the possible differences between cases and controls and rescaling genes in a similar space, regardless the original explained R^2. We compute genes T-scores using two different strategies depending on the cohort sample size for computational feasibility: if sample size is lower than $10,000$ (e.g small multiple cohorts in CARDIoGRAM or PGC-SCZ), T-scores are computed via moderate t-statistic from `eBays` function in `limma` `R` package [169]; otherwise (e.g. large-genomic single cohort UK Biobank) a ordinary t-statistic is computed being more computationally efficient since it does not require to estimate prior for variance from the overall gene distribution.
In both cases, suppose the trait of interest is binary with $\tilde{M}_0$ controls and $\tilde{M}_1$ cases in a certain cohort made of $\tilde{M}$ samples. Controls are bootstrapped L times to randomly select a reference sets for each repetition $l = 1 \dots, L$ composed of $Q\%$ of $\tilde{M}_0$ controls indicated with $\mathcal{S}^l_{ref}$. The remaining control samples together with all the cases form instead a comparison set for each repetition $\mathcal{S}^l_{comp}$ of dimension $\tilde{M}_1 + \frac{100-Q}{100}\tilde{M}_0$. The computation of T-scores is based on the following algorithm:

Algorithm 1 Gene T-scores computation

Ensure: $\tilde{M}_0$ n. of controls, $\tilde{M}_1$ n. of cases
  for $l = 1, \dots, L$ bootstrap repetition **do**
    $Q\%$ of $\tilde{M}_0$ controls as reference $\mathcal{S}^l_{ref}$ set;
    remaining $(100 - Q)\%$ of $\tilde{M}_0$ controls plus $\tilde{M}_1$ cases as comparison $\mathcal{S}^l_{comp}$ set;
    for $s \in \mathcal{S}^l_{comp}$ **do**
      for $g \in \mathcal{G}_{rel}$ **do**
        in gene g compute t-statistic of sample s versus all samples in $\mathcal{S}^l_{ref}$: $T^l(s, g)$
      end for
    end for
  end for
  Average scores across repetition $T(s, g) \leftarrow \frac{1}{L}\sum_{l=1}^{L} T^l(s, g)$

The output is a matrix T of dimensions $M \times |\mathcal{G}_{rel}|$, having in each column the converted T-scores across all samples for a certain gene. With this strategy, T-scores are also computed for controls since in each bootstrap iteration a certain amount of controls is included in the comparison set. We set as default number of repetition $L = 40$ and the percentage of retained controls for the reference set $Q = 80$, specifics of each analysis are described in section 4.1.
The central step in algorithm 1 differs depending on the magnitude of $\tilde{M}$:

- **Case $\tilde{M} \leq 10,000$:** T-scores are computed as moderate t-statistics via `limma` package comparing each sample $s \in \mathcal{S}^l_{comp}$ with all individuals in $\mathcal{S}^l_{ref}$. The derivation of moderate t-statistic is explained in detail in [170]. It was originally developed to detect differentially expressed genes in designed micro-array experiments with

arbitrary number of treatments and uses an empirical Bayes approach to shrinkage the estimation of sample variances to a pooled estimates based on the overall gene distribution, leading to more stable results. In our case, the derivation of T-scores from moderate t-statistic is achieved via the following steps:

1. a binary design $\tilde{M} \times |\mathcal{S}^l_{comp}| + 1$ matrix (De) is created with the first column indicating all the samples in the reference set $\mathcal{S}^l_{ref}$ and any other column pointing to each sample in the comparison set $\mathcal{S}^l_{comp}$.
2. `lmFit` function is applied to the overall imputed gene expression matrix $\tilde{M} \times |\mathcal{G}_{rel}|$ that fits for each given gene a linear model based on the aforementioned design matrix. This step assumes $E(\widetilde{\boldsymbol{W}}_g) = De\boldsymbol{\alpha}_g$ and the linear model estimates the regression coefficients $\hat{\boldsymbol{\alpha}}_g$, the sample variance s^2_g as approximation of residual variance σ^2_g and obtain the covariance matrices as $var(\hat{\boldsymbol{\alpha}}_g) = s^2_g(De^T De)^{-1}$.
3. The coefficients for a specific contrast are extracted via `contrasts.fit` function with contrasts being a vector of length $|\mathcal{S}^l_{comp}|$ and each entry indicating the difference between a sample in the comparison set and all samples in the reference set. Practically, the contrast matrix considered C is of dimension $\tilde{M} \times |\mathcal{S}^l_{comp}|$ with each column being 1 in correspondence of the considered sample $s \in \mathcal{S}^l_{comp}$, -1 for all the samples in $\mathcal{S}^l_{ref}$ and 0 otherwise. The coefficients for a specific contrast (i.e. sample s difference to the reference set) are extracted as $\hat{\boldsymbol{\beta}}_g = C^T\hat{\boldsymbol{\alpha}}_g$ and the estimated covariance matrix is $var(\hat{\boldsymbol{\beta}}_g) = s^2_g C^T(De^T De)^{-1}C$.
4. `eBayes` function is applied to compute moderate t-statistics for a contrast referring to a sample s and a gene g as

$$T(s,g) = \frac{\hat{\beta}_{gs}}{\tilde{s}_g\sqrt{v_{gs}}} \qquad \text{with} \quad \tilde{s}^2_g = \frac{d_0 s^2_0 + d_g s^2_g}{d_0 + d_g} \tag{3.19}$$

   where v_{gs} are the diagonal entry of $C^T(De^T De)^{-1}C$ matrix also referred to unscaled standard deviations and $\tilde{s}^2_g$ being a modification of the sample variance that assumes an inverse χ^2 prior distribution for the σ^2_g with mean s^2_0 and degrees of freedom d_0. $\tilde{s}^2_g$ is then the posterior values for the residual variances where d_g is the residual degrees of freedom for gene g (in our case $\tilde{M} - |\mathcal{S}^l_{comp}| - 1$). The hyper-parameters s^2_0 and d_0 are estimated based on residual sample variances from all genes (see [170] for details).

- **Case $\tilde{M} > 10,000$:** Let R_l bet the cardinality of reference set $\mathcal{S}^l_{ref}$. For a gene $g \in \mathcal{G}_{rel}$ we denote with μ^g_l and σ^g_l the mean and standard deviation respectively of

imputed gene expression for all samples in $\mathcal{S}^l_{ref}$:

$$\mu^g_l = \frac{1}{R_l} \sum_{r \in \mathcal{S}^l_{ref}} \widetilde{W}^g_r$$
$$\sigma^g_l = \sqrt{\frac{1}{R_l - 1} \sum_{r \in \mathcal{S}^l_{ref}} \left(\widetilde{W}^g_r - \mu^g_l\right)}$$

T-score for a sample $s \in \mathcal{S}^l_{comp}$ and gene g is defined as

$$T^l(s, g) = \frac{\widetilde{W}^g_s - \mu^g_l}{\sigma^g_l / \sqrt{R_l}} \tag{3.20}$$

From an implementation point of view, we used a nested parallelization for this step, parallelizing externally over genes in $\mathcal{G}_{rel}$ (since the combined information is not required) and internally over individuals in $\mathcal{S}^l_{comp}$, making it particularly efficient for UK Biobank data set composed of $\sim 340,000$ individuals.

Regardless $\tilde{M}$ magnitude, gene T-score gives the information of how much a sample s in comparison set is deviating from the overall distribution of the samples in reference set. Of note, the usage of contrasts to indicate a difference with samples from a control group (i.e. $\mathcal{S}^l_{ref}$ obtained as a subset of non-affected individuals for a certain trait) is not mandatory, although preferred to enhance differences connected to disease presence. Indeed the entire methodology can be extended to a scenario of no case/control separation (e.g. a data set with deep phenotyping as UK Biobank) in which reference/comparison separation would be randomly defined from all the individuals. In addition, T-scores computation with known phenotype can be extended to continuous trait dividing in individuals in having "low" values and "high" values (see section 3.2.4).

Finally, to derive individual level **pathway-scores**, suppose DB_{Pa} is a database of pathways where each entry is a group of genes being part or having a role in a meaningful biological entity. We used as default pathway databases Reactome [75] and Gene Ontology (GO) [76] collections (see section 2.3.1).
For each pathway $\mathcal{P}$ in a collection, individual level pathway-scores are derived for all samples as the mean across gene T-scores for genes that belong to a certain pathway and are reliably imputed via PriLer. Namely, let $n_{\mathcal{P}}$ be the number of genes belonging to pathway $\mathcal{P}$ and $n_{\mathcal{P},\mathcal{G}}$ the number of genes also being reliably predicted in the tissue considered ($n_{\mathcal{P},\mathcal{G}} \leq n_{\mathcal{P}}$), we then define

$$PaSc(m, \mathcal{P}) = \frac{1}{|\mathcal{P} \bigcap \mathcal{G}_{rel}|} \sum_{g \in \mathcal{P} \bigcap \mathcal{G}_{rel}} T(m, g)$$
$$\text{with } m = 1, \ldots, \tilde{M} \tag{3.21}$$
$$\mathcal{P} \in DB_{Pa} \text{ and } \mathcal{P} \bigcap \mathcal{G}_{rel} \neq \varnothing$$

Note that no filtering is required for the inclusion of genes such as a certain significance level of association with the phenotype of interest. However, to obtain a pathway score we require that the gene-set is built on more than 2 genes ($n_{\mathcal{P},\mathcal{G}} \geq 2$), meaning pathways do not need to be complete with $n_{\mathcal{P},\mathcal{G}} = n_{\mathcal{P}}$ to have an associated a score. Additionally, we avoid redundant pathways derived from the exact same genes ($n_{\mathcal{P},\mathcal{G}} = n_{\mathcal{Q},\mathcal{G}}$) by selecting the ones that overall are built on the lower amount of genes, i.e. if $n_{\mathcal{P}} < n_{\mathcal{Q}}$ then pathway $\mathcal{P}$ is retained. In this way, a gene-set more specific and closer to the complete information is retained, being $\frac{n_{\mathcal{P},\mathcal{G}}}{n_{\mathcal{P}}} > \frac{n_{\mathcal{Q},\mathcal{G}}}{n_{\mathcal{Q}}}$. The output of (3.21) is then a matrix of dimensions $\tilde{M} \times \widetilde{DB}$ with $\widetilde{DB}$ being the final set of pathways that satisfy the aforementioned conditions.
Pathway-scores can be completely agnostic to the trait they will be tested, in case T-scores are computed without the cases-controls information or they are computed using the info of a specific trait but then tested for associations with other phenotypes (see section 4.1 for application). Hence, a unique derivation leads to individual level pathway-scores that can be tested in a systematic way for a range of traits. As already explained, T-scores usage in the computation of pathway scores is crucial as it allows each gene to have a similar contribution in the pathway computation due to the new scaling space. This would otherwise not be possible in the case of imputed gene expression because the resulting gene variance depends on the one explained by PriLer model.

3.2.2 Genes and pathways association with a phenotype

Consider gene T-scores and pathway-scores matrices matrices, we indicate with $\boldsymbol{T}^g = (T(1,g),\dots,T(\tilde{M},g))$ and $\boldsymbol{PaSc}^{\mathcal{P}} = (PaSc(1,\mathcal{P}),\dots,PaSc(\tilde{M},\mathcal{P}))$ the vector of T-scores for a gene $g \in \mathcal{G}_{rel}$ and pathway-scores for a gene-set $\mathcal{P} \in \widetilde{DB}$ respectively, across $\tilde{M}$ samples in the new cohort, corresponding to the columns of matrices T (algorithm 1) and $PaSc$ (formula (3.21)). Let $\boldsymbol{y} = (y_1 \dots, y_{\tilde{M}})$ a trait of interest measured in the new cohort, we separately test the dependency between a gene/pathway (briefly called $\boldsymbol{x}$) with trait $\boldsymbol{y}$ via Generalized Linear Model (GLM) while correcting for possible covariates such as ancestry, sex and age (Fig. 3.1 Module 2). Note that although PriLer model account for known covariates, some effects can still be present especially due to ancestry differences in train reference panel and imputed cohorts, hence an additional correction while testing for association is required. GLM are a generalization of ordinary linear regression allowing the response variable (trait) to have an error distribution different from the normal one and the relationship between response and independent variables to be not linear [171].
Let $\tilde{Z} = [\tilde{\boldsymbol{Z}}^1|\dots|\tilde{\boldsymbol{Z}}^{\tilde{D}}]$ be the $\tilde{M} \times \tilde{D}$ matrix of known covariates.
GLM hypotheses require 1) the existence of an invertible function, named link function g, such that $g(\mu_m) = x_m\beta + (\tilde{Z}\boldsymbol{\alpha})_m$ with $\mu_m = \mathbb{E}(y_m)$ the expected value of response variable; 2) $\boldsymbol{y}$ distribution being in the exponential family i.e., the density function can be expressed in the following form

$$f(y_m|\theta_m,\phi) = \exp\left(\frac{y_m\theta_m - b(\theta_m)}{a(\phi)} + c(y_m,\phi)\right) \tag{3.22}$$

where $b(\cdot), c(\cdot,\cdot)$ and $a(\cdot)$ are specified functions determined by the distribution, $\phi \in \mathbb{R}^+$ is the so called dispersion parameter and $\theta \in \mathbb{R}$ is the natural parameter. The regression coefficients are estimated using Maximum Likelihood Estimation (MLE) via iterative least reweighted least square and computed applying `glm` `R` function.
Depending on trait nature, we account for three possible scenarios.

- $(y_1 \dots, y_{\tilde{M}})$ is a **continuous** vector being a realization of $\boldsymbol{y}$ response variable that follows a normal distribution with

$$\mathbb{E}(\boldsymbol{y}) = \boldsymbol{\mu} = \boldsymbol{x}\beta + \tilde{Z}\boldsymbol{\alpha},$$
$$Var(\boldsymbol{y}) = \sigma^2 I.$$

  Although the problem reduces to a linear regression, this formulation still goes under GLM class with link function being the identity function. Indeed, it can be shown that the normal density function can be written in form (3.22)

$$f(y_m|\mu_m, \sigma_2) = \exp\left(\frac{y_m\mu_m - \frac{1}{2}\mu_m^2}{\sigma_2} + \left[-\frac{1}{2}\left(\frac{x^2}{\sigma^2} + log(2\pi\sigma^2)\right)\right]\right)$$

  Regression coefficients estimates are found via `glm` function with `family = "gaussian"` and `link = "identity"` in `R` obtaining an estimate for regression coefficients $(\hat{\beta}, \hat{\boldsymbol{\alpha}})$. Note that also the dispersion parameter (equivalent to σ^2 variance), is unknown a prior and it is estimated from the data. Focusing our attention on regression coefficient $\hat{\beta}$ from gene/pathway feature, the estimation of corresponding standard error $S.E.(\hat{\beta})$ depends on the estimated variance $\hat{\sigma}^2$. Consequently, p-value testing $H_0 : \hat{\beta} = 0$ is computed based on $\frac{\hat{\beta}}{S.E.(\hat{\beta})}$ that follows a Student's t-distribution with $\tilde{M} - \tilde{D} - 2$ degrees of freedom under the null hypothesis, also known as t-statistic.

- $(y_1 \dots, y_{\tilde{M}})$ is a **binary** vector e.g. 0 for non affected and 1 for affected individuals. Let $p \in [0, 1]$ be the probability of an individual being affected, we assume that $\boldsymbol{y}$ response variable follows a Bernoulli distribution such that

$$\mathbb{E}(\boldsymbol{y}) = p,$$
$$g(p) = log\left(\frac{p}{1-p}\right),$$

  with $g : [0, 1] \longrightarrow \mathbb{R}$ an invertible link function. This scenario is part of GLM family because the probability mass function (equivalent to the density function for discrete

variable) of a Bernuolli distribution can be expressed in the following form

$$f(y_m = k|\theta, \phi) = p^k(1-p)^{1-k} = \exp\left(\frac{k\theta - b(\theta)}{a(\phi)} + c(k, \phi)\right)$$

$$\begin{aligned} \text{with} \quad & \theta = log\left(\frac{p}{1-p}\right), \\ & b(\theta) = log(1 + e^{\theta}), \\ & c(k, \phi) = log\binom{\phi}{k}, \\ & a(\phi) = 1/\phi, \\ & \phi = 1. \end{aligned}$$

The problem also known as binary logistic regression is solved again using `glm` function with `family = "binomial"` and `link = "logit"`. Note that, different from the Gaussian case, the dispersion parameter ϕ is known and equal to 1. This corresponds to the hypothesis that the variance of the response variable does not exceed the nominal variance $p(1-p)$, implying that the actual variance σ^2 is known. In this case, under the null hypothesis $H_0 : \hat{\beta} = 0$ the ratio $\frac{\hat{\beta}}{S.E.(\hat{\beta})}$ (called Z-statistic) follows a normal distribution with mean 0 and standard deviation 1 from which p-values are extracted. In addition, from regression coefficient $\hat{\beta}$ the odds ratio of having a certain trait for a one-unit increase in T-score/pathway-score is extracted as $OR = \exp(\hat{\beta})$.

- $(y_1 \ldots, y_{\tilde{M}})$ is a **ordinal categorical** vector of classes from 1 to K, namely it assumes integer values on an arbitrary scale for which only the relative ordering among them is significant, for instance the output of a cognitive test with 0 indicating individuals failing the test, 1 individuals passing the test at the second attempt and 2 individuals passing the test at the first attempt. For simplicity we assume y is non-decreasing in order with $y_m \leq y_{m+1}$. Ordinal regression can be performed using a GLM that fits both a coefficient β for gene expression, coefficient vector α for covariates and a set of thresholds $\theta_1, \ldots, \theta_K$ to a data set. Let $\mathbb{P}(\boldsymbol{y} \leq i)$ be the cumulative response probability having observed gene expression vector $\boldsymbol{x}$ and covariate matrix $\tilde{Z}$. The ordinal logistic regression solves

$$\mathbb{P}(\boldsymbol{y} \leq i) = \sigma(\theta_i - \boldsymbol{x}\beta + \tilde{Z}\boldsymbol{\alpha})$$

with σ the inverse link function. We specifically use a ordered logit model in which σ is the standard logistic function, hence

$$\sigma(\theta_i - \boldsymbol{x}\beta + \tilde{Z}\boldsymbol{\alpha}) = \frac{1}{1 + \exp^{-(\theta_i - \boldsymbol{x}\beta - \tilde{Z}\boldsymbol{\alpha})}}$$

Note that the problem can be also formulated in the following form

$$log\left(\frac{\mathbb{P}(\boldsymbol{y} \leq i)}{\mathbb{P}(\boldsymbol{y} > i)}\right) = \theta_i - \boldsymbol{x}\beta - \tilde{Z}\boldsymbol{\alpha}$$

with the log odds of the observed variable being less or equal than a particular category defined as linear combination of intercept term for that class and the considered gene and covariates. This model follows the parallel line assumption (proportional-odds model) in which the intercepts θ_i are different for each class but the slopes (β and $\boldsymbol{\alpha}$) all equal and independent from it.
Practically, the ordered logistic regression is solved via `polr` function from the `MASS` package with specific `Hess=TRUE` to retrieve the observed Hessian matrix computed from the optimization procedure, then used to get standard errors. P-values testing $H_0 : \hat{\beta} = 0$ is computed via `coeftest` function of `lmtest` package assuming $\frac{\hat{\beta}}{S.E.(\hat{\beta})}$ follows the normal distribution under the null hypothesis (z-test) and hence defining the ratio as Z-statistic.

Regardless the origin of the response variable, we broadly refer to *Z-statistic* as the ratio between the coefficient and its standard error estimation

$$Zst = \frac{\hat{\beta}}{S.E.(\hat{\beta})} \tag{3.23}$$

that on the one hand represents a measure of the precision with which the regression coefficient is not null, on the other hand gives information in term of gene/pathway direction effect i.e. whether an increase in T-scores or pathway-scores lead to an increase in observed trait or vice-versa.

The availability of large data set composed of a single cohort is very rare. Commonly genetic data focusing on a certain trait are composed of multiple cohorts collected at different sites, for instance PGC 36 cohorts for SCZ and CARDIoGRAM 9 cohorts for CAD. Under this circumstance, we apply a **meta-analysis** approach implemented in `R` following GWAMA method for GWAS results [22] to summarize the evidence of association coming from all the available cohorts. There is no need of additional alignment procedure at this stage because all the cohorts considered are already aligned for variant/strand/allele frequency in the pre-processing step (see section 4.1) and hence harmonized.
Let C be the number of cohorts, $\hat{\beta}^c$ and $S.E.(\hat{\beta}^c)$ for $c = 1, \ldots, C$ the effect and standard error of a certain gene/pathway (notation dropped for simplicity) with the trait of interest. The combined effect across all cohorts is computed via fixed effects meta-analysis combining genes effects weighted by the inverse of their variance:

$$\hat{B} = \frac{\sum_{c=1}^{C} \hat{\beta}^c w^c}{\sum_{c=1}^{C} w^c} \tag{3.24}$$

with $w^c = \frac{1}{(S.E.(\hat{\beta}^c))^2}$ and $S.E.(\hat{B}) = \sqrt{\frac{1}{\sum_{c=1}^{C} w^c}}$ the standard error of combined effect. The overall association of gene/pathway with the trait of interest is tested based on $\frac{\hat{B}}{S.E.(\hat{B})}$ following a normal standard distribution under the null hypothesis. In addition, to test for consistency of effects across cohorts, we compute Cochran's statistic as

$$Q = \sum_{c=1}^{C} w^c (\hat{B} - \hat{\beta}^c)^2 \tag{3.25}$$

with $Q \sim \chi^2(C-1)$ under the null hypothesis of consistency. If Q associated p-value is lower than a certain threshold (default 0.001), then we reject the consistency hypothesis and reformulate combined estimate using a random-effect model. In particular, we substitute w^c in (3.24) with $\tilde{w}^c = \frac{1}{\tau^2 + (S.E.(\hat{\beta}^c))^2}$ and

$$\tau^2 = \max\left[0, \frac{Q - (C-1)}{\sum_{c=1}^{C} w^c - (\sum_{c=1}^{C} (w^c)^2 / \sum_{c=1}^{C} w^c)}\right]$$

being the random-effect variance component used to inflate the standard error of the estimated effect in each cohort. The corresponding overall effect divided by its standard error $\sqrt{\frac{1}{\sum_{c=1}^{C} \tilde{w}^c}}$ still follows a standard normal distribution under the null hypothesis of no effect [172], hence we perform a similar test of the fixed-effect meta-analysis while accounting for heterogeneity across studies.

Since we are testing all reliable genes or all detected pathway in a database for a certain tissue, we perform multiple testing corrections to adjust p-values and correct for the occurrence of false positives. In particular, we apply Benjamini-Hochberg (BH) procedure [173] to control for False Discovery Rate (FDR) i.e. the expected proportions of detected associations that are false (type I error). We chose as default threshold to define an association significant FDR $\leq 5\%$ guaranteeing that among all associations called true only 5% are truly null. Note that, we decide to control for FDR and not family-wise error rate (FWER) for example via Bonferroni procedure for two main reasons. Firstly, FWER procedures control for the probability of having one or more false positives out of all the hypothesis tests conducted leading to more conservative results and less type I errors, however at the expense of higher rate of false negative hence less power. Secondly, the LD structure of the genome implies a correlation among imputed genes located in the same loci as well as correlation among imputed pathways for gene-sets sharing a high number of genes or even genes in the same loci. FWER procedure such as Bonferroni correction is even more conservative in the presence of correlated tests and hence increases false negative rates [174], contrary to BH procedure that was shown to still control for FDR under positive regression-dependency conditions [175], and conserved theoretical FDR in GWAS simulation under linkage disequilibrium presence [118].

In summary, in each tissue we obtain individual scores reflecting genes and pathway

activities (T and $PaSc$ matrices respectively) and for each phenotype available on the cohort(s) we separately tested the association with detected genes or pathways, hence performing TWAS or PALAS respectively and corrected for multiple tests via BH procedure. In addition, in case of a multiple-cohorts study the results are summarized via meta-analysis. Let Ph be the number of phenotypes tested, the outcome can be indicated with 4 genes and 4 pathways matrices of dimensions $|\mathcal{G}_{rel}| \times Ph$ or $|\widetilde{DB}| \times Ph$ respectively, being

1. B: effect sizes of association derived from GLM regression,
2. SE: standard errors for the aforementioned regression coefficients,
3. Zst: Z-statistics as defined in (3.23),
4. PV: nominal p-values testing gene/pathway effect of the phenotype.

3.2.3 Genetic correlation and Mendelian randomization

Besides evaluating the relationship of genes and pathways with trait of interest (CAD or SCZ), we also leverage the rich collection of UK Biobank phenotypes to investigate the correlation and causality of endophenotypes related to the trait with the trait itself based on genes/pathways mediation. Indeed, we are interested in answering the following questions:

1. is there any connection between the genetic basis of a trait (such as CAD) and an endophenotype characteristic of that trait (such as LDL)?
2. is endophenotype presence casual or protective of the trait and viceversa via imputed genes/pathways genetic instruments?

To achieve this goal, we first compute a correlation based on Z-statistics for associations (3.23) and for the correlated endophenotype-trait pairs we additionally performed Mendelian Randomization [37] to investigate causal mechanisms. This follow-up analysis is not part of the *CASTom-iGEx Module 2* github repository and is further available at gitlab.mpcdf.mpg.de/luciat/castom-igex_mr.

In details, for a certain tissue we initially remove redundant genes and pathways in order to avoid a spurious result only based on usage of the same relevant variant across genes, LD structure or gene co-regulation as well as pathways composed of a similar set of genes [62]. Genes whose TSS is distant less than 250kb are randomly pruned to ensure that they are not regulated by the same set of variants with filtered set denoted as $\mathcal{G}_{pr} \subset \mathcal{G}_{rel}$. For pathways instead, Reactome and GO databases are combined together (briefly indicated as $\widetilde{DB}$) and pathways are pruned based on the Jaccard index of shared

genes reliably imputed for that tissue:

$$JI(\mathcal{P},\mathcal{Q}) = \frac{|(\mathcal{P}\bigcap\mathcal{Q})\bigcap\mathcal{G}_{rel}|}{|(\mathcal{P}\bigcup\mathcal{Q})\bigcap\mathcal{G}_{rel}|} \tag{3.26}$$

with $\mathcal{P},\mathcal{Q} \in \widetilde{DB}$. The final set $\widetilde{DB}_{pr} \subset \widetilde{DB}$ is composed of randomly pruned pathways such that $JI(\mathcal{P},\mathcal{Q}) \leq 0.3$ for each possible combination. Note that in both cases, correlation among genes/pathways scores is still possible due to LD structure and gene co-regulation that goes beyond the gene proximity and it then propagates to pathways that do not share the same set of genes. However, this dependency will be additionally accounted for in MR analysis via an estimate of Pearson correlation from a subset of randomly selected individuals (e.g. $5,000$ controls for CAD from UK Biobank data set). Suppose $\boldsymbol{Zst}$ is the vector of length $|\mathcal{G}_{pr}|$ or $|\widetilde{DB}_{pr}|$ containing Z-statistic associations for pruned genes or pathways and with the trait of interest (e.g. CAD or SCZ). For the same set of genes/pathways, let Ph be the number of endophenotypes considered (e.g. LDL, Lymphocyte count, etc.) and $\boldsymbol{Zst}^p$ for $p = 1,\dots,Ph$ a column of overall association matrix as described in 3.2.2. The first goal is to understand whether genetic contribution in form of genes or pathway lead to a relationship between trait and endophenotype. This is achieved via Spearman's rank correlation coefficient

$$r^p = \frac{\text{cov}(\boldsymbol{rg},\boldsymbol{rg}^p)}{\sqrt{\text{var}(\boldsymbol{rg})}\sqrt{\text{var}(\boldsymbol{rg}^p)}} \quad p = 1,\dots,Ph \tag{3.27}$$

with $\boldsymbol{rg}$ and $\boldsymbol{rg}^p$ the rank i.e. relative position label within $\boldsymbol{Zst}$ and $\boldsymbol{Zst}^p$ respectively, $\text{cov} : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$ sample covariance operator $\text{cov}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$ and $\text{var} : \mathbb{R}^n \longrightarrow \mathbb{R}^+$ sample variance $\text{var}(\boldsymbol{x}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$. To test the significance of observed correlation, a permutation test [176] is performed shuffling $\boldsymbol{Zst}^p$ vector $N = 10,000$ times and computing r_i^p for each permutation i to obtain a p-value of observed Spearman correlation as the frequency of observing an higher correlation under null hypothesis of no association:

$$\text{p-value}^p = \frac{|\{i = 1,\dots,N \mid r_i^p < -|r^p| \vee r_i^p > |r^p|\}|}{N} \quad p = 1,\dots,Ph \tag{3.28}$$

Note that we use Spearman correlation instead of Pearson correlation to capture relationship that are not linear but still monotonic. In the event of a significant correlation between trait and endophenotype determined by p-value$^p \leq 0.05$, the next step is to examine whether the relationship is causal or not, meaning whether an increase in endophenotype due to changes in genetic variables leads to higher likelihood of developing that trait and vice-versa.

Mendelian Randomization (MR) technique is thus applied to answer these questions, practically via `MedenlianRandomization` `R` package [177]. Multiple strategies have been proposed to perform MR analysis [37], nevertheless the common aim is to estimate the causal effect of an exposure E on an outcome O using possibly multiple instrumental

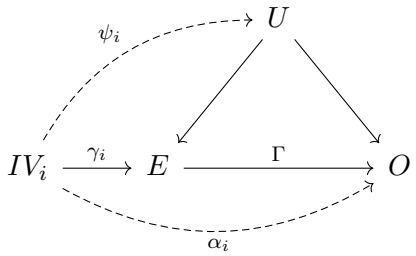

Fig. 3.6.: Mendelian Randomization diagram showing the causal and parametric relationship between instrumental variable IV_i, exposure E and outcome O. Solid line define IV_i as satisfying H1-H3 assumptions and dashed lines indicate a violation of those. U refers to unmeasured confounders that influence both outcome and exposure.

variables (IVs) for E. Different from so far MR applications that regard SNPs as IVs leveraging GWAS summary statistics or individual dosages, we consider the effect of genes and pathways to the corresponding exposure/outcome pair estimated via TWAS and PALAS. Among the possible MR techniques, we use inverse-variance weighted (MR-IVW) method [11] that combines the ratio estimates of casual effect of E on O using each IV in a meta-analysis model. In order to asses the causal role of an exposure, MR-IVW method relies on the following assumptions for IVs:

H1: IVs are associated with the exposure,

H2: IVs are independent of the outcome given the exposure (exclusion restriction),

H3: IVs are independent of the additional factors that confound exposure-outcome relationship.

Fig. 3.6 shows an illustrative diagram for a single IV_i with U combining all unmeasured confounders, solid lines indicating IV_i satisfy the aforementioned hypothesis and dashed lines representing a violation. In particular, assumptions on IVs will be not satisfied in the presence of horizontal pleiotropy i.e. IV_i affects the outcome via multiple phenotypes not related to the exposure with the effect indicated as α_i (violation of H2) or IV_i affects exposure-outcome confounders with strength ψ_i (violation of H3). Conversely, if assumptions H2 and H3 hold for all variants, their pleiotropic effect on the outcome is null. Nevertheless, MR-IVW with multiplicative random-effects still leads to a consistent estimation of the causal effect if the pleiotropy is independent of the IV-exposure associations, also known as Instrument Strength Independent of Direct Effect (InSIDE) assumption, meaning $\psi_i = 0$ but α_i can be different from zero as long as sample covariance between α_i and γ_i is zero or tends to zero for an increasing number of considered variants [97]. Thus we use MR-IVW with multiplicative random effect version [97] that partially correct for pleiotropy in order to preserve the validity of MR method. In general, MR-IVW technique is suited for two-sample analysis in which the estimation of IVs-exposure and IVs-outcome effects come from two non-overlapping data sets and are assumed not to be independent. However, it has been shown that it can be safely applied to one-sample studies, where same samples are used to estimate IVs-exposure and IVs-outcome effects, with large

sample size such as UK Biobank, even in presence of substantial correlation between exposure and outcome as a result of confounding [100]. Due to the presence of correlated genes/pathways and the possible bias in the result, we use a multiplicative random-effect MR-IVW extension that takes into account correlation of IVs via generalized weighted linear regression [178], all implemented in `mr_ivw` function.
Briefly, let N be the number of genes or pathways included in the analysis ($|\mathcal{G}_{pr}|$ or $|\widetilde{DB}_{pr}|$), let $\hat{\boldsymbol{\beta}}^E = \left(\hat{\beta}_1^E, \ldots, \hat{\beta}_N^E\right)$ and $\hat{\boldsymbol{\beta}}^O = \left(\hat{\beta}_1^O, \ldots, \hat{\beta}_N^O\right)$ be their estimated effects on exposure E and outcome O respectively, from TWAS or PALAS. Fixed-effect IVW method can be regarded as a weighted linear regression of $\hat{\boldsymbol{\beta}}^O$ on $\hat{\boldsymbol{\beta}}^E$ with no intercept term using as weights the variance and correlation estimate matrix Ω with each entry defined as $\Omega_{i_1,i_2} = SE\left(\hat{\beta}_{i_1}^O\right) SE\left(\hat{\beta}_{i_2}^O\right) \rho_{i_1,i_2}$ and ρ_{i_1,i_2} the estimated correlation between IV_{i_1} and IV_{i_2} [178]:

$$\begin{aligned} \Gamma &= \left(\left(\hat{\boldsymbol{\beta}}^E\right)^T \Omega^{-1} \hat{\boldsymbol{\beta}}^E\right)^{-1} \left(\hat{\boldsymbol{\beta}}^E\right)^T \Omega^{-1} \hat{\boldsymbol{\beta}}^O \\ SE(\Gamma) &= \sqrt{\left(\left(\hat{\boldsymbol{\beta}}^E\right)^T \Omega^{-1} \hat{\boldsymbol{\beta}}^E\right)^{-1}} \end{aligned} \tag{3.29}$$

Random-effect IVW extension allows for balanced pleiotropy i.e. it assumes pleiotropic effects across genes/pathways are random ($\sum_{i=1}^N \alpha_i = 0$) and InSIDE assumption of independence between pleiotropy and IV-exposure effects ($cor(\gamma_i, \alpha_i) = 0$) [97]. Hence, $SE(\Gamma)$ is modified to account for heterogeneity of ratio estimates due to pleiotropy based on Cochran's Q statistic, similarly to (3.25):

$$SE(\Gamma) = \sqrt{\left(\left(\hat{\boldsymbol{\beta}}^E\right)^T \Omega^{-1} \hat{\boldsymbol{\beta}}^E\right)^{-1} \frac{\left(\hat{\boldsymbol{\beta}}^O - \Gamma\hat{\boldsymbol{\beta}}^E\right)^T \Omega^{-1} \left(\hat{\boldsymbol{\beta}}^O - \Gamma\hat{\boldsymbol{\beta}}^E\right)}{N-1}} \tag{3.30}$$

The second term in (3.30) allows the variance of Γ causal effect to increase when heterogeneity is detected.
As already mentioned, we systematically test for causal association in a tissue between a trait and $\widetilde{Ph}$ related endophenotypes that are correlated with each other at a nominal p-value threshold of 0.05, obtaining Γ_p causal estimate and $SE(\Gamma_p)$ standard error for $p = 1, \ldots, \widetilde{Ph}$. The significance of MR estimate is then assessed via Z-test assuming $\frac{\Gamma_p}{SE(\Gamma_p)}$ follows a standard normal distribution under the null hypothesis of no association. Finally, the resulting p-values are adjusted for multiple test correction via BH procedure independently for each tissue.

3.2.4 Discussion

In the second step of CASTom-iGEx, we use PriLer prediction models to impute tissue-specific gene expression on large-scale genotype-only cohorts, convert the imputed expression into gene T-scores in order to rescale genes in a common space as well as enhance

differences among cases and controls, and collapse these scores into pathway information for each subject in a study. As a result, genes and pathways can be singularly tested for association with traits of interest using GLM (TWAS and PALAS respectively). Finally, these associations are exploited via correlation analysis and Mendelian Randomization technique to reveal candidate mechanisms in endophenotypes that mediate the occurrence of a certain trait due to aggregation of genetic signals into multiple genes and pathways.
As already pointed out, the conversion of imputed gene expression into T-scores is crucial to transform the available data into a space that does not depend on the variance explained by the model and gives the same relevance to all genes. In 3.2.1 we only define T-scores in the eventuality of a binary trait, however a natural extension for continuous traits (e.g. Fluid Intelligence score) would be to assign as "controls" individual inside the interquartile range (IQR) of $25 - 75$ percentiles distribution and as "cases" the more extreme cases outside IQR.
Of note, the approach we employ to compute individual level pathway-scores averages the information coming from singular genes without using specific gene weights nor taking into consideration the interaction among genes such as co-expression. An improvement of such an approach would then be an integration with tissue-specific information of co-expression or transcription factors regulation among those. On the other hand, pathway-scores can be also derived as targeted polygenic risk scores, weighting genes by strength of association with a certain trait. This approach was recently developed in a method called PRSet [91] that computes PRS across multiple biological pathways from variants. The major advantage of our method with respect to previously developed techniques for pathway discovery is the creation of individual-level scores, not only for a systematic association with traits but also for subsequent investigations such as case stratification. Different from MAGENTA which uses a hyper-geometric test for gene enrichment overlapping with GWAS hits [78] or more sophisticated methods that even include co-expression databases such as DEPICT [83], we do not require any p-value cutoff neither on SNPs nor on genes but rather give the opportunity to even small effect genes to contribute to pathway composition. Likewise, a method relying on GWAS summary statistics, PASCAL [80], does not impose any p-value threshold. However, it maps considered variants into genes based on their position rather than the corresponding regulation and does not provide insights on the sign of association or tissue specificity. Our PALAS methodology follows into the category of "self-contained" tests (i.e. testing the null hypothesis of no gene in the pathway is associated with the trait) compared to "comparative" methods (i.e. testing the null hypothesis of genes in the pathway being as strongly associated with the trait as other genes), and are generally more powerful [179].
As regards TWAS and PALAS analyses, it is important to stress that the p-value threshold we will refer to is only used to report significant results and to compare SNP dosages, gene T-scores, and pathway-scores associations from the same pool of individuals to study the aggregation mechanisms. In addition, when using gene T-scores and pathway-scores for patient stratification, we include all the available information without any filtering based on the p-value. A cut-off is only used for Mendelian Randomization application to ensure

H1 requirement is met for all instrumental variables (section 3.2.3).
For subsequent correlation and causality analyses, the comparison among a trait and its related endophenotypes can be performed both in a one-sample (same set of individuals to estimate exposure- and outcome-IVs associations) or two-sample scenario (two different data sets with no or very few overlapping individuals). In the second case, the optimal solution would be to work with previously harmonized data sets in terms of variants when performing correlation and MR analysis. In this scenario, the considered SNPs and indels are available in both data sets with the same REF/ALT annotation and similar ALT allele frequency. However, it is still possible to apply these methods to summary statistics TWAS and PALAS for two not harmonized data sets that have been filtered to retain genes and pathways exhibiting similar behavior. In particular, suppose two PriLer models on a reference panel such as CMC are separately trained, independently harmonizing variants of data set 1 and data set 2 with the reference panel and leading to PriLer models 1 and 2. We subsequently predict gene expression and pathway-scores on the reference panel based on models 1 and 2 and keep only common reliable genes and common detected pathways with Pearson correlation higher than 0.8 among the 2 PriLer models. In this way, we ensure that imputed expression and pathway information do not drastically differ even under the circumstance that the prediction models are based on different SNPs.
Note that, the approach we adopt to estimate correlation from imputed expression and pathway information is a naive one and does not take into consideration any additional correlation due to LD structure or pathway composition, although an initial pruning of genes/pathways regulated by the same entities is performed. Indeed, from GWAS summary statistics different methods have been developed that account for LD structure and are not biased by sample overlap, for instance LD score [94]. Nonetheless, our aim in deriving correlation from T-scores and pathway-scores is mainly a pre-processing step for Mendelian Randomization analysis, compared to the general aim of detecting endophenotypes potentially causal to the trait of interest.

As already mentioned, in the past years multiple approaches have been proposed to perform MR analysis. Among all, MR-IVW with multiplicative random-effects has been suggested for primary analysis being able to account for variant heterogeneity in causal estimates and for its efficiency under valid IVs [37]. In order to guarantee H1 hypothesis and avoid weak instrument inclusion, we only consider genes/pathways that are significantly associated with the exposure E at 0.05 FDR threshold. However, ensuring H2 and H3 to be satisfied is more complicated. In general, multiplicative random-effect MR-IVW method still give unbiased estimate when H2 is violated but InSIDE and balanced pleiotropy assumptions are met. Under H3 violation or directional pleiotropy, estimates for Γ are biased, increasing with the correlation among instrument strength and direct effect due to pleiotropy (see [97] for details). Other developed methods can lead to unbiased estimator under certain hypotheses violations. For example, Mendelian Randomization through Egger regression (MR-Egger) [99], can still give proper estimates under balanced as well as directional pleiotropy, meaning $\sum_{i=1}^{N} \alpha_i \neq 0$, fitting a linear regression of $\hat{\boldsymbol{\beta}}^O$ on

$\hat{\beta}^E$ but with intercept different from zero that estimates the actual pleiotropy observed. Nevertheless, this methodology was not recommended in the scenario of large one-sample application, i.e. exposure-IV and outcome-IV estimates from the same data set, since it was biased in the presence of correlation due to confounding that can not be accounted for, different from MR-IVW with random-effects [100]. Because this scenario will happen when considering UK Biobank for CAD analysis and to maintain consistency, we decide to only apply MR-IVW with random-effects for each tested causal trait-endophenotype pair, while cautioning on the exploratory nature of this analysis. Indeed, this type of analysis was more recently defined as a "Joint association study" rather than Mendelian Randomization study, including genome-wide associations without restricting to specific genes or pathways [180]. Despite its exploratory nature, it was argue that this type of approach is still able to provide "a suggestive evidence of a causal effect" in the presence of a non-null finding. The usage of genes and pathways instead of variants as instrumental variables drastically increases the interpretability of the results, identifying shared genes and mechanisms between a trait and an endophenotype that can exploit a causal role. On the other hand, we also increase the possible pleiotropy for IVs, as it is unlikely to imagine a pathway not being associated with any other confounder of exposure or outcome, unless extremely specialized and composed of few genes. MR-IVW with random-effects can control for a pleiotropic scenario when balanced, and integrated with an approach that accounts for IVs correlation, represented the most reliable and suitable choice for our application.

3.3 Genetically informed patient stratification

After having characterized genes and pathways associated to a certain disease as well as identified endophenotypes contributing to the disease etiology, the last module of CASTom-iGEx aims at stratifying patients leveraging tissue-specific gene T-scores (Figure 3.1 Module 3). In particular, input data used to cluster patients is solely derived from genetic information and intelligently rescaled based on Z-statistic association of genes with the disease to avoid ancestry or high variance phenotypes such as height to drive the stratification. Additionally, this clustering module detects differences in clinical variables and endophenotypes (if available) or derives gene risk-scores (gene-RS) imitating the actual endophenotype to suggest possible differences in disease features. Finally, these differences can be connected to changes not only of genes characterizing patients groups but also biological pathways taking advantage of the individual pathway-scores. To validate and use these results on new cohorts, we also implemented in CASTom-iGEx the possibility to project new individuals on existing clustering structure. Thus, our pipeline can be used to infer the likelihood of a new patient to be in a certain disease group that is then connected to specific endophenotypes trajectory and biological pathways activity.

3.3.1 Clustering via community detection

In order to stratify patients, we apply a graph-based clustering technique inspired from PhenoGraph approach [181] which was developed for single-cell data and it is suited for large scale data sets. This strategy was used both in the context of a large-scale single cohort such as CAD cases from UKBB and multiple cohorts combined such as SCZ cases in PGC cohorts.
Let M_c be the number of patients, a patient graph is indicated with the notion $G = (V, E)$, with $V = \{x_i | i = 1, \dots, M_c\}$ the set of vertices or nodes representing affected individuals and E the set of edges $E \subset \{(x_i, x_j) | x_i, x_j \in V \wedge i \neq j\}$ representing unordered pair of nodes to which is assigned a weight $S(i, j)$, i.e. the similarity between patient x_i and x_j. The considered genes from which clustering is derived are imputed from genetic information, hence during pre-processing step we aim at reducing ancestry contribution and remove redundant information of co-localized genes due to LD structure. Thus, for each tissue the following pre-processing steps are performed:

1. genes T-scores correlation among considered patients in computed (Pearson's corr.) and **genes are clumped** at 0.9 based on TWAS p-value with the trait of interest (e.g. CAD or SCZ). In particular, clumping procedure is performed sorting the reliably imputed genes from the most to the least significant. The first gene in this list ("current list") is considered and all the other genes with an absolute correlation higher than the desired threshold (0.9) are added to the "remove list", updating the "current list" of genes to be considered removing both the most significant gene and the correlated ones. This procedure is repeated until "current list" coincide with an empty list, hence all the genes have been considered, and the final set of genes is reached subtracting those in "remove list" from the initial set. This procedure ensures to maintain the most significant associations and no other genes correlated with those at a predefined threshold.

2. Let $\mathcal{G}_{cl}$ be the set of filtered genes, we **standardize** each T-score $g \in \mathcal{G}_{cl}$ subtracting the average and dividing per sample standard deviation across cases, namely

$$\boldsymbol{R}^g := \frac{\boldsymbol{T}^g - \mu_g}{\sigma_g} \tag{3.31}$$

   with $\mu_g = \frac{1}{M_c} \sum_{i=1}^{M_c} T(i, g)$ and $\sigma_g = \sqrt{\frac{1}{M_c - 1} \sum_{i=1}^{M_c} (T(i, g) - \mu_g)^2}$ sample mean and standard deviation.

3. Each standardized gene T-score $\boldsymbol{R}^g$ is then **corrected for L principal components (PCs)** to reduce ancestry contribution in the derived clustering structure, considering the residuals ($\boldsymbol{E}^g$) of the following linear model

$$\boldsymbol{R}^g \sim PC_1 + \cdots + PC_L. \tag{3.32}$$

4. Finally, each corrected gene T-score $\boldsymbol{E}^g$ is **multiplied by gene** g **Z-statistic** derived from TWAS of the trait of interest (Zst_g), leading to corrected and rescaled T-score

$$\widetilde{\boldsymbol{T}}^g := Zst_g \cdot \boldsymbol{E}^g. \tag{3.33}$$

The final object is then a matrix $\widetilde{T}$ of dimension $M_c \times |\mathcal{G}_{cl}|$ with columns composed of PCs-corrected and TWAS-rescaled T-scores and vector rows for each affected individual indicated with $\widetilde{\boldsymbol{T}}_i$. Thus, i) we enhance differences between patients via normalization, ii) we adjust ancestry information in the form of PCs in each gene T-score and iii) we give higher priority in defining clustering structure to genes that are more associated with trait of interest via TWAS-rescaling. Note that, PCs correction does not entirely remove their association with the final clustering structure but drastically reduces it (see section 4.3.7). In addition, TWAS-rescaling ensures a higher modularity and more defined and densely connected clusters compared to the same pre-processing procedure without rescaling (see section 4.3.8).
In the SCZ application composed of multiple cohorts, the different data sets are combined together via juxtaposition. The pre-processing steps are similar are described before for the joint gene T-score matrix, even PCs correction since they were estimated from the merged cohorts. However, we add an additional initial step to control for cohort heterogeneity, i.e. outliers removal. In particular, all cases across cohorts are combined together using juxtaposition and steps from 1 to 4 are performed, outliers are then detected as individuals that deviate beyond median ± 6 standard deviations in the first 2 Uniform Manifold Approximation and Projection (UMAP) [182] components of the newly derived T-score matrix $\widetilde{T}$. These individuals are then removed from further clustering analysis and pre-processing steps are repeated on the new set of samples.

After pre-processing steps, a sparse similarity matrix for each pair of samples is built based on Shared Nearest Neighbor (SNN) set derived from exponential similarity kernel of $\widetilde{T}$. In particular, for each pair of patients (x_i, x_j)

1. the euclidean distance of transformed gene T-scores is computed

$$E(i,j) = \sqrt{\sum_{g \in \mathcal{G}_{cl}} (\widetilde{T}(i,g) - \widetilde{T}(j,g))^2}, \tag{3.34}$$

2. a custom scaling parameter is derived that takes into consideration the local density structure and sparsity of the data derived from previously computed euclidean distance

$$\sigma_{i,j} = \frac{E(i,j) + \frac{1}{K}\sum_{x_l \in \mathcal{E}^K_{x_i}} E(i,l) + \frac{1}{K}\sum_{x_m \in \mathcal{E}^K_{x_j}} E(i,m)}{3} \tag{3.35}$$

   with $\mathcal{E}^K_{x_i}$ the set of closest K patients to sample x_i based on (3.34),
3. using the proper scaling parameter (3.35), exponential similarity kernel is computed

$$H(i,j) = \exp\left(-\frac{E^2(i,j)}{\mu\sigma_{i,j}}\right), \tag{3.36}$$

giving an initial similarity matrix among x_i and x_j that is not sparse but it already takes into consideration the local density of the data. Exponential similarity kernel was already used for similarity network fusion strategy that clustered cancer patient based on an integration of data modality [183]. This measure was shown to be robust to hyper-parameters settings μ and K, hence we decided to fix $\mu = 0.5$ and $K = 30$.

4. In order to retrieve only local interactions information and create a sparse similarity matrix, exponential kernel (3.36) is used to find the set of K nearest neighbour for a patient x_i indicated with $\mathcal{H}_{x_i}^K$. The fraction of SNN between patient x_i and x_j is then computed as the Jaccard Index between the two patients nearest neighbours, i.e.

$$S(i,j) = \frac{|\mathcal{H}_{x_i}^K \bigcap \mathcal{H}_{x_j}^K|}{|\mathcal{H}_{x_i}^K \bigcup \mathcal{H}_{x_j}^K|}. \tag{3.37}$$

We fix again $K = 30$, on the one hand to be consistent with previously defined exponential kernel hyper-parameter, on the other hand because it gives a good compromise between being small enough to prevent large neighborhoods and sufficiently large to valuate local geometry.

Sparse similarity defined in (3.37) represents the weights for edges set E in the unordered patient graph G, with an edge between patient x_i and x_j only present if they share at least one neighbour and maximal if neighbours are exactly the same.
Finally, clustering based on the aforementioned graph structure is performed via Louvain method [184] implemented in the `igraph` R package [185], that detects communities maximizing the graph modularity. The Louvain method adapts well to our sparse graph structure with a reasonable computational time for our large patient network. Briefly, community detection involve the partition of a graph into community of densely connected nodes, leaving the nodes between different communities only sparsely connected. The quality of this partition is measured via modularity Q, a scalar between -1 and 1 that assess the density of connection inside the communities compared to between communities. In our case, being $S(i,j)$ the weights of unordered patient graph G, we define with $k_i = \sum_{j=1}^{M_c} S(i,j)$ the sum of weights edges attached to patient x_i, $m = \frac{1}{2}\sum_{i=1}^{M_c}\sum_{j=1}^{M_c} S(i,j)$ total weights edges sum and C_i the community to which patient x_i is assigned. Then, the modularity of a partition is determined as

$$Q = \frac{1}{2m}\sum_{i=1}^{M_c}\sum_{j=1}^{M_c}\left(S(i,j) - \frac{k_i k_j}{2m}\right)\delta(C_i, C_j) \tag{3.38}$$

where δ is the Kronecker function being 1 when x_i and x_j are assigned to the same community and 0 otherwise. In Louvain clustering, the modularity is the objective function

to be optimized via a hierarchical and agglomerative algorithm. It consists of two phases repeated iteratively, with the starting point the assignment of a community for each node of the graph (i.e. patient).

phase I : For each node x_i and its neighbours x_j, the gain in modularity when removing x_i from its community and by placing it in the community of x_j is evaluated. Algorithm efficiency partly reside on the simplicity in modularity difference formula when moving a node x_i to community C. Indeed, let $k_{i,in} = \sum_{x_j \in C} S(i,j)$ be the weights sum for edges from node x_i to any other node in C, then the gain is computed as

$$\Delta Q_{x_i \to C} = \left[\frac{\sum\limits_{x_j \in C} k_{j,in} + 2k_{i,n}}{2m} + \left(\frac{\sum\limits_{x_j \in C} k_j + k_i}{2m} \right)^2 \right] + \\ - \left[\frac{\sum\limits_{x_j \in C} k_{j,in}}{2m} - \left(\frac{\sum\limits_{x_j \in C} k_j}{2m} \right)^2 - \left(\frac{k_i}{2m} \right)^2 \right].$$

If there is a community for which this gain is positive, the node x_i is reassigned to the community with maximum gain, otherwise it stays in the original one. The first phase is concluded after considering repeatedly and subsequently all nodes and no further improvement is achieved, leading to a local maximum in modularity.

phase II : A new graph is built whose nodes are the communities found in phase I and weights among two new nodes are given by the sum of weights for edges between nodes in the corresponding two communities i.e. $S_{updated}(1,2) = \sum_{x_i \in C^1} \sum_{x_j \in C^2} S(i,j)$, while edges between nodes of the same community give self-loops for the corresponding community in the new network.

These two phases are iterated until no further improvement is achieved in term of modularity. Hence, the number of final communities detected is not decided a prior but detected by Louvain method in a unsupervised manner. However, hyper-parameter K of number of nearest neighbour can still influence the final number of groups detected since it defines the initial structure of the graph decided by the sparsity in the similarity measure.

In order to validate and extend our clustering structure to new patients, we implement a projection method similarly to PhenoGraph [181] that uses the percentage of SNN to predict groups on external cohorts not used to derive communities. This projection method is applied when predicting clustering on CARDIoGRAM cohorts based on UK Biobank for CAD or on scz_boco_eur cohort in PGC when clustering based on all the other cohorts for SCZ (see chapter 4). Note that, pre-processing steps 2 to 4 are initially performed in the new cohort in which a computed clustering structure is projected, excluding step 1 of gene filtering since the same gene set of the computed clustering is retained. Practically, we projected European cohorts onto a clustering structure based on the same ethnicity (or

a subset of that) as we initially filtered per ancestry (see section 4.1). Theoretically, this method can be extended to multi-ethnicity structure but remaining ancestry discrepancies should be investigated.

As previously mentioned, for a new cohort only genes used in the clustering model ($\mathcal{G}_{cl}$) are considered and subsequently standardized, corrected for PCs and rescaled across patients, using the same TWAS Z-statistic applied in the model clustering but mean and standard deviation as well as PCs adjustment computed from the new cohort. Let $G = (V, E)$ the be graph composed of $V = \{x_i | i = 1, \ldots, M_c\}$ nodes on which P communities $C_1, \ldots, C_P$ have been detected based on S similarity matrix. Cluster assignment is also expressed as a $M_c \times P$ matrix H such that $H(i, p) = 1$ is patient x_i belongs to community C_p and 0 otherwise. We define with $V_u = \{x_i | i = M_c + 1, \ldots, M_c + N_c\}$ the new patient cohort to which community have not been assigned. For each new patient, a class label is determined based on the likelihood that a random walk originated from that sample will arrive first at a labelled sample of community C_p. In order to do that, a new graph $\tilde{G} = (\tilde{V}, \tilde{E})$ is built for all $M_c + N_c$ patients and applied in a series of random walk simulations that spread the label information from clustered nodes V to unclustered V_u. The solution to this problem is based on discrete potential theory and the probability of a new patient to be assigned to a community C_p is computed as the solution of linear equations [186]. In short, for the partially labelled graph $\tilde{G}$, we define the $(M_c + N_c \times M_c + N_c)$ sparse weighted adjacency matrix $\tilde{S}$ as in (3.37) and the diagonal degree matrix $\tilde{D}$ with entries $\tilde{D}(i, i) = \sum_{j=1}^{M_c+N_c} \tilde{S}(i, j)$. Hence, we can compute the graph Laplacian of the enlarged sample network

$$\tilde{\mathcal{L}} = \tilde{D} - \tilde{S}$$

that can be decomposed as

$$\tilde{\mathcal{L}} = \begin{pmatrix} \mathcal{L} & B \\ B^T & \mathcal{L}_u \end{pmatrix}$$

where $\mathcal{L}$ and $\mathcal{L}_u$ refer to the clustered nodes V and to be labelled ones V_u respectively. Finally, the probability of a random walk starting from a node in V_u to first arrive at a certain node in V is computed through the solution of

$$\mathcal{L}_u \mathcal{P} = -B^T H \tag{3.39}$$

where $\mathcal{P}$ is the solution $N_c \times P$ matrix representing the probability for each new individual x_i in V_u of being assigned to a community C_p. Hence, we assign the new individual x_i to the community $C_{\hat{p}}$ such that the inferred probability is maximal, i.e. $\hat{p} = \text{argmax}_{p=1,\ldots P} \mathcal{P}(i, p)$.

3.3.2 Characterization of genes and pathways different trajectories

After the detection of cases communities, we investigate which genes are driving and involved in clustering structure as well as find differences in pathway scores, not only for the tissue used in detecting the clustering but also across all available tissues. In order to test this, we use a tissue-specific one-vs-all group approach comparing the distribution of a certain gene/pathway in a group versus the remaining patients. Namely, we apply Wilcoxon-Mann-Whitney (WMW) via `rstatix` R package [187], a widely applied non-parametric method that tests the probability of two random variables being greater than each other. The advantage in using `rstatix` package is the computation of estimates for each test as well as confidence interval (briefly explained below). Prior to WMW testing, each gene T-score is standardized and corrected for PCs, as described in pre-processing step 2 and 3 of section 3.3.1, to test effects that have been already accounting for ancestry and to give comparable WMW estimates, whereas the TWAS-rescaling step is not necessary since each gene is considered separately.
Let $\mathcal{C}$ be the cluster partitioning patients in V composed of P groups: $C_1, \dots, C_P$ such that $C_p \bigcap C_q = \emptyset$ when $p \neq q$ and $\bigcup_{p=1}^{P} C_p = V$. We briefly denote with $\boldsymbol{F}_q$ the vector of gene T-score for a gene g or pathway-score for a pathway $\mathcal{P}$ across samples in C_q cluster, hence

$$\boldsymbol{F}_q := \boldsymbol{T}_q^g = \{T_i^g | x_i \in C_q\}$$
or
$$\boldsymbol{F}_q := \boldsymbol{PaSc}_q^{\mathcal{P}} = \{PaSc_i^{\mathcal{P}} | x_i \in C_q\}.$$

Similarly, we indicate with $\boldsymbol{F}_{-q}$ the same feature across all the other patients not in C_q group, for instance $\boldsymbol{F}_{-q} = \{T_i^g | x_i \in V \setminus C_q\}$. WMW is a non-parametric test having as null hypothesis that two populations come from the same distribution, in particular the probability of one being greater than the other is the same as the opposite case probability. In our setting, we are hence testing

$$H1: \ \mathbb{P}(\boldsymbol{F}_q > \boldsymbol{F}_{-q}) \neq \mathbb{P}(\boldsymbol{F}_q < \boldsymbol{F}_{-q}).$$

This test is suited for continuous response but can be still applied to ordinal variables. However, if responses are continuous and the alternative hypothesis is reduced to a location shift, WMW test detects whether medians of the two distributions are different. Thus, the estimates computed via `rstatix` package (Hodges–Lehmann estimate) represents the median of all possible differences between a sample in $\boldsymbol{F}_q$ and $\boldsymbol{F}_{-q}$. When shape and dispersion of $\boldsymbol{F}_q$ and $\boldsymbol{F}_{-q}$ distributions are different, WMW can still give a significant p-value showing a difference in distribution that it would not be related to a difference in medians [188]. We choose WMW test and not most commonly used t-test because not all

the imputed genes have a normal distribution as this depends on the number of regulatory SNPs decided via PriLer.
Since we are testing all reliable genes and detected pathways in a tissue, for each group C_p p-values from WMW are subsequently corrected for multiple comparison to control for FDR via Benjamini-Hochberg procedure. In order to obtain a cluster-specific summary of associated genes based on genomic location, for each group we combine significant genes detected among all tissues into loci based on physical location, i.e. TSS window is enlarged to 200kb each side and 2 genes are merged if their distance of TSS window is lower than 1Mb. For pathways instead, we perform an additional filtering step prior to WMW testing. In particular, for each tissue we only consider pathways composed of at least 3 genes and no more than 200 genes, both for the original gene-set $\mathcal{P}$ and the gene-set obtained intersecting with reliable tissue-specific genes $\mathcal{P} \cap \mathcal{G}_{rel}$. Pathways are then clumped, similarly to what is described in pre-processing step 1 of clustering (see section 3.3.1), in decreasing order according pathway coverage ($|\mathcal{P} \cap \mathcal{G}_{rel}|/|\mathcal{P}|$) and number of genes considered to compute the pathway ($|\mathcal{P} \cap \mathcal{G}_{rel}|$) and obtaining a final set of pathways having pairwise Jaccard Index not exceeding 0.2 (see formula (3.26)).
Finally, when projecting the clustering on new cohort, we are interested in internally validating the results based on cluster-relevant genes signature. Thus, for each detected patient group q

1. we consider only genes that are cluster-relevant (FDR ≤ 0.01) in the model composed of V patient nodes, namely $\mathcal{G}_q$;
2. we compute WMW vector estimates of length $|\mathcal{G}_q|$ separately for patients in the model clustering and external cohort (V_u new patients): $\boldsymbol{W}^q$ and $\boldsymbol{W}^q_{new}$ respectively;
3. we compute Spearman correlation for those two estimate vectors derived from model and new cohort: cor$(\boldsymbol{W}^q, \boldsymbol{W}^q_{new})$.

Furthermore, to make sure that cluster-specific gene signature is not only based on a single locus, we compute the number of reproduced loci in the new cohort using previously described loci summary of cluster-relevant genes. In particular, for each relevant locus we only retain the most cluster-specific and significant gene g and consider it replicated when the sign of WMW estimate coincide among the two data sets $W^q_g \cdot W^q_{new,g} > 0$ and additionally at the nominal level when significance of gene g for group q in the new data sets exceed 0.05.

3.3.3 Detection of differences in endophenotypes and treatment responses

Similarly to the investigation of differences in genes and pathway, we aim at understanding whether the detected group of patients have different trajectories in disease characteristics and treatment responses. Provided that endophenotype and clinical vari-

ables are available on the same data set the clustering was performed (or projected to) e.g. for UK Biobank, we apply again a one-vs-all comparison strategy to detect group specific differences via generalized linear models. Among the provided phenotypes, we use the following filter and transformations:

- phenotypes registered on less than 100 individuals are removed;
- binary and ordinal phenotypes being not zero in less than 50 individuals are removed;
- continuous phenotypes are standardized, subtracting mean and dividing by standard deviation ($\frac{\boldsymbol{Ph}-\mu}{\sigma}$);
- ordinal phenotypes with less than 10 individuals in the base category in either the considered group C_p or the remaining samples C_{-p} are removed.

Let $\boldsymbol{Ph} = (Ph_1, \dots, Ph_{M_c})$ the phenotypes to be tested available for the clustered patients, $\boldsymbol{H}^p$ a binary M_c-vector that represents membership of patients to cluster C_p and $Z = [\boldsymbol{Z}^1 | \dots | \boldsymbol{Z}^D]$ the matrix of D known covariates usually including principal components from genotype data, age, sex and phenotype-specific confounders (specifics for CAD and SCZ application in Tab. B.6, B.7, B.11). The family used in GLM depends on the nature of $\boldsymbol{Ph}$, i.e. continuous, categorical ordinal or binary. Thus, for each endophenotype considered $\boldsymbol{Ph}$ and group C_p we are testing

$$\boldsymbol{Ph} = \beta^p \boldsymbol{H}^p + \gamma^1 \boldsymbol{Z}^1 + \dots + \gamma^D \boldsymbol{Z}^D + \boldsymbol{E} \tag{3.40}$$

with $\boldsymbol{E}$ error vector of assumed distribution depending on $\boldsymbol{Ph}$ data type and GLM solution computed as described in TWAS and PALAS 3.2.2 section via `glm` or `polr` functions.
The estimated regression coefficient associated with $\boldsymbol{H}^p$, $\hat{\beta}^p$, gives the impact of cluster p compared to all other patients in the considered endophenotype distribution after adjusting for possible covariates. For each cluster C_p, derived p-values are then corrected for multiple testing via BH procedure across all considered endophenotypes.

When treatment annotation is available together with phenotype information for patients, we examined whether patient groups show different treatment outcome based on a certain response phenotype, for example different LDL reduction rate due to cholesterol lowering medication. In order to test that, let $\boldsymbol{Ph}(C_p)$ be the vector of response phenotype evaluated on samples in C_p cluster, standardized if continuous and excluded if less than 300 values were available.
Firstly, for each group C_p the medication effect on response endophenotype is estimated via the following GLM:

$$\boldsymbol{Ph}(C_p) = \beta^p_{Me} \boldsymbol{Me}(C_p) + \gamma^1_p \boldsymbol{Z}^1(C_p) + \dots + \gamma^D_p \boldsymbol{Z}^D(C_p) + \boldsymbol{E}$$

with $\boldsymbol{Me}(C_p)$ the binary vector of treatment Me indicating whether an individual is assuming that medicine and restricted to patients in C_p. Similarly as before, $\boldsymbol{Z}^d$ are

additional covariates evaluated in that group that include also other treatment binary categories. Hence, we define with $\hat{\beta}^{p}_{Me}$ the regression coefficient estimated via GLM delineating the effect on phenotype Ph of medication Me in group C_p.
Secondly, we apply Z-test to evaluate differences in responses between each pair of patient groups C_p and C_q based on previously computed regression coefficients [189]

$$Z(p,q) = \frac{\hat{\beta}^{p}_{Me} - \hat{\beta}^{q}_{Me}}{\sqrt{\left(S.E.\hat{\beta}^{p}_{Me}\right)^2 + \left(S.E.\hat{\beta}^{q}_{Me}\right)^2}} \tag{3.41}$$

that follows a standard normal distribution under the null hypothesis of no differences. Resulting p-values are then corrected for multiple testing across all endophenotypes but separately for each pairwise comparison (C_p, C_q) and medication Me considered.
For instance, a significant difference between two patient groups with $Z(p,q) > 0$ and $\hat{\beta}^{p}_{Me}, \hat{\beta}^{q}_{Me} > 0$ indicates that among cases in group C_p individuals taking medications Me will have an increase in term of response phenotype (compared to cases in C_p not taking it) higher than the response to the medication observed in cases group C_q.

3.3.4 Risk scores computation to mimic not available endophenotypes

It is common for large-scale genotype-only data set to not include any additional endophenotype information or clinical variable besides the trait under study, such as PGC cohorts for SCZ. Although clustering of patients can still be achieved, it would be impossible to test the hypothesis of different endophenotype trajectory lacking additional data on individuals. Hence, we establish a strategy to assign to each patient a endophenotypic specific score derived from genetic data using tissue-specific imputed gene expression called gene risk score (gene-RS).
Suppose two data sets composed of genomic information are available, the first one includes only genotype-data and a trait of interest (e.g. PGC for SCZ) and the second one is composed of genotype and additional endophenotypes related to the trait in the first one (e.g. UK Biobank deep phenotyping including fluid intelligence score or lymphocyte counts). After PriLer tissue-specific models have been estimated and gene T-scores computed on the two data sets, for each tissue gene-endophenotype association is estimated on the second data set via TWAS (as described in section 3.2.2) obtaining for each gene $g \in \mathcal{G}_{rel}$ and endophenotype Ph the association Z-statistic Zst^{Ph}_{g}. To avoid redundant information and control for LD structure, genes are clumped at squared Pearson correlation of 0.1 with genes decreasingly ordered according R^2 PriLer imputation estimation (resulting set of genes indicated with $\tilde{\mathcal{G}}_{rel}$). To estimate correlation among imputed genes, we used UK Biobank subset of samples as described in section 3.2.3. Gene T-scores are then separately corrected for PCs via a linear model, similarly to the pre-processing step 3 of clustering (section 3.3.1). Afterward, for a certain endophenotype

gene-RS is computed on the first data set samples as the weighted sum of gene T-scores $\boldsymbol{T}^g$ multiplied by TWAS estimated in the second data set

$$\boldsymbol{RS}^{Ph} = \sum_{g \in \tilde{\mathcal{G}}_{rel}} Zst_g^{Ph} \boldsymbol{T}^g \tag{3.42}$$

Suppose M is the total number of samples in the first data set, we then obtain an individual risk score that imitate the not provided endophenotype. It is now possible to use the same strategy deployed in section 3.3.3 to test group-specific difference in $\boldsymbol{RS}^{Ph}$, prior to a standardization of each gene-RS across the considered samples giving mean zero and standard deviation of one.
In case of multiple cohorts such as PGC and CARDIoGRAM, we used two different strategies. In the first scenario (i.e. PGC) we perform clustering combining all cohorts together and correcting for PCs that have been computed merging all samples. Hence, the gene-RS are computed in a similar way correcting for PCs and standardizing after all cohorts combination, and cluster differences are tested as usual (section 3.3.3). In the second scenario (i.e. CARDIoGRAM), gene-RS are computed and standardizing separately as well as cluster differences tested separately for each cohort, having PCs computed independently for each cohort and cluster assignment obtained as a projection to validated CAD partition on UK Biobank. A summary result to estimate patient group-specific differences based on gene-RS for a certain endophenotype is then obtained via meta-analysis, same as in section 3.2.2.
Because $\boldsymbol{RS}^{Ph}$ is only an estimate of the actual endophenotype, we also define a measure of confidence of the observed group-specific difference besides p-value significance from (3.40). The confidence of group-specific effect to be possible on the actual endophenotype depends on

1. the reliability on gene-RS to predict the actual phenotype which depends on the number of samples in the second data set available to estimate Zst_g^{Ph} as well as variance explained by genetic components (in our case imputed gene expression) with respect to the total variance of Ph, also known as genetic heritability;
2. the effect size of the group-specific difference.

Hence, we define a non-negative Cluster Reliable Measure (CRM) for a endophenotype Ph and a cluster C_p as

$$CRM^{Ph}(p) = F_{Ph} \cdot |\hat{\beta}^p| \tag{3.43}$$

where $\hat{\beta}^p$ is solution of (3.40) estimating group-specific effects on gene-RS and F_{Ph} is the F-test statistic that indicate gene-RS performance when modeling the actual phenotype. F-test is computed entirely on the second data set (e.g. UK Biobank) and estimates the improvement of adding gene-RS as predictor of the actual endophenotype instead of the covariates only. In particular, suppose $\boldsymbol{Ph}$ is the phenotype vector across M available samples, $\boldsymbol{RS}^{Ph}$ is the gene risk-score based on (3.42) with gene T-scores $\boldsymbol{T}^g$ and Z-statistic

Zst_g^{Ph} computed on the same data set (e.g. UK Biobank) and Z is the matrix of D covariates (usually PCs, age and sex). We apply the same pre-processing used on an external cohort and correct each gene T-score via linear regression for PC1-10 across all UK Biobank samples available for a certain phenotype, compute gene-RS and perform standardization. Consider the two nested models

$$\text{Model 1: } \boldsymbol{Ph} \sim \boldsymbol{RS}^{Ph} + Z$$
$$\text{Model 2: } \boldsymbol{Ph} \sim Z$$

Partial F-statistic comparing linear model 1 against linear model 2 is obtained via `anova` `R` command computing

$$F_{Ph} = \frac{(RSS_2 - RSS_1)/(D + 2 - D - 1)}{RSS_1/(M - D - 1)}$$

that follows a F-distribution with $(D + 2 - D - 1, M - D - 1)$ degrees of freedom under the null hypothesis of model 1 not providing a significantly better performance than model 2 and with RSS indicating the residual sum of squared of a linear model. It is important to note that we use F-statistic and not simply the coefficient of determination of R^2 given by $R^2_{model1} - R^2_{model2}$ since it is inflated for phenotype available for a reduced number of samples (see Fig. 4.60A-B and section 4.4.7).

In summary, CRM is an unbounded score that represents the level of confidence we can assign to each group in trusting that group-specific endophenotype difference would still hold if given the chance to measure the actual phenotype. As we do not know a priori a threshold for CRM above which we can consider a group-specific association reliable, we validate our approach using as ground-truth the UK Biobank clustering and actual endophenotypes detected and as prediction the projected clustering structure onto CARDIoGRAM cohorts and gene-RS differences for the same endophenotypes. We hence calibrated a cut-off of 610 that gives a precision ≥ 0.85 and it is subsequently applied for SCZ gene-RS results (see section 4.3).

3.3.5 Discussion

The third step CASTom-iGEx involves the actual clustering of patients based on genetically derived features i.e. gene T-scores. Our pipeline additionally detects genes and pathways responsible for the obtained tissue-specific clustering structure, creating a summary of associated gene loci and considering a set of highly informative and non-redundant pathways. When additional endophenotype information on the same patient set is available, the detected groups are tested for differences in those endophenotypes or even treatment responses. Otherwise, large-scale genomic data sets with deep phenotyping such as UK Biobank can be leveraged to build gene-RS, mimicking the actual endophenotype to investigate plausible differences and endophenotype trajectories. Finally, our pipeline

allows predicting group membership on external cohorts, with the final aim of suggesting affected pathways and pharmaceutical strategies based on relevant phenotype trajectories in new patients.
The clustering strategy that we apply is built on TWAS-rescaled gene T-scores that gives more relevance to genes related to the trait of interest. We would like to stress that, UMAP methodology is not used to group patients into a new embedded space but exclusively to 1) represent data in a lower dimensional space, 2) detect outliers when multiple cohorts are concatenated in a unique data set.
Note also that, our endophenotype detection strategy based on gene-RS is built on two different data sets: one used for Z-statistic estimation and another one for risk score prediction. With a similar approach used in Mendelian Randomization, in case these two data sets are not harmonized per variants, imputed gene expression based on the two different sets of variants still needs to be correlated, therefore we once more filter genes with a correlation lower than 0.8 (see section 3.2.4).
Regardless of the subsequent analyses, we perform as initial pre-processing the exclusion of genes that are highly correlated via clumping (see section 3.3.1). The threshold we fixed of 0.9 only removes genes that are actual repetition, for instance due to LD structure. Different values can be investigated and we also explore an almost no correlation setting with 0.1 threshold for SCZ (see section 4.4.6 for details). Specifically in that situation, outlier detection due to multiple cohorts combination is performed as a union across all the tested tissues and filtering options, to make possible the comparison of clustering structures.
Since our clustering strategy is solely based on genetically derived information, one can expect patient partition to be driven by the largest source of variation i.e. ancestry or baseline endophenotype differences such as height or blood type. However, the rescaling of gene T-scores based on TWAS disease associations gives higher priority to genes diseases related, reducing the contribution from the aforementioned sources. Indeed, we show that ancestry information is not driving the clustering structure compared to gene contribution. Nevertheless, there can still be a significant difference in PCs distribution, despite being reduced via gene T-score PCs correction. In this scenario, we additionally show that the overlap between tissue-derived and ancestry-derived clustering is minimal and does not influence the observed group-specific endophenotype differences (see section 4.3.7 and 4.4.8). As a matter of fact, highly cluster-relevant genes can be detected based on WMW test p-value, expecting values equivalent to zero. However, additional genes (and pathways) significant in a certain cluster at FDR ≤ 0.01 give insights into clustering further trajectory and contributors.
As regards the actual clustering strategy we applied, building a graph based on local density structure via shared nearest neighbor is also the core of the PhenoGraph approach [181]. Different from Levine et al., we define the set of neighbors based on Gaussian kernel (3.36) and not simply euclidean distance (3.34) which computes a sample similarity using its local density via customized standard deviation. The only hyper-parameter for clustering is the number of nearest neighbors considered K. As already mentioned, we

set $K = 30$ because is a good compromise to evaluate local geometry but avoid large neighborhoods, on top of the fact that PhenoGraph was shown to be robust for the choice of K.

Finally, we would like to stress that over the past 15 years, personalized medicine has been a target of multiple studies and method development, in particular for complex diseases (see section 2.5). The majority of techniques applied for patient stratification rely on polygenic risk scores that combine SNP specific association from GWAS with a disease and individual dosages. It is worth noting that, for complex diseases such as SCZ and CAD, so far PRS failed to separate cases and controls distribution, with an accuracy that cannot be applied at the clinical level [73, 74]. Indeed, we also do not aim to separate affected and not affected individuals based on a gene-RS but to give relevant information on group-specific trends to aid a treatment therapy decision for different patient groups. Finally, an example of detecting clustering evidence for complex diseases based on GWAS has been recently developed in [110]. However, there are core differences between BUHMBOX and CASTom-iGEx method. BUHMBOX does not cluster a priori patients but tests whether observed heterogeneity in individuals with complex diseases is driven by a subgroup of individuals having a certain genetic correlation only from SNPs associated with a related endophenotype. Instead, CASTom-iGEx investigates whether actual (or predicted) endophenotypes vary among genetically-derived clusters, not based solely on SNPs but on aggregated regulatory effects for gene expression.

4 Application of CASTom-iGEx

In this section, we show the potential of our newly developed pipeline CASTom-iGEx in identifying possible biological mechanisms underlying complex diseases, decomposing it to putative endophenotype intermediate mechanisms, and retrieving groups of patients associated to specific pathway and endophenotypic trajectories. After presenting the advantages and reliability of our improved gene expression modeling strategy (PriLer) in inferring genetically based genes distribution, we apply CASTom-iGEx to two differently characterized complex diseases having in common an inherited polygenicity: coronary artery disease (CAD) and schizophrenia (SCZ). These results are additionally validated on external cohorts both in terms of genes/pathways identified and clustering structure results. This chapter is divided as follows:

Section 4.1 Initially, we describe the data sets included in this thesis. In particular, we outline the pre-processing steps performed, the rationale behind data sets harmonization and matching for each trait under investigation, and the phenotype definition as well as normalization in the UK Biobank cohort.

Section 4.2 We then describe PriLer results built on GTEx and CMC reference panels, evaluate tissue-specificity of selected prior weights, compare the results with elastic-net and previously developed methods in terms of prediction performances and robustness, and show the impact of sample size in selecting reliable genes and shaping explained variance.

Section 4.3 Afterward, PriLer models built on GTEx reference panel harmonized with UK Biobank as well as CARDIoGRAM cohorts are leveraged across 11 CAD related tissues to impute gene expression and pathway scores. CASTom-iGEx application to CAD allowed to

1. retrieve existing knowledge and highlight new putative mechanisms based on variants aggregation;
2. investigate type 1 error calibration for our TWAS and PALAS strategy as well as the effect from correlated genes in pathway association;
3. indicate possible intermediate endophenotype contributing to disease etiology and pointing at putative relevant genes and pathways;
4. cluster CAD cases from UK Biobank and find differences in terms of endophenotypes possibly related to disruption into specific genes/pathways and group-specific differences in treatment responses;

5. compare our gene derived clustering structure to stratification obtained from principal components only, hence investigating ancestry contribution to patient stratification
6. assess the genes, pathways, and endophenotype associations under the null hypothesis of no genetically derived partition via randomized clustering.

Section 4.4 Similarly, we applied CASTom-iGEx to PGC cohorts for SCZ, building PriLer models across 10 SCZ related tissues on harmonized reference panels (CMC or GTEx) with 36 European PGC cohorts variants. As before

1. we identified potentially dis-regulated pathways and differentially predicted genes;
2. we performed bidirectional Mendelian Randomization leveraging UK Biobank TWAS and PALAS summary statistics on SCZ related endophenotypes to understand putative causal ones as well as phenotypes variability due to SCZ genetic predisposition;
3. we jointly clustered 35 cohorts based on gene-level T-scores applying two filtering strategies (clumping at 0.9 and at 0.1 correlation to reduce MHC contribution) and identified group-specific pathways;
4. we used CAD UK Biobank and CARDIoGRAM as a proof of principle in defining cluster-reliable measures (CRM) for gene-risk score (gene-RS) differences as a proxy of endophenotype changes and defining a reliable cut-off to be applied for SCZ.
5. we built gene-RS on clustered cohorts leveraging SCZ related endophenotypes and we detected differences in derived phenotype-specific gene-RS, defining reliable ones based on CRM cut-off, with the goal of creating cluster-specific phenotype profile for each group;
6. we compared our gene derived clustering structure with partition derived from PCs, showing a minimal overlap.

4.1 Data description and pre-processing

We differentiate between two types of data sets: reference panels used to build PriLer models and genotype-only data sets to impute gene expression. In Fig. 4.1 it is summarized how reference panels and genotype-only data sets are matched and harmonized for each trait analysis. In particular, with the harmonization of genotype data, we indicate that variants for two or more data sets are filtered to include only SNPs and indels in common (i.e. same position and REF/ALT annotation) and such that the ALT allele frequency among all possible data sets pair is not bigger than 0.15. We hence ensure that the genetic data considered among different cohorts have a uniform allele frequency as well

as configuration.

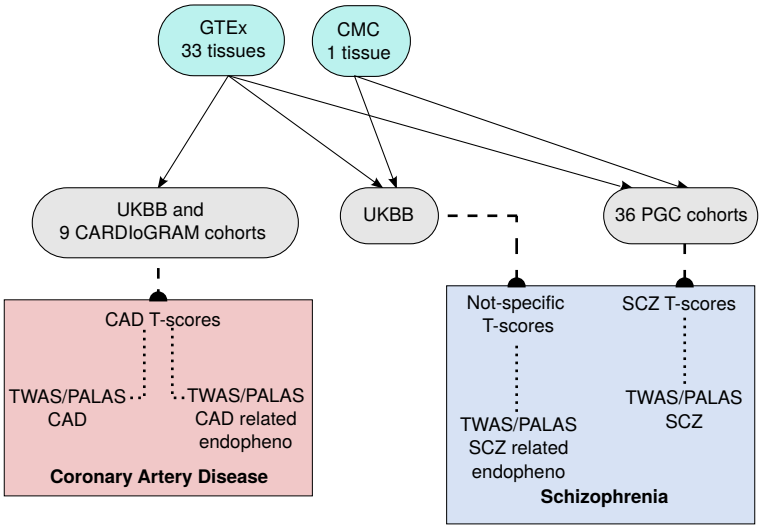

Fig. 4.1.: The diagram describes for each disease under investigation, which reference panels are used to build PriLer models and how different data sets are combined. First level block in light blue indicates reference panels, second level block in grey indicates genetic-only data sets. Each arrow among the two levels symbolizes an harmonization in terms of SNPs and indels between data sets. The last level is trait-specific and give insights in binary phenotype (if any) used to compute T-scores from which associations with genes/pathways are derived. CAD: coronary artery disease, SCZ: schizophrenia, PGC: Psychiatric Genomic Consortium, UKBB: UK Biobank, CMC: CommonMind Consortium

CAD analysis is based on the genetic harmonization of GTEx reference panel with UK Biobank (UKBB) cohort (used for discovery) and 9 CARDIoGRAM cohorts (used for replication). After PriLer model derivation from 11 CAD-related tissues in GTEx, genes are imputed on UKBB and CARDIoGRAM cohorts and T-scores are separately computed for each cohort. Due to the high number of participant in UKBB, we used the "large sample size" scenario as described in section 3.2.1 to compute gene T-scores and included as reference set 30% of the non-affected individuals by CAD with a total size of $92,784$ samples, bootstrapping across 10 folds. Gene T-scores so derived are also used to perform TWAS and PALAS analysis for CAD-related endophenotypes from UKBB, necessary in the Mendelian Randomization application. Instead for the CARDIoGRAM cohorts, we use the "small sample size" scenario creating a reference set composed of 80% of controls and randomly repeating the partitioning 40 times.

SCZ analysis instead relies on the harmonization of 36 PGC cohorts with GTEx or CMC reference panels (separately) in order to create a total of 10 tissue-specific gene expression models. T-scores are then computed separately for each cohort with 80% of controls used as reference set and 40 bootstrap repetitions. Since in SCZ analysis we leveraged UKBB rich phenotype collection, UKBB genotype-only data set is separately harmonized with GTEx or CMC (not jointly with 36 PGC cohort, see Fig. 4.1), imputing gene expression on the same SCZ-related tissues and converting it into gene T-scores that are built in a trait non-specific manner. Hence in this case, reference sets to compute gene T-scores are obtained randomly sampling 20% of the UKBB individuals across 10 repetitions, with each

of the repetition including $68,190$ participants.

4.1.1 Reference panels

PriLer prediction models for gene expression (section 3.1) are built on matched data sets of genotype individual dosages and gene expressions, shortly referred as reference panel. In particular, we used **GTEx v6p** [166] composed of 449 donors for a total of 7051 samples across 44 non-diseases post-mortem tissues and cell lines, and **Common Mind Consortium (CMC) Release1** [54] composed of 592 individuals with RNA-Seq data extracted from post-mortem Dorsolateral Prefrontal Cortex (DLPC) tissue. GTEx project, launched by the National Institutes of Health (NIH) in September 2010, is an ongoing effort that provide open-access data to build a comprehensive public resource to study tissue-specific gene expression and regulation. GTEx genotype, RNA-sequencing and additional covariate information was obtained via dbGaP accession number phs000424.v7.p2 including data for v7, however we applied PriLer to v6p in order to leverage SNPs Array data for genotype that was only available for that release. Similarly, CMC is a public–private partnership that collected autopsies of individuals with and without severe psychiatric disorders such as schizophrenia in order to create functional genomic data in specif brain region. Data for this publication were obtained from NIMH Repository & Genomics Resource, a centralized national biorepository for genetic studies of psychiatric disorders. Although GTEx included around 100 individuals having samples in 11 brain regions, we also used CMC in creating gene expression models being the largest existing collection of collaborating brain banks. Next, we briefly describe **genotype pre-processing** for the two reference panels. For GTEx, we started from GTEx_Analysis_2015-01-12_OMNI_2.5M_5M_450Indiv_chr1to22_genot_imput_info04_maf01_HWEp1E6_ConstrVarIDs.vcf.gz VCF table including genotyped and imputed SNPs and indels for 450 GTEx individuals using 1000 Genomes Project I vs3 as imputation panel. Among 450 individuals, 182 were genotyped on Illumina's HumanOmni5-Quad array and 268 on the HumanOmni2.5-Quad array. This given VCF file was already filtered for variants with minor allele frequency (MAF) < 0.01, imputation INFO score < 0.4 and Hardy-Weinberg Equilibrium (HWE) p-value < 0.000001 (see supplementary information [166]). The CMC genotype data instead, is obtained via Synapse portal at the controlled access data location `https://www.synapse.org/#!Synapse:syn3275221`. It was performed on the Illumina Infinium HumanOmniExpressExome 8 v1.1b and 668 samples were imputed with 1000 Genomes Phase I integrated panel (see Online Methods [54]). For both reference panels, we performed the following QC steps using PLINK software [19]:

1. we consider only 22 autosomal chromosomes,
2. REF and ALT alleles for both SNPs and indels are aligned to human reference genome hg19,
3. position with more than 1 alternative allele option are removed,

4. variants with INFO score < 0.8 are filtered,
5. non-common variants with MAF < 0.05 are filtered,
6. variants deviating from HWE with a p-value < 0.00005 are filtered.

Reference Panel	Tissue	Tissue short name	n. individuals	n. QCed genes	n. QCed variants
CMC	Dorsolateral Prefrontal Cortex	DLPC	478	15578	6491178
GTEx	Adipose Subcutaneous	AS	242	25971	6486416
	Adipose Visceral Omentum	AVO	164	25139	
	Adrenal Gland	AG	105	23624	
	Artery Aorta	AA	185	24274	
	Artery Coronary	AC	99	23880	
	Artery Tibial	AT	239	24335	
	Brain Caudate basal ganglia	BCbg	90	24512	
	Brain Cerebellar Hemisphere	BCH	77	23762	
	Brain Cerebellum	Bce	93	24570	
	Brain Cortex	BC	81	24110	
	Brain Frontal Cortex BA9	BFCB	77	23765	
	Brain Hippocampus	BHi	74	23723	
	Brain Hypothalamus	BHy	72	24426	
	Brain Nucleus accumbens basal_ganglia	BNabg	81	24386	
	Cells EBV-transformed lymphocytes	CE	94	21779	
	Colon Sigmoid	CS	118	24051	
	Colon Transverse	CT	145	25354	
	Esophagus Gastroesophageal Junction	EGJ	115	23575	
	Esophagus Mucosa	EM	224	25038	
	Esophagus Muscularis	EMusc	199	24360	
	Heart Atrial Appendage	HAA	151	23666	
	Heart Left Ventricle	HLV	172	22681	
	Liver	L	94	22158	
	Lung	Lu	241	27372	
	Muscle Skeletal	MS	297	22942	
	Pancreas	P	132	23153	
	Skin Not Sun Exposed Suprapubic	SnotSun	173	25922	
	Skin Sun Exposed Lower leg	SSun	252	26582	
	Small Intestine Terminal Ileum	SITI	74	25010	
	Spleen	Sp	79	24354	
	Stomach	S	144	24861	
	Thyroid	T	233	27305	
	Whole Blood	WB	280	22805	

Tab. 4.1.: Overview reference panels as input for PriLer after QC steps, matching with GWAS summary statistics and including only individuals with European ancestry.

In addition, because we included GWAS summary statistics as optional prior information in PriLer for disease-related tissues, genotype data of GTEx was matched with GWAS results for CAD from [118] and SCZ from [141]. As regards CMC instead, we matched imputed genotype with GWAS for SCZ due to its only usage in SCZ analysis (Fig. 4.1). In particular, matching GWAS summary statistics with a reference panel was performed by retaining solely SNPs having same position as well as REF/ALT annotation and indels having same position and length, when the latter is available. Finally, genotype imputation probability (oxford format) for each allele combination was then converted to a unique dosage value in $[0, 2]$ range where 0 refers to homozygous REF, 1 to heterozygous REF/ALT and 2 to homozygous ALT configurations. The final number of variants used in PriLer models after quality control steps and GWAS matching was respectively $6,486,416$ for GTEx and $6,491,178$ for CMC (see Tab. 4.1).

As regards, **RNA-sequencing (RNAseq) pre-processing**, we applied similar QC steps in the context of eQTL analysis from GTEx and CMC consortia respectively. In particular, for GTEx we started from phe000006.v2 data i.e. gene reads counts and Reads per Kilobase Million (RPKM)
(GTEx_Data_20150112_RNAseq_RNASeQCv1.1.8_gene_reads.gct.gz and
GTEx_Data_20150112_RNAseq_RNASeQCv1.1.8_gene_rpkm.gct.gz) that include gene expressions across 54 tissues for a total of 551 individuals. We excluded poor quality samples from v7 that were annotated with `SMAFRZE == 'EXCLUDE'`, considered only samples included in genotype data and excluded tissues annotated having less than 70 samples as well as sex-specific tissues (Testis, Vagina, Ovary, Uterus, Prostate) reducing to 39 tissues in 441 individuals. Afterward, following GTEx guidelines for eQTL analysis in v6p release [166], for each tissue

1. genes having RPKM > 0.1 in at least 10 individuals and number of reads ≥ 6 in at least 10 individuals were kept,
2. gene RPKM expression values were quantile normalized to the average empirical distribution observed across samples, and for each gene expression was inverse quantile normalized to a standard normal distribution across samples.

In addition, we excluded 4 tissues (Breast Mammary Tissue, Cells Transformed Fibroblasts, Nerve Tibial and Pituitary) due to the lack of matching with available prior information (see below for prior acquisition).
For CMC data instead, we used the already processed RNAseq data available at `https://www.synapse.org/#!Synapse:syn5607698` that correspond to "SVA corrected excluded ancestry" format for 592 individuals. Briefly, gene reads count were normalized with `voom` R package [190] without covariates to compute log Counts per million mapped reads (CPM) and only genes with at least 1 CPM in at least 50% of the samples were kept. Next, known and hidden covariates computed via surrogate variable analysis (SVA) are considered and voom is applied again to estimate confidence weights for each normalized observed read count by residualizing on the covariates. Finally, gene expression is adjusted for those hidden and known covariates by weighted-linear regression and adjusted expression is obtained as corresponding residuals. Note that we used the specific version "SVA corrected excluded ancestry" meaning that gene expression has been already adjusted for surrogate variables but not for ancestry that is included as covariates in PriLer model (see below), as Fromer et al. suggest for CMC eQTL analysis (see Online Methods [54] for details).
Finally, for both GTEx and CMC filtered genes were annotated via Ensembl on GRCh37 with `biomaRt` [191] to retrieve transcription starting site position (TSS) of each gene obtained as the starting site of the first transcript. Therefore, TSS corresponds to start position or end position, depending on forward/reverse strand location of that gene. The retrieved number of genes across all considered tissues is shown in Tab. 4.1.

As regards **covariates** included in PriLer models, we adhered to the guidelines of eQTL analysis in GTEx and CMC respectively. For GTEx reference panel, the following features were used as covariates in gene expression models: sex, genotype array platform, PEER components and first 3 principal components (PCs) derived from genotype data. In particular, PEER (probabilistic estimation of expression residuals) method [192] detects hidden batch effect and other potential confounders explaining the majority of gene expression variability and is computed independently for each considered tissue from normalized expression matrices. The number of PEER factors (F) must be decided a prior and we determined it as a function of tissue sample size (M_t) following GTEx approach that aimed at maximizing cis-eQTL discovery:

$$F = \begin{cases} 15 & \text{if } M_t < 150, \\ 30 & \text{if } 150 \leq M_t < 250, \\ 35 & \text{if } M_t \geq 250. \end{cases}$$

Instead, PCs are directly provided from GTEx, being computed on the 450 donor using EIGENSTRAT [193] implemented in Ricopili. The number of top PCs added as covariates was a priori set to 3 as they captured the majority of the population structure among GTEx individuals.
For CMC reference panel instead, since gene expression is already corrected for known and hidden cofactors, we only used as covariates in PriLer models 5 ancestry components directly provided which were computed via GemTools [194] on a set of high quality autosomal SNPs from pre-imputation genotype data. The number of components was suggested from CMC eQTL analysis in [54] as it was sufficient to describe the ancestry space.

Furthermore, we focused our analysis on individuals with European ancestry. Specifically, we build PriLer models on 377 donors from GTEx with reported race "white" and 478 donors (212 controls and 266 patients diagnosed with SCZ and schizoaffective disorders) from CMC with "Caucasian" reported ethnicity, see Tab4.1 for tissue specific sample distribution.

As regards variant-specific **prior information** to be incorporated in PriLer model, we used GWAS summary statistics for CAD and SCZ as well as epigenetic open-chromatin region information. In particular,

- GWAS summary statistics for CAD [118] and for SCZ [141] are binarized using 0.05 and 0.01 nominal p-values respectively. Two different thresholds are applied in order to obtain a comparable number of SNPs having GWAS specific prior (Fig. 4.2). This caution is taken since the initial number of variant intersecting a certain prior increases the prior weight starting value, although relevant prior will be eventually associated with an higher prior weight (see section 4.2.2).

- One-hot encoded open chromatin regions derived form ATAC-seq and ChIP-seq H3k27ac and specific for cell lines and tissues are used as epigenetic prior. ChIP-seq H3k27ac data is obtained from ENCODE and Epigenome Roadmap Project, ATAC-Seq profiles for heart related related tissues are obtained from [195] (GSE72696) and ATAC-Seq profiles for brain related tissues are extracted from human postmortem prefrontal cortex neuronal cells in [196] (GSE83345). The full sample list and GEO accession number is shown in Tab. B.1 and all these annotations can be downloaded from gitlab.mpcdf.mpg.de/luciat/castom-igex (refData/prior_features/).

Among all the prior features, we additionally modified 2 ATAC-Seq brain related prior features FPC_neuronal_ATAC_R2 and FPC_neuronal_ATAC_R4 because of their reduced number of putative genome-wide gene regulatory elements (GREs) in comparison to H3K27ac features, specifically $44,475$ and $34,883$ versus mean number of GREs $128,817.3$ across all cell types in H3k27ac prior. Thus, each GRE of the ATAC-Seq brain related prior features is enlarged by median length of GREs in H3k27ac data (i.e. $1,192$ bp). As a consequence, the number of variants intersecting those priors increases, becoming more comparable with H3k27ac data and hence suffering less of a lower prior weight assignment at the initial step (see section 4.2.2).
The resulting tissue-specific binary prior matrix contains column-wise prior features considered for that tissue, and has entry 1 in correspondence of a variant that intersects an open chromatin region for a certain cell type or passes a nominal p-value GWAS threshold (see section 3.1.1). The complete table with tissue-specific selection of priors is shown in Tab. B.2.

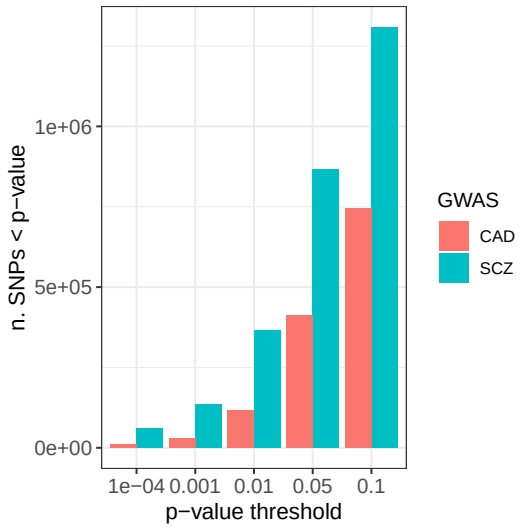

Fig. 4.2.: Number of GWAS hits for different p-value thresholds for CAD and SCZ, the number of variants passing a certain threshold is comparable in CAD for p-value = 0.05 and in SCZ for p-value = 0.01

Finally, as explained in section 3.1.2, prior weights are computed on a subset of all expressed genes called **cis-heritable** genes and estimated from GCTA software [32].

This list is obtained from `http://gusevlab.org/projects/fusion/` database of TWAS method [9] both for GTEx and CMC reference panels. We used the at the time latest results on GTEx involving genes with heritability p-value ≤ 0.01 estimated from GTEx v7 (`http://gusevlab.org/projects/fusion/weights/GTEX7.txt`).

4.1.2 Genotype-only data sets

Imputation of gene expression on genotype-only cohorts based on tissue-specific PriLer model first requires

1. a quality control step on genotype-only cohorts,
2. harmonization with a reference panel in order to build gene expression models on a set of variants available for both data sets.

As shown in Fig. 4.1, we used and separately pre-process 3 genotype-only data sets: UK Biobank (UKBB) cohort that contains a large collection of phenotypes for circa $500,000$ individuals, 9 CARDIoGRAM cohorts composed of patients and controls for CAD and 36 European cohorts from PGC including patients and non-affected individuals for SCZ (Tab. 4.2).

Data set	Data set usage	N. cohorts	N. samples	N. cases	Reference panel matched with	N. final variants
UK Biobank	SCZ related phenotypes	1	340,939	-	GTEx	5,728,140
					CMC	5,774,100
UK Biobank	CAD/ CAD related phenotypes	1	340,939	19,026/-	GTEx	4,257,718
CARDIoGRAM	CAD	9	26,681	13,279		
PGC	SCZ	36	55,419	24,764	GTEx	5,912,207
					CMC	5,934,252

Tab. 4.2.: Specifics of genotype-only data sets for which gene expression is imputed, n. of final variants refers to final set of SNPs and indels after QC steps and harmonization with reference panels.

In details, **UK Biobank** is a large long-term prospective biobank study that collected (at the time of writing this thesis) genetic and deep phenotyping data on approximately $500,000$ individuals at 22 sites in United Kingdom of age between 40 and 69 at recruitment [17]. For our analysis, we had access under application numbers 34217 and 25214 and downloaded imputed data from the 3^{rd} release as described in Resource 668. Similarly to reference panels QC steps, we first aligned REF/ALT allele to hg19 considering only autosomal chromosomes. We then excluded individuals that withdraw consent and with non-white British ancestry (info retrieved from ukb_sqc_v2.txt file). As regards post-imputation QC steps, variants to be excluded were found via QCTOOL v2 based on the following criteria:

1. call rate < 0.98,

2. INFO score < 0.8,
3. non-common with MAF < 0.05,
4. deviating from HWE with p-value $< 10^{-6}$,
5. multi-allelic positions.

Afterward, already computed kinship matrix (ukbA_rel_sP.txt file) was used to detect relatives up to 3^{rd} degree and the largest amount of not related individuals was retained following UKBB guidelines [17]. Finally, individuals with not concordant genotypically inferred and submitted sex as well as poor quality samples due to heterozygosity and missing rates detected as outliers were excluded (info stored in ukb_sqc_v2.txt file, Resource 531). On this set of high-quality variants, allele combination probabilities were then converted into dosages via PLINK, similarly to reference panels. In order to harmonize UKBB high-quality variants with after QC variants in GTEx and CMC, matched SNPs and indels having ALT frequency differences > 0.15 were excluded. The resulting set is composed of $340,939$ individuals with imputed genotype dosages for $5,728,140$ and $5,774,100$ variants when matching with GTEx and CMC panels respectively.

CARDIoGRAM consortium, is a collaborative effort that combines data from multiple large scale genetic studies to identify risk loci for coronary artery disease and myocardial infarction. We used a subset of 9 case/control cohorts of European ancestry among the available ones which could be accessed through the collaboration with Prof. Dr. med. Heribert Schunkert (a PI of the consortium). In the following list it is indicated the number of cases/controls in each cohort:

- German Myocardial Infarction Family Studies I (GerMIFSI) [197]: $622/1521$,
- German Myocardial Infarction Family Studies II (GerMIFSII) [198]: $1188/1238$,
- German Myocardial Infarction Family Studies III (GerMIFSIII) [199]: $1048/1419$,
- German Myocardial Infarction Family Studies IV (GerMIFSIV) [118]: $940/1128$,
- German Myocardial Infarction Family Studies V (GerMIFSV) [200]: $2392/1537$,
- LUdwigshafen RIsk and Cardiovascular Health Study (LURIC) [201]: $2085/591$,
- Cardiogenics (CG) [202]: $366/401$,
- Wellcome Trust Case Control Consortium (WTCCC) [202]: $1884/2871$,
- Myocardial Infarction Genetics Consortium (MIGen) [202]: $2827/2909$.

The following quality steps before imputation on the genotype data were performed separately for each cohort:

1. samples with call rate < 0.98 were excluded,
2. variants with call rate ≤ 0.98 were excluded,

3. variants with MAF < 0.01 were excluded,
4. samples with discordant recorded and genotype-derived sex were excluded,
5. samples detected as outliers based on two top dimensions from multidimensional scaling (MDS) i.e. deviating beyond mean ± 5 standard deviation were excluded,
6. individuals with relatives up to the fourth-degree based on identity-by-descent (IBD) matrix where excluded (PI_HAT ≥ 0.0625),
7. samples with heterozygosity rate beyond mean ± 3 s.d. were excluded,
8. variants with HWE p-value $< 10^{-6}$ were discarded.

Afterward, for each cohort imputation was performed using Haplotype Reference Consortium panel on the Sanger Imputation Server (`https://www.sanger.ac.uk/science/tools/sanger-imputation-service`). We then performed the following post-imputation QC steps independently for each cohort:

1. variants with call rate ≤ 0.98 were excluded,
2. variants with MAF < 0.05 were excluded,
3. variants with HWE p-value $< 10^{-6}$ were excluded,
4. variants with INFO imputation score < 0.8 were excluded,
5. variants having multi-allelic positions were excluded,
6. related samples with PI_HAT ≥ 0.0625 based on IBD analysis were discarded.

Since individuals could overlap among cohorts, we performed IBD analysis inter-cohorts using pre-imputation QCed set of variants and removed individuals with up to forth degree relatives (PI_HAT ≥ 0.0625) favouring samples annotated as cases and/or with a lower SNPs missing rate. This yielded to a total of $26,681$ non related and high quality samples among which $13,279$ were affected by coronary artery disease. The final set of variants instead was extracted as the harmonized set across all 9 cohorts, also matching with UKBB and GTEx data sets and for which ALT frequency differences for any possible pair of data sets in 9 cohorts plus UKBB plus GTEx did not pass 0.15 threshold. Hence, the number of variants harmonized across CARDIoGRAM cohorts, UKBB and reference panel GTEx used to study CAD is composed of $4,257,718$ SNPs and indels.

Finally, as regards **PGC** cohorts, we request access to wave 2 composed imputed genotype data for 36 European ancestry cohorts with phenotypic information of SCZ affected individuals and controls [141]. We adhered to PGC guidelines and performed the following post-imputation QC steps for each cohort:

1. excluded variants with MAF < 0.01,
2. excluded variants with INFO imputation score < 0.6,

3. excluded variants in multi-allelic position,
4. excluded variants missing in at least 20 samples, i.e. genotype certainty < 0.8,
5. removed individuals with not available diagnosis and having related and/or duplicated samples.

MAF and INFO filtering threshold were lowered compared to previous data sets to not excessively penalize variants filtering due to matching of high number of heterogeneous cohorts. Hence, we also increase the set of variants for reference panels (CMC and GTEx) to be harmonized with PGC cohorts including SNPs and indels with INFO ≥ 0.6 and MAF ≥ 0.01 computed on Caucasian individuals only. After harmonization and including only variants with ALT frequency differences among each possible pair of data set plus a reference panel ≤ 0.15, the final set is composed of $5,912,207$ and $5,934,252$ SNPs and indels when matching with GTEx and CMC respectively, across $55,419$ individuals.

A summary of number of variants used to build and impute gene expression PriLer models, number of individuals and data set usage for genotype-only data set is shown in Tab. 4.2.

4.1.3 Phenotypes in UK Biobank

Different from CARDIoGRAM and PGC data sets that include only the phenotypic information related to the disease under investigation, UK Biobank (UKBB) contains deep phenotyping resources about lifestyle and health conditions. Leveraging this rich collection, we used UKBB in 3 different contexts: to define CAD phenotype, to extract CAD related endophenotypes and to extract SCZ related endophenotypes (Fig. 4.1).
Coronary artery disease diagnosis was determined using the stricter definition (CAD HARD) as described in [120] which combines self-reported questionnaire answers (data field 20002) on heart attack/myocardial infarction, percutaneous transluminal coronary angioplasty (PTCA) +/- stent, coronary artery bypass grafts (CABG) and triple heart bypass as well as hospital episodes ICD10 or ICD9 coded (data fields 41270 and 41271) on myocardial infarction and ischaemic heart diseases (I21-I24 or 410-412), old myocardial infarction (I25.2) and OPCS-4 codes (data fields 41272) for procedures as CABG (K40-K46) and PTCA (K49-K50, K75).
All the other phenotypes available under application numbers 34217 and 25214 were instead processed using PHESANT software [203]. PHESANT is a tool specifically tailored for UK Biobank that performs phenome scans testing the association of a trait with a comprehensive set of phenotypes. We use PHESANT to process UKBB phenotypes in an automatic manner via a rule-based system that determine the appropriate coding and consequently conversion of each phenotype. Specifically, UKBB categorizes phenotypes as either continuous, integer, categorical (single) or categorical (multiple) and according to this initial state and actual data distribution, PHESANT converts them as continuous,

ordered categorical, unordered categorical or binary after an initial filtering step that removes constant phenotypes or recorded ones in less than 500 samples. Continuous variables and integer ones with more than 20 distinct values are inverse-rank normalized and annotated as continuous. Integer variables with less than 20 values, categorical single variables with a natural ordering and continuous ones with more than 20% of individuals having the same value are annotated as ordered categorical. Categorical single without a natural order are annotated as unordered categorical and integer as well as categorical single with 2 distinct values and categorical multiple are converted into a binary variable per category and annotated as binary. Based on PHESANT assignment, we applied the correct GLM during trait association of genes and pathways i.e. Gaussian when trait is continuous, ordinal logistic regression when trait is ordinal categorical, and binary logistic when trait is unordered categorical or binary (section 3.2.2). Hence, PHESANT processed UKBB phenotypes are used in the subsequent analysis, unless differently stated such as in the hypothesis-driven endophenotype analysis for CAD clustering (see section 4.3.6). Finally, blood biochemistry phenotypes Lymphocyte-to-Monocyte ratio (LMR), Platelet-to-Lymphocyte (PLR), Neutrophil-to-Lymphocyte ratio (NLR) and Eosinophil-to-Lymphocyte ration (ELR) were derived by us from the original lymphocyte (data-field 30120), monocyte (data-field 30130), platelet (data-field 30080), neutrophil (data-field 30140) and eosinophil (data-field 30150) counts transforming them in the proper phenotypes ratio and processing the output via PHESANT.

The complete list of UKBB phenotypes used in this thesis for correlation and Mendelian Randomization analysis are in Tab. B.4 (CAD) and Tab. B.8 (SCZ). For cluster-specific endophenotype analysis, details can be found in Tab. B.6, B.7 (CAD) and Tab. B.11 (SCZ). The tables include UKBB data-field of the considered phenotypes and the covariates applied to correct for.

4.2 PriLer benchmark and validation

We initially created PriLer gene expression models across 34 tissues (Tab. 4.1) without intersecting with additional genotype-only data set to evaluate and benchmark our new methodology. First, we observe the variance captured by PriLer models and its dependency from sample size and number of priors. Then, we investigated the relevance of inferred prior features weights via simulation of random prior features using as example artery coronary GTEx tissue. Furthermore, we benchmarked our new methodology against elastic-net regression (directly built in PriLer pipeline, Fig. 3.5) also estimating the robustness of PriLer in selecting reg-SNPs via re-sampling. Finally, we compare our results with two widely used methods for gene expression cis-effects modeling: Fusion [9] and prediXcan [10].

4.2.1 PriLer explained gene expression variability

After applying PriLer pipeline as described in section 3.1 to 34 tissues, we defined for each tissue t a set of reliable genes $\mathcal{G}^t_{rel}$ based on R^2 estimation from genetic component that should explain at least 0.01 on the final model and more than 0 variability on the test folds from cross-validation (see def. (3.17)). We observed that the cardinality of $\mathcal{G}^t_{rel}$ largely depends on the number individuals in the tissue-specific model (Fig. 4.3A) with a Pearson correlation of 0.8538 ($P = 1.4^{-10}$) that becomes even more emphasized when considering the fraction of reliable genes with respect to the total number of QCed genes (Fig. 4.3B, corr.$= 0.9485$, $P = 1.6^{-17}$). Instead, there is no evidence of relationship between number of training samples and number of SNPs regulating at least 1 gene (i.e. not null regression coefficient), from now on referred as reg-SNPs (Fig. 4.3C, corr.$=-0.0720$, $P = 0.69$).

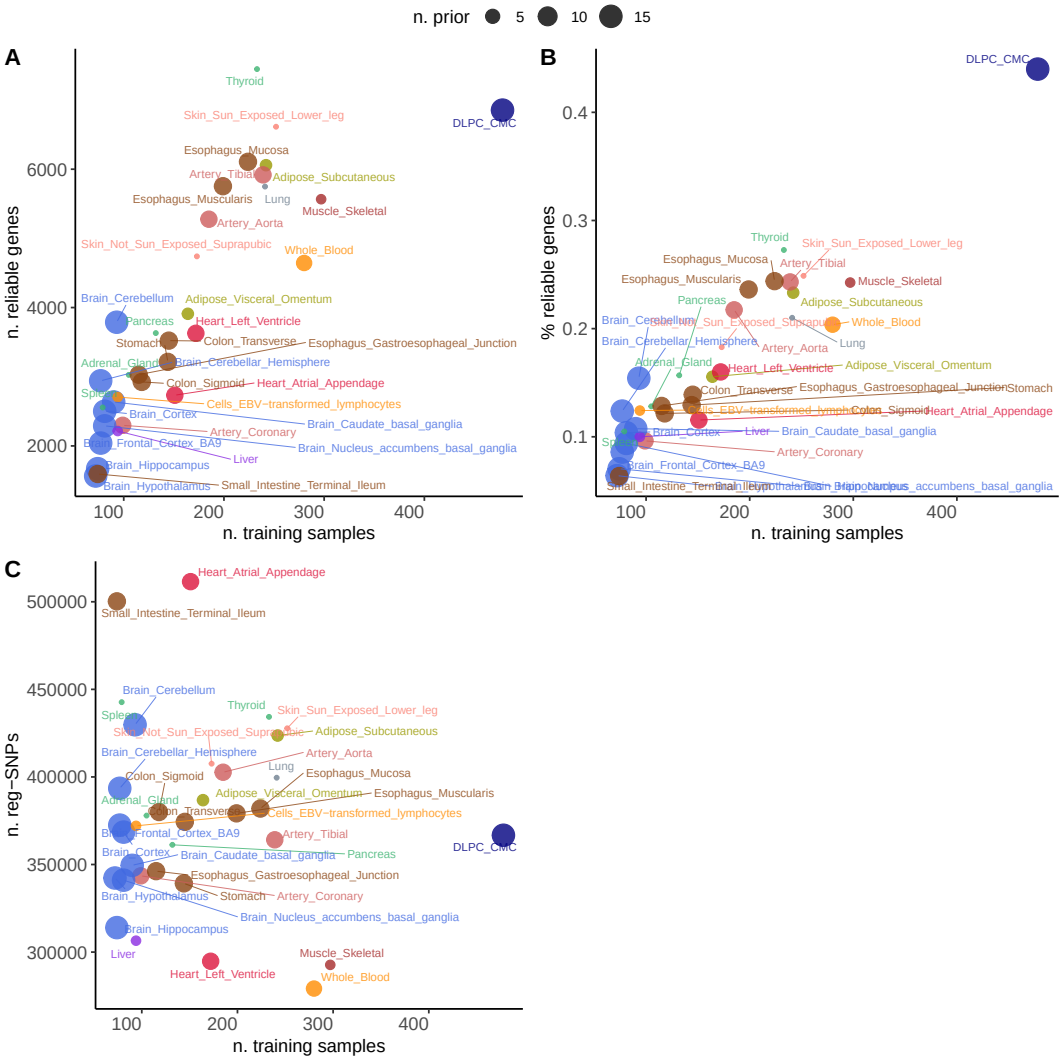

Fig. 4.3.: Across 34 tissue-specific PriLer models, comparison of n. of individuals in the model and **(A)** n. of reliable genes, **(B)** fraction of reliable genes with respect to total number of QCed genes, **(C)** n. of reg-SNPs. The dot size refers to the number of prior in a tissue-specific model and dot color to macro tissue categories

Details on tissue-specific models from PriLer can be found in Tab. 4.3

Tissue	n. individuals	n. prior	n. QCed genes	n. reliable genes	n. reg-SNPs	fraction reg-SNPs with prior
DLPC_CMC	478	15	15578	6854	366706	0.367
Adipose_Subcutaneous	242	3	25971	6058	423635	0.259
Adipose_Visceral_Omentum	164	3	25139	3910	386744	0.277
Adrenal_Gland	105	1	23624	3027	377947	0.235
Artery_Aorta	185	7	24274	5277	402688	0.385
Artery_Coronary	99	7	23880	2298	343528	0.392
Artery_Tibial	239	7	24335	5918	364109	0.419
Brain_Caudate_basal_ganglia	90	15	24512	2635	349538	0.396
Brain_Cerebellar_Hemisphere	77	15	23762	2941	393519	0.391
Brain_Cerebellum	93	15	24570	3788	429782	0.365
Brain_Cortex	81	15	24110	2501	368497	0.386
Brain_Frontal_Cortex_BA9	77	15	23765	2041	372502	0.38
Brain_Hippocampus	74	15	23723	1671	313967	0.432
Brain_Hypothalamus	72	15	24426	1565	342292	0.373
Brain_Nucleus_accumbens_basal_ganglia	81	15	24386	2290	341053	0.369
Cells_EBV-transformed_lymphocytes	94	2	21779	2706	372078	0.262
Colon_Sigmoid	118	8	24051	2925	379960	0.331
Colon_Transverse	145	8	25354	3522	374357	0.305
Esophagus_Gastroesophageal_Junction	115	8	23575	3030	346204	0.286
Esophagus_Mucosa	224	8	25038	6107	381943	0.31
Esophagus_Muscularis	199	8	24360	5754	379200	0.295
Heart_Atrial_Appendage	151	7	23666	2733	511496	0.338
Heart_Left_Ventricle	172	7	22681	3628	294768	0.451
Liver	94	2	22158	2215	306512	0.33
Lung	241	1	27372	5749	399559	0.204
Muscle_Skeletal	297	2	22942	5566	292706	0.337
Pancreas	132	1	23153	3631	361244	0.227
Skin_Not_Sun_Exposed_Suprapubic	173	1	25922	4740	407576	0.134
Skin_Sun_Exposed_Lower_leg	252	1	26582	6614	427755	0.138
Small_Intestine_Terminal_Ileum	74	8	25010	1594	500251	0.286
Spleen	79	1	24354	2556	442687	0.199
Stomach	144	8	24861	3215	339228	0.297
Thyroid	233	1	27305	7447	434307	0.184
Whole_Blood	280	6	22805	4644	279175	0.357

Tab. 4.3.: Summary of PriLer output on 34 tissues

To investigate the relationship between sample size and final set of reliable genes without the tissue-specific complexity, we additionally performed a down-sampling analysis on DLPC tissue creating a total of 5 different models built on randomly extracted $50, 100, 150$ non-affected individuals (*Control50, Control100, Control150*), the entire set of non-affected individuals (*ControlAll*) and all the available samples (*All*). The number of reliable genes increases with sample size, although not linearly and possibly reaching a plateau (Fig. 4.4 A). Instead, the number of reg-SNPs is relatively stable except for *Control50* being based on almost double the amount of reg-SNPs compared to the other DLPC models (Fig. 4.4 B), probably due to an overfitting in model selection when sample size is too small. Note that 50 samples is an extreme case that implies PriLer model inside the inner-CV is built on only 32 samples. However, we never reach this situation since we initially filtered tissues with less than 70 samples.

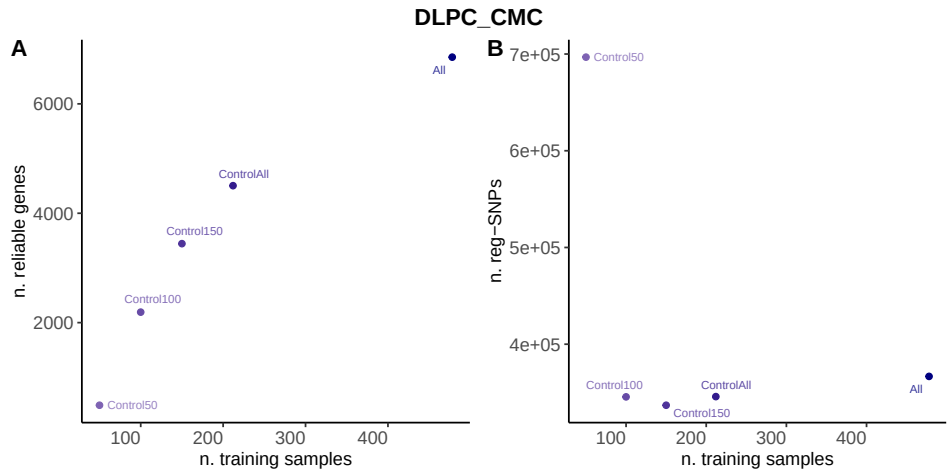

Fig. 4.4.: Down-sampling of DLPC, comparison of n. of individuals across 5 PriLer models with increasing sample size and **(A)** n. of reliable genes, **(B)** n. of reg-SNPs

As regards the variability captured by PriLer model for each reliable gene, we are interested in R^2_g metric representing the variability of gene expression explained by genetic components (see def. (3.14)), from now on simply indicated as R^2. In particular, Fig. 4.5 shows R^2 distribution for reliable genes in a certain tissue considering the final gene expression model **(A)** and out-of-sample R^2 estimated as average across test folds in the CV setting (see def. (3.15)) **(B)**. As expected, estimates of R^2 on test folds, R^2_{test}) are lower than the ones built on final model. In addition, R^2 evaluation from the final model tends to increase with a decrease in sample size (Pearson corr. between n. samples and median $R^2 = -0.6469$, P$= 3.5^{-5}$), possibly showing an overfit in model evaluation nonetheless attenuated when estimating on test folds (corr$= -0.4062$, P$= 0.02$). Note that, R^2 estimates exceed 1 for a small number of genes, indeed R^2_g computation does not assure an upper bound to 1 that however is satisfied when summing all the 3 components in def. (3.14).

As before, we investigated R^2 distribution when PriLer model has a varying sample size in the context of DLPC tissue. This down-sampling analysis confirms an over estimate in R^2 when reducing sample size, particularly in the limit case of 50 individuals (Fig. 4.6 A), that is attenuated for R^2_{test} (Fig. 4.6 B). In addition, when only considering the 264 genes simultaneously reliable across all 5 models, the overfit in final model due to lower sample size is still observable (Fig. 4.6 C), however in this case PriLer better captures gene expression variability when estimated on test folds with increasing in sample size (Fig. 4.6 D). Hence, the higher R^2_{test} distribution that we observed when including all reliable genes (Fig. 4.6 B), is related to the different gene set that increases in size with the number of individuals and hence includes more genes with lower cis-SNPs variability.

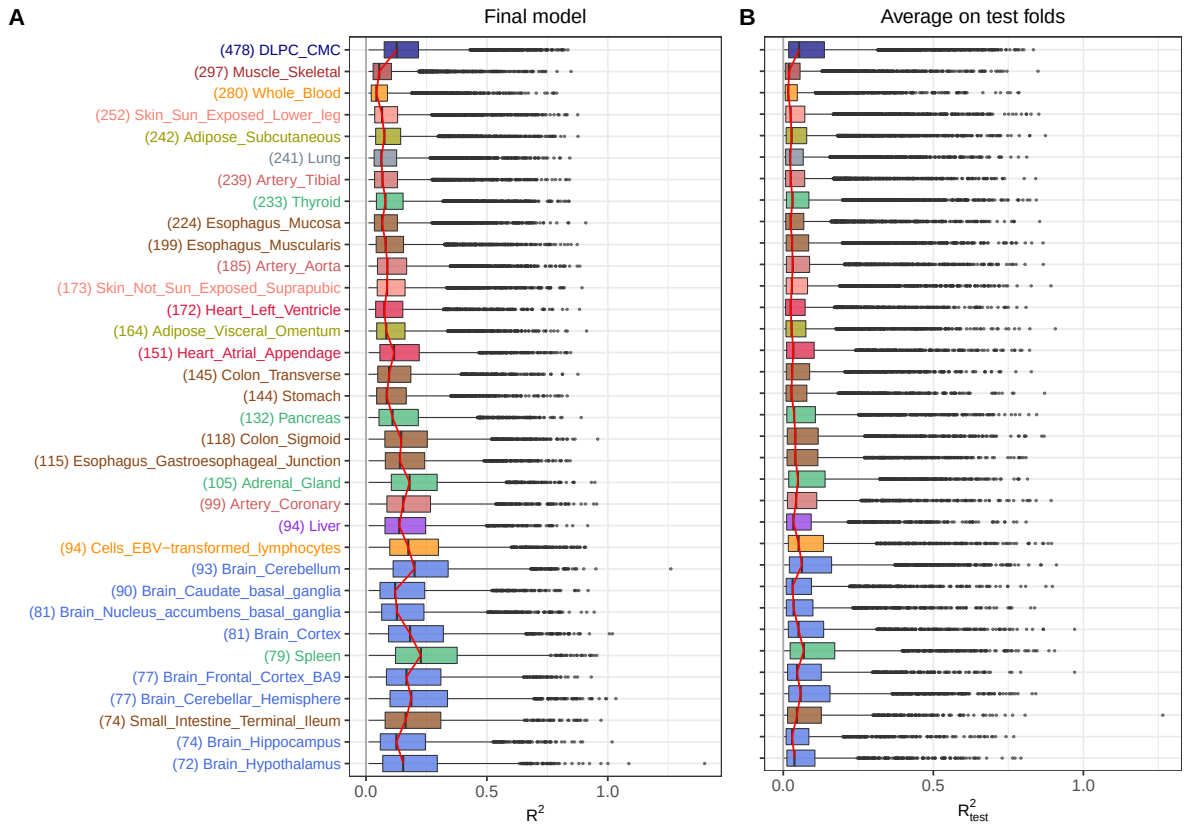

Fig. 4.5.: For each tissue-specific PriLer model, **(A)** R^2 and **(B)** R^2_{test} distribution being the average R^2 across CV test folders, for reliable genes. Tissues are sorted for sample size in descending order (also shown in brackets) and red line connecting median is displayed.

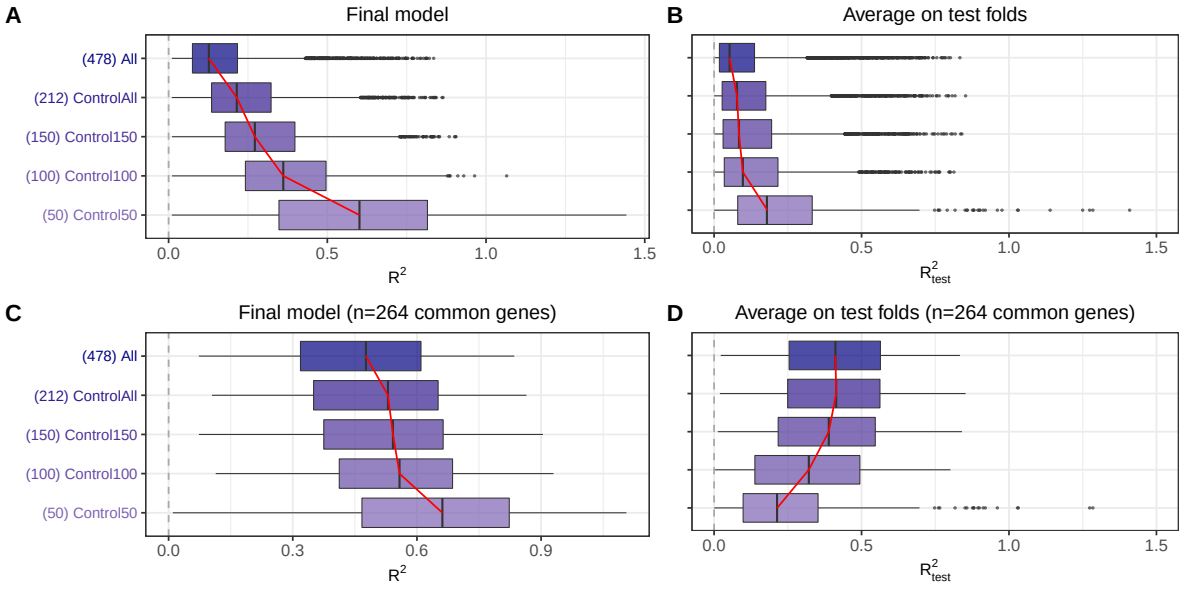

Fig. 4.6.: Down-sampling of DLPC, **(A)** R^2 and **(B)** R^2_{test} distribution being the average R^2 across CV test folders for reliable genes, and **(C)** R^2 and **(D)** R^2_{test} distribution when including only reliable genes in common across the 5 models. Models are sorted for sample size in descending order and red line connecting median is displayed. X-axes are capped at 1.5, removing 5 and 1 genes for Control50 in (B) and (D) panels respectively.

Since the final set of reliable genes is composed of both cis-heritable and non cis-heritable genes a priori defined via GCTA software (see section 3.1.2), we explicitly observed the differences in predictive performances of the two gene classes (Fig. 4.7). As explained in section 3.1.2, prior weights and consequentially variant-specific prior coefficients are estimated solely based on cis-heritable genes but a gene expression prediction model is built also for non cis-heritable ones imposing the already computed prior coefficients. As expected, the predominant class among reliable genes is the cis-heritable one (Fig. 4.7B) and the variance explained by the genetic component for non cis-heritable genes (R^2_{test}) is significantly lower across all tissues (Fig. 4.7A, Wilcoxon-Mann-Whitney $P \leq 2.21^{-49}$).

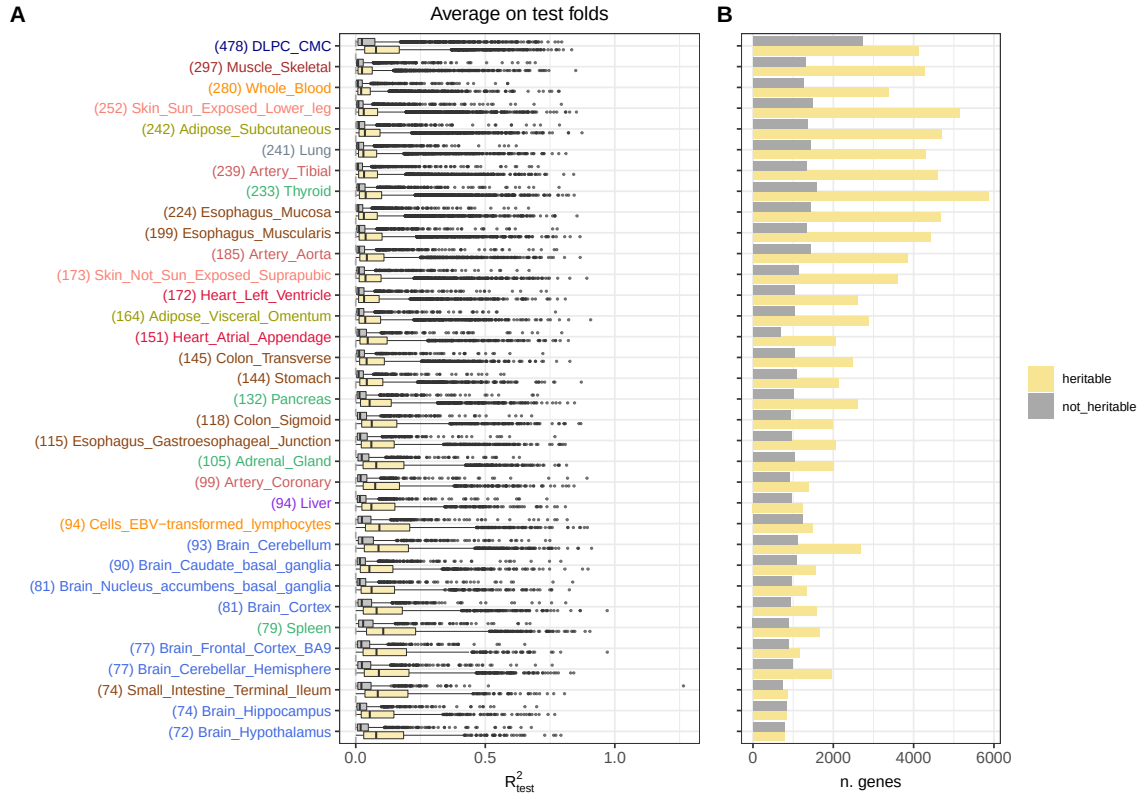

Fig. 4.7.: (Adapted from Trastulla et al., in prep.) For each tissue-specific PriLer model, **(A)** R^2_{test} distribution of reliable genes divided per cis-heritable and non cis-heritable status and **(B)** the corresponding number of genes. Tissues are sorted per sample size in descending order (also shown in brackets).

4.2.2 Prior weights validation via simulation

In each tissue-specific PriLer model, prior weight γ_k associated with k^{th} prior feature (e.g. intersection with a tissue specific open chromatin region) are automatically learned by the algorithm (section 3.1.1). A high value assigned to γ_k indicates that feature k increases the likelihood of variants intersecting that feature to be regulatory for gene expression. Hence, prior weight resulting values and intermediate ones from iterative steps provide additional knowledge in tissue-specific regulatory mechanisms. To validate this notion, we considered artery coronary tissue as example that includes 7 baseline priors (Tab. B.2) and

we randomly simulated prior features emulating open chromatin cell-type specific prior *heart_left_ventricle* in two contexts: randomly selecting variants and randomly selecting gene regulatory elements (GREs). In the first case, we aim at resembling a prior that is not biologically meaningful but contains the same amount of information or twice of an existing one. In the second case, we aim at resembling ChIP-Seq H3k27ac data used to build the baseline prior information in order to capture intrinsic genome structure such as LD.

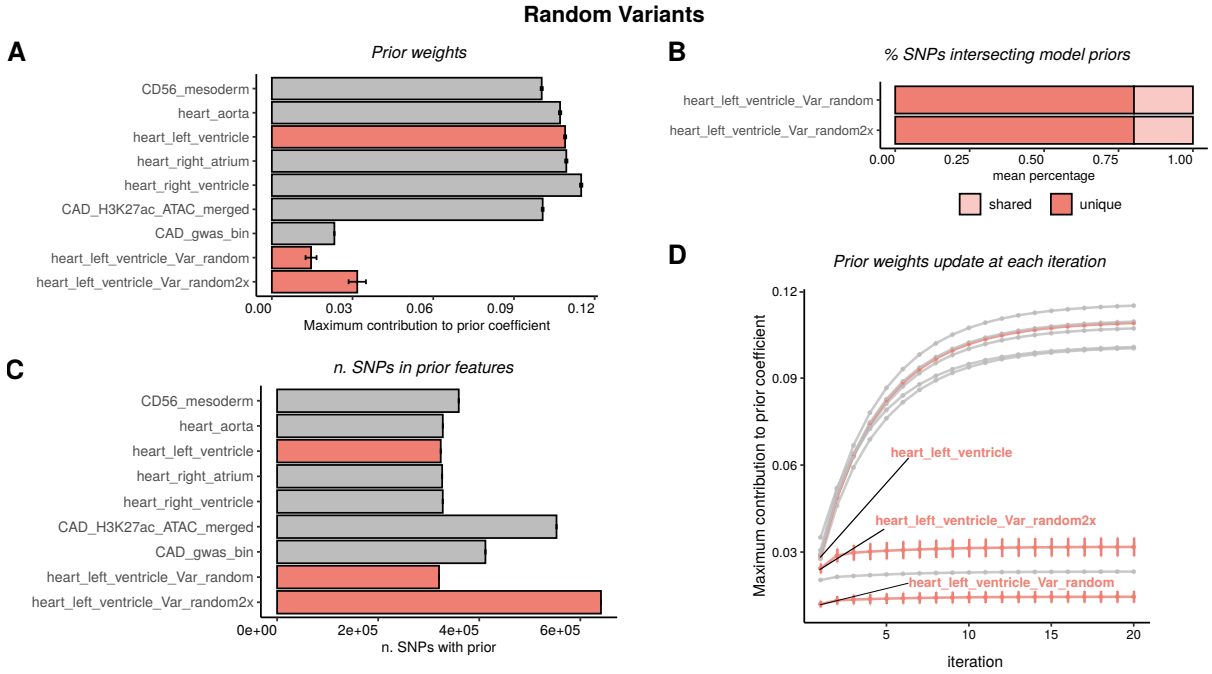

Fig. 4.8.: (Adapted from Trastulla et al., in prep.) Two prior features *Var_random* and *Var_random2x* are simulated 50 times in the same size/twice of prior *heart_left_ventricle* (pink). PriLer model for artery coronary tissue is created incorporating the other 6 baseline prior features. **(A)** Mean $\pm$ SD computed prior weights for each prior feature, **(B)** mean percentage of variants in random priors features that are in common (shared) or do not overlap (unique) with baseline features, **(C)** mean $\pm$ SD number of variants intersecting a prior features, **(D)** mean $\pm$ SD prior weight update at each iterative step.

Initially, we created two random prior features choosing n variants in the same size or twice of *heart_left_ventricle*, called with suffix *Var_random* and *Var_random2x* respectively, repeating the process 50 times. The overall count of variants intersecting a certain prior feature is shown in Fig. 4.8C. The resulting prior weights (Fig. 4.8A) were always lower than the baseline *heart_left_ventricle* (mean $\pm$ s.d $= 0.109 \pm 4.01^{-4}$) and close to zero, however not significantly different from it (mean $\pm$ s.d $= 0.0145 \pm 2.04^{-3}$ and 0.0317 ± 3.22^{-3} for *Var_random* and *Var_random2x* respectively). Indeed, we observed that the intersection among randomly selected variants and variants having at least one baseline prior was not null and on average 20% were shared (Fig. 4.8B). It becomes evident that the initial estimate of prior weight for a feature that intersects even a reduced amount of reg-SNPs will not be precisely zero. Instead, the more variants a prior feature intersects (prior size), the higher the likelihood to cross a reg-SNPs just by chance, leading to the extreme case of maximum weight assigned to a prior that cover the entire genome. On the

other hand, weights for not relevant priors do not increase in the iterative procedure, as is shown in Fig. 4.8D where *Var_random* and *Var_random2x* remained stable after a slight growth at iteration 1 instead of a 3-fold increase that can be observed for the original *heart_left_ventricle*. We conclude that variants in *heart_left_ventricle* open chromatin region led to a better performance in predicting gene expression contrary to the ones intersecting the randomly generated priors.

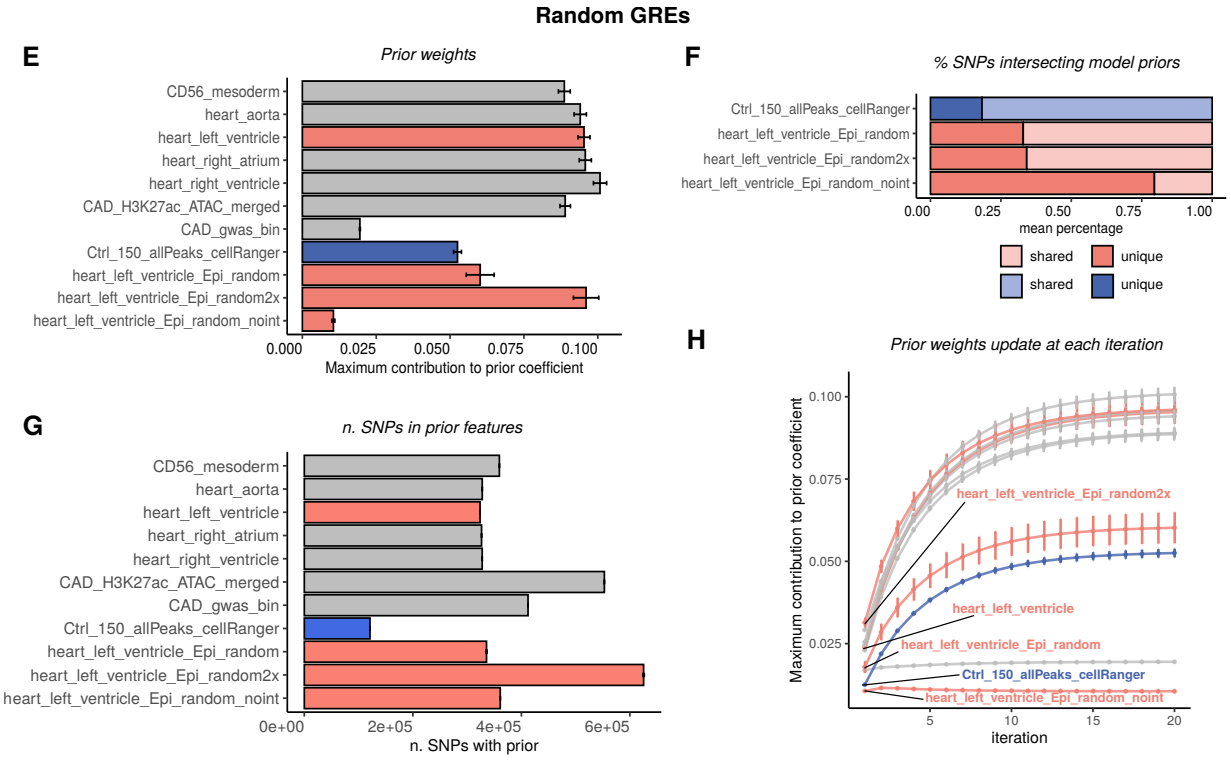

Fig. 4.9.: (Adapted from Trastulla et al., in prep.) Open chromatin regions (i.e. gene regulatory elements GREs) are randomly extracted 50 times among ChIP-Seq H3k27ac, in the same number or twice as *heart_left_ventricle* and intersected with variants to create *Epi_random* and *Epi_random2x* respectively. In addition, *Epi_random_noint* is created sampling 50 times GREs among the ones not in 6 baseline priors for artery coronary tissue in the same size as *heart_left_ventricle* (pink). We additionally include a brain tissue related prior *Ctrl_150_allPeaks_cellRanger* (blue). **(E)** Mean ± SD computed prior weights for each prior feature, **(F)** mean percentage of variants in random priors features and brain related that are in common (shared) or do not overlap (unique) with baseline features, **(G)** mean ± SD number of variants intersecting a prior features, **(H)** mean ± SD prior weight update at each iterative step.

Secondly, we randomly extracted open chromatin regions (i.e. GREs) from ChIP-Seq H3k27ac data across all tissues and cell-type origin and repeated the selection 50 times. To maintain the comparison with *heart_left_ventricle* prior, we selected n and $2n$ GREs with n being the number of open chromatin region in the original *heart_left_ventricle*, and then intersected them with variant position to create the binary prior features *Epi_random* and *Epi_random2x* respectively. Additionally, we introduced a prior feature used in PriLer brain tissue models *Ctrl_150_allPeaks*. Because the percentage of variants of these 3 additional priors shared with the baseline ones was higher than 65% (Fig. 4.9F), we also generated a random prior called *Epi_random_noint* selecting GREs among the ones not used in the

baseline prior features but still in the same number as *heart_left_ventricle*, reducing the overlap to 20% (Fig. 4.9F). In simulating GREs, the number of variants with *Epi_random* or *Epi_random_noint* prior was similar to the original *heart_left_ventricle* and twice for *Epi_random2x* across 50 repetitions (Fig. 4.9G), however not precisely the same as in random variant simulation scenario. As expected from the high percentage of shared variants, the prior weights estimates for *Epi_random*, *Epi_random2x* and *Ctrl_150_allPeaks* were very different from zero (mean $\pm$ sd $= 0.06 \pm 0.005$, 0.096 ± 0.004, 0.052 ± 0.001) and even close to the actual one (0.095 ± 0.002) in 2x case (Fig. 4.9E). On the other hand, *Epi_random_noint* that shares only 20% of variants with baseline priors had estimated prior weight close to zero (0.01 ± 0.0004, Fig. 4.9E), remaining stable during the iterative process (Fig. 4.9H). In addition, despite *Epi_random2x* having twice size and consequentially higher prior weight assignment at iteration 0, it converges to the same value of *heart_left_ventricle* (Fig. 4.9H). Finally, we observe that the learned relevance for brain-related prior *Ctrl_150_allPeaks* achieved high values (mean $\pm$ sd $= 0.05 \pm 3.33^{-4}$) even showing an incremental growth in the iterative updates (Fig. 4.9H). However, this evidence of relevance was confounded by the high percentage of shared variants with baseline priors (82%, Fig. 4.9F).
We can conclude that prior weights are representative of tissue-specific gene expression regulation but weights could be inflated when priors are composed of a similar variant set than the actual relevant features. Nonetheless, with a marginal sharing of variants, weights for non relevant priors are close to zero both in a complete random sampling (Fig. 4.8) and when LD structure is accounted for (Fig. 4.9).

4.2.3 Comparison of PriLer with elastic-net

As explained in section 3.1, our new methodology PriLer is an extension of elastic-net regression that includes prior information on single SNPs and indels. Hence, a natural benchmark is a comparison between PriLer and the built-in elastic-net (enet) regression models from PriLer pipeline (Fig. 3.5) across the 34 GTEx and CMC tissues.
The cardinality of the tissue-specific set composed of reliable genes $\mathcal{G}^t_{rel}$ was similar between the two methods but always higher in PriLer (Fig. 4.10A) with an average of 85.94 more genes per tissue (SD 47.39) and a total of $2,992$ additional genes. As pointed out in section 4.2.1, the number of reliable genes depends on training tissue sample size. However, the increase in the number of reliable genes in PriLer was not determined by sample size (Pearson correlation$= 0.20$), nor by the number of priors included in PriLer models (corr.$= 0.16$).
Considering only reliable genes in PriLer, we compared model performances based on out-of-sample variance explained by genotype as average across test folders R^2_{test} (3.15) and we recorded the number of genes with better performances in PriLer (Fig. 4.10B). Across all tissues, more than 50% of the genes had significantly better performances in PriLer, with a range going from 61% for small intestine (one-sided sign test P$= 1.55^{-17}$) to

53% for skin not exposed (P$= 2.11^{-4}$), and overall better performance for 56.9% of genes across all tissues (P$= 1.48^{-323}$). Part of this improvement is explained by the number of prior features used in a tissue model (correlation between the fraction of improved genes in PriLer and n. of prior features$= 0.48$) but only marginally inversely dependent on the number of training samples (corr.$= -0.28$).
When looking at the extent of differences for explained variance R^2 between PriLer and enet, we noticed similar ranges for the two methodologies both for R^2 on the complete training samples and on test folders from CV (Fig. 4.11). Better performances were more pronounced on the final model and tended to increase as the training sample size decreased (Fig. 4.11A), reflecting an advantage in using PriLer when the sample size is limited. An increase in R^2_{test} was less evident (Fig. 4.11B), nevertheless significantly different from zero when testing via WMW paired test (P$< 5^{-12}$ for all tissues).
Beyond model performances in terms of variance explained, we noticed that the number of reg-SNPs across all genes detected in PriLer was always lower than those from elastic-net with an average of $43,014$ fewer variants (SD $14,530$) for a total of $1,462,466$ (Fig. 4.10C). This decrease was also connected to the number of prior features in a tissue model (corr.$= -0.68$). As expected, reg-SNPs selected in PriLer were composed of a higher fraction of variants intersecting any prior information for prior features in corresponding tissue model (Fig. 4.10D, mean $\pm$ SD$= 11\% \pm 3.32\%$).
Lastly, we examined PriLer robustness in selecting reg-SNPs compared to the built-in elastic-net, down-sampling whole blood tissue (n.samples $= 280$) 10 times to create new reference panels made of 100 individuals. To evaluate the results, we computed the Jaccard index (JI) between each possible pair of repetition of reg-SNPs selection using a binary representation of all variants with 1 indicating the variant regulates at least a gene and 0 otherwise. We observed that PriLer JI distribution was significantly increased representing a better agreement in selecting regulatory variants (WMW P$= 2.8^{-14}$).

In conclusion, when comparing PriLer to the built-in elastic-net without prior information, our new methodology achieves better results in terms of prediction performances evaluated via cross-validation with a smaller selection but more biologically meaningful and robust variants.

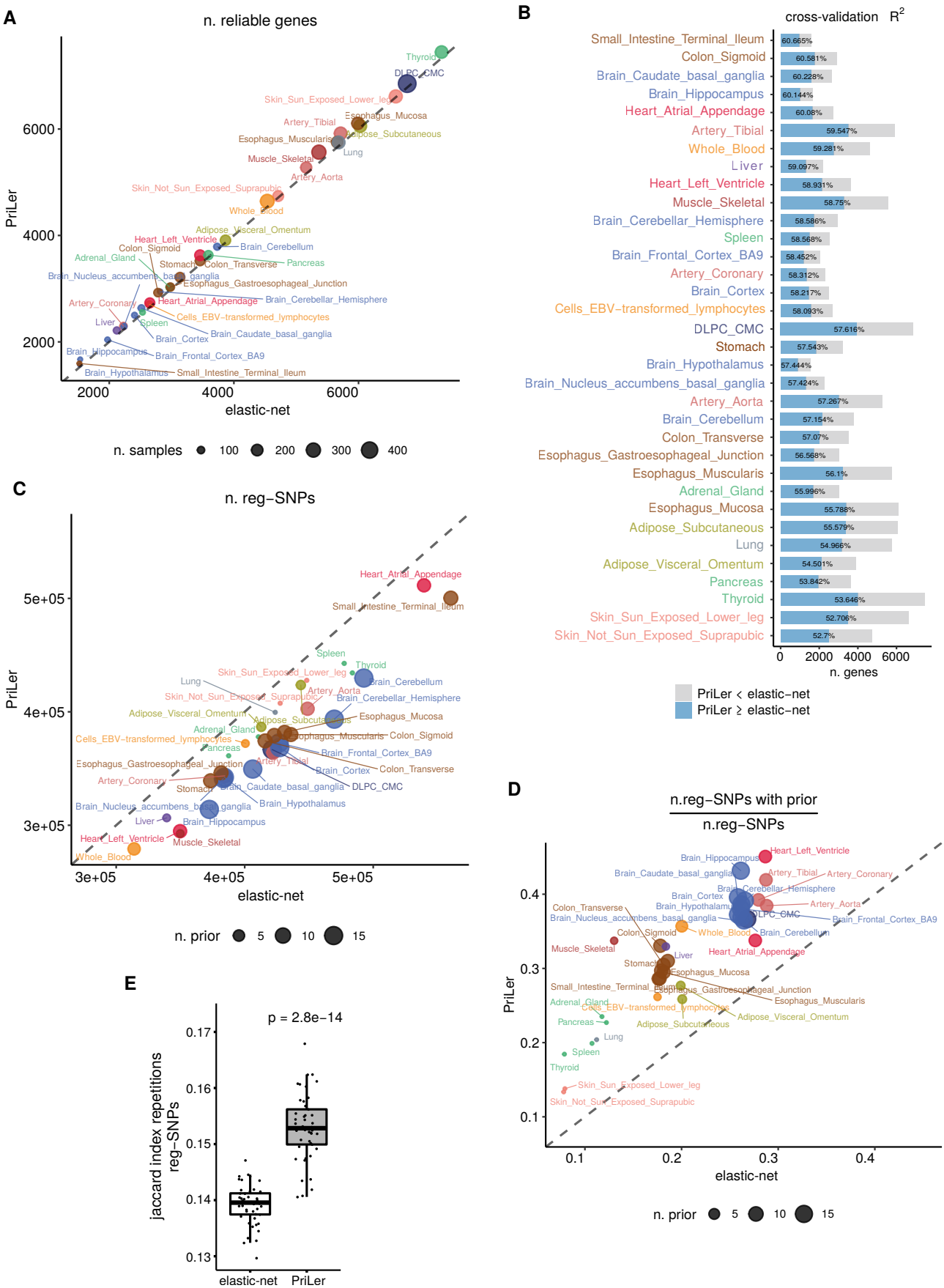
A
n. reliable genes
PriLer
elastic-net
n. samples
B
cross-validation R2
n. genes
PriLer < elastic−net
PriLer ≥ elastic−net
C
n. reg−SNPs
n. prior
D
n.reg−SNPs with prior
n.reg−SNPs
E
p = 2.8e−14
jaccard index repetitions reg−SNPs
elastic-net
PriLer

Fig. 4.10.: (From Trastulla et al., in prep.) Comparison of PriLer and elastic-net regression across 34 tissues in GTEx and CMC reference panels. **(A)** Number of reliable genes ($R^2_{test} > 0$ and $R^2 > 0.01$); **(B)** number of genes among the reliable ones for PriLer tissue-specific models with better prediction performance measured in CV test folders (R^2_{test}) in PriLer compared to elastic-net (blue, percentage shown) and vice-versa (grey) with tissues ordered by decreasing percentage of genes with better performances in Priler; **(C)** number of reg-SNPs (variant that regulates at least one gene in a certain tissue model); **(D)** fraction of reg-SNPs intersecting at least 1 prior feature for that tissue-specific model; **(E)** reg-SNPs robustness in whole blood tissue measured via Jaccard index of down-sampled models composed of 100 individuals 10 times (via bootstrapping), comparing each pair of repetition for PriLer and elastic- net models (p-value on top: Wilcoxon-Mann-Whitney test).

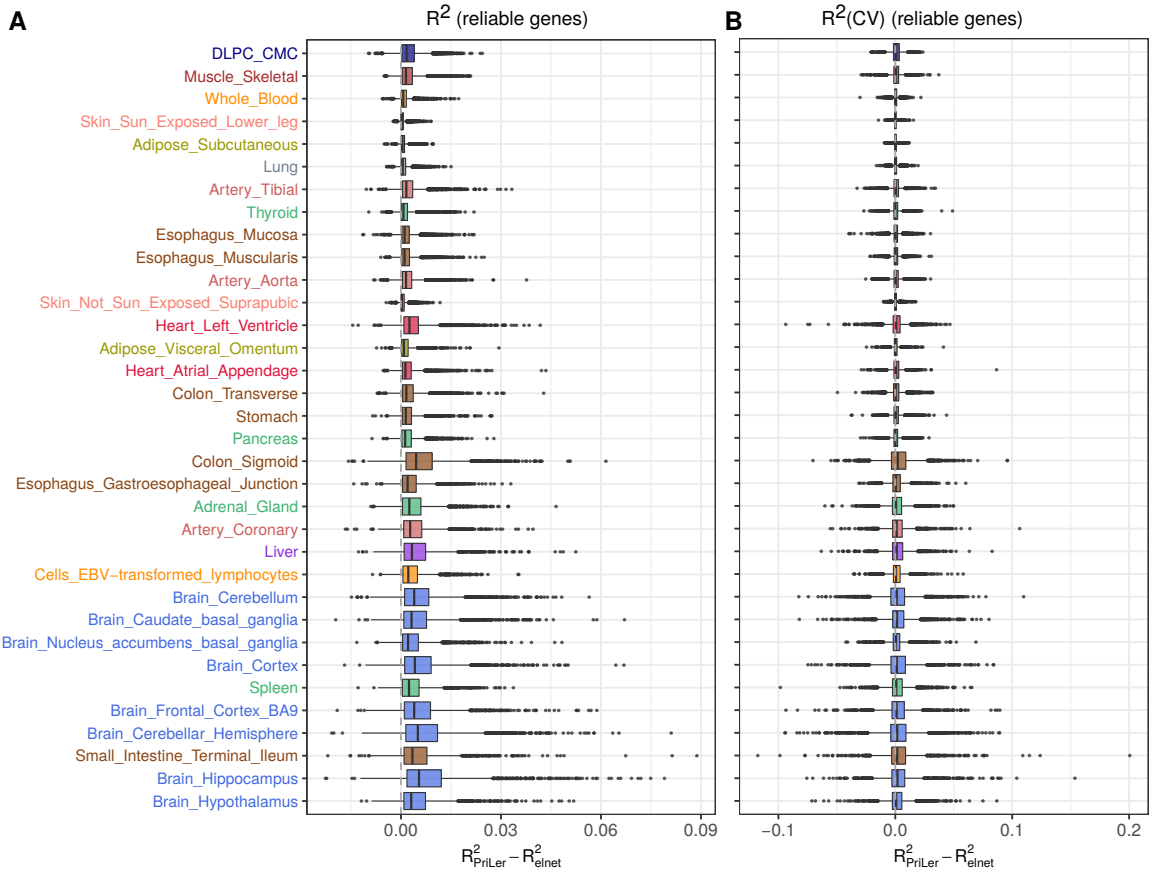

Fig. 4.11.: For genes reliable in PriLer model, Y-axis indicates the tissue-specific model considered in decreasing ordered with respect to their sample size; X-axis shows the difference between PriLer and elastic-net in term of **(A)** R^2 in final gene-expression models and **(B)** average R^2_{test} across test folders

4.2.4 Comparison of PriLer with Fusion and PrediXcan

Ultimately, we compared PriLer results in terms of model performances and regulatory variants selection with state-of-the-art methods developed to infer gene expression from cis-genetic features: **TWAS** [9] and **prediXcan** [10]. In this section we refer with **TWAS** method to what is now called Fusion. The authors changed subsequently the name of their method [9] to avoid ambiguity with the now called TWAS referring to genome-wide gene-trait testing.

We directly downloaded summary results of TWAS built on GTEx v6p and CMC public available at `https://data.broadinstitute.org/alkesgroup/FUSION/WGT/GTEx.ALL.tar` and `https://data.broadinstitute.org/alkesgroup/FUSION/WGT/CMC.BRAIN.RNASEQ.tar.bz2` respectively. For prediXcan, we downloaded results built on GTEx v6p from `https://s3.amazonaws.com/predictdb2/deprecated/download-by-tissue-HapMap/`[1] and on CMC from `https://github.com/laurahuckins/CMC_DLPFC_prediXcan/blob/master/DLPFC_oldMetax.db.tar.gz` which refers to a study of Huckins et al. [153] building gene expression models on DLPC in CMC via prediXcan.
PriLer, TWAS, and prediXcan methodology were built on the same tissue and databases versions (GTEx v6p and CMC release 1), nevertheless differences in pre-processing steps led to a different set of samples and variants, although overlapping. Having restricted our analysis to Caucasian individuals only, the number of training samples building PriLer models is generally lower (Fig. C.1, C.2) with an average decrease of 19 (SD$=18$) and 22 (SD$=17$) samples with respect to TWAS and prediXcan, thus reducing the overall power.
To compare gene expression model performances between PriLer and previously developed methods, we evaluated PriLer prediction results via squared Pearson correlation between actual gene expression adjusted for covariates and predicted gene expression from genotype information only (see def. (3.16)) across cross-validation test folders combined together ($corr^c_{test}$). We used squared $corr^c_{test}$ instead of previously reported R^2_{test} to apply the same strategy considered in TWAS and prediXcan to evaluate model performances.
For each tissue, we filtered for genes having at least two variants in the 200kb TSS gene cis-window that were also available in TWAS or prediXcan summary statistics. Comparing PriLer and TWAS (Fig. 4.12A), PriLer achieved better prediction performance in 76.6% of genes out of $68,891$ across all tissues, although 80.5% of gene expression models in PriLer were built on more reg-SNPs than those built on TWAS. The increase in both model prediction performances and n. reg-SNPs was consistent across all tissues (Fig. 4.13), with percentage of improved genes for squared $corr^c_{test}$ ranging from 60.5% of small intestine terminal ileum (73.3% genes using more reg-SNPs) to 91.8% of whole blood (83.6% genes using more reg-SNPs). Observing squared $corr^c_{test}$ tissue-specific distribution (Fig. C.3), PriLer model performances were always significantly higher than TWAS, with paired WMW p-value $\leq 10^{-5}$ across all tissues. As already pointed out, better prediction performances are also at the expense of a significantly higher number of reg-SNPs in PriLer (Fig. C.5, paired WMW p-value $\leq 10^{-5}$). Since TWAS results are obtained from the best model between 5 different ones among which "best eQTL", the decrease in reg-SNPs selection for TWAS can be related to a not null number of models defined via "best eQTL", predicting gene expression solely from one cis-variant with most significant association to gene expression variability.

[1] link originally used in 2019 now deprecated, corresponding results can be found at `https://zenodo.org/record/3572842/files/GTEx-V6p-HapMap-2016-09-08.tar.gz?download=1`

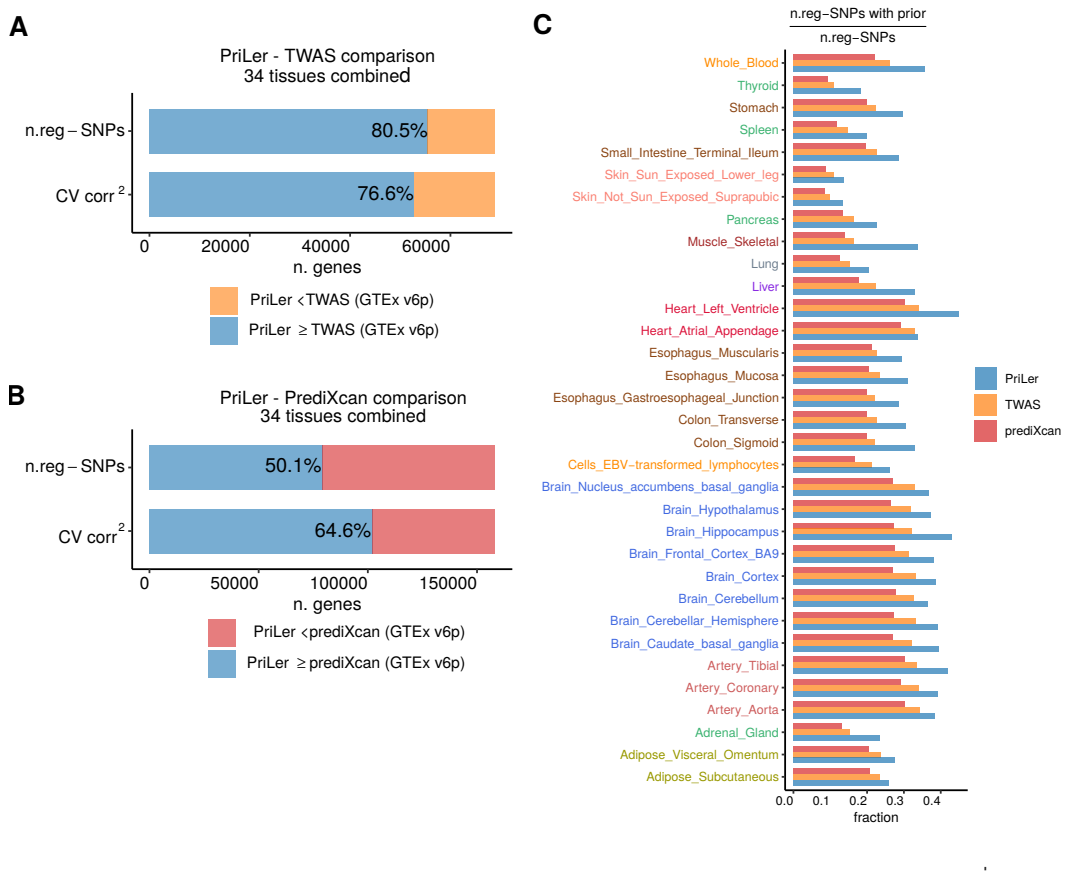

Fig. 4.12.: (Adapted from Trastulla et al., in prep.) Comparison of PriLer to TWAS and prediXcan methods built on 33 tissues from GTEx v6p and DLPC tissue from CMC. **(A-B)** Number and percentage of genes with better performances in PriLer in terms of CV squared correlation (computed combining test folders) and higher number of reg-SNPs in final gene expression models across all tissues, with respect to TWAS (A) and prediXcan (B). Genes included are those in common between the two methods considered and having at least two variants in 200kb TSS window. **(C)** Fraction of reg-SNPs having at least a prior information for the tissue-specific prior features selected for PriLer model, across PriLer, TWAS, and prediXcan.

We noticed a similar scenario when comparing PriLer to prediXcan: combining all tissues 64.6% of genes out of $158,249$ reached a higher $corr^c_{test}$ in PriLer yet, different from the comparison with TWAS, only 50.1% of those gene expression models were based on a higher number of reg-SNPs in PriLer (Fig. 4.12B). As shown in Fig. 4.14, in almost all tissues more than half of the genes achieved better performances in PriLer with the maximum reached by 71% in cells EBV transformed lymphocytes and the only exception being small intestine terminal ileum with 48% of genes being improved in PriLer. On the other hand, only 11 out 34 tissues had more than 50% of gene expression models built on a higher number of reg-SNPs in PriLer, varying from 65.7% of dorsolateral prefrontal cortex to 38% of hypothalamus. These results were congruent with the tissue-specific distribution of $corr^c_{test}$ and n. of reg-SNPs: PriLer showed significantly better model estimates with paired WMW p-value $\leq 10^{-5}$ for almost all tissues. The only exception was small intestine terminal ileum for which WMW $P = 2.2^{-5}$ and the median of correlation differences (PriLer minus prediXcan) was lower than zero (Fig. C.4), indicating an overall better

performance in prediXcan. In addition, the same 11 tissues having more than half of gene expression models based on a higher number of reg-SNPs, had also a reg-SNPs distribution not significantly lower in PriLer than prediXcan (Fig. C.6).

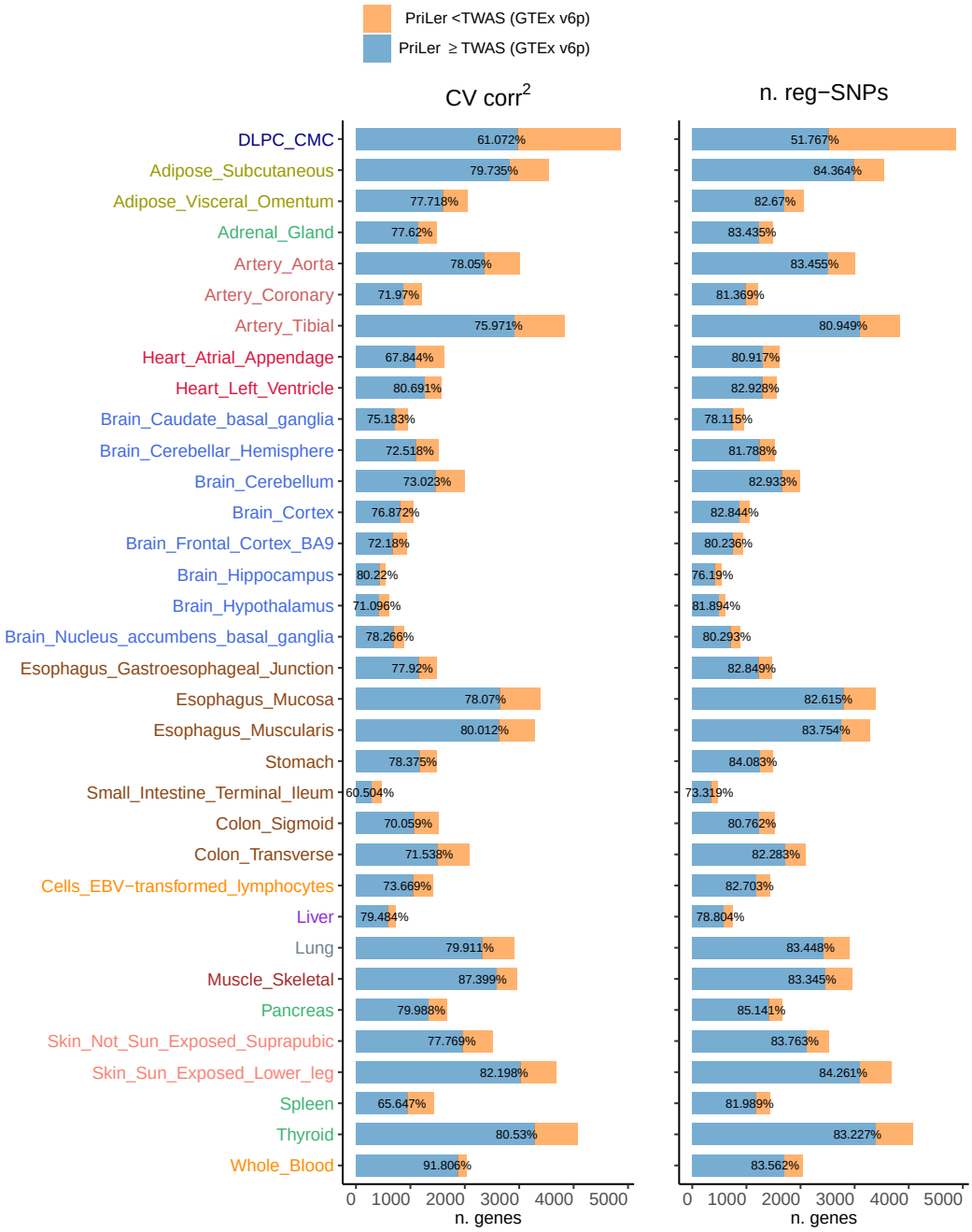

Fig. 4.13.: Comparison of PriLer with TWAS results built on GTEx v6p and CMC. Left barplot: across each tissue (y-axis), x-axis shows the number of genes with better performances in PriLer (blue) or TWAS (orange) in terms of CV squared correlation (computed combining test folders). Right barplot: x-axis shows the number of genes with higher number of reg-SNPs in final gene expression models in PriLer (blue) or TWAS (orange). Labelled percentage text refers to percentage of genes with better performances in PriLer. Genes included are those in common between PriLer and TWAS and having at least two variants in 200kb TSS window.

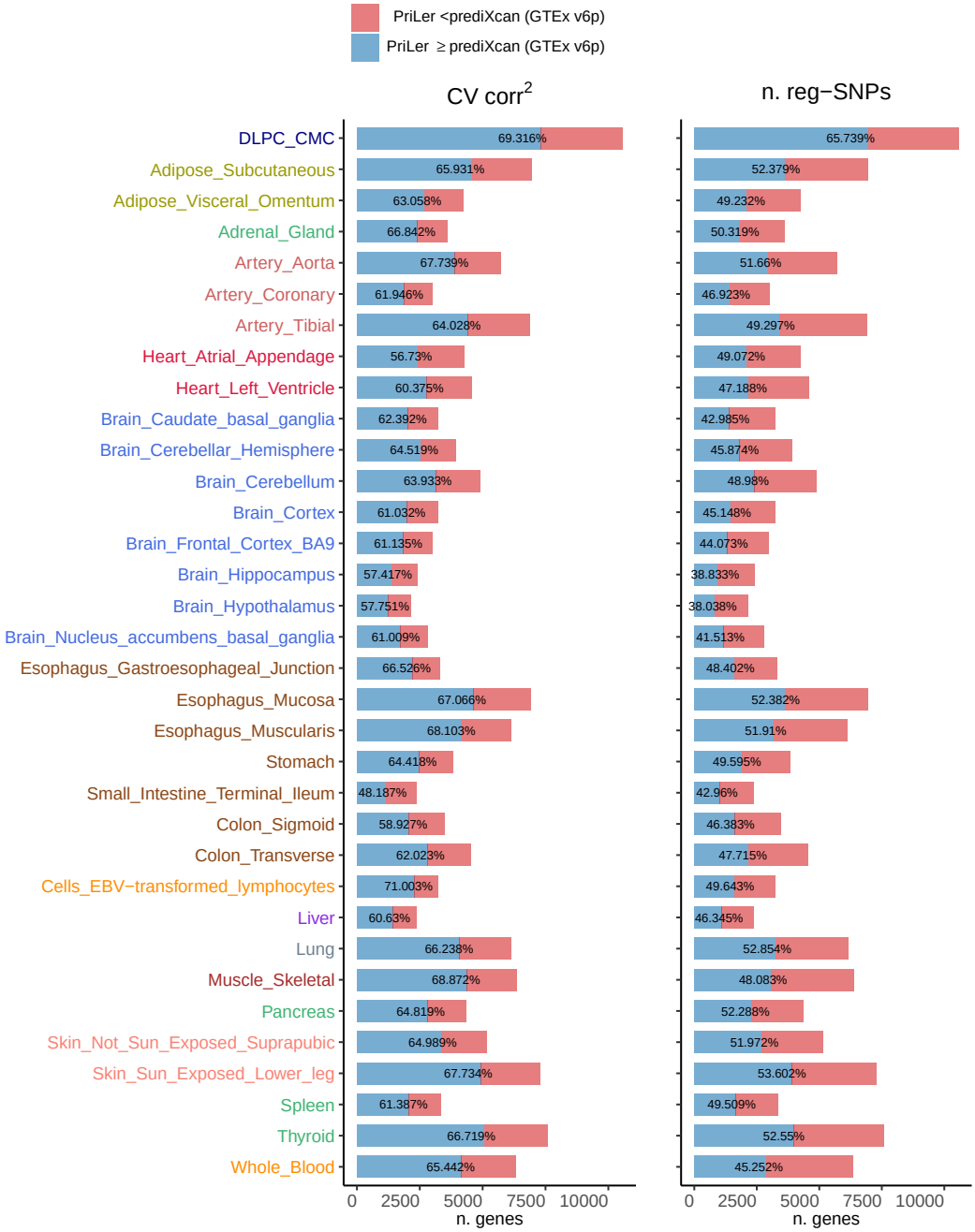

Fig. 4.14.: Comparison of PriLer with prediXcan results built on GTEx v6p and CMC. Left barplot: across each tissue (y-axis), x-axis shows the number of genes with better performances in PriLer (blue) or prediXcan (red) in terms of CV squared correlation (computed combining test folders). Right barplot: x-axis shows the number of genes with higher number of reg-SNPs in final gene expression models in PriLer (blue) or prediXcan (red). Labelled percentage text refers to percentage of genes with better performances in PriLer. Genes included are those in common between PriLer and prediXcan and having at least two variants in 200kb TSS window.

Furthermore, the fraction of reg-SNPs that intersected prior features used in the corresponding tissue-specific PriLer model was always higher in PriLer, followed by TWAS and then prediXcan (Fig. 4.12C). This improvement in biological relevance was similar to what we observed in the comparison with elastic-net (Fig. 4.10D).

Therefore, we showed that PriLer generally exceeds both TWAS and prediXcan in terms of

prediction performances. In addition, when PriLer used a number of reg-SNPs systemically higher such as in the comparison with TWAS, there was an even higher improvement in prediction performances. Finally, the selected reg-SNPs by PriLer were more relevant from a biological regulation perspective than those identified by TWAS and prediXcan, showing evidence of an higher intersection with gene regulatory elements (GREs).

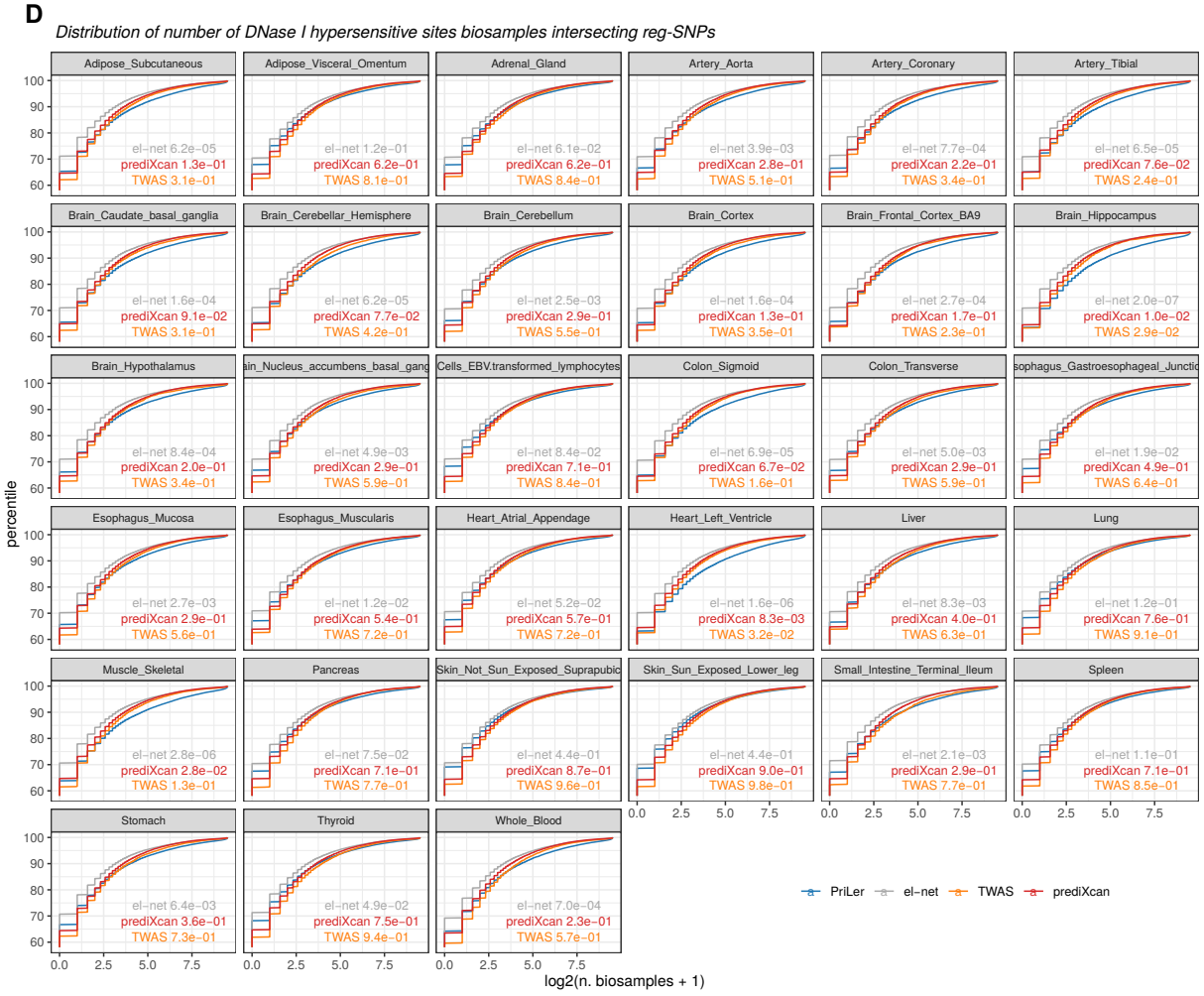

Fig. 4.15.: (Adapted from Trastulla et al., in prep.) **(D)** Enrichment of reg-SNPs in DNase I hypersensitive sites (DHSs) from external sources not used as prior features. The cumulative distribution (y-axis) of reg-SNPs intersecting with a certain number of DHS biosamples (x-axis) is shown for all the 4 methodologies: PriLer, elastic-net, TWAS and prediXcan. Distribution differences between PriLer and the other methods are tested with Kolmogorov-Smirnoff test (p-value shown).

The GREs considered to investigate biologically meaningfulness of reg-SNPs were the same used to train PriLer models. Thus, we also externally validated reg-SNPs selection to examine the location in biologically relevant portions of the genome using DNase I hypersensitive sites (DHSs). We leveraged the recently annotated map of DHSs across 733 human biosamples covering 438 tissue and cell types described in [204]. Comparing PriLer with elastic-net, TWAS, and prediXcan, we considered reg-SNPs of reliable genes for the first two methods and reg-SNPs from summary statistics of the

last two methods and we intersected their genomic location with annotated DHSs. For each variant regulating at least one gene, we considered the DHS that the reg-SNP overlapped with and counted the number of biosamples sharing that DHS. The percentages of reg-SNPs intersecting at least one biosamples were systematically higher in PriLer compared to elastic-net across all tissues but never compared to TWAS and only for 3 tissues compared to prediXcan (Tab. B.3). We then explored the actual number of biosamples rather than the intersection with at least one, and observed the cumulative distribution of reg-SNPs intersecting more than a certain number of biosamples with pairwise differences between PriLer and the other methodologies, tested via Kolmogorov-Smirnoff (`alternative="greater"`, nominal p-value reported in each tissue-specific box), as shown in Fig. 4.15D. PriLer reg-SNPs intersected an overall higher number of biosamples DHSs than elastic-net across 23 out 33 tissues (BH corrected p-value ≤ 0.05). Relaxing the corrected p-value threshold to 0.35, PriLer leaded to a significant improvement also when compared to prediXcan in hippocampus, left ventricle and muscle skeletal tissues (nominal p-value ≤ 0.03). These 3 tissues also reported the most significant differences in TWAS comparison, although not passing the after correction threshold (nominal p-value ≤ 0.13). Consistently with this result, we observed that hippocampus, left ventricle and muscle skeletal tissues are among the top 4 tissues with strongest improvement in the fraction of reg-SNPs with prior information in PriLer versus the other methodologies, specifically with elastic-net differences > 0.16 (Fig. 4.10D), TWAS and prediXcan differences > 0.11 and > 0.15 respectively (Fig. 4.12C). Thus, the superiority of PriLer in selection biologically meaningful variants is externally validated compared to elastic-net but only for a reduced number of tissues compared to the other two methods, possibly due to a particularly appropriate inclusion of prior features.

To sum up, our newly developed gene expression imputation method leads to an improvement in term of explained gene variance, in the majority of the cases reducing the number of reg-SNPs although more biologically relevant, is more robust in selecting regulatory variants and gives insight into gene regulation via prior features assigned weights.

4.3 Coronary Artery Disease

We first applied CASTom-iGEx to coronary artery disease. In particular, we built PriLer gene expression models on GTEx for 11 CAD related tissues after variant matching and harmonization with UKBB, composed of $19,024$ cases and $321,916$ non-affected individuals, and 9 CARDIoGRAM cohorts, composed of $13,279$ cases and $13,402$ non-affected individuals (Fig. 4.1, Tab. 4.2). The CAD related tissues were obtained from GTEx reference panel and included 2 adipose tissues (subcutaneous and visceral omentum), adrenal gland, 2 artery tissues (aorta and coronary), 2 colon tissues (sigmoid and transverse), 2 heart

tissues (atrial appendage and left ventricle), liver, and whole blood. Across this study we considered UKBB as the discovery set and CARDIoGRAM cohorts as the replication set.

4.3.1 Associated genes and pathways

After converting imputed gene expression to gene T-scores and pathway-scores, TWAS and PALAS analyses were performed as described in section 3.2.2. In particular, we included in the logistic regression sex and the first 10 PCs as covariates. The PCs were computed separately for UKBB and each CARDIoGRAM cohorts.
TWAS output testing genes association with CAD can be observed in Fig. 4.16A, explicated in the form of genome-wide Z-statistics (see def. (3.23)). We identified 383 significant genes across 11 tissues at the tissue-specific FDR threshold of 0.05, corresponding to 180 unique genes of which 163 outside the extended major histocompatibility complex (MHC) (chr6:26Mb-34Mb). The majority of the CAD associated genes were detected in only one tissue (Fig. 4.16B), with aorta and left ventricle showing the highest number of significant genes (Fig. 4.16C). This was related to the initial number of PriLer reliable genes tested, thus a consequence of the number of samples in GTEx reference panel, showing a Spearman correlation with the number CAD associated genes of 0.69. Nevertheless, when considering tissue-specific genes (i.e. CAD related genes only detected in a tissue), aorta and liver exhibited the highest percentage ($> 40\%$, Fig. 4.16D). In addition, we observed a high concordance in terms of Spearman correlation of gene Z-statistics for closely related tissues such as atrial appendage and left ventricle or aorta and artery coronary (cor> 0.85, Fig 4.16E red heatmap), although based on the limited shared subset of genes as measured via Jaccard index (Fig. 4.16E green heatmap). Importantly, the same genes across 11 tissues were also imputed on replication CARDIoGRAM cohorts based on the same harmonized SNP set and then tested for association combining the results via meta-analysis. To test the replicability of the findings built on UKBB, we evaluated the fraction of associated genes in UKBB having the same direction effect in CARDIoGRAM meta-analysis. With positive or negative direction effect for a gene we refer to an increase or decrease (respectively) of the genetically predicted gene expression being associated with higher susceptibility to CAD. The significance of this fraction is assessed via one-sided sign test under the null hypothesis of replication fraction equal to 0.5. All tissues were significant at least at the nominal level of 0.05 (Fig. 4.16F), with replicability of 82% ($P = 4.35^{-38}$) when combining all tissues together. In addition, 50% of the genes detected from UKBB were at least significant at the nominal level in CARDIoGRAM, with tissue-specific percentages varying from 34% of atrial appendage to 57% of artery aorta.

Taking a closer look at CAD associated genes, many of the detected association were in loci already identified from GWASs.

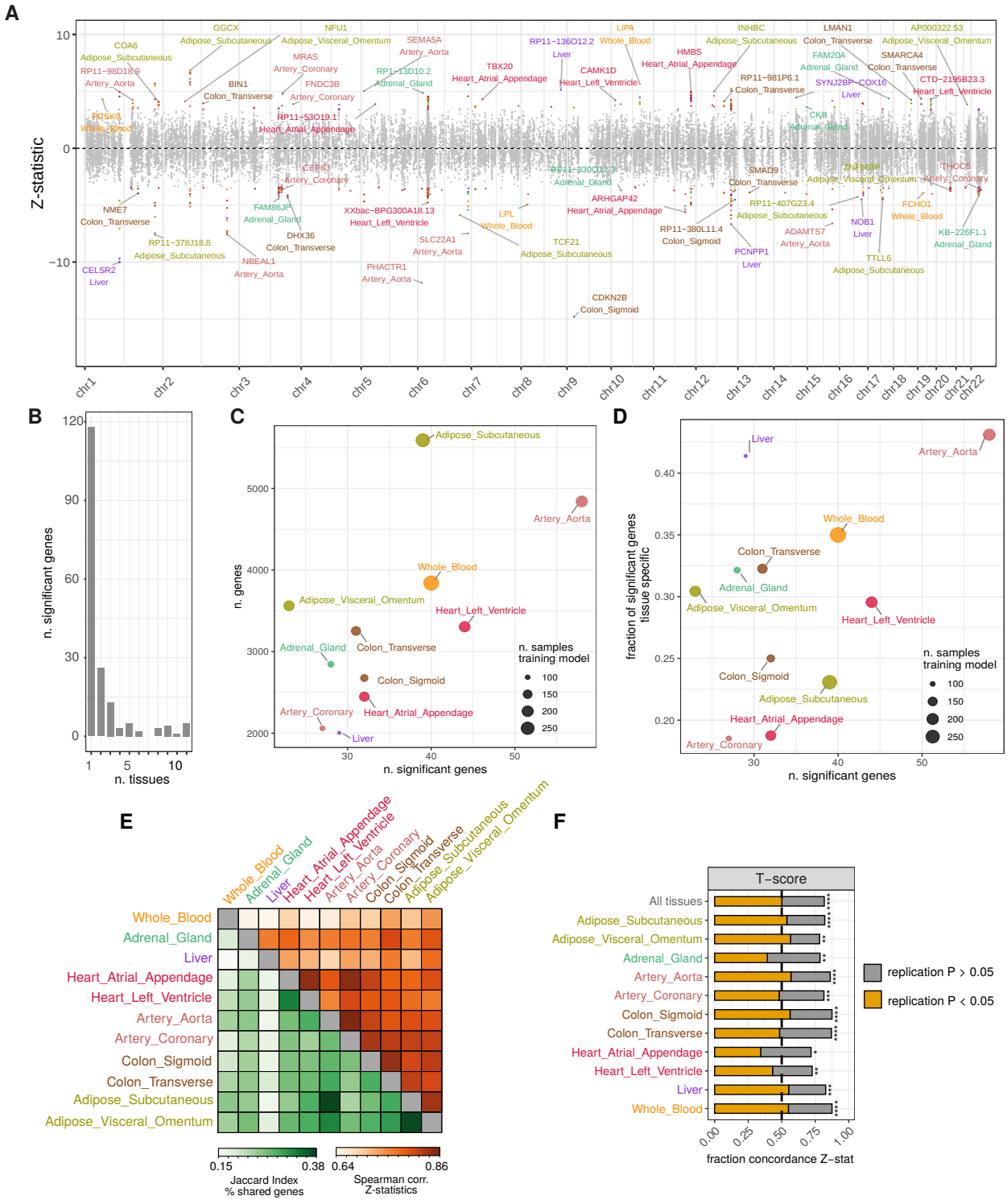

Fig. 4.16.: (Adapted from Trastulla et al., in prep.) **(A)** Manhattan plot showing Z-statistic across 11 tissues, colored dots indicate genes significant at tissue-specific FDR level of 0.05. **(B)** Number of significant genes detected in one or more tissues. **(C,D)** Number of CAD significant genes versus **(C)** number of tested reliable genes (predicted in PriLer) or **(D)** fraction of significant genes uniquely detected in a tissue, dot size refers to the number of PriLer training sample in the GTEx reference panel. **(E)** Lower-triangular (green): percentage of imputed genes that are in common between 2 tissues (Jaccard index), upper-triangular (orange): Spearman correlation of CAD Z-statistics among shared genes. **(F)** Reproducibility of gene levels T-scores with discovery UKBB samples and replication from the meta-analysis of CARDIoGRAM cohorts. X-axis shows the fraction of significant genes in UKBB that have the same effect sign (Z-stat) in CARDIoGRAM meta-analysis, p-values are computed from one-sided sign test ($* = P \leq 0.05$, $** = P \leq 0.01$, $*** = P \leq 0.001$, $**** = P \leq 0.0001$). The bar in yellow represents the fraction of genes concordant that are also significant at the nominal p-value threshold of 0.05.

In particular, we combined TWAS summary statistics across all tissues in loci, recursively merging significant genes with $[\text{TSS} - 200\text{kb}, \text{TSS} + 200\text{kb}]$ window distant less than 1Mb. The 200kb threshold is related to the cis-window applied for building gene expression model in PriLer, whereas 1Mb distance is chosen to combine significant genes possibly arising from LD structure. In this way, the 383 significant genes merged into 83 loci. In addition, we compared these significant loci with two GWAS summary statistics:

case (1) a meta-analysis of UK Biobank SOFT CAD GWAS (includes individuals with fatal or nonfatal myocardial infarction, percutaneous transluminal coronary angioplasty or coronary artery bypass grafting, chronic ischemic heart disease and angina) with CARDIoGRAMplusC4D 1000 Genomes-based GWAS and the Myocardial Infarction Genetics and CARDIoGRAM Exome [120] downloaded from www.CARDIOGRAMPLUSC4D.ORG;

case (2) custom GWAS performed via logistic regression implemented in PLINK2 software [205] (`plink2 -glm` option) using the same case/control individuals, covariates, and variants as considered is our TWAS and PALAS analysis, from now own defined as "matched GWAS".

While case (1) allowed to compare our results with the (at the time) state-of-the-art CAD genetic associations, the case (2) GWAS analysis was built on the same individuals and set of variants in our analysis. This enabled the comparison in prediction performances among GWAS, TWAS and PALAS and the investigation of aggregation effects of variants into meaningful biological entities such as genes and pathways. Since we used an FDR correction strategy in our TWAS and PALAS analysis, GWAS p-values were adjusted with the same Benjamini-Hochberg (BH) procedure [173], both in case (1) and (2). Among the 83 loci we identified in our TWAS analysis, 33 loci did not intersect any FDR significant SNP of case (1) amounting to 92 genes. Of these genes, 59 were also replicated in the CARDIoGRAM cohorts in term of effect size sign concordance and 23 additionally at the nominal 0.05 significance, merging into 11 loci (details in Tab. 4.4).
Conversely, the remaining 50 loci out of 83 that intersected a significant GWAS result, included many well-known CAD risk genes (Fig. 4.16A), among which sortilin 1 (SORT1, same locus as CELSR2 shown in figure), Cyclin Dependent Kinase Inhibitor 2B (CDKN2B) and Phosphatase And Actin Regulator 1 (PHACTR1), that corresponded to the top 3 significant genes in our analysis. These associated genes were highly tissue specific as SORT1 and PHACTR1 are only imputed in liver and aorta tissues respectively while CDKN2B is predicted in colon sigmoid and whole blood but only significant in the former (Z-stat$= -14.81$, P$=1.26^{-49}$ in colon sigmoid, Z-stat$= 2.47$, P$= 0.01$ in blood). SORT1 implication for CAD has been functionally validated via in-vivo experiments and connected to LDL clearance [206]. In addition, the region 9p21 of CDKN2B was long identified as a GWAS hit and a decrease in CDKN2B has been connected to proliferation of vascular cells and increased mortality in mouse models [207], however the exact mechanisms remain unclear. From our analysis, CDK2NB PriLer gene expression model in colon sigmoid

assigned a non zero coefficient to 51 variants, with the strongest influence from rs597816 variant ($\beta = -0.076$, Fig. 4.17A) that also showed a GWAS p-value (from the case (2) matched setting) $< 10^{-20}$ and partially explained the high level of significance reached in TWAS. Nevertheless, the same was not observed in whole blood (Fig. 4.17B), in which the variants with top PriLer coefficients were not GWAS significant and the aforementioned rs597816 was not included in the model ($\beta = 0$), thus indicating that the tissue-specific regulation influence the phenotypic association. Interestingly, in both tissues the top relevant variants in PriLer model did not intersect any gene regulatory region (GRE) used as prior, indicating that the prior information did not force the intersecting variants to have strongest impact on gene expression.

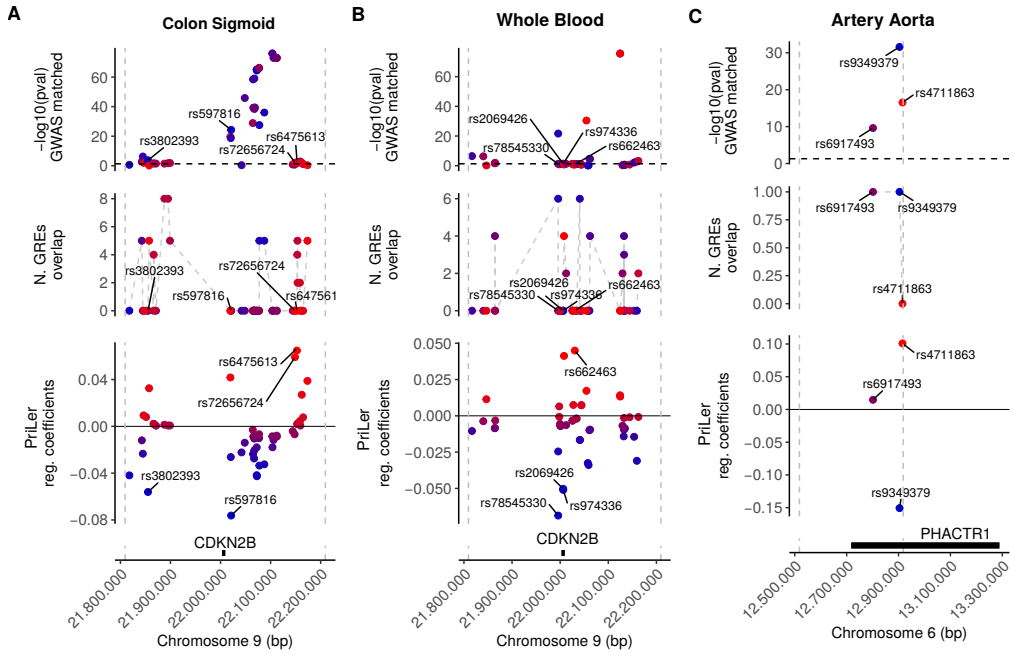

Fig. 4.17.: PriLer models for CDKN2B in colon sigmoid **(A)** and whole blood **(B)** and PHACTR1 in artery aorta **(C)**, with each dot representing a variant with PriLer regression coefficient different from zero and the corresponding genomic position shown in the x-axis. Panel from the bottom to the top: 1) genomic position of considered gene with dashed lines representing TSS $\pm 200kb$ window, 2) regression coefficient from PriLer gene expression model, 3) number of GREs in the PriLer model that a variant intersects (tissue-specific selection in Tab. B.2), 4) -log10 p-value from our GWAS summary statistics with matched individuals and variants. The color code of each dot reflects PriLer coefficient values and the labelled SNPs correspond to the top 4 in PriLer coefficient absolute values.

In addition, PHACTR1 was identified from previous GWAS as overlapping gene with rs9349379 associated variant [120], hypothesizing that PHACTR1 changes might lead to impairments in vascular pathobiology. In our TWAS analysis, the same variant rs9349379 showed the strongest impact on PHACTR1 regulation, that also overlapped with CAD_H3K27ac_ATAC epigenetic prior (Fig. 4.17C) and induced a decrease in imputed gene expression for CAD patients.

Regarding the newly identified genes (Tab. 4.4), among those replicated at the nominal

level in the CARDIoGRAM meta-analysis we identified Nucleoside diphosphate kinase 7 (NME7) and Iron-Sulfur Cluster Scaffold (NFU1). Despite being a reliable gene in 8 tissues (excluded whole blood and atrial appendage), NME7 reached significance only for aorta and the two colon tissues, with a decrease of imputed gene expression in CAD patients. Variants in NME7 (and ATP1B1) locus where recently identified in a trans-ancestry GWAS [208] and NME7 was further prioritize in the largest CAD GWAS to date that included more than a million participants [74]. This gene is a γ-tubulin ring complex component that facilitate microtubule nucleation of this complex and is involved in ciliary signaling and protein trafficking to primary cilia [209]. Instead, NFU1 is only reliably imputed in adipose visceral in which the increase was significant in CAD patients. NFU1 is a mitochondrial iron-sulfur scaffold protein implicated in iron-sulfur assembly and transfer to lipoic acid synthase. Mutation in NFU1 were linked to mitochondrial dysfunction and pulmonary hypertension in patients as well as CRISPR-Cas9 rat models [210]. For both showcase genes, their role in the context of cardiovascular disease needs further investigation and experimental validation.

Chrom	Gene	Loci	Tissue	UKBiobank (Discovery)		CARDIoGRAM (Replication)	
				P-value	Z	P-value	Z
chr1	ATP1B1	chr1:168.9-169.5Mb	Adrenal_Gland	3.134178e-04	3.603950	6.782846e-03	2.707322
chr1	NME7	chr1:168.9-169.5Mb	Artery_Aorta	8.315375e-05	-3.935124	3.383281e-03	-2.930582
chr1	NME7	chr1:168.9-169.5Mb	Colon_Sigmoid	4.248872e-04	-3.524123	3.826846e-03	-2.892093
chr1	NME7	chr1:168.9-169.5Mb	Colon_Transverse	8.488568e-05	-3.930170	4.136809e-03	-2.867537
chr2	MAP3K2	chr2:127.7-128.3Mb	Artery_Aorta	4.979708e-04	3.481846	8.391695e-06	4.454939
chr2	BIN1	chr2:127.7-128.3Mb	Colon_Transverse	6.926833e-05	3.978785	1.034511e-02	2.564075
chr2	ERCC3	chr2:127.7-128.3Mb	Heart_Atrial_Appendage	5.738399e-04	-3.443687	3.874994e-05	-4.114809
chr2	NFU1	chr2:69.5-69.9Mb	Adipose_Visceral_Omentum	3.145251e-05	4.162687	2.532944e-02	2.236341
chr3	ALG1L	chr3:125.4-126Mb	Adrenal_Gland	1.435744e-04	-3.801928	1.423083e-02	-2.451383
chr3	ALG1L	chr3:125.4-126Mb	Artery_Coronary	6.725689e-05	-3.985787	1.205337e-02	-2.510578
chr3	ALG1L	chr3:125.4-126Mb	Colon_Sigmoid	3.044326e-05	-4.170125	1.855013e-02	-2.354450
chr3	RP11-124N2.1	chr3:125.4-126Mb	Colon_Transverse	3.808274e-04	3.553026	2.454542e-02	2.248483
chr3	ALG1L	chr3:125.4-126Mb	Heart_Atrial_Appendage	1.201518e-04	-3.845816	7.904283e-03	-2.656131
chr3	ALG1L	chr3:125.4-126Mb	Heart_Left_Ventricle	1.097006e-04	-3.868069	1.130471e-02	-2.533136
chr3	RP11-666A20.3	chr3:125.4-126Mb	Heart_Left_Ventricle	4.708662e-04	-3.496805	3.572013e-02	-2.100099
chr3	ALG1L	chr3:125.4-126Mb	Whole_Blood	2.867626e-04	-3.626974	2.947294e-02	-2.177102
chr3	FNDC3B	chr3:171.6-172Mb	Artery_Coronary	9.041403e-05	3.914974	3.722798e-02	2.083255
chr7	IGFBP3	chr7:45.8-46.2Mb	Artery_Aorta	4.427863e-04	-3.513175	4.022620e-02	-2.051419
chr12	OR7E47P	chr12:52.3-52.7Mb	Heart_Left_Ventricle	1.380409e-04	3.811652	3.040734e-02	2.164743
chr15	CTD-2262B20.1	chr15:85.8-86.2Mb	Adipose_Subcutaneous	2.504570e-04	3.661792	7.969579e-03	2.653356
chr16	RP11-407G23.4	chr16:57.1-57.5Mb	Adipose_Subcutaneous	2.064329e-05	-4.257817	9.554376e-03	-2.591553
chr17	FAM20A	chr17:66.3-66.8Mb	Adrenal_Gland	1.070258e-05	4.402469	2.738668e-02	2.205962
chr20	OSER1	chr20:42.6-43Mb	Whole_Blood	2.784586e-04	-3.634557	4.683332e-02	-1.987804

Tab. 4.4.: Tissue specific significant genes for CAD in discovery cohort UKBB grouped into loci with overlapping variants from GWAS [120] not significant at FDR 0.05 threshold and replicated in the external cohort (CARDIoGRAM) in terms of concordance of sign and nominal p-value ≤ 0.05

We then aggregated the tissue-specific gene T-scores into pathway-scores for each individual and performed PALAS to detect the pathways associated with CAD phenotype (see section 3.2). In particular, we considered 3 biological pathway databases: Reactome [75], Gene Ontology [76] and WikiPathways [77] and corrected p-values separately for each tissue and pathway database, considering as significant those with FDR ≤ 0.05.

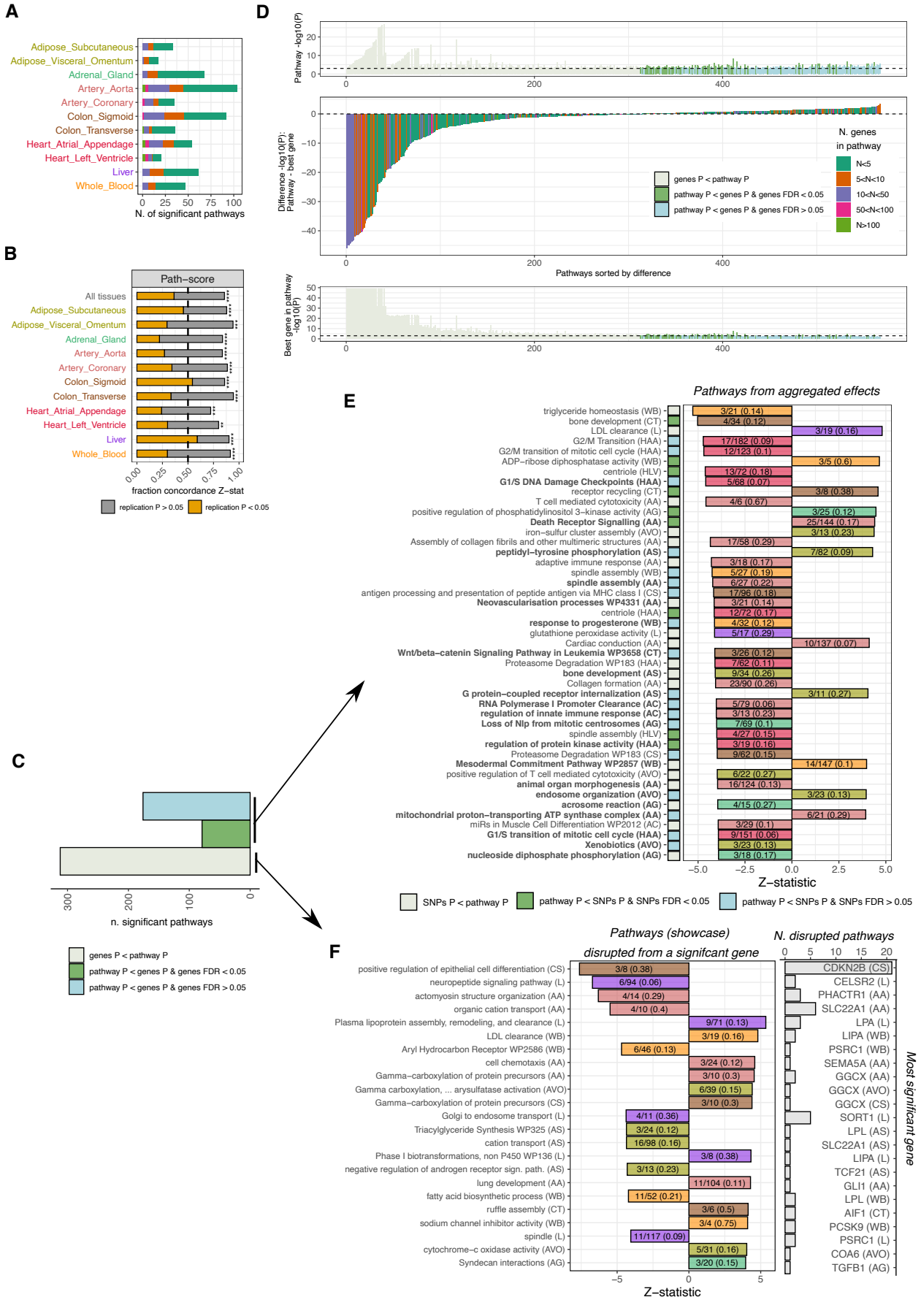
A
Adipose_Subcutaneous
Adipose_Visceral_Omentum
Adrenal_Gland
Artery_Aorta
Artery_Coronary
Colon_Sigmoid
Colon_Transverse
Heart_Atrial_Appendage
Heart_Left_Ventricle
Liver
Whole_Blood
0 25 50 75 100
N. of significant pathways
B
Path−score
All tissues
fraction concordance Z−stat
replication P > 0.05
replication P < 0.05
C
n. significant pathways
genes P < pathway P
pathway P < genes P & genes FDR < 0.05
pathway P < genes P & genes FDR > 0.05
D
Pathway -log10(P)
Difference -log10(P): Pathway − best gene
N. genes in pathway
N<5
5<N<10
10<N<50
50<N<100
N>100
Pathways sorted by difference
Best gene in pathway -log10(P)
E
Pathways from aggregated effects
triglyceride homeostasis (WB)
bone development (CT)
LDL clearance (L)
G2/M Transition (HAA)
G2/M transition of mitotic cell cycle (HAA)
ADP−ribose diphosphatase activity (WB)
centriole (HLV)
G1/S DNA Damage Checkpoints (HAA)
receptor recycling (CT)
T cell mediated cytotoxicity (AA)
positive regulation of phosphatidylinositol 3−kinase activity (AG)
Death Receptor Signalling (AA)
iron−sulfur cluster assembly (AVO)
Assembly of collagen fibrils and other multimeric structures (AA)
peptidyl−tyrosine phosphorylation (AS)
adaptive immune response (AA)
spindle assembly (WB)
spindle assembly (AA)
antigen processing and presentation of peptide antigen via MHC class I (CS)
Neovascularisation processes WP4331 (AA)
centriole (HAA)
response to progesterone (WB)
glutathione peroxidase activity (L)
Cardiac conduction (AA)
Wnt/beta−catenin Signaling Pathway in Leukemia WP3658 (CT)
Proteasome Degradation WP183 (HAA)
bone development (AS)
Collagen formation (AA)
G protein−coupled receptor internalization (AS)
RNA Polymerase I Promoter Clearance (AC)
regulation of innate immune response (AC)
Loss of Nlp from mitotic centrosomes (AG)
spindle assembly (HLV)
regulation of protein kinase activity (HAA)
Proteasome Degradation WP183 (CS)
Mesodermal Commitment Pathway WP2857 (WB)
positive regulation of T cell mediated cytotoxicity (AVO)
animal organ morphogenesis (AA)
endosome organization (AVO)
acrosome reaction (AG)
mitochondrial proton−transporting ATP synthase complex (AA)
miRs in Muscle Cell Differentiation WP2012 (AC)
G1/S transition of mitotic cell cycle (HAA)
Xenobiotics (AVO)
nucleoside diphosphate phosphorylation (AG)
Z−statistic
SNPs P < pathway P
pathway P < SNPs P & SNPs FDR < 0.05
pathway P < SNPs P & SNPs FDR > 0.05
F
Pathways (showcase)
disrupted from a signifcant gene
N. disrupted pathways
positive regulation of epithelial cell differentiation (CS)
neuropeptide signaling pathway (L)
actomyosin structure organization (AA)
organic cation transport (AA)
Plasma lipoprotein assembly, remodeling, and clearance (L)
LDL clearance (WB)
Aryl Hydrocarbon Receptor WP2586 (WB)
cell chemotaxis (AA)
Gamma−carboxylation of protein precursors (AA)
Gamma carboxylation, ... arysulfatase activation (AVO)
Gamma−carboxylation of protein precursors (CS)
Golgi to endosome transport (L)
Triacylglyceride Synthesis WP325 (AS)
cation transport (AS)
Phase I biotransformations, non P450 WP136 (L)
negative regulation of androgen receptor sign. path. (AS)
lung development (AA)
fatty acid biosynthetic process (WB)
ruffle assembly (CT)
sodium channel inhibitor activity (WB)
spindle (L)
cytochrome−c oxidase activity (AVO)
Syndecan interactions (AG)
CDKN2B (CS)
CELSR2 (L)
PHACTR1 (AA)
SLC22A1 (AA)
LPA (L)
LIPA (WB)
PSRC1 (WB)
SEMA5A (AA)
GGCX (AA)
GGCX (AVO)
GGCX (CS)
SORT1 (L)
LPL (AS)
SLC22A1 (AS)
LIPA (L)
TCF21 (AS)
GLI1 (AA)
LPL (WB)
AIF1 (CT)
PCSK9 (WB)
PSRC1 (L)
COA6 (AVO)
TGFB1 (AG)
Most significant gene
Z−statistic

Fig. 4.18.: (Adapted from Trastulla et al., in prep.) **(A)** Number of significant pathways associated with CAD (tissue-specific FDR≤ 0.05) for each tissue, with the bar color coded according to the number of genes in the pathway also reliably predicted in that tissue (T-score genes). **(B)** Reproducibility of pathway scores associations with discovery UKBB samples and replication from the meta-analysis of CARDIoGRAM cohorts. X-axis shows the fraction of significant pathways in UKBB that have the same effect sign (Z-stat) in CARDIoGRAM meta-analysis, p-values are computed from one-sided sign test ($* = P \leq 0.05$, $** = P \leq 0.01$, $*** = P \leq 0.001$, $**** = P \leq 0.0001$). The bar in yellow represents the fraction of pathways concordant in that are also significant at the nominal p-value threshold of 0.05. **(C)** Number of significant pathways (FDR≤ 0.05) that include at least one gene more significant than the pathway (ivory), all genes in the pathway less significant but with at least one gene having FDR ≤ 0.05 (green), and all genes in the pathway less significant and not passing FDR 0.05 threshold (light blue). **(D)** For each significant pathway, the central panel shows the difference of $-\log_{10}(P)$ between the pathway and the most significant gene included, sorted from the smallest to the highest on the x-axis and color coded according the number of T-score genes for that pathway. Top and bottom panels refer to pathway and most significant gene $-\log_{10}(P)$ respectively, with the color referring to pathway classification as in (C). Dashed horizontal line on top and bottom panels correspond to $P = 0.001$. **(E)** Among pathways more significant than any included gene (green and light blue from (C)), prioritization of more reliable ones that include more than 5 genes or more than 2 when the coverage of the total pathway genes is $\geq 10\%$, less than 200 genes in both original genes and T-score genes pathway and p-value ≤ 0.0001. X-axis shows the PALAS Z-statistic color coded by tissue origin, with each pathway barplot including the gene pathway coverage. The pathway name in bold indicate those without any significant gene at the FDR 0.05 level and the acronym in brackets refers to the initials of the tissue considered. The square next to each pathway name is color coded according the comparison of pathway significance with matched GWAS result, similarly to (C), where a SNP in a pathway means that the SNP was included from PriLer model on a gene in that pathway. **(F)** Similar to (E), pathways with at least one gene more significant (ivory in (C)) filtered using the same prioritization criteria as in (E) and showing only one exemplar pathway per most significant gene (number of pathways including per gene on the right), with the showcase selected as having the highest coverage.

This procedure identified a total of 567 significant pathways across all tissues, 351 from GO, 143 from Reactome, and 73 from WikiPathways, with artery aorta reporting the highest number of associations followed by colon sigmoid (Fig. 4.18A). The majority of significant pathways were built upon less than 5 genes. Pathways are defined as a collection of genes regardless of their tissue of origin and we refer to these genes as "original genes" (indicated with $\mathcal{P}$ in section 3.2.1). Nevertheless, the individual scores are tissue-specific, depending on which genes were reliably imputed in that tissue (referred to as "T-score genes", $\mathcal{P} \bigcap \mathcal{G}^t_{rel}$) as well as the gene tissue-specific prediction. Of note, the high number of significant pathways in artery aorta is connected to a large number of significant genes in that tissue (Fig. 4.16C), and not due to the single strongest hit (PHACTR1) being included in 4 only pathways. On the other hand, out of 92 significant pathways in colon sigmoid 38 included CDKN2B, the most significant gene from TWAS. Similarly to genes, pathways detected in UKBB discovery cohort were replicated on CARDIoGRAM, with 86% of the results reproduced in terms of the direction of effect size (one-side sign test $P = 10^{-63}$) and 36% significant at the nominal level in the replication cohorts, with tissue-specific percentages varying from 54% of colon sigmoid to 22% of adrenal gland (Fig. 4.18B) and generally lower than gene T-scores percentages of replicability at the nominal level.

Because CASTom-iGEx computes sample-level scores for pathways and tests for association with the same procedure as TWAS, it is now possible to compare p-values between a pathway and the genes included (T-score genes). In this way, we can now discern whether

association signals from many genes (even weak ones) cooperate together and boost the effect size observed in the aggregated pathway (Fig. 4.18C,E), or contrarily the aggregation into pathway includes noisy and/or discordant information, and a single gene (or few) in that pathway reaches a better significant driving the disruption (Fig. 4.18C,F). We found that 312 (55%) of significant pathways included at least one gene more significant than the level reached by the pathway itself (class I pathways), with the remaining 255 (45%) showing an aggregation effect (class II pathways). Class II can be further split into pathways containing at least one gene FDR significant from TWAS (79) and those that are entirely composed of small effect non significant genes (176) that would be missed in a significant cut-off strategy. Comparing the actual level of significance between pathways and the corresponding best (most significant) T-score gene (Fig. 4.18D), we observed that the differences in significance was wider for pathways in class I, because of the disruption induced by genes in CDKN2B or SORT1 loci. This difference was more pronounced in pathways consisting of an higher number of T-score genes, possibly due to a discordant effect between the "leading" gene and the other genes in the pathway, nevertheless reaching significance. On the other hand, small effect genes aggregated for class II pathways led to an increase in pathway significance generally higher when composed of genes not passing FDR 0.05 threshold (Fig. 4.18D).

In order to inspect the nature of the detected CAD-associated pathways, we further prioritized our results selecting only those pathways including > 5 T-score genes or ≥ 3 in case of a ≥ 0.1 coverage, with the coverage for a pathway referring to the fraction of original genes in that pathway being also reliably imputed in the tissue considered and referred to as T-score genes ($\frac{n_{\mathcal{P},\mathcal{G}}}{n_{\mathcal{P}}}$, see section 3.2.1). In addition, we considered only pathways with less than 200 genes in both original and T-score genes and having PALAS $P \leq 10^{-4}$, thus focusing on more reliable associations. Applying this filtering strategy to class II significant pathways, we prioritized 45 pathways (Fig. 4.18E), of which 21 (with text in bold) can be considered as "novel" as they did not include any FDR significant gene. Thus, we can assume that the observed effect on CAD phenotype arise from a mechanism of aggregation that would missed via a p-value gene filtering approach. Some of these pathways are related to well-known CAD pathomechanisms such as *LDL-clearance* in liver [211], *collagen formation* [212] and *neovascularisation processes* in artery aorta. Of note, the *neovascularisation process* from WikiPathways was also externally replicated at the nominal level (Z-stat $= -2.878$, P$= 0.004$). In addition, we applied the same filtering prioritization strategy to class I pathways. Here the underlying hypothesis is that the pathway disruption is operated by a single highly significant gene, not necessarily cooperating with other genes in altering the pathway activity. This strategy led to 63 reliable associations due to the dysregulation of 23 highly significant genes for which a pathway exemplar is shown in Fig. 4.18F, selected as the one with highest coverage. As already noticed, CDKN2B highly significant gene accounted for the majority of dysregulated pathways (> 20), implicated in mechanisms such as the *epithelial cell differentiation* (Fig. 4.18F) and *oxidative stress induced senescence*. Similarly, we observed a significance in *Golgi to endosome transport* pathway related to SORT1 decreased expression in liver and lipid

metabolism pathways due to well-known LPA, LIPA, LPL CAD-associated genes both in liver and whole blood. Additionally, *gamma-carboxylation* pathways in multiple tissues were associated with CAD due to GGCX gene, a gene located in a already identified CAD GWAS locus (2p11.2), related to vitamin K-dependent coagulation [213]. Moreover, *actomyosin structure organization* was altered by PHACTR1 decrease and *cytochrome-c oxidase activity* in adipose visceral omentum was related to COA6 increase. Interestingly, COA6 gene was located in a novel loci not detected by a recent GWAS [120] nor the matched GWAS analysis we performed, but it was not replicated at the nominal level in CARDIoGRAM cohorts (Z-stat$= 0.52$, P$= 0.6$). Nevertheless, impaired activity of *cytochrome-c oxidase* was connected to myocardial insufficiency and dilated cardiomyopathy via a reduction in mitochondrial respiration [214], hence showing possible connections to CAD mechanisms.

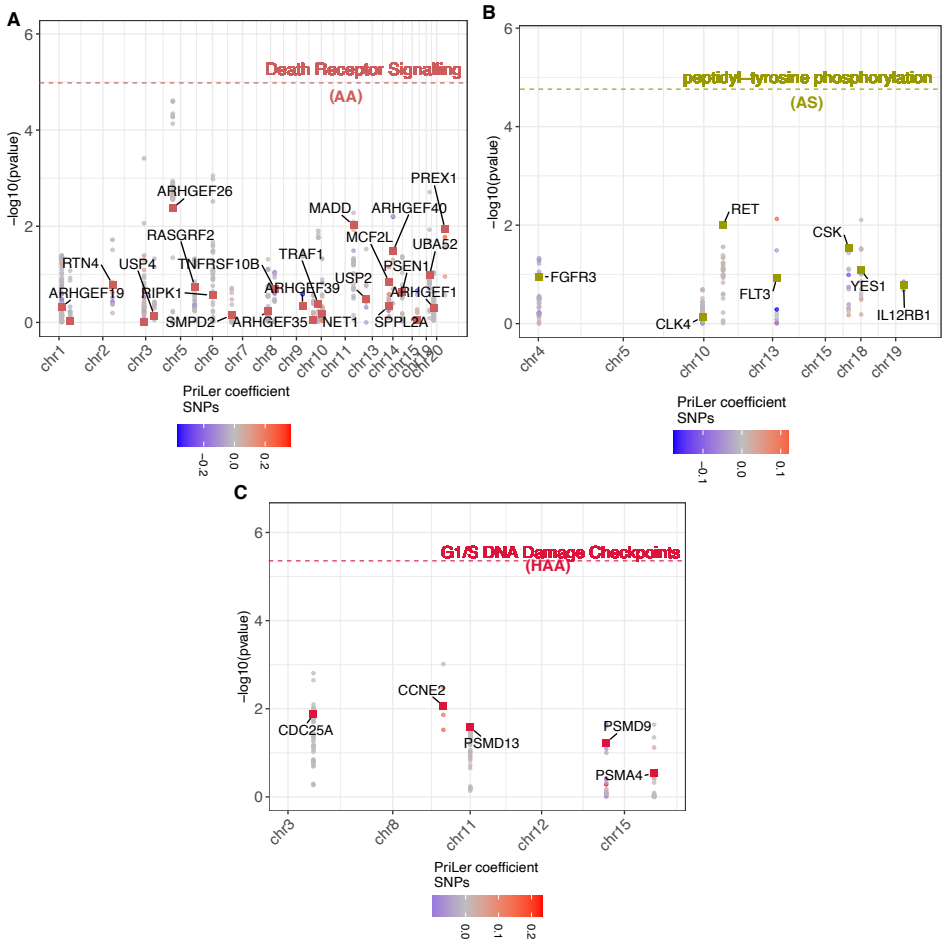

Fig. 4.19.: (Adapted from Trastulla et al., in prep.) Selection of pathways more significant than single genes: **(A)** Reactome *Death Receptor Signaling* in artery aorta, **(B)** GO *peptidyl-tyrosine phosphorylation* in adipose subcutaneous, and **(C)** Reactome *G1/S DNA Damage Checkpoints* in heart atrial appendage. Each panel shows $-\log_{10}(P)$ from TWAS of the genes included in that pathway (colored squares) and from matched GWAS of SNPs regulating those genes (dots, color reflecting PriLer regulatory coefficients on the bottom). The dashed line corresponds to $-\log_{10}(P)$ of the considered pathway from PALAS

Taking a closer look to pathways arising from an aggregation of effects, we observed three examples for which the pathway significance greatly exceeded the single gene level (Fig. 4.19), namely *death receptor signaling* in artery aorta, *peptidyl-tyrosine phosphorylation*

in adipose subcutaneous and *G1/S DNA damage checkpoints* in atrial appendage. The Manhattan plots in Fig. 4.19 also includes the level of significance from the matched GWAS, depicting only those variants that are used in the PriLer gene expression model of the T-score genes in the considered pathways. *Death receptor signaling* was constructed from 25 genes and reached a level of significance almost twice as the most relevant gene ARHGEF26. Nevertheless, the variants regulating ARHGEF26 were more significant than the gene itself and reached a level close to the PALAS p-value (Fig. 4.19A). Interestingly, ARHGEF26 did not pass FDR 0.05 significance level (Z-stat$= 2.86$, P$= 4.2$e-03, FDR$= 0.15$), but it was prioritized in the most recent CAD GWAS supported from multiple predictors and not simply based on its genomic location [74]. On the other hand, *peptidyl-tyrosine phosphorylation* and *G1/S DNA damage checkpoints* pathways greatly exceeded the level of significance of both genes and variants related to the pathway (Fig. 4.19B-C), revealing an aggregation of small effects into these pathways and suggesting possible pathomechanisms related to cell growth.

Since we considered the same set of individuals and variants, we can compare the 3 strategies GWAS, TWAS and PALAS, going from the single SNP association to aggregation effect on meaningful biological entities (Fig. 4.20). In particular, we hierarchically investigated pathways compared to the included genes (Fig. 4.20A), genes compared to SNPs in corresponding PriLer models (Fig. 4.20B) and pathways compared to SNPs in PriLer models of the corresponding genes (Fig. 4.20C). In particular, we found that the majority of the significant pathways included both genes and SNPs more significant (295, Fig. 4.20D), as well as the majority of the genes was less significant than the corresponding variants. This highlighted a limitation of our strategy that entirely rely on cis-regulation for predicting gene expression, while SNP can affect the phenotype via trans-regulatory mechanisms that are not captured in PriLer gene imputation and hence pathway scores. Nevertheless, we highlighted 104 pathways that are more significant than both genes and SNPs and included not significant FDR corrected genes and variants, thus indicating an aggregation mechanism. Examples of these latter class of pathways not detectable both from genes or SNPs filtering strategy are indeed the aforementioned peptidyl-tyrosine phosphorylation and G1/S DNA damage checkpoints (Fig. 4.18E).

In summary, these results demonstrate that some of the significant pathomechanisms are driven by many variants with small effect that aggregate into individual pathway levels. On the other hand, pathway dysregulation is also observable at the individual level due to a single gene effect, that still perturbs the entire mechanism although it might be not collaborative with the other genes in the pathway, .

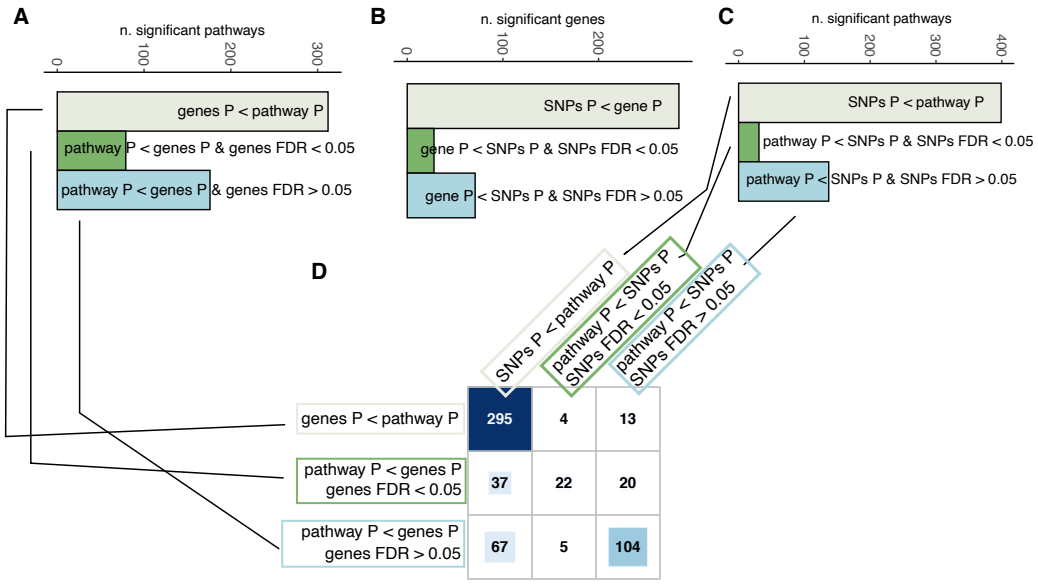

Fig. 4.20.: (Adapted from Trastulla et al., in prep.) Comparison of significance between **(A)** TWAS and PALAS, **(B)** matched GWAS and TWAS, **(C)** matched GWAS and PALAS. **(A)** Number of significant pathways that include at least one T-score gene more significant (ivory), all genes less significant and at least one passing FDR 0.05 threshold (green), and all genes less significant and not passing multiple-testing correction (light green). **(B)** Same as (A) but between SNPs and genes, with SNPs associated to a gene as those having a non-zero coefficient in the PriLer gene expression models. **(C)** Same as (A) but between SNPs and pathways, with SNPs associated to a pathway as those regulating (PriLer coeff. non-zero) the T-score genes included in the pathway. **(D)** Contingency table for significant pathway division in (A) and (C).

Before moving on towards the genetic relationship of CAD with related phenotypes, we show in the next paragraph that our CASTom-iGEx approach is well calibrated under null-hypothesis of no associations and that the pathways significance are not driven by gene correlation nor LD structure.

4.3.2 P-value calibration under null-hypothesis

To observe whether our approach provided well-calibrated p-values both at the gene and pathway-score levels, we considered whole blood as exemplar tissue that included $3,840$ genes, 902 Reactome pathways and $2,803$ GO pathways and we simulated random phenotypes 50 times. In detail, we created binary vectors that resembled CAD phenotype keeping the same case/control size, i.e. $19,026$ cases and $321,913$ controls. To create random phenotypes that included as much as possible the same confounders of the actual CAD classification, we selected the same number of females/males and the same age as what was observed in the actual CAD phenotype among the case/control classes. We then performed TWAS and PALAS and tested for associations between the randomly created phenotype and gene T-scores and pathway-scores previously computed for CAD (i.e. considering as reference set a subset of individuals non-affected by CAD, see section 3.2.1). Finally, multiple-testing correction is performed via BH procedure, correcting

for each simulation separately. Combing all the simulations, we observed that p-value distribution approximates a uniform distribution in $(0, 1)$ range (Fig. 4.21A-C), validated also via Kolmogorov-Smirnoff test that compared a random uniform distribution with the simulated one form gene associations (P$= 0.17$), pathway associations in Reactome (P$= 0.87$) and pathway associations in GO (P$=0.52$). The same conclusions can be drawn from quantile-quantile plots in Fig. 4.21D-F, with the expected distribution of p-value extracted from a uniform one. It is possible to observe that the association with the actual CAD phenotype greatly diverged from the simulated ones, and that few genes/pathways passed FDR 0.05 threshold in the simulated phenotypes (blue points), nevertheless they remained in the 95% confidence intervals of the standard uniform order statistics that follows a beta distribution. We can then conclude that CASTom-iGEx strategy for TWAS and PALAS returns well-calibrated p-values.

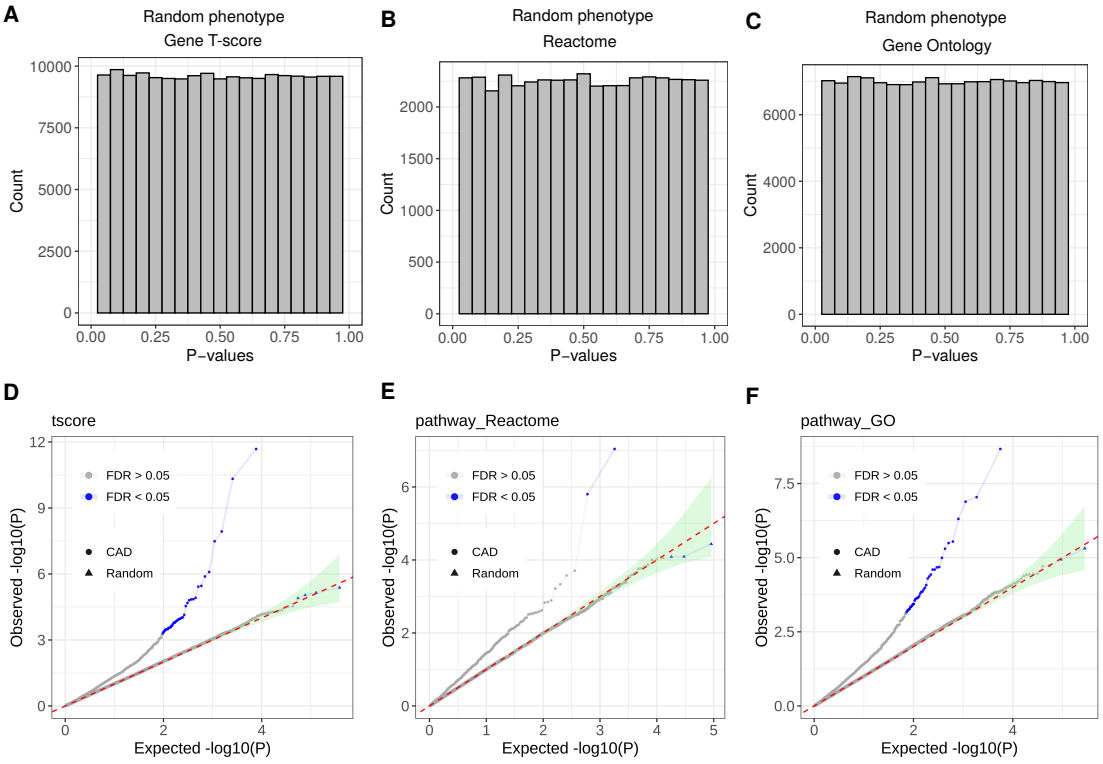

Fig. 4.21.: (From Trastulla et al., in prep.) P-value distribution from TWAS and PALAS (Reactome and GO databases) in whole blood for random phenotype across 50 simulations matching cases/controls sizes and age/sex distributions. **(A-C)** Count of p-values in binned intervals for associations with random phenotypes of **(A)** gene T-scores, pathway-scores in **(B)** Reactome and **(C)** GO. **(D-F)** Expected and observed distribution of p-value for CAD (dot) and random phenotype simulations (triangle) associations with dashed line indicating the diagonal and shaded green area representing 95% confidence interval from beta distribution for **(D)** gene T-scores, **(E)** pathway-scores in Reactome and **(F)** GO. Blue points indicates genes that are significant at 0.05 FDR level, correcting CAD and each simulation separately.

4.3.3 Gene correlation effect on the improvement of pathway significance

Next, we investigated whether genes correlation and LD structure were connected to the increase observed in pathway-score significance. To this aim, we performed three analyses:

Study 1 Simulation of gene T-scores with correlated genes and consequential simulation of pathway-scores and phenotype. Here the aim was to understand to what extent the correlation among genes, supposedly all relevant in the phenotype etiology, is affecting the level of significance in the pathway analysis.

Study 2 Simulation of pathway structure from actual gene T-scores in whole blood, creating gene-sets composed of 3 or more genes located in the same loci, for a total of 46 simulated pathways. In this case, the goal was to understand how the loci structure can influence the pathway significance.

Study 3 Estimation of relationship between pathway significance increase and average gene correlation across all detected pathways with $n_{\mathcal{P},\mathcal{G}} \geq 2$, to observe the actual relevance and extent of genes correlation in pathway significance.

For **Study 1**, we simulated $N = 10$ gene T-scores for $M = 10,000$ samples $(\boldsymbol{T}_1, \ldots, \boldsymbol{T}_N)$ following a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \Sigma)$. The covariance $N \times N$ matrix Σ was defined a priori with diagonal variance entries set to 1 and $\Sigma_{i,j}$ (namely the correlation between gene i and j) randomly assigned either from a uniform distribution in varying ranges (Fig. 4.22A), or fixing the uniform distribution interval to $[0.9, 1]$ but varying the number of correlated genes from none to all (Fig. 4.22B). Consequentially, pathway-scores across M samples where computed as described in (3.21), considering a single pathway as composed of the 10 simulated genes. Finally, a binary phenotype was simulated across the M samples from gene T-scores via a Bernoulli distribution with probability p from

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \boldsymbol{T}_1 + ... + \beta_N \boldsymbol{T}_N,$$

equal effect sizes $\boldsymbol{\beta} = (0.15, \ldots, 0.15)$ and intercept

$$\beta_0 = log\left(\frac{p_{CAD}}{1-p_{CAD}}\right),$$

with p_{CAD} the number of fraction of CAD cases in the UKBB cohort. Keeping the same number and parameters, we repeated this simulation 100 times. Note that, the decision to keep effect sizes fixed as $\boldsymbol{\beta} = (0.15, \ldots, 0.15)$ implies that all the genes are always relevant to the phenotype, even when they are correlated and that the effect size is always concordant in the direction, and hence not canceling out the effects between relevant genes in the pathway-score computation (see section 4.4.2 for details). With the variation

of the correlation ranges consistent across all the gene pairs, we observed that the average improvement of significance for pathway-scores with respect to the genes in the pathway decreased with the increase of the correlation intervals (Fig. 4.22A), reaching a peak when a mild correlation is observed across genes ($0.1 \leq \text{corr} \leq 0.2$).

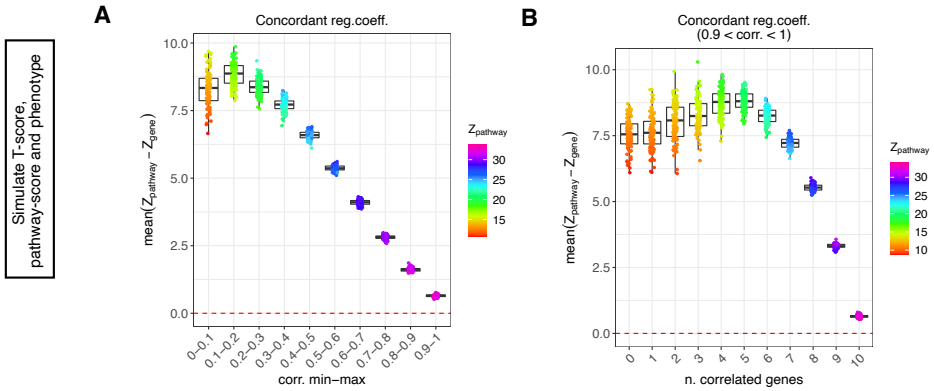

Fig. 4.22.: Simulations of 10 gene T-scores from a normal distribution $\mathcal{N}(0, \Sigma)$, phenotype simulated from a logistic regression model with all genes contributing equally $\beta_m = 0.15$ and pathway-scores simulated as individual-level means of T-scores, repeated 100 times. **(A)** Covariance matrix defined from a uniform distributions with values extracted between the ranges on the x-axis and y-axis showing average differences of Z-statistics between simulated pathway and gene. Each dot represent a simulation with color code indicating pathway Z-statistic. **(B)** Similar to (A) but number of correlated genes showed on the x-axis with values extracted from a uniform distribution in $[0.9, 1]$.

Despite the Z-statistic increase for higher ranges, the improvement is minimal and close to zero when genes T-scores are almost identical ($0.9 \leq \text{corr} \leq 1$). In addition, the improvement in pathway significance quickly decreased when keeping the same correlation sampling range ($0.9 \leq \text{corr} \leq 1$) but varying the number of correlated genes from 5 to 10 while it slowly increased between 0 and 5 (Fig. 4.22B). Having assumed that all the genes have impact in the phenotype distribution, there was an increase in the pathway significance with respect to the singular genes when the correlation was mild or limited to few ones, that however was close to zero when all the genes included are highly correlated. For **Study 2**, we only considered actual genes in whole blood that were showing a certain level of significance i.e. nominal TWAS p-value ≤ 0.01 and created simulated gene-sets from those genes that were also in the same loci and had the same effect size sign in CAD associations (all Z-stat genes > 0 or < 0), again to avoid a compensatory effect for gene relevance. This procedure led to a total of 46 simulated pathways with the number of genes included varying from 3 to 7. Although all the genes were in the same loci, the increase in pathway significance was dependent and inversely proportional to the estimated average genes correlation (Fig. 4.23A), with almost no increase for pathways that included highly correlated genes, in accordance to *Study 1*. This resulted in a general lower significance of pathways composed of correlated genes (Fig. 4.23B). We can then concluded that the gene correlation due to the regulation from the same variants (or in LD with them) rather than the vicinity of genomic coordinates is relevant in the observed pathway significance and that genes in the same loci not correlated do still lead to an improvement in the

information captured by the pathway scores.

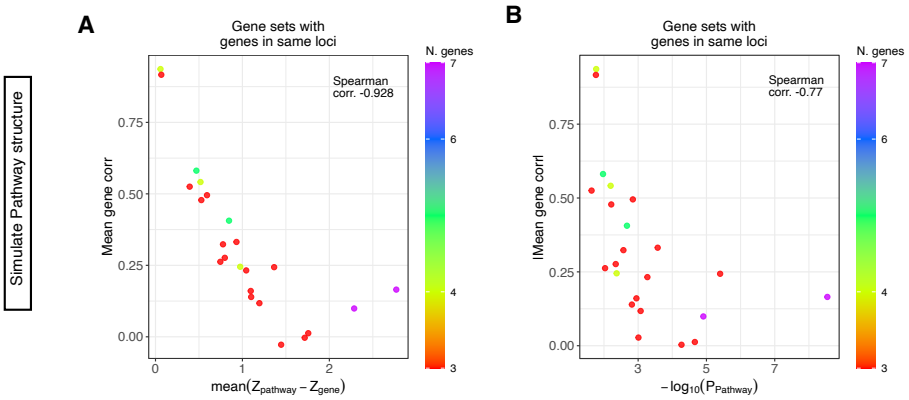

Fig. 4.23.: (Adapted from Trastulla et al., in prep.) Simulation of pathway structure using genes in whole blood from the same locus, having effect size sign concordant and TWAS nominal p-value < 0.1 for a total of 46 pathways. **(A)** Each point represent a simulated pathway with color code indicating the number of genes. X-axis indicates the average differences in Z-statistic between pathway and included genes and y-axis shows the absolute value of mean correlation among genes in a pathway. **(B)** Similar to (A) but with x-axis showing -$\log_{10}(P)$ from PALAS

For **Study 3**, we finally considered the actual pathway-scores and increase or decrease in pathway association level compared to the average genes correlation included in the pathways. Across all the pathways databases, there was no rank correlation between average differences of significance in pathways versus genes and genes correlation (Fig. 4.24, absolute Spearman corr. < 0.045).

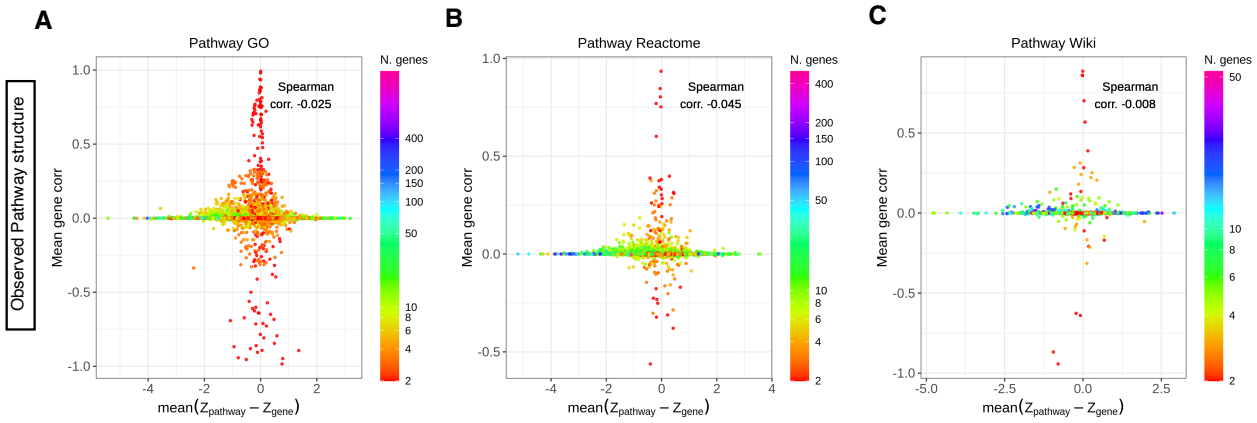

Fig. 4.24.: (Adapted from Trastulla et al., in prep.) Relationship of pathway significance improvement with respect to the included gene depending on gene correlation. Each point represents a pathway in **(A)** GO, **(B)** Reactome and **(C)** WikiPathways with color code indicating the number of genes included in the pathway-score calculation. In all the panels, x-axis shows the average differences in Z-statistics between a pathway and the corresponding genes and y-axis shows the mean gene correlation for those genes.

Indeed, we observed that pathways with highly correlated genes ($> |0.5|$), usually including less than 4, showed only marginal improvement in pathway significance. In contrast, pathways with a striking effect of increased significance were those formed by more than 10 genes and having and average correlation around zero.

Hence, we conclude that the increase in pathway relevance with respect to single genes became maximal when the correlation among genes was minimal. Overall, genes correlation due to LD structure did not increase pathway significance nor pathway improvement compared to single genes. Finally, observing actual pathway structures, the gene-sets with best improvement were formed by not correlated genes.

4.3.4 Phenotypic interpretation of genes and pathways

In the next step, we sought to pinpoint disease relevant endophenotypes and clinical features associated with the specific biological pathways subject to excessive CAD genetic liability. To that end, we leveraged the rich collection of phenotypes with potential relevance to CAD within UKBB (316 in total, Tab. B.4) and performed tissue-specific correlation analysis and Mendelian Randomization (MR) with respect to CAD associated genes or pathways (see section 3.2.3).

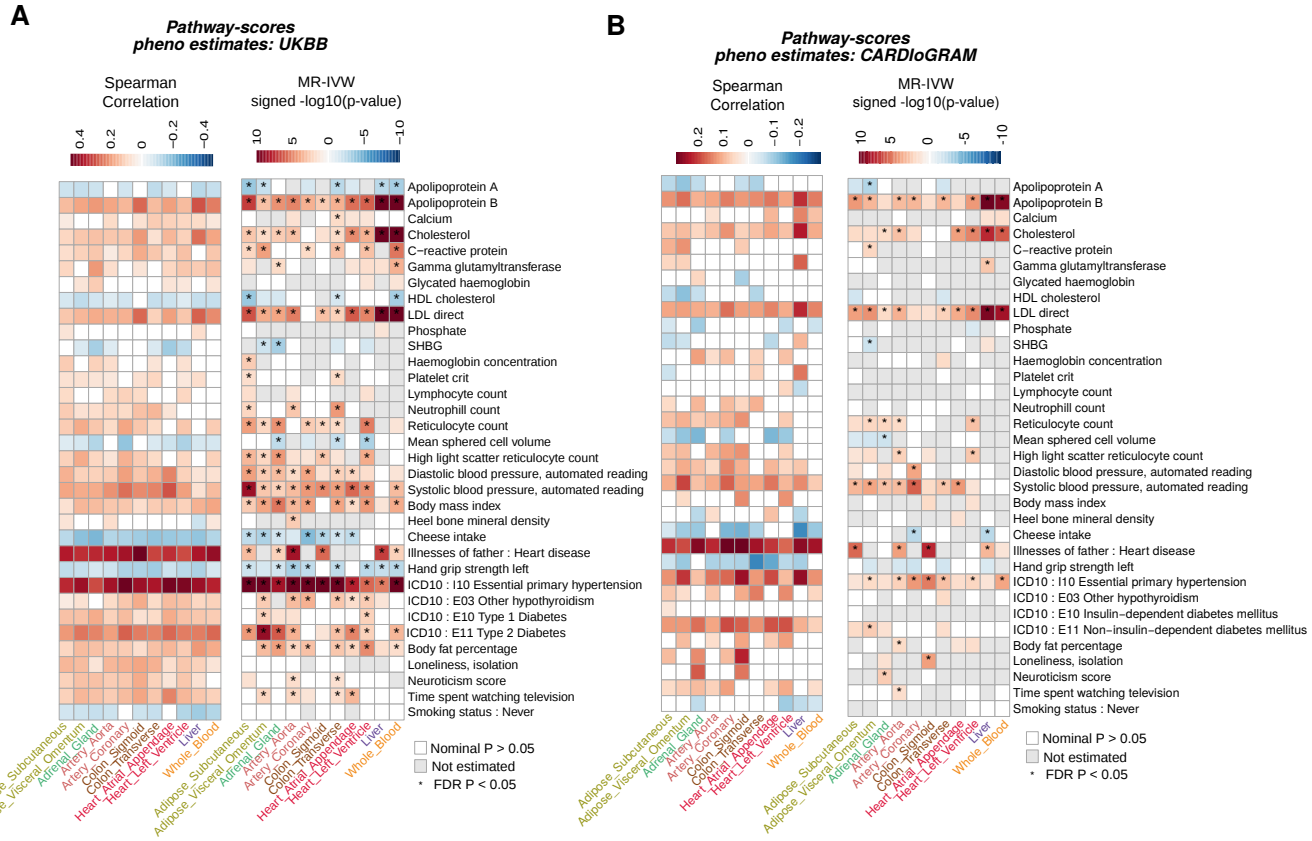

Fig. 4.25.: Correlation and causality of CAD and CAD related phenotypes from combined GO and Reactome (randomly pruned based on Jaccard index < 0.3). Heatmap on the left shows tissue specific Spearman correlation of Z-statistics between CAD and selected phenotypes (rows), white cells indicate nominal permutation p-value > 0.05. Heatmap on the right shows -$\log_{10}$ p-value with the sign indicating the direction of estimate from MR-IVW for correlated phenotypes (not grey cells), white cells indicate nominal MR-IVW p-value > 0.05, * in correspondence of 0.05 FDR p-values (tissue specific correction). **(A)** CAD effect and endophenotype effects both from UKBB. **(B)** Replication with CAD effect from CARDIoGRAM and endophenotype effects from UKBB.

Here the aim is to understand which endophenotypes are genetically similar to CAD and among those identify endophenotypes that exhibit a causal or protective role via MR, with endophenotype regarded as exposure (e.g. LDL), CAD as outcome, and genes or pathways as instrumental variables. In Fig. 4.26 and 4.25 is shown the estimated Spearman correlation and signed MR significance from MR-IWV method for a selection of CAD related endophenotypes using genes T-scores or pathway-scores are instrumental variables, respectively.

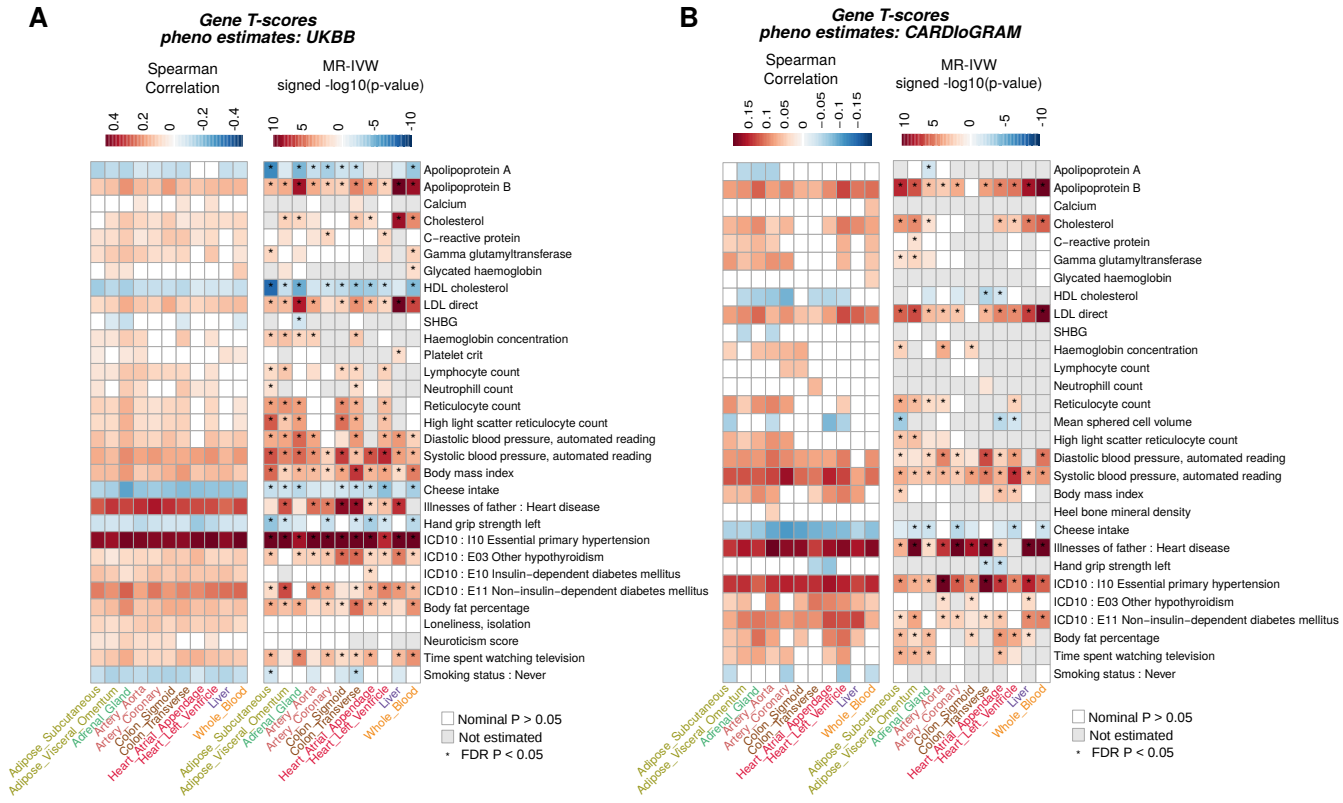

Fig. 4.26.: Correlation and causality of CAD and CAD related phenotypes from gene T-scores (randomly pruned based on TSS $> 250kb$). Heatmap on the left shows tissue specific Spearman correlation of Z-statistics between CAD and selected phenotypes (rows), white cells indicate nominal permutation p-value > 0.05. Heatmap on the right shows -$\log_{10}$ p-value with the sign indicating the direction of estimate from MR-IVW for correlated phenotypes (not grey cells), white cells indicate nominal MR-IVW p-value > 0.05, * in correspondence of 0.05 FDR p-values (tissue specific correction). **(A)** CAD effect and endophenotype effects both from UKBB. **(B)** Replication with CAD effect from CARDIoGRAM and endophenotype effects from UKBB.

This analysis identified multiple well-established physiological parameters associated with CAD, such as LDL cholesterol ("LDL direct" [211]), apolipoprotein B [215], blood pressure ("systolic/diastolic, automated reading") and hypertension ("ICD10 : I10") [216]. The aforementioned examples were consistent across genes and pathways, replicated via CARDIoGRAM CAD estimates and indicating both a positive correlation effect as well as a significant causal role in CAD etiology, with the only exception of diastolic blood pressure that exhibited a reduced consistency of association across all tissues, at least

for replication in CARDIoGRAM via pathway-scores (Fig. 4.25B). Moreover, a protective role of HDL cholesterol was identified in whole blood, adipose subcutaneous and colon transverse tissues from pathway levels estimated in UKBB (MR-IVW estimate P= 5.9^{-4}, 2.5^{-4} and 0.021, Fig. 4.25A). A similar trend, however not significant was observed when considering CARDIoGRAM CAD replication data as outcome (Fig. 4.25B). HDL protective role was also observed in gene T-scores across all tissues, however replicated passing FDR 0.05 significance only in colon transverse and atrial appendage (Fig. 4.26). The effect of increasing HDL as CAD therapeutic are still controversial [217] and further investigation should go toward the direction of multi-variable analysis to account for the pleiotropic effect of genes and pathway in the lipid landscape [218].

In addition, this analysis supported the notion of a partially genetically mediated inflammatory contribution to CAD, based on elevated C-reactive protein levels as an endpoint of genetically increased susceptibility to the activation of inflammatory processes [129]. In particular, we detected a significant causal effect on CAD from C-reactive protein using pathways as genetic instruments in multiple tissues (Fig. 4.25A), with the strongest effect in whole blood (P= 7.83e-06) that included 311 pruned pathways associated with C-reactive protein. Importantly, this causal effect was also replicated for adipose visceral from 266 pathways using UKBB and CARDIoGRAM CAD estimates (P discovery= 6.83e-05, P replication= 5.58e-03, Fig. 4.25). This association in adipose visceral could still be observed from genes variation, although with a smaller effect and not significant for CAD estimated from UKBB (P discovery= 0.028, P replication= 0.020, Fig. 4.26). Importantly, our strategy allowed to identify underling pathways and genes that regulate both CAD and a mediating endophenotype. Indeed, key molecular gene-sets influencing both C-reactive protein and CAD in adipose visceral were lysosome, inflammatory response and post-translation protein modification pathways (Fig. 4.27A) and driver genes of the observed causal effect could be identified in LIPA, VPS13C and RPM6 (Fig. 4.27B). Similarly, for a established effect such as LDL and CAD in liver, we identified as exemplar pathways clathrin-coated pit, neuropeptide signaling, and endosome membrane (Fig. 4.27C), and among pivotal genes SORT1, PCSK9, and AGPAT4 (Fig. 4.27D). Of note, the heterogeneity in CAD-endophenotype relationship tested via Cochran's Q statistic was always significant in all 4 panels of Fig. 4.27 (Q-stat P $< 10^{-6}$), indicating a pleiotropic effect that require further investigation.

Interestingly, we also found intermediate phenotypes that, despite being genetically correlated through pathway association levels with CAD, did not show a casual significant effect in CAD etiology such as loneliness or glycated hemoglobin (Fig. 4.25A). In addition, this analysis identified a causal effect between time spent in television watching and CAD through pathway effects, which was also replicated in artery aorta (P discovery= 6.3e-03 and P replication= 1.2e-02, Fig. 4.25) and recently supported by SNP-based genetic studies [219]. Surprisingly, a protective role for cheese consumption in artery coronary was found from pathways (P discovery= 6.84e-05 and P replication= 2.68e-03), and confirmed from genes (P discovery= 0.05 and P replication= 2.38e-03). This was in accordance with a detected non-linear inverse association with cardiovascular disease risk from a

meta-analysis of prospective studies [220].

In summary, these analyses identified key disease relevant endophenotypes associated with CAD and connected them to the underlying genes and pathways. Thus, these results exemplify the utility of the CASTOM-iGEx approach to directly identify and interpret the (endo-) phenotypic impact of disease associated genes and pathways.

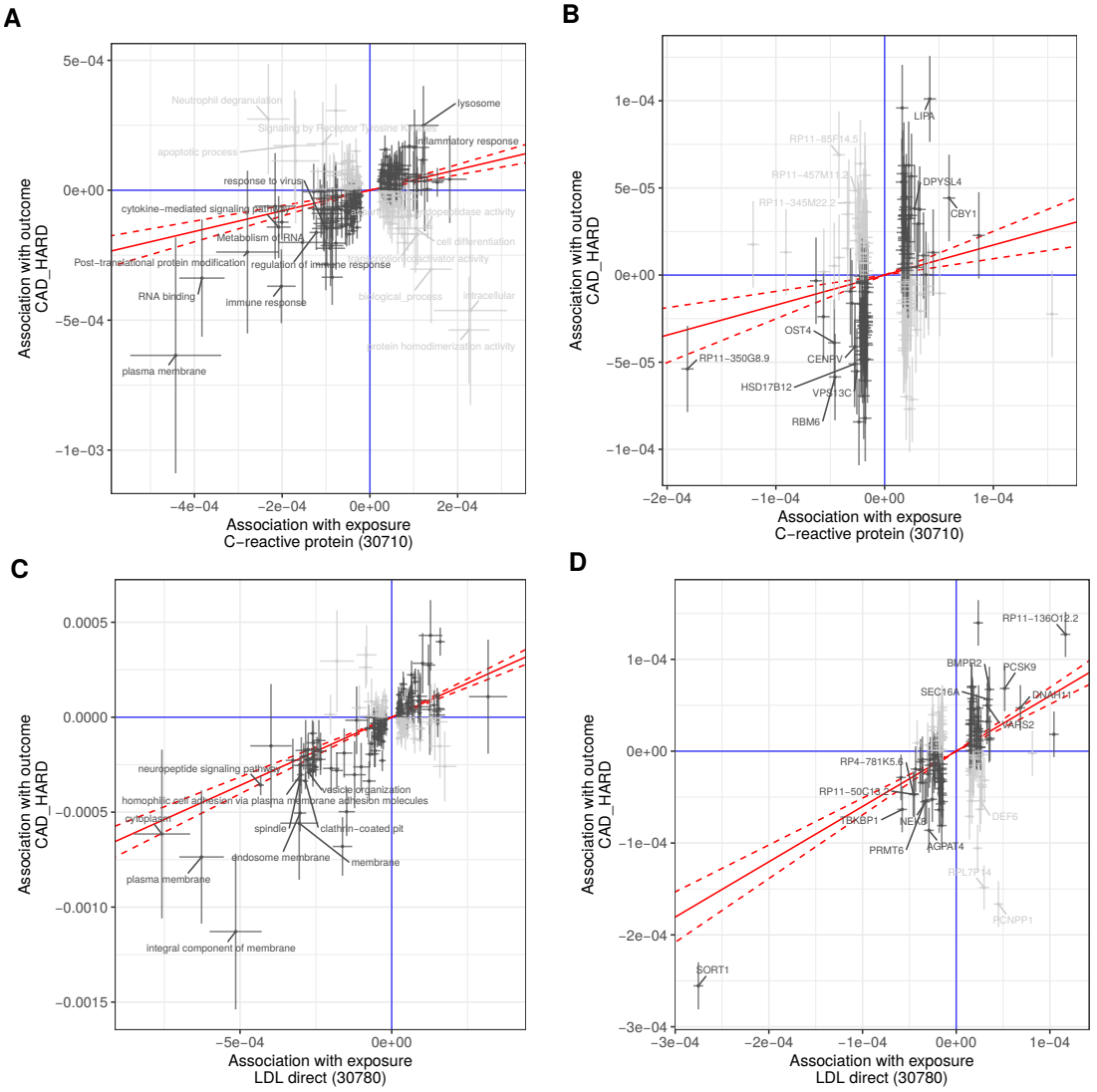

Fig. 4.27.: Scatter plots of the effect sizes with 95% confidence interval for **(A-B)** C-reactive protein in adipose visceral omentum or **(C-D)** LDL direct in liver from UKBB (x-axis) and CAD from UKBB (y-axis). **(A,C)** panels show MR-IVW from pathway-scores and **(B,D)** from gene T-scores. In each panel, the red line represents the causal estimate using the IVW with 95% confidence interval, in black and gray pathways/genes with association concordant and discordant in sign respectively between CAD and the endophenotype.

4.3.5 Patients stratification from imputed gene expression

Coronary artery disease is characterized by a heterogeneity both at the genetic and symptom levels [135]. With the last module of CASTom-iGEx methodology (Fig. 3.1), we

aim at testing the hypothesis of clinical features of a patient being connected to differences in genetic liability. Hence, we stratified CAD patients using the imputed gene expression profiles, with a strategy that gives higher relevance to CAD associated genes, partitioning cases in the genetic disease context.
We primarily focused on CAD UKBB cohort as it includes the majority of individuals, nevertheless we further show that the pipeline is extendable to predict clustering status to external cohorts (of similar ancestry), e.g. cases from CARDIoGRAM cohorts. In particular, we considered gene T-scores and applied the clustering pipeline described in section 3.3.1 to cluster individuals ($n = 19,024$), separately across the 11 CAD related tissues. The clustering detected 3 to 5 groups across tissues, with a variety of patients distribution per group that went from 278 in adrenal gland to 8642 in visceral omentum (Fig. 4.28A). We also observed differences in Louvain modularity, representing a internal quality metric of the density of connection inside the groups compared to between groups (details in eq. (3.38)), with the best result achieved in liver (0.41). In addition, clustering structure largely overlapped across tissues (Fig. 4.28B) with the only exception of colon sigmoid. Indeed, the clustering in that tissue was largely influenced by the strongest genes associated with CAD i.e. CDKN2B that is nevertheless only imputed in colon sigmoid.

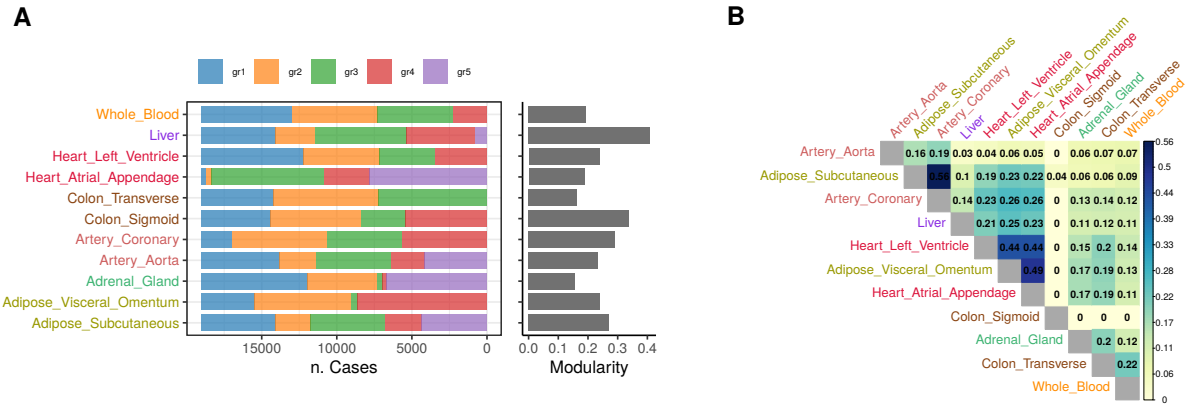

Fig. 4.28.: (From Trastulla et al., in prep.) **(A)** Proportion of CAD cases from UKBB in each tissue-specific grouping (left) and modularity from Louvain clustering (right). **(B)** Pairwise normalized mutual information (NMI) for each tissue combination of tissue-specific clustering.

Due to the relevance of liver in CAD pathophysiology, the strongest separation metric for clustering, and a good concordance in clustering structure with the other tissues, we focused on the stratification arising from liver for the investigation of genetic features and endophenotype liability characteristic of each group.

4.3.6 Patients stratification in liver

The clustering was performed on $19,024$ individuals affected by CAD in UKBB using $1,892$ gene T-scores from liver that are clumped at 0.9 correlation, standardized, corrected for the first 10 PCs and multiplied by CAD Z-statistic obtained via TWAS (see section 3.3.1

for details). Of note, the PCs correction step was essential to reduce the impact of ancestry in the final clustering structure (see section 4.3.7), whereas the multiplication of each gene by CAD Z-statistic conferred a higher relevance to CAD-related genes as well as a better community partition in terms of density (see section 4.3.8).
This procedure identified 5 distinct groups (Fig. 4.29A) as a result gene combination from multiple loci. In particular, we investigated the predicted gene expression differences across all tissues (not simply the tissue considered for clustering) via Wilcoxon-Mann-Whitney (WMW), testing one group at a time versus the remaining patients and correcting for multiple testing for each tissue and each group separately (see section 3.3.2).

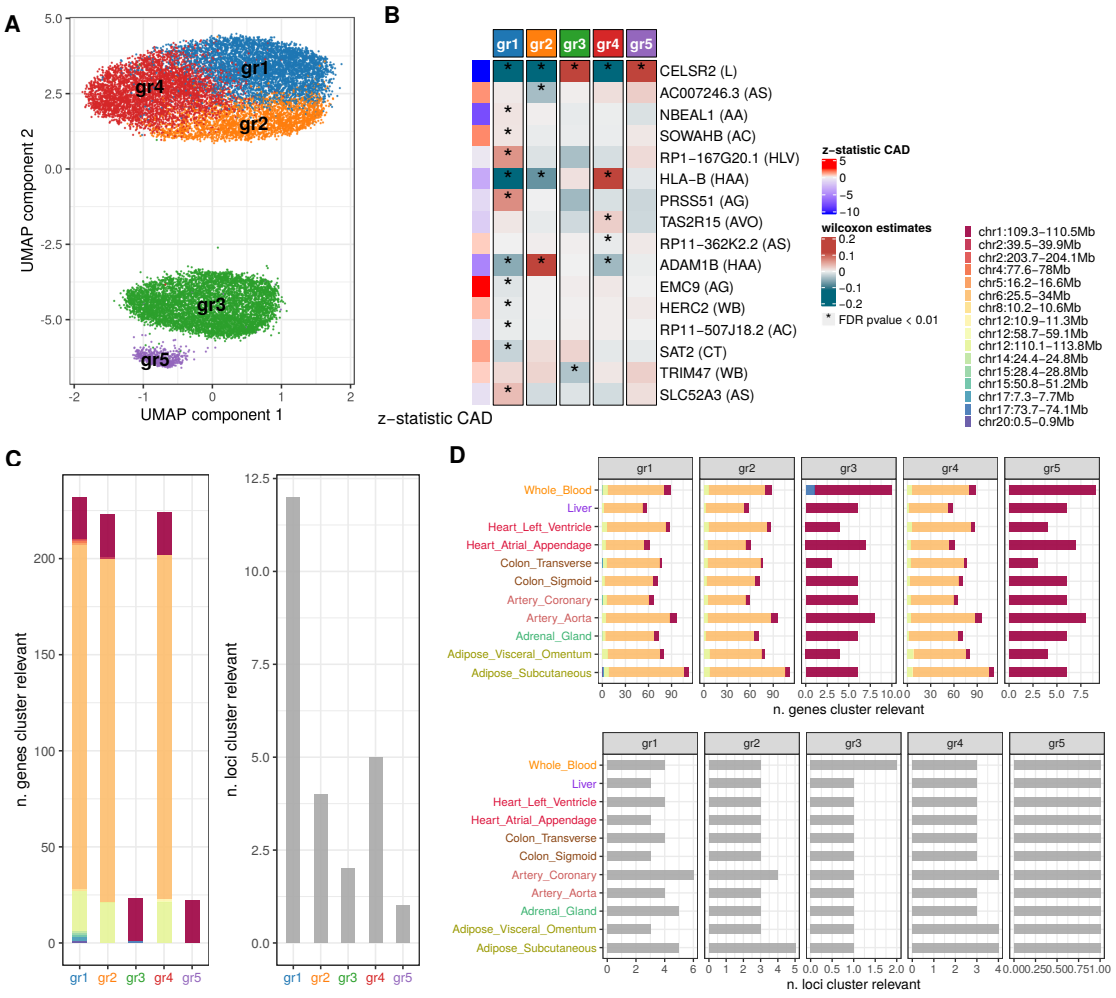

Fig. 4.29.: (From Trastulla et al., in prep.) **(A)** First 2 components of uniform manifold approximation and projection (UMAP) from gene T-scores in Liver normalized across CAD patients corrected for PCs and multiplied by Z-statistic CAD associations. Each dot represents a patient colored by the cluster membership. **(B)** WMW estimates (capped) for the most group-specific significant gene in the 16 associated loci, parenthesis refers to the tissue considered (acronyms refer to the initial of the tissue name). Row annotation on the left indicate the corresponding CAD Z-statistics from TWAS. **(C-D)** Number significant genes and loci (tissue specific FDR ≤ 0.01) associated with each group from Mann-Whithney-Wilcoxon test of group id versus remaining patients **(C)** combing all tissues and **(D)** tissue specific.

We considered as cluster-specific genes those satisfying FDR ≤ 0.01, using a more stringent

threshold to reduce false positive associations (see section 4.3.9). Out of the $36,397$ tested genes, we identified 887 cluster-specific ones which collapsed into 50 tissue-specific loci (Tab. B.5), and 236 unique genes across all tissues which merged into 16 loci (Fig. 4.29B). The highest number of genes and loci was detected in gr_1, followed by group gr_4 and gr_2 both at the overall and tissue-specific level (Fig. 4.29C-D). On the other hand, gr_3 and gr_5 were mostly driven by genes in SORT1 locus (chr1:109.3-110.5Mb) apart from chr17:73.7-74.1Mb locus that is associated with gr_3 in whole blood. The specific configuration of imputed gene expression leads to the resulting clustering structure, thus even correlated genes in a locus can have an independent impact in the final clustering structure (see section 4.4.5). Nevertheless, to provide a general overview, we reported in Fig. 4.29B a single exemplar for each of the 16 cluster-specific loci across all tissues. The most prominent differences are CELSR2 in liver (similar trend as SORT1, $P <$ 1e-258 for all groups), HLA-B in MHC locus in atrial appendage for gr_4 and gr_5 (P $= 0$) and ADAM1B in chr12:110.1-113.8Mb for gr_2, gr_1 and gr_4 (P $<$ 2e-132). Interestingly, SORT1 in liver had a positive WMW estimate in gr_5 and gr_3 but higher in the former (WMW estimates $= 3.41$ and $= 1.74$ respectively). Decrease in SORT1 are associated with an increased risk of CAD (Fig. 4.29B row annotation), that thus conferred a lower severity to gr_3 and gr_5, as we will investigate in detail in the endophenotype analysis.

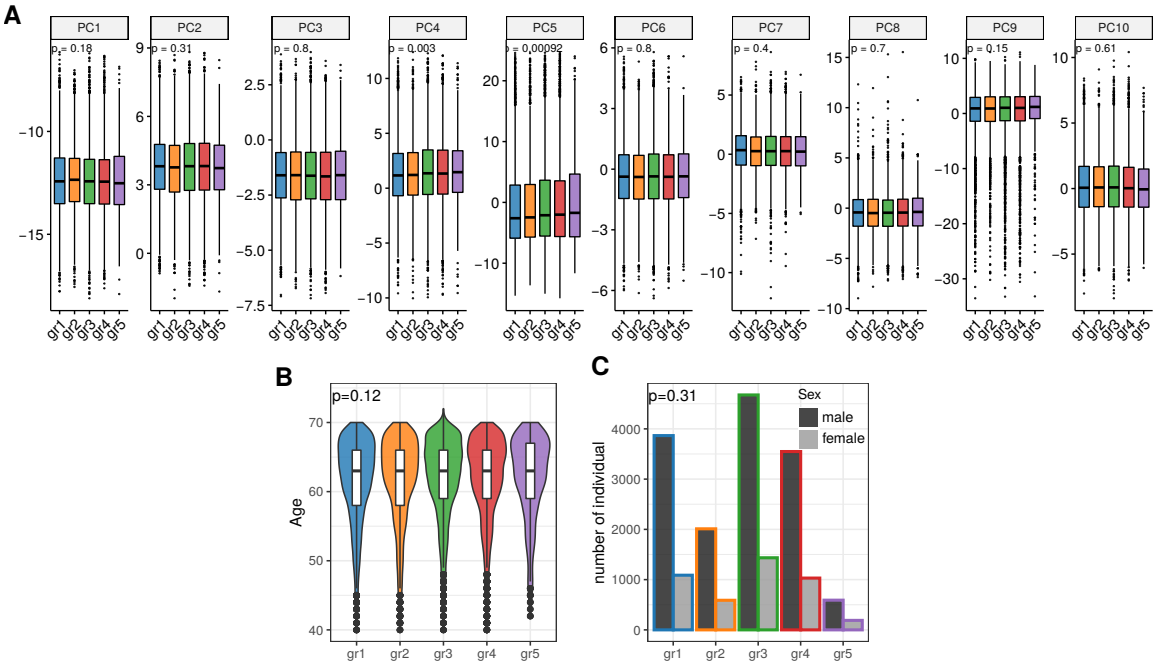

Fig. 4.30.: (Adapted from Trastulla et al., in prep.) **(A)** Distribution of UKBB PCs from 1 to 10 for each CAD liver cluster (p-values from Kruskal-Wallis test). **(B)** Distribution of age and sex for each CAD liver cluster (p-values from Kruskal-Wallis and Chi-squared test respectively).

As regards genes with smaller effect sizes, we found TRIM47 in whole blood associated with gr_3 (P $= 1.96$e-05), EMC9 in adrenal gland and SAT2 in colon transverse for gr_1 (P $= 1.2$e-04 and 5e-05, respectively). TRIM47 intersects variants associated to LDL from GWAS [221] as well as from our methodology (FDR $= 0.12$, Z-stat $= 2.36$ for LDL in whole

blood), suggesting that gr_3 is characterized by differences in LDL metabolism not only driven by SORT1 locus. Moreover, the clustering structure was not a consequence of age and sex among CAD patients (Fig. 4.30B-C) and a mild association was only detected with PC4 and PC5 (Fig. 4.30A). Note that, adjusting gene T-scores for PCs was essential in drastically reducing PC4 and PC5 association with clustering structure (Fig. 4.35), nevertheless a mild effect persisted. Thus, we additionally clustered patients solely based on PCs to investigate the overlap with the liver tissue-specific clustering. Despite the two patients partitions not being completely independent, cluster-specific endophenotypes observed in each clustering were divergent and we concluded that patients ancestry was not driving the observed clustering structure, nor the pathophysiological consequences (more details in section 4.3.7).

To evaluate the reproducibility of our clustering structure, we projected the gene-level T-scores from 9 CARDIoGRAM cohorts into the partition built on UKBB (see section 3.3.1). After having predicted the grouping label for each cohort separately, we computed the fraction of CAD affected individuals in CARDIoGRAM that were attributed to each cluster and compared it to the actual fraction in the UKBB clustering, finding a concordant partition (Fig. 4.31A).

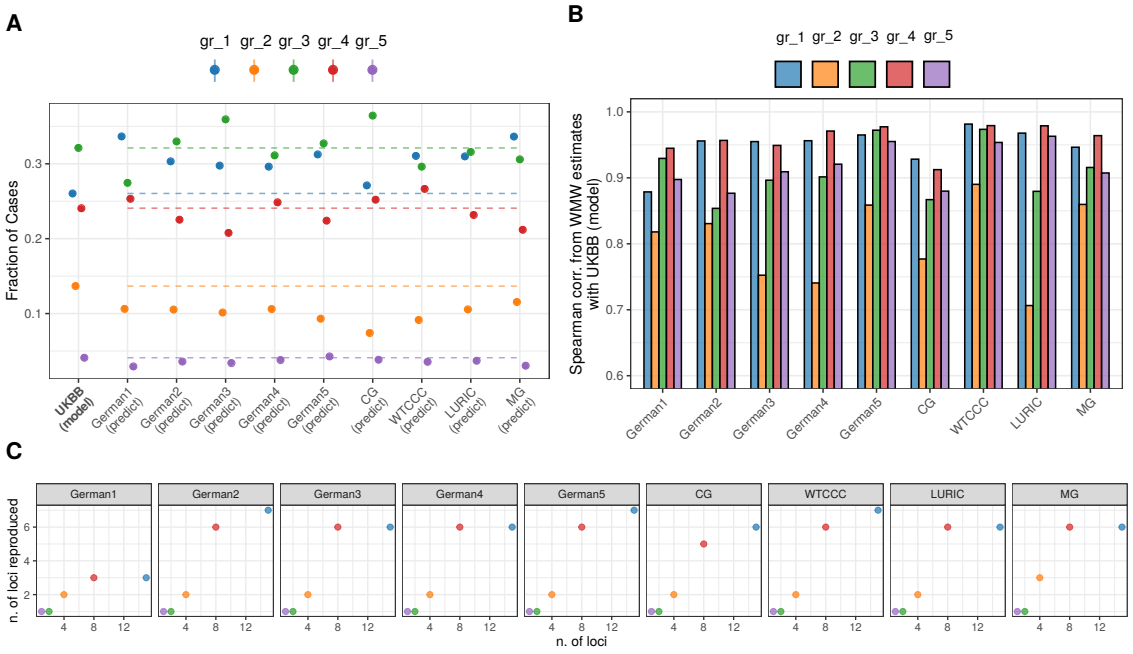

Fig. 4.31.: (Adapted from Trastulla et al., in prep.) For each cohort in CARDIoGRAM, prediction of liver clustering structure on 9 external cohorts. **(A)** Y-axis indicates the fraction of CAD patients assigned to each group in UKBB data set and each CARDIoGRAM cohort for which the clustering structure was projected. **(B)** For each group, Spearman correlation of WMW estimates in UKBB and each external cohort using genes that are significantly associated with that group in UKBB. **(C)** Reproducibility of group-specific loci on predicted groups in external cohorts, the x-axis shows the number of loci across all tissues associated with each group in UKBB, the y-axis shows how many of these loci have the same sign and are significant at the nominal level of 0.05, using as exemplar the strongest association of the WMW estimates in the predicted clustering structure.

In addition, confronting the gene expression profiles from WMW group-specific estimates

for cluster significant genes, we found that the structure predicted in CARDIoGRAM was concordant with UKBB across all the cohorts (Spearman correlation > 0.7, Fig. 4.31B). Finally, this significance was not driven by a single locus, but consistency in terms of WMW estimates was observed in multiple cluster-specific loci detected in UKBB (Fig. 4.31C). These analyses highlight the possibility to generalize the identified CAD patient stratification even among cohorts of variable European ancestries genotyped with different platforms, as the case of CARDIoGRAM cohorts.

Apart from individual genes, we additionally investigated the differences in biological processes characteristic of the 5 identified groups, testing group-specific distribution in pathway-scores. To avoid redundant information, for each tissue we first reduced the tested pathways, clumping genes-sets in GO and Reactome with Jaccard similarity > 0.2, while giving priority to gene-sets with highest coverage and number of gene T-scores (see section 3.3.2). We thus tested $7,978$ filtered pathways across 11 tissues via WMW with the same strategy used for genes (see section 3.3.2). This resulted in $1,321$ significant associations across all groups and tissues. Because pathways in different tissues could potentially include different genes, we then removed significant pathways shared among tissues but having a non concordant WMW estimation sign. This led to $1,140$ significant results ($gr_1 = 482$, $gr_2 = 58$, $gr_3 = 56$, $gr_4 = 488$, $gr_5 = 56$) for a total of 271 unique pathways, with artery aorta and liver showing the highest number of associations (Fig. 4.32A). In addition, the cumulative number of group-specific associations increased faster when decreasing the p-value threshold for gr_1 and gr_4 (Fig. 4.32B) since these groups are more heavily affected by changes in the MHC-locus, hence including a high number of genes and consequentially gene-sets. In details (Fig. 4.32C), we observed an increase in pathway-activity levels for Golgi Associated Vesicle Biogenesis in liver driven by the increase in SORT1 expression for gr_3 and gr_5 (estimates$= 0.92$ and 1.62 respectively, P $=0$ for both) and a significant lower distribution for the other groups. In the next paragraph we will show that this increase in golgi associated vesicles was connected to a relative reduction in term of LDL circulating in blood, concordant with the notion that vesicles filled with LDL are taken up by the cells via receptor-mediated endocytosis mechanisms. In addition, N-acetyltransferase activity pathway was increased solely in gr_3 (P=5.5e-05, estimate $= 0.064$) from the cumulative aggregation of the single genes (ELP3, NAT10, SAT2, NAGS) that reach a group-specific significance (genes P≥ 2.84e-04, estimates ≤ 0.03) and none of them is in proximity of the SORT1 locus, highlighting a different mechanism characteristic of that group that only arises as a cumulative effect. Of note, 41 pathways out of the total group-specific associations achieved a higher significance than the single genes. Moreover, we found that gr_2 was significantly decreased in alcohol metabolic process in left ventricle (estimate$= -0.62$, P$= 2.64$e-173) due to ALDH2 gene (locus chr12:110.1-113.8Mb) whereas gr_1 and gr_4 showed an opposite significant effect (estimates> 0.14, P< 1.91e-14). Interestingly from our PALAS analysis, this pathway decrease was mildly associated with LDL increase in liver (Z-stat$= -1.47$, P$= 0.14$) but with a stronger effect in other tissues such as artery aorta (Z-stat$= -4.16$, P$= 3.10$e-5). Indeed, we will show in the next paragraph that this was actually reflected in a increase in terms of measure LDL

for gr_2. Finally, group-specific pathways only in gr_1 and gr_4 were of opposite effect, mostly related to inflammatory mechanisms and lipid metabolisms. These pathways conferred a general higher risk to gr_1 due to an association sign mostly concordant to CAD risk (4.32C), nevertheless with some exceptions such as 3-hydroxyacyl-CoA dehydrogenase activity (CAD Z-stat$= 1.23$). We thus demonstrated that differences in genetic variants converged and impaired specific molecular mechanisms in non-overlapping group of patients.

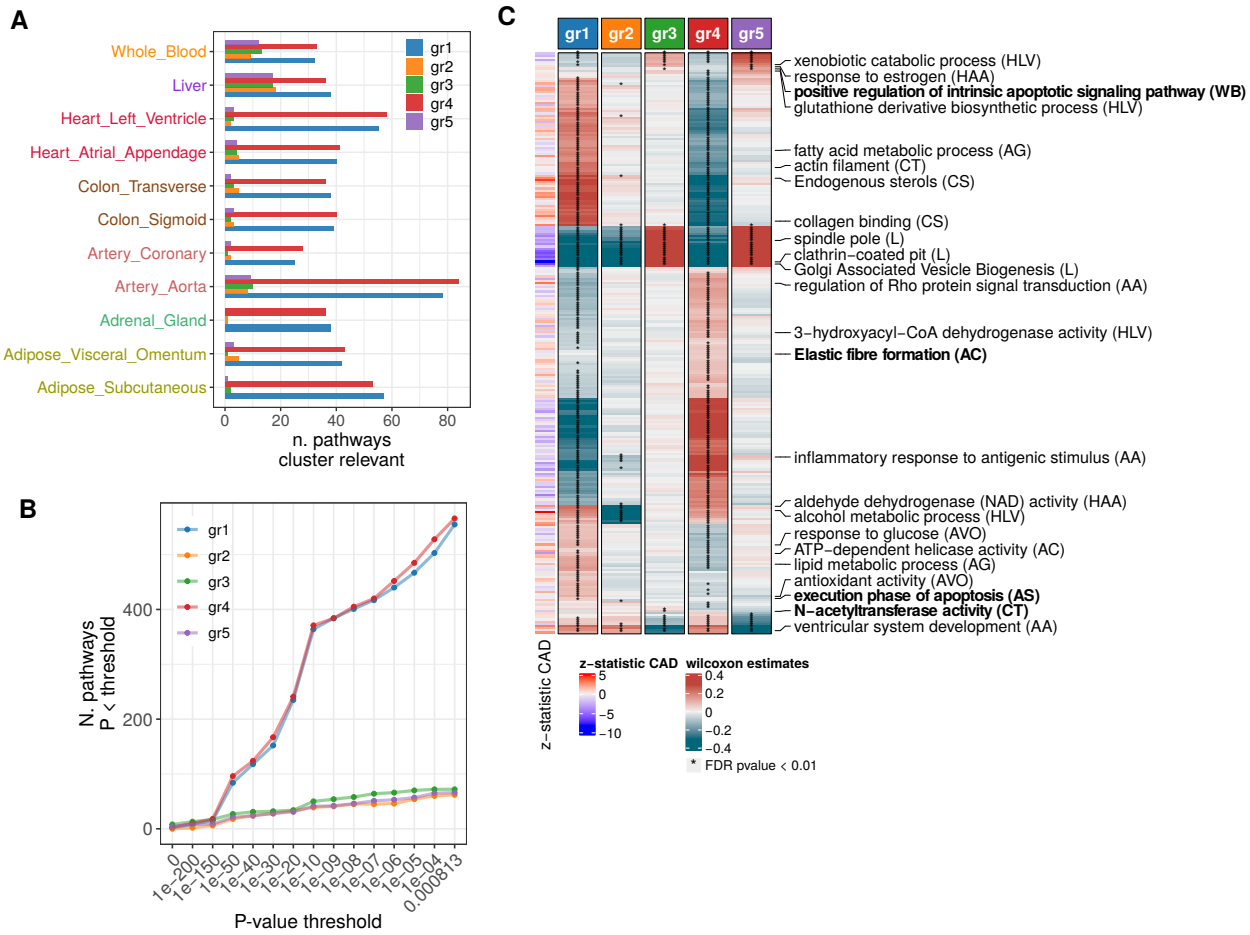

Fig. 4.32.: (Adapted from Trastulla et al., in prep.) **(A)** Number of significant pathways (tissue specific FDR ≤ 0.01) associated with each group from WMW test of group id versus remaining patients. The included pathways are both from Reactome and GO and filtered such that Jaccard Similarity ≤ 0.2, retaining the pathways with highest coverage and removing significant pathways having discordant WMW estimates across tissues. **(B)** For each group, number of significant pathways (y-axis) passing the WMW p-value threshold (x-axis). **(C)** WMW estimates (capped) for 271 significant pathways (rows) in each group versus the rest test (column), considering only the most significant tissue per pathways when repeated. The names on the row are a selection of significant pathways, parenthesis refers to the tissue considered (acronyms indicates the initial of the tissue name). Row annotation on the left refers to the corresponding CAD Z-statistics from PALAS. Names in bold indicate that the pathway reaches a higher significant than any of the genes in it, for at least one group.

Subsequently, we investigated whether the observed differences in genetic liability distribution did actually have an impact on measured endophenotypes and hence had a connection to different pathomechanisms in CAD patients. To test this hypothesis, we leveraged again the deep phenotypying available for UKBB data set and tested 637

collected endophenotypes across 19 categories (Tab. B.6) as well as 33 hypothesis-driven endophenotypes that correspond to clinical phenotypes (Tab. B.7) using the original value not processed by PHESANT tool [203] (see section 4.1.3). The cluster-specific differences in endophenotypes were detected via GLM testing gr_i versus all remaining affected individuals and correcting for PCs, age, sex as well as medications depending on the phenotype class (see Tab. B.6 for details) as described in section 3.3.3. We thus identified 22 significant cluster-specific endophenotype associations (FDR ≤ 0.05) in a total of 14 phenotypes (Fig. 4.33A).

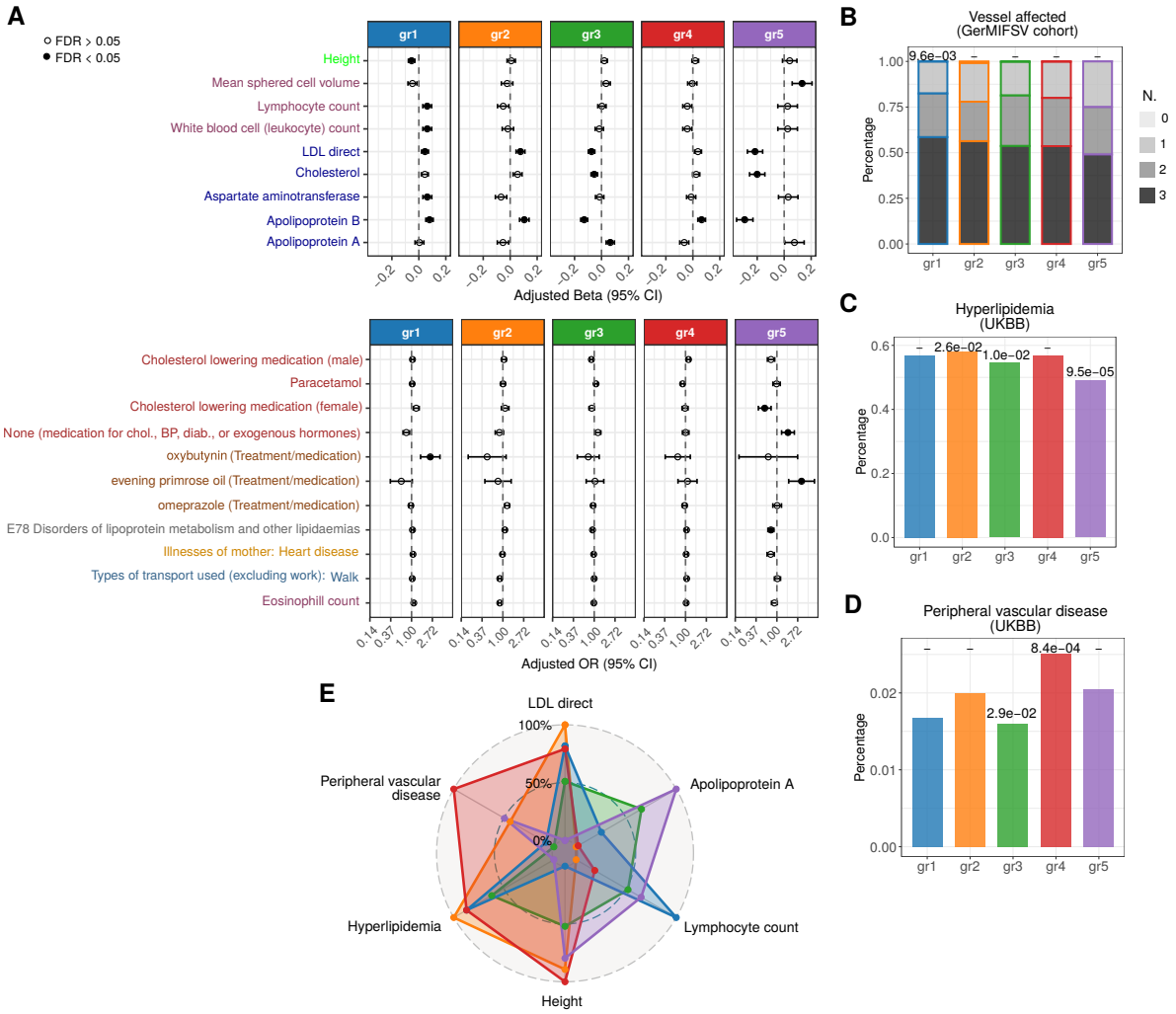

Fig. 4.33.: (Adapted from Trastulla et al., in prep.) **(A)** Among 637 tested endophenotype from UKBB, forest plot shows significantly different ones (FDR ≤ 0.05 or p-value ≤ 0.001) in at least one group (gr_i versus remaining patients) using Generalized Linear Model (GLM), indicating regression coefficient (β_{GLM}) with 95% Confidence Interval (CI). Full dot means that β_{GLM} is significant after BH correction performed separately for each group across all the endophenotype. **(B)** CAD severeness across projected clusters in German5 based on percentage of patients with a certain number of vessel affected. **(C-D)** Percentage of patients with certain comorbidities/severeness in UKBB clustering, nominal p-values from group-wise GLM shown on top, - means nominal p-value > 0.05. **(E)** Mean value of selected group-specific endophenotype in each group rescaled to 0-100 range.

In accordance with pathway score and gene T-score distributions, CAD patients in gr_3 and

gr_5 showed an overall decrease in LDL levels ($\beta_{GLM} = -0.07, -0.22$ and P= 2.13e-09, 6.33e-14 respectively), together with cholesterol and apolipoprotein B. This tendency was in general stronger for gr_5 and it was also reflected in a significant lower assumption of CAD related medications ($OR_{GLM} = 1.70$, P= 0.0006) and comorbidity of lipoprotein metabolism disorders ($OR_{GLM} = 0.75$, P= 9.45e-05), reflecting a healthier status. Note that, evening primrose oil medication can have an impact in reducing the observed LDL values in gr_5, but the significant increase in odds ratio was based on a total of 98 individuals assuming that medication, of which 12 in gr_5, hence not sufficient to explain the observed overall decrease. On the other hand, LDL values were significantly increased for gr_1 and gr_2, with stronger effects in the latter ($\beta_{GLM} = 0.046, 0.076$ and P= 4.99e-04, 5.19e-06 respectively). This was consistent with the decrease in alcohol metabolic process and its inverse relationship to LDL explained before. Finally, gr_1 was characterized by an increase in inflammation related phenotypes such as leukocyte ($\beta_{GLM} = 0.062$, P= 2e-04) and lymphocyte counts ($\beta_{GLM} = 0.062$, P= 2e-04), hence connected to impaired immune related pathways. Interestingly, gr_1 was also characterized by a shorter height tendency ($\beta_{GLM} = -0.054$, P= 1.37e-05) proven to be inversely related to CAD risk increase [222]. This together with LDL increase and inflammation phenotype highlighted a higher phenotype severity in gr_1. We then specifically focused on 33 clinically established CAD connected features among the detected groups (Tab. B.7) as well as available CAD clinical phenotyping on GerMIFSV cohort from CARDIoGRAM. In particular, the clustering structure on GerMIFSV cohort was projected from the UKBB model as previously described (Fig. 4.31). We thus found that gr_1 exhibited an increase in term of number of vessel affected (Fig. 4.33B) and thus confirming a severity also detectable from projecting clustering structure into an external cohort. In the hypothesis-driven endophenotype analysis from UKBB, we focused on results nominal at $P < 0.01$ and considered as reliable those results passing a permutation p-value threshold of 0.1 among 50 random clustering repetition (see section 4.3.9), identifying two cluster-specific comorbidities: hyperlipidemia (Fig. 4.33C) and pheripheral vascular disease (Fig. 4.33D). The former (hyperlipidemia) was significantly less frequent in gr_5 ($OR_{GLM} = 0.75$, P= 9.5e-05), showing evidence of decrease also in gr_3 ($OR_{GLM} = 0.92$, P= 1e-02) and an opposite effect in gr_2 ($OR_{GLM} = 1.1$, P= 2.6e-02), concordant with the previously observed LDL differences. The latter (pheripheral vascular disease) was instead indicating a significant co-occurrence in gr_4 ($OR_{GLM} = 1.46$, P= 8.36e-04), pointing toward a possible pathomechanisms for that subset of samples. Overall, these results suggested that gr_1 and gr_5 represent two extreme in term of CAD severity from most to least severe in term of endophenotype manifestation (summary in Fig. 4.33E).

Lastly, we took advantage of the medication information available from UKBB deep phenotyping and investigated whether the response to a certain treatment was variable across the groups (see section 3.3.3). In this case, we considered 87 endophenotypes included in 8 macro categories as response (namely Arterial stiffness, Blood biochemistry, Blood count, Blood count ratio, Blood pressure, Body size measures, Hand grip strength and Impedance measures) and observed the group-specific variability comparing patients

with and without prescription of 17 medications, accounted simultaneously (Aspirin, Blood pressure medication, Calcium, Cholesterol lowering medication, Folic acid or Folate (Vit B9), Glucosamine, Ibuprofen (e.g. Nurofen), Insulin, Iron, Paracetamol, Selenium, Vitamin A, Vitamin B, Vitamin C, Vitamin D, Vitamin E, Zinc). First of all, we focused on LDL response to cholesterol lowering medication as a well-known mechanism and observed that indeed CAD patients assuming these medications exhibited a great decrease in LDL values across all groups (Fig. 4.34A-B). Although, only significant at the nominal level, we observed that the reduction in gr_5 was less pronounced, especially when compared to gr_4 or gr_2 (Z-test comparison P$= 0.015, 0.027$ respectively, Fig. 4.34A). This finding was in accordance with results from pharmacogenomic studies that connected minor allele T of rs646776 in SORT1 locus to a higher reduction of LDL induced by statin medications [223]. Indeed, the higher expression of SORT1 characteristic of gr_5 was connected to decreases in T allele dosages of rs646776 via PriLer gene expression model (PriLer $\beta = -0.25$) and hence explaining the lower LDL reduction after statin usage. Note that the lower reduction of gr_5 was compensated by a general decrease of LDL values regardless the medication assumption and that anyway stabilized to similar values as the other CAD patient groups (Fig. 4.34B).

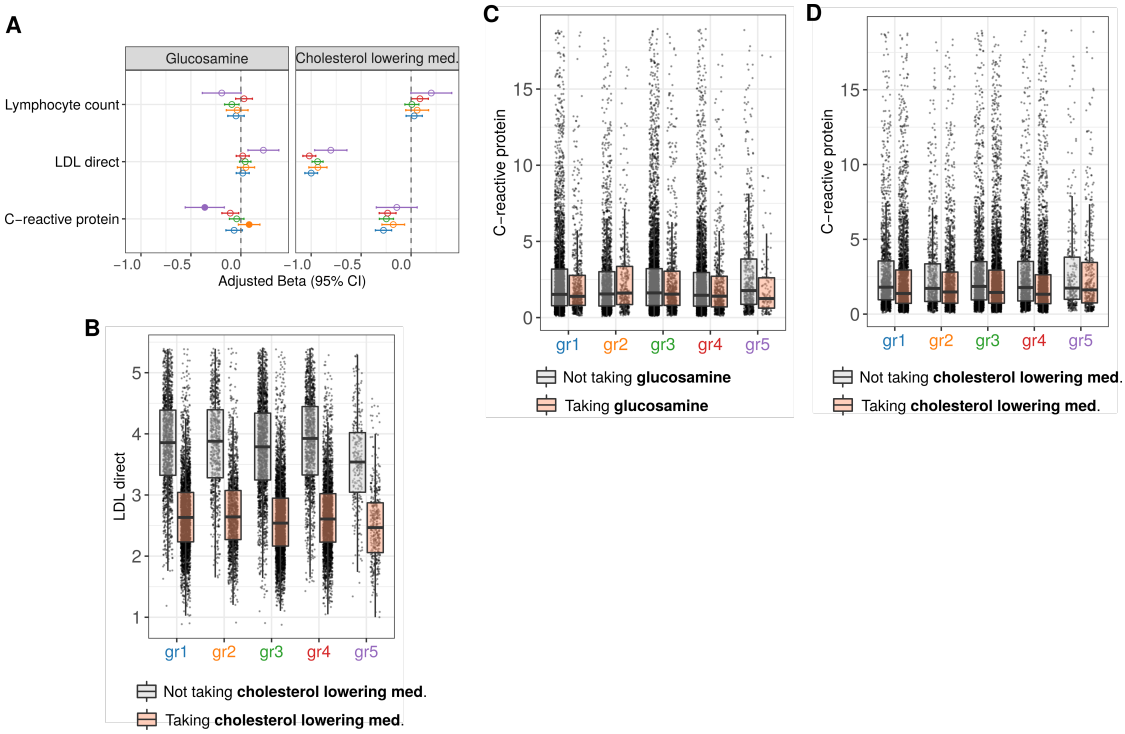

Fig. 4.34.: (Adapted from Trastulla et al., in prep.) **(A)** Treatment response showing the effect of glucosamine (right) and cholesterol-lowering medications (left) in each group for selected phenotypes. X-axis shows regression coefficient with 95% CI from GLM in each group, full dots indicate groups that are significantly different in a pairwise comparison after BH correction (pairwise comparison-specific and treatment-specific), tested using Z-test for comparing regression coefficients. **(B-D)** Treatment response of a medication on a endophenotype in each group. On the right of each panel, it is shown the distribution of original endophenotype values in each group when taking or not the medication, y-axis is cropped at 98 percentile values. **(B)** Cholesterol lowering medications effects on LDL direct, **(C)** glucosamine supplements on C-reactive protein, **(C)** cholesterol lowering medications effects on C-reactive protein.

We then took into consideration the most significant group-specific treatment response i.e. C-reactive protein changes due to glucosamine assumption (Fig. 4.34A,C), with CAD patients in gr_5 showing a significant decrease after assumption, especially when compared to gr_2 (Z-test comparison P= 9.68e-05). In particular, our strategy detected a sub-population of CAD patients with an estimated reduction in CRP levels after glucosamine assumption of 30% ($1 - e^{-\beta_{GLM}}$), higher than what is observed in other groups and even more than what was previously estimated in regular users (17%, [224]). Interestingly, the well-known CRP reduction effect from statin [225] was observed in all groups but gr_5 (Fig. 4.34A,D), thus suggesting a tailored precision medicine strategy to reduce inflammation and CAD predisposition.
In conclusion, CASTom-iGEx patient stratification methodology detected distinct patient groups exhibiting different genetic liabilities that then translated into divergences in clinical parameters as well as medication responses.

Before moving forward to CASTom-iGEx application to schizophrenia, we discuss in detail the ancestry contribution in clustering definition (section 4.3.7), the usefulness of TWAS scaling for phenotype Z-statistic (section 4.3.8) and the calibration of p-values and control for false positives in detecting cluster-specific genes and pathways as well as endophenotypes.

4.3.7 Ancestry contribution to clustering

Because gene T-scores were directly computed from genotype dosages, we investigated whether the clustering structure was emerging from ancestry differences. In our pre-processing strategy prior to grouping patients, we specifically corrected each gene for PCs via linear regression. This correction greatly reduced the association between the final clustering and PCs as well as assessment centre, which is related to the individuals population of origin (Fig. 4.35). Indeed, we observed PCs associations passing from highly to barely significant nominal p-values in almost all tissues via Kruskal-Wallis and χ^2 test. Nevertheless, PCs 4 and 5 still showed a nominal level of association with the patient stratification on adipose subcutaneous, artery coronary and liver tissues.
Thus, we further investigated the ancestry contribution to the tissue grouping comparing the detected liver clustering with the clustering output from the available PCs in UKBB, i.e. from 1 to 40 (Fig. 4.36). In this case, we simply standardized each PCs to have mean 0 and standard deviation 1 and applied the PhenoGraph algorithm based on shared nearest neighbours as described in 3.3.1. We thus detected 7 different groups (Fig. 4.36A). When compared to the clustering obtained from tissue-specific imputed gene expression, as expected we found a minimal significance, nevertheless with p-value > 0.05 for the adipose tissues, liver and whole blood (Fig. 4.36B, left).

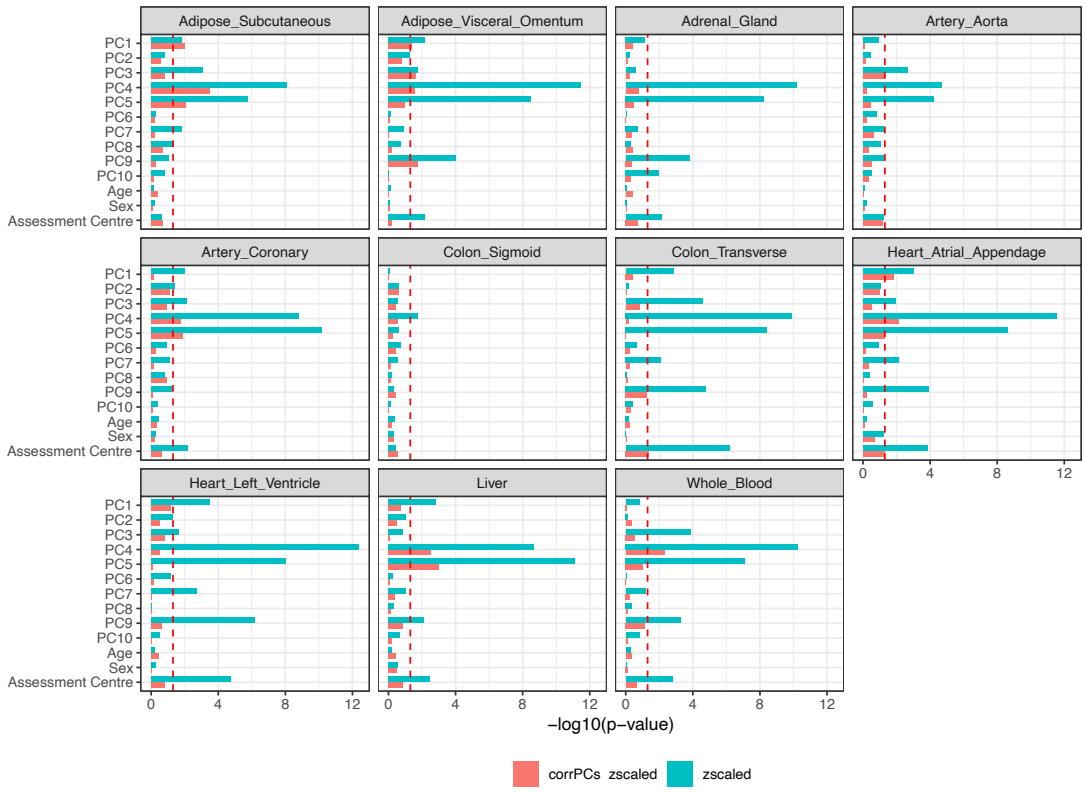

Fig. 4.35.: Differences in confounders association between cluster with correction for PCs preprocessing and TWAS-rescaled (corrPCs zscaled) and solely TWAS-rescaled (zscaled) across all tissues. X-axis shows -$\log_{10}$(p-value) from Kruskal-Wallis test (PCs from 1 to 10 and Age) and χ^2 test (Sex and Assessment Centre). The red dashed line in each plot correspond to nominal p-value $= 0.05$.

Indeed, we estimated the overall extent of overlap in clustering structure via Normalized Mutual Information (NMI) (Fig. 4.36B, right), being consistently lower than 0.00071, although reaching the highest value in liver. We then created $10,000$ random partitions of samples, keeping the same number of patient in each group as liver, to compare the NMI between PCs and the actual liver clustering to the NMI between PCs and the random assignments. We actually observed that the PCs-liver NMI was higher than random clustering-PC NMI in 97.2% of the simulations (Fig. 4.36C), hence sharing some minimal structure that would not emerge by chance. In order to understand which groups from the two configuration would share a number of individuals higher than by chance, we then computed for each pair of group in liver and PCs clustering the odds ratio from Fisher's Exact test (Fig. 4.36D), building a contingency table of gr_i and not gr_i in liver and gr_j and not gr_j in PCs for each combination of $i = 1, \ldots, 5$ and $j = 1, \ldots, 7$). A significant enrichment was exclusively detected between gr_1 in liver and gr_7 in PCs (OR$= 1.12$, P$= 0.0006$), possibly due to a high number of individuals coming from Reading and Birmingham surroundings (Fig. 4.36E). Finally, we investigated whether this reduced overlap and ancestry contribution to liver clustering was affecting the observed endophenotype differences. In this case, we considered for each endophenotype tested for a group-specific effect the best p-value result across all liver groups and all PCs groups (Fig. 4.36F). To perform cluster-specific endophenotype analysis in PCs

clustering, we used the GLM approach as described in section 3.3.3 but only corrected for age and sex as covariates. Importantly, the clustering structures in liver and PCs led to different endophenotypes significance, showing no correlation between -$\log_{10}$(p-value), with "Places of birth in UK" the most significant endophenotypes in PCs clustering that were not associated with liver structure. Nevertheless, 3 endophenotypes passed FDR significance 0.05 in both configurations: height, comorbidity with lipidaemias and aspartate aminotransfarase. Hence, we examined whether these common associations were due to the mildly overlapping gr_1 in liver and gr_7 in PCs, observing which groups were associated with those endophenotype differences (Fig. 4.36G). Comorbidity with lipidaemias had opposite effects in two non-overlapping groups (gr_5 in liver and gr_2 in PCs), height was different both in gr_1 liver and gr_7 in PCs but with opposite effect (thus discordant with respect to the enrichment) and aspartate aminotransfarase was decreased in gr_1 liver and increased in gr_3 PCs, hence the only result concordant with observed individuals depletion (Fig. 4.36D, OR$= 0.86$ and P$= 0.0002$). We finally concluded that there was a minimal impact from PCs on the tissue-specific clustering structure that could not be removed even after correcting for PCs. This ancestry contribution practically translated into one group in liver clustering having an overlaps with a group from PCs clustering more than what would be observed by chance. However, the endophenotype differences were mostly not affected by any ancestry background and we can conclude that what we observed in the actual clustering was due to a genetic liability and not emerging from European ancestry differences.

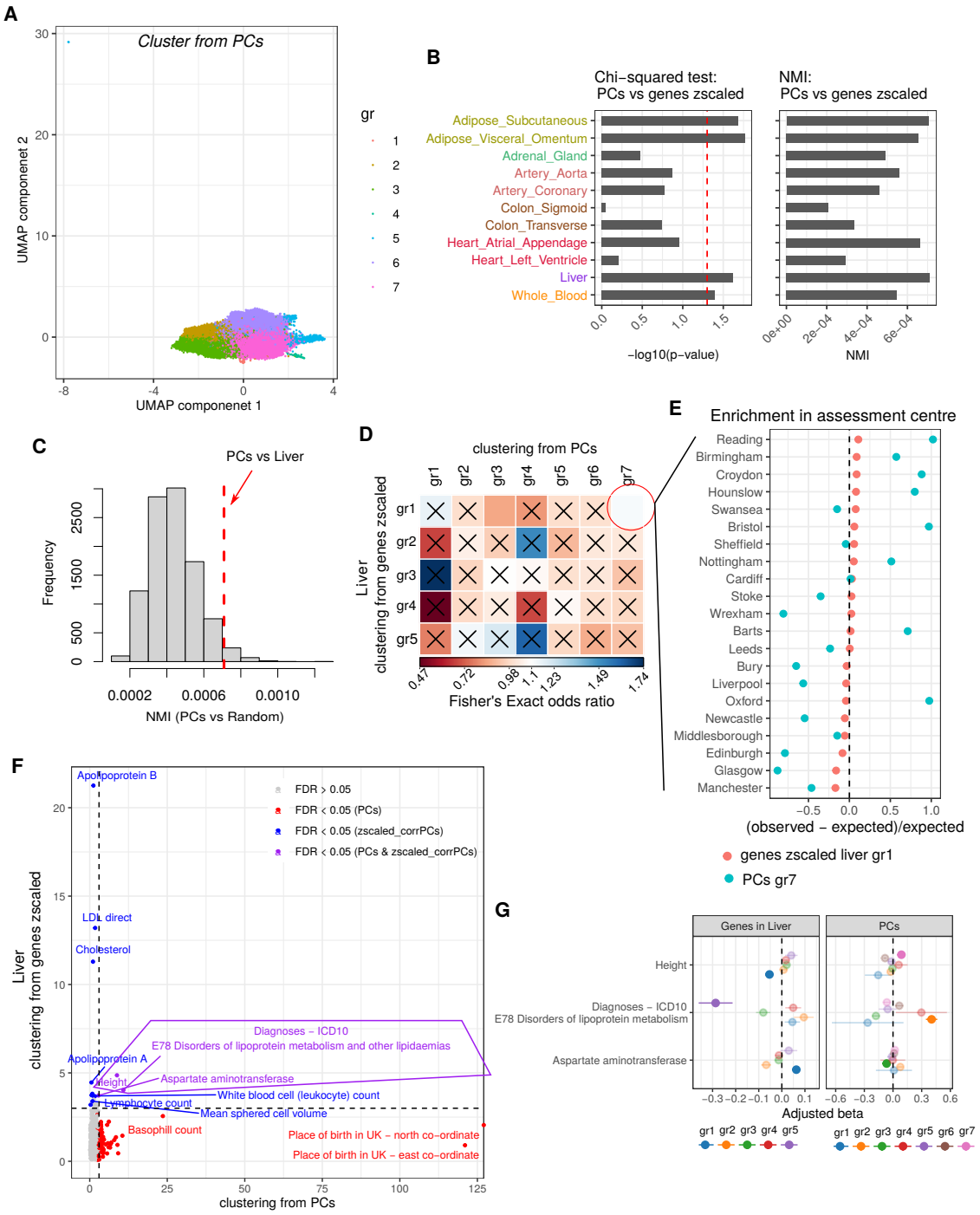

Fig. 4.36.: (Adapted from Trastulla et al., in prep.) **(A)** UMAP of CAD cases based on the first 40 UKBB PCs (standardized), color refers to the assigned PCs clustering. **(B)** Comparison between PCs clustering and grouping from gene T-scores corrected for PCs, standardized and TWAS-rescaled ("genes zscaled") in term of -$\log_{10}$ p-value of χ^2-test (left, dashed line refers to nominal p-value 0.05) and NMI (right), for each tissue. **(C)** Histogram of NMI between cluster from PCs and $10,000$ randomly assigned groups with the same size as liver clustering, the dashed line refer to the NMI comparing PCs and the actual liver clustering. **(D)** Pairwise Fisher's Exact test between a group detected in PCs clustering (columns) and a group detected in liver clustering (rows), heatmap indicates the computed odds ratio with $\times$ highlighting a non-significance at the nominal level of 0.01. **(E)** Investigation of enrichment in assessment centre for gr_1 in liver and gr_7 in PCs clustering. Considering the centre assignment versus a group assignment (gr_i or not gr_i), x-axis indicates the fraction of (observed - expected)/expected counts as computed from the χ^2 statistic across the centres (y-axis). **(F)** Each dot represent a tested endophenotype and indicates the -$\log_{10}$ p-value of the most significant group-specific difference in PCs (x-axis) and liver (y-axis) clustering. Dashed lines refer to p-value $= 0.001$ and color reflects the FDR significance threshold. **(G)** Forest plot of group-specific differences for the 3 endophenotypes significant in both PCs and liver cluster, x-axis shows the regression coefficient from GLM testing gr_i vs all remaining. Not shaded dots indicate groups with most significant association in term of p-value.

4.3.8 Comparison genes TWAS-rescaling and non-scaling strategies

Our clustering strategy in based on a feature pre-processing that 1) standardize each gene separately to mean 0 and standard deviation 1, 2) correct for first 10 PCs via linear regression and 3) finally multiplies each gene by the Z-statistic of the phenotype of interest (here CAD) in the considered tissue obtained via TWAS. Here, we compared this strategy (from now on called "zscaled") with the same two steps but without multiplying per CAD Z-statistic (called "original"). First of all, we observed that the concordance in clustering structure between these two pre-processing procedures varied across tissues, with a reduced normalized mutual information (NMI) but still indicative of a minimal level of overlap, especially for adrenal gland and heart atrial appendage (Fig. 4.37A). More importantly, the number of detected groups per tissue greatly varied between zscaled and original configuration, going from between 3 to 5 in zscaled to always higher than 10 in original (Fig. 4.37B, left). In addition, the cluster modularity from Louvain method was always increased for zscaled (Fig. 4.37B, right), indicative of a better defined structure and connectivity inside a group when specifically pointing towards phenotype relevant features.

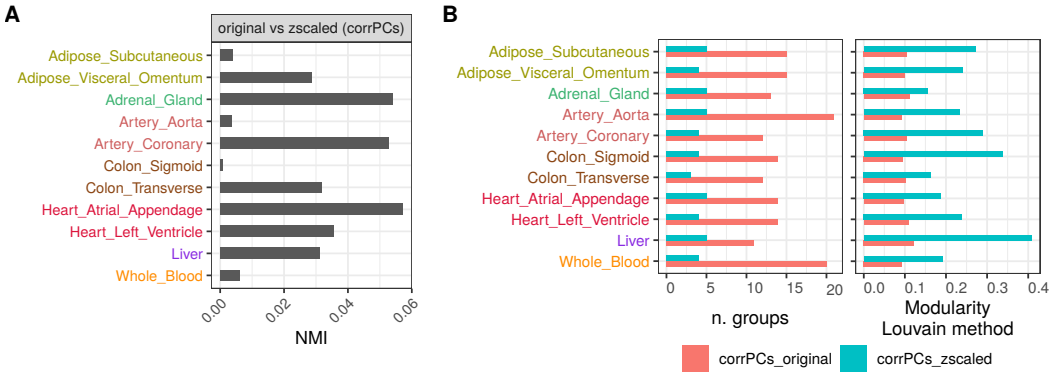

Fig. 4.37.: Comparison between clustering performed from gene T-scores corrected for PCs and normalized gene-wise ("original") and corrected for PCs, normalized gene-wise and multiplied per Z-statistic association from CAD TWAS ("zscaled") **(A)** NMI between the two versions for each tissue-specific clustering. **(B)** Comparison in terms of number of detected groups (left) and modularity score from Louvain clustering (right) for each tissue-specific clustering

4.3.9 Endophenotypes and features association with random clustering

Lastly, we explored the p-value calibration on cluster-specific genes and pathways with the null hypothesis of no cluster structure obtained from imputed gene expression as well as the endophenotype association when individuals are partitioned at random. To achieve that, we randomly partitioned CAD patients 50 times, keeping the same number of individuals in each cluster as obtained from the actual liver clustering. The random

partitions and the actual clustering were independent (Fig. 4.38A), with the only exception of repetition 2 showing a nominal significance (P= 0.0063) but not passing FDR correction among the 50 repetitions (FDR= 0.32). We then considered only the first 10 random repetitions due to computational time and resources issues and tested group-specific differences in genes across all tissues (section 3.3.2).

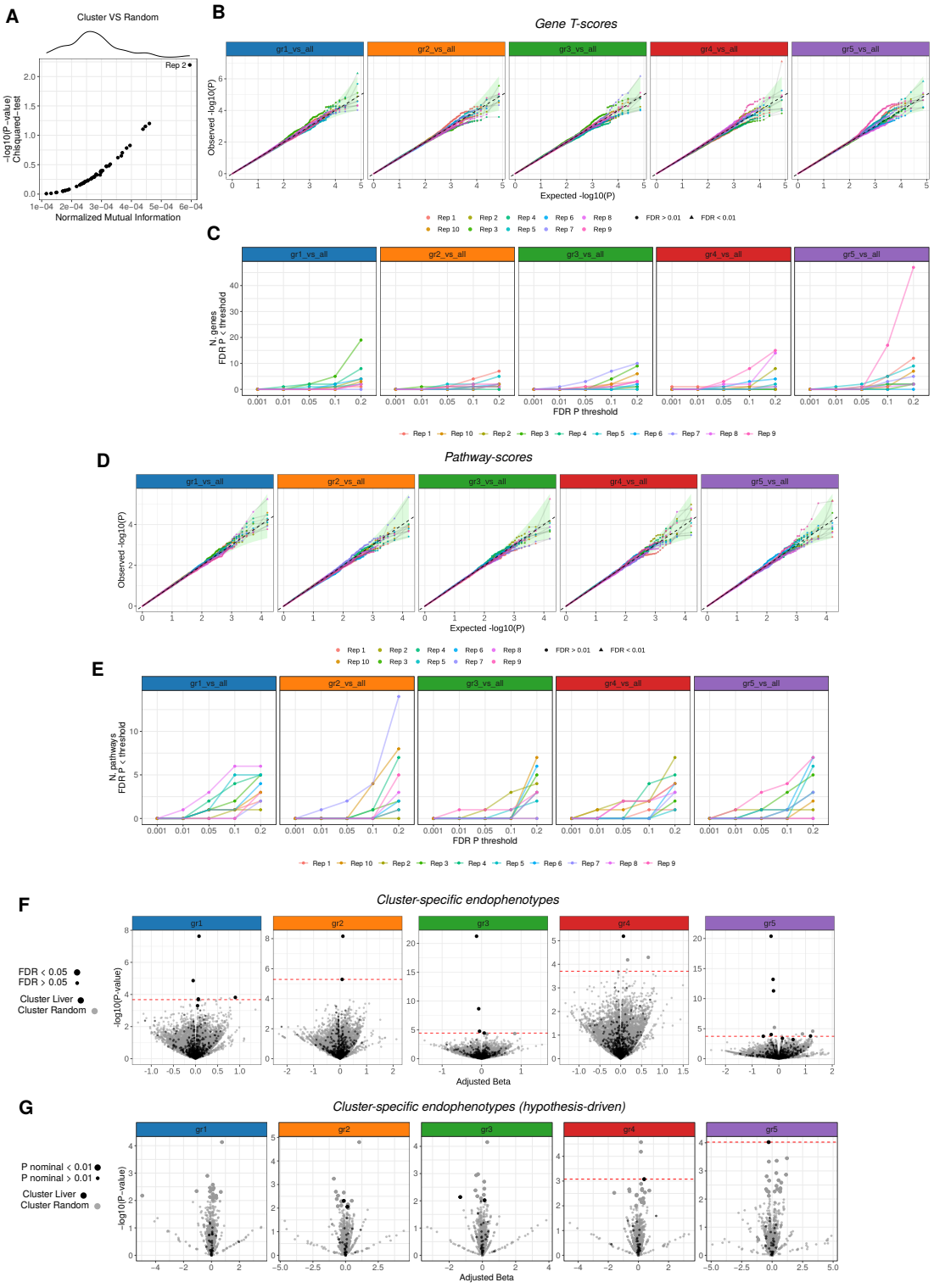

Fig. 4.38.: (Adapted from Trastulla et al., in prep.) Random partition repeated 50 times of CAD patients following the same structure of liver grouping. **(A)** X-axis shows NMI and y-axis shows -$\log_{10}$ p-values from χ^2 statistics between actual liver clustering and random partitions. **(B)** Quantile-quantile plot for group-wise specific (gr_i vs all remaining) testing gene T-scores association via WMW across all tissues. The expected p-values are from uniform distribution and the dashed line indicates the diagonal and shaded green area representing 95% confidence interval from beta distribution. Each line refers to one of the 10 simulations. **(C)** For each of the 10 random clustering, number of significant association passing FDR threshold (y-axis) at varying FDR levels (x-axis), with correction performed tissue-wise and group-wise. **(D)** Same as (B) but testing pathway-scores differences for the selected gene-sets (JS ≤ 0.2). **(E)** Same as (C) but testing pathway-scores differences. **(F)** Volcano plot of cluster-specific endophenotype differences, x-axis shows β regression coefficient from GLM referring to gr_i vs remaining cases features and y-axis shows corresponding -$\log_{10}$ p-value. Each grey dot is a tested endophenotype among 637 UKBB phenotypes for a random clustering configuration out of 50 repetitions, each black dot refers instead to the endophenotype testing on the actual liver clustering, in both cases the size indicates the significance after correction (FDR ≤ 0.05). The dashed horizontal line indicates 0.1 threshold for p-value permutation correction considering the 10 random partitions, for each group separately. **(G)** Same as (F) but considering the 33 raw hypothesis-driven endophenotypes and size indicates to nominal p-value ≤ 0.01

The QQ-plot of WMW p-value distributions did not deviate from the expected one, i.e. uniform in $[0, 1]$ range (Fig. 4.38B), with the exception of some repetitions such as Rep9 in gr_5 showing an inflated distribution. When counting the number of group-specific genes passing a certain FDR threshold, a 0.01 FDR upper bound was identifying only 1 gene significant in 1 out of 10 repetitions (Fig. 4.38C) and hence decided to use this stricter threshold in defining group-specific genes, instead of the otherwise used 0.05. Similarly, we computed WMW test to detect group-specific pathways (pruned at Jaccard Similarity ≤ 0.2) for the 10 random clustering repetitions among all tissues. P-value distribution for pathways was again following a uniform distribution (Fig. 4.38D), with less inflation than what was observed for genes, probably due to reduction of highly correlated pathways via clumping. As before, we decided to fix at 0.01 the FDR threshold to call for group-specific pathways (Fig. 4.38E), hence limiting the false discovery to 1 pathway in each group for 1 repetition (from gr_1 to gr_3) or for 2 repetitions (gr_4 and gr_5).
Finally, we considered the 50 random clustering and searched for endophenotype differences in each cluster via GLM as described in section 3.3.3 for 637 UKBB phenotypes (Tab. B.6, Fig. 4.38F) and 33 hypothesis-driven ones (Tab. B.7, Fig. 4.38G), comparing the effect size (β_{GLM}) and corresponding p-value from the random clustering configurations with the actual liver clustering. Highly significant results were only reached for the actual liver clustering, nevertheless 7 out $158,324$ tests in the random clustering pass FDR threshold of 0.05 (Fig. 4.38F). Having now the possibility to compute a permutation p-value from the 50 random repetitions (dashed line in Fig. 4.38F), we observed that solely 5 out of the 22 associations did not pass a permutation correction of $P \leq 0.1$ but were significant from FDR ≤ 0.05, among which LDL increase in gr_1, increased mean cell sphered volume in gr_5 and higher percentage of people not taking any medication CAD related in gr_5. Instead, in the hypothesis-driven analysis, we could not rely on a nominal p-value ≤ 0.01 as significant results due to the high number of associations that would have been called significant with that threshold in the random partitions (Fig. 4.38G). Thus, we considered

as reliable significant associations the 2 results passing permutation p-value ≤ 0.1, that are actually shown in Fig. 4.33C-D. We concluded that, our gene and pathway associations analysis was well calibrated, however we still applied a stricter FDR cut-off of 0.01 to reduce false positive results. In addition, the endophenotypes associations observed in the actual liver clustering greatly exceeded the significance from the random partitions in the general endophenotype analysis. Finally, we only regarded as significant results those passing p-value correction via permutation of 0.1 in the hypothesis-driven endophenotypes to reduce false positive results.

4.4 Schizophrenia

In this section, we focus on SCZ and assess genetically derived features associated with this complex disease as well as patient stratification obtained from CASTom-iGEx application and the corresponding credible endophenotype characterization. Conversely from CAD, there is limited knowledge on possible pathomechanisms and genetic trajectory that contribute to disease etiology and possible differentiate endophenotypes among patient groups.
The application of CASTom-iGEx pipeline included an initial build of PriLer gene expression models on GTEx for 9 SCZ related tissues (8 brain tissue regions: caudate basal ganglia, cerebellar hemisphere, cerebellum, cortex, frontal cortex BA9, hippocampus, hypothalamus, nucleus accumbens basal ganglia, and 1 immune related tissue: cell EBV transformed lymphocytes) and on DLPC from CMC after variant matching and harmonization with 36 European PGC cohorts (wave 2) for a total of $24,764$ cases and $30,655$ controls (Fig. 4.1, Tab. 4.2). We additionally considered CMC data set as replication cohort specifically for DLPC tissue as it included 212 controls and 266 patients diagnosed with SCZ and schizoaffective disorders.

4.4.1 Associated genes and pathways

Because PGC data set is compose of multiple cohorts, TWAS and PALAS were performed separately for each cohort and overall associations were obtained via meta-analysis (see section 3.2.2). For each cohort, we included as covariates in the logistic regression PCs from 1 to 7, PC9, PC15 and PC18. This PCs choice was applied following the corresponding GWAS publication [141] to account for the PCs referring to the highest variability and additional ones associated with the disease status. TWAS results in form of genome-wide Z-statistics can be observed in Fig. 4.39A. We identified $1,274$ significant genes across 10 tissues at the tissue-specific FDR threshold of 0.05, corresponding to 768 unique genes, 655 of which located outside the MHC locus. Similarly to CAD, SCZ associated genes where mostly detected in a single tissue (Fig. 4.39B), with the highest number of significant genes identified in DLPC (Fig. 4.39C), related to the number of reliable genes tested and

sample training size.

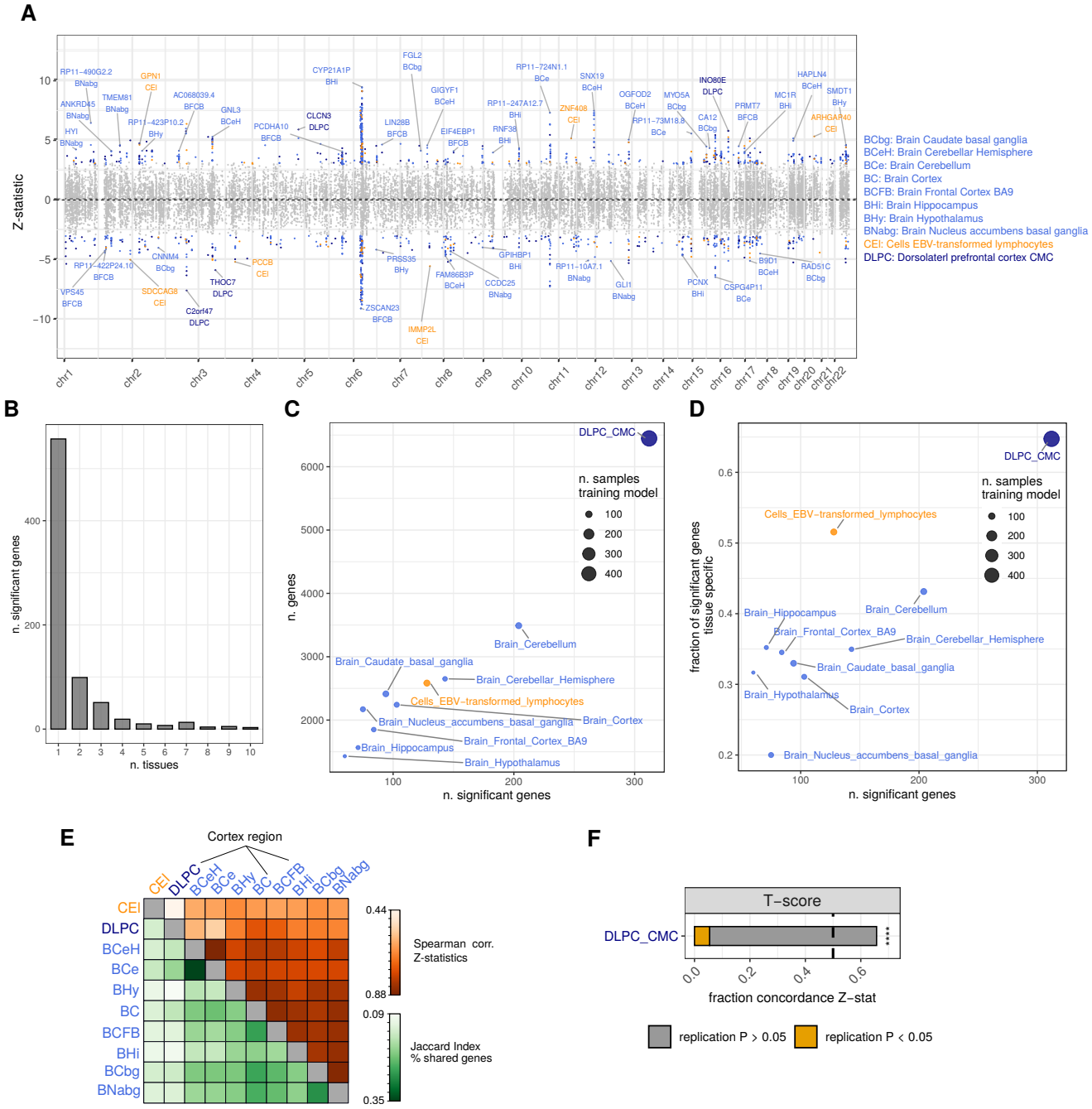

Fig. 4.39.: (Adapted from Trastulla et al., in prep.) **(A)** Manhattan plot showing Z-statistic across 10 tissues, colored dots indicate genes significant at tissue-specific FDR level of 0.05. **(B)** Number of significant genes detected in one or more tissues. **(C,D)** Number of SCZ significant genes versus **(C)** number of tested reliable genes (predicted in PriLer) or **(D)** fraction of significant genes uniquely detected in a tissue, dot size refers to the number of PriLer training sample in GTEx reference panel. **(E)** Lower-triangular (green): percentage of imputed genes that are in common between 2 tissues (Jaccard index), upper-triangular (orange): Spearman correlation of CAD Z-statistics among shared genes. **(F)** Reproducibility of gene levels T-scores with discovery PGC samples and replication from CMC data set. X-axis shows the fraction of significant genes in PGC that have the same effect sign (Z-stat) in CMC cohort, p-values are computed from one-sided sign test ($**** = P \leq 0.0001$). The bar in yellow represents the fraction of genes concordant that are also significant at the nominal p-value threshold of 0.05

Indeed, there was a clear correlation between number of samples in reference panel and number of SCZ associated genes (Spearman corr.$= 0.78$). In addition, DLPC together with transformed lymphocytes exhibited the highest percentage of tissue-specific genes ($> 50\%$, Fig. 4.39D), namely only associated with SCZ in that tissue, reflecting immune related expression differences with respect to brain regions. Considering Spearman correlation of gene Z-statistics (Fig. 4.39E red heatmap), we observed a high concordance for brain related tissues in GTEx (cor. > 0.55), with the highest concordance observed for two replicate regions, cerebellum and cerebellar hemisphere (cor. $= 0.88$), also reflected in the largest shared subset of genes (Jaccard index $= 0.35$, Fig. 4.39E green heatmap). On the other hand, DLPC was less correlated with the other brain regions, possibly due to a different reference panel. However, the tissue signature was preserved as observed from its highest correlations with the two GTEx cortex region (cortex and frontal cortex BA9, cor$= 0.75$ and $= 0.74$ respectively). Notably, when comparing our results with previous TWAS for SCZ [71, 153] we found 102 genes among our 655 SCZ related ones that were also present in at least one TWAS previous output, out of 268 detected. Furthermore, we tested the replicability of our result considering CMC reference panel that included SCZ affected individuals and controls, imputing gene expression from the model trained on the same cohort and applying our TWAS methodology. In DLPC tissue, we found that 66% of the significant genes out of 312 showed the same Z-statistic sign when tested in CMC data set (one-sided sign test P$=$ 1.5e-08, Fig. 4.39F). Nevertheless, only 5.4% of those 312 associated genes where also significant in CMC at the nominal level, reflecting the critical difference in terms of sample size for discovery and replication data sets (only 478 individuals in the latter).

Many of the detected associations were in loci already identified from GWAS, considering as term of comparison summary statistics from PGC wave 2 [141] that mostly overlaps with the data set we leveraged for this analysis. To assess the extent of novel results obtained via CASTom-iGEx, we first combined TWAS summary statistics across all tissues in loci recursively merging genes with $[\text{TSS} - 200\text{kb}, \text{TSS} + 200\text{kb}]$ window distant less than 1M (similarly to CAD), reaching a total of 242 loci. We then adjusted GWAS summary statistics with BH procedure to be consistent with our methodology and identified 32 genes (30 without repetition) in 24 loci that did not intersect a GWAS significant variant (Tab. 4.5).

Generally, highly significant genes intersected well known SCZ-associated loci from GWAS such as MHC locus, SNX19 and C2orf47 regions (Fig. 4.39A). An example of putative gene with evidence of phenotypic impairment is indeed complement C4A in MHC locus, for which increased expression was associated with an increased SCZ risk ($4 \leq$ Z-stat ≤ 9 in the 7 tissues for which is reliable, namely all but frontal cortex B49, hypothalamus and nucleus accumbens basal ganglia). In particular, from in-vivo experiments C4A was found to reduced cortical synapse density and to alter mouse behaviour and functions such as social behaviour, spatial working memory and anxiety phenotypes, that resembled

SCZ negative symptoms [150, 151]. When observing the PriLer regression coefficient regulation for C4A in DLPC (Fig. 4.40A) we can particularly appreciate the complexity of SNP configuration in predicting C4A gene expression that involved 98 SNPs. The top 2 regulatory variants did not correspond to the most significant GWAS associations in the gene cis-window, and only one of the 2 was overlapping all gene regulatory elements used as priors (rs116026314).

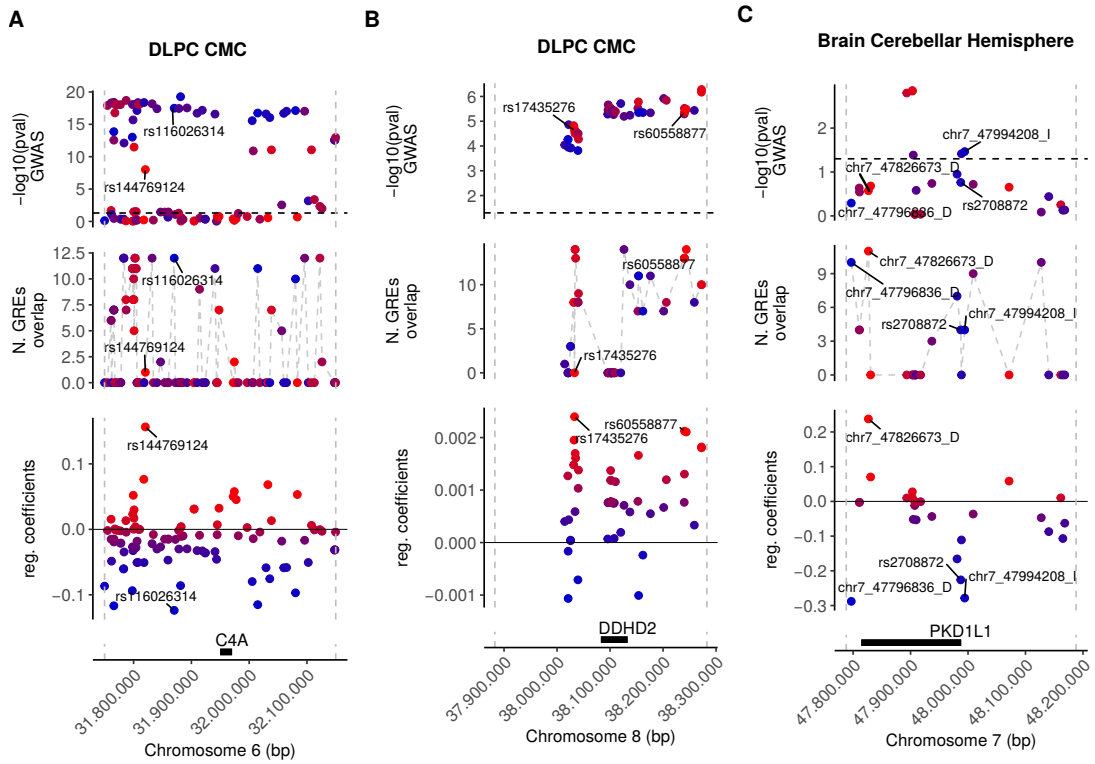

Fig. 4.40.: PriLer models for **(A)** C4A in DLPC , **(B)** DDHD2 in DLPC, and **(C)** PKD1L1 in cerebellum hemisphere, with each dot representing a variant with PriLer regression coefficient different from zero and the corresponding genomic position shown in the x-axis. Panel from the bottom to the top: 1) genomic position of the gene with dashed lines representing TSS $\pm 200kb$ window, 2) regression coefficient from our gene expression model color, 3) number of GREs in the PriLer model that a variant intersects (tissue-specific selection in Tab. B.2), 4) -$\log_{10}$ p-value from PGC2 GWAS summary statistics [141]. The color code of each dot reflects PriLer coefficient values and the labelled SNPs correspond to the top PriLer coefficient in absolute values.

This underlined again that PriLer model does not force an association due to GREs presence, rather it takes into account the best fitting configuration. Another interesting candidate we identified as significant was DDHD2 in DLPC, cerebellum and transformed lymphocytes with $-4.23 \leq$ Z-stat ≤ -3.71 (Fig. 4.40B for PriLer regulation in DLPC). This gene was initially identified from exome sequencing as a de novo mutation in SCZ individuals [226] and additionally detected in a recent TWAS for neuropsychiatric disorders [227], validating the hypothesis of a common-rare variant convergence in SCZ [73].

Regarding newly identified associations (Tab. 4.5), the increase of Polycystin 1 Like 1, Transient Receptor Potential Channel Interacting (PKD1L1) in cerebellar hemisphere was associated with higher SCZ risk via a regulation of 26 cis-variants, among which 3 indels represented the most relevant regulatory variants (Fig. 4.40C). PKD1L1 is a

component of a ciliary calcium channel and has been identified from GWAS to overlap with variants associated with anxiety disorders [228]. Notably, we also identified Myeloid Leukemia Factor 2 (MLF2) in cerebellum and transformed lymphocytes whose expression was negatively associated with SCZ. Despite not being identified so far as SCZ related, the gene region overlaps a variant associated with response to paliperidone in SCZ treatment in a pharmacogenomic study [229] (however not passing genome-wide significance, P= 6e-06). For both mentioned genes, their role in the context of schizophrenia needs further investigation.

Chrom	Gene	Loci	Tissue	PGC (Discovery)		CMC (Replication)	
				P-value	Z	P-value	Z
chr2	STEAP3	chr2:119.8-120.2Mb	DLPC_CMC	2.173198e-03	3.065482	0.17579658	-1.3538113
chr2	TTN	chr2:179.5-179.9Mb	DLPC_CMC	1.753626e-03	-3.129066	0.67446598	-0.4200267
chr3	MRPS25	chr3:14.9-15.3Mb	Brain_Cortex	1.131593e-03	3.255584	NA	NA
chr3	IGSF11	chr3:118.7-119.1Mb	Brain_Cerebellum	2.921136e-03	-2.975917	NA	NA
chr3	RP11-553K23.2	chr3:139.1-139.5Mb	DLPC_CMC	1.652047e-03	-3.146558	0.91809521	-0.1028334
chr4	AC021860.1	chr4:38.5-38.9Mb	Brain_Cerebellum	2.725549e-03	-2.997107	NA	NA
chr4	CEP135	chr4:56.6-57Mb	DLPC_CMC	1.586136e-03	-3.158444	0.47240793	-0.7185667
chr6	RP1-101K10.6	chr6:153.1-153.5Mb	Brain_Cerebellar_Hemisphere	2.344453e-03	-3.042727	NA	NA
chr7	DPY19L1P1	chr7:32.3-33Mb	DLPC_CMC	9.768379e-04	3.297114	0.04617969	-1.9937472
chr7	AC018641.7	chr7:32.3-33Mb	Brain_Caudate_basal_ganglia	2.915439e-04	-3.622700	NA	NA
chr7	AC018641.7	chr7:32.3-33Mb	Brain_Frontal_Cortex_BA9	5.442828e-04	-3.457960	NA	NA
chr7	RP11-225B17.2	chr7:32.3-33Mb	Brain_Nucleus_accumbens_basal_ganglia	7.129961e-04	3.384532	NA	NA
chr7	PKD1L1	chr7:47.6-48.2Mb	Brain_Cerebellar_Hemisphere	1.855135e-03	3.112494	NA	NA
chr7	LINC00525	chr7:47.6-48.2Mb	Brain_Cerebellum	2.142654e-03	3.069712	NA	NA
chr8	POLB	chr8:42-42.4Mb	Brain_Hippocampus	2.674889e-04	-3.644907	NA	NA
chr8	RP11-1023P17.2	chr8:53-53.4Mb	Cells_EBV-transformed_lymphocytes	2.117321e-03	3.073263	NA	NA
chr10	NMT2	chr10:15-15.4Mb	Brain_Cerebellar_Hemisphere	2.391460e-03	-3.036747	NA	NA
chr10	CDHR1	chr10:85.8-86.2Mb	Brain_Frontal_Cortex_BA9	1.749773e-03	-3.129713	NA	NA
chr10	FRAT2	chr10:98.9-99.3Mb	DLPC_CMC	2.405818e-03	-3.034942	0.20948361	-1.2549866
chr11	IGHMBP2	chr11:68.5-68.9Mb	Brain_Cerebellar_Hemisphere	1.689124e-03	-3.140061	NA	NA
chr12	MLF2	chr12:6.7-7.1Mb	Brain_Cerebellum	2.123934e-03	-3.072332	NA	NA
chr12	MLF2	chr12:6.7-7.1Mb	Cells_EBV-transformed_lymphocytes	8.473889e-05	-3.930587	NA	NA
chr12	LEMD3	chr12:65.4-65.8Mb	Brain_Cerebellar_Hemisphere	2.575922e-03	3.014277	NA	NA
chr14	RP11-407N17.5	chr14:39.5-39.9Mb	Brain_Frontal_Cortex_BA9	2.178312e-03	3.064779	NA	NA
chr15	AP4E1	chr15:51-51.4Mb	DLPC_CMC	1.400150e-03	3.194620	0.44923863	0.7566850
chr17	TOB1-AS1	chr17:48.7-49.2Mb	Brain_Cortex	1.662417e-03	-3.144727	NA	NA
chr17	RP11-700H6.4	chr17:48.7-49.2Mb	Brain_Frontal_Cortex_BA9	6.205975e-04	-3.422449	NA	NA
chr19	TMEM91	chr19:41.6-42.1Mb	DLPC_CMC	1.340849e-04	3.818832	0.77533413	0.2854047
chr19	EXOSC5	chr19:41.6-42.1Mb	Brain_Cerebellum	2.559290e-03	-3.016242	NA	NA
chr19	CCDC97	chr19:41.6-42.1Mb	Brain_Hypothalamus	1.002937e-03	-3.289701	NA	NA
chr21	TCP10L	chr21:33.8-34.2Mb	DLPC_CMC	7.324729e-04	3.377127	0.09948149	-1.6473726
chr21	BRWD1	chr21:40.5-40.9Mb	DLPC_CMC	2.150610e-03	3.068605	0.77595625	-0.2845926

Tab. 4.5.: Tissue specific significant genes for SCZ in discovery cohorts PGC grouped into loci with overlapping variants from GWAS [141] not significant at FDR 0.05 threshold, with replication only for DLPC tissue from the external CMC cohort that includes schizo-affective and non-affected individuals.

We subsequently performed PALAS from the computed tissue-specific pathway-scores. As before we considered 3 biological pathways databases, Reactome [75], Gene Ontology [76] and WikiPathways [77]. We additionally included "CMC geneSet" collection obtained as SCZ hypothesis driven gene-sets defined in [54] and hence only computed for DLPC tissue in CMC data set. We identified $1,578$ significant pathways across all tissues ($1,080$ unique) of which 255, 692 and 125 in Reactome, GO and WikiPathways respectively and 8 in CMC geneSet. The majority of these associations were detected for DLPC tissue, followed by transformed lymphocyte and cerebellum (Fig. 4.41A) and mostly composed on pathways with less than 5 genes, in terms of T-score genes available to compute the pathway-score. Similarly to genes, we tested replicability in CMC data set considering

significant pathways in DLPC from PGC for GO and Reactome collections and found that 65% out of 261 pathways were concordant in term of Z-statistic sign (one-sided sign test P= 1.08e-06) and 4.2% were also significant at the nominal level $P \leq 0.05$ (Fig. 4.41B). Similarly to CAD, it was possible to compare p-values of a pathway with those from genes included in that pathway. Hence, we can now identify gene-sets with an increased significance due to the aggregation of genes relevance or vice-versa, pathways mostly disrupted from a single gene that surpass the overall pathway significance (Fig. 4.41C). We detected 980 (62%) of significant pathways that included at least one gene more significant than the level reached by the pathway itself (class I pathways), while the remaining 598 (38%) showed an aggregating genes mechanism (class II pathways). Among class II pathways, 316 contained at least one gene significant at FDR 0.05 threshold, whereas 282 were composed in their totality of small effect non-significant genes.

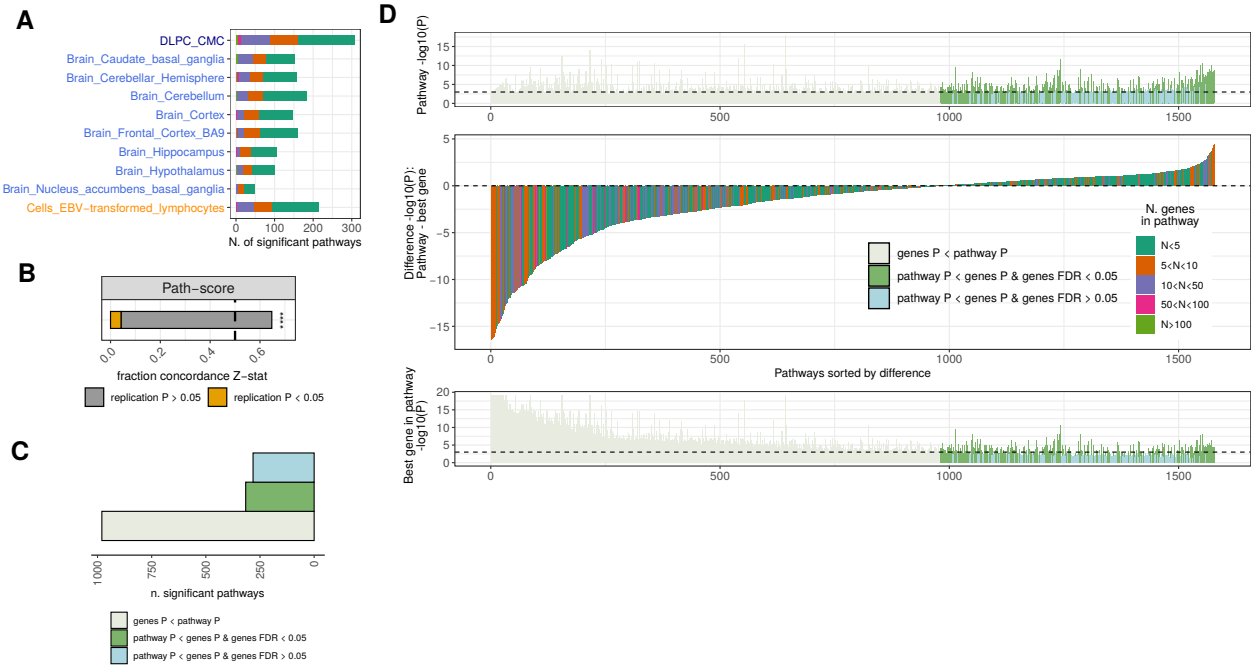

Fig. 4.41.: (Adapted from Trastulla et al., in prep.) **(A)** Number of significant pathways associated with SCZ (tissue-specific FDR≤ 0.05) for each tissue, with the bar color coded according the number of genes in the pathway also reliably predicted in that tissue (T-score genes). **(B)** Reproducibility of pathway scores associations from PGC discovery data set and replication from CMC cohort. X-axis shows the fraction of significant genes in PGC that have the same effect sign (Z-stat) in CMC cohort, p-values are computed from one-sided sign test ($* * ** = P \leq 0.0001$). The bar in yellow represents the fraction of pathways concordant that are also significant at the nominal p-value threshold of 0.05. **(C)** Number of significant pathways (FDR≤ 0.05) that include at least one gene more significant than the pathway (ivory), all genes in the pathway less significant but with at least one gene having FDR ≤ 0.05 (green), and all genes in the pathway less significant and not passing FDR 0.05 threshold (light blue). **(D)** For each significant pathway, the central panel shows the difference of -$\log_{10}(P)$ between the pathway and the most significant gene included, sorted from the smallest to the highest on the x-axis and color coded according the number of T-score genes for that pathway. Top and bottom panels refer to pathway and most significant gene -$\log_{10}(P)$ respectively, with the color referring to pathway classification as in (C). Dashed horizontal line in top and bottom panels correspond to $P = 0.001$.

The level of significance compared to the corresponding best T-score gene included in the pathway was strikingly wide for pathways in class I (Fig. 4.41D). Conversely, small

effect genes aggregate for class II pathways led to an increase in significance generally higher when including at least one significant genes and for pathways constituted of 5 to 10 genes.

To prioritize SCZ-associated gene-sets, we applied the same filtering strategies as CAD including only pathways with > 5 T-score genes or ≥ 3 in case of a coverage ≥ 0.1 (i.e. $\frac{n_{\mathcal{P},\mathcal{G}}}{n_{\mathcal{P}}}$), less than 200 genes in both original and T-score genes and PALAS $P \leq 10^{-4}$.

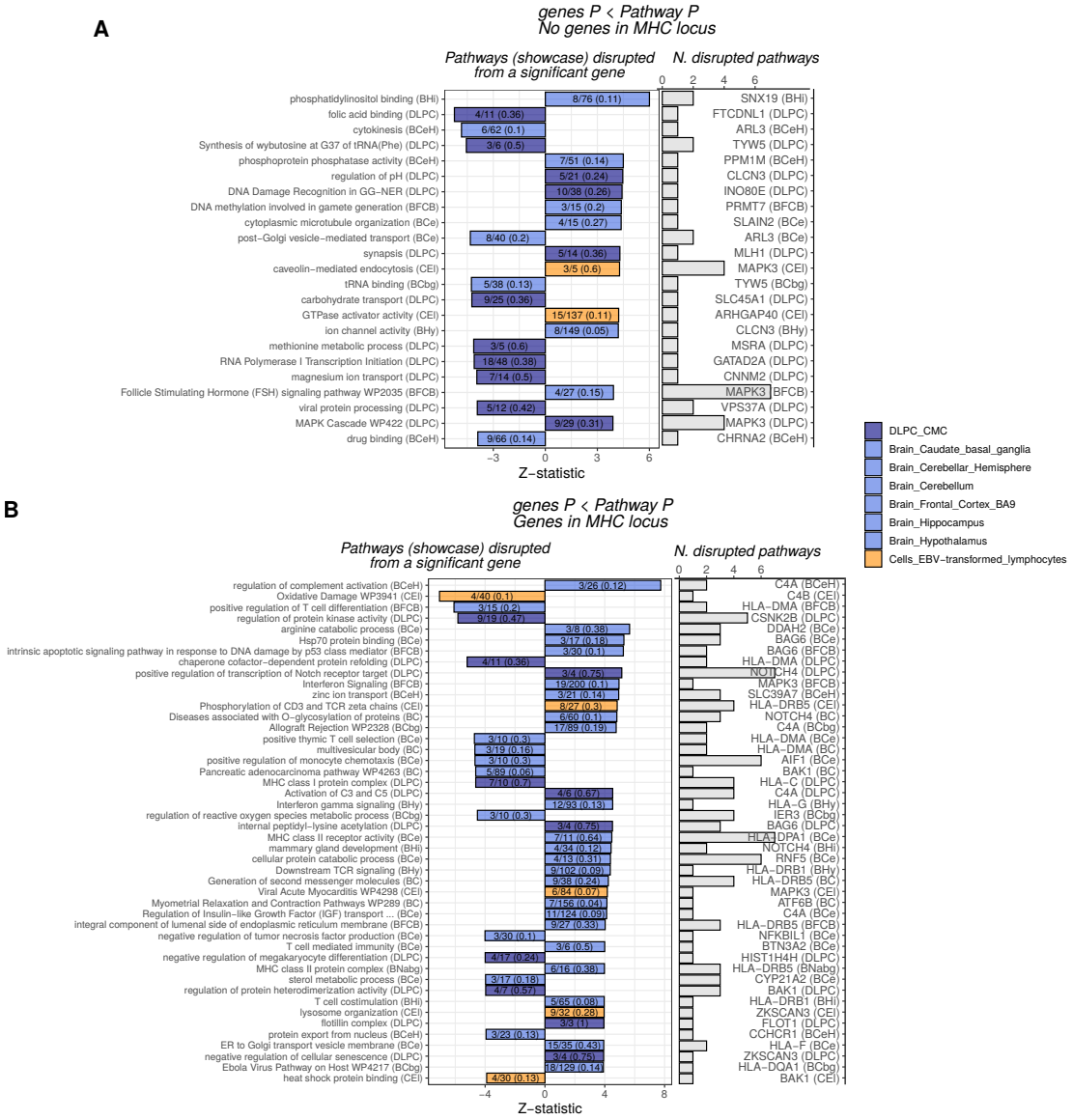

Fig. 4.42.: (Adapted from Trastulla et al., in prep.) Significant pathways with at least one gene more relevant than the pathway itself (ivory bar in Fig. 4.41C). Prioritization of high confidence results that include more than 4 genes or more than 2 in case of a coverage $\geq 10\%$, less than 200 genes in both original genes and T-score genes pathway and p-value ≤ 0.0001. X-axis shows the PALAS Z-statistic color coded by tissue origin, with each pathway barplot including the gene pathway coverage. The acronym in brackets refer to the initials of the tissue considered. Showed only one exemplar pathway per most significant gene (number of pathways including that gene on the right), with the showcase selected as having the highest coverage. In case of the same pathway present across multiple tissue, only the most significant one is kept. **(A)** Selected pathways that do not include any genes in MHC locus, **(B)** or include at least one gene in MHC locus.

In addition, we distinguished between pathways including no genes in MHC locus (Fig. 4.43A, 4.42A) and at least one gene in MHC locus (Fig. 4.43B, 4.42B), to investigate the impact of genes outside the strongest TWAS hit. For significant pathways in class I, we prioritized 172 pathways (Fig. 4.42) of which 47 and 125 without any gene in MHC or with at least one gene in MHC, respectively. We regarded class I pathways as those being disrupted by a single highly significant gene and possibly not cooperating with any other gene in the pathway. Indeed, the disrupted pathways were associated to 29 and 52 significant genes for which a pathway exemplar is shown in Fig. 4.42, selected based on the highest coverage and uniqueness across tissues (23 and 47 respectively). Outside MHC locus (Fig. 4.42A), MAPK3 accounted for the majority of associated pathways (> 15 in total across all tissues), involved in mechanisms such as *MAPK Cascade* and *FSH signaling pathway*. This was in line with the hypothesis of SCZ pathophysiology is related to abnormalities in integration of signaling mediated by multiple neurotransmitter receptors [230]. Interestingly, CLCN3 gene was significantly increasing *ion channel activity pathway* in hypothalamus which indeed represents a long recognised hypothesis arisen from GWAS [231]. As regards class I pathways including genes in MHC locus (Fig. 4.42B), the majority implied immune related mechanisms with NOTCH4 and CA4 gene being the most disruptive genes across multiple tissues. Notably, *Regulation of Insulin-like Growth Factor* pathway was also significantly increased by C4A changes implying a pleiotropic effect. In addition, we identified pathways not previously reported through GWAS nevertheless showing evidence from additional lines of investigation, such as *Oxidative Damage* and *Disease associated with O-glycosylation of proteins*. In particular, *Oxidative Damage* pathway was perturbed in transformed lymphocytes due to C4B decrease, encoding the basic form of complement factor 4 as C4A; whereas *O-glycosylation of proteins* was impaired by NOTCH4 in cortex.

As regards class II pathways arising from an aggregation of effects, we prioritized 45 and 41 gene-sets that were composed on no genes in MHC locus or at least one gene, respectively (Fig. 4.43). Among those, 7 were regarded as novel as the did not include any gene passing FDR significance and they were found in pathways not including any genes in MHC locus (Fig. 4.43A). Some of these pathways were again related to well-known SCZ pathomechanisms such as immune related pathways and complement system when including genes in MHC locus (Fig. 4.43B). In addition, pathways such as *ErbB signaling pathway* and *mTOR signaling* related to myelination [232] and *regulation of neuronal apoptotic process* were identified from genes outside the MHC locus (Fig. 4.43A). For instance, the latter significant pathway is connected to the hypothesis that increased sensitivity to apoptosis might induce synaptic or dendritic neuronal loss in SCZ patients [233]. Notably, we also find evidence of aggregation in *Adipogenesis* pathway. In SCZ contest, this pathway was associated with adipose tissue dysfunction from the decrease in adiponectin, with a consequential down-stream impairments for increased C-reactive protein and fasting glucose [234].

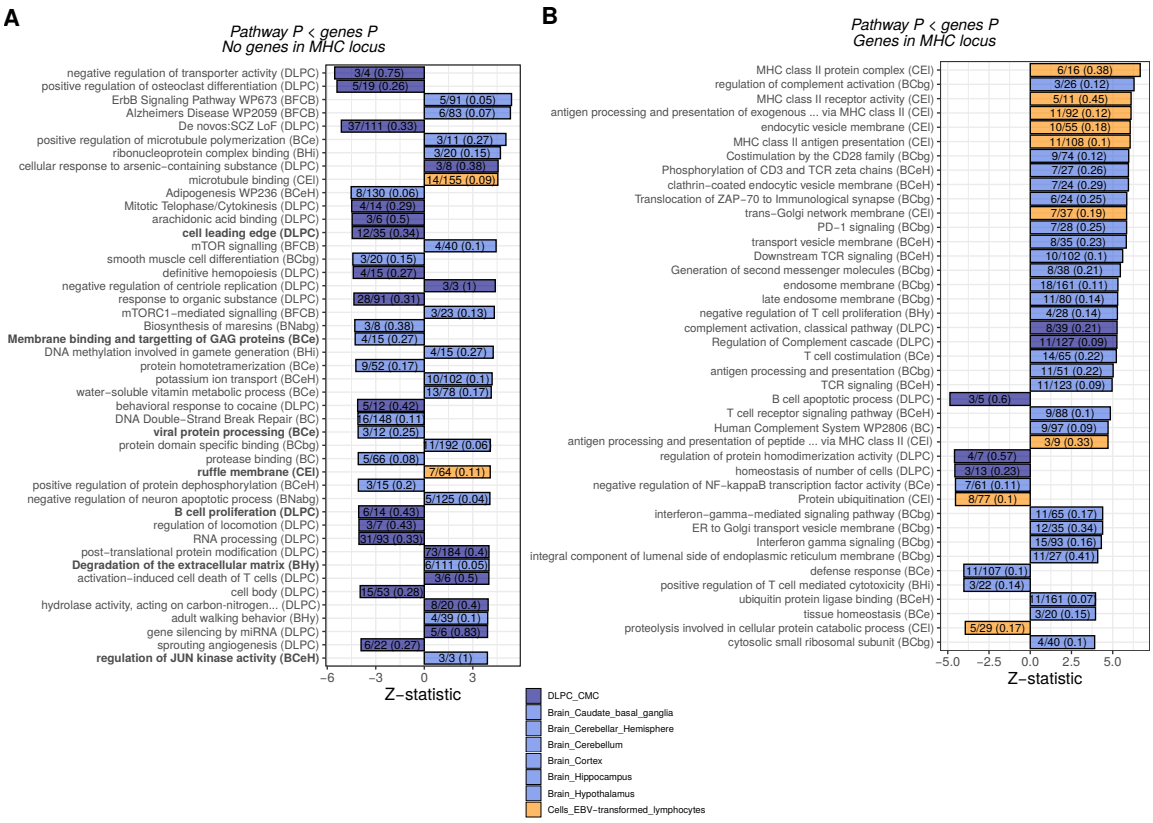

Fig. 4.43.: (Adapted from Trastulla et al., in prep.) Pathways more significant than any included gene (green and light blue from Fig. 4.41(C)), prioritization as described in Fig. 4.42. X-axis shows the PALAS Z-statistic color coded by tissue origin, with each pathway barplot including the gene pathway coverage. The pathway names in bold indicate those without any significant gene at the FDR 0.05 level and the acronym in brackets refer to the initials of the tissue considered. **(A)** Selected pathways that do not include any genes in MHC locus, **(B)** or include at least one gene in MHC locus.

Furthermore, we investigated 3 examples of pathways arising from an aggregation of genes effects and hence exceeding any gene-level (Fig. 4.44), namely *cell leading edge*, *De Novos: SCZ loss of function (LoF)* in DLPC and *calcium ion transmembrane transport* in cerebellar hemisphere. The first two gene-sets were also prioritized among class II pathways not including any genes in MHC locus, however the latter was still significant at FDR 0.05 threshold but did not pass the prioritization strategy due to $P < 10^{-4}$. The Manhattan plots in Fig. 4.44 show TWAS results for all the genes in the tissue considered but highlighted those also included in the pathway under consideration. Different from CAD, we did not perform GWAS using exactly same variants and individuals but we still compared our results to SCZ GWAS from wave 2 [141] (bottom part) and overlay on those variants the PriLer regulatory coefficient outcome. Strikingly, *cell leading edge* pathway (Fig. 4.44A) computed from 12 out 35 genes reached a level of significance more than twice of the most significant gene in that pathway (in $\log_{10}$ scale), i.e. Steroid Receptor RNA Activator 1 (SRA1), underlying a cooperative effect among genes. In addition, no but one genes would be identified as relevant for SCZ from GWAS study (bottom panel), with the only exception of SRA1 when using a non stringent cut-off. Notably, the leading SNP

in SRA1 cis-regulatory region from GWAS had the highest impact in gene regulation from PriLer coefficient.

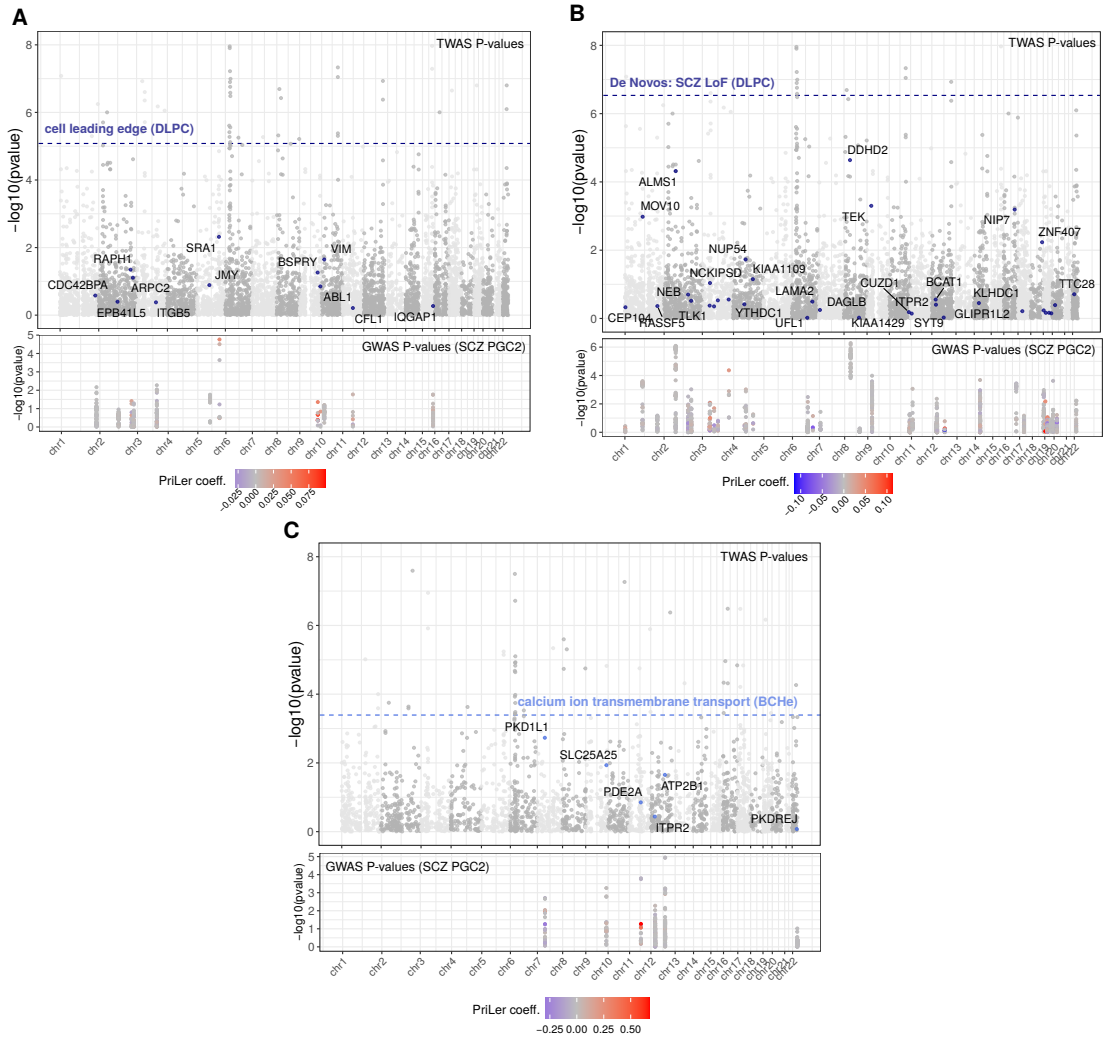

Fig. 4.44.: (Adapted from Trastulla et al., in prep.) Selection of pathways reaching a better significance than the single genes: **(A)** GO "cell leading edge" in DLPC, **(B)** CMC Gene Sets "De Novos: SCZ LoF" in DLPC and **(C)** GO "calcium ion transmembrane transport" in brain cerebellar hemisphere. Each top panel shows -$\log_{10}(P)$ from TWAS for all genes tested in that tissue, with colored and labelled ones referring to those included in that pathway. The dashed line corresponds to -$\log_{10}(P)$ of the considered pathway from PALAS. Each bottom panel shows GWAS p-value from [141] of SNPs regulating those genes (color reflecting PriLer regulatory coefficients).

Moreover, de novo loss-of-function pathway in DLPC was also indicating an aggregation mechanisms from the 37 out of 111 included genes (Fig. 4.44B). This gene-set structure was obtained from [54] and was identified collapsing the results of rare variants exome sequencing of multiple SCZ family studies. Despite being composed of SCZ significant genes such as DDHD2, the level of significance reached by the pathway was still exceeding any gene. Most importantly, this finding supports the hypothesis of agreement between rare and common variants that affect the same genes and hence could be related to analogous pathomechanisms [73]. The cumulative behaviour of genes effect will be investigated in the next section using de novo loss-of-function gene-set as a case study. Finally, the novel

identified SCZ related gene, PKD1L1 in cerebellar hemisphere, was part of *calcium ion transmembrane transport* pathway (Fig. 4.44C) and contributed to increase the overall pathway significance together with other 5 genes composing the pathway (out of 53). Disruption in calcium ion-channel were long identified as plausible etiology of SCZ from GWAS [141].

In conclusion, the genes and pathways identified via CASTom-iGEx are in line with widely studied and previously reported results and suggested new possible genes and biological mechanisms, even arising from a cooperation of effects from multiple genes.

Before moving on towards the genetic relationship of SCZ with related phenotypes, we show in the next paragraph the mechanism of cumulative genes effect on pathway level and its dependence from the same sign of TWAS associations.

4.4.2 Incremental effect from pathway-scores

Here, we studied the effect of aggregation of genes effect using as showcase the significant De novo LoF gene-set in DLPC. As previously mentioned, this gene-set is a collection of genes harboring rare variants detected in probands from multiple SCZ family studies. In DLPC, this pathway was composed of 35 genes and reached a significance of $P = 2.92\text{e-}07$, improved with respect to any gene level (genes $P \geq 2.29\text{e-}05$). To understand the impact of a single gene on the total pathway association, we sorted the 35 genes in the considered pathway incrementally with respect to SCZ Z-statistic and added one gene at a time to the gene-set structure, computing at each increment the gene-set association with SCZ (i.e. PALAS Z-statistic and p-value, Fig. 4.45A). First of all, we observed that the majority of the genes in the gene-set (28 out of 35) is negatively associated to SCZ (Z-stat < 0). Notably, an increment in the pathway level corresponding to the best incremental gene-set configuration was achieved when adding same directional effect genes even with very low effect (i.e. until nominal $P < 0.1$), arguing for the importance of the small effect variants in SCZ architecture. In addition, significant opposite sign association can disrupt the overall pathway signal. For instance, the overall pathway significance drastically decreased when adding ALMS1 gene that was positively and significantly associated with SCZ, hence with an opposite sign with respect to the majority of genes (Fig. 4.45A). On the other hand, genes with a negative Z-statistic but not associated with SCZ even at the nominal level contributed only to noise increase to the gene-set signal and thus slowly decreased the overall level of significance (from NEB to ULF1 genes). Importantly, the considered genes were independent and mostly located in different loci (Fig. 4.45B) with only ALS2CL and NCKIPSD both in 3p21.31 and indeed showing the highest interaction (Corr.$= -0.03$). Note that the gene correlation in this case was computed as weighted average correlation across all cohorts multiplied by the

root-effective sample size N_f, namely

$$N_f = \frac{4}{\left(\frac{1}{N_{cases}} + \frac{1}{N_{controls}}\right)}$$

Thus, the increment in the pathway compared to the single genes that we observed can be a consequence of different patients' liability that converge into the same genetic mechanism.

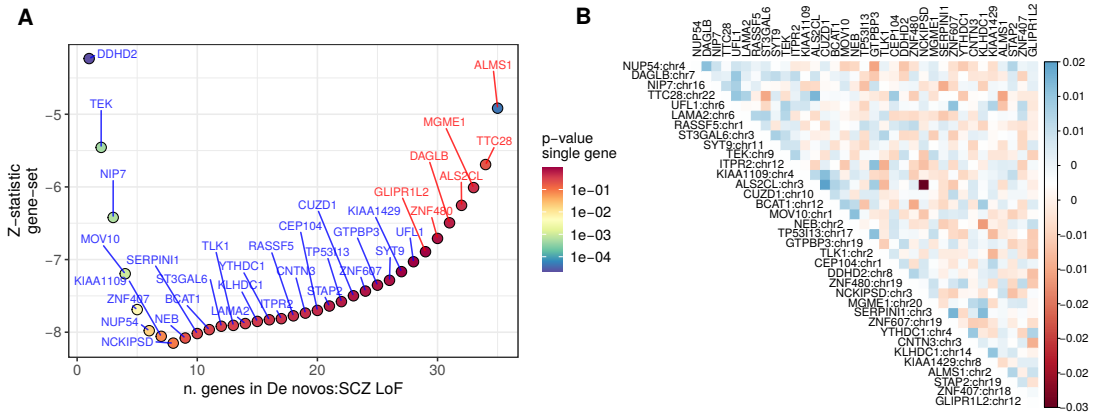

Fig. 4.45.: **(A)** Incremental significance on pathway level from genes in "De novos: SCZ LoF", with each gene added one-by-one to compute a pathway-score going from lowest to highest Z-statistic value. X-axis shows the number of genes composing the incremental pathway, y-axis shows the Z-statistic level reached with that pathway configuration and the labelled dot indicates the gene that is added at that step. The dot color code represent the actual TWAS p-value of that added gene and the label color code indicates the gene-specific Z-statistic sign (blue is negative and red is positive). **(B)** Heatmap of the pairwise genes correlation built on PGC cohorts.

4.4.3 Phenotypic interpretation of genes and pathways

Given the pronounced heterogeneity in symptoms, clinical manifestation and disease course of SCZ, we next sought to leverage the CASTom-iGEx pipeline to identify endophenotypes associated with the genetic basis of SCZ, contributing to this heterogeneity. To that end, we once again leveraged the rich collection of phenotypes within UKBB with potential relevance to SCZ (144 in total, Tab. B.8) and performed tissue-specific correlation analysis and Mendelian Randomization (MR) between SCZ and the considered endophenotype associated pathway, using a bidirectional approach for MR strategy. Namely, we aimed at identifying endophenotypes genetically similar to SCZ and among those 1) detect endophenotypes that exhibit a causal or protective role via MR, with endophenotype regarded as exposure (e.g. lymphocyte count) and SCZ as outcome or 2) detect endophenotypes that are a consequence of SCZ predisposition, with SCZ regarded as exposure and endophenotype as outcome (e.g. fluid intelligence) (see section 3.2.3). Note that, UKBB and SCZ results were obtained from a non-harmonized set of SNPs between the two data set and hence can present a different gene expression regulation and consequentially pathway-score distribution. For this reason, we first restricted to genes and pathways that

were reliable and generally available in both data set as well as with a Pearson correlation exceeding 0.8 (Pearson) for imputed gene expression and pathway scores (Fig. 4.46), see section 3.2.4 for details.

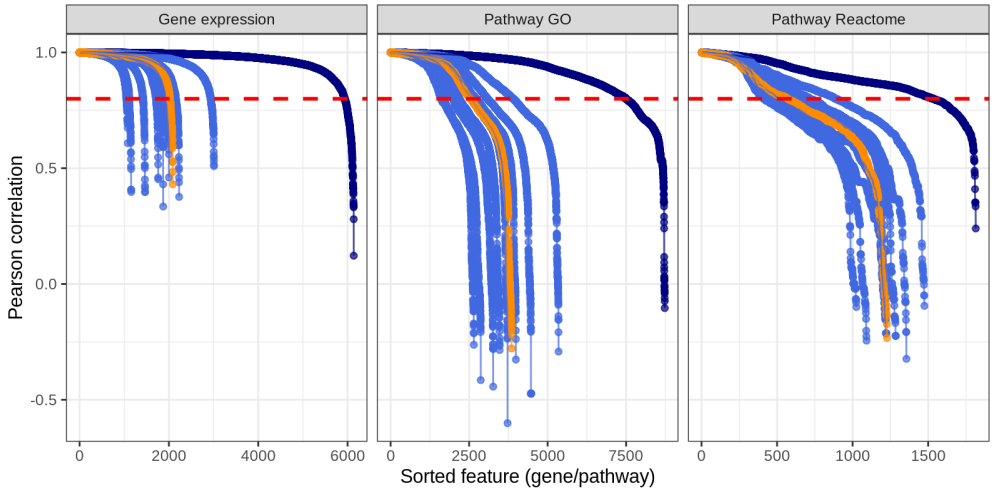

Fig. 4.46.: Sorted genes and pathway based on correlation between scores computed on UKBB and on PGC data set. Connected dots by a line correspond to a specific tissue. Dashed horizontal line corresponds to applied cut-off of 0.8 for considering genes/pathways in further analyses, i.e. Mendelian Randomization and clustering.

In Fig. 4.47 is shown the estimated Spearman correlation and signed MR significance from MR-IWV method for a selection of SCZ related endophenotypes using genes T-scores or pathway-scores as instrumental variables, panel A and B respectively. Of note, the MR-IVW methodology was here operating in a 2-sample setting, with exposure and outcome estimated from two non-overlapping data sets.

Using this strategy, we identified the genetic predisposition to SCZ to also mediate a reduction of fluid intelligence (Fig. 4.47). This was particularly evident from 118 pruned pathways in DLPC associated with SCZ (P$=$ 8.09e-07) and only significant at the nominal level in the same tissue from 101 pruned genes associated with SCZ (P$=$ 0.03). Of note, a reverse effect of lower intelligence increasing the predisposition for SCZ was only detected when considering pathway-scores, but less significant for DLPC than the opposite MR direction (P$=$ 0.005). Looking into mediating mechanisms, key molecular gene-sets influencing both SCZ and fluid intelligence in DLPC were nervous system development, axon terminus and folic acid binding (Fig. 4.48A) and driver genes of the observed causal effect could be identified in C2orf47, ZSCAN23 and ALMS1P (Fig. 4.48B). Of note, the heterogeneity in SCZ-endophenotype relationship tested via Cochran's Q statistic was always significant both when considering genes and pathways (Q-stat P$=$ 2.57e-30 and P$=$ 1.34e-12 respectively), indicating a pleiotropic effect.

Fig. 4.47.: Correlation and causality of SCZ and SCZ related phenotypes in UKBB from **(A)** gene T-score and **(B)** pathway-scores association with genes pruned based on TSS$> 250kb$ and pathways pruned based on Jaccard similarity > 0.3. In both panels, heatmap on the left shows tissue specific Spearman correlation of Z-statistics between SCZ and selected UKBB phenotypes (rows), heatmaps in the middle/right show $-\log_{10}$ p-value from MR-IVW with the sign indicating the direction of estimate for correlated phenotypes (not gray cells), with the middle panel referring to MR results when SCZ is the exposure and right panel when SCZ is the outcome

In general, these results were consistent with previous GWAS based analysis [36, 235] that report shared genetic influences of SCZ and intelligence and MR significant association in

both directions.

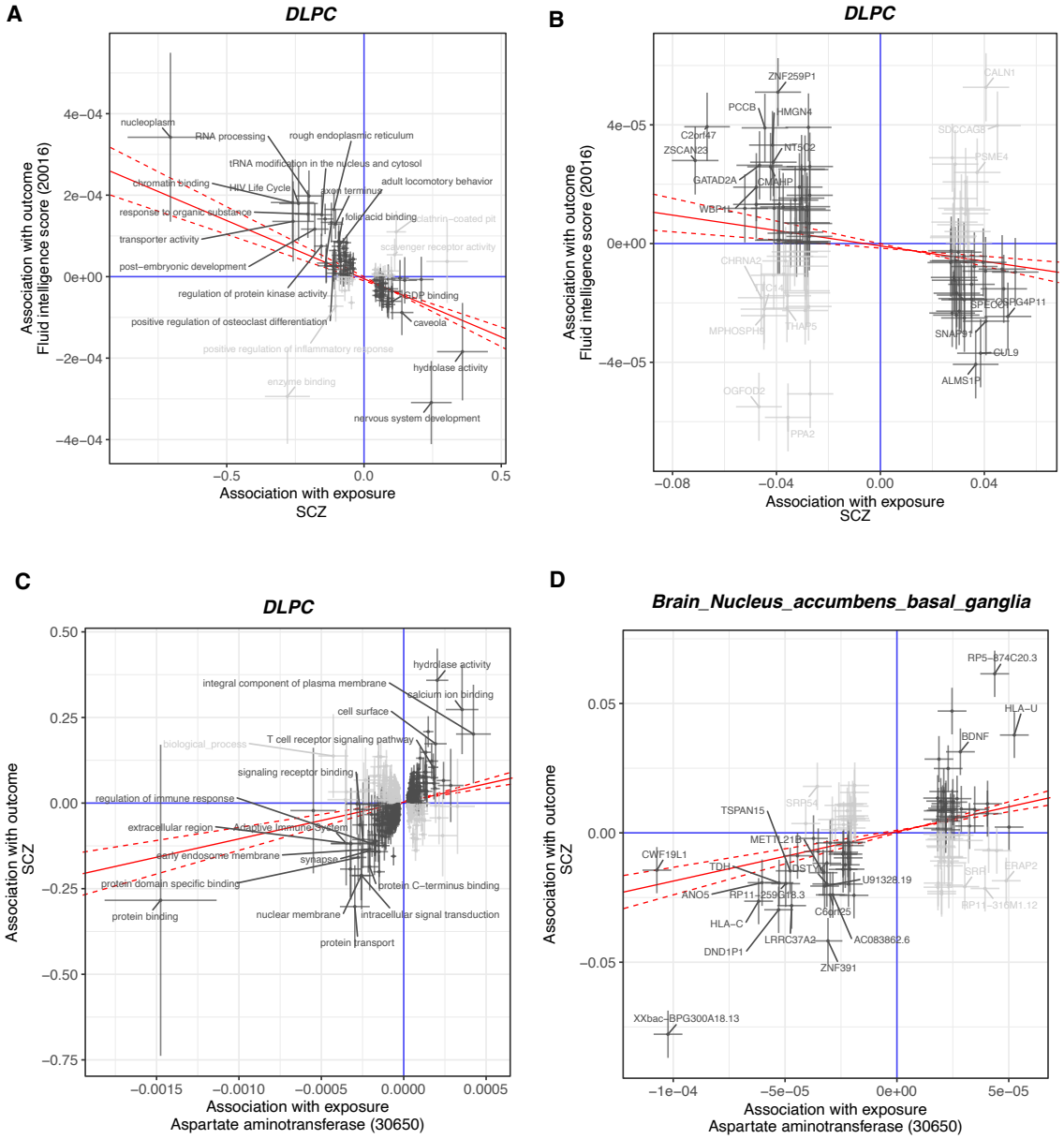

Fig. 4.48.: Scatter plots of the effect sizes with 95% confidence interval for **(A-B)** Fluid intelligence score as outcome in DLPC, **(C)** aspartate aminotransferase as exposure in DLPC and **(D)** aspartate aminotransferase as exposure in nucleus accumbes basal ganglia. **(A,C)** panels show MR-IVW from pathway-scores and **(B,D)** from gene T-scores. In each panel, the red line represents the causal estimate using the IVW with 95% confidence interval, in black and gray pathways/genes with association concordant and discordant in sign respectively between SCZ and the endophenotype.

In addition, genetically mediated pathway de-regulation in SCZ had a casual effect on increasing dysfunctions for visual memory and prospective memory with the strongest significance increase for 'Time to complete round' in the pairs matching test (P= 6.6e-10 in DLPC, Fig. 4.47B). Note that, this memory phenotype and SCZ were showing a mediating mechanisms both using gene t-score and pathway-score, and in the case of genes being confirmed across all tissues. These results provided further evidence for a genetic

contribution to overall decreased cognitive performance observed even in drug-naïve SCZ patients, confirmed at the early stage of the disease [236].
Among phenotypes with bidirectional causal effect, we identified aspartate aminotransferase (AST) in caudate basal ganglia, significant in both directions and both when considering genes (Fig. 4.47A, SCZ exposure/outcome P= 0.0062 / = 0.0064) or pathways (Fig. 4.47B, SCZ exposure/outcome P= 4.68e-06 / = 0.006) as instrumental variables. Regarding AST as mediator causal effect (G $\rightarrow$ AST $\rightarrow$ SCZ), highly significant tissues were DLPC when considering pathways (Fig. 4.48C) and nucleus accumbens basal ganglia when considering genes (Fig. 4.48D). From 543 pathways associated with AST in UKBB, the MR-IVW estimate effect (P= 0.00051) included mediating pathways with concordant effect, among which calcium ion binding, synapse, and adaptive immune system. In addition, from the 106 genes associated with AST in basal ganglia, mostly non-coding genes are responsible for the observed predisposition to SCZ (P= 7.4e-05). As in the previous example, heterogeneity of genetic instruments was significant (Cochran's Q statistic P$<$ 3e-22), possibly representative of a pleiotropic mechanisms. Although AST is considered mostly as liver injury biomarker, it is also found in brain and catalyses aspartate and alpha-ketoglutarate conversion to oxaloacetate and glutamate, and thus with a possible connection to the *glutamate hypothesis* for SCZ [237].

In summary, these observations further underscore that polygenic risk factors for SCZ converge onto distinct biological processes, affecting distinct endophenotypes relevant to symptoms and clinical presentation of SCZ.

4.4.4 Patients stratification from imputed gene expression

Similarly to CAD, we searched for evidence of distinct genetic liability profiles in individuals affected by SCZ related to their clinical heterogeneity [238], applying the third module of our CASTom-iGEx pipeline (Fig. 3.1). In particular, we considered SCZ patients among the 35 PGC2 cohorts (collectively called PGC2) and left out 1 cohort composed of $1,773$ cases for validation purposes (scz_boco_eur). We investigated 2 different filtering strategies for genes, removing genes with |corr.| > 0.9 in the first scenario and > 0.1 in the second scenario, such that the MHC locus contribution would be drastically reduced. Before proceeding to the patient stratification, we first removed individuals that were outliers (as described in 3.3.1) in at least one tissue and in at least one configuration, going from $22,991$ to $22,732$ affected individuals. This strategy detected from 3 to 7 clusters across tissues and filtering strategies, with varying size from 211 to $12,488$ when genes |corr.| ≤ 0.9 and from 7 to $13,398$ when genes |corr.| ≤ 0.1 (Fig. 4.49A). The clustering structure largely overlapped in the permissive filtering strategy (Fig. 4.49B, lower triangular), but greatly reduced with the strict filtering threshold (Fig. 4.49B, upper triangular), indicative of the lower influence of MHC locus in defining the stratification and favoring tissue-specific genes. In addition, the two filtering strategies partially overlapped

(Fig. 4.49B, diagonal) leading to an intra-tissue NMI between 0.24 and 0.34 for all the tissues but DLPC (NMI$= 0.05$).

In the next sections, we will display in detail the results obtained stratifying patients based on DLPC tissue, as one of the most relevant tissue for SCZ and well characterized due to the large sample size in the reference panel CMC. Different from CAD, endophenotype and clinical information were not available in the PGC cohorts, nevertheless we detected plausible group-specific phenotypic differences approximating SCZ related endophenotypes via gene risk-scores (gene-RS), estimated from UKBB phenotypes (see section 3.3.4). To ensure the reliability of group-specific differences found via gene-RS, in section 4.4.7 we validated this approach in CAD, leveraging CARDIoGRAM cohorts for which the endophenotype was not available (as for PGC), and compared their group-specific gene-RS results with the actual differences detected in UKBB endophenotypes.

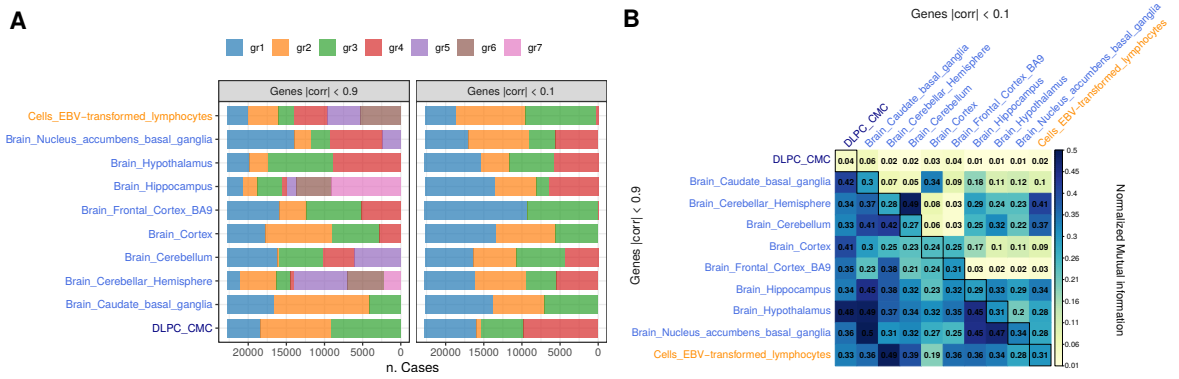

Fig. 4.49.: (From Trastulla et al., in prep.) **(A)** Proportion of individuals in each tissue-specific cluster among SCZ patients considering 2 filtering strategies: genes are clumped based on imputed R^2 removing genes with Pearson correlation higher than 0.9 (left) and higher than 0.1 (right). **(B)** Normalized mutual information (NMI) for each pair of tissue-specific clustering, among the 2 filtering strategies. Lower triangular matrix refers to absolute gene correlation < 0.9, upper triangular matrix to absolute correlation < 0.1, with the diagonal showing intra-tissue NMI between the two filtering strategies.

4.4.5 Patients stratification in DLPC

The clustering was performed combining all the 35 cohorts together and considering $5,678$ gene T-scores from DLPC, clumped at 0.9 correlation, standardized, corrected for 10 PCs (computed merging all cohorts) and finally multiplied by SCZ Z-statistic previously obtained via TWAS, leading to the identification of 3 clusters (Fig. 4.50A). Applying the same FDR threshold of 0.01 as in CAD, we identified 594 cluster-specific genes out of $26,836$ tested across the 10 tissues. Among those, 92% are located in the MHC locus and largely overlap between groups (404 in common), compared to the cluster-specific genes outside the MHC locus that were mostly unique for each cluster (Fig. 4.50B). Of note, the clustering was largely driven by a specific configuration of genes inside the MHC locus, as the most significant genes with WMW p-values around 0 were located in the extended MHC region, in contrast to the associations outside MHC locus starting from a p-value $>$ 1e-10

(Fig. 4.50C). Collapsing the 594 cluster-specific genes, we found 55 loci tissue-specific that correspond to 34 overall loci (Tab. B.9). The number of associated genes was almost constant across the groups due to the spread overlap in MHC locus, nevertheless the highest number of loci was detected in gr_1 (Fig. 4.50F), with a varying tissue specificity that did not solely consider DLPC (Fig. 4.50G). Reducing the complexity of each locus to a single candidate exemplar (Fig. 4.50D), we observed that the most prominent difference is C4A in DLPC (P$< 2.1e-174$), with gr_1 WMW estimate of opposite direction to its SCZ Z-statistic sign (WMW est $= -2.31$), indicating a down-regulation possibly preventive of severe cognitive symptoms (see below). Among additional loci associated with DLPC clustering, M-Phase Phosphoprotein 10 (MPHOSPH10, in 2p13.3) was associated with gr_2 and gr_3 in DLPC tissue, with an opposite effect of -0.02 (P$= 2.87e-05$) and 0.02 (P$= 1.18e-03$) respectively. Despite MPHOSPH10 not being associated with SCZ (Z-statistic$= 0.35$), variants overlapping this gene region were previously found related to intelligence from GWAS [235]. Another example is Aldehyde Dehydrogenase 18 Family Member A1 (ALDH18A1, 10q24.1) with differences detected in brain cerebellum solely for gr_1 (WMW est = 0.07, P$= 1.1e-3$), found to be negatively associated with fluid intelligence score phenotype via TWAS in UKBB (Z-statistic$= -3.58$, FDR $= 0.016$). Both examples suggest other mechanisms regardless MHC locus that could impact cognitive functions manifestations. Comparing the WMW estimates with the TWAS SCZ associations across all 10 tissues (Fig. 4.50E), it resulted that gr_1 represented a group of individuals affected by SCZ but with lower genetic liability towards it (Spear. corr. $= -0.94$), opposite to individuals in gr_3 with higher SCZ genetic liability (corr.$= 0.89$) and gr_2 being in-between having a variable concordant/discordant relation to SCZ associated genes (corr$= 0.15$).
Due to its relevance, we specifically zoomed into the MHC region and observed the cluster-specific genes in DLPC at the individual level (Fig. 4.51), filtering out in the figure genes with correlation higher than 0.7 to avoid highly redundant information. As already mentioned, C4A plays a pivotal role in separating gr_1 and the other two groups. Because HLA-DMA is negatively correlated with C4A (corr. $= -0.53$), a similar pattern in group separation can be observed for HLA-DMA, but with opposite signs in associations. However, genes such as ZSCAN23 correlated to C4A with a lower magnitude (corr. $= -0.18$) and, importantly, exhibited a different stratification criteria, with gr_3 having lower WMW estimates approximation as opposed to the other two groups and in the same direction of SCZ Z-statistic.

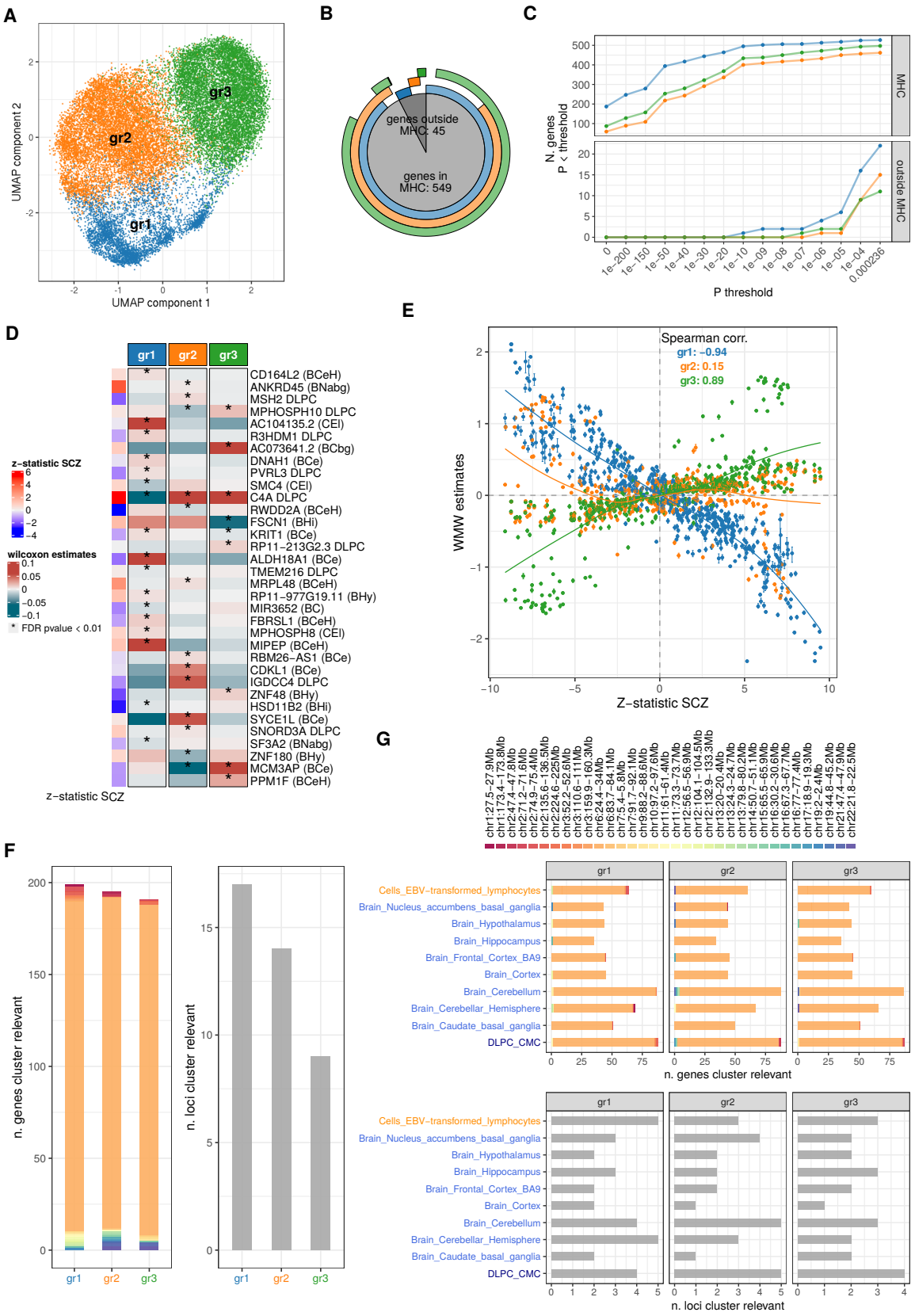
A
UMAP component 2
UMAP component 1
gr1
gr2
gr3
B
genes outside MHC: 45
genes in MHC: 549
C
N. genes P < threshold
MHC
outside MHC
P threshold
0
1e−200
1e−150
1e−50
1e−40
1e−30
1e−20
1e−10
1e−09
1e−08
1e−07
1e−06
1e−05
1e−04
0.000236
D
gr1
gr2
gr3
z−statistic SCZ
wilcoxon estimates
* FDR pvalue < 0.01
CD164L2 (BCeH)
ANKRD45 (BNabg)
MSH2 DLPC
MPHOSPH10 DLPC
AC104135.2 (CEI)
R3HDM1 DLPC
AC073641.2 (BCbg)
DNAH1 (BCe)
PVRL3 DLPC
SMC4 (CEI)
C4A DLPC
RWDD2A (BCeH)
FSCN1 (BHi)
KRIT1 (BCe)
RP11−213G2.3 DLPC
ALDH18A1 (BCe)
TMEM216 DLPC
MRPL48 (BCeH)
RP11−977G19.11 (BHy)
MIR3652 (BC)
FBRSL1 (BCeH)
MPHOSPH8 (CEI)
MIPEP (BCeH)
RBM26−AS1 (BCe)
CDKL1 (BCe)
IGDCC4 DLPC
ZNF48 (BHy)
HSD11B2 (BHi)
SYCE1L (BCe)
SNORD3A DLPC
SF3A2 (BNabg)
ZNF180 (BHy)
MCM3AP (BCe)
PPM1F (BCeH)
E
Spearman corr.
gr1: −0.94
gr2: 0.15
gr3: 0.89
WMW estimates
Z−statistic SCZ
G
chr1:27.5−27.9Mb
chr1:173.4−173.8Mb
chr2:47.4−47.8Mb
chr2:71.2−71.6Mb
chr2:74.9−75.4Mb
chr2:135.6−136.5Mb
chr2:224.6−225Mb
chr3:52.2−52.6Mb
chr3:110.6−111Mb
chr3:159.9−160.3Mb
chr6:24.4−34Mb
chr6:83.7−84.1Mb
chr7:5.4−5.8Mb
chr7:91.7−92.1Mb
chr9:88.2−88.6Mb
chr10:97.2−97.6Mb
chr11:61−61.4Mb
chr11:73.3−73.7Mb
chr12:56.5−56.9Mb
chr12:104.1−104.5Mb
chr12:132.9−133.3Mb
chr13:20−20.4Mb
chr13:24.3−24.7Mb
chr13:79.8−80.2Mb
chr14:50.7−51.1Mb
chr15:65.5−65.9Mb
chr16:30.2−30.6Mb
chr16:67.3−67.7Mb
chr16:77−77.4Mb
chr17:18.9−19.3Mb
chr19:2−2.4Mb
chr19:44.8−45.2Mb
chr21:47.4−47.9Mb
chr22:21.8−22.5Mb
F
n. genes cluster relevant
n. loci cluster relevant
Cells_EBV−transformed_lymphocytes
Brain_Nucleus_accumbens_basal_ganglia
Brain_Hypothalamus
Brain_Hippocampus
Brain_Frontal_Cortex_BA9
Brain_Cortex
Brain_Cerebellum
Brain_Cerebellar_Hemisphere
Brain_Caudate_basal_ganglia
DLPC_CMC
n. genes cluster relevant
n. loci cluster relevant

Fig. 4.50.: (Adapted from Trastulla et al., in prep.) **(A)** First 2 components of uniform manifold approximation and projection (UMAP) from gene T-scores in DLPC standardized across SCZ patients, corrected for PCs, and multiplied by Z-statistic SCZ associations. Each dot represents a patient colored by the cluster membership. **(B)** Pie chart representing the number and fraction of cluster-specific genes, defined as WMW FDR ≤ 0.01 corrected in each tissue separately. Genes are divided based on their overlap with MHC locus (chr6:26Mb-34Mb). The outer circles represent number of cluster-specific genes in each group, color coded as in (A), showing groups overlap among associated genes. **(C)** For each group, number of group-specific genes (y-axis) passing the WMW p-value threshold (x-axis), divided per genes intersecting and outside MHC locus (top and bottom panels, respectively). **(D)** WMW estimates (capped) for the most group-specific significant gene in the 34 associated loci, parenthesis refers to the tissue considered (acronyms refer to the initial of the tissue name). Row annotation on the left indicate the corresponding SCZ Z-statistics from TWAS. **(E)** Group-specific genetic liability with respect to SCZ, each dot is a gene associated with a group (FDR ≤ 0.01). X-axis shows SCZ Z-statistic estimates and y-axis indicates group-specific WMW estimates with 95% CI. **(F-G)** Number significant genes and loci (tissue specific FDR ≤ 0.01) associated with each group from WMW test of group id versus remaining patients **(F)** combing all tissues and **(G)** tissue specific.

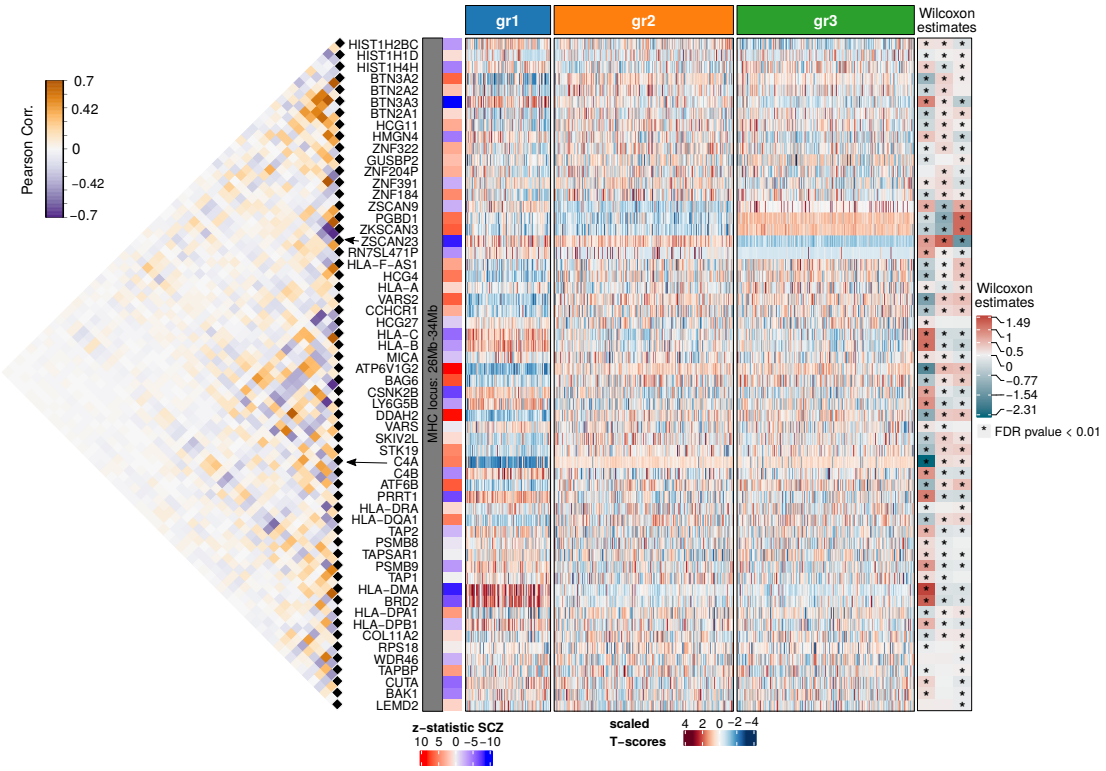

Fig. 4.51.: (Adapted from Trastulla et al., in prep.) DLPC differences for genes in MHC locus. Central heatmap indicates the standardized gene T-scores with each row being a gene in MHC locus (pruned at 0.7 Pearson correlation keeping strongest Z-statistic with respect to SCZ, n. genes$= 58$) and each column a SCZ patient, combining all cohorts and ordering according cluster membership. On the left it is shown the corresponding triangular correlation matrix for the selected genes estimated from SCZ patients. The small heatmap on the right shows a summary of group-wise differences represented as WMW estimates when comparing gr_i against remaining patients and the asterisk indicates whether the gene is significantly different in that group after tissue and group specific BH correction.

A similar scenario was observed for BTN3A3 gene, only marginally correlated with the aforementioned ZSCAN23 and C4A (corr. $= 0.37$ and -0.26). These observations underlie

the importance of MHC locus, despite the challenges in properly estimating it via genotyping arrays [239] due to LD across HLA and non-HLA genes. In particular, we observed here that the contribution to the clustering structure of the entire locus was not simply driven by a single exemplar gene that recapitulate the entire structure, rather by almost independent signals. Nevertheless, the proposed heterogeneity identified to which the MHC locus also contribute needs further investigation via observed endophenotypes (here not available) to evaluate the extent to which imputation in genotyping arrays recapitulate an existing configuration.

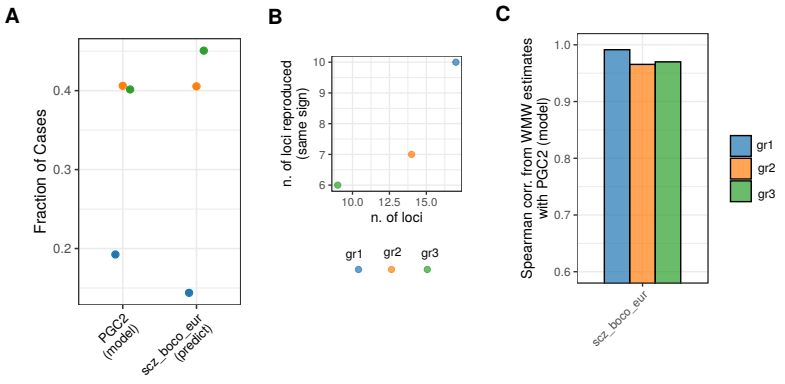

Fig. 4.52.: (Adapted from Trastulla et al., in prep.) Prediction of DLPC clustering structure on scz_boco_eur external cohort. **(A)** Y-axis indicates the fraction of SCZ patients assigned to each group in PGC2 data set and the left-out cohort for which the clustering structure was projected. **(B)** Reproducibility of group-specific loci on predicted groups in external cohort, the x-axis shows the number of loci across all tissues associated with each group in PGC2 data set, the y-axis shows how many of these loci have the same sign using as exemplar the strongest association of the WMW estimates in the predicted clustering structure. **(C)** Spearman correlation of WMW estimates in PGC2 and the external cohort only from genes that are significantly associated with that group (considering all tissues).

To evaluate reproducibility, we projected the gene-level T-scores from the left out scz_boco_eur cohort into the partition built on the other 35 PGC2 cohorts (see section 3.3.1). The fraction of SCZ affected individuals assigned to each group in the left out cohort was similar to the partition built on PGC2 (Fig. 4.52A). In addition, the percentage of loci reproduced in the external cohort was $\geq 50\%$ in each group (Fig. 4.52B). Here we used as metric the concordances in WMW estimate sign and not the nominal significance due to the reduced sample size in the left-out cohort $(1,773)$, thus indicating that the projection involved multiple loci. Finally, comparing the gene expression profiles from WMW group-specific estimates for cluster significant genes, we found that the structure predicted in scz_boco_eur was concordant with PGC2, with Spearman corr. > 0.95 in each group (Fig. 4.52C).

Moreover, we investigated the influence of ancestry information captured via principal components to the clustering structure and we found PC2, PC1 and PC4 with significant changes in the stratified patients (Fig. 4.53A), despite the PCs correction applied prior to clustering detection (see section 3.3.1). Hence, we cannot completely exclude an ancestry contribution to the SCZ partition, however it is clear that genes configuration rather than PC distribution was the driven force defining the clustering structure (Fig. 4.51). In

addition, we compared DLPC clustering with a partition obtained solely from PCs and found non-overlapping endophenotype differences, similarly to CAD. This indicated that the possible contribution from ancestry to the DLPC cluster was nevertheless not influencing the group-specific endophenotypic characteristics (more details in section 4.4.8). Finally, having combined multi-cohorts together in PGC2 clustering, we found that cohort structure was significantly associated with clustering structure (χ^2-test P$= 2.83\text{e} - 15$). However, analyzing the specific group-cohort division (Fig. 4.53), we found that the fraction of samples assigned to each group across all cohort was mostly constant, with exception in enrichment of gr_2 with individuals from scz_ajsz_eur cohort or gr_1 with individuals from scz_irwt_eur cohort, hence once again a confounder not totally removed that cannot be regarded as the driving force providing the observed stratification.

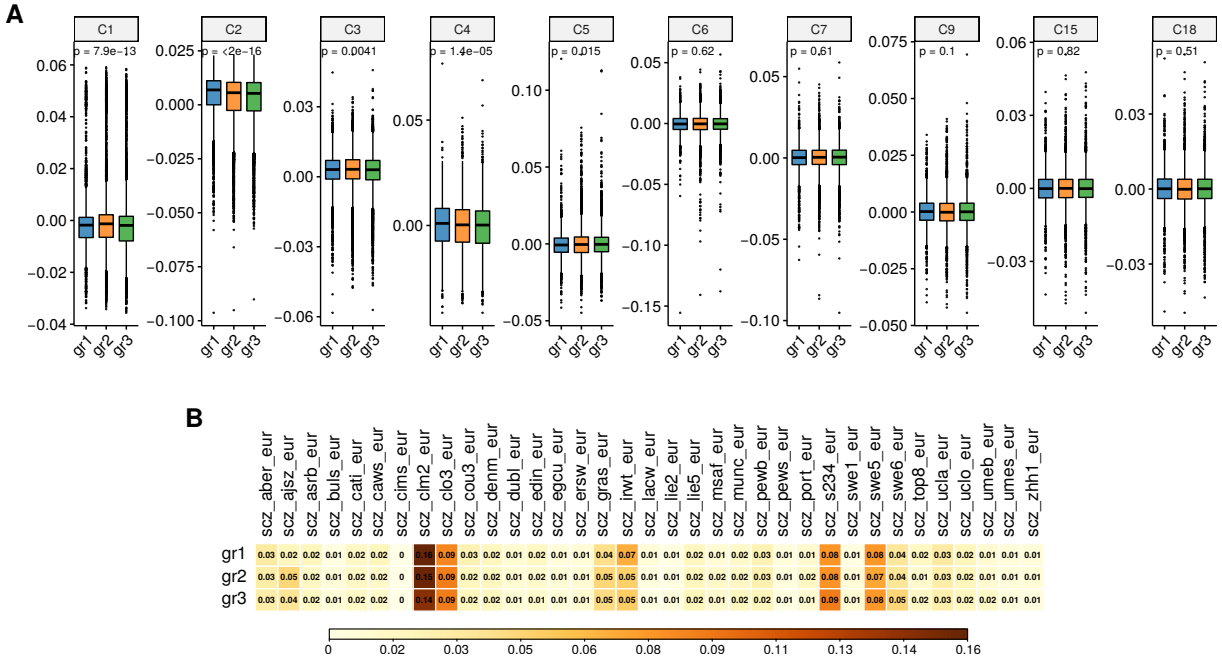

Fig. 4.53.: (Adapted from Trastulla et al., in prep.) **(A)** Distribution of PGC2 10 PCs (same as those used in TWAS and PALAS) for each SCZ DLPC cluster (p-values from Kruskal-Wallis test). **(B)** Contingency table of group and cohort structure with each square referring to the fraction of patients in a group (rows) belonging to a certain cohort (columns). Each row sums to 1.

Subsequently, we detected cluster-specific biological processes leveraging individual pathway-scores. As for CAD, to avoid redundant information we first removed genes-sets with number of gene T-scores lower or equal than 3 and higher than 200 among both GO and Reactome and collapsed the 2 databases clumping pathways with JS > 0.2 while giving priority to gene-sets with highest coverage and number of gene T-scores. We hence tested $6,120$ pathways across 10 tissues with WMW test, resulting in 991 significant associations (FDR ≤ 0.01) across all groups and tissues. We then removed pathways shared among tissues but showing a non concordant association sign, which led to 900 significant results: 360, 258, and 282 for gr_1, gr_2 and gr_3 respectively. In total, we detected 256 unique non-tissue specific pathways, with DLPC tissue up to 239 associations (Fig. 4.54A).

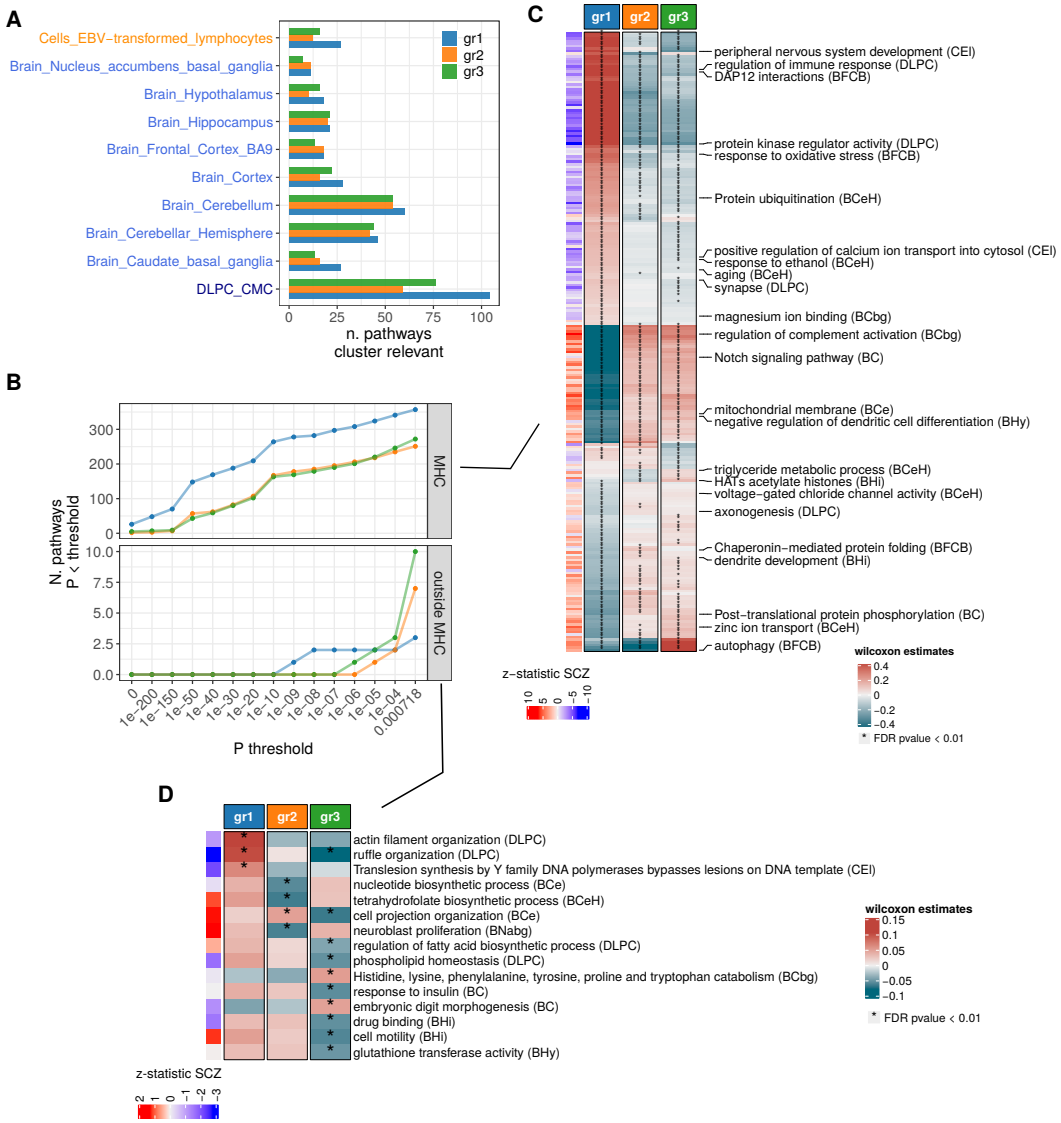

Fig. 4.54.: (Adapted from Trastulla et al., in prep.) **(A)** Number significant pathways (tissue specific FDR ≤ 0.01) associated with each group from WMW test of group id versus remaining patients. The included pathways are both from Reactome and GO and filtered such that Jaccard Similarity ≤ 0.2, retaining the pathways with highest coverage and removing significant pathways having discordant WMW estimates across tissues. **(B)** For each group, number of significant pathways (y-axis) passing the WMW p-value threshold (x-axis), split in pathways that include at least one gene in MHC (top panel) and no gene in MHC (bottom panel). **(C-D)** WMW estimates (capped) for 241 significant pathways (rows) testing each group versus the rest (column) and considering only the most significant tissue per-pathways when repeated. The names on the row are a selection of significant pathways, parenthesis refers to the tissue considered (acronyms indicates the initial of the tissue name). Row annotation on the left refers to the corresponding SCZ Z-statistics from PALAS. Heatmap in (C) refers to pathways including at least one gene in MHC, heatmap in (D) refers to pathways including no genes in MHC.

Similarly to PALAS analysis for SCZ, we differentiated between pathways that included at least one gene or no genes within the MHC locus. We observed a higher significance for the first class of gene-sets (Fig. 4.54B), as expected from cluster-specific gene associations. In particular, a total of 20 associations referred to pathways not including any gene in MHC locus (Fig. 4.54D), among which 15 were also more significant than the WMW association

reached by any genes included in the pathway. For instance, *glutathione transferase activity* computed from GSTM4, HPGDS, GSTT2B, GSTT2 and GSTT1 genes is decreased in gr_3 (P=$5.94e-4$, estimate $=-0.05$). Glutathione is an antioxidants able to prevent oxidative stress and from a meta-analysis individuals affected by schizophrenia exhibited lower levels glutathione ([240]), reinforcing the higher severity of gr_3 in SCZ symptoms not due to MHC locus mechanisms. In addition, *cell projection organization* tested in brain cerebellum and composed of TTC30A, BOC, TSC1 and CLUAP1 genes showed an opposite effect in gr_2 and gr_3 (estimates= 0.05 and -0.06, P =$2.94e-4$ and $4.47e-6$), but no significant difference in gr_1, nevertheless whether this could be connected to differences in axon prolongation needs further investigation. Considering pathways that include at least one gene in MHC, the number of associations drastically increased to 880 (Fig. 4.54C), with gr_1 having an overall opposite WMW estimation sign than SCZ Z-statistic from PALAS, once again connected to a lower genetic liability in SCZ risk driven by specific gene configuration. We detected differences in immune related pathways such as *regulation of immune response* measure in DLPC, with positive estimates for gr_1 (estimate= 0.5, P= $2.57e-189$) as opposed to gr_2 and gr_3 (estimates <-0.16, P$<1e-33$) and similar but with opposite sign for *Notch signaling pathway*. In addition, *voltage-gated chloride channel activity* in cerebellar hemisphere and *axonogenesis* in DLPC were significantly decreased in gr_1 (estimates <-0.08, P$<4e-7$). Finally, the *autophagy* gene-set in frontal cortex BA9 was associated with all 3 groups (P$<6.07e-96$) but with a different configuration, having both gr_1 and gr_2 with negative estimates (estimates <-0.36) and opposite to SCZ Z-statistic PALAS (Z-stat= 2.4). Of note, *autophagy* gene-set in frontal cortex is composed of 5 genes among which only ZKSCAN3 is a cluster-specific gene having concordant WMW estimates between gr_1 and gr_2 (Fig. 4.51). Autophagy is indeed essential for neuronal survival and alterations in the mechanisms can lead to neuronal death and neurodegeneration [241].

Subsequently, we investigated whether the detected groups of SCZ patients led to endophenotypic changes. Different from CAD for which we were able to use the deep phenotyping information collected from UKBB, PGC cohorts did not include any phenotype details. Hence, we formulated a strategy to mimic the UKBB SCZ-related phenotypes into PGC patient cohorts via Gene risk-score (gene-RS), creating an approximation of the actual phenotype that was not collected. The gene-RS computed at the individual level was estimated from UKBB data set with the same concept of polygenic risk-scores, however built from imputed gene expression (see section 3.3.4 for details). Thus, we utilized gene-RS as phenotype proxy and we tested for group-specific differences across all 35 PGC2 cohorts with the same approach used for CAD and correcting in the endophenotype model for PCs. In section 4.4.7 details are shown on the derived Cluster Reliable Measure (CRM) that was evaluated on CAD for which we had available endophenotypes. This measure is devised to define a set of highly reliable endophenotype group-specific associations that are highly likely to be observed in the corresponding measured phenotype. Thus, we tested $1,000$ SCZ-related endophenotypes in the form of gene-RS spanning 27 categories (Tab.

B.11) and corrected for the same 10 PCs used in TWAS and PALAS analyses of SCZ. We thus identified 72 significant (FDR ≤ 0.05) and reliable cluster-specific endophenotypes (based on CRM) that differ in at least one group of SCZ patients (Fig. 4.55B).

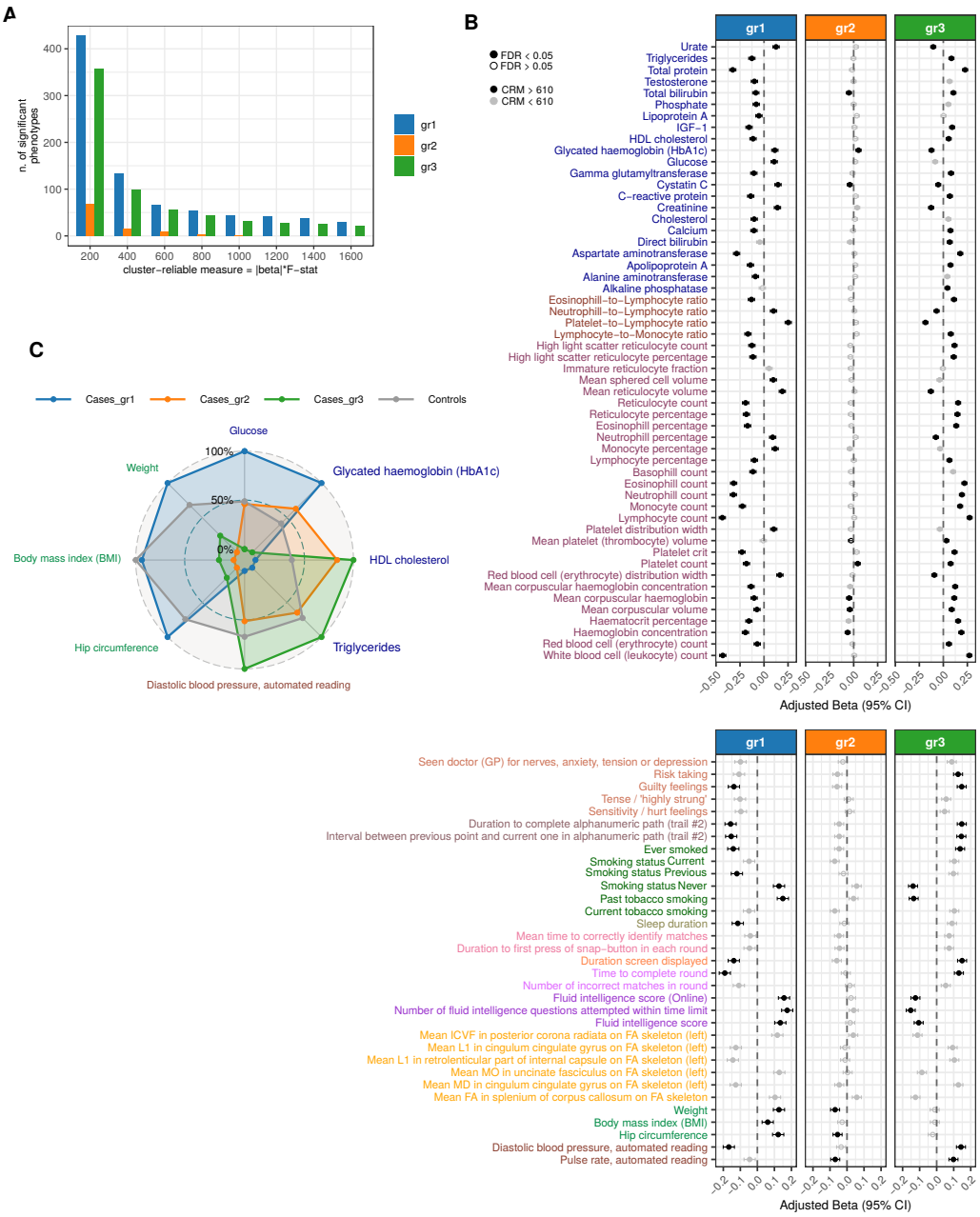

Fig. 4.55.: (Adapted from Trastulla et al., in prep.) **(A)** Number of significant gene-RS endophenotype differences in each group with CRM higher than the threshold displayed on the x-axis. **(B)** Forest plot of gene-RS endophenotypes with FDR ≤ 0.05 and CRM > 500 in at least one group, indicating regression coefficient with 95% CI for the grouping variable (β_{GLM}). Full dot means that β_{GLM} is significant after BH correction performed separately for each group across all the endophenotype, black dot means that the group-specific association is also reliable based on CRM threshold (610). Top panel is specific for blood count and blood biochemistry UKBB phenotype classes. **(C)** Group-specific radar chart for Metabolic Syndrome. Mean value of group-specific gene-RS endophenotype related to metabolic syndrome across all cohorts. Grey radar chart refers to all control combined in PGC cohorts. In each SCZ group plus controls is rescaled to 0-100 range.

In general, the choice of CRM threshold was obviously changing the final number of

reliable cluster-specific associations, decreasing it as the CRM increased and hence was more stringent (Fig. 4.55A). We chose 610 as the threshold leading to a precision > 0.85 when benchmarked on CAD (see section 4.4.7), which retrieved in SCZ application on DLPC 134 associations, 67, 10 and 57 respectively across the 3 groups. Generally, patients in gr_1 showed evidence of a lower SCZ severity considering phenotypes representing inflammatory states and cognitive functions. In particular, leukocyte counts were reliably and significantly lower ($\beta = -0.43$, P$= 4.11e-145$, CRM $= 6215$), and generally by any immune system component in term of counts, from basophill to lymphocytes (Fig. 4.55B). Nevertheless, when considering percentages and not absolute counts, neutrophill and monocyte percentages were increased in gr_1 ($\beta > 0.09$, P$< 1e-07$, CRM > 1054) and consequently the neutrophill-to-lymphocyte ratio (NLR, $\beta = 0.1$, P$= 1e-9$, CRM $= 1197$) and platelet-to-lymphocyte ration (PLR, $\beta = 0.25$, P$= 1.1e-50$, CRM $= 4025$). The general lower inflammatory state of gr_1 is also observed from decreased C-reactive protein liability ($\beta = -0.14$, P$= 1.9e-16$, CRM$= 1812$). Although still under debate, a large meta-analysis concluded that C-reactive protein is increased in SCZ patients compared to healthy subjects [242]. However, the same was deducted for NLR, showing evidence of increase in SCZ compared to healthy controls [243] and a correlation with SCZ severeness in drug-free patients [244]. However, this is in contrast with our observations. NLR inferred score in the form of gene-RS was significantly increased in controls compared to all the other cases ($\beta = 0.05$, P$= 1.8e-8$) at the limit of the CRM imposed threshold (CRM $= 605$) and gr_1 distribution of NLR was not significantly different from controls (P$= 0.07$). In addition, gr_1 showed a decreased liability of developing depression or anxiety related disorders despite not passing CRM threhsold ($\beta = -0.099$, P$= 4.23e-09$, CRM$= 521$) and better performances in cognitive tests such as higher fluid intelligence (FI, $\beta = 0.13$, P$= 1.4e-15$, CRM$= 848$), lower necessary time in pairs matching for testing visual memory ("Time to complete round" $\beta = -0.19$, P$= 8.94e-30$, CRM$= 1260$) or trail making lower time for testing executive function ("Duration to complete alphanumeric path (trail #2)" $\beta = -0.16$, P$= 1.1e-20$, CRM$= 740$). Finally, despite not passing the fixed CRM threshold, we found significant diffusion brain magnetic resonance imaging such as "Mean FA in splenium of corpus callosum on FA skeleton" significantly increased in gr_1 ($\beta = 0.120$, P$= 8.29e-10$, CRM$= 436$) and with an opposite effect in gr_3. This was validated by a fractional anisotropy study for different brain regions of interest that found significantly decreased regions in SCZ compared to healthy controls such as the aforementioned in splenium of corpus callosum [245]. Hence, we concluded that gr_1 exhibited a "healthier" state reflective of the lower SCZ genetic liability (Fig. 4.50E), nevertheless with impairment such as NLR charahcterstic of SCZ patients that require further investigations. On the other hand, gr_3 was complementary to gr_1 representing a more pathological state with higher inflammatory markers and cognitive dysfunction, while gr_2 was associated with a lower number of gene-RS endophenotypic differences and overall represented an intermediate state (Fig. 4.55B). Furthermore, we observed that individuals in gr_1 were characterized of a higher predisposition to metabolic syndrome (MetS) than the other groups and controls (Fig. 4.55C) with a specific configuration of

the 5 risk factors necessary to define the disease [246]. In particular, gr_1 compared to all the other patients had reduced HDL estimates ($\beta = 0.12$, P$= 4.13e-13$, CRM$= 1485$), elevated glucose ($\beta = 0.10$, P$= 3.8e-10$, CRM$= 727$) and elevated hip circumference ($\beta = 0.12$, P$= 4.13e-13$, CRM$= 1485$) as well as BMI ($\beta = 0.06$, P$= 3.1e-04$, CRM$= 765$), although waist circumference would be the preferred measurement for the definition. In addition, significant impairment persisted when comparing gene-RS results on gr_1 with gene-RS across all controls for glucose and HDL cholesterol (P$< 7.4e-3$) although not passing CRM threshold (CRM> 472) and not satisfied for hip-circumferences (P$= 5.7e-2$). In addition, we found a significant decrease in insulin-like growth factor-1 ($\beta = -0.16$, P$= 2e-20$, CRM$= 2192$) for gr_1, in concordance with observed deficiency in patients with metabolic syndrome [247]. Interestingly, SCZ patients have a two-fold risk or more of developing MetS compared to the general population, pointing at pleiotropic genetic factors [248].

Fig. 4.56.: (Adapted from Trastulla et al., in prep.) **(A)** Meta-analysis of generalized linear model testing group-specific differences in gene-RS for a subset of cognitive performance phenotypes. Each dot shows the adjusted regression coefficient which is full if FDR ≤ 0.05 and color according CRM threshold: the darkest grey, the most reliable the results. **(B)** Group-specific radar chart for cognitive performance phenotypes. Mean value of group-specific (and controls) gene-RS endophenotype related to cognitive performances phenotypes, rescaled to 0-100 range. Color code refers to the cognitive test classes in (A).

Given the relevance of cognitive performances in SCZ [238], we specifically focused on gene-RS mimicking cognitive endophenotypes registered in UKBB (Fig. 4.56). In particular, 28 cluster-specific changes were significant at FDR ≤ 0.05 but only 6 of this association

also passed the CRM threshold of 610, mostly due to the reduced power in term of sample size from UKBB registered phenotypes (Fig. 4.56A). As previously mentioned, gr_3 showed evidence of greater cognitive impairments compared to the other groups, for instance in term of

- fluid intelligence: "score", $\beta = -0.11$, P$= 6.5e-15$, CRM$= 665$;
- executive function registered in the form of trail making: "Duration to complete alphanumeric path (trail #2)" $\beta = 0.15$, P$= 1.3e-27$, CRM$= 694$;
- visual memory in the form of pairs matching test: "Time to complete round" $\beta = 0.13$, P$= 5.8e-22$, CRM$= 860$;
- prospective memory from ordinal phenotype with 1=correct at first attempt, 2=correct at second, 3=wrong, $\beta = 0.08$, P$= 4.3e-10$, CRM$= 339$;
- complex processing speed in the form of symbol digit substitution test: "Duration to entering the value", $\beta = 0.08$, P$= 2.4e-08$, CRM$= 63$;
- and processing speed in form of reaction time: "Mean time to correctly identify matches", $\beta = 0.07$, P$= 5.7e-08$, CRM$= 508$.

Most importantly, these results were still valid when comparing gr_3 with gene-RS in healthy controls (Fig. 4.56B, P$< 6.8e-10$). On the other hand, an opposite effect was reached for individuals in gr_1 in the cluster-specific comparison (P< 0.015). Similarly, gr_2 had opposite but milder effects compared to gr_3, with the exclusion of fluid intelligence, pairs matching and symbol substitution tests, none of which reaching significance after correction (P> 0.06). For these 3 specific examples, we observed that both gr_2 and gr_3 were impaired in the sense of lower performances when comparing with healthy controls (Fig. 4.56B) with P$< 6.7e-10$, excluding "Duration to entering value" between gr_2 and healthy controls (P$= 0.3$). In addition, gr_2 showed lower performances in term of working memory measure via numeric memory test ("Maximum digits remembered correctly", $\beta = -0.06$, P$= 1.37$e-05, CRM$= 263$) and even lower than healthy subjects ($\beta = 0.06$, P$= 3.4$e-06, CRM$= 387$). We also noticed that FI score is comparable between gr_1 and healthy controls (P$= 0.66$), however different performances were registered for the 13 tests of which FI score is a summary, with controls having better performances versus all the clusters for FI 3,4,6,7,8,10,12 (Fig. 4.56B). We hence highlighted that each of the identified group tend to specific cognitive impairments, with gr_1 and gr_3 at the extremes from a general higher to lower performances. In light of the other SCZ related endophenotypes such as inflammatory markers, we can conclude that gr_3 represented the group with greatest liability of severeness (Fig. 4.55B). Nevertheless, gr_1 patients are more at risk of developing metabolic syndrome (with a specific configuration of symptoms) and having higher neutrophill-to-lymphocyte ratio, whereas gr_2 showed specific cognitive impairments such as working memory not detected in any other group.

Summarising, we found distinct subgroups of SCZ patients arising from differences

in genetic configurations that exhibited different molecular mechanisms converging to distinct clinical endophenotypes and SCZ-related characteristics.

4.4.6 Patients stratification in DLPC reducing MHC contribution

Because MHC locus was the major driver in the DLPC clustering (92% of the cluster-specific genes), we decided to reduce its contribution via a strict filtering strategy that only kept genes with correlation < 0.1. In the specific context of DLPC, this led to a clustering structure discordant but not completely different to the previous one, with NMI $= 0.05$ (Fig. 4.49B). Leveraging 2571 genes with a pairwise correlation not exceeding 0.1, the patients were now partitioned in 4 groups that included from 2% to 43% of the total SCZ affected individuals. We identified 636 cluster-specific genes (FDR≤ 0.01) across all tissues and (288 unique) that span 151 tissue-specific loci for a total of 59 loci, the majority of which (47) associated to gr_2 (Fig. 4.57A). Different from the previous configuration, 69% of the total significant genes were located in the MHC locus, mostly affecting gr_3 and gr_4 (Fig. 4.57B). Instead, gr_2 configuration was dependent more significantly on genes located outside MHC locus. Of note, this cluster of individual was the smallest in the detected partition (composed of 525 samples) that however did not generalize when projecting into the external cohort scz_boco_eur, as no sample was predicted to belong to that group (Fig. 4.58A). On the other hand, the remaining group exhibited a high consistency in the external validation, with all 6 loci associated to gr_1 replicated in sign concordance, and 6 and 5 out of 8 loci replicated for gr_3 and gr_4 respectively (Fig. 4.58B) and a correlation > 0.8 in terms of cluster-specific gene effects (Fig. 4.58C). Looking closely at genes driving the clustering structure (Tab. B.10), C2orf47 located in 2q33.1 is associated with all groups but gr_2, corresponding to an increase in gr_3 and gr_4 and a decrease for gr_1. This gene involved in calcium import in the mitochondrion, is located in one of most significant hit outside MHC for SCZ and was strongly negatively associated with the disease from our TWAS analysis (Z-stat $= -7.6$ in DLPC). In addition, genes in MHC locus were significantly different across all group. BTN3A3 was the exemplar gene in the locus with strongest WMW estimates and was actually included in the DLPC genes used to obtained the clustering structure. Being negatively associated with SCZ (Z-stat$= -8.35$), the group-specific distribution of BTN3A3 were negative for all clusters but gr_3, indicating a "healthier" state for gr_3 considering simply this gene and not the general MHC locus configuration. Finally, gr_2 and gr_4 are significantly associated with MPHOSPH9 in 12q24.31 and DOC2A in 16p11.2, in both cases with gr_2 and gr_4 individuals showing a higher expression and lower expression respectively (Tab. B.10).These two exemplar genes were both associated with SCZ from our TWAS (P$= 4.2e-07$), but with an opposite effect. In general, a specific configuration of genes defined the clustering structure, located in multiple loci and without a clear direction towards SCZ "more severe" or "more healthy-like" states.

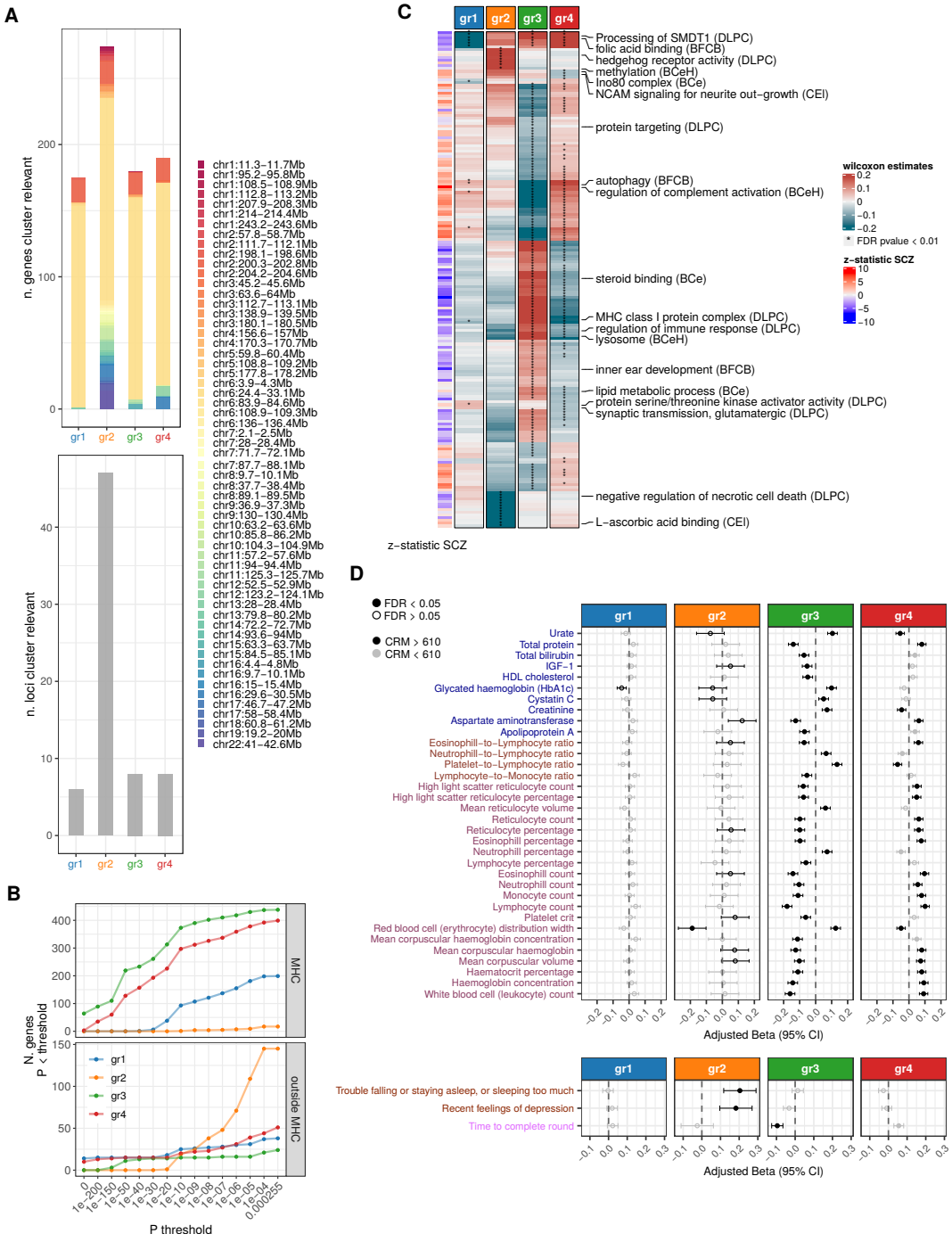

Fig. 4.57.: (From Trastulla et al., in prep.) **(A)** Number significant genes and loci (tissue specific FDR ≤ 0.01) associated with each group from WMW test of group id versus remaining patients combing all tissues. **(B)** For each group, number of group-specific genes (y-axis) passing the WMW p-value threshold (x-axis), divided per genes intersecting and outside MHC locus (top and bottom panels, respectively). **(C)** WMW estimates (capped) for 74 significant pathways (rows) in each group versus the rest test (column) considering only the most significant tissue per-pathways when repeated. The names on the row are a selection of significant pathways, parenthesis refers to the tissue considered (acronyms indicates the initial of the tissue name). Row annotation on the left refers to the corresponding SCZ Z-statistics from PALAS. **(D)** Forest plot of gene-RS endophenotypes with FDR ≤ 0.05 and CRM > 610 in at least one group, indicating regression coefficient with 95% CI for the grouping variable (β_{GLM}). Full dot means that β_{GLM} is significant after BH correction performed separately for each group across all the endophenotype, black dot means that the group-specific association is also reliable based on CRM threshold (610). Top panel is specific for blood count and blood biochemistry UKBB phenotype classes.

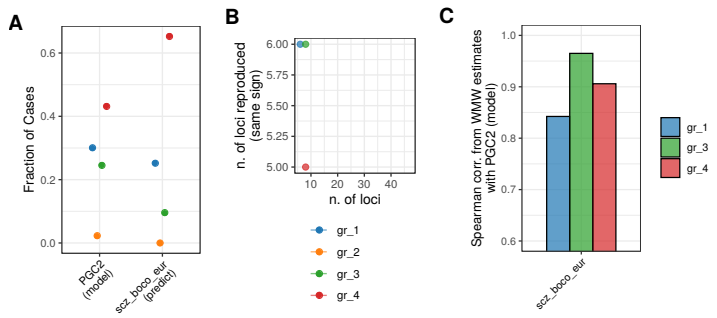

Fig. 4.58.: (Adapted from Trastulla et al., in prep.) Prediction of DLPC clustering structure with genes $|\text{corr.}| < 0.1$ filtering on scz_boco_eur external cohort. **(A)** Y-axis indicates the fraction of SCZ patients assigned to each group in PGC2 data set and the left-out cohort for which the clustering structure was projected. **(B)** Reproducibility of group-specific loci on predicted groups in external cohort, the x-axis shows the number of loci across all tissues associated with each group in PGC2 data set, the y-axis shows how many of these loci have the same sign using as exemplar the strongest association of the WMW estimates in the predicted clustering structure. **(C)** Spearman correlation of WMW estimates in PGC2 and the external cohort only from genes that are significantly associated with that group (considering all tissues).

Of note, the individuals partition was only mildly associated with PCs distribution (Fig. 4.59A), lower than what we observed in the previous configuration with 0.9 genes correlation threshold (Fig. 4.53A). This might be related to the reduced contribution from MHC locus whose variability was partially captured in PCs. However, ancestry in form of PCs clearly did not drive the clustering configuration compared to the genes and pathway contributions.

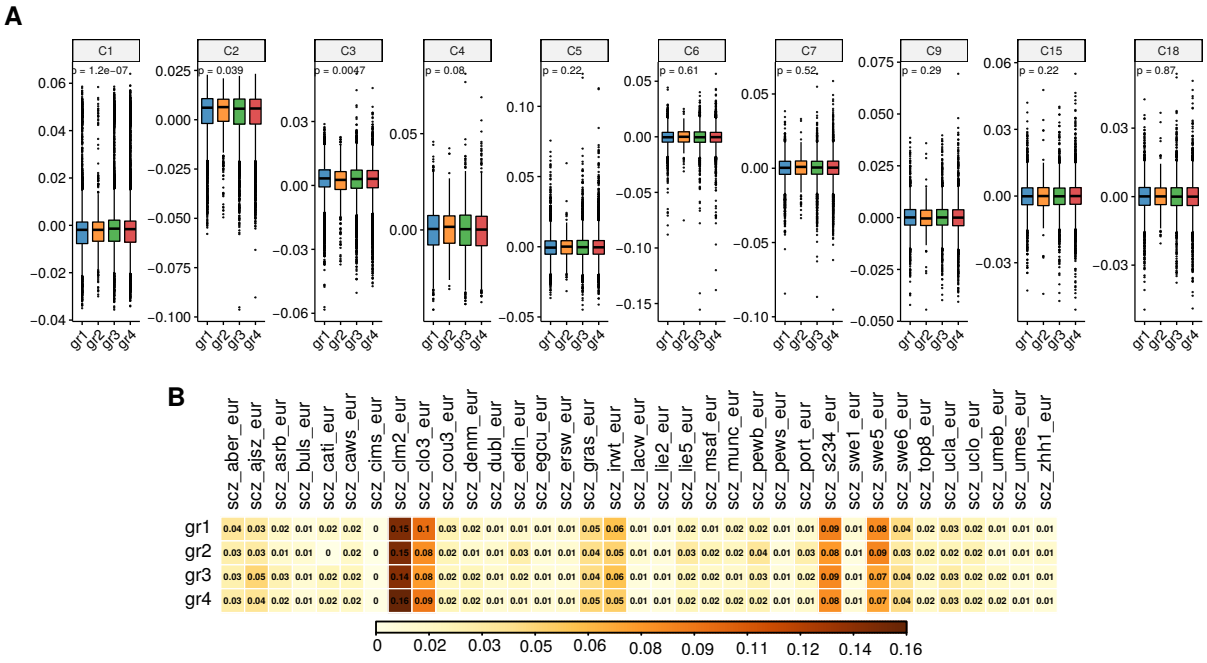

Fig. 4.59.: **(A)** Distribution of PGC2 10 PCs (same as those used in TWAS and PALAS) for each SCZ DLPC cluster filtering genes at $|\text{corr.}| < 0.1$ (p-values from Kruskal-Wallis test). **(B)** Contingency table of group and cohort structure with each square referring to the fraction of patients in a group (rows) belonging to a certain cohort (columns).

This effect was also present in the form of cohort information, for which there was a

significant association with clustering structure (χ^2 P$= 2.1\text{e}-09$) due to an enrichment of certain groups such as gr_3 in scz_ajsz_eur (Fig. 4.59B), nevertheless without being a driver in defining patients partition.
The differences in gene expression converged into differential pathway associations, for a total of 418 group-specific and tissue-specific significant associations (FDR ≤ 0.01), among which 407 were concordant across tissues in case of repetition (21, 23, 214 and 149 respectively in gr_1 to gr_4) and finally collapsed into 180 unique pathways across tissues (Fig. 4.57C). Differences in pathways such as *Processing of SMDT1* and *folic acid binding*, with an increase in gr_3 and gr_4 and a decrease in gr_1 in the same direction of SCZ liability were related to differences in C2orf47 locus. Instead, differences in pathways like *autophagy*, *regulation of complement activation* and *regulation of immune response* were related to differences in the MHC locus, with individuals in gr_3 at a lower risk considering PALAS results. We also detected milder differences in gr_4 not related to MHC locus such as *synaptic transmission, glutamatergic*, with the highest contribution from ALS2 gene (WMW est $= -0.05$, P$= 2.5\text{e}-06$). Finally, the highly variable gene configuration of gr_2 led to two major block of pathways significantly different with respect to all the other groups: increased effect such as in *hedgehog receptor activity* and a decreased effect such as in *L-ascorbic acid binding*. The first example is a regulator of oligodendrocyte production as well as dopaminergic neuron development [249], whereas the second example is related to vitamin C absorption.

To investigate the effect at the endophenotypic level, we applied the same strategy described before and detected gene-RS differences across groups for SCZ related endophenotypes (Fig. 4.57D). Similarly to gr_1 in the context of 0.9 correlation threshold (sec. 4.4.5), gr_3 was exhibiting a lower SCZ liability due to MHC locus genes. Conversely, this led to lower values of inflammatory markers such as leukocyte count ($\beta = -0.16$, P$= 3.2\text{e}-25$, CRM$= 2314$) and better cognitive performances in term of visual memory test, i.e. pairs matching test "time to complete round" ($\beta = -0.1$, P$= 6\text{e}-10$, CRM$= 629$). Nevertheless, results observed before such as fluid intelligence score increase and C-reactive protein decrease were not significant and reliable in this setting, although with a similar trend to 0.9 configuration ($\beta = 0.05$ and -0.04, P$= 0.001$ and 0.01, CRM$= 306$ and 500 respectively). In addition, gr_4 represented a group at higher risk of SCZ, opposite to gr_3, with increased leukocyte count ($\beta = 0.9$, P$= 8.6\text{e}-12$, CRM$= 1322$) and lower memory performances however not reliable ($\beta = 0.05$, P$= 3.6$e-05,CRM$= 364$). Finally, although showing differential genes and pathway distribution, there was no significant and reliable endophenotype in gr_1, while the discovery was limited to 3 in gr_2 due to a reduced sample size. Interestingly, we found that gr_2 had a higher liability in depressive symptoms and sleeping related dysfunctions ("Recent feelings of depression" $\beta = 0.18$, P$= 3.6\text{e}-05$, CRM$= 652$). Notably, L-ascorbic acid binding associated with gr_2 has been found related to negative symptoms in SCZ patients, with vitamin C administration improving the PANSS score related to negative symptoms [250].

In summary, reducing the MHC locus contribution we detected groups strongly driven by genes outside that locus and consequently specific pathway configuration. Although we found similar endophenotypic differences with respect to what was shown in the previous paragraph, here we also detect a group with evidence of negative symptoms only, without additional cognitive impairments.
We conclude the chapter with two benchmarks, the first related to the gene-RS strategy to approximate endophenotype and the second to ancestry contribution to the observed clustering structure comparing DLPC groups with the clustering obtained from PCs.

4.4.7 Gene risk-score to approximate endophenotypes

Because endophenotypes information was not available on PGC cohort, we approximated it via gene-RS as described in section 3.3.4. In particular, we leveraged UKBB deep phenotyping and benchmarked our approach on CAD external cohorts CARDIoGRAM, projecting the clustering results from UKBB CAD partition and comparing the found differences in terms of gene-RS with the ground truth of the actually observed cluster-specific endophenotypes.
We first build gene-specific weights across 10 CAD tissues to compute gene-RS (i.e. Z-statistics) via TWAS for 369 CAD-related phenotypes in UKBB. Then, to evaluate the heritability of the phenotype in being explained by gene-RS, we computed gene-RS on the same samples for which the actual phenotype is available and from which gene weights were obtained (namely UKBB cohort). We finally estimated both R^2 and F-statistic on as described in section 3.3.4. Indeed, we observed that R^2 estimates were inflated for phenotypes registered in fewer samples (Fig. 4.60A, left-panel) and decided to evaluate the goodness of the model via F-statistic (Fig. 4.60A, right-panel). The same trend was observed for the 1000 endophenotype in 11 tissues chosen among UKBB phenotypes as SCZ-related to test differences in PGC clusters in the form of gene-RS (Fig. 4.60B). Afterward, we computed gene-RS also on the 9 external CARDIoGRAM cohorts, together with the projection of tissue-specific clustering based on UKBB CAD patients partition. Having a patient stratification and an approximation of the phenotypic characterization in each cohort, we were now able to estimate cluster-specific differences in term of gene-RS and we obtained a summary result via a meta-analysis. We finally computed the CRM for an endophenotype difference tested in a group as the product of the group-specific regression coefficient and phenotype F-statistic (3.43). Considering all tissues together and filtering for cluster-specific results passing FDR ≤ 0.05, we evaluated the performances of gene-RS in finding the same results as the actual phenotype in term of precision. In particular, having fixed a certain CRM (thr), the precision was computed as the fraction of cluster-specific gene-RS differences concordant in regression coefficient sign with the actual endophenotype difference, over the total number of gene-RS phenotypes passing a

certain CRM threshold:

$$\text{Precision}_{thr} = \frac{|\{\text{pheno}|\text{sign}(\beta_{pheno} \cdot \beta_{gene-RS}) > 0 \text{ AND } CRM_{gene-RS} > thr\}|}{|\{\text{pheno}|CRM_{gene-RS} > thr\}|} \quad (4.1)$$

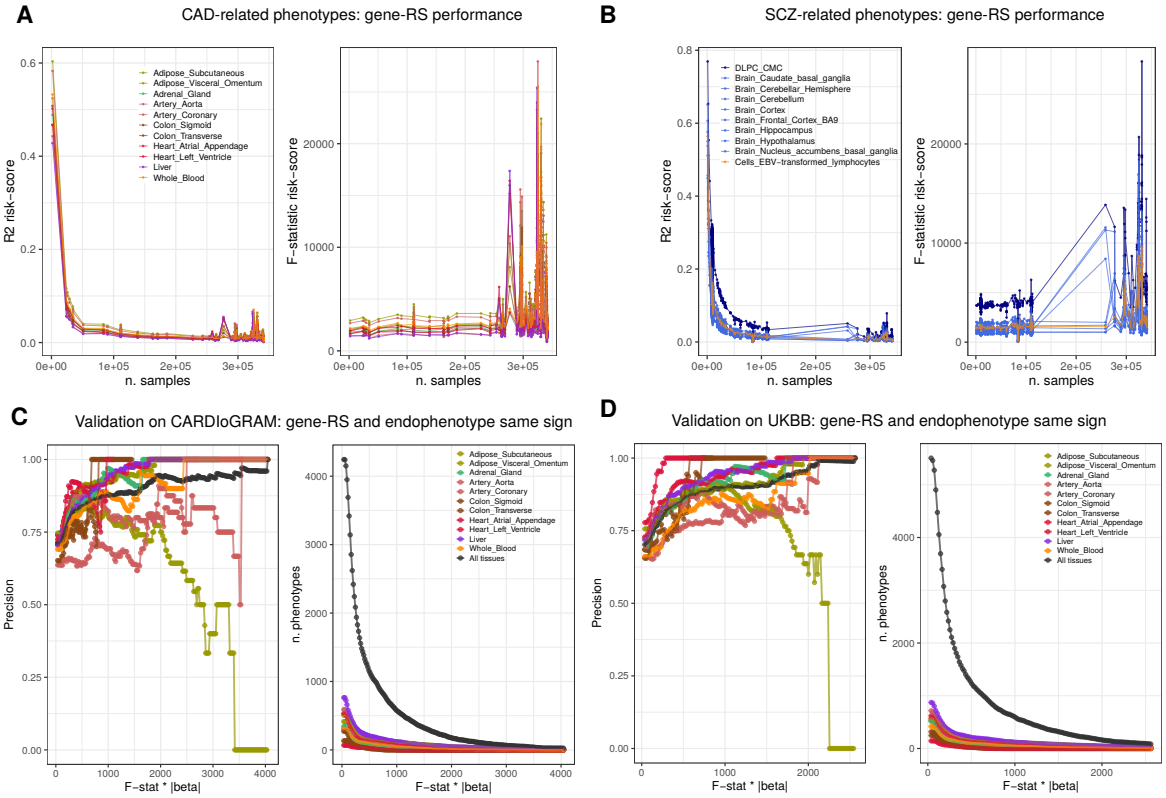

Fig. 4.60.: **(A-B)** Left panel shows R^2 estimates for gene-RS (y-axis) in predicting the actual UKBB phenotype versus the number of available individuals for that phenotype (x-axis). R^2 is computed from nested model subtracting R^2 evaluated on the complete model from R^2 evaluated on covariates only (phenotype specific PCs as used in TWAS and PALAS). Right panel shows F-statistic from nested model (y-axis) compared to the number of samples (x-axis) later used as a component of cluster-reliable measure (CRM). **(A)** Distribution of R^2 and F-statistic for 369 CAD related endophenotypes in 11 tissues and **(B)** for 1000 SCZ related phenotypes in 10 tissues. **(C-D)** CRM validation in CAD: meta-analysis gene-RS differences built on **(C)** 9 CARDIoGRAM cohorts or **(D)** UKBB, the first case being external cohorts and the second case being the same samples used to build TWAS coefficients and to estimate R^2. CRM on the x-axis is compared to the actual endophenotypic differences detected in each tissue-specific clustering. Y-axis of the left panel shows the precision computed as the fraction of group-specific gene-RS differences that have the same sign of GLM regression coefficient in the actual endophenotype analysis, among all the endophenotypes passing CRM threshold. Y-axis of the right panel shows the number of phenotypes considered having group-specific reliable difference with that threshold. The black line represents a summary combining all tissues.

The idea behind this strategy is to decide a cut-off above which it might be possible to trust the results from gene-RS as reliable and likely to happen also for the actual endophenotype. Results for varying CRM thresholds are shown in Fig. 4.60C (left panel) as well as the final number of retrieved associations when considering that threshold (right panel). From all tissues together, the CRM cut-off of 610 leads to a precision higher than 0.85 and a total of 1002 reliable and significant cluster-specific associations. Nevertheless, even a much

lower threshold of 265 would have a precision > 0.8, although increasing the number of associations to 1828 and hence the uncertainty. Of note, we could also compare the cluster-specific results from gene-RS and endophenotype as computed and registered in the same set, namely UKBB (Fig. 4.60D), observing a very similar trend in terms of the number of associations retrieved at a given cut-off.

4.4.8 Ancestry contribution to clustering

Similarly to CAD (section 4.3.7), we investigated whether the clustering structure detected for SCZ patients was emerging from ancestry differences. Although we corrected for PCs as pre-processing step, there still was a significantly different distribution among groups in PCs from 1 to 4 (Fig. 4.53A). Thus, we specifically studied the overlap between tissue-specific partitions (genes $|\text{corr.}| < 0.9$) and the clustering structure obtained from the available PCs in PGC cohorts (from 1 to 20) via PhenoGraph algorithm (section 3.3.1), prior to feature standardization of each PC to mean 0 and standard deviation 1. We thus identified 9 groups (Fig. 4.61A) with a significant overlap to the partitions detected in tissues (Fig. 4.61B left, χ^2-test P < 1e-10). This was expected from the PCs distribution in DLPC clustering, nevertheless we can here appreciate the extent of overlap via NMI that was extremely reduced and did not exceed 0.005 (Fig. 4.61B right). Focusing on DLPC tissue, the small NMI from the comparison with PCs clustering (0.002) was still always higher than the NMI between PCs cluster and $10,000$ random partitions (Fig. 4.61C), hence sharing a structure not emerging by chance. Pairwise odds ratio of enrichment from Fisher's Exact test between groups from DLPC and groups from PCs (Fig. 4.61D) found 10 pairs with significant enrichment or depletion (P< 0.01) with the strongest result for enrichment of DLPC gr_2 in PCs gr_5 (P$= 8.9\text{e}-15$) and consequential depletion of DLPC gr_1 in PCs gr_5 (P$= 1.2\text{e}-11$). Giving the evidence of a sharing structure among the two partitions, we finally investigated whether it would imply a similarity in endophenotypic characteristic for each group. As for CAD, we considered for each endophenotype tested (in the form of gene-RS) the best p-value result across all DLPC groups and all PCs groups (Fig. 4.61E). The level of associations greatly varied in magnitude between DLPC and PCs cluster, nevertheless we identified 6 endophenotypes passing FDR 0.05 threshold in both partitions. Relaxing the FDR threshold to 0.1, we observed in details which group had the highest association with the considered endophenotype differences in both partitions (Fig. 4.61F). We observed that platelet crit is significantly decreased in DLPC gr_1 and PCs gr_2, nevertheless there is no enrichment between these two groups, while PCs gr_1, enriched in DLPC gr_1, showed an opposite effect ($\beta = 0.02$, Fig. 4.61D). LDL direct and apolipoprotein B are decreased in DLPC gr_1 as well as PCs gr_3 but there is no evidence of individuals overlap and similarly for the diffusion magnetic resonance imaging (dMRI) phenotypes "Mean L1 in fornix cres+stria terminalis on FA skeleton (left)", "Mean L3 in cingulum hippocampus on FA skeleton (right)" and "Mean MD in fornix cres+stria terminalis on FA skeleton (left)".

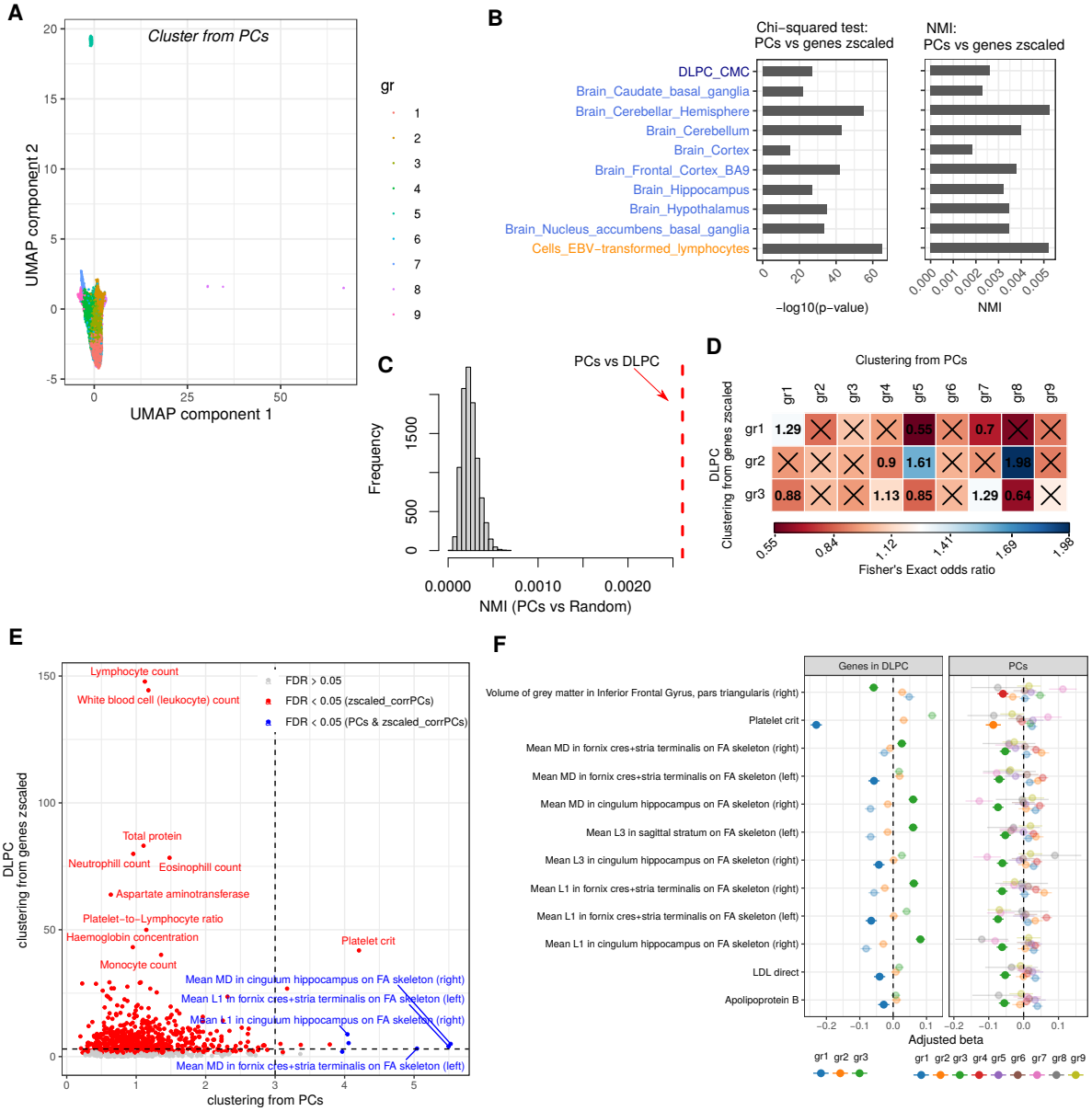

Fig. 4.61.: (Adapted from Trastulla et al., in prep.) **(A)** UMAP of SCZ cases (outliers from genes excluded) based on the first 20 PGC PCs (standardized), color refers to the assigned PCs clustering. **(B)** Comparison between PCs clustering and grouping from gene T-scores corrected for PCs and normalized ("genes zscaled") in term of -$\log_{10}$ p-value of χ^2-test (left) and NMI (right), for each tissue. Genes are clumped at the default threshold of 0.9 correlation. **(C)** Histogram of NMI between cluster from PCs and $10,000$ randomly assigned groups with the same size as DLPC clustering (genes corr. < 0.9), the dashed line refers to the NMI comparing PCs and the actual DLPC clustering. **(D)** Pairwise Fisher's Exact test between a group detected in PCs clustering (columns) and a group detected in DLPC clustering (rows), heatmap indicates the calculated odds ratio with $\times$ highlighting a non-significance at the nominal level of 0.01. **(E)** Each dot represents a tested endophenotype and indicates the -$\log_{10}$ p-value of the most significant group-specific difference in PCs (x-axis) and DLPC (y-axis) clustering. Dashed lines refer to p-value $= 0.001$ and color reflects the FDR significance threshold. **(F)** Forest plot of group-specific differences for endophenotypes significant in PCs cluster at FDR 0.1 threshold, x-axis shows the regression coefficient from GLM testing gr_i vs all remaining. Not shaded dots indicate groups with most significant association in term of p-value.

The remaining dMRI phenotypes instead, included as most significant groups gr_3 of both

DLPC and PCs clusters but with opposite effect that again were not overlapping. However, PCs gr_4 was enriched in DLPC gr_3 and was concordant in those phenotype associations, similarly to "Volume of grey matter in Inferior Frontal Gyrus" but with a significantly decreased distribution in PCs gr_4 and DLPC gr_3.

In conclusion, tissue clustering structures include some effect from ancestry information that cannot be removed via PCs correction during pre-processing. This effect was stronger than what we observed for CAD, possibly due to the higher relevance of MHC locus in SCZ of which PCs also capture the variability. Nevertheless, the contained overlap of DLPC stratification with PCs clustering does not drive the endophenotypic differences observed in the form of gene-RS, and could only interfere with 6 phenotypic differences due to an enrichment of individuals in DLPC gr_3 with individuals in PCs gr_4, hence here limited. We conclude that the emerged stratification represent a genetic liability in SCZ rather than the individuals ancestry background.

5 Discussion

The work presented in this thesis represents to our knowledge the first comprehensive pipeline that creates cis-regulatory models to impute gene expression, converts individual genetic associations into meaningful biological entities, and stratifies patients based on their genetic liability profiles. These interpretable features can be leveraged to identify putative causal or protective pathomechanisms and intermediate phenotypes, directly suggesting testable hypotheses on the underlying biological mechanisms. Moreover, the individual level imputed gene expression profiles are used for patient stratification and reveal distinct molecular pathways activity profiles across the inferred patient groups and as well as differences in endophenotype profiles and clinical outcomes.
During the application to coronary artery disease and schizophrenia, we found well-known biological pathways involved in each disease as well as less recognized or putative novel mechanisms. Interestingly, we revealed an aggregation mechanism of small effect genetic variants onto specific pathways, leading to impairment at the individual level that drives the disease association. Finally, the detected groups based on distinct genetic liability profiles were associated with differences in endophenotype profiles and treatment responses. These groups were also interpreted in the context of biological pathways, decomposing each group scenario in a configuration of perturbed pathways, characteristic of the different observed pathomechanisms.
In conclusion, this pipeline will be a valuable tool for the scientific community to obtain insights into a complex disease pathomechanisms and propose ad-hoc treatment strategies for genetically different groups of patients.

5.1 Integration of epigenetic information to model gene expression

The first module of our integrated pipeline constructs tissue-specific gene regulatory models from cis-variants. From a modeling perspective, PriLer is an extension of elastic-net regression that additionally integrates a priori knowledge on variants, reducing the penalty for SNPs that are more likely to be relevant from a biological point of view. Theoretically, the prior information of variants can be from any biological source. However, we used epigenetic information and GWAS associations due to their relevance in gene expression regulation. The overall importance of a variant as the result of the corresponding prior information is weighted by the relevance of each prior feature (e.g. cell type-specific

open chromatin states). PriLer does not assign a fixed relevance to prior features but automatically learns it in an iterative procedure that takes into account the regulatory role of the corresponding intersecting variants in gene expression (Fig. 3.3). This methodology is integrated into the CASTom-iGEx framework. Nevertheless, it can also be used as a stand-alone tool for creating gene expression models from common cis-variants in matching reference panels, with a customizable cis-window around genes TSS.

In this thesis, we initially created gene regulatory models from GTEx v6p [166] and CMC Release 1 [54] reference panels, without harmonizing the genotype data sets with any other target genotype-only data set, to evaluate PriLer at the best of its capability and providing the biggest possible set of variants among which the gene expression can be modeled. The evaluation is based on in-sample and out-of-sample R^2 metrics referring to the variance explained by the genetic component, the former computed on the entire available set and the latter as the average across CV-folds (see 3.1.4). As expected, the number of genes that can be reliably predicted from their cis-genetic effects ($R^2 \geq 0.01$ and $R^2_{CV} > 0$) depended on the tissue sample size, but not the number of reg-SNPs, i.e. SNPs that regulate at least one gene (Fig. 4.3), highlighting a complexity specific for each tissue regardless the number of individuals available. This was indeed confirmed in a down-sampling analysis of DLPC tissue from CMC, in which the number of reliable genes increased with the sample size but the n. reg-SNPs remained stable, apart from a clear overfit when an extremely low sample size was considered (50 individuals) (Fig. 4.4). In addition, in-sample R^2 increased with the tissue sample size, indicating an expected overfit when evaluating the performance on the entire set of samples that is however not observed in the left-out partitions (Fig. 4.5). Strikingly, out-of-sample R^2_{CV} increased with the sample size when considering the common set of genes in DLPC (Fig. 4.6). This further validates that an increase in power through an increase in the number of individuals leads to a more accurate estimation of the gene expression regulation (Fig. 4.6).

Importantly, the integration of prior knowledge in the elastic-net framework gives insight into the (epigenetic) prior relevance in regulating gene expression in a tissue-specific context. Indeed, when simulating prior features via the random sampling of variants, the final correspondent weights computed by PriLer were the lowest and proximal to zero (Fig. 4.8). In addition, a similar outcome was observed when simulating prior features randomly sampling from observed gene regulatory regions (GREs) of different cell types, hence keeping a plausible biological structure. However, an almost null assigned prior weight was only registered when the considered GREs were not present in the prior features of the baseline model, thus actually biologically relevant (Fig. 4.9). This is connected to two critical aspects of the PriLer setup. Firstly, random prior features will obtain approximately null weights only when the overlap of variants with an actually relevant prior is almost null. Conversely, if the sharing is substantial, PriLer will detect that prior feature as important to a certain extent, because some of those variants will still regulate the expression of genes. Thus, to increase the interpretability of the results in

revealing gene regulation mechanisms, one possibility is to create a "shared" prior feature that includes the variants overlapping all the prior information and the complementary cell-type specific ones, excluding the shared portion. Secondly, the higher the number of variants intersecting a certain binary feature, the higher will be the computed weight, as was observed when increasing the random prior sample size. This happens because of the higher probability of intersecting a variant that is actually relevant for gene regulation in a prior feature with higher dimensionality. A strategy for proper rescaling based on the priors dimension can alleviate this problem. Addressing these issues is important to gain a better understanding of gene regulatory mechanisms. Nevertheless, this was not the focus of our study as we aimed at building improved gene expression models via the selection of more robust and meaningful variants. Crucially, PriLer reduces the penalization that a variant undergoes in each gene expression model based on a priori (properly weighted) information. However, it does not force the selection of those variants in the final gene expression model that will be determined by the specific gene context (Fig. 4.17, 4.40). Notably, it is crucial to include plausible regulatory mechanisms in the set of prior features. Prior feature weights will never be zero unless they do not intersect any reg-SNPs, despite PriLer assigning to random non-meaningful features a low weight. Hence, "non-plausible" features will still minimally contribute to the final gene expression models. This is a consequence of the L2 penalty used to control γ size (see (3.4)). To perform prior features selection, a natural extension is the introduction of an L1 penalty that will shrink non-relevant feature weights to zero and possibly improve the reg-SNP selection.

PriLer implementation includes a built-in elastic-net model for each gene, with the primary aim of finding an optimal setting of hyper-parameters α and λ (Fig. 3.5). Regardless, it can be also used to estimate the improvement reached when including prior information in the gene expression modeling. In particular, we noticed an increase in out-of-sample R^2_{CV} for the majority of genes across all tissues ($> 50\%$, Fig. 4.10B). This highlighted that the prior features are useful overall but not for certain genes whose regulation might be related to biological components not included as priors. Importantly, PriLer reached better performances than elastic-net while utilizing a reduced set but more biologically relevant of reg-SNPs (Fig. 4.10B-D). In addition, the selection of reg-SNPs aided by external meaningful information led to a more robust selection compared to elastic-net, as observed in a down-sampling analysis in whole-blood comparing Jaccard similarity of reg-SNPs selection (Fig. 4.10E).

Finally, PriLer was compared against the two state-of-the-art methods for modeling gene expression from cis-components: prediXcan [10] and Fusion [9], the latter method also called TWAS. Both methods are based on an expression modeling separately constructed for each gene, with prediXcan using elastic-net and TWAS choosing between the best performing model among cis-eQTL, BLUP (best linear unbiased predictor), BLSMM (Bayesian sparse linear mixed models), LASSO (least absolute shrinkage and selection

operator), and elastic-net (see section 2.2.3). In their application on GTEx v6p tissues and CMC data sets, PriLer outperformed both prediXcan and TWAS in terms of average cross-validation squared correlation, with a better variability explained by PriLer in more than 64% and more than 76% of the total genes, respectively (Fig. 4.12A-B). However, the higher improvement when compared to TWAS was also related to a higher number of reg-SNPs systematically used across all tissues. Conversely, about 50% of the gene models used a higher number of reg-SNPs in PriLer compared to prediXcan, and even less than half of them for specific tissues such as hippocampus and hypothalamus (Fig. 4.14), underlying an improvement in explained gene variability regardless the higher number of features selected. Notably, the TSS window we used in PriLer was particularly smaller than the ones used in prediXcan and TWAS, respectively 400kb size compared to 1Mb. Similarly to elastic-net, PriLer selected a higher percentage of biologically relevant reg-SNPs (Fig. 4.12C). In this case, we considered their intersection with the tissue-specific prior information in the form of gene regulatory regions and found prediXcan as systematically showing the smallest percentages. Finally, reg-SNPs biological relevance was estimated using an external gene regulatory reference not applied in the gene expression model, namely accessible chromatin zones measured as DNase I hypersensitivity sites [204]. PriLer reg-SNPs intersected an overall higher number of biosamples DHSs than elastic-net in the majority of tissues, but not compared to prediXcan or TWAS. Indeed, better performances in PriLer were only observed for a selection of tissues among which hippocampus, left ventricle, and muscle skeletal (Fig. 4.15). Notably, the fraction of TWAS reg-SNPs intersecting at least one DHSs biosample was always higher across all tissues (Tab. B.3), possibly related to the model choice in TWAS of "best eQTL" which are particularly enriched in regulatory regions [46].

In conclusion, our proposed methodology for gene expression modeling outperforms existing ones, leading to increase interpretability in terms of regulatory mechanisms and a more biologically meaningful and robust selection of variants. PriLer can be further extended including an L1 penalty term for prior weights sparsity to increase the interpretability of the learned prior relevance and predictions. In addition, trans-effect can be modelled together with cis-effects introducing chromatin three-dimensional interactions [251] and/or regulatory effects of transcription factors expression on target genes [48].

5.2 Gene expression perturbation by disease-related genetic mechanisms

The tissue-specific gene expression models from PriLer are then used to impute gene changes onto large-scale genotype-only cohorts. The obtained gene expression reflects the perturbation of transcription activity that is mediated by the cis-variants alleles configuration. Thus, it is possible to test the association of the genetically mediated expression with

the disease risk and identify putative causal genes. Indeed, TWAS associations cannot be considered as unquestionable causal genes as spurious results can arise from LD structure or shared GWAS hits with actual causal genes [62].

Here, we tested the association of imputed genes with disease status in the context of two complex diseases: coronary artery disease (CAD) and schizophrenia (SCZ). We identified many well-known genes from GWAS, some of them also validated via in-vivo experiments, underlying the fidelity of our analyses. For instance, the lower expression of SORT1 gene in the liver was significantly associated with an increased risk of CAD (Fig. 4.16A). Functional in-vivo experiments [25, 125, 206] confirmed that an increased expression of SORT1 gene in the liver and hence the production of sortilin protein, a lysosomal sorting protein, boosted the clearance of LDL in blood. This mechanism was mediated by the binding of intracellular APOB-containing particles in the Golgi apparatus and extracellular LDL in the plasma membrane. Thus, a decrease in SORT1 expression implies a reduction in LDL clearance and consequently a higher risk of CAD. Liver-specific SORT1 was not the only gene significantly associated with CAD in the 1p13.3 locus. Significant decreases of PSRC1 and CELSR2 imputed expression in the liver were as well associated with CAD risk. PSRC1 was also reliably imputed and significantly associated with CAD in whole blood, contrary to CELSR2 whose expression in visceral omentum, colon sigmoid and artery aorta did not lead to changes in CAD risk. These associations in the liver were indistinguishable from SORT1, with a correlation higher than 0.9. In future analyses, fine-mapping strategies such as FOCUS [72] can aid in discerning causality in a locus via posterior probability estimation. On the other hand, we also detected CAD-associated genes in well-known risk loci but without properly understood pathomechanisms, e.g. CDKN2B and PATCHR1. The region 9p21 of CDKN2B was long identified as a GWAS hit, with rs4977574 SNP increasing susceptibility 1.30 times for heterozygotes A/G and 1.54 times for homozygotes G/G configurations in Caucasian population [252]. This SNP is located in the 58-kb LD block on 9p21 which overall confers a markedly increased risk of CAD. Deletion of the orthologous 70-kb non-coding interval on a mouse model implied a decrease in expression of the nearby genes CDKN2B and CDKN2A, together with the proliferation of vascular cells and increased mortality, and pointed to an impairment in cell cycle and cellular proliferation mechanisms [207]. Alteration of CDKN2B and CDKN2A nearby genes is also mediated by altered expression of the non-coding RNA ANRIL, considered the target genes of 9p21 risk variants, nevertheless the exact mechanisms remain unclear [253]. In our TWAS analysis, both CDKN2A and ANRIL were not tested in any tissue as they were not reliably predicted in the GTEx reference panel, hence the mediation via ANRIL could not be proved. Instead, we found that the genetic component of CDKN2B expression mediated by 51 variants (even outside the 58-kb region) was negatively associated with CAD. This association was only present in the context of colon sigmoid and not whole blood, due to a tissue-specific regulation (Fig. 4.17A-B). Finally, we identified PHACTR1 as significantly associated in the artery aorta mediated by three cis-variants among which rs9349379, well-known from previous GWAS (Fig. 4.17C). Nevertheless, a recent study by Gupta et al. [254] applied CRISPR-editing technology to stem cell-derived endothelial cells targeting

rs9349379 and detected a gene 600 kb upstream of PHACTR1, EDN1, as transcriptionally regulated by rs9349379, whereas no effect on PHACTR1 gene expression was detected.

In the context of SCZ, we similarly retrieved associations established from previous GWASs, some of which were experimentally validated. The most prominent one is C4A increased expression consistent across all tissues, which was recently fine-mapped for SCZ in the MHC locus [150]. In a mouse model with human C4A, it was found that overexpressing C4A diminished the cortical synapse density and increased microglia engulfment [151]. Interestingly, the highest levels of C4 protein were observed during development, but microglia engulfment due to C4A overexpression peaked during adolescence and a consequential reduction in synapse density was present only after adolescence. This mechanism underlies a cumulative cascade of events consistent with the developmental nature of SCZ [151]. In addition, C4A transgenic mice exhibited abnormal behaviors including social one, deficiency of spatial working memory, and increase anxiety. These phenotypes resembled the negative symptoms observed in individuals affected by SCZ. From a regulatory point of view, the cis-component change in expression was explained by an intricate configuration of 98 alleles, including significant GWAS hits (Fig. 4.40A). Another striking example is the association of SCZ with the decrease in DDHD2 expression. This gene was already identified from GWASs simply based on genomic location [73]. A decrease in DDHD2 cis-regulated expression in SCZ patients was also detected from recent TWASs from brain models of human fetuses [227, 255]. DDHD2 is a brain triglyceride hydrolase whose deleterious mutation is associated with complex hereditary spastic paraplegia. The double knock-out of DDHD2 in mice showed an increase in triglycerides in the central nervous system, with lipid droplets mostly localized in the intracellular compartments of neurons, concomitant to an observed phenotype of impairment in motor coordination and long-term spatial memory [256]. In the context of SCZ, DDHD2 was initially detected as carrying de novo mutations from exome sequencing [226]. Thus, DDHD2 association from common variants was in accordance with the convergence of common-rare variants in SCZ. Indeed, from the latest GWAS and exome-sequencing studies [73, 144], it was observed that genes with damaging ultra-rare mutations are enriched for GWAS associated common variants, and vice versa fine-mapped genes from GWAS are enriched for mutated genes [73]. Notably, this convergence indicates that the altered function of these genes affects both individuals carrying rare mutations and those having the pathogenic common alleles configuration. Importantly, this interplay was validated not simply via DDHD2 but from the overall significant association of the *De Novos: SCZ loss of function (LoF)* pathway (Fig. 4.44B).

The strongest signals from our TWAS analyses were already included in disease-associated loci from GWASs with comparable sample sizes. On the contrary, novel putative causal genes were characterized by a lower Z-statistic in absolute value but below FDR 0.05 threshold (Tab. 4.4, Tab. 4.5). In particular, we highlighted two novel associations for CAD: a decrease in NME7 in the aorta and colon tissues and an increase of NFU1 in the

adipose visceral omentum. As already mentioned, TWAS analysis can only point to putative causal genes. Nevertheless, NME7 was also prioritized in the latest GWAS for CAD [74] via fine-mapping and other complementary predictors supporting orthogonal lines of evidence. In detail, NME7 was the closest gene to rs61806987 associated variant, it was harboring protein-altering variants, it was a cis-eQTL and, it was assigned the highest score in the locus via PoPs [84]. From a functional point of view, NME7 is part of the ciliome and takes part in the control of the microtubule-organizing center. A complete knock-out of NME7 in rats model proved to be unsustainable with life itself, leading to premature death [257]. Nevertheless, a heterozygotic state with only one copy deleted developed normally, but affected carbohydrate and lipid metabolism increasing body weight and insulin levels and decreasing glucose tolerance [258]. Instead, NFU1 was not mentioned as a causal gene in the latest GWAS. However, patients with a point mutation in NFU1 and rats model via CRISPR-Cas9 showed mitochondrial dysfunction leading to pulmonary hypertension [210]. In particular, James et al. showed that the mutation in NFU1 was linked to pulmonary arterial hypertension through dysregulation of the antioxidant system in the mitochondria and increased reactive oxygen species levels [210]. Further investigations are necessary for the prioritization of this gene in the contest of CAD.
In the context of SCZ, we pointed to two newly identified genes: an increase in PKD1L1 in the cerebellar hemisphere and a decrease in MLF2 in the cerebellum and transformed lymphocytes. PKD1L1 is part of the ciliary calcium channel controlling calcium concentration within primary cilia without affecting cytoplasmic calcium concentration. Although it was so far not directly related to SCZ, PKD1L1 was identified as the closest gene to variants associated with anxiety disorders from a GWAS [228]. From our PALAS analysis, *calcium ion transmembrane transport* gene-set in the cerebellar hemisphere was considered as significant mostly from PKD1L1 contribution (Fig. 4.44C). Finally, MLF2 is involved in transcription regulation and was originally known to be associated with myeloid leukemia. In the context of psychiatric disorders, MLF2 was found to overlap with variants mildly associated with response to paliperidone in SCZ treatment ($P >$ e-06) [229]. In addition, MLF2 was prioritized in a GWAS with more than $1,200$ rats [259] for novel object interaction test, a predictor of addiction-like traits, consistent with MLF2 association with smoking behaviors in humans.

In conclusion, the TWAS analyses performed here from PriLer gene expression models are consistent with GWAS of comparable sample size in terms of associated genes with CAD and SCZ. This is particularly true for strong associations such as SORT1 for CAD or C4A for SCZ. Nevertheless, novel putative causal genes detected here, such as NME7 for CAD, have been further prioritized in more recent GWASs with increased sample size. In the specific example of NME7, in-vivo studies pointed out the connection between decreased NME7 and CAD pathomechanisms. This highlights the plausibility of our findings and the necessity to further validate new candidates in an experimental setting.

5.3 Convergence of small effects into biological pathways

Besides the identification of associated genes with complex diseases, one of the main benefits of CASTom-iGEx is the possibility to measure pathway activity levels for every single individual computing a sample-specific score. Thus, it is possible to perform PAthway Level Association Study (PALAS) with a framework similar to GWAS and TWAS for each computed pathway. Importantly, this allows comparing the significance reached at the pathway level to the one observed for genes and consequentially variants involved in the pathway. On the one hand, this strategy for pathway association bypasses any a priori filtering based on a p-value threshold, hence maintaining all the possible effects even those with supposedly marginal. On the other hand, the significance of the association for pathways is comparable with the ones from involved genes or variants. This is possible because all these associations are computed on the same set of individuals and hence the statistical power will only depend on the effect size and sample variance rather than changes in the sample size. In particular, we distinguished between two classes of significant pathways: those including at least one gene more significant than the level reached by the pathway itself (**class I**) and those more significant than any gene in the pathway (**class II**). The first class is indicative of pathways disrupted by the genetic perturbation of usually only one gene, the second class instead highlights a cumulative mechanism from genes that combined increase the pathway relevance.
In the two applications of CASTom-iGEx to CAD and SCZ, we found that the majority of pathways fell under the class I category (Fig. 4.18C, 4.41C). These pathways were disrupted by the strong TWAS associations such as CDKN2B, SORT1, and PCSK9 for CAD (Fig. 4.18F) or genes in the MHC locus, SNX19, and MAPK3 for SCZ (Fig. 4.42). On the other hand, class II pathways were composed of genes with weaker effects (Fig. 4.18D, 4.41D) and they can be further divided into pathways with at least one significant gene passing FDR 0.05 threshold (green) and those composed entirely of not significant genes (light blue). Hence, the aggregation of effects that we observe at the pathway level arises only from small effect genes, highlighting the importance of weak associations usually not prioritized. In the context of CAD, we additionally performed GWAS using the same case-control division from UKBB as well as the same variants involved in the gene expression imputation. This further analysis allowed the comparison of the cumulative phenomenon at multiple levels, going from variants to genes to pathways (Fig. 4.20). In particular, we detected 104 pathways that exceeded the significance of both variants and genes, additionally not passing the FDR 0.05 and hence solely arising from a cumulative effect. Nevertheless, we observed that the majority of associated pathways were less significant than both variants and genes. In addition, most of the variants were more significant than any gene they regulated or any pathway they were involved in. This indicates that either the estimation we performed of gene regulation does not capture

properly the reality (for example due to biases in the reference panels), or/and the variants might operate via additional mechanisms other than cis-regulation. Indeed, GWAS signal of associated variants represents the joint cis- and trans- effect of that variant. Moreover, a reduced significance at the pathway level might be indicative of a too simplistic approach of collapsing the gene expression by taking the average. In this context, we observed that the effect sign of genes inside a pathway contributes to the final pathway association. In particular, genes with opposite signs cancel out the signal collapsed at the pathway level (Fig. 4.45). Nevertheless, even genes with very weak effects contribute to the pathway relevance increase, at least until a p-value < 0.1. Above this threshold, genes can be considered as not involved in the complex disease and only increase the noise collected at the pathway level, hence losing relevance. Considering this limitation, a possible extension is to compute pathway-scores weighting genes by their Z-statistics, similar to pathway-specific polygenic risk-score [91] but from the perspective of perturbed genes. We did not include any disease information in the construction of pathway scores to keep the computed pathways unbiased with respect to the disease considered, hence giving a general estimate applicable in multiple contexts. Nonetheless, another extension might take into consideration tissue-specific gene regulation to account for the interactions among genes in a pathway, leveraging large measured gene-expression collections (e.g. GTEx or Brain Atlas projects) and/or known experimental results. Notably, the example shown in (Fig. 4.45) to study the incremental effect did not include any correlated genes. In this context, we examined the impact of gene correlation and genomic location in pathway significance via randomization analyses (section 4.3.3). As expected, when genes were highly correlated with each other there was no or very little increase in the significance of the pathway. On the contrary, a very mild correlation (between 0.1 and 0.2) or a strong one (> 0.9) but shared only in 50% of the genes led to the best increase in pathway significance, despite being very similar to no correlation (Fig. 4.22). When creating gene sets from relevant genes in LD, we observed that the correlation rather than their genomic position was related to the increase in the pathway significance, with an inverse relationship of higher correlation implying an increase in the cumulative effect at the pathway level (Fig. 4.23). These results were finally validated by the actual increase based on the real correlation structure of the biological pathways, with the strongest aggregation of effects observed in pathways without correlated genes (Fig. 4.24).

The novel pathway strategy we implemented in CASTom-iGEx prioritized well-known biological processes involved in the disease etiology as well as novel and controversial ones. For instance, in coronary artery disease (Fig. 4.18E-F) we detected lipid-related mechanisms (*LDL clearance*, *fatty acid biosynthetic process*, *triglyceride homeostasis*, and *Plasma lipoprotein assembly, remodeling, and clearance*), neovascularisation (*Neovascularisation processes* and *Collagen formation*) and inflammatory-related processes (*adaptive immune response* and *regulation of innate immune response*), already detected from GWAS risk loci (Fig. 2.9). Most importantly, these results were also concordant with clinical studies. Both randomized genetic studies and intervention trials have proven that lowering LDL particle

concentration diminished the risk of cardiovascular events [211]. Neovascularisation was long studied as a post-infarct intervention to decrease the apoptosis of myocytes and increase the survival of the myocardium [260]. The clinical evidence of a reduction in inflammation and CAD risk is still an ongoing debate. Nevertheless, C-reactive protein concentration was associated with risk of cardiovascular diseases [129] and recent clinical trials with colchicine, a drug used to treat arthritic conditions that target inflammatory pathways, have been proven successful in reducing future cardiac events and deaths [261, 262]. Interestingly, LDL pathway associations resulted in both class I and class II separation, contrary to neovascularisation and inflammatory pathways that were only part of class II and arose from an aggregation of effects. Indeed, significant associations in class I correspond to gene sets whose activity is disrupted by the perturbation of (usually) a single gene. For example, the strongest associations from CDKN2B correspond to differences in *epithelial cell differentiation* and *oxidative stress induced senescence* pathways. Similarly from PHACTR1, we observed a significant disruption in *actomyosin structure organization*. However, the involvement of these mechanisms in CAD pathophysiology due to the remarkable perturbation of involved genes needs further investigation. Of note, some of the significant pathways might arise simply due to LD, hampering the identification of putative causal mechanisms. For example, *spindle* and *neuropeptide signaling* pathways were related to two genes located in SORT1 locus in liver, CELSR2 and PSRC1 respectively (Fig. 4.18F). Nevertheless, *spindle assembly* pathway was significant in other tissues (whole blood, artery aorta, and heart left ventricle, Fig. 4.18E) and included other genes that were not in LD with PSRC1, hence possibly representing an actual impaired mechanism in CAD not simply arising from LD structure. In the context of pathways resulting from an aggregation of effects, we particularly observed 3 pathways that reached a strikingly higher significance than any of the genes involved, namely death receptor signaling in artery coronary, peptidyl-tyrosine phosphorylation in adipose subcutaneous and G1/S DNA Damage Checkpoints in heart atrial appendage (Fig. 4.19). None of the genes involved in these pathways were significant. Nevertheless, a gene in *Death Receptor Signalling*, ARHGEF26, reached significance and was prioritized in the most recent GWAS for CAD [74]. Multiple programs such as apoptosis and necrosis signaling regulate cell death and these actively mediated cell suicide mechanisms have been found connected to the pathogenesis of myocardial infarction and heart failure [263]. The genetic evidence that we found can help to develop novel drug strategies specific to the genes involved to inhibit this signaling and reduce heart damage. The other 2 pathways highlighted (*peptidyl-tyrosine phosphorylation* and *G1/S DNA damage checkpoints*) were as well composed of not significant genes, not even passing nominal TWAS p-value < 0.01, and are indicative of impairments in cell growth mechanisms that require further validation in the context of CAD.

Regarding schizophrenia, a substantial proportion of pathways were related to immune and inflammatory mechanisms due to the relevance of MHC locus in SCZ etiology (Fig. 4.42B, 4.43B). Apart from established results such as *regulation of complement activation*

[150] or adaptive immune system and T cells involvement [264], newly identified mechanisms including genes in MHC locus involved *Oxidative Damage* and *O-glycosylation of proteins*. In particular, oxidative stress parameters from abnormal serum, plasma, and red blood cells were hypothesized to be biomarkers for SCZ course showing impairment after the first episode of psychosis, independent of undergoing treatment [265]. Instead, genes related to sphingolipid metabolism and N- and O-linked glycan biosynthesis were found differentially expressed in individuals affected by schizophrenia in prefrontal cortical samples [266]. Outside MHC locus (Fig. 4.42A, 4.43A), we identified pathways related to MAPK3 impairment such as *MAPK Cascade* and *FSH signaling pathway*. This gene was already identified in previous TWASs and functionally validated in a zebrafish model showing consequences on neurodevelopmental phenotypes [71, 154]. Abnormal activity of the MAPK- and cAMP-associated pathways in the frontal cortex were previously identified, by studying the proteins involved in those pathways in terms of expression and specific phosphorylation [230]. Interestingly, other signaling pathways for cell proliferation were also detected among class II gene sets, namely *ErbB signaling pathway* and *mTOR signaling*. ErbB receptor activation initiates a cascade of events including the activation of PI3K/Akt/mTOR pathway and Shp2/Erk/MAPK pathway and is involved in the myelination [232]. Myelin dysfunction has been confirmed from brain imaging and post-mortem studies and connected to abnormalities in synaptic formation and function, with consequences at the level of cognitive performances [267]. Furthermore, we identified pathways related to ion channel mechanisms in both class I and class II categories, a long-known hypothesis arising from GWASs. Among them, *ion channel activity pathway* was associated via the perturbation of CLCN3, a gene involved in the regulation of neurotransmitter vesicle turnover and hypothesized to regulate synaptic plasticity from in-vivo study [268]. Interestingly, *calcium ion transmembrane transport* pathway was among class II category, exceeding the significance of all genes including the newly identified PKD1L1 (Fig. 4.44C). Recently, calcium ion-channel were found to affect macroscopic electrical signals observed as an endophenotype in individuals affected by SCZ. In particular, the increase in delta-oscillation power was connected to the altered calcium transporters or voltage-gated ion channel activities [269]. Novel mechanisms arisen from class II pathways not composed of any gene in MHC locus include *Adipogenesis*, *cell leading edge* and *De novos:SCZ Loss of Function* pathways (4.43A). Despite not being directly related to neuronal mechanisms, adipose tissue dysfunction was observed in individuals affected by SCZ in the form of decreased adiponectin (a hormone regulating glucose levels) and down-stream increase of C-reactive protein and fasting glucose [234]. Moreover, *cell leading edge* greatly exceeded the significance level of any gene in the pathway, not even passing FDR 0.05 threshold (Fig. 4.44A). The mechanisms of disruption in the cell leading edge configuration and consequential differences in cell migration for SCZ patient might be connected to irregularities in the total size and regional volume brain and requires further investigation. Finally, *De novos:SCZ Loss of Function* (Fig. 4.44B) significance underlies again the already observed agreement between rare and common variants affecting the same genes and possibly the same mechanisms in SCZ [73, 142].

In conclusion, we implemented a novel method for the identification of associated biological pathways and gene sets inside CASTom-iGEx framework. The computation of individual-level pathway scores allows testing for associations without the application of cut-offs on genes or variants based on their significance. In the light of the same cohort onto which genes and pathways are computed, it is now possible to compare the significance reached at the pathway level with the one observed at the genes (or variants) level. Hence, we can discern between pathways arising from an aggregation of effects or those disrupted by few (very significant) genes. In the CAD and SCZ application, we retrieved many well-established mechanisms as well as novel ones, particularly emerging from the aggregation of individual gene effects such as *Death Receptor Signaling* for CAD and *cell leading edge* for SCZ.

5.4 Linking changes in endophenotypes to their underlying molecular drivers

We previously showed that TWAS and PALAS analysis retrieve meaningful results in terms of CAD and SCZ finding associated pathways and genes. These two analyses are directly integrated into the CASTom-iGEx framework (Fig. 3.1). The TWAS and PALAS summary statistics can be leveraged in a downstream analysis that infers correlation and causal relationship among genetically determined traits, e.g. a disease of interest and related endophenotype. We refer to this type of analysis as Mendelian Randomization (MR) with interpretable instrument variables in the form of genes and pathways. Indeed, to assess the causal role we applied a technique used in MR studies, i.e. the inverse-variance weighted method (MR-IVW) with random effects. [11]. This specific method was chosen among those developed for MR since it provides unbiased estimates when the exclusion restriction hypothesis (H2) is violated and under the presence of balanced pleiotropy. In addition, it can be applied in large-scale one-sample setting such as among phenotypes collected in UKBB data set [100] (see section 3.2.4). Nevertheless, this analysis was mostly exploratory and not focused on known biological mechanisms specific to the exposure-trait relationship investigated. Burgess et al. [180] referred to this approach as a "joint association study" rather than Mendelian Randomization, which can still point to putative causal effects in the presence of non-null findings. In particular, our strategy can be applied to a wide range of related endophenotypes and a trait/disease of interest and is composed of two steps. The first identifies a putative relationship between two traits (e.g. LDL and CAD) via Spearman correlation, pruning genes in proximal genomic regions and pathways sharing 30% of the genes. The second step investigates the causal role of an endophenotype and the disease of interest (or vice versa) via MR-IVW with random effects, additionally accounting for not removed correlation among instrumental variables. Different from the usual MR application on GWAS that uses variants as instrumental variables, we considered genes and pathways associations with endophenotype and the trait of interest. Thus, the

original genetic status is converted into tissue-specific gene expression and biologically meaningful pathways, allowing for a more interpretable outcome that directly points to responsible biological entities in the presence of a causal relationship.

The application of this strategy to CAD or SCZ and the corresponding disease-related endophenotypes identified medically validated results. For instance, the correlation both from genes and pathways between CAD and LDL, apolipoprotein B or blood pressure translated into causal endophenotype - disease relationships (Fig. 4.26, 4.25). As already mentioned, LDL and hypertension are the primary line of intervention for CAD proven to drastically reduce the incidence in a clinical setting [211, 216]. Instead, apolipoprotein B was found as the predominant trait of lipoprotein lipids in CAD etiology in a multi-variable study, resulting as the unique lipid retaining a certain relevance with CAD [215]. This result is also plausible at the level of genes and pathways but not examined in this thesis. Importantly, we were able to directly investigate genes and pathways responsible for these causal effects and found genes that could be (or already are) prioritized as a drug target in the reduction of CAD risk such as SORT1, PCSK9, and AGPAT4 (Fig. 4.27D). On the other hand, we also identified controversial results still under debate. For instance, HDL was detected in this study as negatively correlated with CAD risk and showing a protective role for CAD etiology (Fig. 4.26, 4.25). This was also confirmed in a multi-variable MR study that included HDL, LDL, and triglycerides based on the same set of variants [218]. Burgess et al. found HDL as causally protective of CAD risk in an independent manner with respect to the other two lipid measurements. Similarly, we expect that genes and pathways exert a pleiotropic effect among different lipid classes and hence a multi-variable approach might elucidate possible independent relationships. Nevertheless, clinical trials aiming at increasing HDL had only a modest effect or even failed to show a reduction in cardiovascular events [217, 270], indicating the need for further investigations. Similarly, we detected C-reactive protein (CRP) as having a causal role in CAD manifestation in whole blood and adipose visceral tissues. CRP increase in CAD affected individuals and severity has been observed from epidemiological studies [129], and inflammatory-related interventions are already undergoing clinical trials [261, 262], proven efficacy in at least a percentage of patients. Nevertheless, previous MR studies did not find a causal effect from CRP marker [130]. Notably, Eiriksdottir et al. only considered 4 variants in the CRP gene to perform MR which explained 98% of its expression variability. Instead, we included all the genes whose cis-genetic components were associated with CRP levels, thus also considering possible inflammatory biomarkers. Hence, CRP expression alone might not be a modifiable exposure to reduce CAD risk, however a general inflammatory state represented by CRP increase might lead to an increase in CAD risk. Notably, this causal relationship was driven by biological pathways possibly involved in inflammatory mechanisms such as lysosome, inflammatory response, and post-translation protein modification (Fig. 4.27A). Interestingly, we also detected a causal role of sedentary lifestyle (time spent watching TV) for CAD risk from both genes and pathways, in line with a previous GWAS-based result [219].

In the context of SCZ, we instead studied a bidirectional relationship: genetic features (i.e. genes and pathways) being associated with SCZ endophenotypes mediating a causal effect on SCZ (direct) or vice versa associated with SCZ and mediating a causal effect onto SCZ-related traits (reverse). In particular, we detected established reduction in cognitive performances for which SCZ risk resulted as causal in the reverse setting, such as fluid intelligence score (FIs) and "time to complete round" in the pairs matching test (visual memory performances) (Fig. 4.47). For fluid intelligence, the effect was present also in the direct setting (IVs $\rightarrow$ FIs $\rightarrow$ SCZ) in the context of pathway-scores associations. Instead, the "time to complete round" in the pairs matching test had a bidirectional effect. Namely, even a lower performance in visual memory resulted in a causal SCZ predisposition. Of note, this application allows identifying responsible genes and pathways for both FIs and SCZ, such as C2orf47, ZSCAN23, and ALMS1P drivers (Fig. 4.48B), and nervous system development, axon terminus and folic acid binding relevant pathways (Fig. 4.48A). In general, additional cognitive tests resulted impaired from SCZ genetic risk, although to a lesser extent, such as prospective memory, reaction time, and trail making (Fig. 4.47). This is particularly in accordance with actual studies measuring cognitive performance in drug-naïve SCZ patients that found a general decrease [236]. Notably, risk-taking phenotype and cannabis consumption were only significant in the reverse association, considering SCZ as exposure. On the one hand, this might be related to the reduced sample size in these traits, hence including only a few genes and pathways in the direct analysis. On the other hand, it might imply that these phenotypes are not genetically mediating SCZ but rather the contrary, with SCZ predisposition leading to an increase in risk-taking behavior. Finally, in terms of blood-related markers, our results were in accordance with an MR study based on variants, indicating a causal role of lymphocyte count towards SCZ risk (Fig. 4.47). Notably, neutrophil percentage as well as the neutrophill-to-lymphocyte ratio (NLR) exhibited a protective role for SCZ, opposite to the observation from clinical studies [243] that found an increase of NLR even in first-episode psychosis patients, and hence requires further investigations. Finally, a blood biomarker with a bidirectional effect was aspartate aminotransferase (AST). In the context of brain functions, AST catalyzes aspartate and converts alpha-ketoglutarate to oxaloacetate and glutamate. This mechanism might be connected to the *glutamate hypothesis* for SCZ that assumes the disruption of glutamatergic signaling as core mechanisms of SCZ pathophysiology, for example through N-methyl-d-aspartate receptor (NMDAR) hypofunction model [237] that leads to exceeding glutamate release. Indeed, elevated levels of glutamate in the hippocampus and prefrontal cortex have also been detected in SCZ via magnetic resonance spectroscopy [271]. Whether the hypofunction of NMDAR is caused by an exceed in glutamate due to the increased catalyzes of aspartate by AST requires further investigation. Notably, in the examples that we have shown (Fig. 4.27, 4.48) the heterogeneity in exposure-outcome was always significant when tested via Cochran's Q statistic, highlighting the need for random-effect usage and the presence of pleiotropy. Further approaches might specifically focus on some of the associations found and consider a subset of genes (for instance in a pathway) related to the exposure-outcome. In addition, further MR methodology might be considered, for

instance addressing the presence of unbalanced pleiotropy via MR-Egger [99].

In summary, we leveraged gene and pathway associations obtained via CASTom-iGEx to identify endophenotypes having a protective or causal role for CAD or SCZ. In particular, we investigated the disease etiology mediated by the genetic effect collapsed at the level of genes and pathways, hence considering more interpretable instrumental variables than single genetic variants. We found tissue-specific causal roles, some of which are well established, such as LDL on CAD and SCZ on fluid intelligence score, and others that are observed only at the epidemiological level and require further investigation such as C-reactive protein on CAD and aspartate aminotransferase on SCZ. These results represent a starting point in dissecting the molecular features responsible for these associations that could be operationalized as possible treatment targets.

5.5 Characterization of genetically defined patient subgroups

In the last module of CASTom-iGEx, we leveraged the imputed gene expression computed at the individual level to genetically stratify patients. Our method relies on a community detection technique called Louvain clustering [184] that has the advantage of being particularly efficient in terms of computational time and obtaining good quality community that optimizes modularity (see section 3.3.1). Recently, an extension has been developed called Leiden clustering [272], overcoming the insurgence of disconnected partitions from the Louvain algorithm and reducing the computational time required via a fast local move approach. This novel method is already implemented in `R` and could be easily integrated into our framework. In CASTom-iGEx, communities of patients are detected based on their tissue specific gene T-scores. These are previously corrected for PCs, standardized, and TWAS-scaled via the multiplication of each gene for the TWAS Z-statistic of the disease of interest across all patients. Each of these pre-processing steps is crucial in obtaining meaningful partitions: correcting for PCs reduces the ancestry relevance and contribution to the final clustering structure (Fig. 4.35), standardization enhances the differences among patients and TWAS-scaling gives a higher relevance to genes associated with a certain disease, without any filtering based on p-value threshold. Importantly, the TWAS-scaling step gives rise to better partitions in terms of modularity and more homogeneous cluster sizes (Fig. 4.37). Our clustering strategy is different from a simplistic stratification based on polygenic risk score (PRS) percentiles usually performed in the context of complex diseases [104]. PRS is a unique score computed for each individual as the linear combination of variants alleles multiplied by their effect sizes with respect to a certain disease. We similarly weight each gene by their disease relevance and sign via Z-statistic. However, the clustering is performed in an unsupervised manner and considers all the genes simultaneously, without collapsing single information

in a unique score. This allows retrieving any kind of possible configuration in terms of genetic liability and not necessarily partitioning between low and high-risk individuals. Most importantly, CASTom-iGEx clustering identified groups of patients with 1) distinct cell type-specific molecular pathways distributions as well as 2) distinct endophenotype and clinical profiles in the CAD and SCZ application. To the best of our knowledge, this is the first attempt at unsupervised genetic stratification that converges into molecular and phenotypic differences. A parallel but supervised approach was applied in a recent study by Nguyen et al. [114] in the context of major depression. In particular, the authors performed subtype-specific GWAS for 16 subtypes spanning different symptomatology and found loci specific only to certain subgroups. This study validates again that clinical heterogeneity arises from different genetic liabilities, however in a supervised manner. On the other hand, another recently developed method called BUHMBOX [110] attempted to decompose the source of genetic heterogeneity observed in complex diseases. In particular, they tested whether complex disease genetic heterogeneity is driven by a subgroup of individuals exhibiting a higher genetic correlation with a certain trait or endophenotype, again following a supervised approach. Instead, CASTom-iGEx does not use any information on those endophenotypes for which the differences in stratified patients are tested. For instance, CASTom-iGEx can still detect a distinct endophenotypic pattern when a group is driven only by few loci. This particularly occurs if those loci impact the considered endophenotype. On the contrary, BUHMBOX would not necessarily identify a subgroup heterogeneity in this scenario, depending on the alleles distribution of the other variants associated with this endophenotype. Finally, BUHMBOX does not derive an actual partition given a set of individuals nor uses a definition of groups, contrary to our methodology that actively divide the patient space. Another advantage of our method compared to the previously mentioned studies is the possibility to perform a consistent comparison across multiple observed endophenotypes. Hence, from a unique clustering, we can deduce the downstream distinct pattern across a wide range of measured phenotypes. We also benchmarked our clustering and group-specific endophenotype analysis showing that the problem is well calibrated under the null hypothesis of no genetically derived groups, with cluster-specific genes and pathways associations almost always following a uniform distribution (Fig. 4.38B, D). In addition, the endophenotypes associations for random clustering are rarely significant after FDR correction and in general with a level of significance always lower than the actual genetically derived clustering structure (Fig. 4.38F).

The application of CASTom-iGEx clustering on CAD allowed the testing of cluster-specific patterns across multiple CAD-related phenotypes in the UKBB data set. This was not possible for the SCZ application composed solely of genotype and SCZ status in PGC cohorts. To approximate plausible SCZ-related phenotypes in PGC cohorts based on their genetic heritability, we computed for each individual in PGC a gene risk-score (gene-RS). Similarly to PRS, gene-RS is computed as the weighted sum of associations and observed genetic components at the level of genes. Here the considered gene associations with a trait (Z-statistics) were estimated from UKBB deep phenotyping and used as weights in the PGC imputed gene expression, hence to mimic the genetic component of a phenotype

that was instead not measured in that cohort. The tested cluster-specific differences in terms of endophenotype gene-RS were additionally associated with a measure of reliability (CRM), with the threshold calibrated on the CAD application for which actual endophenotype differences were tested (Fig. 4.60). Moreover, using a label propagation approach, CASTom-iGEx can also project on external cohorts, usually smaller, the clustering structure observed in larger and usually better-characterized cohorts. Applying this strategy to CAD and SCZ, we showed that the projected clusters obtained were similar to the model clustering in terms of concordant genes signatures, loci and fraction of patients per group (Fig. 4.31, 4.52, 4.56). The consistent projection highlighted the generality of our method and a perspective clinical application, training on larger more informative, and characterized cohorts to acquire knowledge on smaller data sets of clinical nature. Finally, the patients' stratifications we found were not defined nor driven by ancestry structure. As already mentioned, ancestry contribution was reduced after PCs correction but not completely removed. Observing the actual PCs across groups, we found that some of the components were significant. However, the PCs distribution was not separating the patient space (Fig. 4.30, 4.53, 4.59), and the significance with the cluster structure was far lower than the actual driver of the clusters i.e. molecular features in the form of genes and pathways (Fig. 4.29, 4.50 4.57). To further validate this, we additionally compared tissue-specific clustering structure with the one derived solely from PCs and found a minimal overlap (NMI < 0.0052, Fig. 4.36B, 4.61B) although not null and outside a randomly assigned clustering structure (Fig. 4.36C, 4.61C). Most importantly, the minimal overlap between tissue-derived and ancestry-derived clustering did not influence the observed group-specific endophenotype differences, which resulted as completely divergent among the two partitions (Fig. 4.36E-F, 4.61E-F). We concluded that the minimal ancestry information still present in the tissue-specific partitions does not compromise the findings on group-specific endophenotypic and biological characterization.

In the application of CASTom-iGEx to CAD, we focused on liver tissue due to its relevance in CAD pathophysiology and role in lipid metabolism. CAD patients in UK Biobank were stratified into 5 groups associated to 236 genes (FDR 0.01) that merged into 16 unique loci (Fig. 4.29) and led to distinct molecular pathways (Fig. 4.32). For instance, patients in gr_3 and gr_5 showed increased pathway activity related to *Golgi Associated Vesicle Biogenesis*, driven by differences in SORT1 imputed expression and with a stronger effect size for gr_5. This pathway perturbation was concomitant with a decrease in LDL and apolipoprotein B measured levels (Fig. 4.33A). The observed phenotype was in accordance with an increase in the vesicular transport of cholesterol between the endoplasmatic reticulum and the plasma membrane through the Golgi apparatus [273]. In addition, *N-acetyltransferase activity* was increased solely in gr_3 for the cumulative effect of single genes outside SORT1 locus, nevertheless with a marginal relevance compared to the previously mentioned pathways. Interestingly, this pathway is also related to dyslipidemia (high LDL, low HDL) from in-vivo studies [274] concordant with the lipid-related phenotype of the group. The remaining groups exhibited an opposite effect for

Golgi Associated Vesicle Biogenesis and a corresponding increase in LDL and apolipoprotein B. Moreover, genetic liability of gr_2 converged into decreased *alcohol metabolic process* related to ALDH2 changes, whereas gr_1 and gr_4 showed a significant opposite effect, hence translating in the highest increase for LDL and hyperlipidemia in gr_2 (Fig. 4.33E). Inflammatory mechanisms and lipid metabolisms were perturbed specifically in gr_1 and gr_4 with opposing effect and gr_1 effect signs mostly concordant with CAD risk (Fig. 4.32C). The general higher risk observed at the pathways level for gr_1 was concomitant with higher severity from an endophenotypic perspective, with both increased inflammation and lipid profiles (Fig. 4.32A). Interestingly, individuals in gr_1 had the tendency to be shorter in height. The genetic association between shorter height and an increased risk of CAD was already noticed and partially explained by the pleiotropic effect of variants associated with shorter height and an adverse lipid profile [222]. Here, gr_1 had a comprehensive genetic CAD liability rather than simply lipid distribution, highlighting additional shared genetic mechanisms associated with inflammation. Strikingly, the higher risk of individuals in gr_1 was externally validated in the GermanV CARDIoGRAM cohort in terms of the increased number of affected vessels (Fig. 4.32B). Furthermore, the hypothesis-driven analysis detected group-specific differences in two clinical phenotypes: reduction of hyperlipidemia comorbidity in gr_3 and gr_5 (in accordance with their phenotype profiles) and increased number of patients with peripheral vascular disease in gr_4 (Fig. 4.32C-D). Finally, the medication information available on the UKBB data set allowed us to investigate cluster-specific medication responses, comparing individuals in a group based on their medication assumption. As expected, statin usage generally reduced LDL values, nevertheless with a lower effect in gr_5 (Fig. 4.34A-B). This finding is consistent with pharmacogenomic studies that found SNP rs646776 in SORT1 associated with a decreased effect [223]. However, the diminished efficacy of statin medication was compensated by a general lower LDL distribution of gr_5 patients, hence the actual LDL values after statin assumption were similar across all groups (Fig. 4.34B). In addition, glucosamine assumption reduced CRP levels specifically in gr_5 individuals, whereas no reduction effect was observed in the other groups, contrary to cholesterol-lowering medication usages (Fig. 4.34C-D). The anti-inflammatory effect of statin medications is known [275], contrary to the less studied anti-inflammatory effect of glucosamine, a natural supplement used to alleviate pain in people with osteoarthritis. Thus, this result highlights a possible cost-effective therapeutic strategy to decrease CRP and inflammatory states for genetically defined sub-population of patients that might juxtapose ongoing trials [261, 262].

In the SCZ application of CASTom-iGEx, we focused on DLPC tissue as the most reliably imputed tissue and its implication in SCZ pathophysiology. We considered two filtering strategies for gene correlation: genes were clumped at 0.9 (as default) and at 0.1 to reduce MHC contribution. In the default correlation filtering, SCZ patients were stratified into three groups and mostly driven by differences in MHC genes (Fig. 4.50). Indeed, 92% of the associated genes were located in the extended MHC locus with a striking significance of p-value $<$ 1e-200 (Fig. 4.50B-C). However, the MHC locus was not the

only contributor and associated genes across the entire genome were grouped in 34 loci (Fig. 4.50D), with certain cluster-specific genes outside MHC still related to cognitive functions such as MPHOSPH10 and ALDH18A1. Importantly, genes configuration specific to each cluster highlighted that the three groups were characterized by an SCZ genetic liability from low in gr_1 to intermediate in gr_2 to high in gr_3 individuals (Fig. 4.50E). In a detailed analysis of the MHC locus, we found that the dense correlation structure due to LD could at least differentiate between three cluster-specific genes with strong effects but reduced interconnection: C4A, ZSCAN23, and BTN3A3 (Fig. 4.51). Hence, the contribution to the clustering structure of the entire locus was not simply recapitulating in a single exemplar gene but by marginally connected signals. We caution that these results necessitate further validation via high-density platforms such as the Immunochip platform that contains a dense panel of SNPs from the MHC locus [276]. Notably, both transcription factors ZSCAN23 and ZKSCAN3 (corr.$= -0.7$), strongly associated with patients stratification, were identified as druggable nevertheless unexplored targets for SCZ that require further functional characterization [277]. The cluster-specific genes converged into immune-related pathways, ion channels, axonogenesis, and autophagy (Fig. 4.54), with a group-specific effect sign concordant with low-risk, intermediate, and high-risk configuration of gr_1, gr_2, and gr_3 respectively. Most importantly, these distinct molecular features were concomitant with different SCZ-related phenotypic liability measured via gene-RS (Fig. 4.55). Individuals in gr_1 showed an increased genetic loading for better cognitive performances and lower liability for inflammatory markers and risk for depression or anxiety-related disorders. In addition, gene-RS for diffusion brain magnetic resonance imaging such as fractional anisotropy of splenium of corpus callosum was increased in gr_1 and showed an opposite effect in gr_3, in accordance with MRI studies that found this region decreased in SCZ patients compared to healthy individuals [245]. Strikingly, while these results were conformed to the overall lower SCZ risk of individuals in gr_1, these individuals were also associated with a higher predisposition of developing metabolic syndrome (MetS) (Fig. 4.55C). MetS is prevalent in people with SCZ compared to the general population, even in drug-naive patients [248]. This prevalence might arise from an interplay between pleiotropic genetic factors and environmental exposures such as increased smoking, unhealthy lifestyle, and obesity from anti-psychotic drugs [139]. Our results point towards a sub-group of SCZ patients at higher risk of developing MetS for which a specific selection of treatments should be prescribed to minimize metabolic impact. In terms of cognitive performances, individuals in gr_3 were at greater risk of impairments, in particular regarding fluid intelligence, executive function, visual memory, perspective memory, and processing speed, both across SCZ patients and when compared to healthy controls (Fig. 4.56). On the other hand, individuals in gr_2 had a lower working memory liability, not observed in any other group, and even lower than the liability observed in healthy subjects. Generally, we discover group-specific cognitive impairments with gr_1 and gr_3 at the extremes from high to low performances. Together with the inflammatory markers, negative symptoms, and diffusion MRI phenotypes, we conclude that individuals in gr_3 were at the greatest likelihood of SCZ severeness.

Finally, we stratified SCZ patients considering a lower correlation threshold for genes (0.1), to reduce MHC locus contribution. In this case, we detected groups strongly driven by genes outside the extended MHC region (Fig. 4.57B), for a total of 59 cluster-specific loci. These differences converged into distinct biological pathways (Fig. 4.57C) such as *Processing of SMDT1* and *folic acid binding* increased in gr_3 and gr_4 and decreased in gr_1 related to C2orf47 distribution; *autophagy* and immune-related pathways with opposite effect in gr_3 and gr_4 from genes in MHC; *synaptic transmission, glutamatergic* decrease in gr_4 with the highest contribution from ALS2 gene; *hedgehog receptor activity* increase and *L-ascorbic acid binding* decrease specific of gr_2. Contrary to the previous findings, the groups had a mixture effect size that did not lead to a clear division between healthy-like and high-risk groups. In terms of immune-related endophenotypic liabilities, the groups-specific profiles of gr_3 and gr_4 were similar to the 0.9 correlation threshold (Fig. 4.57D). In addition, individuals in gr_2 showed a higher liability of negative symptoms related to depression but without any other cognitive impairment. Interestingly, these changes are concomitant with a group decrease of *L-ascorbic acid binding* pathway that is related to absorption of vitamin-C, whose administration was shown to increase negative symptoms in SCZ patients [250].

In summary, individual-level genes and pathways are operationalized in the CASTom-iGEx pipeline to stratify patients according to their composite liability. This analysis identified distinct group-specific pathways associated with distinct endophenotype profiles, clinical outcomes, and treatment responses. Our approach provides a direct biological interpretation of responsible pathways in patient stratification and hence testable hypotheses on potential pathomechanisms leading to the phenotypical manifestation.

5.6 Conclusions

The CASTom-iGEx framework developed here is a valuable tool to incorporate genetic signals from large-scale studies to obtain insights into potential pathomechanisms and discover intermediate disease-related phenotypes arising from shared biological mechanisms. Furthermore, CASTom-iGEx patients' stratification based on their composite gene-specific liabilities paves the way to a precision medicine strategy with consequences at the clinical level. For instance, the stratification of SCZ patients identified a high metabolic risk group that might be treated with specific antipsychotic drugs to reduce possible lethal comorbidities. The sub-group of CAD patients at higher overall risk might be advised of early intervention in reducing external risk factors. CAD sub-group benefiting from glucosamine administration might be treated with this supplement to reduce the inflammatory state instead of advising for more expensive and possibly less effective treatment. The next step in this direction would be a validation of the mentioned results via clinical randomized control trials in terms of personalized treatments from genetic patient stratification. In

addition, there is the need of integrating these genetic-based insights with additional deep patient phenotyping, including multi-omic characterization, imaging, and external/environmental factors. Notably, the identified genes and pathomechanisms might serve as a starting point for therapeutic targets and require further validation via experimental systems such as CRISPR-Cas9. A proof-of-principle in this direction regards a viral injection targetting hepatic CAD-associated PCSK9 via CRISPR–Cas9 that led to a reduction in cholesterol up to 40% in a mouse model [278]. In addition, the interpretability of results provided by biological pathways might lead to transversal therapies that would be impossible to identify by inspecting single genes, especially those with weak effects. For instance, possible targets might act to modify the general pathway activity via a combination of targetable genes. Finally, the results we elucidate here are consistent and reproducible across different cohorts. Nevertheless, we specifically selected individuals of European ancestry due to the possible differences in gene expression prediction models across populations [279]. We envision that the increasing availability of population-specific genetic and phenotypic studies will allow the construction of ancestry-specific gene expression models onto which CASTom-iGEx can be directly applied.

Bibliography

1. Klein, R. J., Zeiss, C., Chew, E. Y., *et al.* Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science* **308,** 385–389 (Apr. 2005).
2. Levey, D. F., Stein, M. B., Wendt, F. R., *et al.* Bi-ancestral depression GWAS in the Million Veteran Program and meta-analysis in >1.2 million individuals highlight new therapeutic directions. *Nature Neuroscience* **24,** 954–963 (2021).
3. Loos, R. J. *15 years of genome-wide association studies and no signs of slowing down* Dec. 2020.
4. Gallagher, M. D. & Chen-Plotkin, A. S. The Post-GWAS Era: From Association to Function. *American Journal of Human Genetics* **102,** 717–730 (2018).
5. Visscher, P. M., Wray, N. R., Zhang, Q., *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics* **101,** 5–22 (July 2017).
6. Buniello, A., MacArthur, J. A. L., Cerezo, M., *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research* **47,** D1005–D1012 (Jan. 2019).
7. Yang, J., Benyamin, B., McEvoy, B. P., *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42,** 565–569 (2010).
8. Genkel, V. V. & Shaposhnik, I. I. Conceptualization of Heterogeneity of Chronic Diseases and Atherosclerosis as a Pathway to Precision Medicine: Endophenotype, Endotype, and Residual Cardiovascular Risk. *International Journal of Chronic Diseases* **2020,** 1–9 (Feb. 2020).
9. Gusev, A., Ko, A., Shi, H., *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics* **48,** 245–252 (2016).
10. Gamazon, E. R., Wheeler, H. E., Shah, K. P., *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics* **47,** 1091–1098 (Aug. 2015).
11. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology* **37,** 658–665 (Nov. 2013).
12. Smith, G. D., Ebrahim, S., Lewis, S., *et al.* Genetic epidemiology and public health: hope, hype, and future prospects. *The Lancet* **366,** 1484–1498 (Oct. 2005).
13. Manolio, T. A., Collins, F. S., Cox, N. J., *et al. Finding the missing heritability of complex diseases* Oct. 2009.

14. MacArthur, J., Bowler, E., Cerezo, M., *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). eng. *Nucleic acids research* **45,** D896–D901 (Jan. 2017).

15. Lee, J. J., Wedow, R., Okbay, A., *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics* **50,** 1112–1121 (2018).

16. Uffelmann, E., Huang, Q. Q., Munung, N., *et al.* Genome-wide association studies. *Nature Reviews Methods Primers* **1,** 59 (Dec. 2021).

17. Bycroft, C., Freeman, C., Petkova, D., *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562,** 203–209 (Oct. 2018).

18. Welter, D., MacArthur, J., Morales, J., *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research* **42,** D1001–D1006 (Jan. 2014).

19. Purcell, S., Neale, B., Todd-Brown, K., *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81,** 559–575 (2007).

20. Loh, P. R., Tucker, G., Bulik-Sullivan, B. K., *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics* **47,** 284–290 (2015).

21. Jiang, L., Zheng, Z., Qi, T., *et al.* A resource-efficient tool for mixed model association analysis of large-scale data. *Nature Genetics* **51,** 1749–1755 (2019).

22. Mägi, R. & Morris, A. P. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* **11,** 288 (2010).

23. Andersson, R., Gebhard, C., Miguel-Escalada, I., *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507,** 455–461 (2014).

24. Maurano, M. T., Humbert, R., Rynes, E., *et al.* Systematic localization of common disease-associated variation in regulatory DNA. eng. *Science (New York, N.Y.)* **337,** 1190–1195 (Sept. 2012).

25. Musunuru, K., Strong, A., Frank-Kamenetsky, M., *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466,** 714–719 (2010).

26. Slatkin, M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future (2008).

27. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics* **39,** 906–913 (2007).

28. Bush, W. S. & Moore, J. H. Chapter 11: Genome-Wide Association Studies. *PLOS Computational Biology* **8,** e1002822– (Dec. 2012).

29. Marouli, E., Graff, M., Medina-Gomez, C., *et al.* Rare and low-frequency coding variants alter human adult height. *Nature* **542,** 186–190 (2017).

30. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics* **19,** 491–504 (2018).

31. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying Causal Variants at Loci with Multiple Signals of Association. *Genetics* **198,** 497–508 (Oct. 2014).

32. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics* **88,** 76–82 (2011).

33. Benner, C., Spencer, C. C. A., Havulinna, A. S., *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32,** 1493–1501 (May 2016).

34. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **82,** 1273–1300 (Dec. 2020).

35. Ellinghaus, D., Jostins, L., Spain, S. L., *et al.* Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nature Genetics* **48,** 510–518 (2016).

36. Smeland, O. B., Bahrami, S., Frei, O., *et al.* Genome-wide analysis reveals extensive genetic overlap between schizophrenia, bipolar disorder, and intelligence. *Molecular Psychiatry* **25,** 844–853 (2020).

37. Burgess, S., Davey Smith, G., Davies, N. M., *et al.* Guidelines for performing Mendelian randomization investigations. *Wellcome Open Research* **4,** 186 (Nov. 2019).

38. Wray, N. R., Goddard, M. E. & Visscher, P. M. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Research* **17,** 1520–1528 (Oct. 2007).

39. King, E. A., Wade Davis, J. & Degner, J. F. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS Genetics* **15** (2019).

40. Gandal, M. J., Leppa, V., Won, H., Parikshak, N. N. & Geschwind, D. H. The road to precision psychiatry: translating genetics into disease mechanisms. *Nature Neuroscience* **19,** 1397–1407 (2016).

41. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38,** e164–e164 (Sept. 2010).

42. Farh, K. K.-H., Marson, A., Zhu, J., *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518,** 337–343 (2015).

43. Kundaje, A., Meuleman, W., Ernst, J., *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).

44. Dunham, I., Kundaje, A., Aldred, S. F., *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

45. Bock, C., Halbritter, F., Carmona, F. J., *et al.* Quantitative comparison of DNA methylation assays for biomarker development and clinical applications. *Nature Biotechnology* **34,** 726–737 (2016).

46. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369,** 1318–1330 (Sept. 2020).

47. Greenwald, W. W., Li, H., Benaglio, P., *et al.* Subtle changes in chromatin loop contact propensity are associated with differential gene regulation and expression. *Nature Communications* **10,** 1054 (2019).

48. Cano-Gamez, E. & Trynka, G. *From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases* May 2020.

49. Hu, X., Kim, H., Stahl, E., *et al.* Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. *American Journal of Human Genetics* **89,** 496–506 (2011).

50. Trynka, G., Sandor, C., Han, B., *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genetics* **45,** 124–130 (2013).

51. Iotchkova, V., Ritchie, G. R., Geihs, M., *et al.* GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. *Nature Genetics* **51,** 343–353 (Feb. 2019).

52. Finucane, H. K., Bulik-Sullivan, B., Gusev, A., *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* **47,** 1228–1235 (2015).

53. Finucane, H. K., Reshef, Y. A., Anttila, V., *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nature Genetics* **50,** 621–629 (2018).

54. Fromer, M., Roussos, P., Sieberts, S. K., *et al.* Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature Neuroscience* **19,** 1442–1453 (2016).

55. Yao, C., Chen, G., Song, C., *et al.* Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nature Communications* **9,** 3268 (2018).

56. Hannon, E., Spiers, H., Viana, J., *et al.* Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nature Neuroscience* **19,** 48–54 (2016).

57. Kumasaka, N., Knights, A. J. & Gaffney, D. J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nature Genetics* **48,** 206–213 (2016).

58. Hormozdiari, F., van de Bunt, M., Segrè, A. V., *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *American Journal of Human Genetics* **99,** 1245–1260 (2016).

59. Giambartolomei, C., Vukcevic, D., Schadt, E. E., *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genetics* **10** (2014).

60. He, X., Fuller, C. K., Song, Y., *et al.* Sherlock: Detecting gene-disease associations by matching patterns of expression QTL and GWAS. *American Journal of Human Genetics* **92,** 667–680 (2013).

61. Giambartolomei, C., Liu, J. Z., Zhang, W., *et al.* A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics* **34,** 2538–2545 (2018).

62. Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., *et al.* Opportunities and challenges for transcriptome-wide association studies. *Nature Genetics* **51,** 592–599 (2019).

63. Zou, H. & Hastie, T. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **67,** 301–320 (Mar. 2005).

64. Robinson, G. K. That BLUP is a Good Thing: The Estimation of Random Effects. *Statistical Science* **6,** 15–32 (Feb. 1991).

65. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLOS Genetics* **9,** e1003264 (Feb. 2013).

66. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58,** 267–288 (Mar. 1996).

67. Lloyd-Jones, L. R., Holloway, A., McRae, A., *et al.* The Genetic Architecture of Gene Expression in Peripheral Blood. *The American Journal of Human Genetics* **100,** 228–237 (Feb. 2017).

68. Lamb, J., Crawford, E. D., Peck, D., *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. eng. *Science (New York, N.Y.)* **313,** 1929–1935 (Sept. 2006).

69. Zhang, W., Voloudakis, G., Rajagopal, V. M., *et al.* Integrative transcriptome imputation reveals tissue-specific and shared biological mechanisms mediating susceptibility to complex traits. *Nature Communications* **10** (Dec. 2019).

70. Barbeira, A. N., Dickinson, S. P., Bonazzola, R., *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature Communications* **9,** 1825 (2018).

71. Gusev, A., Mancuso, N., Won, H., *et al.* Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nature Genetics* **50,** 538–548 (2018).

72. Mancuso, N., Freund, M. K., Johnson, R., *et al.* Probabilistic fine-mapping of transcriptome-wide association studies. *Nature Genetics* **51,** 675–682 (Apr. 2019).

73. Trubetskoy, V., Pardiñas, A. F., Qi, T., *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604,** 502–508 (Apr. 2022).

74. Aragam, K. G., Jiang, T., Goel, A., Kanoni, S. & Wolford, B. N. Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. *medRxiv* (2021).

75. Fabregat, A., Sidiropoulos, K., Viteri, G., *et al.* Reactome pathway analysis: A high-performance in-memory approach. *BMC Bioinformatics* **18,** 142 (Mar. 2017).

76. Ashburner, M., Ball, C. A., Blake, J. A., *et al.* Gene ontology: tool for the unification of biology. eng. *Nature genetics* **25,** 25–29 (May 2000).

77. Slenter, D. N., Kutmon, M., Hanspers, K., *et al.* WikiPathways: A multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research* **46,** D661–D667 (Jan. 2018).

78. Segrè, A. V., Groop, L., Mootha, V. K., Daly, M. J. & Altshuler, D. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genetics* **6** (Aug. 2010).

79. Lee, P. H., O'dushlaine, C., Thomas, B. & Purcell, S. M. INRICH: Interval-based enrichment analysis for genome-wide association studies. *Bioinformatics* **28,** 1797–1799 (2012).

80. Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z. & Bergmann, S. Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *PLoS Computational Biology* **12** (2016).

81. Pei, G., Sun, H., Dai, Y., *et al.* Investigation of multi-trait associations using pathway-based analysis of GWAS summary statistics. *BMC Genomics* **20** (2019).

82. De Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLoS Computational Biology* **11,** 1–19 (2015).

83. Pers, T. H., Karjalainen, J. M., Chan, Y., *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nature Communications* **6,** 5890 (2015).

84. Weeks, E., Ulirsch, J., Cheng, N., *et al. Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases* 2020.

85. Wu, C. & Pan, W. Integrating eQTL data with GWAS summary statistics in pathway-based analysis with application to schizophrenia. eng. *Genetic epidemiology* **42,** 303–316 (Apr. 2018).

86. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4,** 44–57 (2009).

87. Pain, O., Pocklington, A. J., Holmans, P. A., *et al.* Novel Insight Into the Etiology of Autism Spectrum Disorder Gained by Integrating Expression Data With Genome-wide Association Statistics. *Biological Psychiatry* **86,** 265–273 (2019).

88. Dall'Aglio, L., Lewis, C. M. & Pain, O. Delineating the Genetic Component of Gene Expression in Major Depression. *Biological Psychiatry* **89,** 627–636 (2021).

89. Su, K., Yu, Q., Shen, R., *et al.* Pan-cancer analysis of pathway-based gene expression pattern at the individual level reveals biomarkers of clinical prognosis. *Cell Reports Methods* **1,** 100050 (2021).

90. Darst, B. F., Koscik, R. L., Racine, A. M., *et al.* Pathway-Specific Polygenic Risk Scores as Predictors of Amyloid-β Deposition and Cognitive Function in a Sample at Increased Risk for Alzheimer's Disease. eng. *Journal of Alzheimer's disease : JAD* **55,** 473–484 (2017).

91. Choi, S. W., Garcia-Gonzalez, J., Ruan, Y., *et al.* The power of pathway-based polygenic risk scores. *PREPRINT (Version 1) available at Research Square* (2021).

92. Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J. & Visscher, P. M. *Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model* June 2018.

93. Watanabe, K., Stringer, S., Frei, O., *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nature Genetics* **51,** 1339–1348 (2019).

94. Bulik-Sullivan, B., Finucane, H. K., Anttila, V., *et al.* An atlas of genetic correlations across human diseases and traits. *Nature Genetics* **47,** 1236–1241 (Nov. 2015).

95. Jordan, D. M., Verbanck, M. & Do, R. HOPS: a quantitative score reveals pervasive horizontal pleiotropy in human genetic variation is driven by extreme polygenicity of human traits and diseases. *Genome Biology* **20,** 222 (2019).

96. Burgess, S., Bowden, J., Fall, T., Ingelsson, E. & Thompson, S. G. *Sensitivity analyses for robust causal inference from mendelian randomization analyses with multiple genetic variants* 2017.

97. Bowden, J., Del Greco M., F., Minelli, C., *et al.* A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization (2017).

98. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genetic Epidemiology* **40,** 304–314 (May 2016).

99. Bowden, J., Smith, G. D. & Burgess, S. Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. *International Journal of Epidemiology* **44,** 512–525 (May 2015).

100. Minelli, C., Del Greco M., F., van der Plaat, D. A., *et al.* The use of two-sample methods for Mendelian randomization analyses on single large datasets. *International Journal of Epidemiology* (Apr. 2021).

101. Zhu, Z., Zhang, F., Hu, H., *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics* **48,** 481–487 (2016).

102. Porcu, E., Rüeger, S., Lepik, K., *et al.* Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nature Communications* **10** (Dec. 2019).

103. Gibson, G. *Rare and common variants: Twenty arguments* 2012.

104. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics* **17,** 392–406 (2016).

105. Khera, A. V., Chaffin, M., Aragam, K. G., *et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations* Sept. 2018.

106. Fahed, A. C., Wang, M., Homburger, J. R., *et al.* Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nature Communications* **11,** 1–9 (2020).

107. Martin, A. R., Gignoux, C. R., Walters, R. K., *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. eng. *American journal of human genetics* **100,** 635–649 (Apr. 2017).

108. Tehranchi, A., Hie, B., Dacre, M., *et al.* Fine-mapping cis-regulatory variants in diverse human populations. *eLife* **8** (eds Morris, A. P. & Wittkopp, P. J.) e39595 (2019).

109. Amariuta, T., Ishigaki, K., Sugishita, H., *et al.* Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements. *Nature Genetics* **52,** 1346–1354 (2020).

110. Han, B., Pouget, J. G., Slowikowski, K., *et al.* A method to decipher pleiotropy by detecting underlying heterogeneity driven by hidden subgroups applied to autoimmune and neuropsychiatric diseases. *Nature Genetics* **48,** 803–810 (July 2016).

111. Howard, D. M., Folkersen, L., Coleman, J. R., *et al.* Genetic stratification of depression in UK Biobank. *Translational Psychiatry* **10** (Dec. 2020).

112. Ozdemir, C., Kucuksezer, U. C., Akdis, M. & Akdis, C. A. The concepts of asthma endotypes and phenotypes to guide current and novel treatment strategies. eng. *Expert review of respiratory medicine* **12,** 733–743 (Sept. 2018).

113. Tromp, J., Ouwerkerk, W., Demissei, B. G., *et al.* Novel endotypes in heart failure: Effects on guideline-directed medical therapy. *European Heart Journal* **39,** 4269–4276 (2018).

114. Nguyen, T. D., Harder, A., Xiong, Y., *et al.* Genetic heterogeneity and subtypes of major depression. *Molecular Psychiatry* **27,** 1667–1675 (Mar. 2022).

115. Wu, H. M., Goate, A. M. & O'Reilly, P. F. Heterogeneous effects of genetic risk for Alzheimer's disease on the phenome. *Translational Psychiatry* **11** (Dec. 2021).

116. Wang, H., Naghavi, M., Allen, C., *et al.* Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet* **388,** 1459–1544 (Oct. 2016).

117. Zdravkovic, S., Wienke, A., Pedersen, N. L., *et al.* Heritability of death from coronary heart disease: a 36-year follow-up of 20 966 Swedish twins. eng. *Journal of internal medicine* **252,** 247–254 (Sept. 2002).

118. Nikpay, M., Goel, A., Won, H. H., *et al.* A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics* **47,** 1121–1130 (2015).

119. Chen, Z. & Schunkert, H. Genetics of coronary artery disease in the post-GWAS era. *Journal of Internal Medicine* **290,** 980–992 (2021).

120. Nelson, C. P., Goel, A., Butterworth, A. S., *et al.* Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nature Genetics* **49,** 1385–1391 (Sept. 2017).

121. Congrains, A., Kamide, K., Oguro, R., *et al.* Genetic variants at the 9p21 locus contribute to atherosclerosis through modulation of ANRIL and CDKN2A/B. *Atherosclerosis* **220,** 449–455 (Feb. 2012).

122. Jarinova, O., Stewart, A. F. R., Roberts, R., *et al.* Functional analysis of the chromosome 9p21.3 coronary artery disease risk locus. eng. *Arteriosclerosis, thrombosis, and vascular biology* **29,** 1671–1677 (Oct. 2009).

123. Khera, A. V. & Kathiresan, S. Genetics of coronary artery disease: Discovery, biology and clinical translation. *Nature Reviews Genetics* **18,** 331–344 (2017).

124. Teslovich, T. M., Musunuru, K., Smith, A. V., *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466,** 707–713 (2010).

125. Strong, A., Ding, Q., Edmondson, A. C., *et al.* Hepatic sortilin regulates both apolipoprotein B secretion and LDL catabolism. *Journal of Clinical Investigation* **122,** 2807–2816 (Aug. 2012).

126. Kessler, T., Vilne, B. & Schunkert, H. The impact of genome-wide association studies on the pathophysiology and therapy of cardiovascular disease. *EMBO Molecular Medicine* **8,** 688–701 (July 2016).

127. Abifadel, M., Varret, M., Rabès, J.-P., *et al.* Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. *Nature Genetics* **34,** 154–156 (2003).

128. Jansen, H., Samani, N. J. & Schunkert, H. *Mendelian randomization studies in coronary artery disease* Aug. 2014.

129. The Emerging Risk Factors Collaboration. C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: an individual participant meta-analysis. *The Lancet* **375,** 132–140 (Jan. 2010).

130. Eiriksdottir, G., Harris, T. B., Launer, L. J., *et al.* Association between C reactive protein and coronary heart disease: Mendelian randomisation analysis based on individual participant data. *BMJ* **342,** 425 (Feb. 2011).

131. Voight, B. F., Peloso, G. M., Orho-Melander, M., *et al.* Plasma HDL cholesterol and risk of myocardial infarction: A mendelian randomisation study. *The Lancet* **380,** 572–580 (2012).

132. Kessler, T., Zhang, L., Liu, Z., *et al.* ADAMTS-7 inhibits re-endothelialization of injured arteries and promotes vascular remodeling through cleavage of thrombospondin-1. *Circulation* **131,** 1191–1201 (2015).

133. Bauer, R. C., Tohyama, J., Cui, J., *et al.* Knockout of Adamts7, a novel coronary artery disease locus in humans, reduces atherosclerosis in mice. *Circulation* **131,** 1202–1213 (2015).

134. Li, L., Chen, Z., von Scheidt, M., *et al.* Transcriptome-wide association study of coronary artery disease identifies novel susceptibility genes. *Basic Research in Cardiology* **117** (Dec. 2022).

135. Patel, K. V., Pandey, A. & De Lemos, J. A. Conceptual framework for addressing residual atherosclerotic cardiovascular disease risk in the era of precision medicine. *Circulation* **137,** 2551–2553 (June 2018).

136. McGrath, J., Saha, S., Chant, D. & Welham, J. Schizophrenia: A Concise Overview of Incidence, Prevalence, and Mortality. *Epidemiologic Reviews* **30,** 67–76 (Nov. 2008).

137. Laursen, T. M., Nordentoft, M. & Mortensen, P. B. Excess Early Mortality in Schizophrenia. *Annual Review of Clinical Psychology* **10,** 425–448 (Mar. 2014).

138. Michalopoulou, P. G., Lewis, S. W., Wykes, T., Jaeger, J. & Kapur, S. Treating impaired cognition in schizophrenia: The case for combining cognitive-enhancing drugs with cognitive remediation. *European Neuropsychopharmacology* **23,** 790–798 (2013).

139. Kahn, R. S., Sommer, I. E., Murray, R. M., *et al.* Schizophrenia. *Nature Reviews Disease Primers* **1** (Nov. 2015).

140. Sullivan, P. F., Kendler, K. S. & Neale, M. C. Schizophrenia as a Complex Trait: Evidence from a Meta-analysis of Twin Studies. *Archives of General Psychiatry* **60,** 1187–1192 (2003).

141. Ripke, S., Neale, B. M., Corvin, A., *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511,** 421–427 (2014).

142. Pardiñas, A. F., Holmans, P., Pocklington, A. J., *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nature Genetics* **50,** 381–389 (Mar. 2018).

143. O'Dushlaine, C., Rossin, L., Lee, P. H., *et al.* Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nature Neuroscience* **18,** 199–209 (2015).

144. Singh, T., Poterba, T., Curtis, D., *et al.* Rare coding variants in ten genes confer substantial risk for schizophrenia. *Nature* **604,** 509–516 (2022).

145. Owen, M. J., Sawa, A. & Mortensen, P. B. *Schizophrenia* July 2016.

146. Purcell, S. M., Moran, J. L., Fromer, M., *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506,** 185–190 (2014).

147. Chang, X., Lima, L. d. A., Liu, Y., *et al.* Common and Rare Genetic Risk Factors Converge in Protein Interaction Networks Underlying Schizophrenia. *Frontiers in Genetics* **9,** 434 (2018).

148. Purcell, S. M., Wray, N. R., Stone, J. L., *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460,** 748–752 (2009).

149. Smyth, A. M. & Lawrie, S. M. *The neuroimmunology of schizophrenia* Dec. 2013.

150. Sekar, A., Bialas, A. R., de Rivera, H., *et al.* Schizophrenia risk from complex variation of complement component 4. *Nature* **530,** 177–183 (2016).

151. Yilmaz, M., Yalcin, E., Presumey, J., *et al.* Overexpression of schizophrenia susceptibility factor human complement C4A promotes excessive synaptic loss and behavioral changes in mice. *Nature Neuroscience* **24,** 214–224 (2021).

152. Mokhtari, R. & Lachman, H. M. The Major Histocompatibility Complex (MHC) in Schizophrenia: A Review. *Journal of Clinical & Cellular Immunology* **07** (2016).

153. Huckins, L. M., Dobbyn, A., Ruderfer, D. M., *et al.* Gene expression imputation across multiple brain regions provides insights into schizophrenia risk. *Nature Genetics* **51,** 659–674 (Apr. 2019).

154. Hall, L. S., Medway, C. W., Pain, O., *et al.* A transcriptome-wide association study implicates specific pre-and post-synaptic abnormalities in schizophrenia. *Human Molecular Genetics* **29,** 159–167 (Jan. 2020).

155. Reay, W. R., Kiltschewskij, D. J., Geaghan, M. P., *et al. Genetic estimates of correlation and causality between blood-based biomarkers and psychiatric disorders* tech. rep. (2022), 8969.

156. Ohi, K., Sumiyoshi, C., Fujino, H., *et al.* Genetic overlap between general cognitive function and schizophrenia: A review of cognitive GWASs. *International Journal of Molecular Sciences* **19** (Dec. 2018).

157. Hubbard, L., Tansey, K. E., Rai, D., *et al.* Evidence of Common Genetic Overlap Between Schizophrenia and Cognition. *Schizophrenia Bulletin* **42,** 832–842 (2016).

158. Ruderfer, D. M., Ripke, S., McQuillin, A., *et al.* Genomic Dissection of Bipolar Disorder and Schizophrenia, Including 28 Subphenotypes. *Cell* **173,** 1705–1715 (June 2018).

159. Lee, S. I., Dudley, A. M., Drubin, D., *et al.* Learning a prior on regulatory potential from eQTL data. *PLoS Genetics* **5,** e1000358 (Jan. 2009).

160. Ardlie, K. G., DeLuca, D. S., Segrè, A. V., *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348,** 648–660 (2015).

161. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. eng. *Journal of statistical software* **33,** 1–22 (2010).

162. Svanberg, K. A Class of Globally Convergent Optimization Methods Based on Conservative Convex Separable Approximations. *SIAM Journal on Optimization* **12,** 555–573 (2002).

163. Steven G. Johnson. *The NLopt nonlinear-optimization package*

164. Hastie, T., Tibshirani, R. & Friedman, J. in *The elements of statistical learning: data mining, inference and prediction* 2nd ed., 241–249 (Springer, 2009).

165. Cawley, G. C. & Talbot, N. L. C. *On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation* tech. rep. (2010), 2079–2107.

166. Aguet, F., Brown, A. A., Castel, S. E., *et al.* Genetic effects on gene expression across human tissues. *Nature* **550,** 204–213 (2017).

167. Zou, H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101,** 1418–1429 (2006).

168. Li, Y. I., Van De Geijn, B., Raj, A., *et al.* RNA splicing is a primary link between genetic variation and disease. *Science* **352,** 600–604 (2016).

169. Ritchie, M. E., Phipson, B., Wu, D., *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43,** e47 (2015).

170. Smyth, G. K. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments: *Statistical Applications in Genetics and Molecular Biology* **3** (2004).

171. Nelder, J. A. & Wedderburn, R. W. M. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)* **135,** 370–384 (1972).

172. Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F. & Botella, J. Assessing heterogeneity in meta-analysis: Q statistic or I2 index? *Psychological methods* **11,** 193–206 (June 2006).

173. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57,** 289–300 (1995).

174. Simes, R. J. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73,** 751–754 (Dec. 1986).

175. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* **29,** 1165–1188 (Aug. 2001).

176. Pitman, E. J. G. Significance Tests Which May be Applied to Samples From any Populations. *Supplement to the Journal of the Royal Statistical Society* **4,** 119–130 (1937).

177. Yavorska, O. O. & Burgess, S. MendelianRandomization: An R package for performing Mendelian randomization analyses using summarized data. *International Journal of Epidemiology* **46,** 1734–1739 (Dec. 2017).

178. Burgess, S., Dudbridge, F. & Thompson, S. G. Combining information on multiple instrumental variables in Mendelian randomization: Comparison of allele score and summarized data methods. *Statistics in Medicine* **35,** 1880–1906 (May 2016).

179. De Leeuw, C. A., Neale, B. M., Heskes, T. & Posthuma, D. The statistical properties of gene-set analysis. *Nature Reviews Genetics* **17,** 353–364 (2016).

180. Burgess, S., Butterworth, A. S. & Thompson, J. R. Beyond Mendelian randomization: how to interpret evidence of shared genetic predictors (2016).

181. Levine, J. H., Simonds, E. F., Bendall, S. C., *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162,** 184–197 (July 2015).

182. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction (2018).

183. Wang, B., Mezlini, A. M., Demir, F., *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods* **11,** 333–337 (2014).

184. Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008,** P10008 (Oct. 2008).

185. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal* **Complex Sy,** 1695 (2006).

186. Grady, L. Random Walks for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28,** 1768–1783 (2006).

187. Kassambara, A. rstatix: Pipe-Friendly Framework for Basic Statistical Tests (2020).

188. Divine, G. W., Norton, H. J., Barón, A. E. & Juarez-Colunga, E. The Wilcoxon–Mann–Whitney Procedure Fails as a Test of Medians. *American Statistician* **72,** 278–286 (2018).

189. Cohen, J., Cohen, P., West, S. G. & Aiken, L. S. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* 3rd, 46–47 (Lawrence Erlbaum Associates, Mahwah, New Jersey, 2003).

190. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* **15,** R29 (2014).

191. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. eng. *Nature protocols* **4,** 1184–1191 (2009).

192. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols* (2012).

193. Price, A. L., Patterson, N. J., Plenge, R. M., *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38,** 904–909 (2006).

194. Klei, L., Kent, B. P., Melhem, N., Devlin, B. & Roeder, K. GemTools: A fast and efficient approach to estimating genetic ancestry (Apr. 2011).

195. Miller, C. L., Pjanic, M., Wang, T., *et al.* Integrative functional genomics identifies regulatory mechanisms at coronary artery disease loci. *Nature Communications* **7,** 12092 (2016).

196. Fullard, J. F., Giambartolomei, C., Hauberg, M. E., *et al.* Open chromatin profiling of human postmortem brain infers functional roles for non-coding schizophrenia loci. *Human Molecular Genetics* **26,** 1942–1951 (2017).

197. Samani, N. J., Erdmann, J., Hall, A. S., *et al.* Genomewide association analysis of coronary artery disease. eng. *The New England journal of medicine* **357,** 443–453 (Aug. 2007).

198. Erdmann, J., Grosshennig, A., Braund, P. S., *et al.* New susceptibility locus for coronary artery disease on chromosome 3q22.3. eng. *Nature genetics* **41,** 280–282 (Mar. 2009).

199. Erdmann, J., Willenborg, C., Nahrstaedt, J., *et al.* Genome-wide association study identifies a new locus for coronary artery disease on chromosome 10p11.23. eng. *European heart journal* **32,** 158–168 (Jan. 2011).

200. Stitziel, N. O., Won, H.-H., Morrison, A. C., *et al.* Inactivating mutations in NPC1L1 and protection from coronary heart disease. eng. *The New England journal of medicine* **371,** 2072–2082 (Nov. 2014).

201. Winkelmann, B. R., März, W., Boehm, B. O., *et al.* Rationale and design of the LURIC study–a resource for functional genomics, pharmacogenomics and long-term prognosis of cardiovascular disease. eng. *Pharmacogenomics* **2,** 1–73 (Feb. 2001).

202. Deloukas, P., Kanoni, S., Willenborg, C., *et al.* Large-scale association analysis identifies new risk loci for coronary artery disease. eng. *Nature genetics* **45,** 25–33 (Jan. 2013).

203. Millard, L. A., Davies, N. M., Gaunt, T. R., Smith, G. D. & Tilling, K. Software application profile: PHESANT: A tool for performing automated phenome scans in UK Biobank. *International Journal of Epidemiology* **47,** 29–35 (Feb. 2018).

204. Meuleman, W., Muratov, A., Rynes, E., *et al.* Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584,** 244–251 (2020).

205. Chang, C. C., Chow, C. C., Tellier, L. C., *et al.* Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience* **4** (Feb. 2015).

206. Chen, C., Li, J., Matye, D. J., Wang, Y. & Li, T. Hepatocyte sortilin 1 knockout and treatment with a sortilin 1 inhibitor reduced plasma cholesterol in Western diet-fed mice. *Journal of Lipid Research* **60,** 539–549 (2019).

207. Visel, A., Zhu, Y., May, D., *et al.* Targeted deletion of the 9p21 non-coding coronary artery disease risk interval in mice. *Nature* **464,** 409–412 (Mar. 2010).

208. Koyama, S., Ito, K., Terao, C., *et al.* Population-specific and trans-ancestry genome-wide analyses identify distinct and shared genetic risk loci for coronary artery disease. *Nature Genetics* **52,** 1169–1177 (Nov. 2020).

209. Liu, P., Choi, Y. K. & Qi, R. Z. NME7 is a functional component of the γ-tubulin ring complex. *Molecular Biology of the Cell* **25,** 2017–2025 (July 2014).

210. James, J., Zemskova, M., Eccles, C. A., *et al.* Single Mutation in the NFU1 Gene Metabolically Reprograms Pulmonary Artery Smooth Muscle Cells. *Arteriosclerosis, Thrombosis, and Vascular Biology* **41,** 734–754 (Feb. 2021).

211. Ference, B. A., Ginsberg, H. N., Graham, I., *et al.* Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement fromthe European Atherosclerosis Society Consensus Panel. *European Heart Journal* **38,** 2459–2472 (Aug. 2017).

212. McCurdy, S., Baicu, C. F., Heymans, S. & Bradshaw, A. D. Cardiac extracellular matrix remodeling: fibrillar collagens and Secreted Protein Acidic and Rich in Cysteine (SPARC). eng. *Journal of molecular and cellular cardiology* **48,** 544–549 (Mar. 2010).

213. De Vilder, E. Y., Debacker, J. & Vanakker, O. M. *GGCX-associated phenotypes: An overview in search of genotype-phenotype correlations* Feb. 2017.

214. Vogt, S., Ruppert, V., Pankuweit, S., *et al.* Myocardial insufficiency is related to reduced subunit 4 content of cytochrome c oxidase. *Journal of Cardiothoracic Surgery* **13** (Sept. 2018).

215. Richardson Id, T. G., Sanderson Id, E., Palmer Id, T. M., *et al.* Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis (2020).

216. The Heart Outcomes Prevention Evaluation Study Investigators. Effects of an Angiotensin-Converting-Enzyme inhibitor, Ramipril, on Cardiovascular Events in High-Risk Patients. *The New England Journal of Medicine* **324,** 145–153 (2000).

217. Katz, P. M. & Leiter, L. A. *Drugs Targeting High-Density Lipoprotein Cholesterol for Coronary Artery Disease Management* Nov. 2012.

218. Burgess, S., Freitag, D. F., Khan, H., Gorman, D. N. & Thompson, S. G. Using multivariable Mendelian randomization to disentangle the causal effects of lipid fractions. *PLoS ONE* **9** (Oct. 2014).

219. Van de Vegte, Y. J., Said, M. A., Rienstra, M., van der Harst, P. & Verweij, N. Genome-wide association studies and Mendelian randomization analyses for leisure sedentary behaviours. *Nature Communications* **11,** 1770 (Dec. 2020).

220. Chen, G. C., Wang, Y., Tong, X., *et al.* Cheese consumption and risk of cardiovascular disease: a meta-analysis of prospective studies. *European Journal of Nutrition* **56,** 2565–2575 (Dec. 2017).

221. Liu, D. J., Peloso, G. M., Yu, H., *et al.* Exome-wide association study of plasma lipids in >300,000 individuals. *Nature Genetics* **49,** 1758–1766 (2017).

222. Nelson, C. P., Hamby, S. E., Saleheen, D., *et al.* Genetically determined height and coronary artery disease. eng. *The New England journal of medicine* **372,** 1608–1618 (Apr. 2015).

223. Postmus, I., Trompet, S., Deshmukh, H. A., *et al.* Pharmacogenetic meta-analysis of genome-wide association studies of LDL cholesterol response to statins. *Nature Communications* **5,** 5068 (2014).

224. Kantor, E. D., Lampe, J. W., Vaughan, T. L., *et al.* Association between use of specialty dietary supplements and c-reactive protein concentrations. *American Journal of Epidemiology* **176,** 1002–1013 (Dec. 2012).

225. Ridker, P. M., Cannon, C. P., Morrow, D., *et al.* C-Reactive Protein Levels and Outcomes after Statin Therapy. *New England Journal of Medicine* **352,** 20–28 (Jan. 2005).

226. Xu, B., Ionita-Laza, I., Roos, J. L., *et al.* De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nature Genetics* **44,** 1365–1369 (Dec. 2012).

227. Hall, L. S., Pain, O., O'Brien, H. E., *et al.* Cis-effects on gene expression in the human prenatal brain associated with genetic risk for neuropsychiatric disorders. *Molecular Psychiatry* **26,** 2082–2088 (2021).

228. Otowa, T., Maher, B. S., Aggen, S. H., *et al.* Genome-Wide and Gene-Based Association Studies of Anxiety Disorders in European and African American Samples. *PLOS ONE* **9,** e112559– (Nov. 2014).

229. Li, Q., Wineinger, N. E., Fu, D.-J., *et al.* Genome-wide association study of paliperidone efficacy. *Pharmacogenetics and Genomics* **27** (2017).

230. Funk, A. J., Mccullumsmith, R. E., Haroutunian, V. & Meador-Woodruff, J. H. Abnormal Activity of the MAPK- and cAMP-Associated Signaling Pathways in Frontal Cortical Areas in Postmortem Brain in Schizophrenia. *Neuropsychopharmacology* **37,** 896–905 (2012).

231. Askland, K., Read, C., O'Connell, C. & Moore, J. H. Ion channels and schizophrenia: a gene set-based analytic approach to GWAS data for biological hypothesis testing. eng. *Human genetics* **131,** 373–391 (Mar. 2012).

232. Mei, L. & Nave, K. A. *Neuregulin-ERBB signaling in the nervous system and neuropsychiatric diseases* July 2014.

233. Parellada, E. & Gassó, P. Glutamate and microglia activation as a driver of dendritic apoptosis: a core pathophysiological mechanism to understand schizophrenia. **11,** 271 (2021).

234. Osimo, E. F., Sweeney, M., de Marvao, A., *et al.* Adipose tissue dysfunction, inflammation, and insulin resistance: alternative pathways to cardiac remodelling in schizophrenia. A multimodal, case–control study. *Translational Psychiatry* **11,** 614 (2021).

235. Savage, J. E., Jansen, P. R., Stringer, S., *et al.* Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nature Genetics* **50,** 912–919 (July 2018).

236. Fatouros-Bergman, H., Cervenka, S., Flyckt, L., Edman, G. & Farde, L. Meta-analysis of cognitive performance in drug-naïve patients with schizophrenia. *Schizophrenia Research* **158,** 156–162 (2014).

237. Olney, J. W. & Farber, N. B. Glutamate Receptor Dysfunction and Schizophrenia. *Archives of General Psychiatry* **52,** 998–1007 (Dec. 1995).

238. Habtewold, T. D., Rodijk, L. H., Liemburg, E. J., *et al.* A systematic review and narrative synthesis of data-driven studies in schizophrenia symptoms and cognitive deficits. *Translational Psychiatry* **10,** 244 (2020).

239. De Bakker, P. I. W., McVean, G., Sabeti, P. C., *et al.* A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nature Genetics* **38,** 1166–1172 (2006).

240. Tsugawa, S., Noda, Y., Tarumi, R., *et al.* Glutathione levels and activities of glutathione metabolism enzymes in patients with schizophrenia: A systematic review and meta-analysis. *Journal of Psychopharmacology* **33,** 1199–1214 (Apr. 2019).

241. Merenlender-Wagner, A., Malishkevich, A., Shemer, Z., *et al.* Autophagy has a key role in the pathophysiology of schizophrenia. *Molecular Psychiatry* **20,** 126–132 (2015).

242. Fernandes, B. S., Steiner, J., Bernstein, H.-G., *et al.* C-reactive protein is increased in schizophrenia but is not altered by antipsychotics: meta-analysis and implications. *Molecular Psychiatry* **21,** 554–564 (2016).

243. Karageorgiou, V., Milas, G. P. & Michopoulos, I. Neutrophil-to-lymphocyte ratio in schizophrenia: A systematic review and meta-analysis. *Schizophrenia Research* **206,** 4–12 (2019).

244. Zhou, X., Wang, X., Li, R., *et al.* Neutrophil-to-Lymphocyte Ratio Is Independently Associated With Severe Psychopathology in Schizophrenia and Is Changed by Antipsychotic Administration: A Large-Scale Cross-Sectional Retrospective Study. *Frontiers in Psychiatry* **11,** 581061 (2020).

245. Kelly, S., Jahanshad, N., Zalesky, A., *et al.* Widespread white matter microstructural differences in schizophrenia across 4322 individuals: Results from the ENIGMA Schizophrenia DTI Working Group. *Molecular Psychiatry* **23,** 1261–1269 (May 2018).

246. Alberti, K. G., Eckel, R. H., Grundy, S. M., *et al.* Harmonizing the metabolic syndrome: A joint interim statement of the international diabetes federation task force on epidemiology and prevention; National heart, lung, and blood institute; American heart association; World heart federation; International. *Circulation* **120,** 1640–1645 (Oct. 2009).

247. Aguirre, G. A., Ita, J. R., Garza, R. G. & Castilla-Cortazar, I. Insulin-like growth factor-1 deficiency and metabolic syndrome. *Journal of Translational Medicine* **14,** 3 (2016).

248. Papanastasiou, E. The prevalence and mechanisms of metabolic syndrome in schizophrenia: a review. *Therapeutic Advances in Psychopharmacology* **3,** 33–51 (2013).

249. Liu, C., Bousman, C. A., Pantelis, C., *et al.* Pathway-wide association study identifies five shared pathways associated with schizophrenia in three ancestral distinct populations. *Translational Psychiatry* **7,** e1037–e1037 (2017).

250. Tanra, A. J., Sabaruddin, H., Liaury, K. & Zainuddin, A. A. Effect of adjuvant vitamin c on brain-derived neurotrophic factor levels and improvement of negative symptoms in schizophrenic patients. *Open Access Macedonian Journal of Medical Sciences* **9,** 353–357 (Jan. 2021).

251. Delaneau, O., Zazhytska, M., Borel, C., *et al.* Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science* **364,** eaat8266 (May 2019).

252. Ruth, M., Alexander, P., Nihan, K., *et al.* A Common Allele on Chromosome 9 Associated with Coronary Heart Disease. *Science* **316,** 1488–1491 (June 2007).

253. Holdt, L. M., Beutner, F., Scholz, M., *et al.* ANRIL expression is associated with atherosclerosis risk at chromosome 9p21. *Arteriosclerosis, Thrombosis, and Vascular Biology* **30,** 620–627 (Mar. 2010).

254. Gupta, R. M., Hadaya, J., Trehan, A., *et al.* A Genetic Variant Associated with Five Vascular Diseases Is a Distal Regulator of Endothelin-1 Gene Expression. *Cell* **170,** 522–533 (July 2017).

255. Walker, R. L., Ramaswami, G., Hartl, C., *et al.* Genetic Control of Expression and Splicing in Developing Human Brain Informs Disease Mechanisms. *Cell* **179,** 750–771 (Oct. 2019).

256. Inloes, J. M., Hsu, K.-L., Dix, M. M., *et al.* The hereditary spastic paraplegia-related enzyme DDHD2 is a principal brain triglyceride lipase. *Proceedings of the National Academy of Sciences* **111,** 14924–14929 (Oct. 2014).

257. Šedová, L., Buková, I., Bažantová, P., *et al.* Semi-Lethal Primary Ciliary Dyskinesia in Rats Lacking the Nme7 Gene. *International Journal of Molecular Sciencies* (2021).

258. Šedová, L., Prochazka, J., Zudová, D., *et al.* Heterozygous Nme7 Mutation Affects Glucose Tolerance in Male Rats. *genes* (2021).

259. Gunturkun, M. H., Wang, T., Chitre, A. S., *et al.* Genome-Wide Association Study on Three Behaviors Tested in an Open Field in Heterogeneous Stock Rats Identifies Multiple Loci Implicated in Psychiatric Disorders. *Frontiers in Psychiatry* **13** (Feb. 2022).

260. Kocher, A. A., Schuster, M. D., Szabolcs, M. J., *et al.* Neovascularization of ischemic myocardium by human bone-marrow–derived angioblasts prevents cardiomyocyte apoptosis, reduces remodeling and improves cardiac function. *Nature Medicine* **7,** 430–436 (2001).

261. Nidorf, S. M., Fiolet, A. T. L., Mosterd, A., *et al.* Colchicine in Patients with Chronic Coronary Disease. *New England Journal of Medicine* **383,** 1838–1847 (Aug. 2020).

262. Tardif, J.-C., Kouz, S., Waters, D. D., *et al.* Efficacy and Safety of Low-Dose Colchicine after Myocardial Infarction. *New England Journal of Medicine* **381,** 2497–2505 (Nov. 2019).

263. Del Re, D. P., Amgalan, D., Linkermann, A., Liu, Q. & Kitsis, R. N. Fundamental mechanisms of regulated cell death and implications for heart disease. *Physiological Reviews* **99,** 1765–1817 (July 2019).

264. Debnath, M. Adaptive Immunity in Schizophrenia: Functional Implications of T Cells in the Etiology, Course and Treatment. *Journal of Neuroimmune Pharmacology* **10,** 610–619 (2015).

265. Flatow, J., Buckley, P. & Miller, B. J. Meta-analysis of oxidative stress in schizophrenia. *Biological Psychiatry* **74,** 400–409 (2013).

266. Narayan, S., Head, S. R., Gilmartin, T. J., Dean, B. & Thomas, E. A. Evidence for disruption of sphingolipid metabolism in schizophrenia. *Journal of Neuroscience Research* **87,** 278–288 (2009).

267. Takahashi, N., Sakurai, T., Davis, K. L. & Buxbaum, J. D. Linking oligodendrocyte and myelin dysfunction to neurocircuitry abnormalities in schizophrenia. *Progress in Neurobiology* **93,** 13–24 (2011).

268. Guzman, R. E., Alekov, A. K., Filippov, M., Hegermann, J. & Fahlke, C. Involvement of CLC-3 chloride/proton exchangers in controlling glutamatergic synaptic strength in cultured hippocampal neurons. *Frontiers in Cellular Neuroscience* **8** (May 2014).

269. Mäki-Marttunen, T., Krull, F., Bettella, F., *et al.* Alterations in Schizophrenia-Associated Genes Can Lead to Increased Power in Delta Oscillations. *Cerebral Cortex* **29,** 875–891 (Feb. 2019).

270. Tariq, S. M., Sidhu, M. S., Toth, P. P. & Boden, W. E. HDL Hypothesis: Where Do We Stand Now? (2014).

271. Merritt, K., Egerton, A., Kempton, M. J., Taylor, M. J. & McGuire, P. K. Nature of glutamate alterations in schizophrenia a meta-analysis of proton magnetic resonance spectroscopy studies. *JAMA Psychiatry* **73,** 665–674 (July 2016).

272. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* **9** (Dec. 2019).

273. Soccio, R. E. & Breslow, J. L. *Intracellular cholesterol transport* July 2004.

274. Hong, K. U., Doll, M. A., Lykoudi, A., *et al.* Acetylator Genotype-Dependent Dyslipidemia in Rats Congenic for N-Acetyltransferase 2. *Toxicology Reports* **7,** 1319–1330 (2020).

275. Antonopoulos, A. S., Margaritis, M., Lee, R., Channon, K. & Antoniades, C. *Statins as Anti-Inflammatory Agents in Atherogenesis: Molecular Mechanisms and Lessons from the Recent Clinical Trials* tech. rep. (2012), 1519–1530.

276. Cortes, A. & Brown, M. A. *Promise and pitfalls of the Immunochip* Feb. 2011.

277. Lago, S. G. & Bahn, S. *The druggable schizophrenia genome: from repurposing opportunities to unexplored drug targets* Dec. 2022.

278. Ding, Q., Strong, A., Patel, K. M., *et al.* Permanent alteration of PCSK9 with in vivo CRISPR-Cas9 genome editing. *Circulation Research* **115,** 488–492 (Aug. 2014).

279. Bhattacharya, A., García-Closas, M., Olshan, A. F., *et al.* A framework for transcriptome-wide association studies in breast cancer in diverse study populations. *Genome Biology* **21** (Feb. 2020).

A

Appendix

A.1 Differentiability and continuity of PriLer objective function

In this section, we show the derivation of partial derivative with respect to prior weights of PriLer function (see section 3.1.1). In particular, to prove the derivation of (3.8), let first compute the derivative of the sigmoid function $s(x) = \frac{1}{1+e^{-x}}$

$$\begin{aligned} s'(x) &= -\left(\frac{-e^{-x}}{(1+e^{-x})^2}\right) = \frac{1+e^{-x}}{(1+e^{-x})^2} - \frac{1}{(1+e^{-x})^2} = \\ &= \frac{1}{1+e^{-x}}\left(1 - \frac{1}{1+e^{-x}}\right) = s(x)(1-s(x)). \end{aligned} \tag{A.1}$$

Since prior coefficient v_p (as defined in (3.3)) is equivalent to $2 - 2s\left(\sum_{k=1}^{K} \gamma_k A_{p,k}\right)$, it follows that

$$s\left(\sum_{k=1}^{K} \gamma_k A_{p,k}\right) = 1 - \frac{v_p}{2}. \tag{A.2}$$

Hence, the partial derivative of the v_p with respect to γ_k can be expressed as

$$\begin{aligned} \frac{\partial v_p}{\partial \gamma_k} &= -2A_{p,k}s\left(\sum_{k=1}^{K} \gamma_k A_{p,k}\right)\left(1 - s\left(\sum_{k=1}^{K} \gamma_k A_{p,k}\right)\right) = \\ &= -A_{p,k}\left(1 - \frac{v_p}{2}\right)v_p = A_{p,k}v_p\left(\frac{v_p}{2} - 1\right). \end{aligned} \tag{A.3}$$

Finally, from (A.3) it directly follow the gradient form of the PriLer objective function with respect to prior weights as defined in (3.8).

To show that PriLer objective function (as defined in (3.7)) is also twice differentiable and continuous, we explicitly derive each entry of the corresponding Hessian matrix $\frac{\partial^2 f}{\partial \gamma_h \partial \gamma_k}$.

From each gradient entry as computed in (A.3), we obtain

$$\begin{aligned}\frac{\partial^2 v_p}{\partial\gamma_h\partial\gamma_k} &= \frac{\partial}{\partial\gamma_h}\left(A_{p,k}v_p\left(\frac{v_p}{2}-1\right)\right) = A_{p,k}\left[\frac{\partial v_p}{\partial\gamma_h}\left(\frac{v_p}{2}-1\right)+\frac{1}{2}v_p\frac{\partial v_p}{\partial\gamma_h}\right] = \\ &= A_{p,k}A_{p,h}\left[v_p\left(\frac{v_p}{2}-1\right)^2+\frac{1}{2}v_p^2\left(\frac{v_p}{2}-1\right)\right] = \\ &= A_{p,k}A_{p,h}v_p\left(\frac{v_p}{2}-1\right)\left(\frac{v_p}{2}-1+\frac{v_p}{2}\right) = \\ &= A_{p,k}A_{p,h}v_p\left(\frac{v_p}{2}-1\right)(v_p-1)\end{aligned} \tag{A.4}$$

It directly follows that (see (3.7) for f definition)

$$\frac{\partial^2 f}{\partial\gamma_h\partial\gamma_k} = \begin{cases}\sum_{n=1}^{N}\left[\sum_{p=1}^{P}A_{p,k}A_{p,h}v_p\left(\frac{v_p}{2}-1\right)(v_p-1)\mathcal{L}(\beta_p^n,\lambda_n,\alpha_n)\right], & \text{if } k\neq h\\ 2E+\sum_{n=1}^{N}\left[\sum_{p=1}^{P}A_{p,k}^2v_p\left(\frac{v_p}{2}-1\right)(v_p-1)\mathcal{L}(\beta_p^n,\lambda_n,\alpha_n)\right], & \text{if } k=h\end{cases}$$

Hence, the Hessian matrix of f exists and is continuous in each entry as a product and linear combination of continuous functions.

A.2 R^2 decomposition for PriLer model with additive confounder effects

Here we explicitly calculate the coefficient of determination R^2 decomposition for PriLer when modeling confounder effects, as defined in (3.14) (see section 3.1.3). First, we can rewrite R^2 definition 3.13 as

$$\begin{aligned}R^2 &= \frac{\|\boldsymbol{Y}-\overline{Y}\|_2^2-\|\boldsymbol{Y}-\hat{\boldsymbol{Y}}\|_2^2}{M\sigma_Y^2} = \frac{\|(\boldsymbol{Y}-\hat{\boldsymbol{Y}})+(\hat{\boldsymbol{Y}}-\overline{Y})\|_2^2-\|\boldsymbol{Y}-\hat{\boldsymbol{Y}}\|_2^2}{M\sigma_Y^2} = \\ &= \frac{\|\hat{\boldsymbol{Y}}-\overline{Y}\|_2^2+2\left\langle\hat{\boldsymbol{Y}}-\overline{Y},\boldsymbol{Y}-\hat{\boldsymbol{Y}}\right\rangle}{M\sigma_Y^2}\end{aligned} \tag{A.5}$$

where the third equality follows from Euclidean norm and Euclidean inner product definition:

$$\|\boldsymbol{x}+\boldsymbol{y}\|_2^2 = \sum_{i=1}^{n}(x_i+y_i)^2 = \sum_{i=1}^{n}(x_i)^2+\sum_{i=1}^{n}(y_i)^2+2\sum_{i=1}^{n}(x_iy_i) = \|\boldsymbol{x}\|_2^2+\|\boldsymbol{y}\|_2^2+\langle\boldsymbol{x},\boldsymbol{y}\rangle$$

Since $\boldsymbol{W}=\boldsymbol{Y}-Z\hat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{V}}=Z\hat{\boldsymbol{\mu}}$, then by definition $\boldsymbol{Y}=\boldsymbol{W}+\widehat{\boldsymbol{V}}$, $\overline{Y}=\overline{W}+\overline{\widehat{V}}$, and $\hat{\boldsymbol{Y}}=\widehat{\boldsymbol{W}}+\widehat{\boldsymbol{V}}$. It follows that

$$\begin{aligned}\|\hat{\boldsymbol{Y}}-\overline{Y}\|_2^2 &= \|\widehat{\boldsymbol{W}}+\widehat{\boldsymbol{V}}-\overline{W}-\overline{\widehat{V}}\|_2^2 = \\ &= \|\widehat{\boldsymbol{W}}-\overline{W}\|_2^2+\|\widehat{\boldsymbol{V}}-\overline{\widehat{V}}\|_2^2+2\left\langle\widehat{\boldsymbol{W}}-\overline{W},\widehat{\boldsymbol{V}}-\overline{\widehat{V}}\right\rangle.\end{aligned} \tag{A.6}$$

In addition, again by definition it results that $\boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{W} - \widehat{\boldsymbol{W}}$ and from the linearity of inner product, we then obtain

$$\begin{aligned}\left\langle \hat{\boldsymbol{Y}} - \overline{Y}, \boldsymbol{Y} - \hat{\boldsymbol{Y}} \right\rangle &= \left\langle \widehat{\boldsymbol{W}} + \widehat{\boldsymbol{V}} - \overline{W} - \overline{\widehat{V}}, \boldsymbol{W} - \widehat{\boldsymbol{W}} \right\rangle = \\ &= \left\langle \widehat{\boldsymbol{W}} - \overline{W}, \boldsymbol{W} - \widehat{\boldsymbol{W}} \right\rangle + \left\langle \widehat{\boldsymbol{V}} - \overline{\widehat{V}}, \boldsymbol{W} - \widehat{\boldsymbol{W}} \right\rangle .\end{aligned} \tag{A.7}$$

Combining (A.6) with (A.7) and keeping in mind the symmetric property of the inner product, we can rewrite R^2 in (A.5) as

$$\begin{aligned}R^2 = &\frac{\|\widehat{\boldsymbol{W}} - \overline{W}\|_2^2 + \|\widehat{\boldsymbol{V}} - \overline{\widehat{V}}\|_2^2 + 2\left\langle \widehat{\boldsymbol{W}} - \overline{W}, \widehat{\boldsymbol{V}} - \overline{\widehat{V}} \right\rangle}{M\sigma_Y^2} + \\ &+ \frac{2\left\langle \widehat{\boldsymbol{W}} - \overline{W}, \boldsymbol{W} - \widehat{\boldsymbol{W}} \right\rangle + 2\left\langle \widehat{\boldsymbol{V}} - \overline{\widehat{V}}, \boldsymbol{W} - \widehat{\boldsymbol{W}} \right\rangle}{M\sigma_Y^2} = \\ &\frac{\|\widehat{\boldsymbol{W}} - \overline{W}\|_2^2 + 2\left\langle \widehat{\boldsymbol{W}} - \overline{W}, \boldsymbol{W} - \widehat{\boldsymbol{W}} \right\rangle}{M\sigma_Y^2} + \frac{\|\widehat{\boldsymbol{V}} - \overline{\widehat{V}}\|_2^2}{M\sigma_Y^2} + \frac{2\left\langle \widehat{\boldsymbol{V}} - \overline{\widehat{V}}, \boldsymbol{W} - \overline{W} \right\rangle}{M\sigma_Y^2} = \\ &R_g^2 + R_c^2 + R_{g,c}^2\end{aligned} \tag{A.8}$$

Appendix Tables

B

B.1 PriLer

# GEO Accession	Sample Name	Type	Name in PriLer
GSM906394	adipose	ChIP-Seq H3K27ac	AdiposeNuclei
GSM916066	adipose nuclei	ChIP-Seq H3K27ac	AdiposeNuclei
GSM896163	adrenal gland	ChIP-Seq H3K27ac	adrenal_gland
GSM1013126	adrenal gland	ChIP-Seq H3K27ac	adrenal_gland
GSM1120339	adrenal gland	ChIP-Seq H3K27ac	adrenal_gland
GSM1160190	adrenal gland, fetal day97 M	ChIP-Seq H3K27ac	adrenal_gland
GSM1273660	AFG	ChIP-Seq H3K27ac	H7_dAFG
GSM1273650	APS	ChIP-Seq H3K27ac	H7_dAPS
GSM773016	brain, angular gyrus	ChIP-Seq H3K27ac	angular_gyrus
GSM1112807	brain, angular gyrus	ChIP-Seq H3K27ac	angular_gyrus
GSM772832	brain, anterior caudate	ChIP-Seq H3K27ac	anterior_caudate
GSM1112811	brain, anterior caudate	ChIP-Seq H3K27ac	anterior_caudate
GSM773011	brain, cingulate gyrus	ChIP-Seq H3K27ac	cingulate_gyrus
GSM1112813	brain, cingulate gyrus	ChIP-Seq H3K27ac	cingulate_gyrus
GSM773020	brain, hippocampus middle	ChIP-Seq H3K27ac	hippocampus_middle
GSM916035	brain, hippocampus middle	ChIP-Seq H3K27ac	hippocampus_middle
GSM1112791	brain, hippocampus middle	ChIP-Seq H3K27ac	hippocampus_middle
GSM772995	brain, inferior temporal lobe	ChIP-Seq H3K27ac	inferior_temporal_lobe
GSM1112812	brain, inferior temporal lobe	ChIP-Seq H3K27ac	inferior_temporal_lobe
GSM773015	brain, mid frontal, Brodmann area 9/46, dorsolateral prefrontal cortex	ChIP-Seq H3K27ac	mid_frontal_lobe
GSM1112810	brain, mid frontal, Brodmann area 9/46, dorsolateral prefrontal cortex	ChIP-Seq H3K27ac	mid_frontal_lobe
GSM997258	brain, substantia nigra	ChIP-Seq H3K27ac	substantia_nigra
GSM1112778	brain, substantia nigra	ChIP-Seq H3K27ac	substantia_nigra
GSM1102782	CD14 primary cells	ChIP-Seq H3K27ac	CD14_primary_cells
GSM1027287	CD19 primary cells	ChIP-Seq H3K27ac	CD19_primary_cells
GSM1058764	CD3 primary cells	ChIP-Seq H3K27ac	CD3_primary_cells
GSM772885	CD34 mobilized primary cells	ChIP-Seq H3K27ac	CD34
GSM772894	CD34 mobilized primary cells	ChIP-Seq H3K27ac	CD34
GSM772963	CD4 memory primary cells	ChIP-Seq H3K27ac	CD45RO_CD4
GSM772997	CD4 memory primary cells	ChIP-Seq H3K27ac	CD45RO_CD4
GSM772835	CD4 naive primary cells	ChIP-Seq H3K27ac	CD45RA_CD4
GSM772934	CD4 naive primary cells	ChIP-Seq H3K27ac	CD45RA_CD4
GSM773004	CD4+ CD25- CD45RA+ naive primary cells	ChIP-Seq H3K27ac	CD25_CD45RA_naive
GSM997239	CD4+ CD25- Th primary cells	ChIP-Seq H3K27ac	CD25_Th
GSM997233	CD4+ CD25+ CD127- Treg primary cells	ChIP-Seq H3K27ac	CD25_CD127_Treg
GSM916026	CD4+ CD25int CD127+ Tmem primary cells	ChIP-Seq H3K27ac	CD25int_CD127_Tmem
GSM997260	CD4+ CD25int CD127+ Tmem primary cells	ChIP-Seq H3K27ac	CD25int_CD127_Tmem
GSM1027288	CD56 primary cells	ChIP-Seq H3K27ac	CD56_primary_cells
GSM772880	CD8 memory primary cells	ChIP-Seq H3K27ac	CD45RO_CD8
GSM772949	CD8 naive primary cells	ChIP-Seq H3K27ac	CD45RA_CD8
GSM772976	CD8 naive primary cells	ChIP-Seq H3K27ac	CD45RA_CD8
GSM1112780	colon smooth muscle	ChIP-Seq H3K27ac	colon_smooth_muscle
GSM1112779	colonic mucosa	ChIP-Seq H3K27ac	colonic_mucosa
GSM1112802	colonic mucosa	ChIP-Seq H3K27ac	colonic_mucosa
GSM1112790	duodenum mucosa	ChIP-Seq H3K27ac	Duodenum_mucosa
GSM1521740	ERG	ChIP-Seq H3K27ac	ERG
GSM1521745	ERG	ChIP-Seq H3K27ac	ERG
GSM906393	esophagus	ChIP-Seq H3K27ac	dTrophoblastesophagus
GSM1013127	esophagus	ChIP-Seq H3K27ac	dTrophoblastesophagus
GSM2199917	FPC_neuronal_ATAC_R2	ATAC-Seq	FPC_neuronal_ATAC_R2
GSM2199919	FPC_neuronal_ATAC_R4	ATAC-Seq	FPC_neuronal_ATAC_R4
GSM910555	gastric	ChIP-Seq H3K27ac	gastric
GSM1013122	gastric	ChIP-Seq H3K27ac	gastric
GSM1013128	gastric	ChIP-Seq H3K27ac	gastric
GSM1227053	gastric	ChIP-Seq H3K27ac	gastric
GSM753425	H1 BMP4 derived mesendoderm cultured cells	ChIP-Seq H3K27ac	dMES

GSM753426	H1 BMP4 derived mesendoderm cultured cells	ChIP-Seq H3K27ac	dMES
GSM864035	H1 BMP4 derived mesendoderm cultured cells	ChIP-Seq H3K27ac	dMES
GSM864799	H1 BMP4 derived mesendoderm cultured cells	ChIP-Seq H3K27ac	dMES
GSM767341	H1 derived mesenchymal stem cells	ChIP-Seq H3K27ac	dMesenchymal
GSM767342	H1 derived mesenchymal stem cells	ChIP-Seq H3K27ac	dMesenchymal
GSM753429	H1 derived neuronal progenitor cultured cells	ChIP-Seq H3K27ac	dNPCs
GSM767343	H1 derived neuronal progenitor cultured cells	ChIP-Seq H3K27ac	dNPCs
GSM818031	H1 derived neuronal progenitor cultured cells	ChIP-Seq H3K27ac	dNPCs
GSM896162	H1 derived neuronal progenitor cultured cells	ChIP-Seq H3K27ac	dNPCs
GSM956008	H1 derived neuronal progenitor cultured cells	ChIP-Seq H3K27ac	dNPCs
GSM1273645	H7_ESC	ChIP-Seq H3K27ac	hESC_H7
GSM1876021	HCASMC_ATAC_Control_D1	ATAC-Seq	CAD_H3K27ac_ATAC_merged
GSM1876025	HCASMC_ATAC_Control_D2	ATAC-Seq	CAD_H3K27ac_ATAC_merged
GSM1876036	HCASMC_ChIP_H3K27ac_D1	ChIP-Seq H3K27ac	CAD_H3K27ac_ATAC_merged
GSM1876037	HCASMC_ChIP_H3K27ac_D2	ChIP-Seq H3K27ac	CAD_H3K27ac_ATAC_merged
GSM1876038	HCASMC_ChIP_H3K27ac_D3	ChIP-Seq H3K27ac	CAD_H3K27ac_ATAC_merged
GSM906392	heart, aorta	ChIP-Seq H3K27ac	heart_aorta
GSM1227055	heart, aorta	ChIP-Seq H3K27ac	heart_aorta
GSM906396	heart, left ventricle	ChIP-Seq H3K27ac	heart_left_ventricle
GSM908951	heart, left ventricle	ChIP-Seq H3K27ac	heart_left_ventricle
GSM1127173	heart, left ventricle	ChIP-Seq H3K27ac	heart_left_ventricle
GSM910557	heart, right atrium	ChIP-Seq H3K27ac	heart_right_atrium
GSM1013124	heart, right ventricle	ChIP-Seq H3K27ac	heart_right_ventricle
GSM1220280	heart, right ventricle	ChIP-Seq H3K27ac	heart_right_ventricle
GSM1112830	hESC-derived CD184+ endoderm cultured cells	ChIP-Seq H3K27ac	HUES64_derived_CD184
GSM1112831	hESC-derived CD184+ endoderm cultured cells	ChIP-Seq H3K27ac	HUES64_derived_CD184
GSM1112824	hESC-derived CD56+ ectoderm cultured cells	ChIP-Seq H3K27ac	CD56_ectoderm
GSM1112829	hESC-derived CD56+ ectoderm cultured cells	ChIP-Seq H3K27ac	CD56_ectoderm
GSM1112825	hESC-derived CD56+ mesoderm cultured cells	ChIP-Seq H3K27ac	CD56_mesoderm
GSM1112832	hESC-derived CD56+ mesoderm cultured cells	ChIP-Seq H3K27ac	CD56_mesoderm
GSM469966	IMR90 cell line	ChIP-Seq H3K27ac	IMR90_cell_line
GSM469967	IMR90 cell line	ChIP-Seq H3K27ac	IMR90_cell_line
GSM1112799	kidney	ChIP-Seq H3K27ac	kidney
GSM1112806	kidney	ChIP-Seq H3K27ac	kidney
GSM1058765	large intestine, fetal day108 M	ChIP-Seq H3K27ac	large_intestine_fetal_day108_M
GSM1112808	liver	ChIP-Seq H3K27ac	liver
GSM1112809	liver	ChIP-Seq H3K27ac	liver
GSM906395	lung	ChIP-Seq H3K27ac	lung
GSM1013123	lung	ChIP-Seq H3K27ac	lung
GSM1273670	MHG	ChIP-Seq H3K27ac	H7_dMHG
GSM1521750	MRG	ChIP-Seq H3K27ac	MRG
GSM1521755	MRG	ChIP-Seq H3K27ac	MRG
GSM1058767	muscle, leg, fetal day110 F	ChIP-Seq H3K27ac	muscle_leg_fetal_day110_F
GSM1160189	muscle, trunk, fetal day115 F	ChIP-Seq H3K27ac	muscle_trunk_fetal_day115_F
GSM1521730	NE	ChIP-Seq H3K27ac	NE
GSM1521735	NE	ChIP-Seq H3K27ac	NE
GSM1876029	Normal_Artery_ATAC_D1	ATAC-Seq	CAD_H3K27ac_ATAC_merged
GSM1876030	Normal_Artery_ATAC_D2	ATAC-Seq	CAD_H3K27ac_ATAC_merged
GSM1876031	Normal_Artery_ATAC_D3	ATAC-Seq	CAD_H3K27ac_ATAC_merged
GSM956009	ovary	ChIP-Seq H3K27ac	ovary
GSM906397	pancreas	ChIP-Seq H3K27ac	pancreas
GSM1013129	pancreas	ChIP-Seq H3K27ac	pancreas
GSM1273665	PFG	ChIP-Seq H3K27ac	PFG
GSM1102784	placenta, day 113	ChIP-Seq H3K27ac	placenta_day_113
GSM910556	psoas muscle	ChIP-Seq H3K27ac	psoas_muscle
GSM1013130	psoas muscle	ChIP-Seq H3K27ac	psoas_muscle
GSM1127171	psoas muscle	ChIP-Seq H3K27ac	psoas_muscle
GSM1227054	psoas muscle	ChIP-Seq H3K27ac	psoas_muscle
GSM1112795	rectal mucosa	ChIP-Seq H3K27ac	rectal_mucosa
GSM1112801	rectal mucosa	ChIP-Seq H3K27ac	rectal_mucosa
GSM1112796	rectal smooth muscle	ChIP-Seq H3K27ac	rectal_smooth_muscle
GSM910559	sigmoid colon	ChIP-Seq H3K27ac	sigmoid_colon
GSM915331	sigmoid colon	ChIP-Seq H3K27ac	sigmoid_colon
GSM916064	skeletal muscle	ChIP-Seq H3K27ac	skeletal_muscle
GSM1058766	small intestine, fetal day108 M	ChIP-Seq H3K27ac	small_intestine_fetal_day108_M
GSM906398	spleen	ChIP-Seq H3K27ac	spleen
GSM1013132	spleen	ChIP-Seq H3K27ac	spleen
GSM1120338	spleen	ChIP-Seq H3K27ac	spleen
GSM1013125	thymus	ChIP-Seq H3K27ac	thymus
GSM1027289	thymus, fetal day110 F	ChIP-Seq H3K27ac	Fetal_Thymus_d110
GSM733771	GM12878	ChIP-Seq H3K27ac	GM12878

Tab. B.1.: GEO accession number and corresponding name for prior features derived from ChIP-Seq H3k27ac and ATAC-seq epigenetic data, the same "Name in PriLer" refers to merged features

Tissue	n. priors	Prior features
Dorsolateral Prefrontal Cortex	15	ERG, MRG, substantia_nigra, anterior_caudate, mid_frontal_lobe, angular_gyrus, cingulate_gyrus, hippocampus_middle, inferior_temporal_lobe, NE, dNPCs, FPC_neuronal_ATAC_R2, FPC_neuronal_ATAC_R4, Ctrl_150_allPeaks_cellRanger, PGC_gwas_bin
Adipose Subcutaneous	3	AdiposeNuclei, dMES, CAD_gwas_bin
Adipose Visceral Omentum	3	AdiposeNuclei, dMES, CAD_gwas_bin
Adrenal Gland	1	adrenal_gland
Artery Aorta	7	heart_aorta, heart_left_ventricle, heart_right_atrium, heart_right_ventricle, CD56_mesoderm, CAD_H3K27ac_ATAC_merged, CAD_gwas_bin
Artery Coronary	7	heart_aorta, heart_left_ventricle, heart_right_atrium, heart_right_ventricle, CD56_mesoderm, CAD_H3K27ac_ATAC_merged, CAD_gwas_bin
Artery Tibial	7	heart_aorta, heart_left_ventricle, heart_right_atrium, heart_right_ventricle, CD56_mesoderm, CAD_H3K27ac_ATAC_merged, CAD_gwas_bin
Brain Caudate basal ganglia	15	ERG, MRG, substantia_nigra, anterior_caudate, mid_frontal_lobe, angular_gyrus, cingulate_gyrus, hippocampus_middle, inferior_temporal_lobe, NE, dNPCs, FPC_neuronal_ATAC_R2, FPC_neuronal_ATAC_R4, Ctrl_150_allPeaks_cellRanger, PGC_gwas_bin
Brain Cerebellar Hemisphere	15	ERG, MRG, substantia_nigra, anterior_caudate, mid_frontal_lobe, angular_gyrus, cingulate_gyrus, hippocampus_middle, inferior_temporal_lobe, NE, dNPCs, FPC_neuronal_ATAC_R2, FPC_neuronal_ATAC_R4, Ctrl_150_allPeaks_cellRanger, PGC_gwas_bin
Brain Cerebellum	15	ERG, MRG, substantia_nigra, anterior_caudate, mid_frontal_lobe, angular_gyrus, cingulate_gyrus, hippocampus_middle, inferior_temporal_lobe, NE, dNPCs, FPC_neuronal_ATAC_R2, FPC_neuronal_ATAC_R4, Ctrl_150_allPeaks_cellRanger, PGC_gwas_bin
Brain Cortex	15	ERG, MRG, substantia_nigra, anterior_caudate, mid_frontal_lobe, angular_gyrus, cingulate_gyrus, hippocampus_middle, inferior_temporal_lobe, NE, dNPCs, FPC_neuronal_ATAC_R2, FPC_neuronal_ATAC_R4, Ctrl_150_allPeaks_cellRanger, PGC_gwas_bin
Brain Frontal Cortex BA9	15	ERG, MRG, substantia_nigra, anterior_caudate, mid_frontal_lobe, angular_gyrus, cingulate_gyrus, hippocampus_middle, inferior_temporal_lobe, NE, dNPCs, FPC_neuronal_ATAC_R2, FPC_neuronal_ATAC_R4, Ctrl_150_allPeaks_cellRanger, PGC_gwas_bin
Brain Hippocampus	15	ERG, MRG, substantia_nigra, anterior_caudate, mid_frontal_lobe, angular_gyrus, cingulate_gyrus, hippocampus_middle, inferior_temporal_lobe, NE, dNPCs, FPC_neuronal_ATAC_R2, FPC_neuronal_ATAC_R4, Ctrl_150_allPeaks_cellRanger, PGC_gwas_bin
Brain Hypothalamus	15	ERG, MRG, substantia_nigra, anterior_caudate, mid_frontal_lobe, angular_gyrus, cingulate_gyrus, hippocampus_middle, inferior_temporal_lobe, NE, dNPCs, FPC_neuronal_ATAC_R2, FPC_neuronal_ATAC_R4, Ctrl_150_allPeaks_cellRanger, PGC_gwas_bin
Brain Nucleus accumbens basal_ganglia	15	ERG, MRG, substantia_nigra, anterior_caudate, mid_frontal_lobe, angular_gyrus, cingulate_gyrus, hippocampus_middle, inferior_temporal_lobe, NE, dNPCs, FPC_neuronal_ATAC_R2, FPC_neuronal_ATAC_R4, Ctrl_150_allPeaks_cellRanger, PGC_gwas_bin
Cells EBV-transformed lymphocytes	2	GM12878, PGC_gwas_bin
Colon Sigmoid	8	colonic_mucosa, Duodenum_mucosa, gastric, large_intestine_fetal_day108_M, rectal_mucosa, rectal_smooth_muscle, small_intestine_fetal_day108_M, CD56_mesoderm
Colon Transverse	8	colonic_mucosa, Duodenum_mucosa, gastric, large_intestine_fetal_day108_M, rectal_mucosa, rectal_smooth_muscle, small_intestine_fetal_day108_M, CD56_mesoderm
Esophagus Gastroesophageal Junction	8	colonic_mucosa, Duodenum_mucosa, gastric, large_intestine_fetal_day108_M, rectal_mucosa, rectal_smooth_muscle, small_intestine_fetal_day108_M, CD56_mesoderm
Esophagus Mucosa	8	colonic_mucosa, Duodenum_mucosa, gastric, large_intestine_fetal_day108_M, rectal_mucosa, rectal_smooth_muscle, small_intestine_fetal_day108_M, CD56_mesoderm
Esophagus Muscularis	8	colonic_mucosa, Duodenum_mucosa, gastric, large_intestine_fetal_day108_M, rectal_mucosa, rectal_smooth_muscle, small_intestine_fetal_day108_M, CD56_mesoderm
Heart Atrial Appendage	7	heart_aorta, heart_left_ventricle, heart_right_atrium, heart_right_ventricle, CD56_mesoderm, CAD_H3K27ac_ATAC_merged, CAD_gwas_bin
Heart Left Ventricle	7	heart_aorta, heart_left_ventricle, heart_right_atrium, heart_right_ventricle, CD56_mesoderm, CAD_H3K27ac_ATAC_merged, CAD_gwas_bin
Liver	2	liver, CAD_gwas_bin
Lung	1	lung
Muscle Skeletal	2	skeletal_muscle, HSMM
Pancreas	1	pancreas
Skin Not Sun Exposed Suprapubic	1	CD56_ectoderm
Skin Sun Exposed Lower leg	1	CD56_ectoderm
Small Intestine Terminal Ileum	8	colonic_mucosa, Duodenum_mucosa, gastric, large_intestine_fetal_day108_M, rectal_mucosa, rectal_smooth_muscle, small_intestine_fetal_day108_M, CD56_mesoderm
Spleen	1	spleen
Stomach	8	colonic_mucosa, Duodenum_mucosa, gastric, large_intestine_fetal_day108_M, rectal_mucosa, rectal_smooth_muscle, small_intestine_fetal_day108_M, CD56_mesoderm
Thyroid	1	HUES64_derived_CD184
Whole Blood	6	CD14_primary_cells, CD19_primary_cells, CD3_primary_cells, CD56_primary_cells, HUVEC, GM12878

Tab. B.2.: Prior features considered in each PriLer tissue specific model when assessing PriLer performances

Tissues	PriLer	el-net	TWAS	prediXcan
Adipose_Subcutaneous	34.63	28.88	37.82	35.34
Adipose_Visceral_Omentum	32.01	29.51	37.36	35.66
Adrenal_Gland	32.18	29.29	36.64	35.49
Artery_Aorta	33.3	29.02	37.5	35.05
Artery_Coronary	33.41	28.51	36.65	34.93
Artery_Tibial	34.86	29.01	37.31	34.93
Brain_Caudate_basal_ganglia	34.46	29	37.54	35.02
Brain_Cerebellar_Hemisphere	34.69	28.86	37.38	35.02
Brain_Cerebellum	33.85	29.37	37.91	35.53
Brain_Cortex	34.66	29.19	37.56	35.42
Brain_Frontal_Cortex_BA9	34.19	28.8	36.25	35.73
Brain_Hippocampus	36.49	29.07	36.68	35.46
Brain_Hypothalamus	33.86	29	37.35	35.33
Brain_Nucleus_accumbens_basal_ganglia	33.16	28.93	37.63	35.37
Cells_EBV.transformed_lymphocytes	31.63	28.84	37.48	35.56
Colon_Sigmoid	35.08	29.31	37.14	35.48
Colon_Transverse	33.25	29.04	37.03	35.21
Esophagus_Gastroesophageal_Junction	32.55	28.92	37.91	35.54
Esophagus_Mucosa	34.21	29.76	38.28	35.66
Esophagus_Muscularis	32.87	29.1	37.35	36.07
Heart_Atrial_Appendage	32.41	29.37	37.14	35.01
Heart_Left_Ventricle	36.77	29.76	37.44	35.41
Liver	33.31	29.32	36.04	35.18
Lung	31.66	29.15	37.98	35.57
Muscle_Skeletal	36.1	29.39	38.4	35.23
Pancreas	32.41	29.55	38.6	35.36
Skin_Not_Sun_Exposed_Suprapubic	30.82	29.27	37.47	35.55
Skin_Sun_Exposed_Lower_leg	31.33	29.82	38.36	35.71
Small_Intestine_Terminal_Ileum	32.89	28.48	37.6	35.39
Spleen	32.35	29.73	38.11	35.72
Stomach	33.23	29.22	37.67	35.53
Thyroid	31.73	28.69	38.06	35.28
Whole_Blood	35.75	30.71	40.36	36.46

Tab. B.3.: Percentage of reg-SNPs intersecting at least one DNase I hypersensitive site across four gene expression imputation methods

B.2 Coronary Artery Disease

Phenotype Class	N.	Phenotype UKBB ids	Covariates
Blood_biochemistry	30	30600,30610,30620,30630,30640,30650, 30660,30670,30680,30690,30700,30710, 30720,30730,30740,30750,30760,30770, 30780,30790,30800,30810,30820,30830, 30840,30850,30860,30870,30880,30890	PC 1-10,Sex,Age, 6154_1,6154_2,6154_3,6154_5, 6154_6,6154_4,6155_3,6155_7, 6155_2,6155_5,6155_4,6155_1, 6155_6,6179_2,6179_1,6179_3, 6179_4,6179_5,6179_6,6153_2, 6153_1,6153_3,6177_1,6177_2, 6177_3
Blood_count	31	30000,30010,30020,30030,30040,30050, 30060,30070,30080,30090,30100,30110, 30120,30130,30140,30150,30160,30170, 30180,30190,30200,30210,30220,30230, 30240,30250,30260,30270,30280,30290, 30300	PC 1-10,Sex,Age, 6154_1,6154_2,6154_3,6154_5, 6154_6,6154_4,6155_3,6155_7, 6155_2,6155_5,6155_4,6155_1, 6155_6,6179_2,6179_1,6179_3, 6179_4,6179_5,6179_6,6153_2, 6153_1,6153_3,6177_1,6177_2, 6177_3
Blood_pressure	3	102,4079,4080	PC 1-10,Sex
Body_size_measures	4	48,49,21001,21002	PC 1-10,Sex
Bone-densitometry_of_heel	6	78,3143,3144,3146,3147,3148	PC 1-10,Sex
Diet	23	1488,1289,1299,1309,1319,1329,1339, 1349,1359,1369,1379,1389,1408,1428_3, 1428_0,1428_1,1428_2,1438,1458,1478, 1498,1528,1548	PC 1-10,Sex
Family_history	40	20107_1,20107_100,20107_2,20107_6, 20107_8,20107_4,20107_3,20107_11, 20107_9,20107_10,20107_13,20107_12, 20107_101,20110_100,20110_1,20110_3, 20110_8,20110_4,20110_5,20110_11, 20110_2,20110_10,20110_9,20110_6, 20110_12,20110_101,20111_100,20111_12, 20111_13,20111_3,20111_8,20111_9,20111_5, 20111_6,20111_10,20111_1,20111_4,20111_2, 20111_11,20111_101	PC 1-10,Sex
Hand_grip_strength	2	46,47	PC 1-10,Sex
ICD10_Circulatory_system	64	41270_I05,41270_I07,41270_I08,41270_I10, 41270_I11,41270_I12,41270_I15,41270_I20, 41270_I21,41270_I22,41270_I23,41270_I24, 41270_I25,41270_I26,41270_I27,41270_I28, 41270_I30,41270_I31,41270_I33,41270_I34, 41270_I35,41270_I36,41270_I37,41270_I38, 41270_I42,41270_I44,41270_I45,41270_I46, 41270_I47,41270_I48,41270_I49,41270_I50, 41270_I51,41270_I60,41270_I61,41270_I62, 41270_I63,41270_I64,41270_I65,41270_I66, 41270_I67,41270_I69,41270_I70,41270_I71, 41270_I72,41270_I73,41270_I74,41270_I77, 41270_I78,41270_I79,41270_I80,41270_I81, 41270_I82,41270_I83,41270_I84,41270_I85, 41270_I86,41270_I87,41270_I88,41270_I89, 41270_I95,41270_I97,41270_I98,41270_I99	PC 1-10,Sex
ICD10_Endocrine	36	41270_E03,41270_E04,41270_E05,41270_E06, 41270_E07,41270_E10,41270_E11,41270_E13, 41270_E14,41270_E16,41270_E20,41270_E21, 41270_E22,41270_E23,41270_E24,41270_E26, 41270_E27,41270_E28,41270_E29,41270_E34, 41270_E46,41270_E53,41270_E55,41270_E61, 41270_E65,41270_E66,41270_E73,41270_E74, 41270_E78,41270_E80,41270_E83,41270_E85, 41270_E86,41270_E87,41270_E88,41270_E89	PC 1-10,Sex
Impedance_measures	5	23099,23100,23101,23102,23105	PC 1-10,Sex
Mental_health	26	1920,1930,1940,1950,1960,1970,1980,1990, 2000,2010,2020,2030,2040,4526,4537,4548, 4559,4570,4581,4609,4620,5375,5386,5663,5674,20127	PC 1-10,Sex
Physical_activity	32	864,874,884,894,904,914,924,943,971,981,991, 1001,1011,1021,1070,1080,1090,2624,2634, 3637,3647,6162_2,6162_3,6162_1,6162_4,6162_100, 6164_4,6164_1,6164_2,6164_100,6164_5,6164_3	PC 1-10,Sex
Smoking	14	1239,1249,1259,1269,1279,2644,2867,3436, 3456,5959,20116_0,20116_1,20116_2,20160	PC 1-10,Sex

Tab. B.4.: UK Biobank phenotypes included in correlation and Mendelian Randomization analysis with respect to CAD. "Covariates" column refers to the confounders used in PALAS and TWAS for phenotypes in that class, with blood biochemistry and count including ids from the "Medication" class.

Loci	ngenes	gene	tissue	best CAD Z-stat	Group significant	best WMW gene	best WMW estimate	best WMW pvalue
chr1:109.4-110.5Mb	6	RP5-1065J22.8,SYPL2, AMIGO1,AC000032.2, RP4-735C1.4,GSTM3	AS	-4.5806886045	gr1,gr2,gr3, gr4,gr5	RP5-1065J22.8	-0.25132,-0.17771,0.48696, -0.27048,0.97473	3.77e-73,3.24e-24,4.8e-234, 4.11e-81,1.16e-105
chr2:39.5-39.9Mb	1	AC007246.3	AS	1.2437692597	gr2	AC007246.3	-0.03977	0.0000181
chr6:25.8-33.7Mb	98	*MHC-locus genes*	AS	-4.2854597528	gr1,gr2,gr4	HLA-B	-0.77594,-0.08691,0.87628	0,1.99e-05,0
chr12:58.7-59.1Mb	1	RP11-362K2.2	AS	0.4684165245	gr4	RP11-362K2.2	-0.00446	0.000179
chr12:110.1-110.5Mb	1	GLTP	AS	-0.3364592496	gr2	GLTP	-0.11059	0.000000201
chr12:112.1-113.5Mb	6	AC003029.1,RP3-462E2.5, ADAM1A,TMEM116, HECTD4,OAS1	AS	5.2893887686	gr1,gr2,gr4	TMEM116	-0.04571,2.36515,-0.03551	3.4e-125,0,6.22e-78
chr17:7.3-7.7Mb	1	TNFSF12	AS	-0.0390065888	gr1	TNFSF12	0.01924	0.0000288
chr20:0.5-0.9Mb	1	SLC52A3	AS	-0.7277762281	gr1	SLC52A3	0.03398	0.0000895
chr1:109.6-110.5Mb	4	CELSR2,AMIGO1, RP4-735C1.4,GSTM3	AVO	-1.6666391221	gr1,gr2,gr3, gr4,gr5	RP4-735C1.4	-0.09594,-0.12862,0.19399, -0.12219,0.32612	3.66e-08,7.67e-09,3.59e-33, 8.32e-12,4.36e-18
chr6:25.5-33.7Mb	68	MHC-locus genes	AVO	-4.3793228088	gr1,gr2,gr4	HLA-B	-0.85333,-0.08019,0.96545	0,2.62e-05,0
chr12:10.9-11.3Mb	1	TAS2R15	AVO	-1.620381613	gr4	TAS2R15	0.02217	0.00016
chr12:110.7-112.7Mb	7	FAM216A,VPS29, FAM109A,AC003029.1, ADAM1A,TMEM116, NAA25	AVO	-6.1460024312	gr1,gr2,gr4	TMEM116	-0.14798,2.15816,-0.09355	3.98e-166,0,3.94e-99
chr1:109.4-110.4Mb	6	TMEM167B,SARS, MYBPHL,GSTM2, GSTM1,AC000032.2	AG	3.3344466291	gr1,gr2,gr3, gr4,gr5	SARS	0.40403,0.3579,-0.69966, 0.44269,-1.2753	2.33e-127,7.25e-64,0, 1.23e-146,2.05e-235
chr6:25.5-33.9Mb	63	MHC-locus genes	AG	-4.1103281393	gr1,gr2,gr4	HLA-DQB1-AS1	-0.46386,-0.07382,0.55972	6.9e-198,7.7e-06,1.08e-256
chr8:10.2-10.6Mb	1	PRSS51	AG	-1.0681284458	gr1	PRSS51	0.06507	0.000144
chr12:112.1-113.8Mb	2	ADAM1A,RP11-545P7.4	AG	3.2311146576	gr1,gr2,gr4	ADAM1A	0.03358,-1.59637,0.02995	4.79e-46,2.92e-212,5.79e-37
chr14:24.4-24.8Mb	1	EMC9	AG	2.3687233506	gr1	EMC9	-0.00907	0.00012
chr1:109.3-110.5Mb	8	CLCC1,WDR47, CELSR2,SYPL2, AMIGO1,GSTM5, RP4-735C1.4,GSTM3	AA	3.4111748284	gr1,gr2,gr3, gr4,gr5	CELSR2	0.02761,0.02442,-0.06886, 0.02981,-1.16541	1.99e-36,2.62e-20,2.37e-104, 2.65e-40,3.39e-102
chr2:203.7-204.1Mb	1	NBEAL1	AA	-7.5643682153	gr1	NBEAL1	0.00876	0.0000756
chr6:25.8-33.9Mb	82	MHC-locus genes	AA	4.5802484347	gr1,gr2,gr4	HCG27	0.52707,0.08608,-0.65365	6.51e-229,1.58e-05,4.73e-297
chr12:112-112.7Mb	5	ALDH2,AC003029.1, MAPKAPK5-AS1, ADAM1A,TMEM116	AA	-5.4413874263	gr1,gr2,gr4	TMEM116	-0.10892,1.95019,-0.0882	1.74e-82,0,4.26e-56
chr1:109.8-110.5Mb	6	SYPL2,AMIGO1, GSTM2,GSTM1, AC000032.2,GSTM5	AC	3.0319888977	gr1,gr2,gr3, gr4,gr5	SYPL2	0.13727,0.10924,-0.26075, 0.13562,-0.60034	3.6e-24,7.55e-11,6.17e-68, 2.29e-22,1.25e-42
chr4:77.6-78Mb	1	SOWAHB	AC	1.3555106515	gr1	SOWAHB	0.00809	0.0000204
chr6:26.2-28.4Mb	6	BTN3A2,RP11-457M11.5, ZNF391,RP1-265C24.5, AL022393.7,RP5-874C20.3	AC	-1.19229995	gr1,gr4	BTN3A2	0.12674,-0.13485	2.23e-55,3.07e-48
chr6:29.4-34Mb	49	MHC-locus genes	AC	-3.751226577	gr1,gr2,gr4	HLA-DRB1	-0.13472,-0.03993,0.28299	1.84e-66,1.94e-07,1.6e-114
chr12:110.1-110.5Mb	1	GLTP	AC	-1.0386371669	gr2	GLTP	-0.01984	0.00000563
chr12:112.1-113.7Mb	4	AC003029.1,ADAM1A, TMEM116,DTX1	AC	-4.9494561622	gr1,gr2,gr4	TMEM116	-0.10179,2.14716,-0.08338	6.93e-83,0,2.89e-61
chr15:50.8-51.2Mb	1	RP11-507J18.2	AC	-0.5857644003	gr1	RP11-507J18.2	-0.00415	0.000214
chr1:109.4-110.5Mb	6	TMEM167B,CELSR2, SYPL2,AMPD2, RP4-735C1.4,GSTM3	CS	4.5836182485	gr1,gr2,gr3, gr4,gr5	CELSR2	0.30218,0.25229,-0.63743, 0.3044,-1.37487	2.56e-122,4.35e-58,0, 2.92e-116,1.52e-242
chr6:26-34Mb	63	MHC-locus genes	CS	4.4873970636	gr1,gr2,gr4	CYP21A2	0.5061,0.1101,-0.65137	1.06e-213,5.09e-08,7.6e-290
chr12:112.1-112.9Mb	3	ADAM1A,TMEM116, RPL7AP60	CS	5.1116835479	gr1,gr2,gr4	TMEM116	-0.23844,1.90541,-0.17158	2.65e-163,0,3.29e-99
chr1:109.9-110.5Mb	3	AMIGO1,RP4-735C1.4, GSTM3	CT	-2.2532826605	gr1,gr2,gr3, gr4,gr5	AMIGO1	-0.07133,-0.06855,0.16263, -0.09108,0.40657	9.36e-11,6.49e-07,3.74e-35, 5.29e-15,1.21e-26
chr6:25.8-33.4Mb	69	MHC-locus genes	CT	4.3126732039	gr1,gr2,gr4	HLA-DQB1-AS1	-0.27096,-0.06134,0.37388	2.42e-112,2.15e-05,6.71e-159
chr12:110.7-112.7Mb	4	VPS29,FAM109A, ADAM1A,TMEM116	CT	3.751344091	gr1,gr2,gr4	TMEM116	-0.17116,2.00887,-0.15359	2.3e-139,0,2.81e-103
chr17:7.3-7.7Mb	1	SAT2	CT	1.0455349273	gr1	SAT2	-0.02237	0.00005
chr1:109.4-110.5Mb	7	WDR47,MYBPHL, SYPL2,GSTM2, GSTM1,RP4-735C1.4, GSTM3	HAA	4.9759940501	gr1,gr2,gr3, gr4,gr5	SYPL2	0.26639,0.24292,-0.53989, 0.27574,-1.11943	6.45e-84,4.18e-45,2.46e-290, 5.4e-86,3.03e-127
chr6:26.2-33.8Mb	50	MHC-locus genes	HAA	-4.1386611718	gr1,gr2,gr4	HLA-B	-0.86251,-0.07401,0.9944	0,1.88e-05,0
chr12:112-112.7Mb	4	ALDH2,ADAM1A, ADAM1B,TMEM116	HAA	-5.0265378079	gr1,gr2,gr4	ADAM1B	-0.05466,2.40601,-0.04308	9.02e-211,0,1.93e-132
chr1:109.8-110.5Mb	4	SYPL2,AMIGO1, RP4-735C1.4,GSTM3	HlV	-2.3961832209	gr1,gr2,gr3, gr4,gr5	SYPL2	0.14649,0.14693,-0.28754, 0.15085,-0.59976	2.49e-22,1.01e-14,1.12e-71, 8.84e-23,3.53e-44
chr5:16.2-16.6Mb	1	RP1-167G20.1	HlV	-0.4308179751	gr1	RP1-167G20.1	0.06094	0.0000727
chr6:25.8-33.7Mb	76	MHC-locus genes	HlV	-4.7479059705	gr1,gr2,gr4	HLA-DRB1	-0.18475,-0.04601,0.281	2.62e-107,1.61e-06,1.78e-158
chr12:111.2-112.7Mb	6	RP1-46F2.2,ALDH2, AC003029.1,RP3-462E2.5, ADAM1A,TMEM116	HlV	-5.8568762711	gr1,gr2,gr4	RP3-462E2.5	0.31558,-1.89141,0.22527	4.36e-189,0,3.67e-126
chr1:109.6-110.3Mb	6	CELSR2,PSRC1,SORT1, SYPL2,ATXN7L2,AMIGO1	Liver	-10.0093350929	gr1,gr2,gr3, gr4,gr5	CELSR2	-0.2424,-0.12874,1.74223, -0.2003,3.42219	0,2.58e-259,0,0,0
chr6:25.7-33.7Mb	50	MHC-locus genes	Liver	-3.9916513583	gr1,gr2,gr4	CYP21A2	0.28199,0.08718,-0.43291	4.89e-67,1.51e-05,2.41e-112
chr12:111.9-112.7Mb	2	PCNPP1,TMEM116	Liver	-6.6447969973	gr1,gr2,gr4	TMEM116	-0.07999,2.30002,-0.06325	8.15e-165,0,2.59e-103
chr1:109.4-110.5Mb	9	TMEM167B,KIAA1324, C1orf194,PSRC1,AMIGO1, GSTM4,GSTM2, GSTM5,GSTM3	WB	-7.0286640474	gr1,gr2,gr3, gr4,gr5	PSRC1	-0.60816,-0.51929,1.15011, -0.58608,2.07084	0,2.69e-167,0,4.81e-306,0
chr6:25.8-33.8Mb	73	MHC-locus genes	WB	4.3734130983	gr1,gr2,gr4	HLA-B	-0.78835,-0.08126,0.92798	0,1.88e-05,0
chr12:110.7-113.5Mb	6	GPN3,RP3-462E2.3, AC003029.1,ADAM1A, TMEM116,OAS1	WB	-4.6227362461	gr1,gr2,gr4	TMEM116	-0.20588,2.0451,-0.17274	3.01e-192,0,2.34e-121
chr15:28.4-28.8Mb	1	HERC2	WB	0.6926708351	gr1	HERC2	-0.00543	0.0000241
chr17:73.7-74.1Mb	1	TRIM47	WB	0.4454657067	gr3	TRIM47	-0.03431	0.0000196

Tab. B.5.: Cluster-specific genes for CAD clustering in Liver grouped into loci. The association is performed in each tissue separately (acronyms refer to the initial letter of the tissue). The order of each comma separated element in the last two columns having Wilcoxon-Mann-Whitney (WMW) estimates corresponds to the order "Group significant" column. The column "best CAD Z-stat" shows the Z-statistic for the most significant gene in that locus with respect to CAD. MHC-locus genes omitted for readability purposes.

Phenotype Class	N	Phenotype UKBB ids	Covariates
Alcohol	26	1558,1568,1578,1588,1598,1608,1618,1628,2664_5, 2664_1,2664_2,2664_3,2664_4,3731,3859_5,3859_1, 3859_2,3859_3,4407,4418,4429,4440,4451,20117_2, 20117_0,20117_1	PC1-10,Age,Sex, 6154_1,6154_2,6154_3,6155_3,6155_2,6155_5,6155_4, 6155_1,6155_6,6179_2,6179_3,6179_4,6179_5,6179_6, 6153_6177_1,6153_6177_2,6153_6177_3
Arterial_stiffness	8	4194,4195,4196,4198,4199,4200,4204,21021	PC1-10,Age,Sex, 6154_1,6154_2,6154_3,6155_3,6155_2,6155_5,6155_4, 6155_1,6155_6,6179_2,6179_3,6179_4,6179_5,6179_6, 6153_6177_1,6153_6177_2,6153_6177_3
Blood_biochemistry	30	30600,30610,30620,30630,30640,30650,30660,30670, 30680,30690,30700,30710,30720,30730,30740,30750, 30760,30770,30780,30790,30800,30810,30820,30830, 30840,30850,30860,30870,30880,30890	PC1-10,Age,Sex, 6154_1,6154_2,6154_3,6155_3,6155_2,6155_5,6155_4, 6155_1,6155_6,6179_2,6179_3,6179_4,6179_5,6179_6, 6153_6177_1,6153_6177_2,6153_6177_3
Blood_count	31	30000,30010,30020,30030,30040,30050,30060,30070, 30080,30090,30100,30110,30120,30130,30140,30150, 30160,30170,30180,30190,30200,30210,30220,30230, 30240,30250,30260,30270,30280,30290,30300	PC1-10,Age,Sex, 6154_1,6154_2,6154_3,6155_3,6155_2,6155_5,6155_4, 6155_1,6155_6,6179_2,6179_3,6179_4,6179_5,6179_6, 6153_6177_1,6153_6177_2,6153_6177_3
Blood_count_ratio	4	LMR,PLR,NLR,ELR	PC1-10,Age,Sex, 6154_1,6154_2,6154_3,6155_3,6155_2,6155_5,6155_4, 6155_1,6155_6,6179_2,6179_3,6179_4,6179_5,6179_6, 6153_6177_1,6153_6177_2,6153_6177_3
Blood_pressure	5	93,94,102,4079,4080	PC1-10,Age,Sex, 6154_1,6154_2,6154_3,6155_3,6155_2,6155_5,6155_4, 6155_1,6155_6,6179_2,6179_3,6179_4,6179_5,6179_6, 6153_6177_1,6153_6177_2,6153_6177_3
Body_size_measures	4	48,49,21001,21002	PC1-10,Age,Sex, 6154_1,6154_2,6154_3,6155_3,6155_2,6155_5,6155_4, 6155_1,6155_6,6179_2,6179_3,6179_4,6179_5,6179_6, 6153_6177_1,6153_6177_2,6153_6177_3
Diet	57	1488,1289,1299,1309,1319,1329,1339,1349,1359,1369, 1379,1389,1408,1418_2,1418_1,1418_3,1418_4,1418_5, 1418_6,1428_3,1428_0,1428_1,1428_2,1438,1448_3, 1448_1,1448_2,1448_4,1458,1468_3,1468_1,1468_2, 1468_4,1468_5,1478,1498,1508_2,1508_1,1508_3, 1508_4,1518,1528,1538_0,1538_1,1538_2,1548,2654_7, 2654_2,2654_4,2654_6,2654_8,2654_9,6144_5,6144_4, 6144_3,6144_1,6144_2	PC1-10,Age,Sex, 6154_1,6154_2,6154_3,6155_3,6155_2,6155_5,6155_4, 6155_1,6155_6,6179_2,6179_3,6179_4,6179_5,6179_6, 6153_6177_1,6153_6177_2,6153_6177_3
Early_life_factors	2	129130	PC1-10,Age,Sex
Family_history	54	1797,1807,1835,1873,1883,3526,4501,5057,20107_1, 20107_100,20107_2,20107_6,20107_8,20107_4,20107_3, 20107_11,20107_9,20107_10,20107_13,20107_12,20107_101, 20110_100,20110_1,20110_3,20110_8,20110_4,20110_5, 20110_11,20110_2,20110_10,20110_9,20110_6,20110_12, 20110_101,20111_100,20111_12,20111_13,20111_3,20111_8, 20111_9,20111_5,20111_6,20111_10,20111_1,20111_4, 20111_2,20111_11,20111_101,20112_100,20112_1,20112_101, 20113_100,20113_1,20113_101	PC1-10,Age,Sex
Hand_grip_strength	2	46,47	PC1-10,Age,Sex, 6154_1,6154_2,6154_3,6155_3,6155_2,6155_5,6155_4, 6155_1,6155_6,6179_2,6179_3,6179_4,6179_5,6179_6, 6153_6177_1,6153_6177_2,6153_6177_3
Height_derived	1	12144der	PC1-10,Age,Sex
ICD9-10_OPCS4	103	41270_D50,41270_D51,41270_D61,41270_D63,41270_D64, 41270_D68,41270_D69,41270_I05,41270_I07,41270_I08, 41270_I10,41270_I12,41270_I20,41270_I21,41270_I22, 41270_I23,41270_I24,41270_I25,41270_I26,41270_I27, 41270_I31,41270_I33,41270_I34,41270_I35,41270_I37, 41270_I38,41270_I42,41270_I44,41270_I45,41270_I46, 41270_I47,41270_I48,41270_I49,41270_I50,41270_I51, 41270_I61,41270_I62,41270_I63,41270_I64,41270_I65, 41270_I67,41270_I69,41270_I70,41270_I71,41270_I72, 41270_I73,41270_I74,41270_I77,41270_I78,41270_I80, 41270_I82,41270_I83,41270_I84,41270_I85,41270_I87, 41270_I89,41270_I95,41270_I97,41270_J02,41270_J06, 41270_J15,41270_J18,41270_J22,41270_J30,41270_J31, 41270_J32,41270_J33,41270_J34,41270_J38,41270_J39, 41270_J40,41270_J43,41270_J44,41270_J45,41270_J47, 41270_J61,41270_J69,41270_J81,41270_J84,41270_J90, 41270_J92,41270_J93,41270_J95,41270_J96,41270_J98, 41270_E03,41270_E04,41270_E05,41270_E10,41270_E11, 41270_E14,41270_E16,41270_E21,41270_E53,41270_E55, 41270_E66,41270_E78,41270_E80,41270_E83,41270_E86, 41270_E87,41270_E88,41270_E89	PC1-10,Age,Sex
Impedance_measures	5	23099,23100,23101,23102,23105	PC1-10,Age,Sex, 6154_1,6154_2,6154_3,6155_3,6155_2,6155_5,6155_4, 6155_1,6155_6,6179_2,6179_3,6179_4,6179_5,6179_6, 6153_6177_1,6153_6177_2,6153_6177_3
Medication	32	2492,6153_2,6153_100,6153_1,6153_4,6153_3,6154_1,6154_2, 6154_100,6154_3,6154_5,6154_6,6154_4,6155_3,6155_100, 6155_7,6155_2,6155_5,6155_4,6155_1,6155_6,6177_1,6177_100, 6177_2,6177_3,6179_2,6179_100,6179_1,6179_3,6179_4,6179_5, 6179_6	PC1-10,Age,Sex
Medications	196	137, all 20003 subclass	PC1-10,Age,Sex
Physical_activity	31	864,874,884,894,904,914,924,943,971,981,991,1011,1021,1070, 1080,1090,2624,2634,3637,3647,6162_2,6162_3,6162_1,6162_4, 6162_100,6164_4,6164_1,6164_2,6164_100,6164_5,6164_3	PC1-10,Age,Sex, 6154_1,6154_2,6154_3,6155_3,6155_2,6155_5,6155_4, 6155_1,6155_6,6179_2,6179_3,6179_4,6179_5,6179_6, 6153_6177_1,6153_6177_2,6153_6177_3
Sleep	7	1160,1170,1180,1190,1200,1210,1220	PC1-10,Age,Sex, 6154_1,6154_2,6154_3,6155_3,6155_2,6155_5,6155_4, 6155_1,6155_6,6179_2,6179_3,6179_4,6179_5,6179_6, 6153_6177_1,6153_6177_2,6153_6177_3
Smoking	39	1239,1249,1259,1269,1279,2644,2867,2877_1,2877_2,2877_3, 2887,2907,2926,3436,3446_1,3446_2,3446_3,3456,3466,3476, 3486,3496,5959,6157_3,6157_100,6157_1,6157_4,6157_2,6158_3, 6158_4,6158_2,6158_100,6158_1,20116_0,20116_1,20116_2,20160,20161,20162	PC1-10,Age,Sex, 6154_1,6154_2,6154_3,6155_3,6155_2,6155_5,6155_4, 6155_1,6155_6,6179_2,6179_3,6179_4,6179_5,6179_6, 6153_6177_1,6153_6177_2,6153_6177_3

Tab. B.6.: UK Biobank phenotypes included in endophenotype analysis testing cluster-specific differences in CAD. "Covariates" column refers to the confounders used in GLM for phenotypes in that class, ids are from "Medication" class.

Phenotype Name	Phenotype UKBB ids	Meaning
BMI	21001	Body mass index
UAP	41270_I200	Unstable angina
Acute_MI	41270_21	Acute myocardial infarction
History_MI	41270_I252	Old myocardial infarction
Coronary_artery_bypass_graft	41272_K40, 41272_K41, 41272_K42	Saphenous vein graft replacement of coronary artery, Other autograft replacement of coronary artery, Allograft replacement of coronary artery
Percutaneous_coronary_intervention	41272_K49, 41272_K75	Transluminal balloon angioplasty of coronary artery, Percutaneous transluminal balloon angioplasty and insertion of stent into coronary artery
History_bleeding	41270_K92, 41270_R04	Other diseases of digestive system, Haemorrhage from respiratory passages
Heart_function_severity	41270_I501	Left ventricular failure
Hypertension	41270_I10, 41270_I11, 41270_I12, 41270_I13, 41270_I15	Essential (primary) hypertension, Hypertensive heart disease, Hypertensive renal disease, Hypertensive heart and renal disease, Secondary hypertension
Hyperlipidemia	41270_E78	Disorders of lipoprotein metabolism and other lipidaemias
Diabetes	41270_E10, 41270_E11, 41270_E12, 41270_E13, 41270_E14	Insulin-dependent diabetes mellitus, Non-insulin-dependent diabetes mellitus, Malnutrition-related diabetes mellitus, Other specified diabetes mellitus, Unspecified diabetes mellitus
T1D	41270_E10	Insulin-dependent diabetes mellitus
T2D	41270_E11	Non-insulin-dependent diabetes mellitus
Medication_Insulin	6153_3, 6177_3	Insulin (female), Insulin (male)
Peripheral_vascular_disease	41270_I702	Atherosclerosis of arteries of the extremities
Cerebrovascular_disease	41270_I672	Cerebral atherosclerosis
Cerabral_stroke	41270_I63, 41270_I64	Cerebral infarction, Stroke not specified as haemorrhage or infarction
Transient_cerebral_ischaemic_attacks	41270_G45	Transient cerebral ischaemic attacks and related syndromes
Chronic_obstructive_pulmonary_disease	41270_J44	Other chronic obstructive pulmonary disease
Chronic_kidney_disease	41270_N18	Chronic renal failure
Dialysis	41270_Z49	Care involving dialysis
Atherosclerotic_heart_disease	41270_I251	Atherosclerotic heart disease
Poor_mobility	41270_Z74	Problems related to care-provider dependency
Pulmonary_hypertension	41270_I27	Other pulmonary heart diseases
LVEF	22420	LV ejection fraction
History_cancer	2453_1, 40006	Cancer diagnosed by doctor, Type of cancer: ICD10
Smoking	41270_F17, 20160_1	Mental and behavioural disorders due to use of tobacco, Ever smoked
Age_angina	3627	Age angina diagnosed
Age_heart_attack	3894	Age heart attack diagnosed
Age_stroke	4056	Age stroke diagnosed
Death_Acute_MI	40001	Acute myocardial infarction
Death_Chronic_ischemic_heart_disease	40001	Chronic ischaemic heart disease
Age_death	40007	Age at death

Tab. B.7.: Original UK Biobank phenotypes included in hypothesis-driven endophenotype analysis testing cluster-specific differences in CAD, corrected for PCs1-10, Age and Sex

B.3 Schizophrenia

Phenotype Class	N.	Phenotype UKBB ids	Covariates
Blood_biochemistry	30	30600,30610,30620,30630,30640,30650, 30660,30670,30680,30690,30700,30710, 30720,30730,30740,30750,30760,30770, 30780,30790,30800,30810,30820,30830, 30840,30850,30860,30870,30880,30890	PC 1-10,Sex,Age
Blood_count	31	30000,30010,30020,30030,30040,30050, 30060,30070,30080,30090,30100,30110, 30120,30130,30140,30150,30160,30170, 30180,30190,30200,30210,30220,30230, 30240,30250,30260,30270,30280,30290, 30300	PC 1-10,Sex,Age
Blood_count_ratio	4	LMR,PLR,NLR,ELR	PC 1-10,Sex,Age
Cannabis_use	2	20453,20454	PC 1-10,Sex,Age
Fluid_intelligence	15	4935,4946,4957,4968,4979,4990,5001, 5012,5556,5699,5779,5790,5866,20016, 20128	PC 1-10,Sex,Age
Mental_health	42	1920,1930,1940,1950,1960,1970,1980, 1990,2000,2010,2020,2030,2040,2050, 2060,2070,2080,2090,2100,4526,4537, 4548,4559,4570,4581,4598,4609,4620, 4631,4642,4653,5375,5386,5663,5674, 6156_13,6156_12,6156_100,6156_11, 6156_15,6156_14,20127	PC 1-10,Sex,Age
Pairs_matching	2	398400	PC 1-10,Sex,Age
Prospective_memory	2	4288,20018	PC 1-10,Sex,Age
Reaction_time	3	403,404,20023	PC 1-10,Sex,Age
Smoking	5	1239,1249,2644,3456,20160	PC 1-10,Sex,Age
Trail_making	8	20147,20148,20149,20155,20156, 20157,20247,20248	PC 1-10,Sex,Age

Tab. B.8.: UK Biobank phenotypes included in correlation and Mendelian Randomization analysis with respect to SCZ. "Covariates" column refers to the confounders used in PALAS and TWAS for phenotypes in that class.

Loci	ngenes	gene	tissue	best SCZ Z-stat	Group significant	Best WMW gene	Best WMW estimate	Best WMW pvalue
chr2:47.4-47.8Mb	1	MSH2	DLPC_CMC	-1.8395323868	gr2	MSH2	0.00784	0.0000823
chr2:71.2-71.6Mb	1	MPHOSPH10	DLPC_CMC	0.3520473266	gr2,gr3	MPHOSPH10	-0.01966,0.01665	2.85e-07,1.47e-05
chr2:135.6-136.5Mb	2	RAB3GAP1,R3HDM1	DLPC_CMC	-0.9431472602	gr1	R3HDM1	0.01065	0.00000205
chr2:224.6-225Mb	1	WDFY1	DLPC_CMC	-1.0951435952	gr3	WDFY1	-0.01318	6.19E-08
chr3:110.6-111Mb	1	PVRL3	DLPC_CMC	-1.3879564135	gr1	PVRL3	0.00597	2.43E-10
chr6:24.4-34Mb	84	*MHC-locus genes*	DLPC_CMC	9.1179242704	gr1,gr2,gr3	C4A	-2.31284,0.17878,0.14103	0,7.58e-269,8.64e-177
chr9:88.2-88.6Mb	1	RP11-213G2.3	DLPC_CMC	-0.0171994994	gr3	RP11-213G2.3	0.01069	0.000000329
chr11:61-61.4Mb	1	TMEM216	DLPC_CMC	0.215306172	gr1	TMEM216	-0.00411	0.0000289
chr15:65.5-65.9Mb	1	IGDCC4	DLPC_CMC	-1.2642548028	gr2	IGDCC4	0.0499	0.0000984
chr17:18.9-19.3Mb	1	SNORD3A	DLPC_CMC	1.0421285199	gr2	SNORD3A	0.00429	0.0000595
chr2:75-75.4Mb	1	AC104135.4	BCbg	-0.0490251206	gr1	AC104135.4	0.06252	0.000142
chr2:224.6-225Mb	1	AC073641.2	BCbg	1.0488189565	gr3	AC073641.2	0.0565	0.0000387
chr6:25-34Mb	50	MHC-locus genes	BCbg	-8.7650278157	gr1,gr2,gr3	IER3	2.10795,-0.08578,-0.08574	0,5.48e-216,1.74e-209
chr1:27.5-27.9Mb	1	CD164L2	BCeH	0.6093725704	gr1	CD164L2	0.00634	0.00000128
chr2:74.9-75.3Mb	1	AC104135.3	BCeH	-0.3962671955	gr1	AC104135.3	0.06489	0.0000991
chr6:25.8-34Mb	65	MHC-locus genes	BCeH	9.1569954703	gr1,gr2,gr3	C4A	-2.20347,0.24164,0.27126	0,3e-219,8.82e-292
chr6:83.7-84.1Mb	1	RWDD2A	BCeH	-2.6887430634	gr2	RWDD2A	-0.00957	0.0000468
chr11:73.3-73.7Mb	1	MRPL48	BCeH	2.6712210593	gr2	MRPL48	0.00619	0.000106
chr12:132.9-133.3Mb	1	FBRSL1	BCeH	-1.0011498674	gr1	FBRSL1	0.01188	0.000132
chr13:24.3-24.7Mb	1	MIPEP	BCeH	1.2891484495	gr1	MIPEP	0.05857	0.0000967
chr22:22.1-22.5Mb	1	PPM1F	BCeH	-1.1368405368	gr3	PPM1F	0.02444	0.000164
chr3:52.2-52.6Mb	1	DNAH1	BCe	-0.2024603626	gr1	DNAH1	0.01081	0.000164
chr6:25.8-34Mb	84	MHC-locus genes	BCe	8.5677584566	gr1,gr2,gr3	C4A	-2.0414,0.35472,0.33328	0,4.58e-267,2.43e-248
chr7:91.7-92.1Mb	1	KRIT1	BCe	-0.9655459604	gr1,gr3	KRIT1	0.00678,-0.00481	7.45e-07,1.22e-05
chr10:97.2-97.6Mb	1	ALDH18A1	BCe	-1.6180039308	gr1	ALDH18A1	0.07352	0.000026
chr13:79.8-80.2Mb	1	RBM26-AS1	BCe	-0.3258249167	gr2	RBM26-AS1	0.00716	0.000118
chr14:50.7-51.1Mb	1	CDKL1	BCe	-0.2470513583	gr2	CDKL1	0.03716	0.000197
chr16:77-77.4Mb	1	SYCE1L	BCe	0.3307331916	gr2	SYCE1L	0.05207	0.000219
chr21:47.5-47.9Mb	1	MCM3AP	BCe	-1.1100190408	gr2,gr3	MCM3AP	-0.05892,0.05559	1.81e-05,5.48e-05
chr6:24.9-33.7Mb	44	MHC-locus genes	BC	-8.5096903884	gr1,gr2,gr3	NOTCH4	-1.91242,0.31281,0.28497	0,8.54e-243,4.19e-221
chr12:104.1-104.5Mb	1	MIR3652	BC	-1.3608713844	gr1	MIR3652	-0.00802	0.0000605
chr2:74.9-75.3Mb	1	AC104135.3	BFCB	-0.283647674	gr1	AC104135.3	0.06382	0.000185
chr2:224.6-225Mb	1	WDFY1	BFCB	-0.5206130105	gr3	WDFY1	-0.0439	0.0000589
chr6:25.8-33.6Mb	44	MHC-locus genes	BFCB	-9.0913025265	gr1,gr2,gr3	HLA-DMA	1.7991,-0.17079,-0.14477	0,1.66e-159,1.1e-115
chr16:77-77.4Mb	1	SYCE1L	BFCB	-0.1599495792	gr2	SYCE1L	0.01571	0.000203
chr6:26-28.5Mb	8	HIST1H2BD,BTN3A2,ZNF391, RP1-265C24.5,ZNF192P1, AL022393.7,ZKSCAN3,ZSCAN31	BHi	-7.5480068674	gr1,gr2,gr3	AL022393.7	0.76927,1.15978,-1.59026	0,0,0
chr6:29.6-33.6Mb	26	MHC-locus genes	BHi	9.4795780052	gr1,gr2,gr3	CYP21A1P	-2.1219,0.2312,0.26086	0,2.6e-215,8.8e-280
chr7:5.4-5.8Mb	1	FSCN1	BHi	1.5396859337	gr3	FSCN1	-0.05445	0.0000872
chr16:67.3-67.7Mb	1	HSD11B2	BHi	-2.5989449838	gr1	HSD11B2	-0.00499	0.000204
chr6:25.9-34Mb	43	MHC-locus genes	BHy	9.4252980261	gr1,gr2,gr3	NCR3	1.93175,-0.12358,-0.09018	0,2.71e-253,3.42e-155
chr12:56.5-56.9Mb	1	RP11-977G19.11	BHy	1.3713440176	gr1	RP11-977G19.11	0.00557	0.0000843
chr16:30.2-30.6Mb	1	ZNF48	BHy	-2.1048370539	gr3	ZNF48	0.01108	0.0000779
chr19:44.8-45.2Mb	1	ZNF180	BHy	0.7607912028	gr2	ZNF180	-0.02512	0.000113
chr1:173.4-173.8Mb	1	ANKRD45	Bnabg	4.1194382573	gr2	ANKRD45	0.00336	0.0000829
chr6:25.8-28.5Mb	11	U91328.19,U91328.22,BTN3A2, BTN2A2,RP11-457M11.5, ZNF204P,ZNF391,RP1-265C24.5, AL022393.7, RP5-874C20.3,ZSCAN31	BNabg	-7.4503374099	gr1,gr2,gr3	RP1-265C24.5	-1.54759,-0.12148,1.60438	0,0,0
chr6:29.6-33.6Mb	31	MHC-locus genes	BNabg	-8.5169850307	gr1,gr2,gr3	XXbac-BPG300A18.13	1.85146,-0.37392,-0.31261	0,1.39e-252,3.66e-189
chr19:2-2.4Mb	1	SF3A2	BNabg	-0.8413159181	gr1	SF3A2	-0.007	2.09E-14
chr21:47.4-47.8Mb	1	FTCD	BNabg	-1.8911644822	gr2	FTCD	-0.03959	0.0000526
chr2:74.9-75.4Mb	3	AC104135.3,AC104135.2,AC104135.4	CEl	0.4123488469	gr1	AC104135.2	0.06891	0.0000377
chr2:224.6-225Mb	1	AC073641.2	CEl	1.1941178054	gr3	AC073641.2	0.05246	0.000183
chr3:159.9-160.3Mb	1	SMC4	CEl	0.8040175276	gr1	SMC4	-0.0169	0.0000973
chr6:24.8-26.9Mb	8	FAM65B,CMAHP,LRRC16A, HIST1H2BH,HIST1H2APS4, BTN3A2,RP11-457M11.2, RP11-457M11.5	CEl	7.4925742529	gr1,gr2,gr3	BTN3A2	-0.67175,0.09297,0.09251	0,3.02e-71,1.12e-66
chr6:27.9-34Mb	51	MHC-locus genes	CEl	9.1408606008	gr1,gr2,gr3	C4A	-2.31136,0.18873,0.20019	0,3.3e-236,1.9e-287
chr13:20-20.4Mb	1	MPHOSPH8	CEl	0.8682809548	gr1	MPHOSPH8	0.0077	0.0000507
chr22:21.8-22.2Mb	1	YDJC	CEl	2.2103045203	gr2	YDJC	0.00947	0.0000352

Tab. B.9.: Cluster-specific genes for SCZ clustering in DLPC grouped into loci (genes $|\text{corr.}| < 0.9$). The association is performed in each tissue separately (acronyms refer to the initial letter of the tissue). The order of each comma separated element in the last two columns having Wilcoxon-Mann-Whitney (WMW) estimates correspond to the order "Group significant" column. The column "best SCZ Z-stat" shows the Z-statistic for the most significant gene in that locus with respect to SCZ. MHC-locus genes omitted for readability purposes.

Loci	N. genes	Gene	Tissue	best SCZ Z-stat	Group significant	Best WMW gene	Best WMW estimate	Best WMW pvalue
chr1:11.3-11.7Mb	1	PTCHD2	DLPC_CMC	3.696229704	gr2	PTCHD2	0.24934	8.49E-08
chr1:95.4-95.8Mb	1	TMEM56	DLPC_CMC	-4.6852591952	gr2	TMEM56	-0.30814	8.87E-12
chr1:112.8-113.2Mb	1	WNT2B	DLPC_CMC	-3.7768452772	gr2	WNT2B	-0.06932	0.0000203
chr1:207.9-208.3Mb	1	CD34	DLPC_CMC	-3.3730082483	gr2	CD34	-0.41774	4.28E-18
chr1:243.2-243.6Mb	1	SDCCAG8	DLPC_CMC	5.0025084846	gr2	SDCCAG8	-0.21226	9.83E-09
chr2:58.3-58.7Mb	1	FANCL	DLPC_CMC	-4.1672663127	gr2	FANCL	-0.36642	6.93E-15
chr2:111.7-112.1Mb	1	BCL2L11	DLPC_CMC	-2.0676770754	gr2	BCL2L11	-0.06843	0.00000356
chr2:200.3-201Mb	4	SEPHS1P6,FTCDNL1, C2orf47,TYW5	DLPC_CMC	-7.5980154953	gr1,gr3,gr4	C2orf47	-1.81462,0.08128,1.56972	0,8.75e-183,0
chr2:202.1-202.8Mb	2	TRAK2,ALS2	DLPC_CMC	2.5217162449	gr1,gr4	TRAK2	0.06789,-0.05298	1.99e-10,9.85e-08
chr3:63.6-64Mb	1	THOC7	DLPC_CMC	-5.8878319289	gr2	THOC7	0.1068	6.64E-09
chr3:138.9-139.5Mb	2	COPB2,RP11-553K23.2	DLPC_CMC	-3.1465576347	gr2	RP11-553K23.2	0.25653	8.65E-09
chr3:180.1-180.5Mb	1	TTC14	DLPC_CMC	-4.9156674725	gr2,gr3	TTC14	-0.01395,0.00687	2.32e-05,7.71e-08
chr4:170.3-170.7Mb	1	CLCN3	DLPC_CMC	5.9135188058	gr2	CLCN3	0.10226	0.0000701
chr5:59.9-60.3Mb	1	ELOVL7	DLPC_CMC	3.8869383531	gr2	ELOVL7	-0.11068	0.000000642
chr5:108.8-109.2Mb	1	MAN2A1	DLPC_CMC	3.1205705562	gr2	MAN2A1	-0.23981	0.000000167
chr6:24.4-33.1Mb	66	MHC locus genes	DLPC_CMC	9.1179242704	gr1,gr2,gr3,gr4	BTN3A3	-0.1716,-0.24291,1.42272,-0.81276	6.52e-43,1.76e-11,0,0
chr6:83.9-84.6Mb	2	ME1,SNAP91	DLPC_CMC	4.5414139273	gr2	SNAP91	-0.30833	1.9E-12
chr6:108.9-109.3Mb	1	ZNF259P1	DLPC_CMC	-4.4349360358	gr2	ZNF259P1	0.1874	0.0000161
chr6:136-136.4Mb	1	PDE7B	DLPC_CMC	-2.4541728646	gr2	PDE7B	-0.22192	0.00000146
chr7:2.1-2.5Mb	1	FTSJ2	DLPC_CMC	-3.8810523441	gr2	FTSJ2	0.29685	7.21E-11
chr7:71.7-72.1Mb	1	CALN1	DLPC_CMC	4.5498010607	gr2	CALN1	0.29122	1.64E-10
chr7:87.7-88.1Mb	1	SRI	DLPC_CMC	2.6881424928	gr2	SRI	0.22316	0.0000012
chr8:9.7-10.1Mb	1	MSRA	DLPC_CMC	-4.5735477558	gr2	MSRA	0.12113	0.0000395
chr8:37.8-38.4Mb	3	BAG4,DDHD2,WHSC1L1	DLPC_CMC	-4.2337775286	gr2	BAG4	-0.19577	0.00000158
chr8:89.1-89.5Mb	1	RP11-586K2.1	DLPC_CMC	-4.4526002155	gr2	RP11-586K2.1	-0.06603	0.000000308
chr9:36.9-37.3Mb	1	RP11-220I1.1	DLPC_CMC	-4.0671843619	gr2	RP11-220I1.1	0.06425	0.00000588
chr9:130-130.4Mb	1	RPL12	DLPC_CMC	3.2888421406	gr2	RPL12	0.13523	0.00000173
chr10:85.8-86.2Mb	1	CDHR1	DLPC_CMC	-2.8629762234	gr2	CDHR1	-0.12828	0.0000591
chr10:104.3-104.9Mb	3	WBP1L,AS3MT,CNNM2	DLPC_CMC	-7.1938119762	gr2	CNNM2	0.2291	0.000000002
chr11:57.2-57.6Mb	1	AP000662.4	DLPC_CMC	-3.9420511472	gr2	AP000662.4	-0.16857	0.0000177
chr11:125.3-125.7Mb	1	CHEK1	DLPC_CMC	2.9584011536	gr2	CHEK1	0.19316	0.00000374
chr12:123.3-124Mb	3	OGFOD2,MPHOSPH9,CDK2AP1	DLPC_CMC	-5.2978323315	gr2,gr4	MPHOSPH9	0.27353,-0.0463	3.22e-12,9.94e-05
chr13:28-28.4Mb	1	POLR1D	DLPC_CMC	-3.8770129504	gr2	POLR1D	-0.21005	1.03E-08
chr14:72.2-72.7Mb	2	RGS6,AC005477.1	DLPC_CMC	-4.6011405521	gr2	RGS6	0.44127	8.51E-23
chr14:93.6-94Mb	1	BTBD7	DLPC_CMC	-2.8532819362	gr2	BTBD7	0.17168	0.00000343
chr15:84.5-85.1Mb	4	EFTUD1P1,CSPG4P11, GOLGA2P7,GOLGA6L4	DLPC_CMC	-5.7182361079	gr2	GOLGA2P7	0.21534	0.00000164
chr16:4.4-4.8Mb	1	CDIP1	DLPC_CMC	3.6716333413	gr2	CDIP1	0.1331	0.00000429
chr16:9.7-10.1Mb	1	RP11-297M9.2	DLPC_CMC	-3.2804132872	gr2	RP11-297M9.2	0.22564	0.00000211
chr16:15-15.4Mb	1	RRN3	DLPC_CMC	-3.7510985393	gr2	RRN3	-0.1396	0.00000745
chr16:29.8-30.3Mb	3	INO80E,DOC2A,MAPK3	DLPC_CMC	5.8131552963	gr2,gr4	DOC2A	0.28012,-0.06377	1.92e-10,2.13e-06
chr17:46.8-47.2Mb	1	ATP5G1	DLPC_CMC	3.0599766961	gr2	ATP5G1	-0.12912	0.00000159
chr18:60.8-61.2Mb	1	BCL2	DLPC_CMC	-2.3663669599	gr2	BCL2	-0.13438	0.0000809
chr19:19.2-19.7Mb	2	MAU2,GATAD2A	DLPC_CMC	-5.2422296188	gr2	MAU2	0.09658	5.34E-09
chr22:41-41.9Mb	4	SLC25A17,XPNPEP3, EP300,RANGAP1	DLPC_CMC	-5.242506119	gr2	SLC25A17	0.27344	7.15E-10

Tab. B.10.: Cluster-specific genes for SCZ clustering in DLPC grouped into loci (genes $|\text{corr.}| < 0.1$). The results shows the association tested in DLPC tissue. The order of each comma separated element in the last two columns with Wilcoxon-Mann-Whitney (WMW) estimates correspond to the order "Group significant" column. The column "best SCZ Z-stat" shows the Z-statistic for the most significant gene in that locus with respect to SCZ. MHC-locus genes omitted for readability purposes.

Phenotype Class	N	Phenotype UKBB ids	Covariates
Alcohol_use	14	20403,20405_0,20405_1,20405_2,20407,20408, 20409,20411_0,20411_1,20411_2,20412,20413, 20414,20416	C1,C2,C3,C4,C5,C6,C7,C9,C15,C18
Anxiety	31	20417,20418,20419,20420,20421,20422,20423, 20425,20426,20427,20428,20429,20505,20506, 20509,20512,20515,20516,20520,20537,20538, 20539,20540,20541,20542,20543,20549_3, 20549_4,20549_1,20550_1,20550_3	C1,C2,C3,C4,C5,C6,C7,C9,C15,C18
Blood_biochemistry	30	30600,30610,30620,30630,30640,30650,30660, 30670,30680,30690,30700,30710,30720,30730, 30740,30750,30760,30770,30780,30790,30800, 30810,30820,30830,30840,30850,30860,30870, 30880,30890	C1,C2,C3,C4,C5,C6,C7,C9,C15,C18
Blood_count	31	30000,30010,30020,30030,30040,30050,30060, 30070,30080,30090,30100,30110,30120,30130, 30140,30150,30160,30170,30180,30190,30200, 30210,30220,30230,30240,30250,30260,30270, 30280,30290,30300	C1,C2,C3,C4,C5,C6,C7,C9,C15,C18
Blood_count_ratio	4	LMR,PLR,NLR,ELR	C1,C2,C3,C4,C5,C6,C7,C9,C15,C18
Blood_pressure	5	93,94,102,4079,4080	C1,C2,C3,C4,C5,C6,C7,C9,C15,C18
Body_size_measures	4	48,49,21001,21002	C1,C2,C3,C4,C5,C6,C7,C9,C15,C18
Cannabis_use	3	20453,20454,20455	C1,C2,C3,C4,C5,C6,C7,C9,C15,C18
Depression	37	20433,20435,20436,20437,20438,20439,20440, 20441,20442,20445,20446,20447,20448,20449, 20450,20507,20508,20510,20511,20513,20514, 20517,20518,20519,20532,20533,20534,20535, 20536_0,20536_1,20536_2,20536_3,20546_3, 20546_1,20546_4,20547_1,20547_3	C1,C2,C3,C4,C5,C6,C7,C9,C15,C18
dMRI_skeleton	432	from 25063 to 25487	C1,C2,C3,C4,C5,C6,C7,C9,C15,C18
Fluid_intelligence	33	4924,4935,4946,4957,4968,4979,4990,5001,5012, 5556,5699,5779,5790,5866,20016,20128,20165, 20167,20169,20171,20173,20175,20177,20179, 20181,20183,20185,20187,20189,20191,20242_0, 20242_1,20242_2	C1,C2,C3,C4,C5,C6,C7,C9,C15,C18
Happiness_and_subjective_well-being	3	20458,20459,20460	C1,C2,C3,C4,C5,C6,C7,C9,C15,C18
Mental_distress	16	20499,20500,20544_6,20544_11,20544_15, 20544_12,20544_1,20544_5,20544_7,20544_4, 20544_3,20544_16,20544_10,20544_13,20544_17, 20544_14	C1,C2,C3,C4,C5,C6,C7,C9,C15,C18
Mental_health	42	1920,1930,1940,1950,1960,1970,1980,1990,2000, 2010,2020,2030,2040,2050,2060,2070,2080,2090, 2100,4526,4537,4548,4559,4570,4581,4598,4609, 4620,4631,4642,4653,5375,5386,5663,5674, 6156_13,6156_12,6156_100,6156_11,6156_15, 6156_14,20127	C1,C2,C3,C4,C5,C6,C7,C9,C15,C18
Numeric_memory	1	20240	C1,C2,C3,C4,C5,C6,C7,C9,C15,C18
Pairs_matching	9	398,399,400,20131,20132,20133,20244_0, 20244_1,20244_2	C1,C2,C3,C4,C5,C6,C7,C9,C15,C18
Prospective_memory	7	4288,4290,4291,4294_1,4294_0,4294_9,20018	C1,C2,C3,C4,C5,C6,C7,C9,C15,C18
Reaction_time	3	403,404,20023	C1,C2,C3,C4,C5,C6,C7,C9,C15,C18
Sleep	7	1160,1170,1180,1190,1200,1210,1220	C1,C2,C3,C4,C5,C6,C7,C9,C15,C18
Smoking	38	1239,1249,1259,1269,1279,2644,2867,2877_1, 2877_2,2877_3,2887,2907,2926,2936,3436, 3446_1,3446_2,3446_3,3456,3466,3476,3486, 3496,5959,6157_3,6157_100,6157_1,6157_4, 6157_2,6158_3,6158_4,6158_2,6158_100, 6158_1,20116_0,20116_1,20116_2,20160	C1,C2,C3,C4,C5,C6,C7,C9,C15,C18
Social_support	8	1031,2110,6160_3,6160_5,6160_100,6160_1, 6160_2,6160_4	C1,C2,C3,C4,C5,C6,C7,C9,C15,C18
Susceptibility_weighted_brain_MRI	15	25026,25027,25028,25029,25030,25031, 25032,25033,25034,25035,25036,25037, 25038,25039,25738	C1,C2,C3,C4,C5,C6,C7,C9,C15,C18
Symbol_digit_substitution	5	20159,20230,20245_0,20245_1,20245_2	C1,C2,C3,C4,C5,C6,C7,C9,C15,C18
T1_structural_brain_MRI	169	from 25001 to 25025, from 25731 to 25735, from 25782 to 25920	C1,C2,C3,C4,C5,C6,C7,C9,C15,C18
Task_functional_brain_MRI	20	12651,25040,25042,25044,25046,25048, 25050,25052,25054,25740,25742,25745, 25761,25762,25763,25764,25765,25766, 25767,25768	C1,C2,C3,C4,C5,C6,C7,C9,C15,C18
Trail_making	12	20147,20148,20149,20155,20156,20157, 20246_0,20246_1,20246_2,20246_3,20247, 20248	C1,C2,C3,C4,C5,C6,C7,C9,C15,C18
Traumatic_events	21	20487,20488,20489,20490,20491,20494, 20495,20496,20497,20498,20521,20522, 20523,20524,20525,20526,20527,20528, 20529,20530,20531	C1,C2,C3,C4,C5,C6,C7,C9,C15,C18

Tab. B.11.: UK Biobank phenotypes included in endophenotype analysis testing cluster-specific differences in SCZ in term of gene-RS mimicking endophenotype values. "Covariates" column refers to the confounders used in GLM for gene-RS in that class.

Appendix Figures

C

C.1 PriLer comparison to state-of-the-art methods

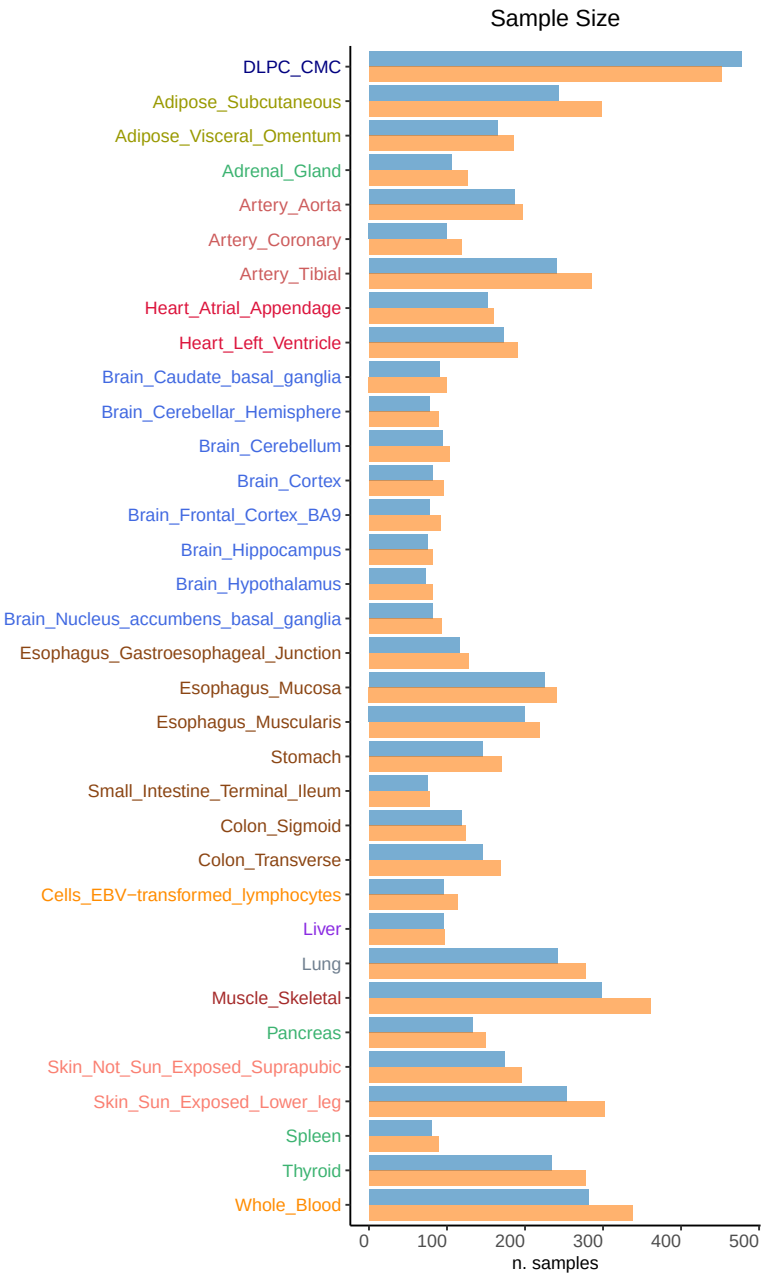

Fig. C.1.: Number of training samples used to create gene expression prediction models in PriLer (blue) and TWAS (orange) for the downloaded summary statistics

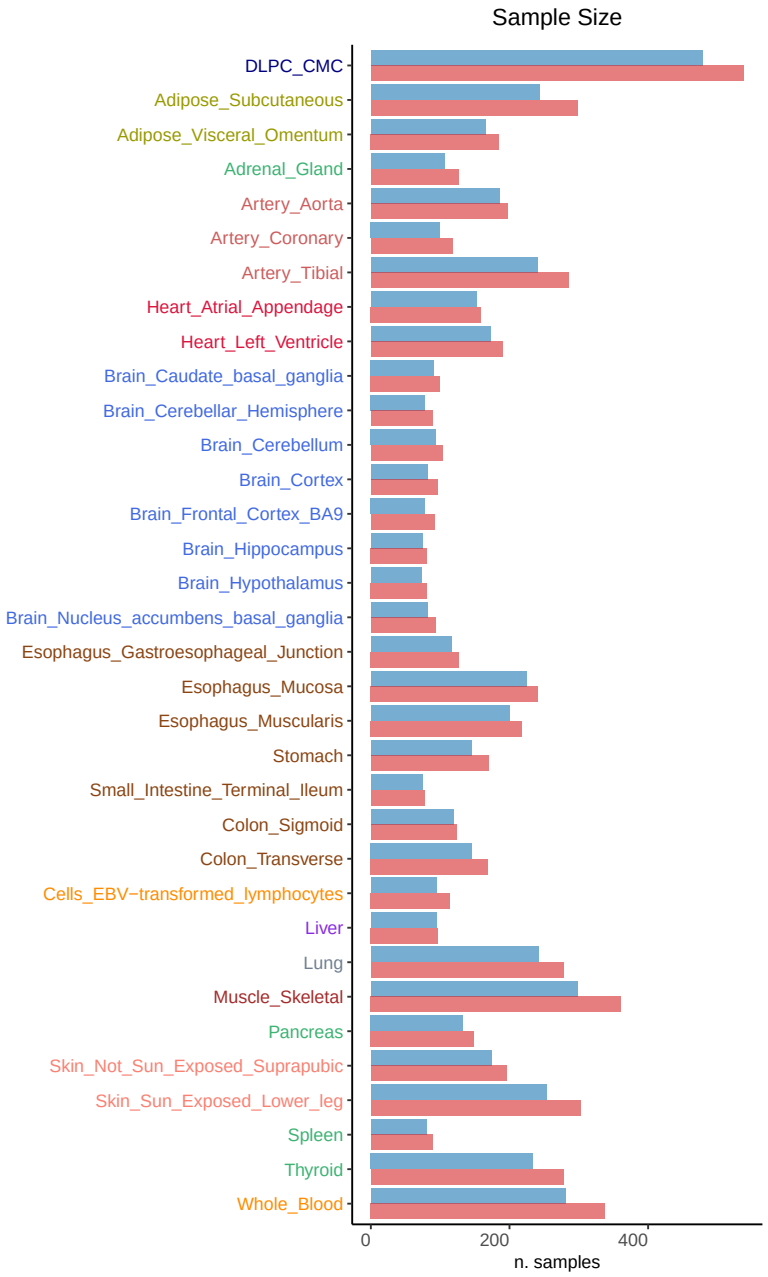

Fig. C.2.: Number of training samples considered to create gene expression prediction models in PriLer (blue) and prediXcan (red) for the downloaded summary statistics

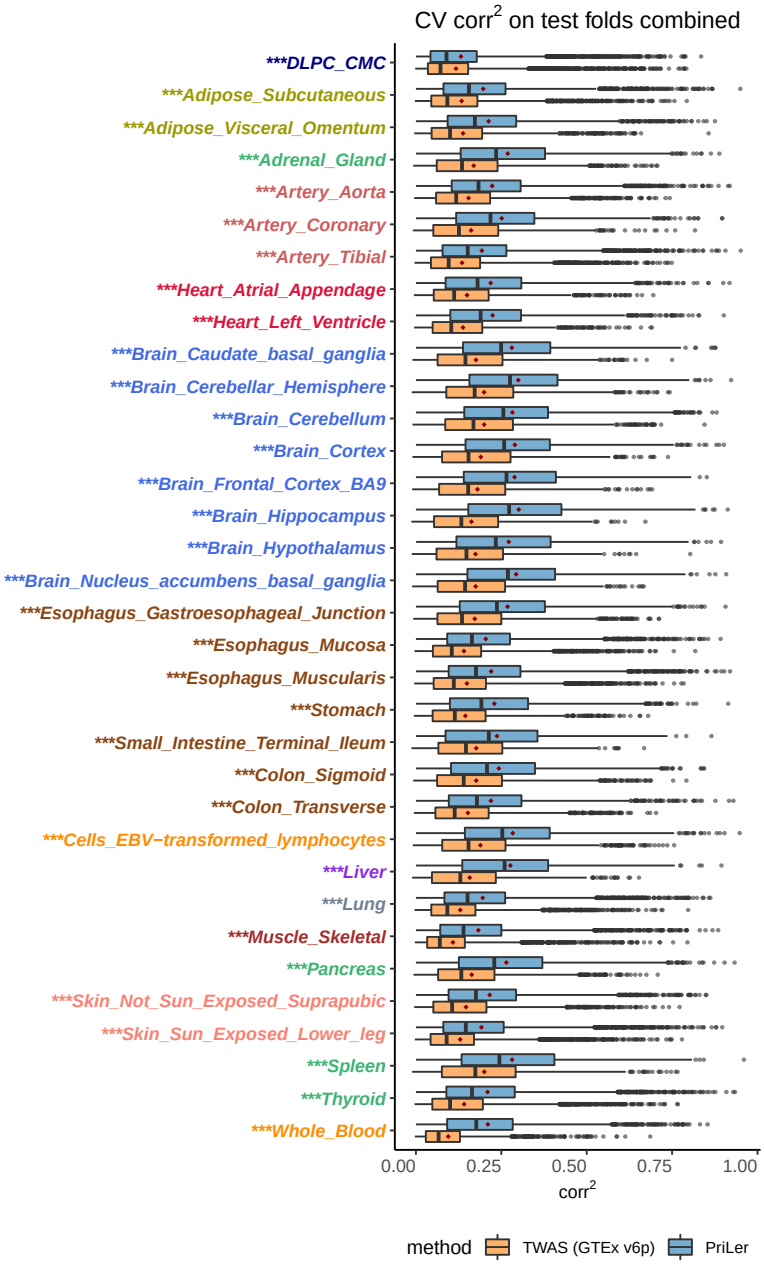

Fig. C.3.: Distribution of CV squared correlation between true and predicted gene expression computed on combined test folders. Each element in a boxplot is a gene in a certain tissue available for both PriLer and TWAS results, and the red dot in each boxplot indicates the mean. Differences in distributions are tested via Wilcoxon-Mann-Whitney test, *: $0.001 < P \leq 0.01$, **: $0.0001 < P \leq 0.001$, ***: $P \leq 0.0001$, tissue label in bold and italic style indicates that the median of gene CV corr^2 differences between PriLer and TWAS is > 0 i.e. PriLer achieves an overall better model performance.

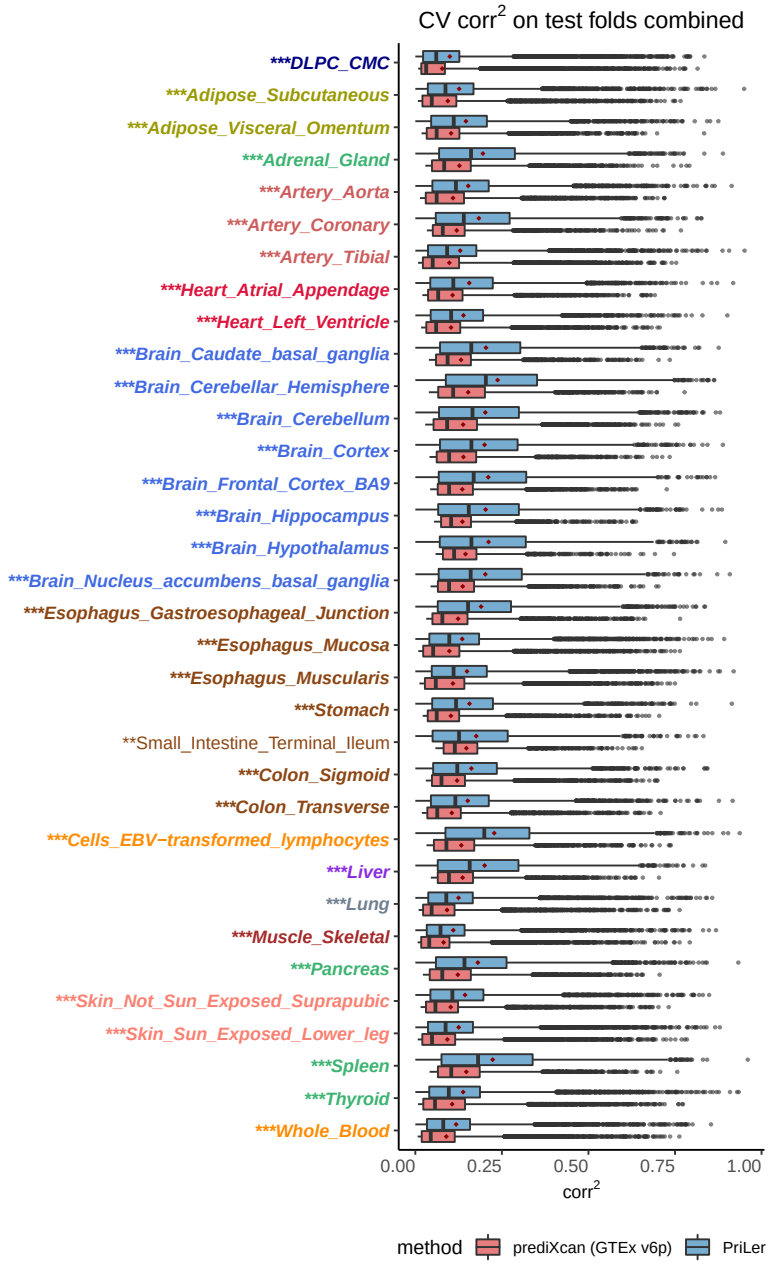

Fig. C.4.: Distribution of CV squared correlation between true and predicted gene expression computed on combined test folders. Each element in a boxplot is a gene in a certain tissue available for both PriLer and prediXcan results, and the red dot in each boxplot indicates the mean. Differences in distributions are tested via Wilcoxon-Mann-Whitney test *: $0.001 < P \leq 0.01$, **: $0.0001 < P \leq 0.001$, ***: $P \leq 0.0001$, tissue label in bold and italic style indicates that the median of gene CV corr^2 differences between PriLer and prediXcan is > 0 i.e. PriLer achieves an overall better model performance.

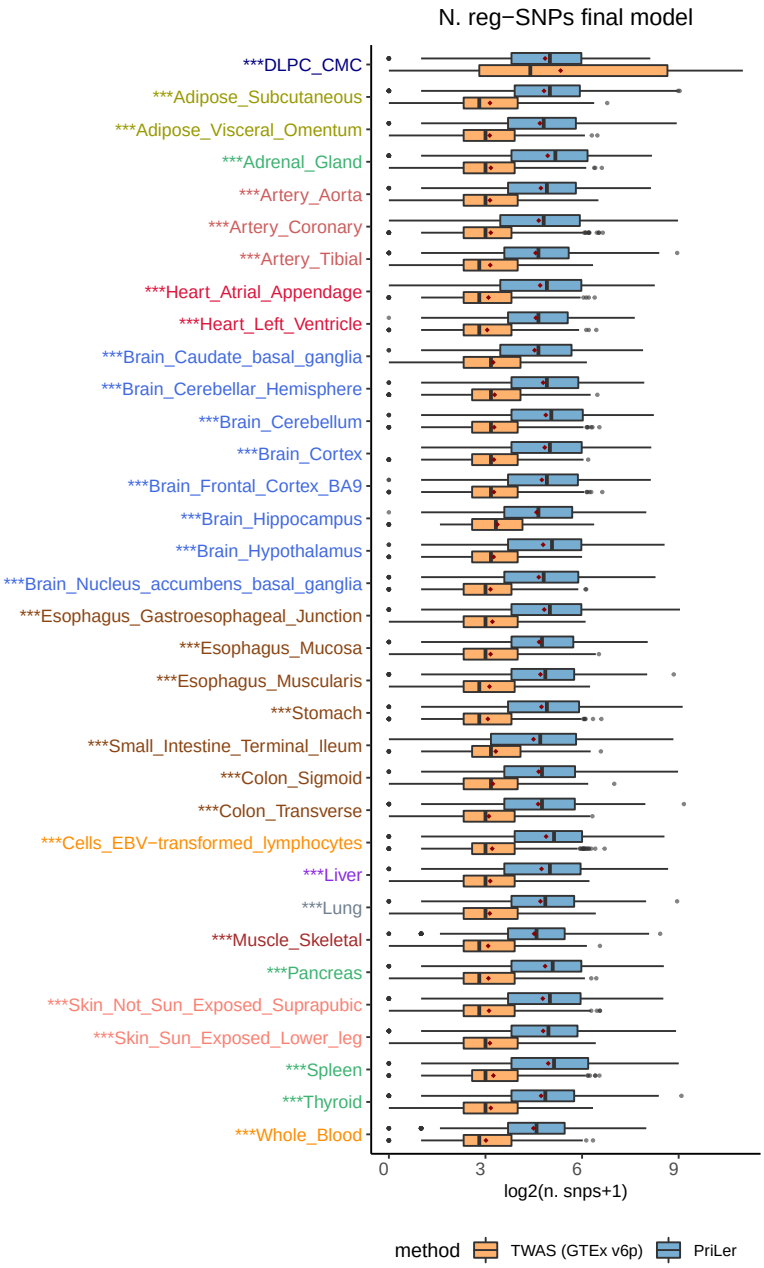

Fig. C.5.: Distribution of n. of reg-SNPs converted in log2-space in the final gene expression model. Each element in a boxplot is a gene in a certain tissue available for both PriLer and TWAS results and the red dot in each boxplot indicates the mean. Differences in distributions are tested via Wilcoxon-Mann-Whitney test, *: $0.001 < P \leq 0.01$, **: $0.0001 < P \leq 0.001$, ***: $P \leq 0.0001$, tissue label in bold and italic style indicates that the median of n. reg-SNPs differences between PriLer and TWAS is < 0 i.e. gene expression in PriLer is modelled using less variants, otherwise text style is plain.

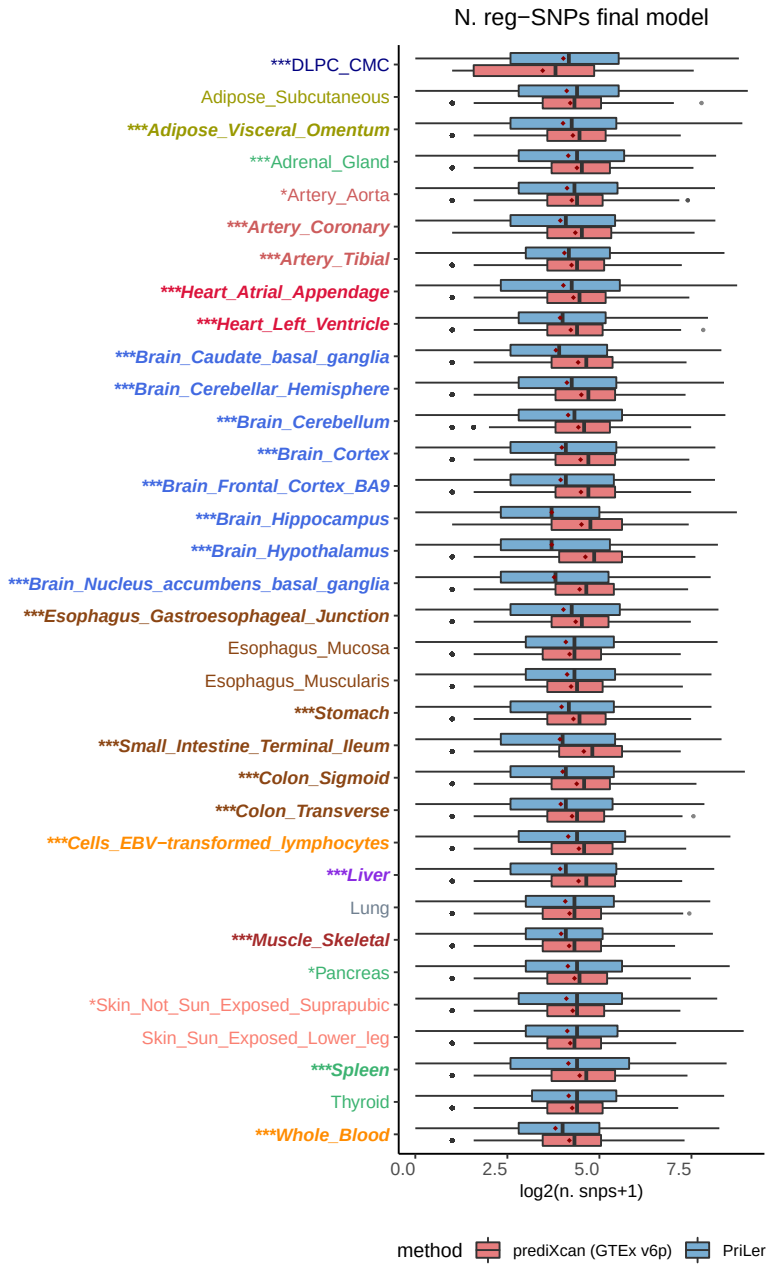

Fig. C.6.: Distribution of n. of reg-SNPs converted in log2-space in the final gene expression model. Each element in a boxplot is a gene in a certain tissue available for both PriLer and prediXcan results and the red dot in each boxplot indicates the mean. Differences in distributions are tested via Wilcoxon-Mann-Whitney test, *: $0.001 < P \leq 0.01$, **: $0.0001 < P \leq 0.001$, ***: $P \leq 0.0001$, tissue label in bold and italic style indicates that the median of n. reg-SNPs differences between PriLer and prediXcan is < 0 i.e. gene expression in PriLer is modelled using less variants, otherwise text style is plain.

List of Figures

List of Tables