GENETIC ARCHITECTURE OF COMPLEX PSYCHIATRIC DISORDERS

DISCOVERIES AND METHODS

by

Zhiyu Yang

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Biological Sciences

West Lafayette, IN

December, 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Peristera Paschou, Chair

Department of Biological Sciences

Dr. Petros Drineas, Co-Chair

Department of Computer Science

Dr. Daisuke Kihara

Department of Biological Sciences and Computer Science

Dr. Michael Gribskov

Department of Biological Sciences and Computer Science

Approved by:

Dr. Jason R. Cannon and Dr. Janice P. Evans

Coffee is overrated. Only junk food brings happiness.

ACKNOWLEDGMENT

First and foremost I would like to thank my supervisors Prof. Peristera Paschou and Prof. Petros Drineas. They are absolutely the most helpful and supportive mentors a gradstudent can expect. As a “woman in science”, I am truly grateful to have Prof. Paschou as a role model in my career development. I would also like to thank my committee members Prof. Michael Gribskov and Prof. Daisuke Kihara, for their suggestions and guidance alone this journey.

I also want to express my gratitude to everyone in Paschou lab: thanks for bearing with me these years and helping me with all sorts of troubles in both research and life. Many thanks to Pritesh Jain for making it possible for me to see the fridge before graduation, Apostolia Topaloudi for those wonderful trips we went on together, and Yin Jin for all the interviewing tips thanks to which I hopefully got a job. It has been a long, strange but full of fun trip because of you guys. Also I must thank Fotis Tsetsos, our colleague and the first friend I made in the US. Fotis was the one introduced me to the fascinating world of human genetics. Without the passion and enthusiasm he showed when I first joined the lab, I could have completely missed the beauty of this research field. Despite his short stay in the lab, he has been influential to me and others in the lab

I want to thank my friends for helping me forget about all gloominess ever so often. Thanks Sanniv for bringing dramas into my dull life, Kaijun for teaching me various things I missed out from college, and Rosanne for inviting me over for all those holiday house parties. And Qian, somehow you manage to appreciate whatever weird stuff I cooked, which really means a lot for me. I would also like to thank those friends back in China who share life, books and video game tips with me through online chatting. I really miss you guys and cannot wait to see you again. Speaking of video games, thanks to Nintendo for not releasing Zelda Breath of the Wild 2, or else there would be no way I could finish this dissertation on time.

Lastly, I would like to sincerely acknowledge my family, for all the unconditional support they provide throughout my life. My father is the main reason I chose to take an academic career path, while my mother on the other hand, teaches me the importance to enjoy life. As for my beloved boyfriend Sam, despite the troubled times, our love for each other grew stronger over the years (hopefully). Knowing that these people will stand by me gives me the courage to move forward fearlessly.

There are still numerous others who have befriended me during these five years that I cannot list. I believe every encounter has a meaning. For whomever I have crossed paths with, I truly appreciate your companionship and you all became a part of me. With that, I would like to wrap up my acknowledgment by thanking Sonia, the cutest hedgehog in the world: your soft belly heals everything!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ADHD attention deficit/hyperactivity disorder

AIC Akaike Information Criteria

AN anorexia nervosa (AN)

ASD autism spectrum disorder

BD bipolar disorder

CD/CV “common disease, common variant” hypothesis

CD/RV “common disease, rare variant” hypothesis

CFI Comparative Fit Index

CNV copy number variants

EFO Experimental Factor Ontology

eQTL expression quantitative trait loci

GO gene ontology

GSMR generalized summary-data-based Mendelian randomization

GWAS genomewide association study

HEIDI heterogeneity in dependent instrument

HPA hypothalamus-pituitary-adrenal

HWE Hardy-Weinberg Equilibrium

IOCDF-GC International Obsessive Compulsive Disorder Foundation Genetics Collaborative

kb kilobase

LD linkage disequilibrium

MAF minor alleles frequency

Mb megabase

MDD major depressive disorder

mtCOJO multi-trait-based conditional and joint analysis

MVP Million Veteran Program

NGS next-generation sequencing

nsSNP non-synonymous single nucleotide polymorphism

OCD obsessive-compulsive disorder

OCGAS OCD Collaborative Genetics Association Studies

OR odds ratio

PheWAS phenomewide association study

PGC Psychaitric Genomics Consortium

PRS polygenic risk score

SCZ schizophrenia

SE standard error

SNP single nucleotide polymorphism

SNV single nucleotide variant

SRMR standardized root mean square residual

TAAICG Tourette association of America International Consortium for Genetics

TS Tourette's syndrome

WTCCC Wellcome Trust Case Control Center

WGS whole genome sequencing

ABSTRACT

Impacting individual's social and physical well-being, psychiatric disorders have been a substantial burden on public health. As such disorders are frequently observed aggregating in families, we can expect a large involvement of heritable components underlying their etiologies. Therefore, studying the genetic architecture and basis is one of the most important aims toward developing effective treatments for psychiatric disorders. The overall objective of this dissertation is to contribute to understanding the genetics of psychiatric disorders. Analyzing summary statistics from genomewide association studies (GWAS) of psychiatric disorders, we mainly present results of two projects. In the first one, we evaluated commonalities and distinctions in genetic risk of four highly comorbid childhood onset neuropsychiatric disorders: attention deficit/hyperactivity disorder (ADHD), autism spectrum disorder (ASD), obsessive-compulsive disorder (OCD) and Tourette's syndrome (TS). Through systematic analysis of genetic architecture and correlation, we confirmed exitance of genetic components shared across ADHD, ASD and TS, as well as OCD and TS. Subsequently, we identified those components at variant, gene, and tissue specificity levels through meta-analyses. Our results pointed toward possible involvement of hypothalamus-pituitary-adrenal (HPA) axis, a human stress response system, in the etiology of these childhood onset disorders. The second project includes the proposition of a novel framework for general GWAS summary statistics-based analyses. Instead of regular odds ratio and standard errors archived in the summary statistics, we proposed a recounstruction approach to rewrite the results in terms of single nucleotide polymorphisms (SNP) allelic and genotypic frequencies. We also put forward three applications built-upon the proposed framework, and evaluated the performance on both synthetic data and real GWAS results of psychiatric disorders for each of them. Through these three applications, we demonstrated that this framework can broaden the scope of GWAS summary statistics-based analyses and unify various of analyses pipelines. We hope our work can serve as a stepping-stone for future researchers aiming at understanding and utilizing GWAS results of complex psychiatric disorders.

1 INTRODUCTION

1.1 Psychiatric disorders and contribution of genetic factors to the outcome

According to the US National Health Expenditure Accounts, health care spending accounts for 17.7% of the overall share of 2019's gross domestic product in US and this cost has been growing steadily over the past decades. The public health burden and associated decrease in quality of life brought on by common and still incurable disorders highlight the need to understanding their etiology, identify susceptible individuals, and develop effective treatments. Caused by the interaction between genetic and environmental factors, complex diseases, including various cancers, psychiatric disorders, cardiovascular diseases and more, predominate public health concerns [1, 2, 3]. However, due to complexity in their etiologies, elucidating the cause of such disorders has proven rather challenging compared to Mendelian and monogenic diseases, which are due to variation at a single genetic [4, 5].

The observation of an aggregation of cases in affected families usually points to the existence of a genetic basis for a specific disorder. On the other hand, visible variation of incidence over a short term (i.e. a few generations) implies environmental factors playing a role [3]. The amount of contribution to disease etiology made by genetic factors in a population is widely measured by disease heritability. Heritability is typically a value between 0 to 1, defined as the proportion of phenotypic variation of a particular trait that can be attributed to genetic variation [6, 3]. Conceptually, this metric provides an indication of how heritable a phenotype is in a specific population: the higher heritability a phenotype has, the higher resemblance we can expect between parents and offspring, and the phenotype can thus be considered more "genetic" in the target population; Mathematically, it is computed as the ratio of variances: genetic variance divided by phenotypic variance [6]. Due to differences in approaches to estimate these two variance terms, heritability for a disease is not always a constant even in the same population. However, there are highly heritable complex diseases long-known for "running in families", and psychiatric disorders make up a predominant category of disorders within this group [7, 8, 9]. For instance, family based studies showed heritability for attention deficit/hyperactivity disorder (ADHD) can be as high as 0.88 [10].For autism spectrum disorder (ASD) heritability has been estimated at ~ 0.9 [11, 12], and even in relatively less heritable psychiatric disorders, such as major depressive disorder (MDD) [8, 9], a heritability of > 0.3 can be observed [13, 14]. Psychiatric disorders are also among those disorders with the greatest impact in terms of both prevalence and patients' disability-adjusted life years [15, 16]. Given this significant societal impact and high estimated heritability, elucidating the genetic components that underlie the etiology of psychiatric disorders has drawn great interest in the genomics era.

1.2 Genetic variants behind complex diseases

The observed variation in the DNA sequence is the focus of genetic research. Assuming a "linear structure", such variation can mainly be divided into two types: point mutations that each takes place at a specific genomic locus, and structural mutations that can involve segments of the DNA sequence.

The former, usually known as single nucleotide polymorphisms (SNP) or single nucleotide variants (SNV), are the simplest form of genetic variation. They represent the occurrence of two (biallelic), or sometimes more than two (multiallelic), different bases at a certain genomic position in the population. SNPs are highly abundant, as they are estimated to occur approximately once in every 1,000 basepairs in the human genome [17, 18]. Most studies on SNPs focus on biallelic ones, as multiallelic SNPs are not as widely seen and are rather complicated to study [19]. Biallelic SNPs, as indicated by their name, present with two alleles (possible bases) in the population: usually the one less commonly observed is called a minor allele, and the other is called a major allele. Based on the frequency of the different observed minor alleles (MAF, by definition MAF ≤ 0.5) in a population, SNPs can be classified as rare (MAF $< 1\%$) and common (MAF $> 1 - 5\%$) [20].

Examples of structural variation are from the chromosomal abnormalities detectable through karyotype analysis (i.e. abnormal chromosome numbers, or translocations of large DNA segments from one chromosome to another) [21, 22], and copy number variants (CNV). CNVs are defined as gain (insertion) and loss (deletion) of DNA segments ranging from kilobases (kb) to megabases (Mb) in size [23, 24]. Another type of structural mutation is called copy neutral variation. This includes behaviors like exchange of DNA segments between two chromosomes (balanced translocations) and DNA segments reversely inserted in the chromosome (inversions). Unlike CNVs, such structural changes in the DNA sequence do not affect the total number of nucleotides in the chromosome [25, 26]. Similar to SNPs, these structural variants can also be categorized into common and rare based on their frequencies of occurrence in a population.

Most of the genetic variants are benign. Some may have an impact on phenotype, but not necessarily deleterious (e.g. genetic variants affecting eye, hair color etc. [27, 28, 29]). However, variations in the DNA sequence can result in changes in protein sequences (non-synonymous mutations) [30], gene expression [31, 32, 33], and various epigenetic regulatory behaviors [34] that lead to an altered gene expression, including DNA methylation [35, 36] and transcription factor binding affinity [37, 38]. Such effects can turn out to be pathogenic and account for disease heritability. Several theories have been developed to characterize the relationship between genetic variants and disease susceptibilities, one of which is the "common disease, common variant" hypothesis (CD/CV). This hypothesis states that the origin of common diseases can be attributed to effects of numerous common genetic variants with low to modest penetration. On the other hand, another theory, known as the "common disease, rare variant" hypothesis (CD/RV) posits the opposite that rare genetic variants, each with a relatively higher risk, make greater contributes to the disease etiologies [39, 40, 41, 42]. Evidence has been found in favor of each theory [43, 44, 45]. However, most studies indicate that the heritability of most complex diseases cannot be explained by solely common or rare variants, neither can it be attributed just to SNPs or structural variants [46, 47, 48]. Each risk factor has its own role to play as part of the disease mechanism, and the phenotypic outcome is believed to result from the joint effect of all kinds of genetic variants [20, 49]. Although the exact genetic architecture for different complex disorders can differ markedly [50], figure 1.1 [20] serves as a classic and well-characterized illustration demonstrating the general relationship between MAF of a SNP and its genetic effect over disease susceptibilities. Such a model is supported by results from analyzing the distribution of odds ratios for common and rare variants [42]. The figure was developed to explain the contribution of SNPs to disease. However, common and rare structural variants can to a great extent behave in a similar manner [20, 51, 52].

1.3 Genetic architecture of psychiatric disorders

Even though genetic architecture varies from disease to disease, if we look specifically into psychiatric disorders, considerable degree of homogeneity can be expected. Most psychiatric disorders, including the aforementioned ADHD, ASD and MDD, turn out to be highly polygenic: meaning their genetic risk can involve a large number of common variants with small effects, as well as some rare variants with large effects and *de novo* variants [16, 53, 54]. This observation motivates researchers to dissect the genetic factors underlying etiology of psychiatric disorders from various angles. Investigating the impact of common SNPs throughout the genome had proven one of the most successful approaches being used to study the genetic factors underlying psychiatric disorders. It has been reported that a large proportion, usually one third to a half, of the heritability of psychiatric disease can be attributed to only common genetic variants [54]. When it further comes down to common SNPs, such ratios will decrease by definition. SNP effects are found to play a big role in outcomes of lots of psychiatric disorders. A study on 4,408,646 Swedish samples reports a family-based heritability of 0.38 for obsessive-compulsive disorder (OCD). Meanwhile, the SNP-based heritability for the same disease is as high as 0.28, indicating more than 70% genetic liability for OCD can be explained by common SNPs [9]. As an example on the lower end, it was found that in ASD, although highly heritable, common SNPs only account for around 15% of the total population observed heritbility [8]. Yet, the importance of SNP effects underlying ASD has been proved by multiple studies [55, 56, 57]. Accomplishments like these reinforce the rationale for scanning common SNPs across the genomes to seek for the ones that play a role in the onset of a psychiatric disorder. This objective can be achieved through carrying out genomewide association studies (GWAS).

1.4 GWAS and its success in psychiatric disorders

Since its first attempt in 2005 [58], GWAS has become one of the standard approaches to unveil the relationship between SNP effects and a phenotype. In a GWAS, association of each SNP in the genome with the phenotype of interest will be examined individually using unrelated samples in a population. This results in millions of hypothesis tests across the genome for humans. As one form of correction for these multiple tests, a genomewide significance threshold of $p < 0.05/10^6 = 5\times10^{-8}$

is widely accepted as a cutoff to distinguish credible associations in European samples (where 10^6 is approximately the number of independent genomic regions in Europeans [59]). Typically, association between a SNP and the trait of interest is evaluated through regressions, with trait measurement as the dependent variable and presence of the effective allele as the independent variable. Most of the time, researchers use linear regression for continuous traits and logistic regression for binary traits, along with appropriate covariates included in the model [60, 61]. The analysis is computationally efficient, and it allows hypothesis-free, unbiased search throughout the genome for potentially causative genetic variants [62, 63]. On the other hand, acquiring the data can be time and cost consuming. As shown in figure 1.1, effect sizes for most common SNPs are low to modest. Therefore, detection of significant and novel associations in a GWAS often requires large sample sizes, typically in the thousands or more, depending on the trait [61]. Experiments in such scale used to be difficult. Thanks to the advancement in technologies, GWAS has been more and more financially feasible in the past decade.

1.4.1 GWAS using SNP arrays

One of the most cost-effective approaches to access genetic variants on the human genome is through whole-genome genotyping using SNP arrays [64]. This technology takes advantage of the fact that SNPs on the genome are not completely independent from each other. Instead, highly dependent genetic variants in a neighbourhood, namely SNPs in linkage disequilibrium (LD), form blocks of few haplotypes (ie combinations of alleles along a DNA strand) [65]. The idea behind genotyping is that, instead of capturing every single variant throughout the genome, based on the LD structures, arrays are developed to capture only the “tagging SNPs” (tagSNPs) that well-characterize the haplotype blocks [66]. Based on this information, the rest of the variants can subsequently be filled out with “predictions” through statistical imputation [67]. This procedure requires high quality reference panels for identifying tagSNPs and accurately imputing the missing SNPs. The former has been achieved by the International HapMap project, which aimed to provide information on common patterns in human DNA sequences through studying the frequencies and correlations of genetic variants across different human populations [68]; As for the latter, the most widely used public reference panel was made available by the 1000 Genomes Project. The reference panel included high-quality haplotypes of 2,504 samples from 26 populations obtained through a combination of sequencing and genotyping techniques, with information of not only SNPs, but also structural variants available [69]. As a result of recombination and mutations, LD structures are population specific [70]. Therefore, choosing appropriate reference panel based on the population distribution of study samples can be crucial [71, 72, 73].

Representing a relatively mature technology, genotyping using SNP arrays has been proven to be extremely economic and reliable [64, 74]. On top of a highly accurate output of the directly genotyped SNPs, pipelines have been developed and refined for stringent quality controls, imputation and association analyses. All these efforts throughout the past decade led to sets of best-practice protocols for SNP array based GWAS [75, 76, 77]. However, as a tradeof for their efficiency, drawbacks for genotyping using SNP arrays are also apparent. First, as mentioned before, the quality of imputation is highly dependent on the reference genome used. For GWAS on populations with well-established reference panels, this should not be as big a problem [78]. However, when it comes to studies on under-represented populations, using a population-specific reference has been demonstrated to be preferable, but is not always feasible [71, 72, 73]. Second, recall that when using SNP arrays, only predefined tagSNPs will be directly genotyped. By the nature of this technology, it will mostly not be able to capture rare variants (MAF $< 1\%$) since they are not likely to be the ones "characterizing the haplotype" [79, 66, 80], nor can they be reliably imputed using most publicly available reference genome. Accurate imputation of low-frequency variants requires large sample size in the panel, since an allele cannot be imputed unless it is first observed in the reference. As example of a precise but highly population-specific reference, the UK10K Cohorts project recently constructed a dataset of 3,781 samples with whole genome sequencing (WGS) data, aiming to characterize rare variants with MAF down to 0.1% only in the British population [81]. Meanwhile, the Haplotype Reference Consortium also created a reference panel of 64,976 human haplotypes, primarily of European ancestries, by unifying WGS data from 20 studies. The panel was also claimed to permit accurate imputation on SNPs with MAF as low as 0.1% [82]. However, both reference panels were generated with European samples. This limitation, may eventually bring us back to the first problem of reference dependency.

1.4.2 GWAS using WGS

To compensate for some limitations of SNP array-based GWAS, an alternative is using WGS data instead of tagSNPs and imputation. WGS, as implied by its name, provides a comprehensive collection of nearly all genetic variants in one's genome, including both common and rare SNPs, as well as structural variants and *de novo* mutations [83]. With the evolution of next-generation sequencing (NGS) technologies, cost of sequencing a complete human genome has decreased drastically in the past decade: from ∼\$100 million in 2004 to today's ∼\$1000, making a nearly 10,000-fold reduction [84]. Unlike the traditional Sanger sequencing [85], NGS allows massive sequencing reactions and base detection to proceed in parallel thus is able to provide high-throughput DNA sequencing [86, 84].

WGS has evident advantages over array genotyping as it captures not only all common SNPs, but also rare variants [64]. Moreover, it also allows analyses on structural and *de novo* variants, which can possibly make up for some "missing heritability" from common SNPs [20, 49]. Therefore, analyzing WGS data is a promising alternative, not only for SNP array-based GWAS, but for research on genetic variants in general. However, considering the cost, we may realize that \$1000 per sample is not as affordable for many labs, given the large sample size a GWAS typically requires. On the other hand, the cost of genotyping using SNP array can be as low as $<$ \$50 per individual [64, 86]. Apart from the cost, another fact to be noted is that fast evolving NGS and its workflows are still facing imperfections, which can lead to a failure in detecting many pathogenic variants [87]. Due to technical challenges, not all genetic variants can be sequenced in a reliable manner [88, 89, 90, 91]. On top of that, the human genome contains "dark" regions that create troubles during sequence assembly or alignments [92, 93]. Another problem for WGS is that, as a relatively newly emerging technology which is still under development, pipelines for analyzing WGS data are not as ready-to-use as the ones made for SNP array based analyses, and they usually come with heavier computational load and higher requirement for bioinformatic proficiency [64].

1.4.3 GWAS in psychiatric disorders

Starting with the first GWAS on bipolar disorder published as one of the seven common diseases studied by the Wellcome Trust Case Control Center (WTCCC) in 2007 [94], GWAS on psychiatric disorders will soon be celebrating its 15-year anniversary. Most of the GWAS have their results archived in NHGRI-EBI GWAS Catalog, a publicly available database recording findings on associations between SNPs and any trait of interest [95]. Taking a systematic look into those records, we can see that by July 2021, ten reported traits, covering 72 child traits according to the Experimental Factor Ontology (EFO) hierarchy, for `mental or behavioural disorders` have been studied in 873 publications. 10,098 associations have been reported. Out of which, 4,624 reached genomewide significance. Due to various limitations mentioned in the previous sections, as yet most GWAS studies are still SNP array genotyping-based. These studies elucidate networks of relationship between common genetic variants and a wide range of phenotypes. Figure 1.2 from [96] captures the tip of the iceberg for such networks. With growth of sample sizes in psychiatric GWAS, the pleiotropic nature of more SNPs is being revealed.

GWAS uncovered the role of common SNPs underlying many psychiatric traits. One of most successful application of GWAS was observed in schizophrenia (SCZ), forfor which more than 200 significant associations in 176 independent genetic loci have been identified as contributing to the genetic risk of SCZ through GWAS and meta-analyzing GWAS results [97, 98, 99, 100, 101]. Part of this achievement can be attributed to the highly heritable nature of SCZ (with family based heritability estimation of ~ 0.8 [102, 103]), especially when SNPs accounts for a good proportion of its heritability (estimated SNP heritability 0.23-0.24 [97]). This success is also due to collaborative efforts and increasing sample sizes. One of the biggest GWAS on SCZ to date included 22,778 cases and 35,362 controls just from East Asia. When meta-analyzed with the European samples, the total sample size was boosted to 56,418 cases and 78,818 controls [97]. Genetic homogeneity in SCZ across populations also helped with this process.

Not all psychiatric disorders can replicate such overwhelming success. For instance, MDD, as mentioned in section 1.1, is not as heritable. One of the most recent GWAS meta-analysis on MDD comprehensively included results from the three biggest MDD GWAS to date [104, 105, 106], encompassing a total of 246,363 cases and 561,190 controls. However, with more than five times the sample size of SCZ, only 102 independent variants were detected [107]. Being one of the leading causes of worldwide disability with a soaring population prevalence [108], MDD will never stop attracting efforts to unravel its genetic and non-genetic risk factors.

Another example is ASD, which is also highly heritable, but not with much heritability conferred by common SNPs (see section 1.3). By more than doubling the sample size of previous GWAS meta-analysis on ASD [55], only five gonomewide loci could be found significantly associated with the trait, with in total 18,381 cases and 27,969 controls used for the analysis [56]. On the other hand, with a slightly higher sample size (20,183 cases and 35,191 controls), 12 independent loci were identified for ADHD [109].

Some disorders have been considered rare for a long time and have thus not attracted as large numbers of researchers and associated increased sample sizes. Tourette's syndrome (TS) is another childhood onset neurodevelopmental disorder with estimated 1% prevalence in children and adolescents, a family-based heritability of 0.77 [110] and a SNP-based heritability of 0.21 [111]. The first GWAS on TS carried out by the Tourette association of America International Consortium for Genetics (TAAICG) failed to identify any significant association, as only 1,285 cases and 4,964 controls were analyzed [112]. In the second attempt, with most of the samples from the first study also included, the total sample size reached 4,819 patients and 9,488 controls, which is still much smaller than most GWAS carried out for adult psychiatric disorders and some for adolescent ones. However, with an evident disadvantage in sample size, still one genomewide significant locus on chromosome 13 was picked up [111].

Overall, with its utility demonstrated by hundreds of studies, GWAS has proven to be one of the most beneficial tools in research on the genetic basis of psychiatric disorders in the past nearly one and a half decades. Encouraged by all these promising results, plans for further expanding the sample sizes, meta-analyses, and mega-analyses are on the table. Given the polygenic nature of psychiatric disorders, very likely what has been discovered so far is merely a tip of the iceberg. We firmly believe that by standing on the shoulder of giants, more will be achieved in the near future.

1.4.4 Limitations of GWAS

GWAS provides an unbiased and efficient way to interrogate millions of common variants spanning the genome. But just like any other methods, it has its own limitations. First of all, GWAS originated on the basis CD/CV hypothesis, although after long years of controversy, most researchers now agree on the polygenic nature of most, if not all, complex diseases [113]. By adopting a population design, the intention of GWAS mostly aims to test the effect of common SNP alleles that present in a considerable proportion of the population. As technology advances, rare variants can be captured by WGS, and some of them may have relatively larger impact on the trait. However, with a significant threshold of $p < 5 \times 10^{-8}$, usually an extremely large sample size will be required to identify a rare variant as genomewide significant. It is encouraged to study rare variants through other more focused and powerful approaches [114, 115, 116].

Secondly, the genomewide significance threshold for GWAS findings, although fully reasonable, can be harsh under many circumstances. Not only for rare variants, but also the detection of common SNPs with low effect size can also be penalized by such threshold. Moreover, emerging WGS results in more independent statistical tests across the genome than before, so the bar can possibly be reset to an even more stringent level [117, 118]. To meet such a rigorous condition of significance, researchers are trying to include more and more samples to boost the analysis power and organize large-consortia GWAS. However, such collaborations can create complications in data sharing, harmonization and merging, especially across multiple ethnicities. Also expanding sample sizes can be challenging for studies targeting isolated populations with unique genetic structure, although high genetic homogeneity may be able to make up for some power loss due to low sample sizes [64, 118, 119, 120].

Last but not least, one of the biggest concerns for GWAS lies in the interpretation of its results. With constantly increasing sample sizes, more and more SNP-phenotype associations are being detected. However, association by itself is far from sufficient for concluding a causal relationship between a trait and SNP effects. Due to LD structures, when a causal mutation is identified by the association test, usually nearby SNPs highly correlated with the variant will also be picked up, especially when the GWAS is carried out through array genotyping and imputation. Therefore, as a downstream step for GWAS, further fine-mapping is required to pinpoint the real causal variant out of all significant ones in a locus [121]. Another type of interpretation difficulty stems from the biological aspect. Around 90% of GWAS findings reside in intronic or intergenic regions, whose functionalities are rather difficult to verify through wetlab experiments [122]. When it comes to connecting genetic variants to biological etiology for complex disorders, we still have a long way to go. It is worth mentioning that there have also been attempts to skip the biological interpretation, and directly utilize results from GWAS in a more practical and data-driven way, such as individual trait outcome or risk prediction. Such attempts include polygenic risk score (PRS), an individual score developed to describe the genetic liability to a trait of interest [123]. However, bounded by SNP heritability, prediction models built upon merely GWAS results do not often achieve satisfying performance [124]. The same also applies to PRS, which in most of the time, only captures a very small amount of phenotypic variance when compared to other risk factors [125, 126, 127].

1.4.5 A peek into the future of GWAS

Despite of all the limitations mentioned in section 1.4.4, it is hard to ignore the huge success it has achieved, in both psychiatric and non-psychiatric disorders. To date, it is still considered one of the most efficient ways to scan for underlying genetic variants responsible for a trait. With advancements in computational power and machine learning techniques, there are researchers calling for the improvement or replacement of GWAS [128, 129, 130]. However, popularity of this method is not likely to fade drastically in the near future, or at least not before any proposed alternatives are proven to be truly sound and more effective.

The pipeline for GWAS is relatively mature, but it will keep evolving to address some of the aforementioned concerns. To meet the stringent genomewide significance threshold, one trend is to expand research collaborations, combining data from different sources through meta or mega-analysis and reach larger sample sizes. Large consortia, such as the Psychaitric Genomics Consortium (PGC) [131], provide the platform for researchers to join forces. Meanwhile, researchers are also refining phenotypic measurements and selection of samples, in order to achieve higher power and result replicability by eliminating noise from the analysis.

Another idea trending in the community is to increase ethnic diversity of GWAS. A statistic done in 2009 showed that by then, more than 95% of GWAS participants are from European ancestry [132]. Ten years later, by July 2021, this ratio moved to 88.65% [133] (Figure 1.3), with the non-European samples predominated by Asians. Due to different LD structures, most GWAS results have low trans-ancestry portability. Among all variants at a significant genomic locus, SNPs showing a consistent effect in various populations are more likely to be causal themselves, rather than being picked up due to being in linkage with the real causal ones [134]. Therefore, more powerful GWAS targeting underrepresented populations are necessary, not only for population specific gene discovery, but also for further fine-mapping.

A third trend that can be observed is the growing use of WGS data in GWAS. This is not surprising, given the cost of NGS has been decreasing drastically in the past decade. In July 2021, out of in total 5,183 publications recorded in GWAS catalog, 15 are carried out, completely or partially, using WGS data. Out of those 15 studies, one was published in 2012, one published in 2014, two in 2015, five in 2017, two in 2019, three in 2020 and one in 2021. As discussed in section 1.4.2, currently technology is still not affordable for WGS studies with large sample sizes. However, with the advancement of NGS, such financial burden will be further reduced. It is reasonable to expect WGS to become more popular in the next decade or so. As an initiative, the UK biobank, which is one of the biggest biomedical data bank containing comprehensive phenotypic and genetic measurements of 502,543 individuals [135], has started their plan on genome sequencing. Last year, they released their first batch of results, including whole-exome sequences of 49,960 study participants [136]. As for biobanks in the US, there has also been news reports about the 125,000th whole human genome sequence being delivered to the Million Veteran Program (MVP) from department of Veterans Affairs [137] in June 2021. Another emerging effort is the All of Us Research Program. This program aims at enrolling at least one million participants across the US and creating the most diverse biomedical database [138]. Such projects make large-scale sequencing studies possible for more researchers, and create tremendous research opportunities for the scientific community.

1.5 Post-GWAS era and its progress in psychiatric disorders

In recent years, with the emergence of biobanks and increased curated accessibility to individual level genotype data, GWAS have been more massively carried out than ever. Numerous GWAS results, namely GWAS summary statistics, are made publicly available, pushing genetic research on complex diseases into a “post-GWAS era”. As mentioned in section 1.4, GWAS has been used as a “quick-and-dirty” approach to identify candidate genetic variants underlying disease etiology, although its results face problems as lack of theoretical interpretability or practical utility. Therefore, most post-GWAS efforts focus on improvement in the following two directions: 1. fine-mapping and digging into disease mechanisms implied by genetic findings; 2. exploring applications of GWAS results through a data-driven manner.

1.5.1 Annotation, fine-mapping and other biological insights from GWAS results

It is not hard to imagine that knowing the risk genetic variants and their genomic positions, in itself, is not as helpful, unless we can also learn their roles in the disease etiology. Therefore, appropriate annotations are necessary as a standard post-GWAS procedure. Genes have long been introduced as "the smallest unit of heredity". Despite of the controversy on that, it is obvious that comparing to individual nucleotide, function of genes have been much thoroughly studied [139, 140, 141]. Thus the most intuitive annotation for a SNP is to annotate it onto a gene by its physical location in the genome. However, as mentioned in section 1.4.4, most GWAS significant SNPs end up in non-coding regions. Even for those directly involved in gene coding, not all mutations are non-synonymous (nsSNPs), meaning many gene-coding SNPs, although can alter mRNA, do not lead to any changes in protein sequences. nsSNPs directly cause alteration in amino acids and are believed to have great impacts on subjects' health [142, 143]. Because of this, they receive more attention in functional studies, which results in abundant tools for nsSNP annotation and phenotypic effect prediction [144, 145, 146, 147]. Fortunately, thanks to purifying selection, nsSNPs tend to be rare, especially the deleterious ones [148, 149]. Since GWAS focus mostly on common SNPs, nsSNPs are rather unlikely to turn out as genomewide significant hits in the analysis. Synonymous and non-coding SNPs, on the other hand, are the predominant GWAS findings. There is growing awareness on important regulatory effects of these SNPs [150, 151, 152]. More databases and tools are developed to annotate non-coding SNPs by their utilities in modulating gene expression [153, 154, 155]. Once SNPs are mapped onto genes, we can subsequently proceed with gene, and even tissue or pathway level of analyses, built on top of SNP-based results from GWAS. Some ideas for gene-based analyses include computing disease-specific gene effects by aggregating the impact of SNPs annotated to each gene [156, 157], or using SNP effects as instrumental variables to look into causal relationship between gene expression and trait of interest through Mendelian randomization [158, 159]. With that, we can further look into disease tissue specificity through studying the enrichment of differentially expressed genes in each tissue [160, 161]. Similarly, enrichment can also be tested for genes in predefined biological pathways [162, 163, 164].

Another direction toward improving the interpretability of GWAS results is to identify causal SNPs within each risk genomic region through fine-mapping. By doing this, we can prioritize SNPs identified in GWAS for further functional validation through wet-lab experiments. This procedure assumes in any associated locus, at least one of the SNPs should be truly responsible for the trait outcome. Other SNPs within the same neighbourhood are also being picked up only because they are in linkage with the causal SNPs [165]. In general, fine-mapping strategies can be classified into two big groups: through functional annotation, and through statistical approaches. For the former strategy, we annotate all candidate SNPs in the region and highlight the functional ones that are more likely to play a role in disease etiology. For the latter, usually a measurement of "causality" will be assigned to each SNP in the region. Such measurement, can be a score calculated from association p-value and regional LD structure [166], coefficient in a penalized multiple regression [167, 168], or a probability of being the causal SNP [169]. Therefore, a Bayesian framework is widely used in statistical fine-mapping [170, 171, 172, 173]. It is also worth noting that although I classified fine-mapping strategies into two big categories, these two branches are not necessarily in an "either-or" relationship. In fact, most fine-mapping studies nowadays combine both strategies: looking into the function of credible causal SNPs obtained through statistical approaches [174, 175]; or directly incorporating functional annotation into SNP causality evaluation [176, 177].

Apart from information on specific genomic loci, by viewing the overall GWAS results, we can also learn valuable lessons regarding biological properties and genetic architecture of a trait. Methods have been developed to estimate phenotype SNP heritability and partitioned heritability by annotations from GWAS summary statistics [178, 179, 180]. Similarity in underlying genetic components between two traits can be evaluated through genetic correlation, which can also be obtained from GWAS summary statistics [181, 182]. Interrogating multiple disorders at the same time, we can study their joint genetic architecture by looking into shared latent genetic factors using structural equation modeling [183]. Furthermore, thanks to the nature of genetic variants, SNPs can be used as perfect instrumental variables, making it possible to infer causality relationships between the genetic risk of a exposure and target disorder through Mendelian randomization [184].

Partially due to controversial classification criteria [185, 186, 187], a transdiagnostic approach is highly encouraged for studying psychiatric disorders [188, 189, 190]. Accordingly, methods have also been developed to accommodate such demand in genetic studies. One of the goals of transdiagnostic research is to find common risk factors shared by multiple disorders that show phenotypic resemblance. In other words, to identify pleiotropic genetic variants which have high impact in multiple psychiatric disorders. Looking into shared genetic risk factors guides us to explore common pathways underlying the etiology of a spectrum of disorders, and provides hints in developing treatments in a more holistic manner. This can be achieved through cross-disorder meta-analysis and evaluating the posterior probability of association for each SNP [191], or more stringent hypothesis testing approaches [192]. Another perspective to view transdiagnostic analysis is that we can utilize information from closely related traits to help with understanding a specific disorder of interest. Although strong assumption on homogeneity in cross-disorder SNP effects is required, such multi-trait joint analysis of GWAS summary statistics indeed provides a great power boost in return [193]. Besides commonalities, sometimes we are also interested in genetic distinctions between two psychiatric disorders. Toward that end, we can run a conditional analysis to get the marginal SNP effects in one disorder conditioned on the effects in the other [184]. An even more straightforward alternative is to run a case-case GWAS, where the healthy controls are replaced by patients from another disorder of interest. Recently, a GWAS summary statistics-based method was also developed for such analysis [194]. However, a weakness of almost all aforementioned methods is that, just like any regular GWAS, usually their results only provide clues about association, without further knowledge on causality. A solution to this problem can be multi-trait colocalization-based methodologies. Such methods combine the idea of variants detection and fine-mapping, and have been used to study whether a disorder of interest has shared or distinct causal genetic factors with an intermediate or related trait [195]. This method has recently been extended from two-trait [196] to multi-trait analysis [197, 195].

1.5.2 Data-driven applications of GWAS findings

Another perspective on interpreting GWAS is to view it as a “feature selection” procedure, where each genetic variant is considered a candidate predictor. Since in GWAS, regressions are carried out with phenotype measurements as dependent variables, intuitively the “features” selected through GWAS will be used to predict individual trait outcomes. A most representative realization of this idea is polygenic risk score (PRS). It is a personalized score describing an individual’s genetic risk to develop a certain trait. Typically, PRS is computed as the weighted mean/sum of presence of risk alleles across one’s genome, where the weights, are usually derived from SNP allelic effects obtained from GWAS results on the same or another relevant trait [123]. In this process, GWAS can be viewed as a discovery, or "training" step if considered from a machine learning point of view, while the scores will subsequently be computed for another group of subjects whose individual level genotype is available and assessed by regressing over their phenotypic outcome. The latter, is equivalent to a validation or "testing" step [198]. As we can easily notice, general protocols for PRS computation and evaluation adopts a machine learning framework. Therefore, we should also expect it to be subject to the rules as well as general weaknesses of machine learning. First, like many machine learning techniques, best practices for PRS are ensured without excessive feature dependency [199]. To achieve this, we usually extract independent SNPs from GWAS results prior to PRS computation [200], or re-estimate SNP weights based on the LD structure using penalized regressions [201] or Bayesian methods [202, 203, 204]. Another principle of machine learning is that to avoid over-fitting, training samples should not be reused anyhow in the testing step. This elementary rule has long been a common sense for data scientists. However, in practice, it is surprisingly difficult to implement for PRS evaluation. The reason is that since "training step" for PRS is a GWAS, it is usually carried out and published as an independent study by another research group. In most of the time researchers computing PRS, on the other hand, can only get access to the GWAS summary statistics. Even though individual-level genotypes for the testing samples are available, as summary statistics preserves privacy, there is no way to verify if any of their testing samples also participate in the original GWAS. Such unknown sample overlap can result in inflated result during performance evaluation [205]. Unfortunately, to the best of my knowledge, there is still no effective solution to this issue. One more universal pitfall in machine learning is that model performance to a great extent depends on the homogeneity of distributions for training and testing data. The statistical learning theory assumes all training and testing data points are independent and identically distributed [206], whereas such assumption can hardly held outside a lab. Reflection of this problem in genetics is population stratification. Due to LD, distribution of genetic effects can vary across populations, leading to low trans-ethnic portability of GWAS results and phenotypic variance explained by PRS [205, 207]. As one of the most popular research topics recent years, statistical geneticists have been putting lots of efforts on improving cross-population PRS prediction performance. Besides GWAS results, most approaches incorporate external information, such as SNP functional annotations [208] or target population specific information [209, 210]. Leveraging fine-mapping is one of the popular ideas toward this aim [211].

PRS is a very special prediction model widely used in genetic epidemiology research. It is nothing but an additive model that may seem too simple to be functional for any data scientist, whereas such simplicity brings along superior interpretability and therefore is highly praised. There are also many other data-driven applications of GWAS aiming at sacrificing such interpretability in exchange for better prediction performance. Most of them directly adopt readily developed machine learning techniques, such as support vector machine [212] or random forest [213]. However, most of such attempts fail to receive as much improvements in prediction accuracy as expected. Nature of complex diseases is partly to blame for such low predictability. Recall that the amount of phenotypic variance that can be explained by genetic variables is measured by disease heritability. Therefore, it makes the upper bound for performance of any trait predictor based entirely on subjects' genetic sequences [124, 205]. Consequently, on top of GWAS, some researchers start to integrate additional information, for instance transcriptional risk factors [214] or association results of other relevant phenotypes [215], to enhance trait predictability. Even though most GWAS results still cannot be translated into a clinically reliable prediction model at this moment, genetic markers are promising predictors for identifying susceptible individuals.

1.5.3 Progress on post-GWAS research in psychiatric disorders

One of the fundamental goals for psychiatric genetic research is to go beyond genetic risk factors and elucidate biological mechanism, which can hopefully result in clinical insights [216]. This can be reflected in annotations and gene, tissue, pathway detection [99, 107, 109, 56, 217], or fine-mapping to unravel possible causal genetic variants [218, 219]. Apart from these standard analyses, many discoveries in psychiatric genetics origin arise from cross-disorder studies. As mentioned in section 1.5.1, transdiagnostic approaches are highly encouraged in psychiatric research. Psychiatric disorders are highly comorbid [220, 221, 222]. Even disorders being classified into different categories can show similarities in symptoms and aggregation in families [223]. By using cross-disorder designs, we can systematically profile commonalities and distinctions in etiologies underlying different psychiatric disorders. The history of genomewide cross-disorder analyses in psychiatric genetics can be traced back to a decade ago, when three of the most powerful GWAS by that time: in SCZ, MDD and bipolar disorder, were combined to find genetic variants playing a role in psychopathology [224]. Around the same time, shared genetic factors between ADHD and ASD [225], as well as OCD and TS [226], also came to researchers' attention. A few years later, the Cross-Disorder Group of PGC published results on jointly analyzing five major psychiatric disorders: SCZ, MDD and bipolar disorder, along with ADHD and ASD. It is still considered one of the biggest and most impactful genomewide cross-disorder studies to date [227]. Fast forward to the present day, a more recent example are the PGC efforts to uncover pleiotropic genetic variants across eight psychiatric disorders [228]. This recent PGC study systematically evaluated pairwise and joint genetic relationships of eight major psychiatric disorders, using the GWAS summary statistics of each trait. It sought genomic loci with pleiotropic effects in all or a subset of the eight disorders, and further investigated shared biological pathways implied by those loci. Another perspective in the analysis of shared genetic risk across psychiatric disorders is cross-disorder polygenic prediction. This is usually evaluated by the amount of phenotypic variance in a target disorder explained by PRS computed using GWAS results from another disorder. One of the earliest realizations of this idea was applied to bipolar disorder and SCZ, where polygenic risk components of SCZ derived from its GWAS results were found contributing to risk specifically of bipolar disorders, but not of many other non-psychiatric traits [229]. Such results provide evidence for shared polygenic basis across psychiatric disorders. In addition to psychiatric disorders, cross-trait analyses can also involve non-psychiatric phenotypes. Studies have shown that cross-trait associations can be uncovered between psychiatric PRSs and certain non-psychiatric traits as well, and vice versa [230]. The recent emergence of biobank data opens up the potential for efficiently studying associations between genetic risk and various phenotypic measurements, through what is called a phenomewide association study (PheWAS) [231]. PheWAS on psychiatric disorders using UK biobank samples have found association between polygenic risk of some major psychiatric disorders associated with a large variety of traits, including multiple sociodemographic, mental and physical health factors [232]. On top of that, cross-disorder causality relationship between psychiatric traits and health risk factors have also been widely reported [184, 233, 234]. Besides focusing on shared etiologies, investigating differential genomic loci for phenotypically related psychiatric disorders is another focus. Despite of considerable amount of homogeneity found across psychiatric disorders, they still show differences in symptoms and classifications. Studying distinctions in their genetic background can help our understanding of biological mechanisms underlying symptoms for specific disorders. Multi-trait conditional analysis has been used to study disorder specific SNPs for bipolar disorder, MDD, SCZ, ADHD and ASD respectively [235]. In another study, the authors tried to identify genetic variants differentiating SCZ and bipolar disorder through carrying out a case-case GWAS [236]. Conventionally such analysis would require the access to individual level genotype data, which may not always be feasible. However, recently a method was developed to extend case-case GWAS to summary statistics-based analysis. The author also applied the method to detect differential SNPs among eight psychiatric disorders [194].

1.6 Dissertation objectives

This dissertation presents discoveries in two "post-GWAS" directions, each constituting one section respectively: One of them focuses on novel discovery in genetics of complex psychiatric disorders through a series of systematic analyses, whereas the other describes the development and evaluation of a new methodological framework that unifies multiple post-GWAS analytical pipelines.

Section 2 presents a cross-disorder analysis on four common childhood-onset neuropsychiatric disorders including ADHD, ASD, OCD and TS. Using the GWAS summary statistics of each individual disorder, we identified shared and distinct genetic risk factors across the four disorders, which further brought about deeper biological insights into their etiologies.

Section 3 presents a novel framework that can be widely adopted by GWAS summary statistics-based analyses. To demonstrate the utility of the framework, we also put forward three applications built upon it, including group-wise PRS, which for the first time being computed without accessing to individual level genotype data. We tested all three applications on both synthetic data as well as real GWAS data. Powerful and robust results indicated great potentials held by the framework.

The objective of both projects was to contribute to better understanding GWAS and their results of complex psychiatric disorders. We hope improved knowledge on the genetic basis of these disorders can provide some directions for further development of advanced therapeutic intervention.

1.7 Figures

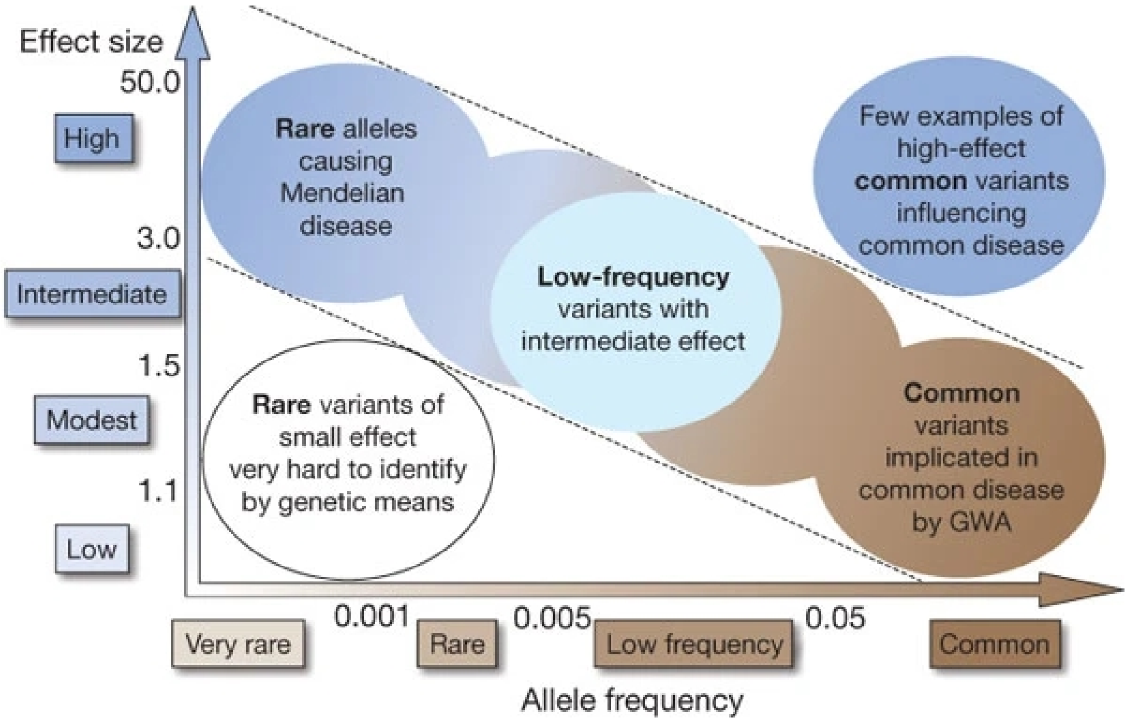

Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Few examples of high-effect common variants influencing common disease
Low-frequency variants with intermediate effect
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
Very rare
0.001
Rare
0.005
Low frequency
0.05
Common
Allele frequency

Figure 1.1: **General genetic architecture of complex diseases, figure cited from [20].**

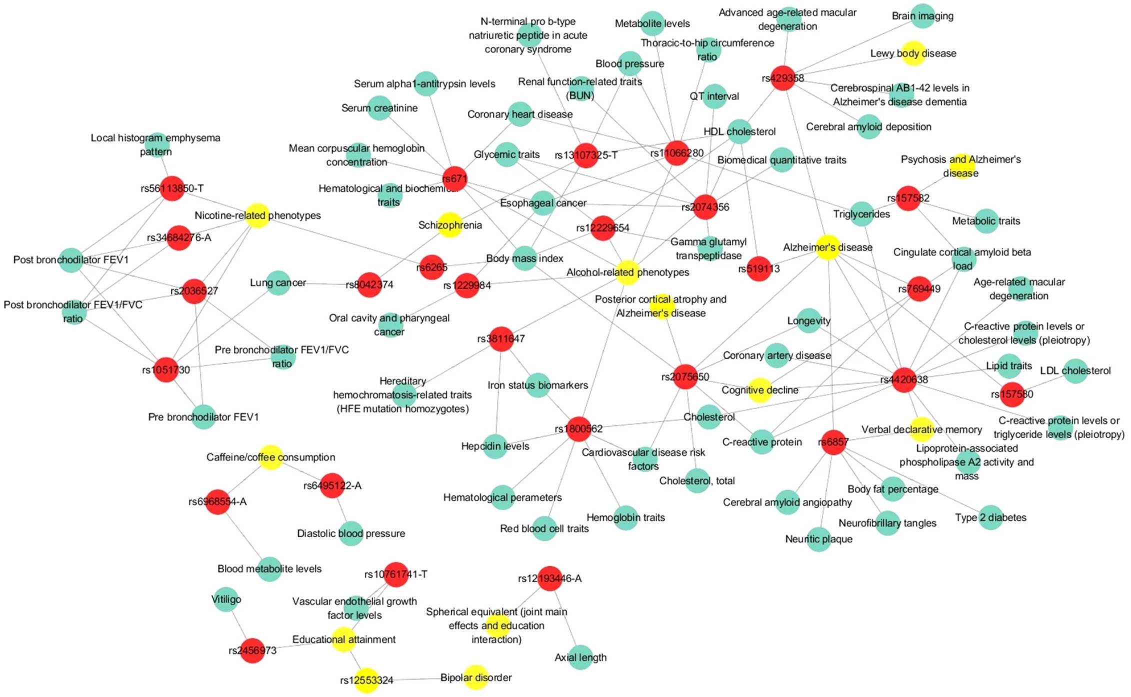

N-terminal pro b-type natriuretic peptide in acute coronary syndrome
Metabolite levels
Advanced age-related macular degeneration
Brain imaging
Thoracic-to-hip circumference ratio
Lewy body disease
Blood pressure
rs429358
Serum alpha1-antitrypsin levels
Renal function-related traits (BUN)
QT interval
Cerebrospinal AB1-42 levels in Alzheimer's disease dementia
Serum creatinine
Coronary heart disease
HDL cholesterol
Cerebral amyloid deposition
Local histogram emphysema pattern
Mean corpuscular hemoglobin concentration
Glycemic traits
rs13107325-T
rs11066280
Biomedical quantitative traits
Psychosis and Alzheimer's disease
rs56113850-T
rs671
Hematological and biochemi... traits
rs157582
Esophageal cancer
rs2074356
Triglycerides
Nicotine-related phenotypes
Schizophrenia
Metabolic traits
rs12229654
rs34684276-A
Gamma glutamyl transpeptidase
Alzheimer's disease
Cingulate cortical amyloid beta load
Post bronchodilator FEV1
rs6265
Body mass index
rs519113
Alcohol-related phenotypes
Lung cancer
rs8042374
rs1229984
rs769449
Age-related macular degeneration
rs2036527
Post bronchodilator FEV1/FVC ratio
Posterior cortical atrophy and Alzheimer's disease
Oral cavity and pharyngeal cancer
Longevity
C-reactive protein levels or cholesterol levels (pleiotropy)
rs3811647
Pre bronchodilator FEV1/FVC ratio
Coronary artery disease
Lipid traits
LDL cholesterol
rs1051730
Hereditary hemochromatosis-related traits (HFE mutation homozygotes)
rs2075650
rs4420638
Iron status biomarkers
Cognitive decline
rs157580
Pre bronchodilator FEV1
Cholesterol
C-reactive protein levels or triglyceride levels (pleiotropy)
rs1800562
Verbal declarative memory
Caffeine/coffee consumption
Hepcidin levels
C-reactive protein
rs6857
Lipoprotein-associated phospholipase A2 activity and mass
Cardiovascular disease risk factors
rs6495122-A
Cholesterol, total
Body fat percentage
rs6968554-A
Hematological perameters
Cerebral amyloid angiopathy
Hemoglobin traits
Type 2 diabetes
Diastolic blood pressure
Red blood cell traits
Neurofibrillary tangles
Neuritic plaque
Blood metabolite levels
rs10761741-T
rs12193446-A
Vitiligo
Vascular endothelial growth factor levels
Spherical equivalent (joint main effects and education interaction)
Educational attainment
rs2456973
Axial length
rs12553324
Bipolar disorder

Figure 1.2: **Network of associations between selected SNPs and phenoypes, figure cited from [96]**. SNPs (in red) in the figure are reported associated (with $p < 2\times10^{-8}$, which is stronger than a genomewide significance) with at least one psychological/behavioral/cognitive phenotypes (in yellow) and one other trait, not necessarily psychological (in turquoise color if not).

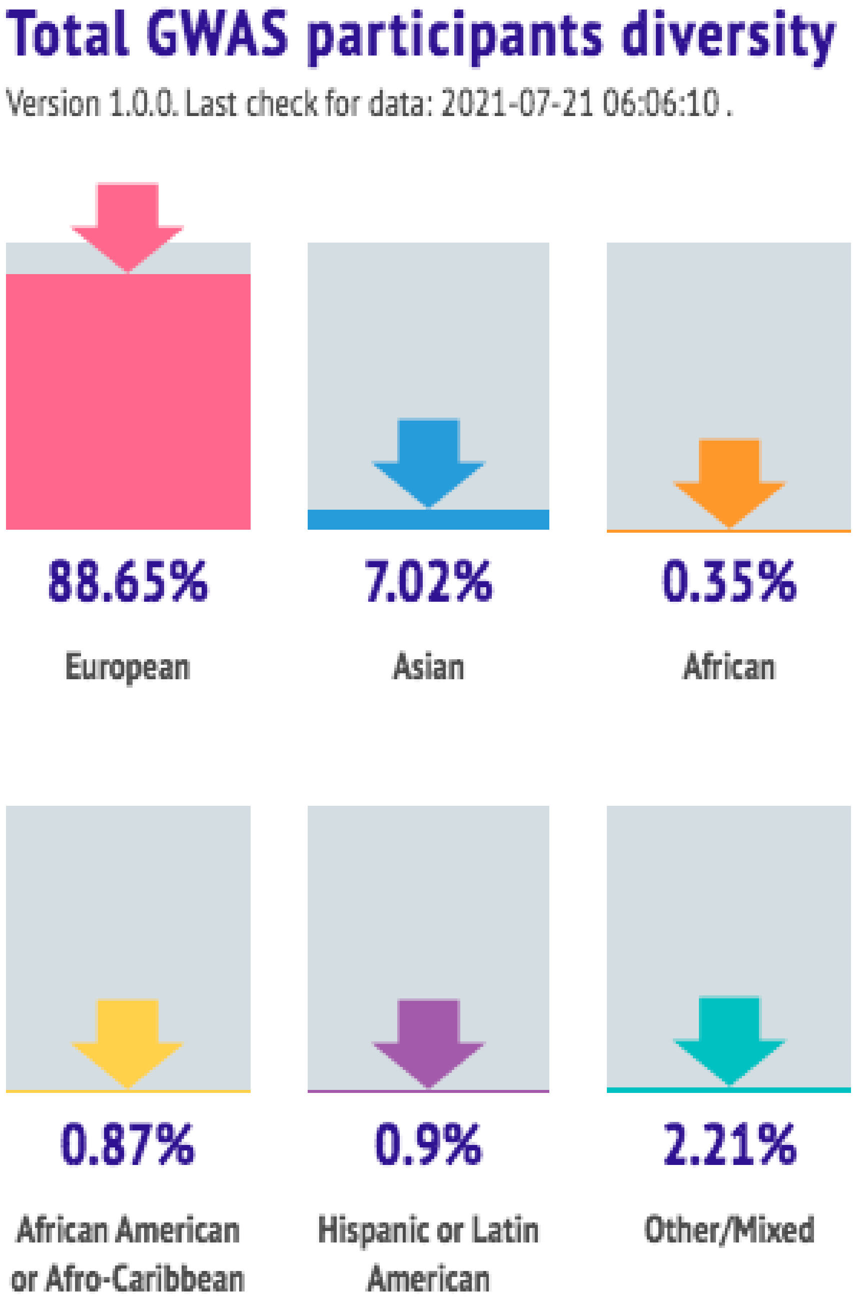


Figure 1.3: **Total GWAS participants diversity reported by GWAS Diversity Monitor [133].** The screenshot was captured on July 26th, 2021.

2 INVESTIGATING SHARED GENETIC BASIS ACROSS TOURETTE SYNDROME AND COMORBID NEURODEVELOPMENTAL DISORDERS ALONG THE IMPULSIVITY-COMPULSIVITY SPECTRUM

2.1 Abstract

Background: Tourette Syndrome (TS) is often found comorbid with other neurodevelopmental disorders across the impulsivity-compulsivity spectrum with Attention Deficit/Hyperactivity Disorder (ADHD), Autism Spectrum Disorder (ASD), and Obsessive-Compulsive Disorder (OCD) as most prevalent. This points to the possibility of a common etiological thread along an impulsivity-compulsivity continuum.

Methods: Investigating the shared genetic basis across TS, ADHD, ASD, and OCD we undertake an evaluation of cross-disorder genetic architecture and systematic meta-analysis, integrating summary statistics from the latest genomewide association studies (GWAS) (93,294 individuals, 6,788,510 markers).

Results: As previously identified, a common unifying factor connects TS-ADHD-ASD, while TS-OCD show the highest genetic correlation in pairwise testing among these disorders. Thanks to a more homogeneous set of disorders and targeted approach that is guided by genetic correlations, we are able to identify multiple novel hits as well as regions that seem to play a pleiotropic role for the specific disorders analyzed here and could not be identified through previous studies. In the TS-ADHD-ASD GWAS SNP-based and gene-based meta-analysis, we uncover 13 genomewide significant regions that host SNPs with a high posterior probability for association with all three studied disorders ($m - value > 0.9$), 11 of which were not identified in previous cross-disorder analysis. On the other hand, we also identify two additional pleiotropic regions in TS-OCD meta-analysis. Through conditional analysis, we highlight genes and genetic regions that play a specific role in a TS-ADHD-ASD genetic factor versus TS-OCD. Cross-disorder tissue-specificity analysis implicates the hypothalamus-pituitary-adrenal gland axis.

Conclusions: Our work underlines the value of re-defining the framework for research across traditional diagnostic categories.

2.2 Introduction

Tourette Syndrome (TS) is a common childhood-onset neuropsychiatric disorder that is often comorbid with other neurodevelopmental disorders along the impulsivity-compulsivity spectrum. In fact, only 10% of TS patients appear as pure TS, while up to 54.3% are also diagnosed with Attention Deficit Hyperactivity Disorder (ADHD), 50% have Obsessive Compulsive Disorder (OCD), and up to 20% have comorbid Autism Spectrum Disorders (ASD) [221, 237, 238]. The high comorbidity rates among these disorders have led to the hypothesis that TS, OCD, ADHD, and ASD might actually lie on an impulsivity-compulsivity continuum, sharing overlapping etiologies that converge in dysfunctional brain circuitries [238]. Here, pursuing a transdiagnostic approach, we seek to identify the common genetic factors and neural underpinnings across this spectrum of phenotypes.

TS, ADHD, ASD and OCD all have a complex and highly heterogeneous genetic architecture with both common and rare genetic variants contributing to their etiology [239, 240, 110, 241]). Over the past few years, twelve genomewide significant loci have been identified for ADHD [109], and five genomewide significant loci were described for ASD [55, 56]. For OCD no genomewide significant loci have been detected to date [242], while one genomewide significant locus was recently reported for TS [111].

Several cross-disorder analyses have previously evaluated the genetic overlap across these disorders revealing broad genetic correlations [228, 243, 244, 245, 227, 246]. Most recently, as part of the Psychiatric Genomics Consortium (PGC), we presented a data-driven meta-analysis of GWAS across eight common psychiatric disorders for which large GWAS data was available. Disorders analyzed included TS, ADHD, ASD, OCD, anorexia nervosa (AN), bipolar disorder (BD), major depression disorder (MDD), and schizophrenia (SZ) [228]. Exploratory factor analysis revealed that early-onset disorders including ADHD, ASD, and TS fell in one of the three identified factors (together with MD which is not typically early-onset). TS was also found weakly correlated in another factor together with compulsive disorders including OCD and AN. AN was however not found to be significantly correlated with TS in pairwise analysis and is not observed frequently in TS patients. This previous eight-disorder GWAS meta-analysis included multiple psychiatric disorders which are not clinically or genetically correlated to TS, thus possibly diluting relevant signals. Although power is high due to overall sample size, the trade-off is increased heterogeneity and thus difficulty to interpret results for one specific set of phenotypes that could be regarded as a group. Factor analysis, tests of pleiotropy and cross-disorder GWAS meta-analysis are all influenced by the input datasets and subject to change based on what disorders are analyzed. Therefore, in order to investigate a specific subset of traits that present with high comorbidity and high genetic correlation more focused cross-disorder studies are warranted.

Here, we build upon the PGC cross-disorder GWAS results as well as the high comorbidity and the existing hypotheses for shared etiology across TS and related disorders across the impulsivity-compulsivity spectrum. Our work highlights variants and genes that may contribute to neurobiology across this spectrum of neurodevelopmental phenotypes many of which could not be previously identified.

2.3 Method

2.3.1 Data sources

Analyses were conducted using summary statistics from GWAS for ADHD, ASD, OCD, and TS as made available by the PGC. For TS, we combined results from the first GWAS on TS, conducted by Scharf et al. [112] and newly collected cases and controls. In total, 4,232 cases and 8,283 ancestry-matched controls were used for the analysis, which resulted in 8,868,895 variants overlapping in the meta-analysis. These summary statistics correspond to the GWAS carried out by Yu et al. [111], excluding samples from the Tic Genetic Consortium. For ADHD, samples were collected by iPSYCH and PGC, with most of the samples genotyped using the Illumina PsychArray. Only samples of European ancestry were included in our analyses, comprising 19,099 cases and 34,194 ancestry-matched controls. In total, 8,047,421 variants overlapping across all cohorts after imputation were analyzed [109]. For ASD, we acquired the summary statistics of 18,382 cases and 27,969 ancestry-matched controls of European ancestry collected by iPSYCH and PGC. Most of the samples were genotyped with the Illumina PsychChip. After meta-analysis, 9,112,387 variants overlapping across sample sources were available [56]. For OCD, we used results from a meta-analysis of GWAS from two consortia: International Obsessive Compulsive Disorder Foundation Genetics Collaborative (IOCDF-GC) [247] and OCD Collaborative Genetics Association Studies (OCGAS) [248], which led to a total of 2,688 affected samples and 7,037 ancestry-matched controls from Europe. Samples were genotyped with multiple different Illumina's BeadChip arrays. After meta-analysis, 8,409,517 variants were found overlap and used for our study [242]. For all data obtained from the PGC, Ricopili pipeline or comparable quality controls were carried out.

2.3.2 Cross-disorder genetic architecture and GWAS meta-analysis

LD-score regression analysis was carried out using the LDSC package [182]. Only common SNPs (MAF > 0.01) with an imputation quality (INFO) score > 0.9 and matched with the provided HapMap3 SNPs reference were analyzed. LD scores estimated for the European samples from the 1000 Genomes phase 3 [69] were used as both the independent variable and the weight for the regression.

2.3.3 Investigating cross-disorder genetic architecture

In order to test for the presence of a common genetic factor that may underlie all traits of interest, we tested the common factor model using Genomic SEM for summary statistics of all disorders showing significant genetic correlation with TS. Prior to that, multivariable LDSC was carried out to obtain the covariance matrices using SNPs that survived the same quality controls for estimation of genetic correlation. Disease population prevalence used for the analysis were as follows: TS: 0.008, ADHD: 0.05, ASD: 0.01, OCD: 0.025. Fitness of model was evaluated using model chi-square, Akaike Information Criteria (AIC), Comparative Fit Index (CFI), and standardized root mean square residual (SRMR).

2.3.4 Causal risk factor inference

To estimate the causative association across traits, we carried out bidirectional generalized summary-data-based Mendelian randomization (GSMR) [184] across all disorders of interest. SNPs that are strongly associated with the exposure ($p < 5 \times 10^{-6}$) were used as genetic instruments. This threshold was chosen so that all the diseases could have more than 10 near-independent genetic instruments ($r^2 > 0.05$) for analyses therefore test power could be granted. A heterogeneity in dependent instrument (HEIDI)-outlier approach was carried out to exclude pleiotropic SNPs ($p_{HEIDI} < 0.01$) that affect the outcome through pathways other than the exposure factor. We used each trait as the target and the other three as exposures respectively and ran 12 independent tests, which made the significant threshold for this analysis $p < 4.17 \times 10^{-3}$ under Bonferroni correction.

2.3.5 GWAS meta-analysis

To investigate the genetic variants underlying the observed overlap, cross-disorder meta-analysis was carried out for TS-ADHD-ASD jointly, as well as pairwise between TS and significantly correlated disorders. SNP-based GWAS meta-analyses was performed using ASSET [249], which takes into account dependency across studies due to sample overlap [250]). For each study, the variants' effect sizes were measured by the logarithm of the odds ratio (OR). The possibility of inflation of results was investigated through observed λ as well as the sample size -corrected value λ_{1000}. Variants with meta-analysis p-values below the genomewide significance threshold ($p < 5 \times 10^{-8}$) were considered significant. To further highlight SNPs that contribute to risk across multiple phenotypes, we estimated the posterior probability of association (referred to as the m-value) with each disorder using a Bayesian statistical framework as implemented by MetaSoft [191]. An m-value threshold of 0.9 has been recommended to predict with high confidence that a particular SNP is associated with a given disorder.

2.3.6 Partitioned heritability analysis

We carried out SNP partitioned heritability analysis and cell type specificity analysis for the GWAS meta-analysis results using the LDSC package as described by Finucane et al. [179]. We investigated the possible enrichment of SNP heritability in 53 non-cell type specific annotation categories (baseline), including 24 main annotations and 29 extended annotations derived from the main annotations as defined in [179]. Even though these annotation categories were not mutually exclusive, we considered 53 as the number of hypothesis tested rather than 24, which gave us a more conservative significant threshold of $p < 9.43 \times 10^{-4}$ after multiple testing correction.

For the cell type specific annotations, we investigated the enrichment of SNP heritability in 13 brain relevant annotations from GTEx using reference created by Finucane et al. while controlling for the 53 baseline categories as defined in [179]. After multiple testing correction, the significant threshold for this analysis was $p < 3.85 \times 10^{-3}$. We also looked into the heritability enrichment in cell type specific chromatin states using cell type specific annotation made publicly available by Finucane et al. The annotation reference included for 489 cell type specific chromatin states. The results were subject to a significant threshold of $p < 1.02 \times 10^{-4}$.

2.3.7 SNP-based Conditional analysis

To further dissect the contribution of genetics on different groups of traits, we carried out multi-trait-based conditional and joint analysis (mtCOJO) [184] to adjust the summary statistics of TS-ADHD-ASD conditioning on TS-OCD and vice versa. Bidirectional causal effects between TS-ADHD-ASD and TS-OCD were first estimated using GSMR with strongly associated SNPs ($p < 1 \times 10^{-5}$). Genetic correlation, SNP-based heritability and potential covariance due to sample overlap were estimated through LD score regression, for which 1000Genomes phase 3 EUR subset was used as reference.

2.3.8 Gene-based cross-disorder GWAS analysis

Gene-based cross-disorder GWAS analysis was carried out using the MAGMA plug-in on the FUMA GWAS annotation platform [251, 252]. For this analysis, variants were mapped onto genes based on their exact physical positions without extended windows and aggregated association p-values were calculated for each gene. Analysis was carried out under a SNP-wise (mean) model. Considering the sample composition, a European ancestry reference from 1000 Genomes phase 3 was used as the reference panel. Analysis was done with the summary statistics of each disorder individually as well as all meta-analysis results obtained. Significance thresholds were set applying Bonferroni correction for each analysis, corresponding to the number of genes being tested.

2.3.9 Gene-property analysis for tissue specificity

To investigate phenotypic tissue specificity, a gene-property analysis testing for the relationship between tissue-specific gene expression and phenotype for associated genes was carried out using MAGMA for meta-analysis results with both GTEx v7 30 and 53 general tissue type expression atlas [141]. Significant thresholds for these analyses were p-value $< 1.67 \times 10^{-3}$ and p-value $< 9.43 \times 10^{-4}$, respectively, under Bonferroni correction. The analysis was done for all meta-analyzed results.

2.3.10 Gene-set analysis

Gene-set analysis was also performed using MAGMA under a default competitive test. Gene sets and gene ontology (GO) terms tested were obtained from MsigDB v 6.1, which contains 10,655 gene sets consistent across multiple sources. Bonferroni correction was applied to calculated association p-values to determine significance.

2.3.11 Results annotation

SNP-based annotation and gene mapping were carried out for significant SNPs with ANNOVAR [253], including functional predictions for all significant non-synonymous mutations using SIFT [254] and PolyPhen-2 [255] plug-ins of ANNOVAR. Regional plots for the top-variants were created for 400 kb windows using the LocusZoom platform [256]. For all significant results from our SNP-based and gene-based meta-analyses, we looked up previously reported associations in the GWAS catalog [95]. Aggregate functional information and tissue expression levels of the genes were acquired from the GeneCards database [139], the GTEx Portal [141], and the Expression Atlas [257]. Annotation of independent genomic risk loci from the FUMA GWAS platform was also adopted under parameters LD $r^2 < 0.6$ for SNPs with association $p < 5 \times 10^{-5}$ and within 1000 kb away from the significant lead-SNP ($p < 5 \times 10^{-8}$). GO-annotation and the over-representation tests were performed using the R package ClusterProfiler v3.0.4 [258]. Genes were mapped onto GO-terms based on org.Hs.eg.db [259]. Enrichment of GO-terms was evaluated through a hypergeometric test [260]. Network plotting was carried out using the built-in function of ClusterProfiler.

2.3.12 Transcriptome-wide association study

Association between the studied disorders and gene expression levels in the brain was evaluated through summary-data-based Mendelian Randomization. The SMR software was used and analysis was performed for each individual disorder as well as using results from our GWAS meta-analyses [159]. We used GWAS summary statistics for each studied disorder (as described above), the LD structure from from 1000 Genomes European reference panel and summary statistics from brain expression quantitative trait loci (eQTL) analysis [261], which quantified the effect of SNPs over gene expression levels in brain tissue [262, 263]. Only variants showing a consistent allele frequency (pairwise MAF difference between datasets no more than 0.20) across all three datasets (GWAS summary statistic, 1000 Genome reference, and eQTL summary statistic) were included in the analysis. All transcript probes with at least one cis-eQTL site showing $p_{eQTL} < 5 \times 10^{-8}$ were taken into consideration. SNPs affecting the same probe with LD $r^2 > 0.9$ or < 0.05 were pruned out from the analyses. Significance thresholds were based on Bonferroni correction for the number of probes tested.

To further verify that the effect of a probe on the trait was mediated by shared causal variants affecting both its expression and the trait rather than different variants in LD, we also carried out the HEIDI test to evaluate the heterogeneity in the effect sizes of SNPs over trait and expression for each probe, evaluated as pHEIDI. As a default of the software, only SNPs with $p_{eQTL} < 1.5654 \times 10^{-3}$ were taken forward for this analysis. Up to top 20 independent SNPs in the cis-eQTL region were used for each tested probe to optimize the test power. A $p_{HEIDI} > 0.05$ indicates the existence of a shared cause underlying the expression level of a transcript probe and the trait, suggesting dysregulation of the transcript is functionally relevant to the trait.

2.4 Results

2.4.1 Architecture of genetic correlations across TS, ADHD, ASD, and OCD

Here, we focus analyses on TS and highly comorbid neurodevelopmental disorders along the impulsivity-compulsivity spectrum. First, to set a foundation for our analysis, we repeated the measurement of genetic overlap across TS, ADHD, ASD, and OCD using LD-score regression (Table 2.1). Our analysis replicated the results from [228]. High genetic correlations were observed between all pairs of disorders, except for ASD and OCD. The highest genetic correlation was found between TS and OCD ($rg = 0.38$, $p = 2 \times 10^{-4}$). Interestingly, a negative genetic correlation was observed between ADHD and OCD ($rg = -0.17$, $p = 0.02$), although it was not significant under Bonferroni correction.

We proceeded with novel analysis that is focused on the specific set of TS-related disorders across the impulsivity-compulsivity spectrum. All of the tests carried out are influenced by the input datasets, thus, when compared to the PGC eight-disorder GWAS meta-analysis [228], results presented here have a direct interpretation for the neurobiology of the specific four disorders of interest. Since ADHD, ASD and OCD showed a high genetic correlation with TS, we tested for the existence of a common genetic factor across these four disorders using Genomic SEM. It should be noted that analysis with four phenotypes only allows the identification of a single factor. Results showed positive loads from ADHD, ASD and TS to the common factor, but not OCD (Figure 2.1, Table 2.4). The highest load was contributed by ADHD. This was in broad concordance with our previous work with eight disorders [228]. Based on these results, and the identified high genetic correlation between TS and OCD in pairwise analyses, we proceeded to pursue further analysis focusing on TS-ADHD-ASD and TS-OCD. In doing so, we aimed to increase homogeneity hoping to identify most relevant genetic signals.

2.4.2 Inferring causal relationships across TS, ADHD, ASD, OCD

To infer the potential causal relationship across the studied traits, we carried out bidirectional GSMR for all pairwise combinations across TS, ADHD, ASD and OCD. Results from this analysis point to broad causality networks across the studied disorders, indicating causal impact of the exposure disorder inducing the outcome disorder while using near-independent SNPs as instruments. After multiple testing correction, the significant threshold was $p < 4.17 \times 10^{-3}$. Under this threshold, our results indicated that being diagnosed with ASD is a causative risk factor for ADHD and vice versa. TS also showed a significant risk effect over OCD and ADHD turned out to be a risk factor for TS. Results can be found in Table 2.5 and Figure 2.1.

2.4.3 Cross-disorder GWAS meta-analysis for TS, ADHD, ASD, OCD

We carried out systematic SNP-based GWAS meta-analyses across TS, ADHD, ASD, and OCD using ASSET [249]. Combining all four datasets described above, 93,294 non-overlapping samples (51,311 controls) were available. We followed a different approach than the PGC eight-disorder meta-analysis study [228] and guided all subsequent analysis based on the genetic architecture of the studied disorders as revealed by Exploratory Factor Analysis rather than analyzing everything jointly. We first pursued meta-analysis of the TS, ADHD and ASD datasets yielding 6,815,585 overlapping SNPs. No obvious inflation was observed ($\lambda_{TS-ADHD-ASD} = 1.20$, $\lambda_{1000} = 1.00$). We identified seven independent regions with high evidence of pleiotropy ($m - value > 0.9$) across all three disorders (Figure 2.2, Table 2.2, Table 2.6, 2.7, 2.8). Despite reduced sample size, thanks to our more focused approach, we were able to identify here six genomewide significant regions harboring highly pleiotropic loci across TS-ADHD-ASD which were not identified as either TS-ADHD-ASD pleiotropic (at m-value threshold > 0.9) or genomewide significant in the PGC eight-disorder analysis (Table 2.2, Table 2.8, 2.9).

Since OCD was the disorder that was most closely correlated with TS but was not found to lie in the TS-ADHD-ASD factor, we also pursued pairwise analysis for the TS and OCD GWAS. 8,112,469 overlapping SNPs were available for analysis across TS and OCD ($\lambda_{TS-OCD} = 1.00$, $\lambda_{1000} = 1.00$). We found 21 genomewide significant variants in a single region (top-result rs140347666 ($p = 5.64 \times 10^{-9}$, $m_{TS} = 1$, $m_{OCD} = 1$); Figure 2.2, Table 2.3, Table 2.6, 2.7, 2.9); all significant SNPs were located in $LINC01122$ on region 2p16.1 and had the same direction of effect. All 21 SNP showed $m - value > 0.9$ for both TS and OCD, indicating high homogeneity across both disorders. This region had not been identified as genomewide significant in the PGC eight-disorder analysis and could be specific to the TS-OCD correlation. However, the PGC eight-disorder meta-analysis [228] had also previously identified six additional regions that were genomewide significant and had $m - value > 0.9$ in both TS and OCD (Table 2.3, 2.8, 2.9).

2.4.4 SNP-based Conditional analysis between TS-ADHD-ASD and TS-OCD

OCD showed a high genetic correlation with TS which cannot be explained by the same latent genetic factor as the group of TS-ADHD-ASD. Therefore, we tried to further explore the group-specific difference between TS-OCD and TS-ADHD-ASD through conditional analysis using mtCOJO. We expected a decreased effect in most of the SNPs after conditioning due to dependency caused by the fact that both groups include TS. However, we also found some SNPs with stronger effects after conditioning, which indicated that they play a role more specific to the particular group thus possibly leading to the differentiation of these two clusters. In the TS-OCD GWAS conditioning on TS-ADHD-ASD analysis only nine significant SNPs in the top region survived (compared to 21 in our original meta-analysis) (Table 2.11). None of them showed an increased effect after conditioning. On the other hand, in the TS-ADHD-ASD conditioning on TS-OCD analysis, 55 SNPs in six genomic regions showed a higher effect despite conditioning (including regions 1p34.1, 1p21.3, 4q24, 5q14.3, 5q21.2, and 10q25.1). These included two extra genomic risk regions that were only now revealed in TS-ADHD-ASD as independent from TS-OCD (region 1p34.1, gene $PTPRF$, $KDM4A$ and $ST3GAL3$; and region 4q24, gene $MANBA$) (Table 2.11). Among the six regions, most of them showed high ADHD-ASD specificity with low m-value for TS. However, we did identify one region hosting SNPs with an increased effect after conditioning, while also having high m-values for all three disorders analyzed (region 5q21.2).

2.4.5 Cross-disorder gene-based association analysis

We proceeded to perform gene-based analysis across the TS-ADHD-ASD and the TS-OCD GWAS meta-analyses as implemented in FUMA [252]. Our gene-based analysis highlighted 18 genes as significantly associated in the TS-ADHD-ASD meta-analysis (Table 2.12). 14 out of the 18 genes (including the top-result $SORCS3$ ($p = 4.97 \times 10^{-10}$) on chromosome 10) can also be picked up even if we only analyze those SNPs with $m - value > 0.9$ for all three disorders. Out of these 14 pleiotropic genes, only one is located in a genomewide significant risk region identified as pleiotropic for all TS, ADHD and ASD from the previous PGC eight-disorder analysis ($SORCS3$) (Figure 2.3, Table 2.2, Table 2.9). The rest of the identified regions could thus be of particular importance for early-onset disorders. On the other hand, for TS-OCD, we tested in total 18,790 genes, out of which four turned out to be significant. Gene $CADM2$ on chromosome 3 was the top one (Figure 2.3, Table 2.3, Table 2.9, 2.12). All the genes showed evidence of a pleiotropic effect, as they were also identified significant when we only analyzed SNPs with $m - value > 0.9$ for both disorders.

2.4.6 Pathway analysis, tissue enrichment, and partitioned heritability analysis

Partitioned heritability analysis revealed enrichment of the cross-disorder GWAS results in conserved regions for both TS-ADHD-ASD and TS-OCD (Table 2.10). Furthermore, we identified significant enrichment in brain frontal cortex cell type in the TS-OCD GWAS. Partitioning heritability by brain-cell-specific chromatin states, we found enrichment in fetal brain, brain germinal matrix and cortex in TS-ADHD-ASD. On the other hand, enrichment in chromatin states specific to the anterior caudate and dorsolateral prefrontal cortex were found in TS-OCD GWAS results (Table 2.10).

In order to better visualize our results while investigating the pathways and interactions among the top risk genes across TS, ADHD, ASD, and OCD, we constructed GO-based networks for the top 200 genes from each gene-based association analysis as well as genes annotated from the SNP-based GWAS meta-analyses. Results are shown in Figure 2.4 and 2.5. Pathways related to neuronal development, axonogenesis, and synaptic structure and organization were highlighted among the most significant in our analysis. These results were further strengthened by tissue specificity analyses, which showed enrichment of our top associated loci in genes expressed in brain tissues (Figure 2.6, 2.7). In the tissue specificity analysis with 53 tissue types (figure 2.6, table 2.13), significant enrichment was found for genes expressed in various brain regions including frontal cortex, basal ganglia, hypothalamus, cerebellum, amygdala, and hippocampus for TS-ADHD-ASD and cortex and frontal cortex for TS-OCD (Figure 2.7). In the 30 tissue types analysis, enrichment in expression in brain and pituitary arose as significant (Figure 2.6) for TS-ADHD-ASD. Interestingly, enrichment in genes expressed in the adrenal gland for TS-ADHD-ASD was also highlighted reaching borderline significance ($p = 1.89 \times 10^{-3}$, with a significance threshold of $p < 1.67 \times 10^{-3}$) (Figure 2.6, 2.7 and Table 2.13).

Next, we incorporated eQTL information into our meta-analyses and performed transcriptome-wide association analyses, aiming to identify genes with expression levels associated across the studied disorders. Results for the TS-ADHD-ASD combined transcriptome-wide analysis are reported in detail in supplementary results Table 2.14. Two transcript probes satisfying the pleiotropy hypothesis were significant, all located on chromosome 17. Among all significant transcripts, the top-result was from the LRRC37A4P probe ($p_{SMR} = 1.38 \times 10^{-6}$, $p_{HEIDI} = 0.10$). This corresponds to the transcript of a pseudogene in region 17q21.3, localizing near $KANSL1$. None of the probes were found significant for TS-OCD.

2.5 Discussion

Motivated by high comorbidity rates across studied phenotypes and a long-standing hypothesis of a shared etiological thread across disorders of the impulsivity-compulsivity spectrum, we present a detailed investigation of the shared genetic basis across TS and often-comorbid ADHD, ASD, and OCD. Our analysis is guided by the genetic architecture across the studied disorders as revealed by exploratory factor analysis as well as genetic correlations. Thus, our findings are not affected by analyzing jointly with disorders that are not genetically or clinically correlated. We confirm the existence of a unifying genetic factor across TS, ADHD and ASD and reproduce the high genetic correlation of TS and OCD that appears to be separate from the TS-ADHD-ASD factor. The identified negative genetic correlation between ADHD and OCD indicates that genetic variants operate in opposite directions in the development of these two disorders. From a clinical perspective, this is quite intuitive since ADHD and OCD may be thought of as lying at opposite extremes of the impulsivity-compulsivity continuum.

The increased power of a trans-diagnostic approach is once again highlighted by the discovery of novel genetic associations, not previously identified in individual GWAS. Furthermore, our study also highlights the value of increasing homogeneity across studied as we are able to identify here multiple novel pleiotropic loci across the disorders of interest that were not identified by the PGC eight-disorder meta-analysis [228] which included the four disorders of interest here. These loci could therefore be considered as specific for the four disorders we focused on. For instance, in the TS-ADHD-ASD meta-analysis, we successfully uncovered 16 LD-independent genomic risk regions (nine through SNP-based and seven through gene-based analysis), 13 of which are highly pleiotropic across all disorders analyzed. 11 were not previously identified as genomewide significant or pleiotropic by the eight-disorder meta-analysis, suggesting a specific role for the disorders that are the focus of our analysis.

The top-significant genomic risk locus showing also high probability for association across TS-ADHD-ASD was in gene LINC00461 on chromosome 5. This gene is highly expressed in brain and visual cortex, and has been previously involved in tumorigenesis [264]. Gene MIR9-2 is also located within gene $LINC00461$. The expression of this microRNA is almost brain-exclusive and has been found crucial during neuronal differentiation [265, 266]. $LINC00461$ was recently reported with high pleiotropic effects across five psychiatric traits [267]. Moreover, in the same study, behavior tests of expression knockdown mice confirmed the critical role it plays in neurodevelopment processes [267]. Interestingly, although this top region on chromosome 5 has also been previously highlighted as genomewide significant by the ADHD individual GWAS as well as results from the PGC eight-disorder GWAS, it was not reported among the most broadly pleiotropic ones and did not have high m-value for TS in that study. This is because of the nature of m-value computation and highlights the importance of fine-resolution cross-disorder comparisons. Since m-value measures the posterior probability of the SNP effect existing in a given disorder, it is subject to the result of meta-analysis, which, is further subject to the data input. Hence, if a SNP effect from the meta-analysis is significantly driven by one or a few disorders which are highly heterogeneous from the others, we may not capture evidence of such effect existing in other disorders even though the overall analysis has an increased power.

Gene-based meta-analysis also proved extremely powerful and led to the identification of multiple novel hits not previously identified by individual GWAS or the PGC eight-disorder meta-analysis [228]. From our TS-ADHD-ASD gene-based analysis, we identified 12 novel genes that could not be identified using the individual disorder summary statistics alone. The top result was $SORCS3$. The effect of this gene remained significant even if we analyze only SNPs with high m-values in all three disorders, indicating a potential pleiotropic effect. This gene encodes a member of the vacuolar protein sorting 10 ($VPS10$) receptor family, which controls intracellular protein signaling in neurons and regulates neuronal viability through many pathways [268]. It is highly expressed in brain tissues [262], and it has been previously implicated in neurological disease including ADHD and ASD etiology [109, 56]. Multiple studies indicate a relationship between $SORCS3$ and the accumulation of amyloid, which is linked to Alzheimer disease [269, 270]. It is also associated with major depression in individuals of European descent [105]. Moreover, its interaction with postsynaptic proteins, such as $PICK1$, indicates that the product of $SORCS3$ regulates glutamate receptor function [271, 272]. As one of the major neurotransmitters in the human brain, the glutamate pathway has long been hypothesized to underlie abnormalities in ADHD, ASD, and TS and is a possible therapeutic target for these disorders [273, 274, 275, 276].

In the case of TS-OCD meta-analysis, we identified three (one through SNP-based analysis and two through gene-based analysis) genomic risk regions, and all of them show pleiotropic effect across TS and OCD. Two of them were not identified by the PGC eight-disorder meta-analysis. On the other hand, the broader study reported six additional risk regions to be pleiotropic across TS and OCD but also other disorders highlighting the trade-off between power and homogeneity and the importance of combining different approaches. We found multiple significant hits on gene $LINC01122$ in chromosome 2 that showed evidence of pleiotropic effect in both disorders. Note that in the original TS GWAS carried out by Yu et al. [111], SNPs in this region were at borderline of genome-wide significance ($\sim 10^{-7}$).

Observing a structure that breaks up the studied TS-related phenotypes in a TS-ADHD-ASD and TS-OCD correlations, we also tried to identify group-wise differentiating effects through a conditional analysis. An intergenic region in 5q21.2 seems of particular importance: not only does this region host SNP with an increased effect in TS-ADHD-ASD conditioning on TS-OCD, but it also shows high posterior probability of association in all three disorders, indicating a group-specific pleiotropic effect. Duplication of the 5q21.2 region has been previously reported as a clinically significant copy number variation (CNV) in schizophrenia [277].

Among the top genes that we found associated in the TS-ADHD-ASD GWAS meta-analysis, we observed enrichment for genes expressed in the brain. Our results provide further support for the involvement of the basal ganglia across all disorders analyzed here. Dysfunction of the basal ganglia has been observed in all four studied disorders [278, 279, 280, 281]. Interestingly, we found significant enrichment in pituitary and hypothalamus expression, and, furthermore, the enrichment of adrenal gland expression was also borderline significant. This observation implicates the involvement of the hypothalamus-pituitary-adrenal (HPA) axis, in accordance with previous clinical studies implicating this system in multiple childhood-onset psychiatric traits including TS and ADHD [282, 283, 284, 285, 286]. The HPA axis plays a critical role in human stress response through the regulation of cortisol secretion [287]. Low-cortisol responsivity to stress was proposed as a biomarker for certain types of ADHD, indicating a possibly altered HPA axis activity in this disorder [288]. Altered cortisol levels among TS individuals have also been reported, with a negative correlation between evening cortisol and patients' tic severity and higher cortisol levels in response to stress [289].

Our analysis provides clues to potential biological distinctions between the studied subgroups of disorders along the impulsivity-compulsivity continuum. while the role of frontal cortex and basal ganglia was highlighted for both TS-ADHD-ASD and TS-OCD, support for HPA axis involvement and significant enrichment of chromatin states in fetal brain cell types was only observed in TS-ADHD-ASD. Our findings thus point to more significant contributions of neurodevelopment and stress-related processes in the TS-ADHD-ASD dimension in comparison to TS-OCD.

Although we provide results on combined datasets of very large size across TS, ADHD, ASD, and OCD, available datasets varied in size for each of the studied disorders. The unbalanced sample size across the studied datasets is one of the limitations of our study. In order to mitigate this problem, we placed emphasis on investigating and reporting the SNP posterior probability of association (m-value) for each disorder providing higher confidence for shared effect across multiple disorders. Existing overlap across the studied samples was relatively small ($< 6\%$ case overlap in the datasets that we studied) and we used ASSET, which takes into account known sample overlap, to control the inflation in meta-analysis results.

In conclusion, through a series of systematic genomewide association meta-analyses we uncovered multiple loci that may underlie biological mechanisms across the TS and its highly comorbid neurodevelopmental disorders along the impulsivity-compulsivity spectrum (ADHD, ASD, OCD). Despite the trade-off in power compared to the PGC earlier meta-analysis across eight disorders [228], we show that, increasing homogeneity when motivated by clinical observations we can identify many additional genomic risk loci that could play a more specific role across clinically correlated phenotypes. The existing evidence for a common genetic background across these highly comorbid disorders highlights what seems to become a recurrent theme across the studies on neuropsychiatric disorders: the importance of thinking across diagnostic boxes when attempting to understand neurobiology. Most importantly moving towards genomic analysis of symptom dimensions across diagnostic categories may prove extremely powerful but would require availability of very large and well-characterized cohorts of patients as well as the harmonization of existing clinical databases spanning the disorder spectrum.

2.6 Acknowledgements

This work was made possible thanks to the concerted efforts of the Psychiatric Genomics Consortium Cross-disorder Working Group, the Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH), the Psychiatric Genomics Consortium ADHD Working Group, the Psychiatric Genomics Consortium ASD Working Group, and the Psychiatric Genomics Consortium TS/OCD Working Group.

Barbara Franke is supported by a personal grant from the Innovation Program of the Netherlands Organisation for Scientific Research (NWO; grant 016-130-669). Her work is also supported by grants from the European Community's Horizon 2020 Programme (H2020/2014 – 2020) under grant agreements n° 667302 (CoCA) and n° 728018 (Eat2beNICE). Dr. Faraone is supported by the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 602805 (Aggressotype), the European Union's Horizon 2020 research and innovation programme under grant agreements No 667302 (CoCA) and 728018 (Eat2beNICE) and NIMH grants 5R01MH101519 and U01 MH109536-01. Dr Paschou is supported by an NSF grant #1715202.

2.7 Tables

*For large Tables 2.4 - 2.13, only legends are presented in this section. Those tables can be downloaded together as spreadsheets from this link.

Table 2.1: **Pairwise genetic correlation.** LD score regression analysis showing pairwise genetic correlation across ADHD, ASD, OCD, and TS. #SNPs = number of overlapping SNPs used in the analysis; Rg = genetic correlation; SE, P = standard error and p-value for Rg; Intercept (SE) = Intercept for genetic correlation and corresponding standard error.

	Disorder pairs					
	ADHD/ASD	ADHD/OCD	ADHD/TS	ASD/OCD	ASD/TS	OCD/TS
#SNPs	1042563	1030018	1062415	1012959	1044625	1100873
Rg	0.35	-0.17	0.26	0.12	0.18	0.38
SE	0.05	0.07	0.06	0.08	0.06	0.1
P	1.33×10^{-11}	0.022	2.05×10^{-05}	0.15	0.0055	0.0002

Table 2.2: **Comparison of results across regions that are shown as genomewide significant and pleiotropic in either the TS-ADHD-ASD GWAS meta-analysis (SNP-based or gene-based) or the PGC eight-disorder GWAS meta-analysis [192].** Regions highlighted in dark blue are identified in both studies. The light blue color highlights novel regions that are only identified as genomewide significant and pleiotropic in this study (ie crossing genomewide significance threshold and m-value¿0.9 across TS-ADHD-ASD). No highlight points to regions were criteria are satisfied only in the PGC eight-disorder GWAS meta-analysis. (Bold font: genomewide significant, *: SNP/Gene that are geneomewide significant and pleiotropic)

Region	SNP-based results	Gene-based results		Results from [192]	Original TS GWAS	Original ADHD GWAS	Original ASD GWAS
	Leading SNP p-value/OR	gene	p-value	Leading SNP p-value/OR	Leading SNP p-val	Leading SNP p-val	Leading SNP p-val
2p15	1.22×10^{-07} / 1.05	*WDPCP *MDH1	$\mathbf{6.08 \times 10^{-08}}$ $\mathbf{7.59 \times 10^{-07}}$	$2.40 \times 10^{-06}/0.94$	9.34×10^{-04}	7.27×10^{-05}	8.61×10^{-06}
2q24.1	1.22×10^{-07} / 1.05	*PKP4	$\mathbf{8.43 \times 10^{-08}}$	2.07×10^{-08} / 1.03	2.30×10^{-05}	3.36×10^{-05}	9.30×10^{-07}
3q13.12	$\mathbf{*1.86 \times 10^{-08}/1.05}$			9.46×10^{-07} / 1.05	9.44×10^{-06}	2.20×10^{-06}	1.51×10^{-04}
3p21.31	1.71×10^{-07} / 0.96	*CCDC36 *USP4 *CCDC71 *NICN1	$\mathbf{1.69 \times 10^{-07}}$ $\mathbf{5.54 \times 10^{-07}}$ $\mathbf{1.00 \times 10^{-06}}$ $\mathbf{1.09 \times 10^{-06}}$	$4.06 \times 10^{-08}/0.94$	1.80×10^{-05}	6.74×10^{-07}	3.51×10^{-05}
4p13	9.29×10^{-05} / 1.04			$\mathbf{*3.00 \times 10^{-10}/0.97}$	2.28×10^{-04}	4.38×10^{-04}	3.08×10^{-05}
4q24	1.51×10^{-07} / 1.05	MANBA *CXXC4	$\mathbf{1.12 \times 10^{-07}}$ $\mathbf{1.35 \times 10^{-06}}$	$1.11 \times 10^{-10}/0.87$	3.03×10^{-04}	6.48×10^{-08}	1.39×10^{-05}
5q14.3	$\mathbf{*2.98 \times 10^{-11}/0.95}$			$\mathbf{1.64 \times 10^{-9}/0.92}$	3.71×10^{-05}	$\mathbf{1.81 \times 10^{-08}}$	1.92×10^{-06}
5q21.2	$\mathbf{*3.56 \times 10^{-11}/1.06}$			$\mathbf{1.55 \times 10^{-16}/1.03}$	8.66×10^{-04}	$\mathbf{1.08 \times 10^{-07}}$	3.25×10^{-07}
7q11.22	7.17×10^{-07} / 0.96	*CALN1	$\mathbf{1.15 \times 10^{-06}}$	3.22×10^{-06} / 0.98	4.01×10^{-04}	3.96×10^{-06}	4.02×10^{-06}
10q25.1	$\mathbf{5.61 \times 10^{-09}/1.06}$	*SORCS3	$\mathbf{4.97 \times 10^{-10}}$	$\mathbf{*9.97 \times 10^{-13}/1.03}$	1.67×10^{-04}	1.76×10^{-08}	1.15×10^{-04}
13q22.3	$\mathbf{*4.03 \times 10^{-09}/0.95}$			1.16×10^{-07} / 0.97	2.46×10^{-03}	5.07×10^{-07}	1.76×10^{-05}
14q13.1	$\mathbf{*3.99 \times 10^{-08}/0.94}$			$\mathbf{5.11 \times 10^{-10}/0.94}$	3.04×10^{-04}	1.02×10^{-05}	3.08×10^{-04}
16p13.3	3.72×10^{-05} / 0.96			$\mathbf{*5.59 \times 10^{-11}/0.97}$	2.81×10^{-05}	1.06×10^{-05}	1.92×10^{-04}
17q21.31	$\mathbf{*3.22 \times 10^{-08}/0.95}$	*WNT3 *KANSL1 *CRHR1 *MAPT	$\mathbf{4.52 \times 10^{-07}}$ $\mathbf{7.98 \times 10^{-08}}$ $\mathbf{2.32 \times 10^{-07}}$ $\mathbf{8.72 \times 10^{-07}}$	$3.28 \times 10^{-06}/0.92$	6.53×10^{-05}	3.79×10^{-04}	2.89×10^{-07}
18q21.2	1.17×10^{-05} / 0.97			$\mathbf{*4.26 \times 10^{-12}}$ / 1.03	3.29×10^{-04}	1.89×10^{-05}	7.05×10^{-05}
20p11.23 - p11.24	$\mathbf{*1.71 \times 10^{-09}/1.05}$	XRN2	$\mathbf{2.33 \times 10^{-09}}$	$\mathbf{2.72 \times 10^{-10}/1.05}$	1.08×10^{-03}	1.33×10^{-06}	2.04×10^{-09}

Table 2.3: **Comparison of results across regions that are shown as genomewide significant and pleiotropic in either the TS-OCD GWAS meta-analysis (SNP-based or gene-based) or the PGC eight-disorder GWAS meta-analysis [192].** Regions highlighted in dark blue are identified in both studies. The light blue color highlights novel regions that are only identified as genomewide significant and pleiotropic in this study (ie crossing genomewide significance threshold and m-value¿0.9 across TS-OCD). No highlight points to regions were criteria are satisfied only in the PGC eight-disorder GWAS meta-analysis. (Bold font: genomewide significant, *: SNP/Gene that are geneomewide significant and pleiotropic)

Region	SNP-based results	Gene-based results		Results from [192]	Original TS GWAS	Original OCD GWAS
	Leading SNP p-value/OR	gene	p-value	Leading SNP p-value/OR	Leading SNP p-val	Leading SNP p-val
1p31.1	$1.32 \times 10^{-04}/1.34$			$*\mathbf{3.63 \times 10^{-11}/1.03}$	1.69×10^{-04}	2.26×10^{-04}
2p16.1	$*\mathbf{5.64 \times 10^{-09}/0.89}$			$\mathbf{2.34 \times 10^{-14}/0.97}$	4.76×10^{-08}	2.03×10^{-04}
3p12.1	$1.06 \times 10^{-06}/1.12$	***CADM2**	$\mathbf{3.99 \times 10^{-07}}$	$5.74 \times 10^{-05}/1.05$	$8.36x10^{-05}$	1.15×10^{-04}
4p13	$4.03 \times 10^{-04}/1.11$			$*\mathbf{5.59 \times 10^{-09}/0.96}$	2.28×10^{-04}	6.04×10^{-05}
6p21.33	$\mathbf{1.48 \times 10^{-07}/0.81}$	***LY6G6F** ***MEGT1** ***APOM**	$\mathbf{7.64 \times 10^{-07}}$ $\mathbf{7.98 \times 10^{-07}}$ $\mathbf{1.54 \times 10^{-06}}$	$*3.63 \times 10^{-14}/0.97$	1.56×10^{-05}	1.55×10^{-05}
14q32.33	$6.22 \times 10^{-04}/0.93$			$*\mathbf{5.20 \times 10^{-09}/1.03}$	6.36×10^{-04}	1.40×10^{-04}
16p13.3	$9.21 \times 10^{-05}/0.92$			$*\mathbf{5.59 \times 10^{-11}/0.97}$	2.81×10^{-05}	9.12×10^{-05}
18q21.2	$5.74 \times 10^{-04}/0.86$			$*\mathbf{4.26 \times 10^{-12}/1.03}$	3.29×10^{-04}	3.99×10^{-04}
22q13.2	$1.23 \times 10^{-05}/0.90$			$*\mathbf{5.36 \times 10^{-14}/1.04}$	6.84×10^{-06}	9.92×10^{-04}

Table 2.4: **Full results for joint genetic architecture analysis using GenomicSEM.** Including model fitness, standardized and unstandardized results. Also see Figure 2.1.

View through this link.

Table 2.5: **Full results for the causality inference using GSMR.** #snps denotes the number of SNPs used for the analysis. Also see Figure 2.2.

View through this link.

Table 2.6: **Summary statistics for all significant results from SNP-based GWAS meta-analyses across TS, ADHD, ASD and OCD.** m-value = Posterior probability for association for each individual disorder; SIFT/Poly1/Poly2 = functional prediction for nonsynonymous exonic SNPs; HetISq = heterozygosity I2 statistic; HetChiSq = heterozygosity chi-square statistic; HetP-Val = heterozygosity test p-value; disorder-OR/P = odds ratio statistic and *p*-value in the original individual disorder GWAS study.

View through this link.

Table 2.7: **Full annotation of top genomic risk regions from SNP-based GWAS meta-analyses.** An asterisk (*) indicates novel LD regions not been reported associated with corresponding traits in published GWAS. rsID = rsID of the leading SNP of the region; p = p-value of the leading SNP from the meta-analysis; Study = Previous studies reporting significant association at this locus; trait = trait reported associated with the locus by the study; reported gene = gene reported by the study; mapped gene = gene mapped onto the reported region.

View through this link.

Table 2.8: **Comparison of statistics for matching SNPs from our TS-ADHD-ASD and TS-OCD SNP-based GWAS meta-analysis results with PGC eight-disorder GWAS meta-analysis.** Table includes leading SNPs in regions with genomewide significant pleiotropic SNPs identified by the SNP-based analysis or the eight-disorder cross-disorder analysis from PGC (corresponds to selected light blue and clear rows in table 2). Red font denotes SNPs found genomewide significant and pleiotropic, and test statistics (p-value, OR and m-values for disorders analyzed) for the same SNP are reported for both studies if available.

View through this link.

Table 2.9: **Detailed results comparison between TS-ADHD-ASD, TS-OCD SNP and gene-based GWAS meta-analysis and PGC eight-disorder GWAS meta-analysis [228].** Regions satisfying at least one of the three following criteria are included: 1. hosting genomewide significant SNPs in the TS-ADHD-ASD, TS-OCD SNP-based analysis; 2. genomewide significant genes from the TS-ADHD-ASD, TS-OCD gene-based analysis; 3. hosting genomewide significant SNPs that also have $m - value > 0.9$ across the disorders of interest here (ie TS-ADHD-ASD, TS-OCD) in the 8-disorder analysis. For each region, the following are shown: region basepair position, number of genomewide significant SNPs with $m - value > 0.9$ in all disorders analyzed, leading SNP, leading SNP p-value, OR and m-values for all disorders of interest from SNP based analysis and Cross-Disorder Group of the Psychiatric Genomics Consortium et al., 2019; significant genes from the gene-based analysis, p-values and whether the gene is still significant when analyzing using only SNPs with $m - value > 0.9$ in all disorders of interest; leading SNP, p-value and OR from the original individual GWAS analyses. Asterisk (*) indicates the region is also highlighted by the TWAS analysis.

View through this link.

Table 2.10: **Partitioned heritability analysis.** Baseline results included 53 non-cell type specific annotations; brain cell types included 13 brain relevant cell type specific annotations; chromatin included results for 489 cell type specific annotation of chromatin states, as described by [179]. Asterisk (*) in the significant column denotes the annotation categories significantly enriched under multiple testing correction.

View through this link.

Table 2.11: **Significant results from conditional analyses (TS-ADHD-ASD conditioned on TS-OCD and vice versa), compared with original meta-analyses results.** Including SNPs that are genomewide significant in either the original meta-analysis results or the conditioned. b, se, $pval$ correspond to beta, standard error and p-value in the original meta-analyses results respectively; b_C, bC_{se}, bC_{pval} correspond to conditioned beta, standard error and p-value. diff = increment (+) or decrement (-) of effect (in terms of z-score) after conditioned.

View through this link.

Table 2.12: **Significant genes from gene-based GWAS analyses.** P-values from individual disorder gene-based analyses are also shown.

View through this link.

Table 2.13: **Significant results from cross-disorder tissue specificity analysis, testing 53 tissue types from GTEx v7 tissue expression atlas.** The significance threshold is set following Bonferroni correction ($p < 9.43 \times 10^{-4}$).

View through this link.

Table 2.14: **Transcriptome-wide analysis.** Significant results from transcriptome-wide analysis, using SMR.

CHR	Gene	TS-ADHD-ASD			
		Beta	SE	p_{SMR}	P_{HEIDI}
17	*LRRC37A4P*	-0.0353	0.0073	1.38×10^{-06}	9.57×10^{-02}
17	*RP11-707O23.5*	0.0335	0.0069	1.26×10^{-06}	6.93×10^{-02}

2.8 Figures

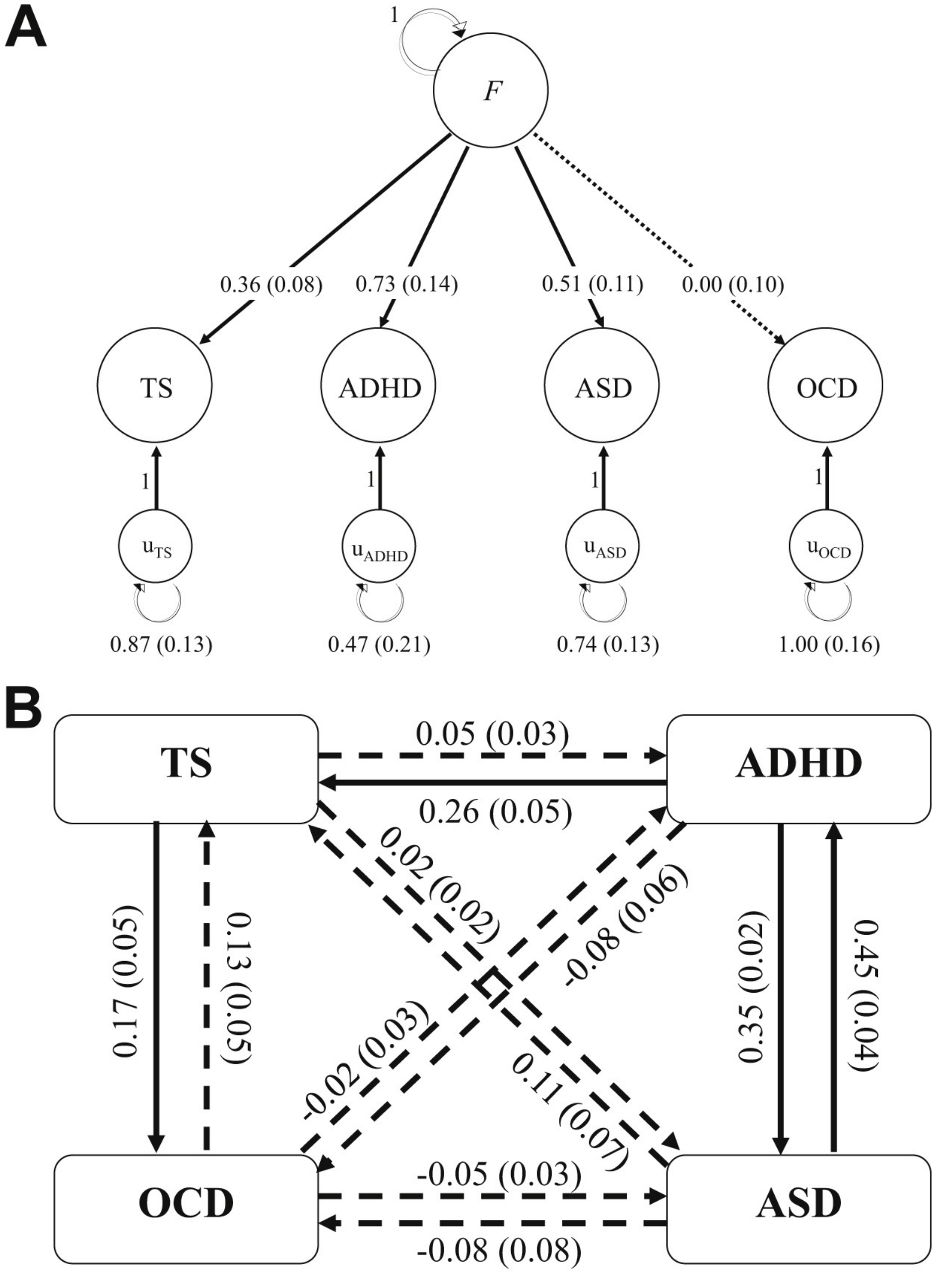


Figure 2.1: **Genetic architecture and causality relationships across disorders of interest.** A. Investigating the existence of a common factor F across all four disorders using Genomic SEM. Path graph shows loads and corresponding standard errors in parenthesis. Circular arrows denote the residual genetic variance not explained by the common factor. Also see table S1. B. Network plot indicating the causality across four disorders estimated using GSMR. Solid arrows indicate a significant causality relationship while dash arrows indicate insignificant relationships. Numbers on the arrow indicate effect size and estimation standard error (in parenthesis). Also see table S2.

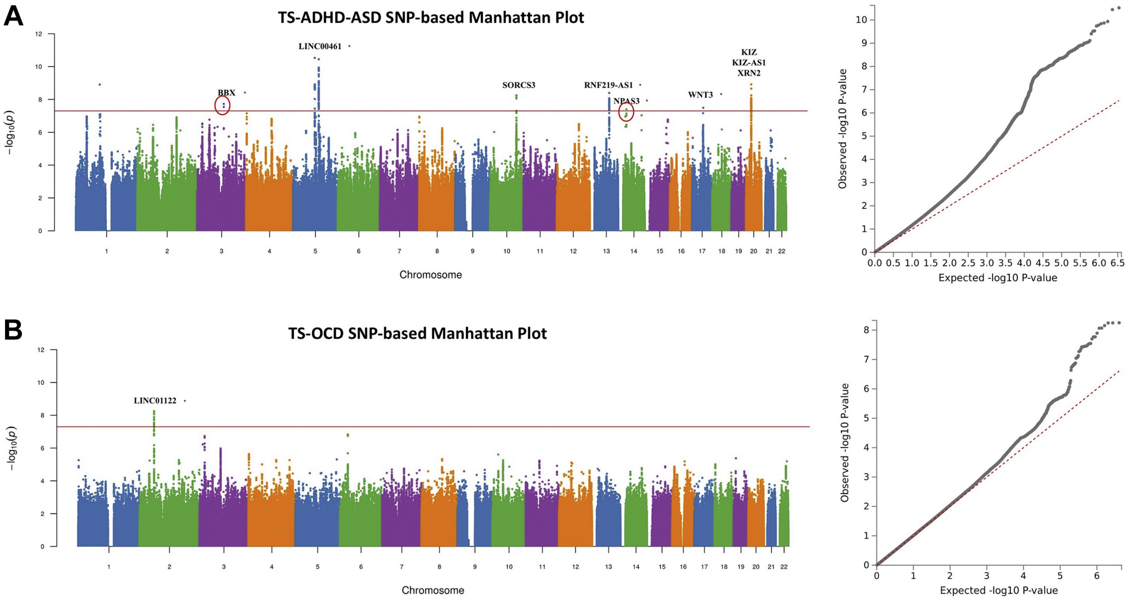


Figure 2.2: **Manhattan plots and QQ plots for cross-disorder GWAS meta-analyses.** An asterisk (*) indicates genes hosting SNPs with $m - value > 0.9$ in all disorders analyzed, and a red circle denotes novel region that was not previously reported associated with disorder of interest. A. TS-ADHD-ASD GWAS meta-analysis; B. TS-OCD GWAS meta-analysis. See also Table S3, S4.

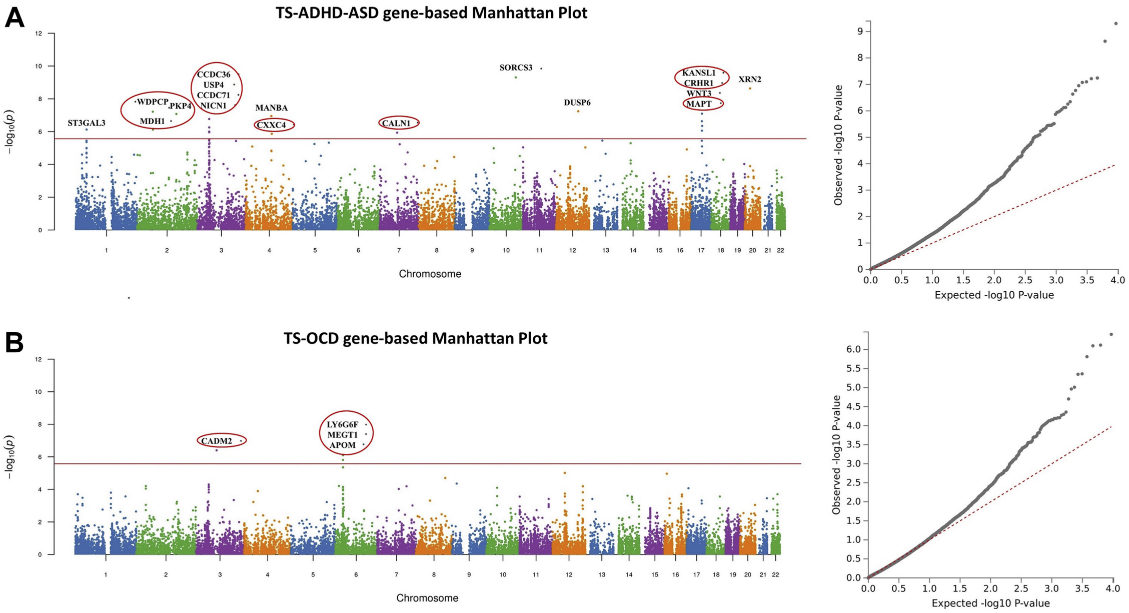


Figure 2.3: **Manhattan plots for gene-based GWAS meta-analyses.** An asterisk (*) indicates genes stay significant when only analyzing SNPs with m-value ¿ 0.9 in all disorders analyzed, and red circle denotes novel genes that could not be picked up though gene-based analysis using summary statistics from individual disorders alone. A. TS-ADHD-ASD gene-based analysis; B. TS-OCD gene-based analysis. See also Table S9.

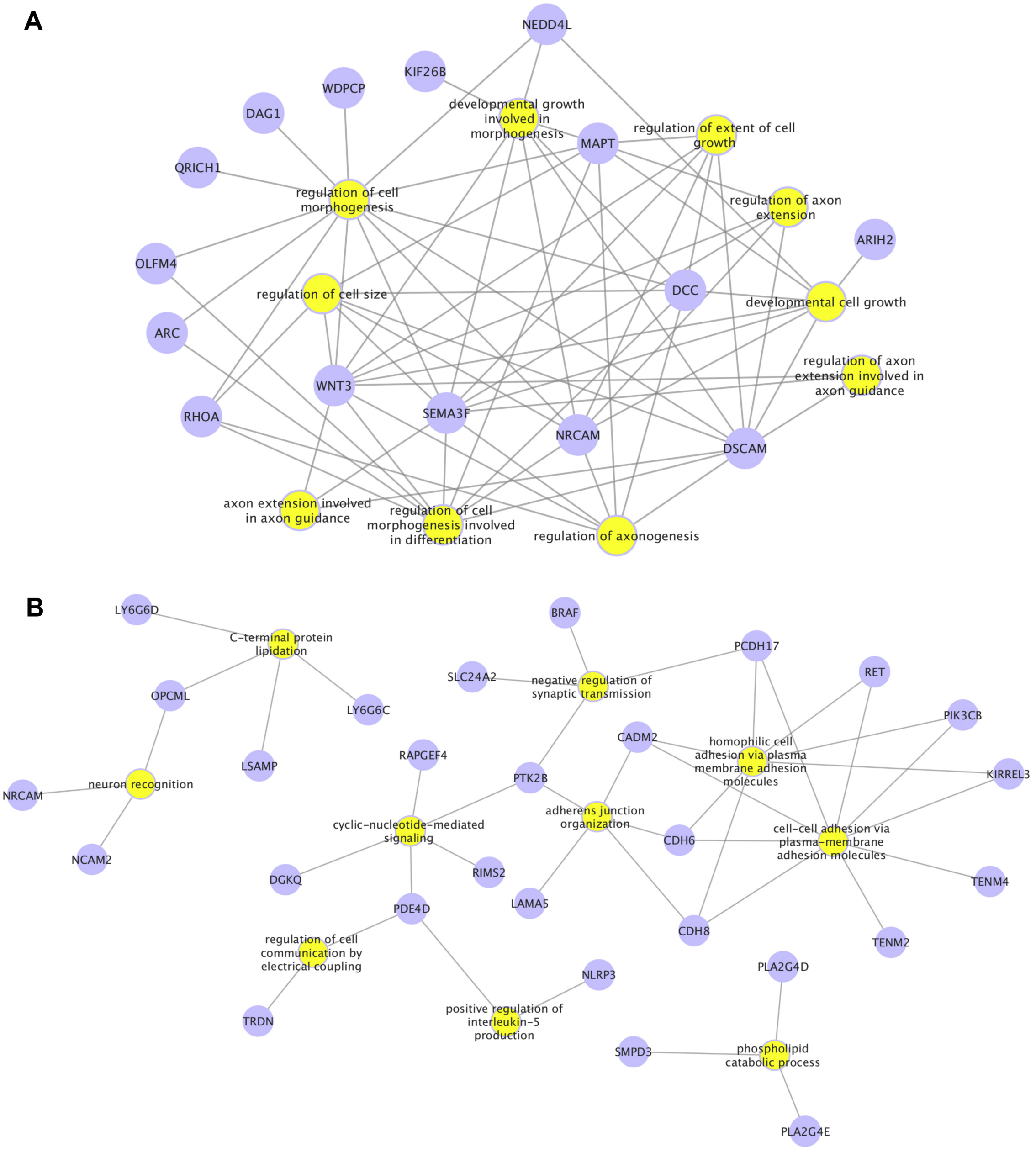

A
NEDD4L
KIF26B
WDPCP
DAG1
developmental growth involved in morphogenesis
regulation of extent of cell growth
MAPT
QRICH1
regulation of cell morphogenesis
regulation of axon extension
ARIH2
OLFM4
regulation of cell size
DCC
developmental cell growth
ARC
regulation of axon extension involved in axon guidance
WNT3
RHOA
SEMA3F
NRCAM
DSCAM
axon extension involved in axon guidance
regulation of cell morphogenesis involved in differentiation
regulation of axonogenesis
B
LY6G6D
C−terminal protein lipidation
BRAF
PCDH17
SLC24A2
negative regulation of synaptic transmission
RET
OPCML
LY6G6C
PIK3CB
CADM2
homophilic cell adhesion via plasma membrane adhesion molecules
RAPGEF4
LSAMP
KIRREL3
PTK2B
NRCAM
neuron recognition
adherens junction organization
cyclic−nucleotide−mediated signaling
CDH6
cell−cell adhesion via plasma−membrane adhesion molecules
NCAM2
RIMS2
DGKQ
TENM4
LAMA5
PDE4D
CDH8
TENM2
regulation of cell communication by electrical coupling
PLA2G4D
NLRP3
TRDN
positive regulation of interleukin−5 production
SMPD3
phospholipid catabolic process
PLA2G4E

Figure 2.4: **Top ten gene networks from top 200 genes annotated from SNP-based GWAS meta-analyses results.** A. TS-ADHD-ASD SNP-based network plot; B. TS-OCD SNP-based network plot.

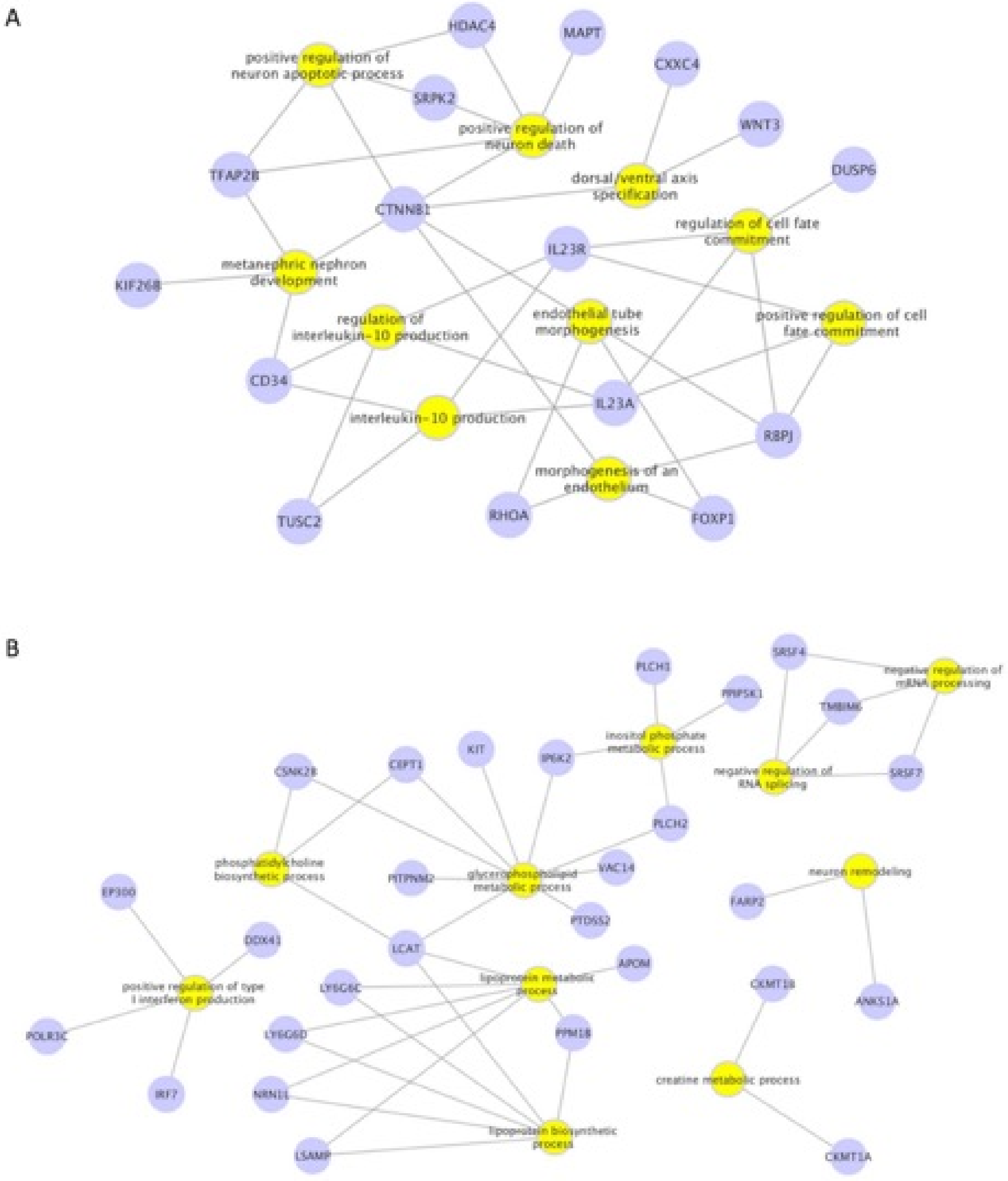


Figure 2.5: **Gene networks plot (gene-based).** Top ten gene networks from top 200 genes from gene-based analysis results. A. TS-ADHD-ASD gene-based network plot; B. TS-OCD gene-based network plot.

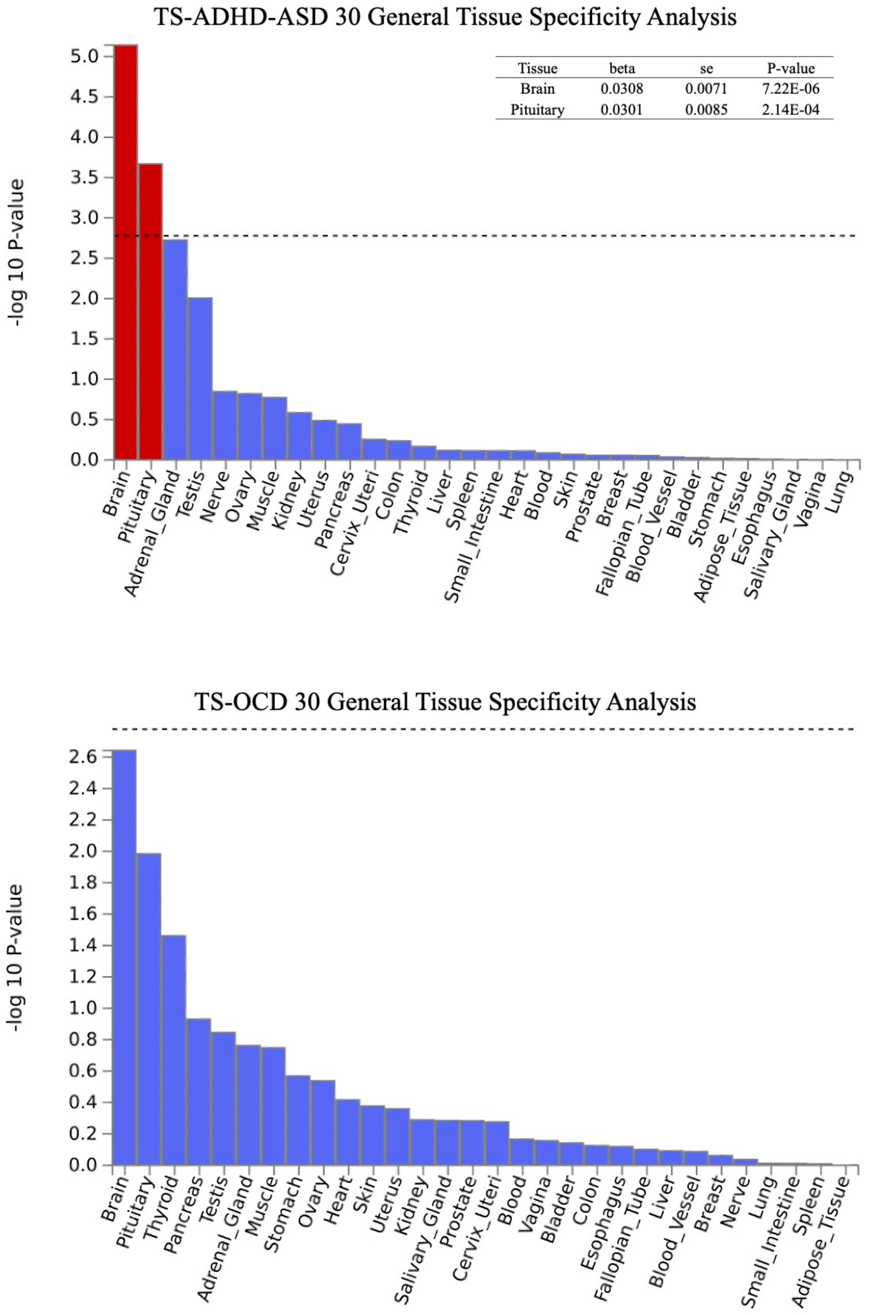
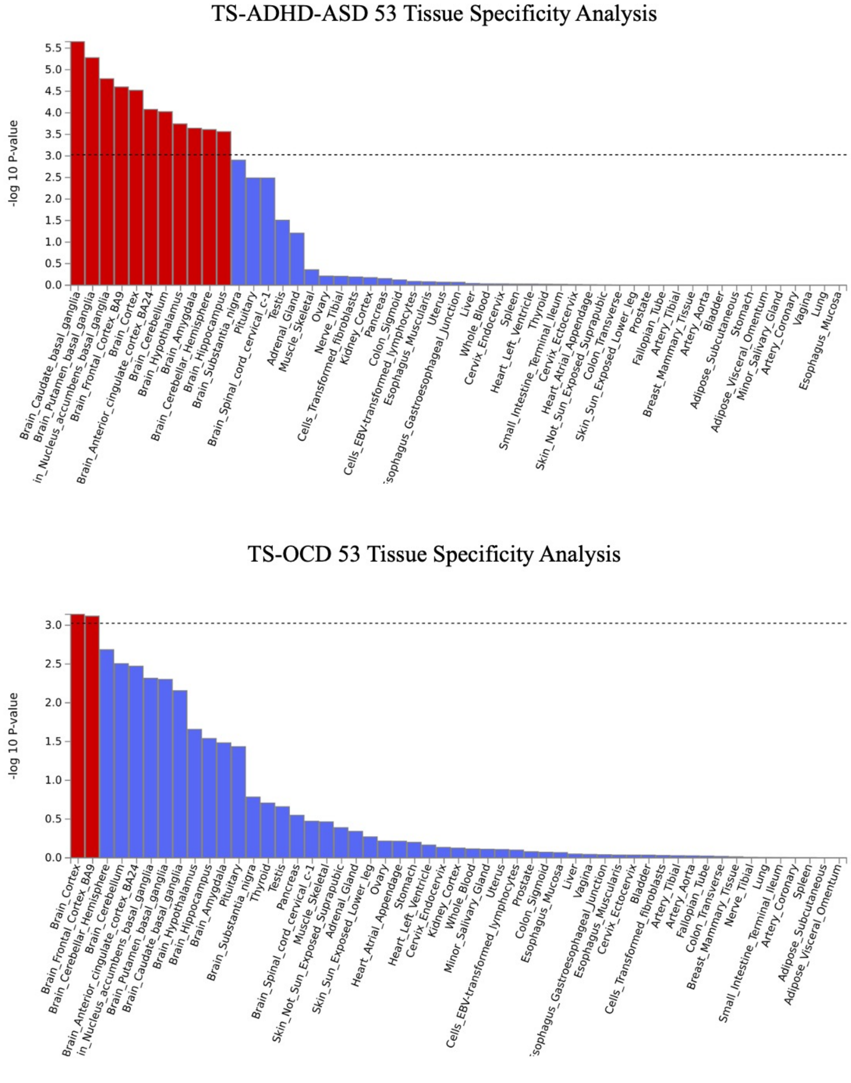


Figure 2.6: **Tissue specificity analyses (30 tissue types).** Cross-disorder tissue specificity analysis testing 30 general tissue types from GTEx v7 tissue expression atlas. Red bar indicates significant enrichment of gene expression in corresponding tissue under Bonferroni correction ($p < 1.67 \times 10^{-3}$). Panel on top right corner of each figure shows detailed statistics for significantly enriched tissue. A. TS-ADHD-ASD cross-disorder tissue specific expression enrichment; B. TS-OCD cross-disorder tissue specific expression enrichment. See also Table S9 and Figure S3


TS-ADHD-ASD 53 Tissue Specificity Analysis

TS-OCD 53 Tissue Specificity Analysis

Figure 2.7: **Tissue specificity analysis (53 tissue types).** Tissue specificity analysis, testing 53 tissue types from GTEx v7 tissue expression atlas. Red bar indicates significant enrichment of gene expression in corresponding tissue under Bonferroni correction (p ¡ 9.43 x 10-4). Panel on top right corner of each figure shows detailed statistics for significantly enriched tissue. A. TS-ADHD-ASD cross-disorder tissue specific expression enrichment; B. TS-OCD cross-disorder tissue specific expression enrichment.

3 RECONSTRUCTING SNP ALLELE AND GENOTYPE FREQUENCIES FROM GWAS SUMMARY STATISTICS

3.1 Abstract

Motivation: The emergence of genomewide association studies (GWAS) has led to the creation of large repositories of human genetic variation, creating enormous opportunities for genetic research and worldwide collaboration. Methods that are based on GWAS summary statistics seek to leverage such records, overcoming barriers that often exist in individual-level data access while also offering significant computational savings.

Results: We propose a novel framework to reconstruct allelic and genotypic frequencies and counts for each SNP from case-control GWAS summary statistics and show how it can broaden the scope of summary statistics based method development. Our framework is simple and efficient with minimal underlying assumptions and can be used to unify common tasks related to GWAS. To this end, we propose here three summary-statistics-based applications implemented in a new software package (ReACt): GWAS meta-analysis (with and without sample overlap), case-case GWAS, and, for the first time, group polygenic risk score (PRS) estimation. We evaluate our methods against the current state-of-the-art on both synthetic data and real genotype data and show high performance in power and error control. Our novel group PRS method based on summary statistics could not be achieved prior to our proposed framework, and we demonstrate here the potential applications and advantages of this approach. Our work further highlights the great potential of summary-statistics-based methodologies towards elucidating the genetic background of complex disease and opens up new avenues for research.

Availability and Implementation: An implementation for ReACt can be found on our github page: `https://github.com/Paschou-Lab/ReAct`.

3.2 Introduction

Genomewide association studies (GWAS) have emerged as a powerful tool, leading to the identification of thousands of common genetic variants that underlie human complex disorders and traits.

They also led to the creation of large repositories of human genetic variation creating enormous opportunities for further analysis. However, sharing and transferring of individual-level genotype data is often restricted due to privacy concerns as well as logistical issues. On the other hand, GWAS summary statistics, typically including information such as odds ratio (OR)/effect size (beta), standard error (SE), p-values, and case/control sample sizes for each SNP being analyzed, are often readily accessible [290]. The availability of such alternative sources of information has spurred intense interest into the development of methodologies seeking to leverage such records effectively in order to retrieve as much information as possible. Besides overcoming barriers in individual-level data access, summary-statistics-based methods also offer advantages in computational costs, which do not scale as a function of the number of individuals in the study [291].

Summary statistics methodologies have been developed to allow a wide array of statistical analyses, including effect size distribution estimation [292, 293]; GWAS meta-analysis and fine mapping [294, 295, 228, 165, 296]; allele frequency and association statistic imputation [297, 298]; heritability and genetic correlation estimation [299, 300, 301, 179]; case-case GWAS [194]; and polygenic prediction [302, 303, 123]. Many of these methods have to incorporate additional information from publicly available sources, such as linkage disequilibrium (LD) statistics from a reference population [299, 297, 183]. Most of the existing methodologies analyzing GWAS summary statistics use the summary statistics (OR, SE, p-value) from the input “as is”, often via relatively complicated estimation and modeling.

In our work, we propose a novel framework that leverages the simple observation that summary statistics information can be expressed as a functions of case/control allele frequencies for each SNP. This allows us to recover case/control allele frequencies from summary statistics by solving a non-linear system of equations. Additionally, if one assumes that the SNPs satisfy Hardy-Weinberg Equilibrium (HWE), the allele frequencies can be used to infer genotype counts. This simple observation allows us to use information from case-control GWAS summary statistics to develop a simple, user-friendly alternative to summary-statistics-based methods for fixed effect meta-analysis and cc-GWAS.

Furthermore, using our framework, we are able to compute group-wise polygenic risk score (PRS) from summary statistics of both a base and a target population. While there have been summary statistics based methods estimating the variance explained by SNPs using results from existing PRS associations [304, 305], to the best of our knowledge, no existing method could return reliable estimates of PRS without any access to individual-level data in the validation cohort prior to our work.

In the remainder of the paper, We describe the mathematical foundations of our framework and its application to fixed effect meta-analysis, cc-GWAS, and group-wise PRS estimation. We demonstrate the performance of the proposed methods using simulated and real data and we compare our approach against current state-of-the-art. Our methods are implemented in a new software package: Reconstructing Allelic Count (ReACt).

3.3 Results

3.3.1 Mathematical foundations

Our framework is motivated by the fact that the summary test statistics from publicly available GWAS can be expressed as a function of allele counts of the effect and the non-effective allele in cases and controls; as a result, the allele counts can be exactly recovered by solving a system of non-linear equations. Interestingly, this rather straight-forward observation has not received much attention in prior work. Additionally, assuming that SNPs included in GWAS studies are in Hardy-Weinberg Equilibrium (HWE), we can also reconstruct the structure of the genotype vectors for publicly available GWAS studies from just summary statistics. We can leverage this information in multiple applications, including: *(i)* the computation of the joint effect of a SNP in a meta-analysis involving multiple studies; *(ii)* to obtain the mean polygenic risk score of cases and controls in a population; and *(iii)* to investigate the genetic differences between traits using a case-case GWAS. All of these can be done using only summary statistics, which circumvents the hassle of individual level data sharing and, as an added bonus, considerably reduces the necessary computational time. We start by introducing some notation that will be useful in this section. Let a and u represent effective and non-effective allele counts respectively; let superscripts $^{\text{cse}}$ and $^{\text{cnt}}$ represent cases and controls respectively; and let OR, SE, and N be the odds ratio, standard error (of $log(OR)$, as presented in most of the GWAS summary statistics), and sample sizes obtained from the summary statistics. Thus, for SNP i, u_i^{cnt} represents the count of the non-effective allele in controls for SNP i; similarly, a_i^{cse} represents the count of the effective allele in cases for SNP i; N^{cse} represents the number of cases, etc. We now note that the allelic effect of SNP i in case-control GWAS summary statistics can be expressed as follows:

$$\begin{aligned} OR_i &= \frac{a_i^{\text{cse}} \cdot u_i^{\text{cnt}}}{a_i^{\text{cnt}} \cdot u_i^{\text{cse}}}, \\ SE_i &= \sqrt{\frac{1}{a_i^{\text{cse}}} + \frac{1}{u_i^{\text{cse}}} + \frac{1}{a_i^{\text{cnt}}} + \frac{1}{u_i^{\text{cnt}}}}. \end{aligned}$$

Additionally, sample sizes can be expressed as:

$$\begin{aligned} 2N^{\text{cse}} &= a_i^{\text{cse}} + u_i^{\text{cse}}, \quad \text{and} \\ 2N^{\text{cnt}} &= a_i^{\text{cnt}} + u_i^{\text{cnt}}. \end{aligned}$$

Therefore, solving the system of the above four non-linear equations allows us to recover the allelic counts of SNP i for effective and non-effective alleles in cases and controls, by solving for the four unknowns a_i^{cse}, a_i^{cnt}, u_i^{cse}, and u_i^{cnt}. Using these counts, we can trivially obtain allele frequencies in case and control groups and, importantly, by further assuming that the SNPs strictly follow HWE, we can even compute the genotypic counts for each genotype from these frequencies. Note that this reverse engineering scheme applies to GWAS summary statistics generated using a χ^2 test or logistic regression, but it does not apply to GWAS summary statistics generated by other methodologies. Furthermore, these frequencies will be different from those observed from individual level data due to model covariates; the recovered frequencies correspond to the allele counts after corrections have been applied. See Section 3.5.1 and 3.7.1 for details.

3.3.2 Fixed effect meta-analysis

Our approach Armed with allelic and genotypic counts, we can provide a new perspective on fixed-effect GWAS meta-analysis. Instead of the conventional inverse-variance weighted meta-analysis, we can now compute the joint effect of a SNP in a meta-analysis using multiple studies by combining the reconstructed allele and genotype counts from each study and run a *complete* logistic regression on each SNP. Thus, we can essentially proceed with the analysis in exactly the same way as standard GWAS (see Section 3.5.2 for details).

As mentioned in Section 3.3.1, we can obtain genotypic counts for any SNP over cases and controls from GWAS summary statistics. Then, combining these counts for all available input studies, along with the trait status, we can carry out a logistic regression for this SNP as follows [306]:

$$\mathsf{Pr}(\mathbf{y}_j = 1|\mathbf{g}_j, \mathbf{s}_j) = S(\beta_0 + \beta_1\mathbf{g}_j + \beta_2\mathbf{s}_j).$$

In the above $\mathbf{y}_j$ denotes the binary trait for the jth individual, $\mathbf{g}_j$ denotes the respective genotype, and $S(\cdot)$ stands for the standard sigmoid function used in logistic regression. Solving for the coefficients β_0, β_1, and β_2 we get the overall SNP effect from the meta-analysis. In order to take into account between-study stratification, we introduce an additional variable $\mathbf{s}_j$ as a covariate, using the overall allele frequencies of each study to estimate it. (See Section 3.5.2 for details.)

Performance evaluation First, we tested the performance of the proposed fixed-effect meta-analysis approach on synthetic data under various conditions. The simulation was carried out using the Balding-Nichols model [307], assuming a minor allele frequency of 0.3. For each setting, we predefined the risk for effective alleles of the causal SNPs by setting $r = 1.15/1.2/1.3$ as well as the level of population stratification between cohorts included in the meta-analysis setting $F_{st} = 0.01/0.05/0.1$. Apart from meta-analyzing mutually exclusive datasets, we also tested the performance of our approach under different extents of sample overlap between the input studies: When generating input summary statistics, we evaluated scenarios where the input studies shared $N_{\texttt{shr}}$ cases and $N_{\texttt{shr}}$ controls, with the value of $N_{\texttt{shr}}$ set to zero, 100, and 500 (see Section 3.6.1 for details). Ideally, the overlapping sample sizes are expected to be input as a parameter by the user (ReACt(Exact) in figures 3.1, 3.2). However, such information is not always available. In those cases, we adopted the approach of estimating unknown sample overlap via Z-scores in input GWAS summary statistics from [308] (ReACt(Est.) in figures 3.1, 3.2). We compared power and type I error rates of our approach vs. state-of-the-art tools that are currently widely used for fixed-effect meta-analysis, namely METAL [309] and ASSET [310]. Since the latest stable release of METAL does not include an implementation for sample overlap correction, we used the GitHub version of METAL from [308]. The performance comparison on the meta-analysis of two studies with even case/control sample sizes is plotted in Figures 3.1, 3.2 and Table 3.15. Performance on meta-analyzing two studies with uneven sample sizes (Table 3.14 in supplementary text) as well as meta-analyzing multiple studies (Table 3.16 in supplementary text) are also tested. Results on synthetic data indicated that our approach has comparable performance with the conventional inverse-variance weighted methods ASSET and METAL, namely

$$\left|\text{Power}_{\textbf{ReACt}} - \text{Power}_{\textbf{ASSET/METAL}}\right| \leq 0.012,$$

when there is no sample overlap. In scenarios where there were samples shared across input studies, our method (regardless of whether the exact size of the sample overlap is known or is estimated) always showed higher power compared to ASSET, namely

$$0.014 \leq \text{Power}_{\textbf{ReACt}} - \text{Power}_{\textbf{ASSET}} \leq 0.219$$

and comparable power to METAL, namely

$$|\text{Power}_{\textbf{ReACt}} - \text{Power}_{\textbf{METAL}}| \leq 0.005.$$

Our advantage in power compared to ASSET was more visible under higher F_{st} values and larger sample overlaps. In terms of type I error rates, we observed that all methods showed good control on the error rates, while ASSET tended to produce more conservative results. Similar observations can also be made when we meta-analyzed multiple studies; see Table 3.16 for details.

Beyond power and type I error, we also analyzed the running time of the different methods (see Table 3.13 in supplementary text). Our C implementation of our method in the ReACt software package has not been highly optimized and yet it has a running time that is comparable to METAL and is much faster than ASSET. We further tested the performance of our method on real genotype data using the UK biobank dataset [135] and analyzing for depressive episode trait. The dataset included a total of 18,368 cases, 312,849 controls, with 640,756 SNPs after quality control (see Section 3.6.1 for details). In this experiment, we treated the top 7 SNPs with p-value stricly less than 10^{-6} from the overall GWAS as "ground truth" and assessed whether various meta-analysis method could pick up these 7 SNPs. Each experiment was carried out over ten iterations: in each iteration, we split the dataset in two equal sized subsets, generated GWAS summary statistics from each of the subsets, and meta-analyzed the resulting summary statistics. We reported average true positive and false positive SNPs counts captured by each method over the ten iterations. Table 3.7 reports our findings and we note that, perhaps due to the lack of stratification, the differences in performance were not as visible as what we observed using synthetic data. We found ReACt(Exact) showing comparable performance with ASSET, whereas ReACt(Est) showed comparable performance with METAL.

3.3.3 Group PRS

Our approach Even though we still cannot compute individual level PRS without access to raw genotypes, we observe that, under the additive model, the mean and standard deviation of PRS for a population are just functions of SNP allele frequencies in the target group (see Section 3.5.3 for details). Therefore, our proposed framework, which returns estimates of allele frequencies for cases and controls using GWAS summary statistics, also allows us to estimate means and standard deviations of PRS for case and control groups using the GWAS summary statistics of the target study. With such information (and a fair assumption of normality in the underlying PRS distribution), we can further run a t-test in order to get a p-value comparing the difference of PRS between cases and controls.

More specifically, in the additive model, the mean and variance of PRS for a population can be expressed as follows:

$$\begin{aligned}\text{mean(PRS)} &= \frac{\sum_{i=1}^{M} S_i p_i}{M}, \text{ and} \\ \text{Var(PRS)} &= \frac{\sum_{i=1}^{M} S_i^2 p_i q_i}{2M^2}.\end{aligned}$$

In the above S_i is the weight of SNP i inferred from the base summary statistics (typically $S_i = \frac{\log(OR_i)}{SE_i}$), M is the total number of SNPs used in the PRS computation, and p_i and $q_i = 1 - p_i$ are allele frequencies of the effective allele and the non-effective allele for SNP i. Therefore, we can simply use the allele frequencies of cases and controls that were computed in Section 3.3.1 in order to get the mean and variance of PRS in cases and controls. See Section 3.5.3 for details.

Performance evaluation We first tested our methods on synthetic data without any confounding factors (ie., no stratification). After generating GWAS summary statistics for synthetic base and target datasets, we compared the estimated group means and standard deviations using our method (which operates on summary statistics) with the real group means and standard deviations of PRS computed from the individual level genotypes using PRSice2 [311]. The results successfully proved that in this scenario our method is extremely accurate. See Table 3.8 which shows typical representative results from our experimental evaluations; essentially identical results were observed in all our experiments on synthetic data.

We further tested our method on real GWAS data, using GWAS summary statistics for MDD [312] as the base study and assessing its predicting power on 18,368 *independent* depressive episode cases and 312,849 ancestry-matched controls in UK biobank. We did not choose the latest MDD GWAS to be a base study because the latest one has included samples from UK biobank. To run ReACt, we generated GWAS summary statistics for the target dataset as described. We compared the estimated PRS statistics using our methods with the real PRS statistics computed using PRSice2. The results are shown in Table 3.9; note that since real GWAS datasets are subject to within study population stratification, we did not expect our method to be as accurate as it was on synthetic data without such stratification. There was, however, very high concordance between the results returned by our methods and ground truth.

Finally, we applied our methods on summary statistics of eight psychiatric disorders. We evaluated their pairwise PRS predictive power by estimating t-test p-values. For this experiment, we took into account potential sample overlap between all pairs of base and target studies; see Section 3.7.2 for details of our sample overlap correction procedure. Results are shown in Table 3.10 and we observe that, in general, our results coincide with pairwise genetic correlation between disorders as discussed in [228].

3.3.4 cc-GWAS

Our approach Similar to our proposed approach for meta-analysis of multiple GWAS datasets using summary statistics, we can also carry out cc-GWAS using regression by simply swapping the labels of the phenotypes. Perhaps the biggest challenge in cc-GWAS is the separation of the differential genetic effects from between-study stratification. To circumvent this issue, we leverage the difference of SNP effects in control groups to estimate the extent of stratification (see Section 3.5.3 for details). Therefore, with a slight modification of the pipeline for meta-analysis of Section 3.5.2, we introduce an alternate approach for cc-GWAS using our framework.

The underlying theory is quite straightforward and allows us to estimate the genetic differences between two traits of interest using their GWAS summary statistics. Using the genotypic counts we can proceed with logistic regression using only the cases from the two studies:

$$\mathsf{Pr}(\mathbf{y}_j^{\mathsf{cse}} = 1|\mathbf{g}_j^{\mathsf{cse}}) = S(\beta_0^{\mathsf{cse}} + \beta_1^{\mathsf{cse}}\mathbf{g}_j^{\mathsf{cse}}).$$

In the above, $\mathbf{y}_j^{\mathsf{cse}}$ is the binary indicator variable denoting which trait case j carries and $\mathbf{g}_j^{\mathsf{cse}}$ is the genotype of this case. We note that in an additive model, the coefficient β_1^{cse} that is part of the output of this regression is a combination of both genetic effects and stratification:

$$\beta_1^{\mathsf{cse}} = \beta_g + \beta_s,$$

where β_g and β_s are the genetic effect and stratification coefficients. We are only interested in the genetic effect β_g and therefore we need to remove β_s. Towards that end, we estimate β_s using the control samples from the input studies; see Section 3.5.3 for details.

Performance evaluation We first tested the performance of our methods on synthetic data. Simulated data were again generated under the Balding-Nichols model, with predefined risks for effective allele of the causal SNPs and the extent of the stratification. Inspired by Peyrot *et al.* [194] we simulated three types of SNPs: *(i)* trait differential SNPs *(ii)* null SNPs; and *(iii)* stress SNPs (see Section 3.6.1 for details). We expect our method to pick up type (i) SNPs and leave the other two. Therefore, in our performance evaluation, we report the power for detecting the type (i) SNPs and type I error rates for picking up type (ii) and (iii) SNPs. Moreover, since we also expect the performance of our method, especially in terms of error control, to vary with sample size, the evaluation was done under different sample sizes in each input study (2,000 cases and 2,000 controls as well as 5,000 cases and 5,000 controls). Power and type I error rates for each type of SNP from the simulation model under each setting are shown in Table 3.11. The method's performance was evaluated for p-values strictly less than $5 \cdot 10^{-5}$. For this threshold, our method showed high power and well-controlled type I errors, especially under for lower values of F_{st}. On the other hand, as expected, as stratification increases between two input studies, the power of our method drop and the type I error rates increased for null SNPs. However, as a general trend, we also see a decrease in such error rates when we increase the control sample size. Meanwhile, slightly higher type I error rates for the stress SNPs are observed.

Next, we evaluated the performance of our method on real GWAS summary statistics and compared our method with the recently released method of [194]. We analyzed BIP [313] and SCZ [314] datasets, for which case-case GWAS with individual level data was available [236]. We filtered out SNPs that showed untrustworthy estimates of the stratification effect ($\mathrm{SE}_s > 0.05$, see Section 3.5.3 for details). This reduced our output size from 8,983,436 SNPs being analyzed to 7,110,776 SNPs. Out of those, our analysis revealed a total of 18 genome-wide significant risk loci, including the two regions identified by [236], namely regions 1q25.1 and 20q13.12). We compared our statistics for SNPs that were also analyzed in [194] and results for this comparison are shown in Table 3.12. The two cc-GWAS methods are mostly comparable. By definition, both we and Peyrot *et al.* [194] only used summary statistics as input, and could not apply the individual level quality control steps of [236]. As a result, both methods identified additional significant loci showing divergent genetic effects between BD and SCZ compared to [236], mainly due to a much larger effective sample size. Results for all genomewide significant risk loci are shown in Table 3.18.

3.4 Discussion

Extracting as much information as possible from easily accessible GWAS summary statistics can help accelerate research that aims to elucidate the genetic background of complex disease, allowing fast sharing of results and datasets while alleviating privacy concerns. Here, we present a simple novel framework to convert SNP statistics from any case-control GWAS back into allelic counts. When summary statistics are generated through simple χ^2 tests, without any correction for model covariates (e.g., stratification) the counts will be exact. In practice however, this backward reconstruction framework returns "pseudocounts" that correspond to corrected SNP effects after, for example, stratification correction. Therefore, results will not be subject to within-study stratification effects, assuming that the input summary statistics have been generated after quality controls. Moreover, given this is just a rewrite of the GWAS results and no real allele frequencies will be recovered, our framework does not affect the differential privacy established by sharing GWAS summary statistics [315]. Our framework simplifies and unifies GWAS meta-analyses, group PRS estimates, and cc-GWAS, both theoretically and experimentally. We implemented the aforementioned three applications in a readily available software package called ReACt.

As one of the most intuitive applications of our framework, we noticed that reconstructing the allelic counts for each SNP allows us to run a full logistic regression model instead of doing the conventional inverse-variance weighted fixed-effect meta-analysis, under the assumption of HWE. As a standard quality control step for GWAS, SNPs severely deviating from HWE should have be filtered out from the input summary statistics. Our approach shows increased power in experiments on synthetic data, especially in cases where there is larger F_{st} difference between the input studies, and provides robust results in real GWAS settings. It should be noted that corrections of sample overlap can now be done in a theoretically more straightforward manner compared to previous methods [250]. Our methods are always comparable (both in accuracy and speed) to the top performing conventional method demonstrating how the reconstructed allelic frequencies can be utilized in various types of analyses, and help reconsider some of the existing problems unifying multiple different applications under a single framework (GWAS meta-analysis, cc-GWAS and group PRS possibilities are presented here).

We further propose a novel perspective on case-case association studies (cc-GWAS), allowing analysis without the need for complicated assumptions or side information apart from sample sizes. To the best of our knowledge, the only publication on summary statistics based case-case GWAS was recently contributed by Peyrot *et al* [194], who also touched the idea of comparing allele frequencies. In our work, we achieve this objective in a straightforward manner: we directly compare the reconstructed allele frequencies of each SNP in two groups of cases, without the requirement to estimate heritabilities or prevalence of disorders as does the method of [194], which are not always possible to obtain for all GWAS results, especially when only a subset of the results are made available. Further, we do not need any extra assumptions on the distribution of SNP effects. ReACt analyzes each SNP independently and, as a result, the analysis is not affected by LD structure or number of causal SNPs underlying each disorder. We evaluated our method on a very general simulation model, which accounts for more universal scenarios and is not subjective to disease prevalence or heritability. The robustness of our approach is demonstrated by its performance on synthetic data in various scenarios. Similar to the existing cc-GWAS analysis tools [194], `ReACt` showed good control of type I errors in null SNPs (type II SNPs) given sufficiently large control sample sizes for both input studies. It also shows slightly higher, but under-controlled, type I errors in the stress test SNPs (type III SNPs). As also pointed out by [194], we do not expect the existence of stress SNPs to be particularly common in practice. We further note that all our experiments on synthetic data were carried out under different levels of population stratification. As expected, our results indicate that the performance of case-case GWAS can be greatly affected by the extent of stratification between the two input studies. We tested the performance of our method for $F_{st} = 0.1$, which is a very high-end estimate of genetic variation across homo sapiens [316]. Even so, our method still showed good power and type I error rates. For higher confidence in results, we suggest larger sample sizes for both cases and controls, especially when there is higher heterogeneity between the population groups of the two studies. A notable difference between our method and the work of [194] is that we do not filter for SNPs showing association due to differential tagging effects. While analyzing such SNPs, our method behaves more like a direct case-case GWAS using individual level data. We note that, just like the other recently described cc-GWAS method, we expect the user to input sample overlaps for cases and controls respectively for more accurate results. In our implementation, we do provide a complementary estimation using Z-scores as proposed by [308]. However, as in reality one can usually expect more controls overlapping than cases, this estimation can lead to a loss in power. Overall, our work can be considered an alternative to [194], offering novel theory and a simpler implementation.

Our framework also introduces a novel perspective on case-control PRS. Conventionally, PRS for a target study is only accessible from individual level genotype data. However, even though getting scores for each individual is not feasible, we notice that if we only focus on the differentiation between cases and controls, the group means and standard errors of PRS can in fact be estimated using only summary statistics of both the base and target studies. With such statistics available, a t-test can be carried out instead place of logistic regression, which is commonly used for predictability evaluation when the individual level PRS are available. It is worth noting that, for case-control studies, t-tests and logistic regression are testing the same hypothesis: whether scores generated from the SNP effect of a base study can differentiate individuals in the target study, or, equivalently, whether the base study can predict the case/control status of samples in the target study. We applied our method to summary statistics of eight psychiatric disorders from PGC for predicting group PRS and found the results in general concordance with the genetic correlation obtained by the work of Lee *et al* [228]. In our work, all evaluations of PRS are based on the simplest p-value based clumping and thresholding (PC+T) approach, which may make the result appear like nothing more than viewing the genetic correlation from another perspective. However, note the methodology underlying the `ReACt` group PRS can be easily adapted to any other PRS computation model, e.g., [317] (SBLUP [215], LDpred [318] , PRS-CS [202], SBayesR [203] etc). To date, most PRS improvements target the selection and prioritization of SNPs or the adjusting of the weights to build a better prediction model using the base study. Our work contributes from a different direction: it allows the user to evaluate the performance of all those models without access to individual level geontype data. Therefore, in order to apply our methodology to different PRS models, one just needs to update the input summary statistics for the base study with the pre-selected SNPs and the updated SNP weights. Moreover, results from group PRS using our approach can be further connected with [305] to quantify the predisposition to a particular disorder that is explained by a certain SNP set.

As discussed earlier in our work, our framework is robust against within-study stratification effects, which means that the group means and standard errors returned are corrected for stratification and can be used directly for within-study comparisons. However, we would like to note that the method is still vulnerable to the common weaknesses of conventional PRS, including differences in population structure between the base and target studies [319]. Users should also keep in mind that general rules of thumb for conventional PRS also apply to our method. For instance, the SNPs used for PRS computations are expected to be independent to a certain extent (clump/prune/LASSO shrink the summary statistics) [123] and as can be observed from the experiments on real data, the predicting power of output PRS will be subject to the power of the base study [304] and the p-value threshold chosen by the user. Practices that are not recommended when running conventional PRS (e.g., using results from a GWAS with really small sample size as the base study [304]) are also not recommended in our setting.

We would like to note future research directions that could further extend our methods. First, the reconstruction scheme that our framework is built upon is based on input summary statistics that are generated using a logistic regression or a χ^2-test. While this is a most common setting, we have not yet explored how to potentially adapt our framework to operate on summary statistics from other models. There exist summary statistics-based methods transforming GWAS results obtained from linear mixed model association to odds ratio [320], and it will be interesting to further explore how such methods could interface with our approach although this is beyond the scope of our present study. Also, in this paper, we presented immediate applications of our framework to common tasks in GWAS analyses. An interesting topic for future work would be to incorporate information beyond GWAS summary statistics. For example, one could consider incorporating external information such as LD structure using LD reference maps; such information could for instance be used to attempt to improve the accuracy of sample overlap estimation and extend the group-PRS applications. Furthermore, we could conceivably move towards haplotype reconstruction opening up new possibilities for research.

In conclusion, we introduce a simple and mathematically elegant framework that may be used to reconstruct allelic counts and genotypes from GWAS summary statistics. This novel framework highlights the power of summary-statistics-based methodology. Future work could lead to additional applications opening up new possibilities in the quest to identify the genetic background of complex disease.

3.5 Theory in details

3.5.1 Our framework

Notation Prior to introducing our methods, we discuss notational conventions. We will reserve the subscript i to denote SNP number: given, say, M SNPs, i will range between one and M. Similarly, we will reserve the subscript ℓ to denote the study number: given L studies from which summary statistics will be meta-analyzed, ℓ will range between one and L. We assume that all L studies released summary statistics on a *common set* of M SNPs. For simplicity, we will first describe our methods for the case $L = 2$ (i.e., when exactly two studies are jointly meta-analyzed) and we will generalize our approach in Section 3.5.2 for $L > 2$.

We will use the three-letter shorthand `cse` for cases and the three-letter shorthand `cnt` for controls. We reserve the variable a to represent counts of the effective allele and the variable u to represent counts of the non-effective allele. We also reserve the variable N to represent counts for the number of cases or controls. Given the above conventions, we now present the following table of allele counts (effective and non-effective allele) for SNP i ($i = 1 \dots M$) in study ℓ ($\ell = 1 \dots L$).

Table 3.1: **Table of allele counts for SNP i ($i = 1 \dots M$) in the ℓ-th GWAS ($\ell = 1 \dots L$).** The total number of cases for the ℓ-th study is $N_\ell^{\texttt{cse}}$ and the total number of controls for the ℓ-th study is $N_\ell^{\texttt{cnt}}$. Clearly, the total number of cases and controls in a study is the same for all SNPs, which is why the variable N does not depend on i. The total number of alleles in cases and controls is equal to twice the number of cases and controls, respectively.

	A_1 (effective allele)	A_2 (non-effective allele)	Number of alleles
Cases	$a_{i\ell}^{\text{cse}}$	$u_{i\ell}^{\text{cse}}$	$2N_\ell^{\text{cse}}$
Controls	$a_{i\ell}^{\text{cnt}}$	$u_{i\ell}^{\text{cnt}}$	$2N_\ell^{\text{cnt}}$

Using the above table, we can also compute the frequencies of the effective or non-effective allele in cases and controls. Table 3.2 summarizes frequency notation for SNP i ($i = 1 \dots M$) in study ℓ ($\ell = 1 \dots L$). Obviously,

Table 3.2: **Notations and definitions of (effective or non-effective) allele frequencies in cases and controls.** The subscripts i and ℓ indicate SNP number and study number, respectively.

$p_{i\ell}^{\texttt{cse}} = \frac{a_{i\ell}^{\text{cse}}}{a_{i\ell}^{\text{cse}} + u_{i\ell}^{\text{cse}}}$	frequency of the *effective allele* A_1 in cases
$p_{i\ell}^{\texttt{cnt}} = \frac{a_{i\ell}^{\text{cnt}}}{a_{i\ell}^{\text{cnt}} + u_{i\ell}^{\text{cnt}}}$	frequency of the *effective allele* A_1 in controls
$q_{i\ell}^{\texttt{cse}} = \frac{u_{i\ell}^{\text{cse}}}{a_{i\ell}^{\text{cse}} + u_{i\ell}^{\text{cse}}}$	frequency of the *non-effective allele* A_2 in cases
$q_{i\ell}^{\texttt{cnt}} = \frac{u_{i\ell}^{\text{cnt}}}{a_{i\ell}^{\text{cnt}} + u_{i\ell}^{\text{cnt}}}$	frequency of the *non-effective allele* A_2 in controls

$$p_{i\ell}^{\texttt{cse}} + q_{i\ell}^{\texttt{cse}} = 1$$

$$p_{i\ell}^{\texttt{cnt}} + q_{i\ell}^{\texttt{cnt}} = 1.$$

Reconstructing allele counts Using Table 3.1, notice that the odds ratio (OR) and its corresponding standard error (SE) for SNP i in study ℓ are given by the following formulas:

$$OR_{i\ell} = \frac{a_{i\ell}^{\text{cse}} \cdot u_{i\ell}^{\text{cnt}}}{a_{i\ell}^{\text{cnt}} \cdot u_{i\ell}^{\text{cse}}}, \tag{1}$$

$$SE_{i\ell} = \sqrt{\frac{1}{a_{i\ell}^{\text{cse}}} + \frac{1}{u_{i\ell}^{\text{cse}}} + \frac{1}{a_{i\ell}^{\text{cnt}}} + \frac{1}{u_{i\ell}^{\text{cnt}}}}. \tag{2}$$

Additionally,

$$2N_{\ell}^{\text{cse}} = a_{i\ell}^{\text{cse}} + u_{i\ell}^{\text{cse}}, \quad \text{and} \tag{3}$$

$$2N_{\ell}^{\text{cnt}} = a_{i\ell}^{\text{cnt}} + u_{i\ell}^{\text{cnt}}. \tag{4}$$

By solving the system of non-linear eqns. (1), (2), (3), and (4), we can recover $a_{i\ell}^{\text{cse}}$, $u_{i\ell}^{\text{cse}}$, $a_{i\ell}^{\text{cnt}}$, and $u_{i\ell}^{\text{cnt}}$ for SNP i in study ℓ. Notice that $OR_{i\ell}$, $SE_{i\ell}$, N_{ℓ}^{cse}, and N_{ℓ}^{cnt} are available from summary statistics. See Appendix 3.7.1 for details on solving the aforementioned system of non-linear equations.

Reconstructing genotype counts Given the reconstructed allele counts of Section 3.5.1, we can now reconstruct genotype counts for SNP i in the ℓ-th study. In order to do this, we need to assume that SNP i is in HWE in both case and control groups of study ℓ. Note that a well-performed GWAS should have SNPs drastically violating HWE filtered out. More precisely, assume that for SNP i in study ℓ we have reconstructed its allele table count (Table 3.1). Then, by assuming that this SNP is in HWE in study ℓ, we can compute the number of cases and controls that exhibit a particular genotype. Recall that there are three possible genotypes: A_1A_1, A_1A_2, and A_2A_2. We will represent each genotype by counting the number of copies of the effective allele in each genotype. Thus, A_1A_1 will correspond to two, A_1A_2 will correspond to one, and A_2A_2 will correspond to zero.

Following our notational conventions from Section 3.5.1, we can now compute the entries in Table 3.3 of genotype counts for SNP i in study ℓ. It is worth noting that

Table 3.3: **Genotype counts for cases and controls for SNP i in study ℓ.** Using the above formulas, we can reconstruct the genotype counts for cases and controls for each of the three possible genotypes.

	A_1A_1 (two copies of A_1)	A_1A_2 (one copy of A_1)	A_2A_2 (zero copies of A_1)
Cases	$N_{i\ell}^{\text{cse}}(2) = (p_{i\ell}^{\texttt{cse}})^2 N_{\ell}^{\texttt{cse}}$	$N_{i\ell}^{\text{cse}}(1) = 2p_{i\ell}^{\texttt{cse}} q_{i\ell}^{\texttt{cse}} N_{\ell}^{\texttt{cse}}$	$N_{i\ell}^{\text{cse}}(0) = (q_{i\ell}^{\texttt{cse}})^2 N_{\ell}^{\texttt{cse}}$
Controls	$N_{i\ell}^{\text{cnt}}(2) = (p_{i\ell}^{\texttt{cnt}})^2 N_{\ell}^{\texttt{cnt}}$	$N_{i\ell}^{\text{cnt}}(1) = 2p_{i\ell}^{\texttt{cnt}} q_{i\ell}^{\texttt{cnt}} N_{\ell}^{\texttt{cnt}}$	$N_{i\ell}^{\text{cnt}}(0) = (q_{i\ell}^{\texttt{cnt}})^2 N_{\ell}^{\texttt{cnt}}$

$$N_{\ell}^{\text{cse}} = N_{i\ell}^{\text{cse}}(0) + N_{i\ell}^{\text{cse}}(1) + N_{i\ell}^{\text{cse}}(2), \tag{5}$$

$$N_{\ell}^{\text{cnt}} = N_{i\ell}^{\text{cnt}}(0) + N_{i\ell}^{\text{cnt}}(1) + N_{i\ell}^{\text{cnt}}(2). \tag{6}$$

Next, we reconstruct the genotype vector for SNP i in study ℓ as follows:

$$\mathbf{g}_{i\ell} = \left[\ \underbrace{0\ldots0}_{N_{i\ell}^{\text{cse}}(0)} \ \underbrace{1\ldots1}_{N_{i\ell}^{\text{cse}}(1)} \ \underbrace{2\ldots2}_{N_{i\ell}^{\text{cse}}(2)} \ \underbrace{0\ldots0}_{N_{i\ell}^{\text{cnt}}(0)} \ \underbrace{1\ldots1}_{N_{i\ell}^{\text{cnt}}(1)} \ \underbrace{2\ldots2}_{N_{i\ell}^{\text{cnt}}(2)} \ \right].$$

Using eqns. (5) and (6), it is easy to conclude that the vector $\mathbf{g}_{i\ell}$ has a total of

$$N_{\ell}^{\text{cse}} + N_{\ell}^{\text{cnt}}$$

entries, which is equal to the number of samples (cases plus controls) included in the ℓ-th study. We can also form the response vector $\mathbf{y}_{\ell}$ for the ℓ-th study, indicating whether a sample is a case (i.e., one) or a control (i.e., zero) as follows:

$$\mathbf{y}_{\ell} = \left[\ \underbrace{1\ldots1}_{N_{\ell}^{\text{cse}}} \ \underbrace{0\ldots0}_{N_{\ell}^{\text{cnt}}} \ \right]. \tag{7}$$

Note that the vectors $\mathbf{y}_{\ell}$ and $\mathbf{g}_{i\ell}$ have the same dimensions (same number of entries). It should be clear that the vector $\mathbf{y}_{\ell}$ *is the same for all SNPs* in the ℓ-th study and hence does not depend on the SNP number i.

We conclude the section by discussing the construction of an indicator vector $\mathbf{s}$ that will denote the study from which a particular sample in our meta-analysis originated. For the sake of simplicity, assume that we meta-analyze summary statistics from two studies ($L = 2$). Then, following the above discussion, we can construct the genotype vectors $\mathbf{g}_{i1}$ and $\mathbf{g}_{i2}$ and concatenate them to construct the overall genotype vector for the i-th SNP in both studies:

$$\mathbf{g}_i = [\mathbf{g}_{i1}\ \mathbf{g}_{i2}] \,.$$

Similarly, we can construct the overall response vector $\mathbf{y}$ for both studies:

$$\mathbf{y} = [\mathbf{y}_1\ \mathbf{y}_2] \,.$$

Notice that the vectors $\mathbf{g}_i$ and $\mathbf{y}$ have the same dimensions (number of entries), equal to the number of samples (cases plus controls) in both studies, i.e., equal to

$$N = N_1^{\text{cse}} + N_1^{\text{cnt}} + N_2^{\text{cse}} + N_2^{\text{cnt}}.$$

We can now construct the indicator vector $\mathbf{s}$ as follows:

$$\mathbf{s} = \left[\begin{array}{cc} \underbrace{0\ldots0}_{N_1^{\text{cse}}+N_1^{\text{cnt}}} & \underbrace{1\ldots1}_{N_2^{\text{cse}}+N_2^{\text{cnt}}} \end{array} \right].$$

Note that a value of zero in $\mathbf{s}$ indicates that the corresponding sample belongs to the first study while a value of one in $\mathbf{s}$ indicates that the corresponding sample belongs to the second study.

3.5.2 Fixed-effect meta-analysis

Logistic regression We run logistic regression for each SNP separately; recall that we number SNPs in our meta-analysis from one up to M. For notational convenience and since we run logistic regression in an identical manner for each SNP, without loss of generality we focus on a single SNP. Let the genotype vector for the selected SNP be denoted by $\mathbf{g}$; let $\mathbf{s}$ be the study indicator vector; and let $\mathbf{y}$ be the response vector, as discussed in the previous section. Recall that all three vectors have the same dimensions (same number of entries), equal to N, namely the total number of cases and controls in both studies. *Notice that we dropped the subscript i from the vector $\mathbf{g}$ for notational convenience, since our discussion in this section will focus on a fixed SNP i, without* *loss of generality.*

Using notation from the previous section, while dropping the subscript i from the genotype vector $\mathbf{g}$, allows us to formulate logistic regression as follows:

$$\mathsf{Pr}(\mathbf{y}_j = 1 | \mathbf{g}_j, \mathbf{s}_j) = S(\beta_0 + \beta_1 \mathbf{g}_j + \beta_2 \mathbf{s}_j), \tag{8}$$

where $S(x) = (1 + e^{-x})^{-1}$ is the sigmoid function; $\mathbf{y}_j$ denotes the jth entry of the vector $\mathbf{y}$; $\mathbf{s}_j$ denotes the jth entry of the vector $\mathbf{s}$; and β_0, β_1, and β_2 are the unknown coefficients of the logistic regression formulation. Here β_0 corresponds to the constant offset, β_1 corresponds to the genotype, and β_2 corresponds to the study-of-origin. We also highlight that $\mathbf{g}_j$ denotes the jth entry of the vector $\mathbf{g}$; recall once again that we dropped the subscript i from the genotype vector in this section. The range for all subscripts j for the above vectors is between one and N.

In order to further describe how logistic regression was implemented in our experiments, it will be convenient to introduce additional notation. Let β be the vector

$$\beta^T = [\beta_0 \ \beta_1 \ \beta_2],$$

and let $\mathbf{x}$ be the vector

$$\mathbf{x}_j^T = [1 \ \mathbf{g}_j \ \mathbf{s}_j].$$

Thus, β is the vector of the (unknown) logistic regression coefficients, while $\mathbf{x}_j^T$ for all $j = 1 \ldots N$ is the vector representing the constant offset, the genotype, and the study origin for the jth sample in our meta-analysis. This allows us to rewrite eqn. (8) as follows:

$$\mathsf{Pr}(\mathbf{y}_j = 1 | \mathbf{g}_j, \mathbf{s}_j) = S(\beta^T \cdot \mathbf{x}_j).$$

We can now compute the negative log-likelihood (NLL) function for β as follows:

$$\begin{aligned} NLL(\beta) &= -\sum_{j=1}^{N} \log(\mathsf{Pr}(\mathbf{y}_j)) = 1|\mathbf{x}_j) \\ &= -\sum_{j=1}^{N} \mathbf{y}_j \log S(\beta^T \cdot \mathbf{x}_j) + (1 - \mathbf{y}_j) \log(1 - S(\beta^T \cdot \mathbf{x}_j)). \end{aligned}$$

Thus, β can be estimated using the Iterative Re-weighted Least Squares (IRLS) algorithm [321] as follows:

Algorithm 1: IRLS for maximum likelihood estimate of logistic regression coefficients

```
Initialize $\beta^0 = [\log(\frac{\bar{y}}{1-\bar{y}})\ 0\ 0]^T$, where $\bar{y}$ is the average of all elements of the vector $\mathbf{y}$;
repeat
    $\eta_j = (\beta^t)^T \mathbf{x}_j, \quad j = 1 \ldots N$;
    $\phi_j = S(\eta_j), \quad j = 1 \ldots N$;
    $d_j = \phi_j(1 - \phi_j), \quad j = 1 \ldots N$;
    $z_j = \eta_j + \frac{\mathbf{y}_j - \phi_j}{d_j}, \quad j = 1 \ldots N$;
    $\mathbf{D} = \text{diag}(d_1, d_2, \ldots, d_N)$;
    $\beta^{t+1} = (\mathbf{X}^T\mathbf{D}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D}\mathbf{z}$;
until convergence;
```

In the IRLS algorithm, we let $\mathbf{D}$ denote the diagonal $N \times N$ matrix whose diagonal entries are $d_1, d_2, \ldots, d_N$; we let $\mathbf{X}$ denote the $N \times 3$ matrix whose rows are the vectors $\mathbf{x}_j^T$ for $j = 1 \ldots N$; and we let $\mathbf{z}$ denote the vector whose entries are the z_j for $j = 1 \ldots N$. Using this notation, the matrix $\mathbf{H} = \mathbf{X}^T\mathbf{D}\mathbf{X}$ is the 3×3 Hessian matrix of this logistic regression problem. The algorithm iterates over $t = 0, 1, 2, \ldots$ and terminates when our convergence criterion, namely the difference $\|\beta^{t+1} - \beta^t\|_1$ [1] drops below the threshold 10^{-4}, which is the same threshold as the one used by PLink [322] for logistic regression.

Note that a drawback for logistic regression is that it can produce anti-conservative results under imbalance, which in our case, includes unbalanced sample sizes in cases and controls, as well as unbalanced sample sizes among input studies. We apply Firth bias-corrected logistic regression test [323, 324] to correct for the estimate under input imbalance (triggered when either the total

[1] This is simply the sum of the absolute values of the three entries of the vector $\beta^{t+1} - \beta^t$.

case/control ratio, or maximum/minimum input sample size ratio is greater or equal to 5 by default). This approach has been reported with stable performance in both balanced and unbalanced studies, as well as with rare SNPs [325].

We conclude this section by discussing how to compute a p-value for the logistic regression formulation of eqn. (8). First, it is well-known that the standard error for the three coefficients of the logistic regression formulation can be computed by using the inverse of the Hessian matrix $\mathbf{H}$. In particular, the standard error for β_0 is equal to $SE_0 = \sqrt{(\mathbf{H^{-1}})_{11}}$; the standard error for β_1 is equal to $SE_1 = \sqrt{(\mathbf{H^{-1}})_{22}}$; and the standard error for β_2 is equal to $SE_2 = \sqrt{(\mathbf{H^{-1}})_{22}}$. As is typical in association studies, we focus on SE_1, the standard error for the vector of genotypes, and compute the respective p-value for the SNP-under-study using the Wald test. More specifically, we find the corresponding p-value of a Z-distribution for the parameter $\left|\frac{\beta_1}{SE_1}\right|$.

Correcting for sample overlap (two studies) Sample overlap between studies can lead to an under-estimation of test statistics variance and results in an inflated test p-value. To prevent this from happening, we will use an "effective sample size" correction as follows. Assume that we are given Table 3.4, which details the number of overlapping samples between the two studies.

Table 3.4: **Number of overlapping cases and controls between the two studies.** For example, the first cell of the table indicates the number of shared cases between the two studies. In practice, the off-diagonal cells of this table are close to zero, since they indicate cases in one study that became controls in the other study and vice-versa. Large numbers in these off-diagonal cells would indicate high heterogeneity across the two studies, in which case a fixed effect meta-analysis is not recommended.

Overlapping	Study 2: Case	Study 2: Control
Study 1: Case	$N_{\texttt{shr}}^{\texttt{cse-cse}}$	$N_{\texttt{shr}}^{\texttt{cnt-cse}}$
Study 1: Control	$N_{\texttt{shr}}^{\texttt{cse-cnt}}$	$N_{\texttt{shr}}^{\texttt{cnt-cnt}}$

Using the counts in Table 3.4, the number of shared cases between the two studies is equal to:

$$N_{\texttt{shr}}^{\texttt{cse}} = N_{\texttt{shr}}^{\texttt{cse-cse}} + \frac{N_{\texttt{shr}}^{\texttt{cse-cnt}} + N_{\texttt{shr}}^{\texttt{cnt-cse}}}{2}. \qquad (9)$$

Notice that if the off-diagonal entries in Table 3.4 are equal to zero then the above number reduces, obviously, to $N_{\mathtt{shr}}^{\mathtt{cse\text{-}cse}}$. Similarly, we have the number of shared controls equal to:

$$N_{\mathtt{shr}}^{\mathtt{cnt}} = N_{\mathtt{shr}}^{\mathtt{cnt\text{-}cnt}} + \frac{N_{\mathtt{shr}}^{\mathtt{cnt\text{-}cse}} + N_{\mathtt{shr}}^{\mathtt{cse\text{-}cnt}}}{2}. \tag{10}$$

Then, the correction is simply carried out by multiplying the case/control sample size of each input study by a "deflation factor" defined as follows:

$$\begin{aligned} \lambda_{\ell}^{\mathtt{cse}} &= \frac{N_{\ell}^{\mathtt{cse}}}{N_{\ell}^{\mathtt{cse}} + N_{\mathtt{shr}}^{\mathtt{cse}}} \\ \lambda_{\ell}^{\mathtt{cnt}} &= \frac{N_{\ell}^{\mathtt{cnt}}}{N_{\ell}^{\mathtt{cnt}} + N_{\mathtt{shr}}^{\mathtt{cnt}}}. \end{aligned}$$

We multiply the sample size for cases (respectively, controls) in each study ℓ by $\lambda_{\ell}^{\mathtt{cse}}$ (respectively, $\lambda_{\ell}^{\mathtt{cnt}}$) before proceeding with the logistic regression described in Section 3.5.2. See [326] for a similar correction strategy. We finally note that in practice the exact number of overlapping samples between two studies is usually not know. In this case, we followed the approach proposed in [308] to estimate the overlapping sample size.

Meta-analyzing multiple datasets We now extend our approach to meta-analyze more than two datasets. The main difference with our previously described approach is the handling of the indicator variable for multiple datasets. We can still reconstruct the genotype count for each input study in exactly the same way as in Table 3.3 as well as the response vector following eqn. (3.5.1). Therefore, when multiple studies are meta-analyzed, $\mathbf{g}_i$ and $\mathbf{y}$ become

$$\begin{aligned} \mathbf{g}_i &= \left[\mathbf{g}_{i1} \dots \mathbf{g}_{iL}\right], \\ \mathbf{y} &= \left[\mathbf{y}_1 \dots \mathbf{y}_L\right]. \end{aligned}$$

The indicator vector $\mathbf{s}$ cannot be binary anymore. Intuitively, one may consider using L binary vectors, each to encode samples from each input study. However, this approach would necessitate up to ${}^{L(L-1)}/_{2}$ vectors to encode pairwise sample overlap. This increases the computational complexity by $O(L^2)$. A simpler alternative is to use categorical variable as the source study indicator. Note that in this case, different rankings of the studies can lead to completely different results. A straightforward idea is to encode the studies using their population allele frequencies, which can be computed via Table 3.1 as follows:

$$I_{i\ell} = \frac{a_{i\ell}^{\text{cse}} + a_{i\ell}^{\text{cnt}}}{a_{i\ell}^{\text{cse}} + a_{i\ell}^{\text{cnt}} + u_{i\ell}^{\text{cse}} + u_{i\ell}^{\text{cnt}}}$$

Note this is encoding also controls for population stratification across multiple sample sources. Then, when analyzing L studies, the indicator vector $\mathbf{s}$ becomes:

$$\mathbf{s} = \left[\underbrace{I_1 \dots I_1}_{N_1^{\text{cse}} + N_1^{\text{cnt}}} \dots \underbrace{I_L \dots I_L}_{N_L^{\text{cse}} + N_L^{\text{cnt}}} \right].$$

We can now proceed with the logistic regression as in Section 3.5.2 . In order to handle sample overlap across multiple studies, we use the subscript $(\cdot)_{\ell_1 \ell_2}$ to denote properties of shared samples between two studies ℓ_1 and ℓ_2. Then, generalizing eqns. (9) and (10), we get, for each pair of input studies ℓ_1 and ℓ_2,

$$\begin{aligned} N_{\ell_1\ell_2}^{\texttt{cse}} &= N_{\ell_1\ell_2}^{\texttt{cse-cse}} + \frac{N_{\ell_1\ell_2}^{\texttt{cse-cnt}} + N_{\ell_1\ell_2}^{\texttt{cnt-cse}}}{2}, \\ N_{\ell_1\ell_2}^{\texttt{cnt}} &= N_{\ell_1\ell_2}^{\texttt{cnt-cnt}} + \frac{N_{\ell_1\ell_2}^{\texttt{cnt-cse}} + N_{\ell_1\ell_2}^{\texttt{cse-cnt}}}{2}. \end{aligned}$$

Finally, for any study $\ell_1 = 1 \dots L$, the sample size correction is

$$\begin{aligned} \lambda_{\ell_1}^{\texttt{cse}} &= \frac{N_{\ell_1}^{\texttt{cse}}}{N_{\ell_1}^{\texttt{cse}} + \sum_{\ell_2 \neq \ell_1}^{L} N_{\ell_1\ell_2}^{\texttt{cse}}}, \\ \lambda_{\ell_1}^{\texttt{cnt}} &= \frac{N_{\ell_1}^{\texttt{cnt}}}{N_{\ell_1}^{\texttt{cnt}} + \sum_{\ell_2 \neq \ell_1}^{L} N_{\ell_1\ell_2}^{\texttt{cnt}}}. \end{aligned}$$

We can now apply $\lambda_{\ell_1}^{\texttt{cse}}$ to correct the sample size for cases in study ℓ_1 and we can apply $\lambda_{\ell_1}^{\texttt{cnt}}$ to correct the sample size for controls and proceed with logistic regression.

3.5.3 PRS and cc-GWAS

Mean PRS for cases and controls Recall that the PRS for the t-th individual in the study is computed as:

$$\mathrm{PRS}_t = \sum_{i=1}^{M} \frac{S_i \cdot g_{it}}{2M}, \tag{11}$$

where g_{it} is the genotype of the i-th SNP for the t-th individual and S_i is the weight for SNP i, which is usually defined as

$$S_i = \log(\mathrm{OR}_i^{\mathtt{base}}),$$

where $\mathrm{OR}_i^{\mathtt{base}}$ is the odds ratio of SNP i in the base summary statistics. Recall from Section 3.5.1 that M is the total number of SNPs. Then, in order to compute the average PRS for, say, cases, we simply need to sum up the individual PRS and average over the number of cases. More precisely,

$$\mathrm{PRS}^{\mathtt{cse}} = \frac{1}{2MN^{\mathtt{cse}}} \sum_{t \in \mathtt{cse}} \sum_{i=1}^{M} S_i \cdot g_{it}.$$

where $N^{\mathtt{cse}}$ is the number of cases in the target study. The above equation can be rewritten as

$$\mathrm{PRS}^{\mathtt{cse}} = \frac{1}{2MN^{\mathtt{cse}}} \sum_{i=1}^{M} S_i \sum_{t \in \mathtt{cse}} g_{it}.$$

Notice that in an additive model, $\sum_{t \in \mathtt{cse}} g_{it}/2N^{\mathtt{cse}}$ is the allele frequency of SNP i over all cases in the target study, which can be computed using only the summary statistics as shown in Section 3.5.1 and Table 3.2. Thus, the mean PRS under an additive model for cases and controls can be computed as follows:

$$\mathrm{PRS}^{\mathtt{cse}} = \frac{\sum_{i=1}^{M} S_i p_i^{\mathtt{cse}}}{M},$$
$$\mathrm{PRS}^{\mathtt{cnt}} = \frac{\sum_{i=1}^{M} S_i p_i^{\mathtt{cnt}}}{M}.$$

Table 3.5: **The probability distribution of g_{it} for SNP i.** In this table, p_i^{cse} denotes the allele frequency of A_1 in cases and $q_i^{\mathsf{cse}} = 1 - p_i^{\mathsf{cse}}$.

$g_{it} = 2$ (two copies of A_1)	$g_{it} = 1$ (one copy of A_1)	$g_{it} = 0$ (zero copies of A_1)
$(p_i^{\mathsf{cse}})^2$	$2p_i^{\mathsf{cse}}q_i^{\mathsf{cse}}$	$(q_i^{\mathsf{cse}})^2$

All relevant information for this computation can be easily obtained from the summary statistics of the base and/or target study.

Estimating the standard deviation of the PRS for cases and controls Interestingly, we can also estimate the standard deviation of the PRS for cases and controls, even Without individual level genotype information, under mild assumptions. First, from eqn. (11), we compute the variance of an individual's PRS as follows:

$$\begin{aligned}\mathrm{Var}(\mathrm{PRS}_t) &= \mathrm{Var}(\sum_{i=1}^{M} \frac{S_i \cdot g_{it}}{2M}) \\ &= \frac{1}{4M^2}\mathrm{Var}(\sum_{i=1}^{M} S_i \cdot g_{it}). \end{aligned} \tag{12}$$

Recall that as a general step prior to the computation of PRS, it is recommended to prune or clump the SNPs used for the PRS computation. Therefore, our first assumption is that the g_{it}'s are pairwise independent. Then, eqn. (12) can be simplified as follows:

$$\begin{aligned}\mathrm{Var}(\mathrm{PRS}_t) &= \frac{\sum_{i=1}^{M}\mathrm{Var}(S_i \cdot g_{it})}{4M^2} \\ &= \frac{\sum_{i=1}^{M} S_i^2 \mathrm{Var}(g_{it})}{4M^2}. \end{aligned} \tag{13}$$

Notice that under an additive model, g_{it} is a discrete random variable that only takes the value zero, one, and two. Consider all cases and, as in Section 3.5.1 , assume that the SNPs are in HWE. Then, the distribution of g_{it} in the cases is presented in Table 3.5. We can now compute the variance of g_{it} in cases as follows:

$$\begin{aligned}\mathrm{Var}(g_{it}) &= \mathrm{E}(g_{it}^2) - (\mathrm{E}g_{it})^2 \\ &= (2p_i^{\mathtt{cse}}q_i^{\mathtt{cse}} + 4(p_i^{\mathtt{cse}})^2) - (2p_i^{\mathtt{cse}}q_i^{\mathtt{cse}} + 2(p_i^{\mathtt{cse}})^2)^2 \\ &= (2p_i^{\mathtt{cse}}q_i^{\mathtt{cse}} + 4(p_i^{\mathtt{cse}})^2) - (2p_i^{\mathtt{cse}}(p_i^{\mathtt{cse}} + q_i^{\mathtt{cse}}))^2 \\ &= 2p_i^{\mathtt{cse}}q_i^{\mathtt{cse}} + 4(p_i^{\mathtt{cse}})^2 - 4(p_i^{\mathtt{cse}})^2 = 2p_i^{\mathtt{cse}}q_i^{\mathtt{cse}}.\end{aligned}$$

Substituting into eqn. (13), we get

$$\mathrm{Var}(\mathrm{PRS}^{\mathtt{cse}}) = \frac{\sum_{i=1}^{M} S_i^2(2p_i^{\mathtt{cse}}q_i^{\mathtt{cse}})}{4M^2}.$$

Similarly, we can compute the estimated variance $\mathrm{PRS}^{\mathtt{cnt}}$ for controls and PRS for the overall population of the target study. To summarize, our estimates are

$$\begin{aligned}\mathrm{Var}(\mathrm{PRS}^{\mathtt{cse}}) &= \frac{\sum_{i=1}^{M} S_i^2 p_i^{\mathtt{cse}}q_i^{\mathtt{cse}}}{2M^2}, \\ \mathrm{Var}(\mathrm{PRS}^{\mathtt{cnt}}) &= \frac{\sum_{i=1}^{M} S_i^2 p_i^{\mathtt{cnt}}q_i^{\mathtt{cnt}}}{2M^2}, \\ \mathrm{Var}(\mathrm{PRS}) &= \frac{\sum_{i=1}^{M} S_i^2 p_i q_i}{2M^2}.\end{aligned}$$

Here p_i is the frequency of allele A_1 for SNP i in all samples of the target study, and can be computed as:

$$\begin{aligned}p_i &= \frac{N^{\mathtt{cse}}p_i^{\mathtt{cse}} + N^{\mathtt{cnt}}p_i^{\mathtt{cnt}}}{N^{\mathtt{cse}} + N^{\mathtt{cnt}}}, \\ q_i &= 1 - p_i.\end{aligned}$$

We can now apply a t-test in order to obtain a p-value for the difference between the PRS distributions in cases and controls. Given the estimated group means and standard deviations for cases and controls, we can further assume that the individual level PRS follow a normal distribution in each group and use the t-test statistic as follows:

$$t = \frac{\text{PRS}^{\mathsf{cse}} - \text{PRS}^{\mathsf{cnt}}}{\sqrt{\text{Var}(\text{PRS})} \cdot \sqrt{\frac{1}{N^{\mathsf{cse}}} + \frac{1}{N^{\mathsf{cnt}}}}}.$$

Finally, the degrees of freedom are given by $df = N^{\mathsf{cse}} + N^{\mathsf{cnt}} - 2$.

cc-GWAS using summary statistics cc-GWAS is a straight-forward approach to investigate the genetic differences between two traits. However, in practice, it is usually challenging and time consuming, due to restrictions in individual level data sharing. Recently, a method for cc-GWAS that relies only on summary statistics has been proposed in [194]. We propose an alternative perspective on summary-statistics-based cc-GWAS framework, using the foundations of Section 3.5.1.

One of the biggest challenges of cc-GWAS is the differentiation of the genetic effects from trait-trait difference and population stratification. Assume that for a fixed SNP, we run logistic regression focusing only on the cases of the two studies. Let $\mathbf{y}_j^{\mathsf{cse}} = 1$ denote that sample j is a case from the first study and let $\mathbf{y}_j^{\mathsf{cse}} = 0$ denote that j is a case from the second study. Let $\mathbf{g}_j^{\mathsf{cse}}$ be the genotype of the j-th case. Then,

$$\mathsf{Pr}(\mathbf{y}_j^{\mathsf{cse}} = 1 | \mathbf{g}_j^{\mathsf{cse}}) = S(\beta_0^{\mathsf{cse}} + \beta_1^{\mathsf{cse}} \mathbf{g}_j^{\mathsf{cse}}). \tag{14}$$

The effect size β_1^{cse} that is the output of logistic regression will include effects from the real genetic differences between trait 1 and trait 2 (β_g) as well as from population stratification (β_s). We can assume that these two effects are independent of each other:

$$\beta_1^{\mathsf{cse}} = \beta_g + \beta_s.$$

Assume that the control samples from studies one and two *do not carry the traits of interest.* Then, we can estimate the effect of population stratification by running another logistic regression, focusing only on controls from the two studies, as follows:

$$\mathsf{Pr}(\mathbf{y}_j^{\mathtt{cnt}} = 1|\mathbf{g}_j^{\mathtt{cnt}}) = S(\beta_0^{\mathtt{cnt}} + \beta_s \mathbf{g}_j^{\mathtt{cnt}}). \tag{15}$$

In the above, $\mathbf{y}_j^{\mathtt{cnt}} = 1$ denotes that sample j is a control from study one, $\mathbf{y}_j^{\mathtt{cnt}} = 0$) denotes that j is a control from study two, and $\mathbf{g}_j^{\mathtt{cnt}}$ denotes the the genotype for the j-th control sample. From this logistic regression, we can get an estimate of the stratification effect β_s. Note that along with β_s, we will also get a standard error for the estimate of stratification SE_s, which essentially corresponds to the sample size of controls in the two input studies. If we do not have a good amount of controls, SE_s will turn out to be large, indicating that the estimate for stratification effect is not reliable and the results from the cc-GWAS should be be interpreted carefully.

If SE_s is small enough, then it is reasonable to assume that the estimate of the stratification effect is credible and we can subsequently treat β_s as a fixed value. Then, the genetic effect from the trait-trait difference that we are interested in is

$$\beta_g = \beta_1^{\mathtt{cse}} - \beta_s. \tag{16}$$

It now follows that the standard error of β_g is

$$\mathrm{Var}(\beta_g) = \mathrm{Var}(\beta_1^{\mathtt{cse}}) \implies \mathrm{SE}_g = \mathrm{SE}_1, \tag{17}$$

using the derivations of Section 3.5.1. Logistic regressions on cases (eqn. (14)) and controls (eqn. (15)) can be carried out as discussed in Section 3.5.2, with minor changes (include only the designated samples; relabel the dependent variable; and remove the indicator variable). By running these two logistic regressions, we can compute $\beta_1^{\mathtt{cse}}, \beta_s, SE_1^{\mathtt{cse}}$, and SE_s. Then, using eqns. (16) and (17), we can compute β_g and SE_g for each SNP. Similarly, we can also compute the corresponding p-value using a Z-distribution for $\left|\frac{\beta_g}{SE_g}\right|$.

3.6 Experiments in details

3.6.1 Data

Synthetic data. We used the Balding-Nichols model [307, 327] for synthetic genotype generation, assuming a minor allele frequency (MAF) of 0.3 for each SNPs and a relative risk r ($r = 1.15/1.2/1.3$) for the effective alleles of the causal SNPs in each population. The simulation was carried out under a range of F_{st} values ($F_{st} = 0.01/0.05/0.1$). For the fixed-effect meta-analysis, we simulated 1,000 cases and 1,000 controls for each input study. A total of 100,000 SNPs were generated, out of which 1,000 are causal SNPs with the predefined risk for the effective alleles. Moreover, on top of the independent populations, we also evaluated the performance of ReACt under the presence of sample overlap by introducing a predefined amount of samples shared between each pair of input studies (100 cases, 100 controls overlap; or 500 cases, 500 controls overlap).

For the cc-GWAS, inspired by [194], we used the same simulation model but introduced three types of SNPs for a thorough evaluation of the method's robustness: *(i)* SNPs with non-zero effect in only one of the studies and zero effect in the other; *(ii)* SNPs with zero effect in both input studies; and *(iii)* SNPs with the same non-zero effect size (predefined r) in both input studies. All of the three types of SNPs would suffer from population stratification at a predefined value of F_{st}. In total, 100,000 SNPs were generated, with 1,000 (for each input study) from type (i), 49,000 from type (ii), and 49,000 from type (iii). To investigate the effect of study sizes, we evaluated the method performance on input studies with 2,000 cases and 2,000 controls each, as well as on studies with 5,000 cases and 5,000 controls each.

Individual level genotype data. We tested the performance of our fixed-effect meta-analysis method and group PRS method on the depressive episode trait in UK biobank dataset [135]. Only independent European ancestry samples identified through PCA and IBD check are included for the analysis. We applied basic quality control filters on those samples, including removing SNPs and samples with a missing rate exceeding 2% or violating the Hardy-Weinberg equilibrium ($p_{HWE} < 10^{-6}$). As a result, 640,756 SNPs and 331,217 samples (18,368 cases and 312,849 controls) survived and were used for the experiment. For the evaluation of the fixed-effect meta-analysis method, we ran a standard GWAS with all samples and treated SNPs with $p < 10^{-6}$ from the results as the “true signals” to be captured. For all GWAS on UB biobank samples, we correct for age, gender, sample collection batch and top 10 PCs obtained using software TeraPCA [328].

Generating summary statistics. For synthetic data and individual level genotypes, summary statistics were generated using PLink [322], correcting for the top ten principal components (PCs) in the case of admixed datasets. For real individual level genotype data, we divided the samples randomly into two equal sized subsets and ran a GWAS on each subset separately to obtain summary statistics for each subset. We performed ten such random iterations in our experimental evaluations. For the fixed-effect meta-analysis, on top of two independent subsets, we also introduced 100/500 sample overlap to investigate the performance of our methods under more challenging scenarios.

Publicly available summary statistics. As part of the performance evaluation for our group PRS method, we used a MDD GWAS summary statistics published in 2013 [312] as the base study, in order to avoid sample overlap with the UK biobank target population. The summary statistics contains in total 1,235,109 SNPs on genome build hg18. After liftover [329] to hg19, 1,234,855 remained for the analysis.

For group PRS and cc-GWAS, we demonstrated the applicability of our methods using publicly available summary statistics. We chose the summary statistics of eight neuropsychiatric disorders made available by the Psychiatric Genomics Consortium (PGC), since the underlying relationships between this set of disorders has been relatively well-studied. Information on the eight summary statistics can be found in Table 3.6.

3.6.2 Evaluation metrics

Fixed-effect meta-analysis. For synthetic experiments, results after performing the meta-analysis were compared with the predefined causal variants. Power and type I error rate under

Table 3.6: **Information on summary statistics for the eight psychiatric disorders used in the experiments.** Note that we used summary statistics only for samples of European ancestry. For MD, we used the summary statistics generated by UK biobank, excluding the 23andMe samples; for BIP, we used the summary statistics including all three patient sub-types.

Disorder	#Cases	#Controls	Total	#SNPs	Reference
obsessive-compulsive disorder (OCD)	2,688	7,037	9,725	8,409,516	[242]
Tourette syndrome (TS)	4,819	9,488	14,307	8,947,432	[111]
eating disorder (ED)	3,495	10,982	14,477	10,641,224	[330]
autism spectrum disorder (ASD)	18,382	27,969	46,351	9,112,386	[56]
bipolar disorder (BIP)	20,352	31,358	51,710	13,413,244	[313]
schizophrenia (SCZ)	36,989	113,075	150,064	9,075,843	[314]
attention-deficit/hyperactivity disorder (ADHD)	19,099	34,194	53,293	8,094,094	[109]
major depression (MD)	69,232	161,009	230,241	9,874,289	[106]

each experimental condition were reported as an average of ten independent repetitions. For real genotype data, in each iteration, we meta-analyzed summary statistics of two subsets using the proposed methods and standard approaches and compared results with the GWAS results on the complete dataset. We again reported results averaged over ten iterations (random splits) showing, on average, how many times a SNP reported as a “true signal” in the overall GWAS got picked up by each meta-analysis method (true positive) as well as how many extra SNPs each method identified (false positive). The performance on real genotype data was also evaluated under 0/100/500 sample overlap. Sample size for each subset under different conditions was 482 cases, 993 controls with no sample overlap; 532 cases, 1043 controls with 100 cases and 100 controls overlap; and 732 cases, 1243 controls with 500 cases and 500 controls overlap.

We compared the performance of ReACt in terms of accuracy as well as running time with METAL [309] and ASSET [310], which are both widely used tools for fixed-effect meta-analysis. Note that the latest stable release of `METAL` does not have the sample overlap correction functionality implemented. Therefore, for performance comparison, we used the *development version* available on GitHub [308].

Group PRS. In order to show that our method outputs reliable estimates of the group-wise statistics for PRS without accessing individual level genotypes, we compared the output of our method to the true group mean and standard deviation computed from the individual level PRS on synthetic data, as described in Section 3.6.1. Performance was evaluated under with a fixed 0.05

F_{st} between the base and target studies. For a pair of base and target studies , we estimated the mean PRS for case/control groups as well as their standard deviation using SNPs with p-values strictly less than $5 \cdot 10^{-5}$ in the summary statistics. We also computed the individual level PRS using PRSise to obtain the true group mean and standard deviation. Our experiments show that our estimates are numerically close to the real values. Next, we evaluated the performance of ReACt on real GWAS datasets, where the individual level genotype of the target study was available. For this experiment, we used an earlier GWAS summary statistics of MDD [312] as the base study (see Section 3.6.1 for details) and cases and matching controls of depressive episode trait in UK biobank as the target population [135]. We clumped the base summary statistics using the European samples from 1000 Genome Project as reference, under parameters `--clump-p1 1 --clump-kb 250 --clump-r2 0.1`. We tested the method and reported results under a range of p-value thresholds $(0.1, 0.01, 0.001, 10^{-4})$. For each threshold, we used only independent SNPs with a p-value smaller than the respective threshold from the base summary statistics for PRS calculation, using both ReACt and PRSice2 [311]. We reported the mean PRS of cases and controls, as well as the resulting p-value from t-test. In the case of PRSice2, we also reported the regression r^2 value and p-value for the PRS predictor with and without correcting for covariates (ie., the top five principal components).

Finally we applied ReACt to summary statistics of eight neuropsychiatric disorders (OCD, TS, ED, ADHD, ASD, BIP, SCZ and MDD, see Section 3.6.1 for details) and reported the pairwise PRS prediction power in terms of t-test p-values for the difference between case/control group PRS means. Prior to the group PRS computation, each base summary statistics was clumped using PLink [322] using parameters `--clump-p1 1 --clump-kb 250 --clump-r2 0.1`, with the European samples from 1000 Genome Project as a reference. All PRS values were estimated using independent SNPs with p-values strictly less than 10^{-5} from the base summary statistics.

cc-GWAS. Out of the three types of SNPs generated for the cc-GWAS evaluation (see Section 3.6.1), we expect ReACt to pick up only type (i) SNPs as they have been designed to be the trait differential SNPs. Therefore, we reported the power of ReACt based on the number of type (i) SNPs that were identified as well as type I error rates for type (ii) SNPs and type (iii) SNPs.

Since the randomness introduced by the simulation could lead to false positives that were not due to the method itself, we filtered out type (iii) SNPs showing extreme differences in effect size between studies, by removing type (iii) SNPs with $|OR_{i1} - OR_{i2}| \geq 0.1$ from performance evaluation. Here OR_{i1} corresponds to the odd ratio for the ith SNP in the first study and OR_{i2} corresponds to the odd ratio for the ith SNP in the other study. Since all three types of SNPs suffered from population stratification, we evaluated the performance of ReACt under a challenging scenario. Besides simulation, experiments using summary statistics for schizophrenia (SCZ) [314] and bipolar disorder (BIP) [331] were also carried out. These two disorders were chosen due to the existence of case-case association study using the individual level genotypes [236]. We tested ReACt using the summary statistics and compared the results with the existing case-case association study between SCZ and BIP to see whether it could detect possible genetic differences between the two disorders. Since no individual level quality control could be carried out, we expected our results to correspond to a case-case GWAS including 36,989 cases from SCZ and 20,352 cases from all three sub-types of BIP (type 1, type 2, and schizoaffective bipolar disorder). For the analysis, we excluded SNPs on the X-chromosome, MHC region (chr6: 25,000,000 - 35,000,000BP), and the inversion on chromosome 8 (chr8: 7,000,000 - 15,000,000BP). As a result, a total of 8,983,436 SNPs shared between both summary statistics were used for the analysis. The results were compared in detail with the results reported by the cc-GWAS in [194].

3.7 Appendix notes

3.7.1 Solving the non-linear system of equations of Section 3.3.1

For notational simplicity, let $a = a_{i\ell}^{\text{cse}}$, $b = u_{i\ell}^{\text{cse}}$, $c = a_{i\ell}^{\text{cnt}}$, and $d = u_{i\ell}^{\text{cnt}}$. We rewrite eqns. (1)-(4) as

$$\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} = w, \text{ with } w = SE_{i\ell}^2, \tag{18}$$

$$a + b = x, \text{ with } x = 2N_\ell^{\text{cse}}, \tag{19}$$

$$c + d = y, \text{ with } y = 2N_\ell^{\text{cnt}}, \text{ and} \tag{20}$$

$$\frac{a \cdot d}{c \cdot b} = z, \text{ with } z = OR_{i\ell}. \tag{21}$$

Our goal is compute values for the four unknowns a, b, c, and d. Combining eqns. (19) and (20), we get

$$a = x - b, \text{ and} \tag{22}$$

$$c = y - d. \tag{23}$$

Substituting eqn. (22) and eqn. (23) into eqn. (21), we get $(x - b)d = zb(y - d)$, which can be rewritten as

$$b = \frac{xd}{yz - zd + d}. \tag{24}$$

Substituting eqn. (24) into eqn. (22), we get

$$a = x - \frac{xd}{yz - zd + d} = \frac{xyz - xzd}{yz - zd + d}. \tag{25}$$

We now note that all four unknowns can be written as functions of d and other known quantities. Substituting eqn. (23), eqn. (24), and eqn. (25) into eqn. (18), we get

$$\frac{1}{\frac{xyz-xzd}{yz-zd+d}} + \frac{1}{\frac{xd}{yz-zd+d}} + \frac{1}{y-d} + \frac{1}{d} = w.$$

Simplifying the above equation, we get

$$\frac{yz - zd + d}{xz(y - d)} + \frac{yz - zd + d}{xd} + \frac{1}{y - d} + \frac{1}{d} = w,$$

which can be further simplified to

$$(wxz + (1 - z)^2) \cdot d^2 + (2yz(1 - z) - wxyz) \cdot d + (yz(x + yz)) = 0. \tag{26}$$

Eqn. (26) is a quadratic equation on d; its real roots (if they exist) are

$$\{d_1, d_2\} = \frac{-(2yz(1 - z) - wxyz) \pm \sqrt{(2yz(1 - z) - wxyz)^2 - 4(wxz + (1 - z)^2)(yz(x + yz))}}{2(wxz + (1 - z)^2)}.$$

Given d, we can immediately compute a, b, and c using eqns. (23), (24), and (25). In order to determine whether d is equal to d_1 or d_2, we first check whether d_1 or d_2 guarantee that a, b, c, and d are all positive numbers. If both d_1 and d_2 satisfy this constraint, then we choose the *largest* of the two roots, as it solves the following trivial minimization problem:

$$\min_{d \in \{d_1, d_2\}} \frac{a+c}{a+b+c+d}.$$

The above choice is based on the assumption that in summary statistics A_1 (whose frequency is equal to the above fraction) typically denotes the effective (minor) allele. Additionally, our code performs a sanity check for allele alignment across studies given the solution d_1 or d_2.

For the sake of completeness, we also prove that it is not possible for both d_1 and d_2 to be negative. First, note that

$$d_1 + d_2 = -\frac{2yz(1-z) - wxyz}{wxz + (1-z)^2} = \frac{yz}{wxz + (1-z)^2} \cdot (wx - 2 + 2z). \tag{27}$$

Using $x = a + b > 0$ and $w = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} > \frac{1}{a} + \frac{1}{b} > 0$, we get

$$wx > (a+b) \cdot (\frac{1}{a} + \frac{1}{b}) = \frac{(a+b)^2}{ab} = \frac{a^2 + 2ab + b^2}{ab} > 2, \tag{28}$$

which implies that $wx - 2 + 2z > 0$. Combining with eqn. (27), we conclude that $d_1 + d + 2$ is non-negative; recall that w, x, y, and z are all non-negative. Additionally,

$$d_1 \cdot d_2 = \frac{yz(x + yz)}{wxz + (1-z)^2} > 0,$$

which implies that d_1 and d_2 must have the same sign, and since their sum is non-negative, they must both be positive. It is a simple exercise to prove that as long as root(s) exist, at least one of them will guarantee that all values for a, b and c will be positive.

One important exception arises when the discriminant in eqn. (26) is negative. In that case, no real roots exist for the quadratic equation. We do note that, theoretically, this should never happen, since the underlying unknown quantities are positive real numbers. However, stratification correction and genotype missingness could force the discriminant to fall below zero. To address this issue, we inflate w (i.e., the square of the standard error for the respective SNP) and recompute the discriminant. More specifically, we iteratively multiply w by 1.001 (a 0.1% inflation) until a non-negative discriminant is obtained or until 50 iterations are reached. The maximum inflation we allow (after the full 50 iterations) is $1.001^{50} - 1 \approx 5\%$. If after 50 iterations we have failed to find a non-negative discriminant we omit this particular SNP from further analyses. Empirically, for most input SNPs, a real root can be found after at most ten iterations.

3.7.2 Correction for sample overlap between the base/target studies for group PRS

The existence of shared samples in base (discovery) and target populations can lead to inflation in association between PRS and the trait of interest in the target population [205, 123]. In our case, such overlap will cause higher levels of significance in the t-test comparing the case and control PRS distribution. So far, for conventional PRS, the most widely accepted approach to address this problem is simply to identify the overlapping individuals and remove them from the target population. However, in practice, this is not always possible since it usually requires additional access to the individual level data of the base population. In this section, we introduce a correction for sample overlap between the base and target populations implemented in ReACt that could alleviate such issues.

In the following, we will use the case group as an example. Assume that the sample size for cases of the target population is $N^{\texttt{cse}}_{\texttt{target}}$, out of which $N^{\texttt{cse}}_{\texttt{shr}}$ are also cases in the base population (overlap). If the probability of each sample being shared between the base and target studies is uniformly distributed in both base and target studies, we would expect the observed mean PRS in target cases $\text{PRS}^{\texttt{cse}}_{\texttt{obs}}$ to be a weighted sum of the mean PRS in base cases $\text{PRS}^{\texttt{cse}}_{\texttt{base}}$ and the mean PRS of cases that only exist in the target population $\text{PRS}^{\texttt{cse}}_{\texttt{target}}$ as follows:

$$\text{PRS}^{\texttt{cse}}_{\texttt{obs}} = \frac{N^{\texttt{cse}}_{\texttt{shr}}}{N^{\texttt{cse}}_{\texttt{target}}} \cdot \text{PRS}^{\texttt{cse}}_{\texttt{base}} + \left(1 - \frac{N^{\texttt{cse}}_{\texttt{shr}}}{N^{\texttt{cse}}}\right) \cdot \text{PRS}^{\texttt{cse}}_{\texttt{target}}.$$

Therefore, the mean PRS for cases only in the target population is:

$$\text{PRS}^{\texttt{cse}}_{\texttt{target}} = \left(\text{PRS}^{\texttt{cse}}_{\texttt{obs}} - \frac{N^{\texttt{cse}}_{\texttt{shr}}}{N^{\texttt{cse}}_{\texttt{target}}} \text{PRS}^{\texttt{cse}}_{\texttt{base}}\right) \cdot \frac{N^{\texttt{cse}}_{\texttt{target}}}{N^{\texttt{cse}}_{\texttt{target}} - N^{\texttt{cse}}_{\texttt{shr}}},$$

where $\mathrm{PRS}^{\mathtt{cse}}_{\mathtt{obs}}$ is the uncorrected group mean computed as described in Section 3.5.3. $\mathrm{PRS}^{\mathtt{cse}}_{\mathtt{base}}$ can be obtained by simply setting the target population to be the same as the base population, using base summary statistics to compute group PRS for the target population. Similarly, we can adjust the variance computation as follows:

$$\mathrm{Var}(\mathrm{PRS}^{\mathtt{cse}}_{\mathtt{obs}}) = \left(\frac{N^{\mathtt{cse}}_{\mathtt{shr}}}{N^{\mathtt{cse}}_{\mathtt{target}}}\right)^2 \cdot \mathrm{Var}(\mathrm{PRS}^{\mathtt{cse}}_{\mathtt{base}}) + \left(1 - \frac{N^{\mathtt{cse}}_{\mathtt{shr}}}{N^{\mathtt{cse}}_{\mathtt{target}}}\right)^2 \cdot \mathrm{Var}(\mathrm{PRS}^{\mathtt{cse}}_{\mathtt{target}}). \tag{29}$$

Therefore, the corrected variance will be

$$\mathrm{Var}(\mathrm{PRS}^{\mathtt{cse}}_{\mathtt{target}}) = \left(\mathrm{Var}(\mathrm{PRS}^{\mathtt{cse}}_{\mathtt{obs}}) - \left(\frac{N^{\mathtt{cse}}_{\mathtt{shr}}}{N^{\mathtt{cse}}_{\mathtt{target}}}\right)^2 \cdot \mathrm{Var}(\mathrm{PRS}^{\mathtt{cse}}_{\mathtt{base}})\right) \cdot \left(\frac{N^{\mathtt{cse}}_{\mathtt{target}}}{N^{\mathtt{cse}}_{\mathtt{target}} - N^{\mathtt{cse}}_{\mathtt{shr}}}\right)^2 \tag{30}$$

Similarly,

$$\mathrm{PRS}^{\mathtt{cnt}}_{\mathtt{target}} = \left(\mathrm{PRS}^{\mathtt{cnt}}_{\mathtt{obs}} - \frac{N^{\mathtt{cnt}}_{\mathtt{shr}}}{N^{\mathtt{cnt}}_{\mathtt{target}}}\mathrm{PRS}^{\mathtt{cnt}}_{\mathtt{base}}\right) \cdot \frac{N^{\mathtt{cnt}}_{\mathtt{target}}}{N^{\mathtt{cnt}}_{\mathtt{target}} - N^{\mathtt{cnt}}_{\mathtt{shr}}} \tag{31}$$

and

$$\mathrm{Var}(\mathrm{PRS}^{\mathtt{cnt}}_{\mathtt{target}}) = \left(\mathrm{Var}(\mathrm{PRS}^{\mathtt{cnt}}_{\mathtt{obs}}) - \left(\frac{N^{\mathtt{cnt}}_{\mathtt{shr}}}{N^{\mathtt{cnt}}_{\mathtt{target}}}\right)^2 \cdot \mathrm{Var}(\mathrm{PRS}^{\mathtt{cnt}}_{\mathtt{base}})\right) \cdot \left(\frac{N^{\mathtt{cnt}}_{\mathtt{target}}}{N^{\mathtt{cnt}}_{\mathtt{target}} - N^{\mathtt{cnt}}_{\mathtt{shr}}}\right)^2 \tag{32}$$

for controls. Then, the corrected p-value will be based on a t-test using the corrected mean and variance and an adjusted degree of freedom:

$$df_{\mathtt{target}} = N^{\mathtt{cnt}}_{\mathtt{target}} + N^{\mathtt{cse}}_{\mathtt{target}} - (N^{\mathtt{cnt}}_{\mathtt{shr}} + N^{\mathtt{cse}}_{\mathtt{shr}}) - 2.$$

This is a straightforward correction on the target PRS using the scores of the base population that one would use if there were no stratification between the base and target populations. In practice, this idealized scenario does not hold. In order to deal with the stratification between the base and target populations, prior to any correction, we shift the scores for base cases and controls by aligning the base population PRS means to the target population as follows:

$$\begin{aligned}
\mathrm{PRS}^{\mathtt{cse*}}_{\mathtt{base}} &= \mathrm{PRS}^{\mathtt{cse}}_{\mathtt{base}} - (\mathrm{PRS}_{\mathtt{base}} - \mathrm{PRS}_{\mathtt{target}}),\\
\mathrm{PRS}^{\mathtt{cnt*}}_{\mathtt{base}} &= \mathrm{PRS}^{\mathtt{cnt}}_{\mathtt{base}} - (\mathrm{PRS}_{\mathtt{base}} - \mathrm{PRS}_{\mathtt{target}}).
\end{aligned}$$

In the above, $\mathrm{PRS}_{\mathtt{base}}$ and $\mathrm{PRS}_{\mathtt{target}}$ are mean PRS for the base and target populations respectively:

$$\begin{aligned}
\mathrm{PRS}_{\mathtt{base}} &= \frac{N^{\mathtt{cnt}}_{\mathtt{base}} \cdot \mathrm{PRS}^{\mathtt{cnt}}_{\mathtt{base}} + N^{\mathtt{cnt}}_{\mathtt{base}} \cdot \mathrm{PRS}^{\mathtt{cnt}}_{\mathtt{base}}}{N^{\mathtt{cse}}_{\mathtt{base}} + N^{\mathtt{cnt}}_{\mathtt{base}}},\\
\mathrm{PRS}_{\mathtt{target}} &= \frac{N^{\mathtt{cnt}}_{\mathtt{target}} \cdot \mathrm{PRS}^{\mathtt{cnt}}_{\mathtt{target}} + N^{\mathtt{cnt}}_{\mathtt{target}} \cdot \mathrm{PRS}^{\mathtt{cnt}}_{\mathtt{target}}}{N^{\mathtt{cse}}_{\mathtt{target}} + N^{\mathtt{cnt}}_{\mathtt{target}}}.
\end{aligned}$$

In practice, we use $\mathrm{PRS}^{\mathtt{cse*}}_{\mathtt{base}}$ and $\mathrm{PRS}^{\mathtt{cnt*}}_{\mathtt{base}}$ instead of $\mathrm{PRS}^{\mathtt{cse}}_{\mathtt{base}}$ and $\mathrm{PRS}^{\mathtt{cnt}}_{\mathtt{base}}$ in equations (29)-(32) for correction. We evaluated the performance of this correction scheme by introducing sample overlaps between the base and target populations using the same simulation model as the one we used to evaluate the performance of our group PRS approach. We computed the real individual level PRS using PRSice2, from which we obtained the inflated PRS descriptive statistics (group mean, standard deviation, and t-test p-value) for all target samples, including the ones that are shared with the base population. We also computed PRS statistics for samples that are present only in the target population as the ground truth. We compared results from our corrected group PRS method to the PRS statistics for the samples that are exclusive to the target population computed using PRSice2. Results on synthetic data demonstrated that our correction can drastically alleviate the inflation in p-values that is the result of sample overlap the between base and target populations. See Table 3.17, which shows representative results from our experimental evaluations. If the number of overlapping samples is unknown to the user, we apply the approach proposed in [308] to get an estimate of the overlapping sample size and we correct the output statistics accordingly. Note that this correction approach is based on the assumption that all samples having an equal probability of being shared between the base and target populations, which might be unrealistic in certain settings.

3.7.3 Speeding up the logistic regression computation

Recall that in section 3.5.2 for any SNP i, if we try to formulate the computation of elements in $\boldsymbol{H}$ and $\boldsymbol{G}$ in one iteration, they will be:

$$\boldsymbol{H}_{uv} = \sum_{j=1}^{N} d_j \cdot \boldsymbol{X}_{ju}\boldsymbol{X}_{jv} \tag{33}$$

and

$$\boldsymbol{G}_{u} = \sum_{j=1}^{N} d_j \cdot z_j \cdot \boldsymbol{X}_{ju} \tag{34}$$

where $u, v \in \{0, 1, 2\}$.

Same as in section 3.5.2, we dropped the subscript i from d_j, z_j and $\boldsymbol{X}$. If we follow these equations, when the sample size $\sum_{\ell=1}^{L} N_\ell$ increases, the computational burden will increase linearly. However, in practice this step can be achieved with an $O(L)$ complexity, as long as we take advantage of the fact that all elements of $\boldsymbol{X}$ are discrete values involving only 0, 1, 2 and study indicators $I_{i\ell}$. This indicates that both $d_j \cdot \boldsymbol{X}_{ju}\boldsymbol{X}_{jv}$ and $d_j \cdot z_j \cdot \boldsymbol{X}_{ju}$ can only take a few possible values. In fact, since there are only $3 \cdot L$ possibilities for $\boldsymbol{X}_{j*}$ (3 different genotypes $\cdot$ L different studies), there are also only 6 possible values for d_j. We denote them as $d_{\ell\mathtt{n}}$, with $\ell \in \{1, \ldots, L\}$ and $\mathtt{n} \in \{0, 1, 2\}$. Therefore, as an example, $d_{1\mathtt{0}}$ will be the value of d_j for a sample j if it belongs to study 1 and has a genotype of A2A2. Similarly, for z_j, since y_j is involved in this computation, we need to consider in total $3L \cdot 2 = 6L$ possible values as $\boldsymbol{y}$ is binary indicator for the phenotypes. We denote those $6 \cdot L$ possible values as $z_{\ell\mathtt{n}}^{\mathtt{cse}}$ for cases and $z_{\ell\mathtt{n}}^{\mathtt{cnt}}$ for controls respectively. Then $z_{1\mathtt{0}}^{\mathtt{cnt}}$ will represent the value of z_j for a sample j if it belongs to study 1, has a genotype of A2A2 and meanwhile is a control. Then we only need to plug in the element of $\boldsymbol{X}$ based on the u, v values of interest.

Noticing this, if we just count the occurrence of those values, the summation can be found out easily. This can be done using the genotype counts that we have already computed in section 3.5.1. Therefore, for any SNP i with occurrence $N_{i\ell}^{\text{cnt}}(n)$ of each genotype n and indicator $I_{i\ell}$ for each input study, in an iteration of the IRLS, we can compute all $d_{\ell\mathtt{n}}$ and $z_{\ell\mathtt{n}}^{\mathtt{cnt}}$ needed for this SNP as described in 1. Then for this iteration, we shall have:

$$\boldsymbol{H}_{00} = \sum_{\ell=1}^{L} \sum_{n=0}^{2} d_{\ell\mathtt{n}} \cdot (N_{i\ell}^{\text{cnt}}(n) + N_{i\ell}^{\text{cse}}(n)) \tag{35}$$

$$\boldsymbol{H}_{01} = \boldsymbol{H}_{10} = \sum_{\ell=1}^{L} \sum_{n=0}^{2} n \cdot d_{\ell\mathtt{n}} \cdot (N_{i\ell}^{\text{cnt}}(n) + N_{i\ell}^{\text{cse}}(n)) \tag{36}$$

$$\boldsymbol{H}_{02} = \boldsymbol{H}_{20} = \sum_{\ell=1}^{L} \sum_{n=0}^{2} I_{i\ell} \cdot d_{\ell\mathtt{n}} \cdot (N_{i\ell}^{\text{cnt}}(n) + N_{i\ell}^{\text{cse}}(n)) \tag{37}$$

$$\boldsymbol{H}_{11} = \sum_{\ell=1}^{L} \sum_{n=0}^{2} n^2 \cdot d_{\ell\mathtt{n}} \cdot (N_{i\ell}^{\text{cnt}}(n) + N_{i\ell}^{\text{cse}}(n)) \tag{38}$$

$$\boldsymbol{H}_{12} = \boldsymbol{H}_{21} = \sum_{\ell=1}^{L} \sum_{n=0}^{2} n \cdot I_{i\ell} \cdot d_{\ell\mathtt{n}} \cdot (N_{i1}^{\text{cnt}}(n) + N_{i1}^{\text{cse}}(n)) \tag{39}$$

$$\boldsymbol{H}_{22} = \sum_{\ell=1}^{L} \sum_{n=0}^{2} I_{i\ell}^2 \cdot d_{\ell\mathtt{n}} \cdot (N_{i1}^{\text{cnt}}(n) + N_{i1}^{\text{cse}}(n)) \tag{40}$$

and

$$\boldsymbol{G}_{0} = \sum_{\ell=1}^{L} \sum_{n=0}^{2} d_{\ell\mathtt{n}} \cdot (z_{\ell\mathtt{n}}^{\mathtt{cnt}} \cdot N_{i\ell}^{\text{cnt}}(n) + z_{\ell\mathtt{n}}^{\mathtt{cse}} \cdot N_{i\ell}^{\text{cse}}(n)) \tag{41}$$

$$\boldsymbol{G}_{1} = \sum_{\ell=1}^{L} \sum_{n=0}^{2} n \cdot d_{\ell\mathtt{n}} \cdot (z_{\ell\mathtt{n}}^{\mathtt{cnt}} \cdot N_{i\ell}^{\text{cnt}}(n) + z_{\ell\mathtt{n}}^{\mathtt{cse}} \cdot N_{i\ell}^{\text{cse}}(n)) \tag{42}$$

$$\boldsymbol{G}_{2} = \sum_{\ell=1}^{L} \sum_{n=0}^{2} I_{i\ell} \cdot d_{\ell\mathtt{n}} \cdot (z_{\ell\mathtt{n}}^{\mathtt{cnt}} \cdot N_{i\ell}^{\text{cnt}}(n) + z_{\ell\mathtt{n}}^{\mathtt{cse}} \cdot N_{i\ell}^{\text{cse}}(n)) \tag{43}$$

Eqn. (35)-(43) grant us fast update of $\boldsymbol{w} = \boldsymbol{H}^{-1}\boldsymbol{G}$ in each iteration. We can do this repeatedly in the IRLS until convergence to get the final result $\widetilde{\boldsymbol{w}}$.

Acknowledgements

This work was supported by National Science Foundation. This research has been conducted using the UK Biobank Resource under Application Number 61553. We would also like to thank our colleague Pritesh Jain for contributing the name ReAct for the tool being developed.

3.8 Tables

Table 3.7: **Performance of fixed-effect meta-analysis on real genotype data.** We applied our method for fixed-effect meta-analysis to the depressive episode trait (ICD F32 Depressive episode) in UK biobank samples and compared the performance of our method vs. ASSET/METAL. SNPs with p-value strictly less than 10^{-6} in the primary GWAS summary statistics using all samples were treated as "true signals". In each iteration of an experiment, we split the dataset evenly into two, generated GWAS summary statistics for each subset, and meta-analyzed the summary statistics using our method and ASSET/METAL. We reported the number of times (out of ten iterations) that a "true signal" got captured using the "significance threshold" $p < 10^{-6}$ by each method under different sample overlap conditions. METAL dev refers to the latest release in GitHub [308]. Two variants of ReACt are tested: Exact and Est, indicating whether the sample overlap was *exactly* known as part of the input or whether it was *estimated*, respectively. Sample overlap indicates the number of cases and controls that were shared between two input studies, ie., 500 sample overlap means that 500 cases **and** 500 controls were shared between the two studies when the split was carried out. The variable P in the table indicates the p-value of the target SNP in the primary GWAS using all samples. *True positive per iteration* reports the average number of SNPs with p-value strictly less than 10^{-6} in the primary GWAS that were captured in one iteration; and *False positive per iteration* reports the average number of extra SNPs being captured in one iteration.

SNP	P	number of times the SNP had p-value $< 10^{-5}$ in meta-analysis									
		no sample overlap[a]		500 sample overlap[b]				1000 sample overlap[c]			
		Exact	ASSET/METAL	Exact	Est.	ASSET	METAL	Exact	Est.	ASSET	METAL
rs60939828	$2.77 \cdot 10^{-9}$	10	10	10	10	10	10	10	10	10	10
rs17487484	$2.61 \cdot 10^{-8}$	10	10	10	10	10	10	10	10	10	10
rs62100766	$1.55 \cdot 10^{-7}$	10	10	9	9	8	9	9	4	4	9
rs4510098	$5.34 \cdot 10^{-7}$	10	10	5	5	5	5	5	4	3	5
rs1079232	$6.69 \cdot 10^{-7}$	2	2	5	4	3	5	3	2	2	3
rs75056899	$7.69 \cdot 10^{-7}$	10	10	3	3	3	3	4	4	4	4
rs12044988	$7.75 \cdot 10^{-7}$	10	10	5	1	1	5	6	4	3	6
True positive per iteration		6.2	6.2	4.7	4.2	4	4.7	4.7	3.8	3.6	4.7
False positive per iteration		0.2	0.2	1.4	0.6	0.4	1.5	1.6	0.5	0.7	1.7

[a] with 9,184 cases and 156,425 controls from each subset

[b] out of 9,434 cases and 156,675 controls from each subset

[c] out of 9,684 cases and 156,925 controls from each subset

Table 3.8: **Estimated and real group mean and standard deviation of PRS for a synthetic target population.** We compared group mean and standard deviation of PRS estimated by ReACt from summary statistics of synthetic base and target studies to the real group mean and standard deviation of individual level PRS obtained using summary statistics of the base and individual level genotype of the target computed by PRSice2. Est stands for estimated. Note that the synthetic data is not subject to clumping since the simulation model does not generate LD structure.

risk	group	`Our Method (ReACt)`		`PRSice2`	
		est. group mean	est. group sd	real group mean	real group sd
1.15	cases	0.0009	0.0078	0.0009	0.0076
	controls	-0.0037	0.0078	-0.0036	0.0081
1.2	cases	0.0016	0.0060	0.0016	0.0059
	controls	-0.0065	0.0060	-0.0064	0.0061
1.3	cases	0.0021	0.0041	0.0021	0.0040
	controls	-0.0125	0.0041	-0.0125	0.0040

Table 3.9: **Estimated and real group mean and standard deviation of PRS for depressive episode cases and controls in UK biobank population.** We assessed the performance of our method using the summary statistics of an independent MDD GWAS as the base study, and the UK biobank samples, including 18,368 cases with depressive episode and 312,849 controls, as the target population. We generated summary statistics for the target populations and estimated group mean PRS and standard deviation of target PRS using ReACt. We computed the individual level PRS for the target study using PRSice2. For both methods, we computed PRS using independent SNPs from the base summary statistics with p-values below various thresholds (P-thres) and compared the performances under each threshold. For ReACt, mean PRS represents the estimated group mean PRS for cases and controls; p-val are the t-test p-values comparing PRS distribution in cases and in controls. For PRSice2, mean PRS represents real group mean PRS computed from individual level data and p-val are the t-test p-values comparing real PRS distribution in cases and in controls; reg. w/o covariate indicates regression results without covariates, which include the regression r^2 value (reg. r^2) and the p-value for the PRS predictor (p-val); reg. w/ top 5PCs indicates the regression results including the top five PCs as covariate, , which also included the regression r^2 value (reg. r^2) and the p-value for the PRS predictor (p-val).

			Our method (`ReACt`)		`PRSice2`					
P-thres	#SNPs	trait	t-test		t-test		reg. w/o covatiate		reg. w/ top 5PCs	
			mean PRS	p-val	mean PRS	p-val	r^2	p-val	r^2	p-val
0.1	4236	cases	-0.0023	$5.50{\cdot}10^{-3}$	-0.0023	$3.97 \cdot 10^{-3}$	$2.48 \cdot 10^{-5}$	$4.18 \cdot 10^{-3}$	$3.54 \cdot 10^{-5}$	$4.14 \cdot 10^{-3}$
		controls	-0.0023		-0.0024					
0.01	594	cases	-0.0036	$1.47{\cdot}10^{-3}$	-0.0032	$1.42 \cdot 10^{-3}$	$3.06 \cdot 10^{-5}$	$1.45 \cdot 10^{-3}$	$4.35 \cdot 10^{-5}$	$1.44 \cdot 10^{-3}$
		controls	-0.0036		-0.0032					
0.001	82	cases	0.0112	$1.09{\cdot}10^{-1}$	0.0147	$1.54 \cdot 10^{-1}$	$6.17 \cdot 10^{-6}$	$1.53 \cdot 10^{-1}$	$3.19 \cdot 10^{-5}$	$1.51 \cdot 10^{-1}$
		controls	0.0112		0.0146					
10^{-4}	10	cases	-0.0244	$9.36{\cdot}10^{-2}$	-0.0247	$1.16 \cdot 10^{-1}$	$7.57 \cdot 10^{-6}$	$1.13 \cdot 10^{-1}$	$2.96 \cdot 10^{-5}$	$1.12 \cdot 10^{-1}$
		controls	-0.0246		-0.0249					

Table 3.10: **Using our method to perform PRS comparisons across eight neuropsychiatric disorders.** We further applied our method to the summary statistics of eight neuropsychiatric disorders from PGC (see table 3.6 for details). For each disorder, we used PGC GWAS summary statistics to compute the group mean and standard deviation of PRS for the other seven disorders. All group PRS were estimated using independent SNPs with $p < 10^{-5}$in the base summary statistics. We report p-values from a t-test comparing the group mean PRS of cases against controls in the target study, and cells with deeper blue colors correspond to lower p-values. The threshold of significance under multiple testing correction is $p < 8.93 \cdot 10^{-4}$.

		Target							
		OCD	TS	ED	ASD	BIP	ADHD	SCZ	MD
Base	OCD	-	$5.71{\cdot}10^{-1}$	$1.26{\cdot}10^{-1}$	$7.83{\cdot}10^{-2}$	$9.51{\cdot}10^{-2}$	$2.64{\cdot}10^{-1}$	$4.44{\cdot}10^{-1}$	$6.81{\cdot}10^{-1}$
	TS	$5.17{\cdot}10^{-2}$	-	$2.31{\cdot}10^{-1}$	$7.78{\cdot}10^{-1}$	$3.05{\cdot}10^{-1}$	$3.57{\cdot}10^{-2}$	$4.50{\cdot}10^{-1}$	$5.40{\cdot}10^{-3}$
	ED	$2.95{\cdot}10^{-1}$	$3.31{\cdot}10^{-1}$	-	$4.83{\cdot}10^{-1}$	$4.29{\cdot}10^{-4}$	$6.28{\cdot}10^{-4}$	$1.89{\cdot}10^{-2}$	$3.27{\cdot}10^{-3}$
	ASD	$9.95{\cdot}10^{-1}$	$7.40{\cdot}10^{-3}$	$9.00{\cdot}10^{-1}$	-	$1.77{\cdot}10^{-1}$	$8.12{\cdot}10^{-4}$	$1.17{\cdot}10^{-1}$	$3.98{\cdot}10^{-13}$
	BIP	$3.54{\cdot}10^{-3}$	$5.82{\cdot}10^{-1}$	$9.84{\cdot}10^{-13}$	$4.03{\cdot}10^{-7}$	-	$1.29{\cdot}10^{-13}$	$1.08{\cdot}10^{-79}$	$1.15{\cdot}10^{-19}$
	ADHD	$2.15{\cdot}10^{-1}$	$1.08{\cdot}10^{-8}$	$2.32{\cdot}10^{-3}$	$2.62{\cdot}10^{-45}$	$9.58\ {\cdot}10^{-2}$	-	$1.37{\cdot}10^{-10}$	$2.88{\cdot}10^{-52}$
	SZC	$3.23{\cdot}10^{-7}$	$9.36{\cdot}10^{-1}$	$4.88{\cdot}10^{-1}$	$1.28{\cdot}10^{-24}$	$1.68{\cdot}10^{-133}$	$2.11{\cdot}10^{-1}$	-	$7.36{\cdot}10^{-94}$
	MD	$5.09{\cdot}10^{-2}$	$4.48{\cdot}10^{-1}$	$3.43{\cdot}10^{-1}$	$2.08{\cdot}10^{-26}$	$5.35{\cdot}10^{-9}$	$6.05{\cdot}10^{-21}$	$6.10{\cdot}10^{-45}$	-

Table 3.11: **Performance of cc-GWAS as implemented in `ReACt` with different sample sizes.** Three types of SNPs have been simulated: *(i)* trait differential SNPs; *(ii)* null SNPs; and *(iii)* stress SNPs. . Under each condition, we simulated individual level genotype with these three types of SNPs for N cases and N controls in each study ($N = 2,000$ and $N = 5,000$) and generated GWAS summary statistics for each study. and generated GWAS summary statistics for each study respectively. We subsequently used the summary statistics to run cc-GWAS in ReACt. We reported the power for detecting type *(i)* SNPs, and false positive rates for picking up type *(ii)* SNPs (Type I err.$^{\texttt{(ii)}}$) and type *(iii)* SNPs (Type I err.$^{\texttt{(iii)}}$) under a significance threshold $p < 5 \cdot 10^{-5}$.

risk	Fst	2,000 cases, 2,000 controls			5,000 cases, 5,000 controls		
		Power	Type I err.$^{\texttt{(ii)}}$	Type I err.$^{\texttt{(iii)}}$	Power	Type I err.$^{\texttt{(ii)}}$	Type I err.$^{\texttt{(iii)}}$
1.15	0.01	$3.67{\cdot}10^{-2}$	$2.65{\cdot}10^{-5}$	$3.16{\cdot}10^{-4}$	$3.51{\cdot}10^{-1}$	$1.84{\cdot}10^{-5}$	$1.87{\cdot}10^{-4}$
	0.05	$3.49{\cdot}10^{-2}$	$9.80{\cdot}10^{-5}$	$5.26{\cdot}10^{-4}$	$3.23{\cdot}10^{-1}$	$6.33{\cdot}10^{-5}$	$3.58{\cdot}10^{-4}$
	0.1	$2.81{\cdot}10^{-2}$	$2.43{\cdot}10^{-4}$	$5.02{\cdot}10^{-4}$	$2.85{\cdot}10^{-1}$	$1.94{\cdot}10^{-4}$	$5.21{\cdot}10^{-4}$
1.2	0.01	$1.54{\cdot}10^{-1}$	$4.69{\cdot}10^{-5}$	$2.47{\cdot}10^{-4}$	$7.16{\cdot}10^{-1}$	$3.47{\cdot}10^{-5}$	$2.03{\cdot}10^{-4}$
	0.05	$1.34{\cdot}10^{-1}$	$1.04{\cdot}10^{-4}$	$5.14{\cdot}10^{-4}$	$6.62{\cdot}10^{-1}$	$8.57{\cdot}10^{-5}$	$3.77{\cdot}10^{-4}$
	0.1	$1.23{\cdot}10^{-1}$	$2.33{\cdot}10^{-4}$	$5.83{\cdot}10^{-4}$	$6.03{\cdot}10^{-1}$	$1.65{\cdot}10^{-4}$	$5.27{\cdot}10^{-4}$
1.3	0.01	$5.85{\cdot}10^{-1}$	$1.63{\cdot}10^{-5}$	$1.57{\cdot}10^{-4}$	$9.68{\cdot}10^{-1}$	$1.43{\cdot}10^{-5}$	$5.46{\cdot}10^{-4}$
	0.05	$5.41{\cdot}10^{-1}$	$5.31{\cdot}10^{-5}$	$4.45{\cdot}10^{-4}$	$9.21{\cdot}10^{-1}$	$7.35{\cdot}10^{-5}$	$5.79{\cdot}10^{-4}$
	0.1	$4.85{\cdot}10^{-1}$	$2.63{\cdot}10^{-4}$	$6.18{\cdot}10^{-4}$	$8.71{\cdot}10^{-1}$	$1.67{\cdot}10^{-4}$	$6.84{\cdot}10^{-4}$

Table 3.12: **Comparison of genomic regions showing significant divergent genetic effects between BD and SCZ as detected by ReACt and ccGWAS by Peyrot et al [194].** We carried out cc-GWAS with ReACt using summary statistics of BD and SCZ and compared our results with the results from Peyrot et al. Only SNPs that are analyzed in both studies are included for the comparison. Genomic regions that are identified to show significant divergent genetic effects between BD and SCZ in either result are shown. CHR, Start and End are chromosomal and base-pair ranges for the region; SNP, BP and p-value (ordinary least squares p-values, P_{OLS}, for ccGWAS by Peyrot et al.) are properties of the leading SNP (if the regions is reported genomewide significant) or statistics for the matching SNP (if the region is not reported as genomewide significant, but is detected by the other method); p-values in red are leading SNPs that are reported genomewide significant by each method; Regions with CHR, Start and End in red are two loci that were also identified by the case-case GWAS using individual level data [236].

	Region		Our method (`ReACt`)			`ccGWAS`		
CHR	Start	End	SNP	BP	p-value	SNP	BP	p-value(P_{OLS})
1	50826176	51118253	rs6682989	50826176	$3.08 \cdot 10^{-8}$	-	-	$6.10 \cdot 10^{-7}$
1	98325796	98559093	rs2660304	98512127	$4.20 \cdot 10^{-9}$	-	-	$2.20 \cdot 10^{-9}$
1	173867252	174643725	rs6701877	174015259	$4.02 \cdot 10^{-8}$	-	-	$5.80 \cdot 10^{-10}$
2	27498734	27752296	rs113954968	27696207	$2.93 \cdot 10^{-8}$	-	-	$1.10 \cdot 10^{-6}$
3	62563175	62583180	rs1993149	62572944	$2.10 \cdot 10^{-8}$	-	-	$8.10 \cdot 10^{-7}$
3	135807609	136597120	rs9866687	94828190	$6.55 \cdot 10^{-7}$	-	-	$4.00 \cdot 10^{-8}$
3	135807609	136597120	rs7372313	135872958	$1.02 \cdot 10^{-8}$	rs1278493	135814009	$1.20 \cdot 10^{-8}$
7	28453906	28484317	rs2192303	28478332	$3.57 \cdot 10^{-8}$	rs7790864	28478625	$2.20 \cdot 10^{-8}$
8	27406353	27453579	rs11778040	27419807	$5.39 \cdot 10^{-7}$	-	-	$4.80 \cdot 10^{-8}$
9	23345347	23362311	rs12554512	23352293	$3.58 \cdot 10^{-10}$	-	-	$4.10 \cdot 10^{-8}$
9	36894685	36963222	rs2039142	36963222	$1.95 \cdot 10^{-8}$	-	-	$2.10 \cdot 10^{-6}$
10	353306	418676	rs35198327	354301	$7.69 \cdot 10^{-9}$	-	-	$1.10 \cdot 10^{-7}$
12	108596308	108633649	rs3764002	108618630	$3.28 \cdot 10^{-9}$	-	-	$6.30 \cdot 10^{-11}$
12	110294902	111212762	rs28637922	110819139	$5.11 \cdot 10^{-10}$	-	-	$8.10 \cdot 10^{-12}$
16	79386766	79463881	rs6564668	79457393	$1.86 \cdot 10^{-8}$	rs9319540	79458022	$3.70 \cdot 10^{-8}$
19	1812521	1866427	rs1054972	1852582	$6.43 \cdot 10^{-8}$	-	-	$1.80 \cdot 10^{-8}$
20	47511792	47938833	rs6095394	47625544	$1.43 \cdot 10^{-9}$	rs11696888	47753265	$1.40 \cdot 10^{-9}$

Table 3.13: **Average running time in seconds for fixed effect meta-analysis for ReACt, METAL, and ASSET.** All experiments were performed at Purdue's Snyder cluster on a dedicated node which features a Haswell processor running at 2.6 GHz with 512 GB of RAM and a 64-bit CentOS Linux 7 operating system. We report average running time in seconds over ten iterations using `ReACt`, `METAL`, and `ASSET`. In the case of METAL we evaluated the performance of the latest release in GitHub [308]. In each iteration, two or four sets of summary statistics (for 100,000 SNPs) were meta-analyzed. Recall that all methods scale as a function of the number of SNPs and is independent of the number of samples, since only summary statistics are used.

	`ReACt`	`METAL`	`ASSET`
2 input studies	2.2s	1.8s	696s
4 input studies	3.1s	3.3s	3715s

Table 3.14: **Performance of fixed-effect meta-analysis with two input studies with uneven case/control sample sizes under different conditions.** We compare power and type I error rate (T1E) of our method meta-analyzing two studies with uneven case/control sample sizes vs. ASSET/METAL for a significance threshold $p < 5 \cdot 10^{-5}$. Study one contains 1500 cases and 500 controls, and study two contains 500 cases and 1500 controls.

risk	Fst	ReACt		METAL/ASSET	
		Power	T1E	Power	T1E
1.15	0.01	4.89E-02	4.24E-05	4.97E-02	4.65E-05
	0.05	5.07E-02	4.65E-05	5.13E-02	4.95E-05
	0.1	4.35E-02	4.04E-05	4.37E-02	4.55E-05
1.2	0.01	1.79E-01	4.75E-05	1.80E-01	5.05E-05
	0.05	1.66E-01	6.36E-05	1.67E-01	6.77E-05
	0.1	1.64E-01	4.44E-05	1.65E-01	4.55E-05
1.3	0.01	6.28E-01	4.24E-05	6.30E-01	4.44E-05
	0.05	5.99E-01	4.85E-05	6.00E-01	4.55E-05
	0.1	5.63E-01	4.65E-05	5.64E-01	4.85E-05

Table 3.15: **Performance of fixed-effect meta-analysis with two input studies under different conditions.** We compare power and type I error rate (T1E) of our method meta-analyzing two studies vs. ASSET/METAL for a significance threshold $p < 5 \cdot 10^{-5}$. METAL dev refers to the latest release in GitHub [308]. Two variants of ReACt are tested: Exact and Est, indicating whether the sample overlap was *exactly* known as part of the input or whether it was *estimated*, respectively. Sample overlap indicates the number of cases and controls that were shared between two input studies. I.e. a sample overlap equal to 100 means that there are 100 cases **and** 100 controls shared between two input studies. Total sample sizes for each input study, including the shared samples, are equal to 2000 when the sample overlap is equal to zero; 2400 when the sample overlap is equal to 100; and 4000 when the sample overlap is equal to 500. In each case, the sample is equally split to cases and controls. Also see figure 3.1 and 3.2.

risk	Fst	overlap	`ASSET`		`ReACt` (Exact)		`ReACt` (Est.)		`METAL` (dev)	
			Power	T1E	Power	T1E	Power	T1E	Power	T1E
1.15	0.01	0	1.04E-01	4.95E-05	1.03E-01	4.85E-05	-	-	1.04E-01	4.95E-05
		100	1.13E-01	4.34E-05	1.27E-01	5.25E-05	1.30E-01	4.85E-05	1.31E-01	5.15E-05
		500	1.69E-01	1.11E-05	2.79E-01	4.75E-05	2.80E-01	4.85E-05	2.80E-01	4.65E-05
	0.05	0	9.66E-02	5.25E-05	9.31E-02	5.25E-05	-	-	9.66E-02	5.25E-05
		100	9.68E-02	3.43E-05	1.19E-01	4.14E-05	1.17E-01	4.14E-05	1.17E-01	4.65E-05
		500	1.53E-01	4.04E-06	2.68E-01	3.84E-05	2.69E-01	3.74E-05	2.67E-01	3.74E-05
	0.1	0	8.65E-02	4.34E-05	8.19E-02	4.04E-05	-	-	8.65E-02	4.34E-05
		100	7.75E-02	3.33E-05	1.05E-01	4.44E-05	1.09E-01	4.65E-05	1.08E-01	5.15E-05
		500	1.24E-01	9.09E-06	2.39E-01	4.65E-05	2.42E-01	4.95E-05	2.41E-01	5.15E-05
1.2	0.01	0	3.21E-01	3.84E-05	3.18E-01	3.74E-05	-	-	3.21E-01	3.84E-05
		100	3.41E-01	3.54E-05	3.82E-01	4.04E-05	3.85E-01	4.04E-05	3.85E-01	4.14E-05
		500	4.95E-01	7.07E-06	6.44E-01	4.04E-05	6.47E-01	4.24E-05	6.46E-01	4.14E-05
	0.05	0	3.13E-01	4.24E-05	3.06E-01	3.94E-05	-	-	3.13E-01	4.24E-05
		100	2.96E-01	4.65E-05	3.59E-01	5.35E-05	3.66E-01	5.35E-05	3.65E-01	5.76E-05
		500	4.47E-01	8.08E-06	6.09E-01	4.85E-05	6.14E-01	5.15E-05	6.11E-01	5.25E-05
	0.1	0	2.83E-01	4.85E-05	2.71E-01	4.44E-05	-	-	2.83E-01	4.85E-05
		100	2.45E-01	4.44E-05	3.28E-01	4.34E-05	3.27E-01	4.55E-05	3.23E-01	4.55E-05
		500	3.95E-01	8.08E-06	5.76E-01	4.75E-05	5.83E-01	4.85E-05	5.80E-01	4.65E-05
1.3	0.01	0	8.00E-01	3.23E-05	7.99E-01	3.23E-05	-	-	8.00E-01	3.23E-05
		100	6.80E-01	3.84E-05	7.36E-01	4.65E-05	7.43E-01	5.15E-05	7.42E-01	5.45E-05
		500	4.90E-01	4.04E-06	6.40E-01	2.42E-05	6.98E-01	5.35E-05	6.97E-01	5.05E-05
	0.05	0	7.82E-01	4.95E-05	7.77E-01	4.44E-05	-	-	7.82E-01	4.95E-05
		100	6.32E-01	3.94E-05	7.48E-01	4.55E-05	7.55E-01	5.25E-05	7.52E-01	5.45E-05
		500	4.99E-01	1.01E-06	6.67E-01	1.31E-05	7.18E-01	4.04E-05	7.16E-01	3.64E-05
	0.1	0	7.32E-01	4.95E-05	7.20E-01	4.44E-05	-	-	7.32E-01	4.95E-05
		100	6.01E-01	3.84E-05	7.67E-01	4.24E-05	7.71E-01	4.65E-05	7.62E-01	5.15E-05
		500	5.49E-01	1.01E-06	7.30E-01	1.31E-05	7.67E-01	3.43E-05	7.63E-01	3.94E-05

Table 3.16: **Performance of fixed-effect meta-analysis with four input studies under different conditions.** We compare power and type I error rate (T1E) of our method meta-analyzing four studies vs. ASSET/METAL for a significance threshold $p < 5 \cdot 10^{-5}$. METAL dev refers to the latest release in GitHub [308]. Two variants of ReACt are tested: Exact and Est, indicating whether the sample overlap was *exactly* known as part of the input or whether it was *estimated*, respectively. Sample overlap indicates the number of cases and controls that were shared between two input studies. I.e. a sample overlap equal to 100 means that there are 100 cases **and** 100 controls shared between two input studies. Total sample sizes for each input study, including the shared samples, are equal to 2000 when the sample overlap is equal to zero; 2400 when the sample overlap is equal to 100; and 4000 when the sample overlap is equal to 500. In each case, the sample is equally split to cases and controls.

risk	Fst	overlap	ASSET		ReACt (Exact)		ReACt (Est.)		METAL (dev)	
			Power	T1E	Power	T1E	Power	T1E	Power	T1E
1.15	0.01	0	4.31E-01	4.75E-05	4.31E-01	4.75E-05	-	-	4.31E-01	4.75E-05
		100	3.19E-01	2.93E-05	4.00E-01	5.15E-05	4.03E-01	5.45E-05	4.03E-01	4.85E-05
		500	2.36E-01	1.01E-06	5.20E-01	4.85E-05	5.27E-01	5.25E-05	5.23E-01	4.85E-05
	0.05	0	4.13E-01	4.34E-05	4.08E-01	4.24E-05	-	-	4.13E-01	4.34E-05
		100	2.49E-01	3.33E-05	3.83E-01	5.25E-05	3.85E-01	5.66E-05	3.78E-01	5.56E-05
		500	2.06E-01	2.02E-06	5.03E-01	5.56E-05	5.14E-01	6.46E-05	5.04E-01	5.25E-05
	0.1	0	3.72E-01	5.35E-05	3.64E-01	4.85E-05	-	-	3.72E-01	5.35E-05
		100	1.90E-01	2.42E-05	3.46E-01	4.55E-05	3.53E-01	5.66E-05	3.41E-01	5.45E-05
		500	1.60E-01	2.02E-06	4.56E-01	5.15E-05	4.66E-01	5.45E-05	4.61E-01	5.35E-05
1.2	0.01	0	7.87E-01	5.15E-05	7.85E-01	5.15E-05	-	-	7.87E-01	5.15E-05
		100	6.48E-01	4.14E-05	7.59E-01	4.85E-05	7.64E-01	5.45E-05	7.59E-01	4.95E-05
		500	6.14E-01	0.00E+00	8.43E-01	5.05E-05	8.49E-01	5.96E-05	8.48E-01	5.25E-05
	0.05	0	7.61E-01	3.43E-05	7.57E-01	3.23E-05	-	-	7.61E-01	3.43E-05
		100	5.26E-01	1.82E-05	7.32E-01	3.54E-05	7.41E-01	4.85E-05	7.33E-01	4.65E-05
		500	5.36E-01	1.01E-06	8.19E-01	2.93E-05	8.28E-01	3.54E-05	8.23E-01	3.23E-05
	0.1	0	7.21E-01	5.15E-05	7.11E-01	5.15E-05	-	-	7.21E-01	5.15E-05
		100	4.22E-01	3.43E-05	6.88E-01	5.35E-05	6.86E-01	5.15E-05	6.76E-01	6.16E-05
		500	4.65E-01	1.01E-06	7.86E-01	4.65E-05	7.91E-01	5.25E-05	7.88E-01	5.15E-05
1.3	0.01	0	9.83E-01	5.45E-05	9.83E-01	5.45E-05	-	-	9.83E-01	5.45E-05
		100	8.59E-01	2.02E-05	9.45E-01	3.23E-05	9.54E-01	4.95E-05	9.50E-01	4.85E-05
		500	6.30E-01	0.00E+00	8.53E-01	5.05E-06	9.12E-01	6.46E-05	9.10E-01	6.87E-05
	0.05	0	9.71E-01	4.65E-05	9.70E-01	4.44E-05	-	-	9.71E-01	4.65E-05
		100	7.68E-01	2.22E-05	9.49E-01	3.23E-05	9.55E-01	5.15E-05	9.50E-01	4.85E-05
		500	6.10E-01	0.00E+00	8.73E-01	1.01E-05	9.23E-01	7.07E-05	9.21E-01	6.67E-05
	0.1	0	9.54E-01	5.66E-05	9.52E-01	4.65E-05	-	-	9.54E-01	5.66E-05
		100	6.91E-01	2.32E-05	9.45E-01	4.04E-05	9.47E-01	4.65E-05	9.40E-01	5.15E-05
		500	6.21E-01	0.00E+00	8.93E-01	1.01E-05	9.27E-01	4.04E-05	9.24E-01	4.55E-05

Table 3.17: **Performance of sample overlap correction for estimating PRS using our method.** Assuming 100 cases and 100 controls shared between base and target studies, we compared the corrected PRS statistics estimated using our method with the real statistics of individual level PRS obtained using PRSice2. Comparison was carried out under various levels of stratification between base and target population ($F_{st} = 0$, 0.05, and 0.1) and p-value thresholds (denoted by P-thres in the table) for SNP selection. For both methods, mean PRS represents the estimated group mean PRS for cases and controls; and p-val are the t-test p-values comparing the resulting PRS distribution in cases and controls. For PRSice2, we computed these statistics for all the samples in the target population, including the samples shared with the base population (denoted by All samples), as well as only for samples that are present exclusively in the target population (denoted by Non-overlapping Samples).

			Our method (`ReACt`)		`PRSice2`			
Fst	P-thres	trait	Corrected statistics		All samples		Non-overlapping Samples	
			mean PRS	p-val	mean PRS	p-val	mean PRS	p-val
0^a	0.05	cases	0.0003	4.07E-05	0.0012	1.09E-54	0.0003	3.59E-07
		controls	0.0000		-0.0009		0.0000	
	0.005	cases	0.0034	1.28E-04	0.0050	6.02E-39	0.0034	1.20E-04
		controls	0.0024		0.0008		0.0025	
	$5 \cdot 10^{-4}$	cases	-0.0030	2.44E-01	-0.0008	8.96E-12	-0.0028	1.47E-01
		controls	-0.0041		-0.0063		-0.0040	
	$5 \cdot 10^{-5}$	cases	0.0441	7.52E-01	0.0471	2.31E-02	0.0449	5.46E-01
		controls	0.0450		0.0419		0.0464	
0.05^b	0.05	cases	0.0000	5.57E-54	0.0002	3.55E-111	0.0001	8.64E-88
		controls	-0.0005		-0.0007		-0.0006	
	0.005	cases	0.0001	4.21E-62	0.0001	5.56E-110	0.0000	3.30E-91
		controls	-0.0019		-0.0025		-0.0024	
	$5 \cdot 10^{-4}$	cases	-0.0063	1.51E-50	-0.0067	1.72E-77	-0.0069	3.61E-70
		controls	-0.0112		-0.0124		-0.0124	
	$5 \cdot 10^{-5}$	cases	-0.0234	4.88E-21	-0.0229	3.21E-32	-0.0232	3.04E-29
		controls	-0.0298		-0.0304		-0.0305	
0.1^c	0.05	cases	0.0001	7.32E-35	0.0004	8.05E-90	0.0004	7.52E-68
		controls	-0.0003		-0.0004		-0.0003	
	0.005	cases	0.0004	2.14E-52	0.0007	8.82E-98	0.0006	3.03E-79
		controls	-0.0014		-0.0017		-0.0015	
	$5 \cdot 10^{-4}$	cases	-0.0048	3.74E-41	-0.0048	6.51E-60	-0.0047	1.32E-52
		controls	-0.0091		-0.0100		-0.0096	
	$5 \cdot 10^{-5}$	cases	0.0109	6.04E-15	0.0087	7.62E-22	0.0088	2.47E-19
		controls	0.0054		0.0021		0.0025	

[a] tested with 550 cases and 550 controls from base and target studies respectively

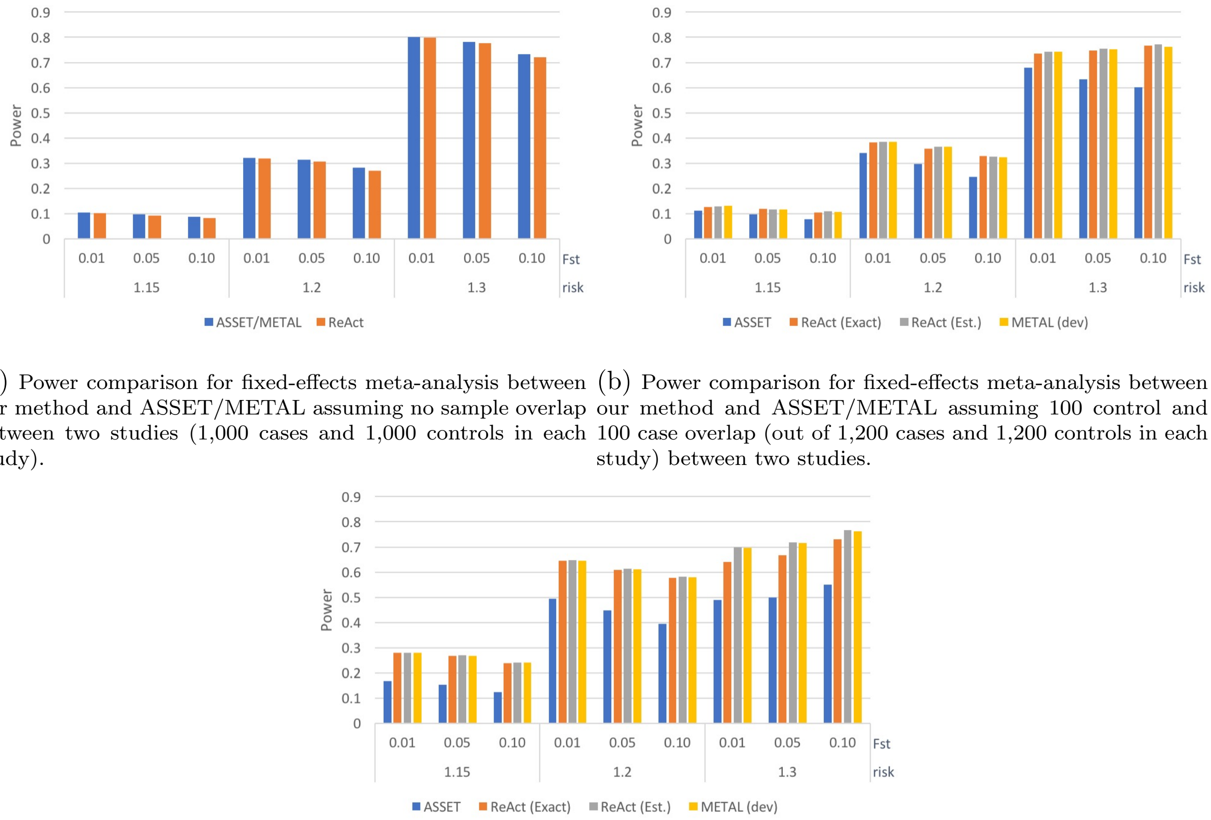
[b] tested with 1,200 cases and 1,200 controls from base and target studies respectively

[c] tested with 1,200 cases and 1,200 controls from base and target studies respectively

Table 3.18: **Using ReACt to run cc-GWAS cross eight neuropsychiatric disorders.** We applied our method for cc-GWAS to the summary statistics of eight neuropsychiatric disorders from PGC. Each spreadsheet reports the genomewide significant trait differential regions for a pair of disorders analyzed. For each genomic region, statistics and annotation for the leading SNP are reported.

This table is large and can be viewed through this link.

3.9 Figures

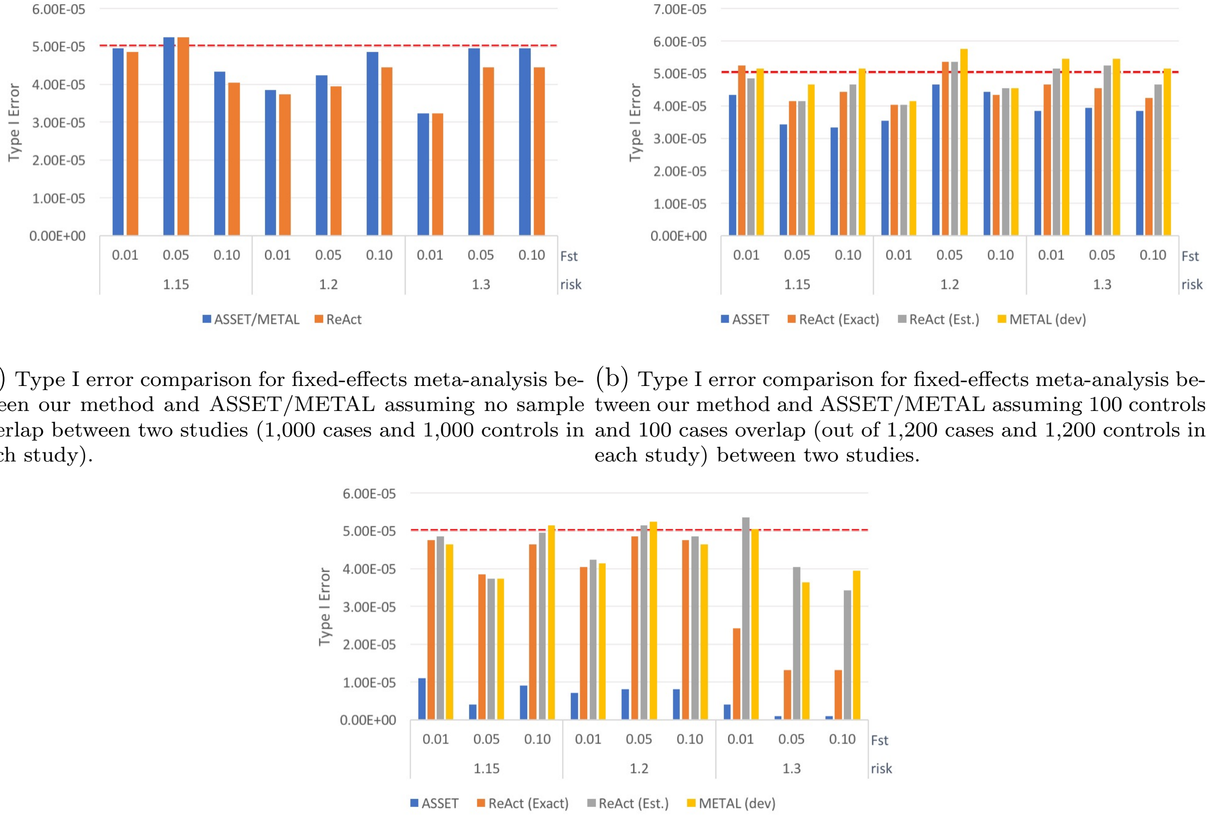


(a) Power comparison for fixed-effects meta-analysis between our method and ASSET/METAL assuming no sample overlap between two studies (1,000 cases and 1,000 controls in each study).

(b) Power comparison for fixed-effects meta-analysis between our method and ASSET/METAL assuming 100 control and 100 case overlap (out of 1,200 cases and 1,200 controls in each study) between two studies.

(c) Power comparison for fixed-effects meta-analysis between our method and ASSET/METAL assuming 500 control and 500 case overlap (out of 2,000 cases and 2,000 controls in each study) between two studies.

Figure 3.1: **Power of fixed-effect meta-analysis with two input studies under different conditions.** We compare the power of our method vs. ASSET/METAL for a significance threshold $p < 5 \cdot 10^{-5}$. METAL dev refers to the latest release in GitHub [308]. Two variants of ReACt are tested: Exact and Est, indicating whether the sample overlap was *exactly* known as part of the input or whether it was *estimated* from the Z-scores [308], respectively. Sample overlap indicates the number of cases and controls that were shared between two input studies, ie., a sample overlap equal to 100 means that that there are 100 cases **and** 100 controls shared between two input studies. Total sample sizes for each input study, including the shared samples, are equal to 2,000 when the sample overlap is equal to zero; 2,400 when the sample overlap is equal to 100; and 4,000 when the sample overlap is equal to 500. In each case, the sample is equally split to cases and controls.



(a) Type I error comparison for fixed-effects meta-analysis between our method and ASSET/METAL assuming no sample overlap between two studies (1,000 cases and 1,000 controls in each study).

(b) Type I error comparison for fixed-effects meta-analysis between our method and ASSET/METAL assuming 100 controls and 100 cases overlap (out of 1,200 cases and 1,200 controls in each study) between two studies.

(c) Type I error comparison for fixed-effects meta-analysis between our method and ASSET/METAL assuming 500 controls and 500 cases overlap (out of 2,000 cases and 2,000 controls in each study) between two studies.

Figure 3.2: **Type I error rate of fixed-effect meta-analysis with two input studies under different conditions.** We compared the type I error rate of our method vs. ASSET/METAL for a significance threshold $p < 5 \cdot 10^{-5}$. METAL dev refers to the latest release in GitHub [308]. Two variants of ReACt are tested: Exact and Est, indicating whether the sample overlap was *exactly* known as part of the input or whether it was *estimated* from the Z-scores [308], respectively. Sample overlap indicates the number of cases and controls that were shared between two input studies, ie., a sample overlap equal to 100 means that there are 100 cases **and** 100 controls shared between two input studies. Total sample sizes for each input study, including the shared samples, are equal to 2,000 when the sample overlap is equal to zero; 2,400 when the sample overlap is equal to 100; and 4,000 when the sample overlap is equal to 500. In each case, the sample is equally split to cases and controls.

4 CONCLUSION

For almost two decades, GWAS has served as one of the most effective tools for uncovering common genetic risk factors underlying complex psychiatric disorders, and the trend is not slowing down. With increased availability of genetic data and massive executions of GWAS, it is time to dig deeper into the biological significance and clinical utility beyond those association results. In this dissertation, we presented in details two "post-GWAS" related projects. One of them focused on unveiling disease mechanism implied by GWAS results, whereas the other is introduced a promising framework that can facilitate various future GWAS summary statistics-based analysis. We demonstrated in this dissertation how much we can already accomplish through using only summary statistics of GWAS, and the potential to achieve even more.

4.1 Summary

In section 2, motivated by high comorbidity rates and symptomatic similarities, we investigated the shared and distinct genetic basis across four common childhood-onset neuropsychiatric disorders. Through analyzing genetic correlation and disease architectures, we found genetic risk factors shared across ADHD, ASD and TS, as well as in between OCD and TS. We subsequently identified those shared genetic factors through systematic meta-analyses and evaluation of posterior probability of association ($m - value$) at SNP, gene and tissue specificity three different levels. Moreover, we looked into the differences in genetic components between these two groups of disorders through conditional analysis. As a results, for each group of disorders, we successfully detected multiple novel genomic regions and genes with pleiotropic effects. Furthermore, our results implied the involvement of HPA-axis, a pathway serving as human central stress response system.

Inspired by the previous project, in section 3, we proposed a novel framework ReACt that can convert the summary statistics of GWAS into the form of SNP allelic frequencies by case control groups. The motivation for this conversion is that in population genetics, many statistics, if not all of them, can be expressed as functions of population SNP allelic/genotypic frequencies, whereas not as many can be written as a function of SNP GWAS statistics in a straightforward manner. To prove the utility of this framework, we put forward three applications: including meta-analysis, case-case GWAS and the one we named group PRS, which was a novel method that could not be achieved before the birth of ReACt. We evaluated the performance of all methods on both synthetic and real GWAS data, and made the implementation publicly available for future improvements.

4.2 Future work

The objective of this dissertation is to contribute to better understanding and utilizing GWAS results for complex psychiatric disorders. We hope that our work can serve as a stepping stone for future researchers going along a similar path.

Our cross-disorder analysis revealed the genetic relationship across TS, ADHD, ASD, and OCD. It will be interesting to further connect those genetic commonalities and distinctions in genetic components with symptom measurements. Instead of categorical diagnosis, symptom-based analyses could be a direction to consider for a refined genetic etiology mapping.

The framework ReACt we proposed is simple in theory, which is part of its advantages and beauty. Moreover, such simplicity also leaves space for further expansion. In this dissertation, we only put forward three ready-to-use, external-support-free applications of this framework, whereas we believe there are numerous more awaiting. A few to be considered includes haplotype-based analysis by integrating LD reference, gene-based analysis and enrichment analysis by incorporating gene annotations and genesets, etc. We look forward to seeing all kinds of creative applications built upon ReACt.

REFERENCES

[1] Jonathan Rees. “Complex disease and the new clinical sciences”. In: *Science* 296.5568 (2002), pp. 698–700.

[2] Anne V Buchanan, Kenneth M Weiss, and Stephanie M Fullerton. “Dissecting complex disease: the quest for the Philosopher’s Stone?” In: *International Journal of Epidemiology* 35.3 (2006), pp. 562–571.

[3] Greg Gibson. “Decanalization and the origin of complex disease”. In: *Nature Reviews Genetics* 10.2 (2009), pp. 134–140.

[4] Stylianos E Antonarakis and Jacques S Beckmann. “Mendelian disorders deserve more attention”. In: *Nature Reviews Genetics* 7.4 (2006), pp. 277–282.

[5] David Botstein and Neil Risch. “Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease”. In: *Nature genetics* 33.3 (2003), pp. 228–237.

[6] Peter M Visscher, William G Hill, and Naomi R Wray. “Heritability in the genomics era—concepts and misconceptions”. In: *Nature reviews genetics* 9.4 (2008), pp. 255–266.

[7] E Bleuler. “Mendelianism in Psychoses, Especially in Schizophrenia”. In: *The Journal of Nervous and Mental Disease* 49.4 (1919), pp. 362–366.

[8] Bart ML Baselmans et al. “Risk in relatives, heritability, SNP-based heritability, and genetic correlations in psychiatric disorders: A review”. In: *Biological Psychiatry* 89.1 (2021), pp. 11–19.

[9] E Pettersson et al. “Genetic influences on eight psychiatric disorders based on family data of 4 408 646 full and half-siblings, and genetic data of 333 748 cases and controls”. In: *Psychological medicine* 49.7 (2019), pp. 1166–1173.

[10] Henrik Larsson et al. “The heritability of clinically diagnosed attention deficit hyperactivity disorder across the lifespan”. In: *Psychological medicine* 44.10 (2014), pp. 2223–2229.

[11] Beata Tick et al. “Heritability of autism spectrum disorders: a meta-analysis of twin studies”. In: *Journal of Child Psychology and Psychiatry* 57.5 (2016), pp. 585–595.

[12] Sven Sandin et al. “The heritability of autism spectrum disorder”. In: *Jama* 318.12 (2017), pp. 1182–1184.

[13] Patrick F Sullivan, Michael C Neale, and Kenneth S Kendler. “Genetic epidemiology of major depression: review and meta-analysis”. In: *American journal of psychiatry* 157.10 (2000), pp. 1552–1562.

[14] Francis J McMahon. *Population-based estimates of heritability shed new light on clinical features of major depression.* 2018.

[15] Spencer L James et al. “Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017”. In: *The Lancet* 392.10159 (2018), pp. 1789–1858.

[16] Patrick F Sullivan and Daniel H Geschwind. “Defining the genetic, genomic, cellular, and diagnostic architectures of psychiatric disorders”. In: *Cell* 177.1 (2019), pp. 162–183.

[17] Ann-Christine Syvänen. “Accessing genetic variation: genotyping single nucleotide polymorphisms”. In: *Nature Reviews Genetics* 2.12 (2001), pp. 930–942.

[18] Barkur S Shastry. “SNPs: impact on gene function and phenotype”. In: *Single Nucleotide Polymorphisms* (2009), pp. 3–22.

[19] Ian M Campbell et al. “Multiallelic positions in the human genome: challenges for genetic analyses”. In: *Human mutation* 37.3 (2016), pp. 231–234.

[20] Teri A Manolio et al. “Finding the missing heritability of complex diseases”. In: *Nature* 461.7265 (2009), pp. 747–753.

[21] Stefan Fröhling and Hartmut Döhner. “Chromosomal abnormalities in cancer”. In: *New England Journal of Medicine* 359.7 (2008), pp. 722–734.

[22] DJ MacIntyre et al. “Chromosomal abnormalities and mental illness”. In: *Molecular Psychiatry* 8.3 (2003), pp. 275–287.

[23] Beata Nowakowska. “Clinical interpretation of copy number variants in the human genome”. In: *Journal of applied genetics* 58.4 (2017), pp. 449–457.

[24] Richard Redon et al. “Global variation in copy number in the human genome”. In: *nature* 444.7118 (2006), pp. 444–454.

[25] Bart Ellenbroek and Jiun Youn. “Chapter 2 - The Genetic Basis of Behavior”. In: *Gene-Environment Interactions in Psychiatry.* Ed. by Bart Ellenbroek and Jiun Youn. San Diego: Academic Press, 2016, pp. 19–46. ISBN: 978-0-12-801657-2. DOI: `https://doi.org/10.1016/B978-0-12-801657-2.00002-1`. URL: `https://www.sciencedirect.com/science/article/pii/B9780128016572000021`.

[26] Brooke Weckselblatt and M Katharine Rudd. “Human structural variation: mechanisms of chromosome rearrangements”. In: *Trends in Genetics* 31.10 (2015), pp. 587–599.

[27] Désirée White and Montserrat Rabago-Smith. “Genotype–phenotype associations and human eye color”. In: *Journal of human genetics* 56.1 (2011), pp. 5–7.

[28] Fan Liu et al. “Digital quantification of human eye color highlights genetic association of three new loci”. In: *PLoS genetics* 6.5 (2010), e1000934.

[29] Pirro G Hysi et al. “Genome-wide association meta-analysis of individuals of European ancestry identifies new loci explaining a substantial fraction of hair color variation and heritability”. In: *Nature genetics* 50.5 (2018), pp. 652–656.

[30] David F Burke et al. “Genome bioinformatic analysis of nonsynonymous SNPs”. In: *BMC bioinformatics* 8.1 (2007), pp. 1–15.

[31] Sanghyeon Kim et al. “Association between SNPs and gene expression in multiple regions of the human brain”. In: *Translational psychiatry* 2.5 (2012), e113–e113.

[32] Karin Fransen et al. “Analysis of SNPs with an effect on gene expression identifies UBE2L3 and BCL3 as potential new risk genes for Crohn’s disease”. In: *Human molecular genetics* 19.17 (2010), pp. 3482–3488.

[33] Barbara E Stranger et al. “Relative impact of nucleotide and copy number variation on gene expression phenotypes”. In: *Science* 315.5813 (2007), pp. 848–853.

[34] Mulin Jun Li et al. “Exploring the function of genetic variants in the non-coding genomic regions: approaches for identifying human regulatory variants affecting gene expression”. In: *Briefings in bioinformatics* 16.3 (2015), pp. 393–412.

[35] Allan F McRae et al. “Identification of 55,000 replicated DNA methylation QTL”. In: *Scientific reports* 8.1 (2018), pp. 1–9.

[36] Jordana T Bell et al. “DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines”. In: *Genome biology* 12.1 (2011), pp. 1–13.

[37] Falk Butter et al. “Proteome-wide analysis of disease-associated SNPs that show allele-specific transcription factor binding”. In: *PLoS Genet* (2012).

[38] Hong-Jian Jin et al. “Identification and validation of regulatory SNPs that modulate transcription factor chromatin binding and gene expression in prostate cancer”. In: *Oncotarget* 7.34 (2016), p. 54616.

[39] Nicholas J Schork et al. “Common vs. rare allele hypotheses for complex diseases”. In: *Current opinion in genetics & development* 19.3 (2009), pp. 212–219.

[40] Sudha K Iyengar and Robert C Elston. “The genetic basis of complex traits”. In: *Linkage Disequilibrium and Association Mapping*. Springer, 2007, pp. 71–84.

[41] Kelly A Frazer et al. “Human genetic variation and its contribution to complex traits”. In: *Nature Reviews Genetics* 10.4 (2009), pp. 241–251.

[42] Walter Bodmer and Carolina Bonilla. “Common and rare variants in multifactorial susceptibility to common diseases”. In: *Nature genetics* 40.6 (2008), pp. 695–701.

[43] Teri A Manolio, Lisa D Brooks, Francis S Collins, et al. “A HapMap harvest of insights into the genetics of common disease”. In: *The Journal of clinical investigation* 118.5 (2008), pp. 1590–1605.

[44] Jonathan K Pritchard. “Are rare variants responsible for susceptibility to complex diseases?” In: *The American Journal of Human Genetics* 69.1 (2001), pp. 124–137.

[45] Jonathan K Pritchard and Nancy J Cox. “The allelic architecture of human disease genes: common disease–common variant... or not?” In: *Human molecular genetics* 11.20 (2002), pp. 2417–2423.

[46] Guhan R Venkataraman and Manuel A Rivas. “Rare and common variant discovery in complex disease: the IBD case study”. In: *Human molecular genetics* 28.R2 (2019), R162–R169.

[47] Noha A Yousri et al. “Whole-exome sequencing identifies common and rare variant metabolic QTLs in a Middle Eastern population”. In: *Nature communications* 9.1 (2018), pp. 1–13.

[48] Greg Gibson. “Rare and common variants: twenty arguments”. In: *Nature Reviews Genetics* 13.2 (2012), pp. 135–145.

[49] Evan E Eichler et al. “Missing heritability and strategies for finding the underlying causes of complex disease”. In: *Nature reviews genetics* 11.6 (2010), pp. 446–450.

[50] Nicholas J Timpson et al. “Genetic architecture: the shape of the genetic contribution to human traits and disease”. In: *Nature Reviews Genetics* 19.2 (2018), pp. 110–124.

[51] Christian R Marshall et al. “Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects”. In: *Nature genetics* 49.1 (2017), pp. 27–35.

[52] Eleanor Wheeler et al. “Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity”. In: *Nature genetics* 45.5 (2013), pp. 513–517.

[53] Patrick F Sullivan, Mark J Daly, and Michael O’donovan. “Genetic architectures of psychiatric disorders: the emerging picture and its implications”. In: *Nature Reviews Genetics* 13.8 (2012), pp. 537–551.

[54] Jacob Gratten et al. “Large-scale genomics unveils the genetic architecture of psychiatric disorders”. In: *Nature neuroscience* 17.6 (2014), pp. 782–790.

[55] The Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium. “Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24. 32 and a significant overlap with schizophrenia”. In: *Molecular autism* 8 (2017), pp. 1–17.

[56] Jakob Grove et al. “Identification of common genetic risk variants for autism spectrum disorder”. In: *Nature genetics* 51.3 (2019), pp. 431–444.

[57] Varun Warrier et al. “A comprehensive meta-analysis of common genetic variants in autism spectrum conditions”. In: *Molecular autism* 6.1 (2015), pp. 1–11.

[58] Robert J Klein et al. “Complement factor H polymorphism in age-related macular degeneration”. In: *Science* 308.5720 (2005), pp. 385–389.

[59] Frank Dudbridge and Arief Gusnanto. “Estimation of significance thresholds for genomewide association scans”. In: *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 32.3 (2008), pp. 227–234.

[60] Andries T Marees et al. “A tutorial on conducting genome-wide association studies: Quality control and statistical analysis”. In: *International journal of methods in psychiatric research* 27.2 (2018), e1608.

[61] Mark I McCarthy et al. “Genome-wide association studies for complex traits: consensus, uncertainty and challenges”. In: *Nature reviews genetics* 9.5 (2008), pp. 356–369.

[62] Psychiatric GWAS Consortium Coordinating Committee. “Genomewide association studies: history, rationale, and prospects for psychiatric disorders”. In: *American Journal of Psychiatry* 166.5 (2009), pp. 540–556.

[63] Rita M Cantor, Kenneth Lange, and Janet S Sinsheimer. “Prioritizing GWAS results: a review of statistical methods and recommendations for their application”. In: *The American Journal of Human Genetics* 86.1 (2010), pp. 6–22.

[64] Vivian Tam et al. “Benefits and limitations of genome-wide association studies”. In: *Nature Reviews Genetics* 20.8 (2019), pp. 467–484.

[65] David E Reich et al. “Linkage disequilibrium in the human genome”. In: *Nature* 411.6834 (2001), pp. 199–204.

[66] Kevin L Gunderson et al. “Whole-genome genotyping”. In: *Methods in enzymology* 410 (2006), pp. 359–376.

[67] Eleonora Porcu et al. “Genotype imputation in Genome-Wide association studies”. In: *Current protocols in human genetics* 78.1 (2013), pp. 1–25.

[68] Richard A Gibbs et al. “The international HapMap project”. In: *Nature* (2003).

[69] 1000 Genomes Project Consortium et al. “A global reference for human genetic variation”. In: *Nature* 526.7571 (2015), p. 68.

[70] Kristin G Ardlie, Leonid Kruglyak, and Mark Seielstad. “Patterns of linkage disequilibrium in the human genome”. In: *Nature Reviews Genetics* 3.4 (2002), pp. 299–309.

[71] Jarmo Ritari et al. “Increasing accuracy of HLA imputation by a population-specific reference panel in a FinnGen biobank cohort”. In: *NAR genomics and bioinformatics* 2.2 (2020), lqaa030.

[72] Meng Lin et al. “Population-specific reference panels are crucial for genetic analyses: an example of the CREBRF locus in Native Hawaiians”. In: *Human molecular genetics* 29.13 (2020), pp. 2275–2284.

[73] Meraj Ahmad et al. “Inclusion of population-specific reference panel from India to the 1000 genomes phase 3 panel improves imputation accuracy”. In: *Scientific reports* 7.1 (2017), pp. 1–8.

[74] Jeffrey Perkel. “SNP genotyping: six technologies that keyed a revolution”. In: *Nature Methods* 5.5 (2008), pp. 447–453.

[75] Stephan Ripke. “Ricopili Pipeline And Standards of GWAS Analyses”. In: *European Neuropsychopharmacology* 29 (2019), S713–S714.

[76] Ben Elsworth et al. “MRC IEU UK Biobank GWAS pipeline version 1”. In: *University of Bristol* (2017).

[77] Max Lam et al. “RICOPILI: rapid imputation for COnsortias PIpeLIne”. In: *Bioinformatics* 36.3 (2020), pp. 930–933.

[78] Jonathan Marchini and Bryan Howie. “Genotype imputation for genome-wide association studies”. In: *Nature Reviews Genetics* 11.7 (2010), pp. 499–511.

[79] Brian L Browning and Sharon R Browning. “Genotype imputation with millions of reference samples”. In: *The American Journal of Human Genetics* 98.1 (2016), pp. 116–126.

[80] Eleftheria Zeggini et al. “An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated data sets”. In: *Nature genetics* 37.12 (2005), pp. 1320–1322.

[81] Jie Huang et al. “Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel”. In: *Nature communications* 6.1 (2015), pp. 1–9.

[82] Shane McCarthy et al. “A reference panel of 64,976 haplotypes for genotype imputation”. In: *Nature genetics* 48.10 (2016), p. 1279.

[83] Pauline C Ng and Ewen F Kirkness. “Whole genome sequencing”. In: *Genetic variation* (2010), pp. 215–226.

[84] Sang Tae Park and Jayoung Kim. “Trends in next-generation sequencing and a new era for whole genome sequencing”. In: *International neurourology journal* 20.Suppl 2 (2016), S76.

[85] Frederick Sanger, Steven Nicklen, and Alan R Coulson. “DNA sequencing with chain-terminating inhibitors”. In: *Proceedings of the national academy of sciences* 74.12 (1977), pp. 5463–5467.

[86] Ayman Grada and Kate Weinbrecht. “Next-generation sequencing: methodology and application”. In: *The Journal of investigative dermatology* 133.8 (2013), e11.

[87] Stephen E Lincoln et al. “One in seven pathogenic variants can be challenging to detect by NGS: An analysis of 450,000 patients with implications for clinical sensitivity and genetic test implementation”. In: *Genetics in Medicine* (2021), pp. 1–8.

[88] Sunguk Shin and Joonhong Park. “Characterization of sequence-specific errors in various next-generation sequencing systems”. In: *Molecular BioSystems* 12.3 (2016), pp. 914–922.

[89] Kensuke Nakamura et al. “Sequence-specific error profile of Illumina sequencers”. In: *Nucleic acids research* 39.13 (2011), e90–e90.

[90] Xiaotu Ma et al. "Analysis of error profiles in deep next-generation sequencing data". In: *Genome biology* 20.1 (2019), pp. 1–15.

[91] Edward J Fox et al. "Accuracy of next generation sequencing platforms". In: *Next generation, sequencing & applications* 1 (2014).

[92] Mark TW Ebbert et al. "Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight". In: *Genome biology* 20.1 (2019), pp. 1–23.

[93] Hayan Lee and Michael C Schatz. "Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score". In: *Bioinformatics* 28.16 (2012), pp. 2097–2105.

[94] Wellcome Trust Case Control Consortium et al. "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls". In: *Nature* 447.7145 (2007), p. 661.

[95] Jacqueline MacArthur et al. "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)". In: *Nucleic acids research* 45.D1 (2017), pp. D896–D901.

[96] Tanya Horwitz et al. "A decade in psychiatric GWAS research". In: *Molecular psychiatry* 24.3 (2019), pp. 378–389.

[97] Max Lam et al. "Comparative genetic architectures of schizophrenia in East Asian and European populations". In: *Nature genetics* 51.12 (2019), pp. 1670–1678.

[98] Antonio F Pardiñas et al. "Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection". In: *Nature genetics* 50.3 (2018), pp. 381–389.

[99] Stephan Ripke et al. "Biological insights from 108 schizophrenia-associated genetic loci". In: *Nature* 511.7510 (2014), p. 421.

[100] Stephan Ripke et al. "Genome-wide association analysis identifies 13 new risk loci for schizophrenia". In: *Nature genetics* 45.10 (2013), pp. 1150–1159.

[101] Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium et al. "Genome-wide association study identifies five new schizophrenia loci". In: *Nature genetics* 43.10 (2011), p. 969.

[102] Rikke Hilker et al. "Heritability of schizophrenia and schizophrenia spectrum based on the nationwide Danish twin register". In: *Biological psychiatry* 83.6 (2018), pp. 492–498.

[103] Patrick F Sullivan, Kenneth S Kendler, and Michael C Neale. "Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies". In: *Archives of general psychiatry* 60.12 (2003), pp. 1187–1192.

[104] David M Howard et al. "Genome-wide association study of depression phenotypes in UK Biobank identifies variants in excitatory synaptic pathways". In: *Nature communications* 9.1 (2018), pp. 1–10.

[105] Craig L Hyde et al. "Identification of 15 genetic loci associated with risk of major depression in individuals of European descent". In: *Nature genetics* 48.9 (2016), pp. 1031–1036.

[106] Naomi R Wray et al. "Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression". In: *Nature genetics* 50.5 (2018), pp. 668–681.

[107] David M Howard et al. “Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions”. In: *Nature neuroscience* 22.3 (2019), pp. 343–352.

[108] World Health Organization et al. *Depression and other common mental disorders: global health estimates*. Tech. rep. World Health Organization, 2017.

[109] Ditte Demontis et al. “Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder”. In: *Nature genetics* 51.1 (2019), pp. 63–75.

[110] Mary M Robertson et al. “Gilles de la Tourette syndrome”. In: *Nature reviews Disease primers* 3.1 (2017), pp. 1–20.

[111] Dongmei Yu et al. “Interrogating the genetic determinants of Tourette’s syndrome and other tic disorders through genome-wide association studies”. In: *American Journal of Psychiatry* 176.3 (2019), pp. 217–227.

[112] Jeremiah M Scharf et al. “Genome-wide association study of Tourette’s syndrome”. In: *Molecular psychiatry* 18.6 (2013), pp. 721–728.

[113] Frank Dudbridge. “Polygenic epidemiology”. In: *Genetic epidemiology* 40.4 (2016), pp. 268–272.

[114] Vikas Bansal et al. “Statistical analysis strategies for association studies involving rare variants”. In: *Nature Reviews Genetics* 11.11 (2010), pp. 773–785.

[115] Alkes L Price et al. “Pooled association tests for rare variants in exon-resequencing studies”. In: *The American Journal of Human Genetics* 86.6 (2010), pp. 832–838.

[116] Dan-Yu Lin and Zheng-Zheng Tang. “A general framework for detecting disease associations with rare variants in sequencing studies”. In: *The American Journal of Human Genetics* 89.3 (2011), pp. 354–367.

[117] Sara L Pulit, Sera AJ de With, and Paul IW de Bakker. “Resetting the bar: Statistical significance in whole-genome sequencing-based association studies of global populations”. In: *Genetic epidemiology* 41.2 (2017), pp. 145–151.

[118] João Fadista et al. “The (in) famous GWAS P-value threshold revisited and updated for low-frequency variants”. In: *European Journal of Human Genetics* 24.8 (2016), pp. 1202–1205.

[119] Eimear E Kenny et al. “Increased power of mixed models facilitates association mapping of 10 loci for metabolic traits in an isolated population”. In: *Human molecular genetics* 20.4 (2011), pp. 827–839.

[120] Catherine Bourgain and Emmanuelle Génin. “Complex trait mapping in isolated populations: Are specific statistical methods required?” In: *European journal of human genetics* 13.6 (2005), pp. 698–706.

[121] David Altshuler, Mark J Daly, and Eric S Lander. “Genetic mapping in human disease”. In: *science* 322.5903 (2008), pp. 881–888.

[122] Lucia A Hindorff et al. “Potential etiologic and functional implications of genome-wide association loci for human diseases and traits”. In: *Proceedings of the National Academy of Sciences* 106.23 (2009), pp. 9362–9367.

[123] Shing Wan Choi, Timothy Shin-Heng Mak, and Paul F O’Reilly. “Tutorial: a guide to performing polygenic risk score analyses”. In: *Nature Protocols* 15.9 (2020), pp. 2759–2772.

[124] Urko M Marigorta et al. "Replicability and prediction: lessons and challenges from GWAS". In: *Trends in Genetics* 34.7 (2018), pp. 504–517.

[125] Anne Marsman et al. "Do current measures of polygenic risk for mental disorders contribute to population variance in mental health?" In: *Schizophrenia Bulletin* 46.6 (2020), pp. 1353–1362.

[126] N Mullins et al. "Polygenic interactions with environmental adversity in the aetiology of major depressive disorder". In: *Psychological medicine* 46.4 (2016), pp. 759–770.

[127] Isabell Brikell et al. "The contribution of common genetic risk variants for ADHD to a general factor of childhood psychopathology". In: *Molecular Psychiatry* 25.8 (2020), pp. 1809–1821.

[128] Hannah L Nicholls et al. "Reaching the end-game for GWAS: machine learning approaches for the prioritization of complex disease loci". In: *Frontiers in genetics* 11 (2020), p. 350.

[129] Faisal Ramzan et al. "Combining Random Forests and a Signal Detection Method Leads to the Robust Detection of Genotype-Phenotype Associations". In: *Genes* 11.8 (2020), p. 892.

[130] Shanwen Sun, Benzhi Dong, and Quan Zou. "Revisiting genome-wide association studies from statistical modelling to machine learning". In: *Briefings in Bioinformatics* 22.4 (2021), bbaa263.

[131] Patrick F Sullivan. "The psychiatric GWAS consortium: big science comes to psychiatry". In: *Neuron* 68.2 (2010), pp. 182–186.

[132] Anna C Need and David B Goldstein. "Next generation disparities in human genomics: concerns and remedies". In: *Trends in Genetics* 25.11 (2009), pp. 489–494.

[133] Melinda C Mills and Charles Rahal. "The GWAS Diversity Monitor tracks diversity by disease in real time". In: *Nature genetics* 52.3 (2020), pp. 242–243.

[134] Urko M Marigorta and Arcadi Navarro. "High trans-ethnic replicability of GWAS results implies common causal variants". In: *PLoS genetics* 9.6 (2013), e1003566.

[135] Cathie Sudlow et al. "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age". In: *PLoS medicine* 12.3 (2015), e1001779.

[136] Cristopher V Van Hout et al. "Exome sequencing and characterization of 49,960 individuals in the UK Biobank". In: *Nature* 586.7831 (2020), pp. 749–756.

[137] John Michael Gaziano et al. "Million Veteran Program: A mega-biobank to study genetic influences on health and disease". In: *Journal of clinical epidemiology* 70 (2016), pp. 214–223.

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

[139] Marilyn Safran et al. "GeneCards Version 3: the human gene integrator". In: *Database* 2010 (2010).

[140] Peter D Stenson et al. "The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine". In: *Human genetics* 133.1 (2014), pp. 1–9.

[141] GTEx Consortium et al. "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans". In: *Science* 348.6235 (2015), pp. 648–660.

[142] Christopher M Yates and Michael JE Sternberg. “The effects of non-synonymous single nucleotide polymorphisms (nsSNPs) on protein–protein interactions”. In: *Journal of molecular biology* 425.21 (2013), pp. 3949–3963.

[143] Leandro M Colli et al. “Burden of nonsynonymous mutations among TCGA cancers and candidate immune checkpoint inhibitor responses”. In: *Cancer research* 76.13 (2016), pp. 3767–3772.

[144] Vasily Ramensky, Peer Bork, and Shamil Sunyaev. “Human non-synonymous SNPs: server and survey”. In: *Nucleic acids research* 30.17 (2002), pp. 3894–3900.

[145] Joke Reumers et al. “SNPeffect v2. 0: a new step in investigating the molecular phenotypic effects of human non-synonymous SNPs”. In: *Bioinformatics* 22.17 (2006), pp. 2183–2185.

[146] Haiming Tang and Paul D Thomas. “Tools for predicting the functional impact of nonsynonymous genetic variation”. In: *Genetics* 203.2 (2016), pp. 635–647.

[147] Rachel Karchin et al. “LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources”. In: *Bioinformatics* 21.12 (2005), pp. 2814–2820.

[148] Jacob A Tennessen et al. “Evolution and functional impact of rare coding variation from deep sequencing of human exomes”. In: *science* 337.6090 (2012), pp. 64–69.

[149] Steffen Schmidt et al. “Hypermutable non-synonymous sites are under stronger negative selection”. In: *PLoS genetics* 4.11 (2008), e1000281.

[150] Christopher M Vockley et al. “Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort”. In: *Genome research* 25.8 (2015), pp. 1206–1214.

[151] Sierra S Nishizaki and Alan P Boyle. “Mining the unknown: assigning function to noncoding single nucleotide polymorphisms”. In: *Trends in Genetics* 33.1 (2017), pp. 34–45.

[152] Ran Elkon and Reuven Agami. “Characterization of noncoding regulatory DNA in the human genome”. In: *Nature biotechnology* 35.8 (2017), pp. 732–746.

[153] Shangwei Ning et al. “LincSNP: a database of linking disease-associated SNPs to human large intergenic non-coding RNAs”. In: *BMC bioinformatics* 15.1 (2014), pp. 1–7.

[154] Yiming Lu et al. “3DSNP: a database for linking human noncoding SNPs to their three-dimensional interacting genes”. In: *Nucleic acids research* (2016), gkw1022.

[155] Simon G Coetzee et al. “FunciSNP: an R/bioconductor tool integrating functional noncoding data sets with genetic association studies to identify candidate regulatory SNPs”. In: *Nucleic acids research* 40.18 (2012), e139–e139.

[156] Benjamin M Neale and Pak C Sham. “The future of association studies: gene-based analysis and replication”. In: *The American Journal of Human Genetics* 75.3 (2004), pp. 353–362.

[157] Jimmy Z Liu et al. “A versatile gene-based test for genome-wide association studies”. In: *The American Journal of Human Genetics* 87.1 (2010), pp. 139–145.

[158] Eleonora Porcu et al. “Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits”. In: *Nature communications* 10.1 (2019), pp. 1–12.

[159] Zhihong Zhu et al. “Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets”. In: *Nature genetics* 48.5 (2016), pp. 481–487.

[160] Guangsheng Pei et al. “deTS: tissue-specific enrichment analysis to decode tissue specificity”. In: *Bioinformatics* 35.19 (2019), pp. 3842–3845.

[161] Eric R Gamazon et al. “Multi-tissue transcriptome analyses identify genetic mechanisms underlying neuropsychiatric traits”. In: *Nature genetics* 51.6 (2019), pp. 933–940.

[162] Weijun Luo et al. “GAGE: generally applicable gene set enrichment for pathway analysis”. In: *BMC bioinformatics* 10.1 (2009), pp. 1–17.

[163] Purvesh Khatri, Marina Sirota, and Atul J Butte. “Ten years of pathway analysis: current approaches and outstanding challenges”. In: *PLoS computational biology* 8.2 (2012), e1002375.

[164] Vijay K Ramanan et al. “Pathway analysis of genomic data: concepts, methods, and prospects for future development”. In: *TRENDS in Genetics* 28.7 (2012), pp. 323–332.

[165] Daniel J Schaid, Wenan Chen, and Nicholas B Larson. “From genome-wide associations to candidate causal variants by statistical fine-mapping”. In: *Nature Reviews Genetics* 19.8 (2018), pp. 491–504.

[166] Christian Benner et al. “Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies”. In: *The American Journal of Human Genetics* 101.4 (2017), pp. 539–551.

[167] Seoae Cho et al. “Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis”. In: *BMC proceedings*. Vol. 3. BioMed Central. 2009, pp. 1–6.

[168] Gao Wang et al. “A simple new approach to variable selection in regression, with application to genetic fine mapping”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82.5 (2020), pp. 1273–1300.

[169] Sarah L Spain and Jeffrey C Barrett. “Strategies for fine-mapping complex traits”. In: *Human molecular genetics* 24.R1 (2015), R111–R119.

[170] Wenan Chen et al. “Fine mapping causal variants with an approximate Bayesian method using marginal test statistics”. In: *Genetics* 200.3 (2015), pp. 719–736.

[171] Nicholas Mancuso et al. “Probabilistic fine-mapping of transcriptome-wide association studies”. In: *Nature genetics* 51.4 (2019), pp. 675–682.

[172] Chris Wallace et al. “Dissection of a complex disease susceptibility region using a Bayesian stochastic search approach to fine mapping”. In: *PLoS genetics* 11.6 (2015), e1005272.

[173] Martijn van de Bunt et al. “Evaluating the performance of fine-mapping strategies at common variant GWAS loci”. In: *PLoS genetics* 11.9 (2015), e1005535.

[174] Kyle J Gaulton et al. “Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci”. In: *Nature genetics* 47.12 (2015), pp. 1415–1425.

[175] Hailiang Huang et al. “Fine-mapping inflammatory bowel disease loci to single-variant resolution”. In: *Nature* 547.7662 (2017), pp. 173–178.

[176] Gleb Kichaev and Bogdan Pasaniuc. “Leveraging functional-annotation data in trans-ethnic fine-mapping studies”. In: *The American Journal of Human Genetics* 97.2 (2015), pp. 260–271.

[177] Gleb Kichaev et al. “Integrating functional data to prioritize causal variants in statistical fine-mapping studies”. In: *PLoS genetics* 10.10 (2014), e1004722.

[178] Doug Speed, John Holmes, and David J Balding. "Evaluating and improving heritability models using summary statistics". In: *Nature Genetics* 52.4 (2020), pp. 458–462.

[179] Hilary K Finucane et al. "Partitioning heritability by functional annotation using genome-wide association summary statistics". In: *Nature genetics* 47.11 (2015), p. 1228.

[180] Doug Speed and David J Balding. "SumHer better estimates the SNP heritability of complex traits from summary statistics". In: *Nature genetics* 51.2 (2019), pp. 277–284.

[181] Huwenbo Shi et al. "Local genetic correlation gives insights into the shared genetic architecture of complex traits". In: *The American Journal of Human Genetics* 101.5 (2017), pp. 737–751.

[182] Brendan Bulik-Sullivan et al. "An atlas of genetic correlations across human diseases and traits". In: *Nature genetics* 47.11 (2015), pp. 1236–1241.

[183] Andrew D Grotzinger et al. "Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits". In: *Nature human behaviour* 3.5 (2019), pp. 513–525.

[184] Zhihong Zhu et al. "Causal associations between risk factors and common diseases inferred from GWAS summary data". In: *Nature communications* 9.1 (2018), pp. 1–12.

[185] Robert F Krueger and Serena Bezdjian. "Enhancing research and treatment of mental disorders with dimensional concepts: toward DSM-V and ICD-11". In: *World Psychiatry* 8.1 (2009), p. 3.

[186] Geoffrey M Reed. "Toward ICD-11: Improving the clinical utility of WHO's International Classification of mental disorders." In: *Professional Psychology: Research and Practice* 41.6 (2010), p. 457.

[187] Oye Gureje et al. "Cultural considerations in the classification of mental disorders: why and how in ICD-11". In: *BMC medicine* 18.1 (2020), pp. 1–2.

[188] Tim Dalgleish et al. "Transdiagnostic approaches to mental health problems: Current status and future directions." In: *Journal of consulting and clinical psychology* 88.3 (2020), p. 179.

[189] Robert F Krueger and Nicholas R Eaton. "Transdiagnostic factors of mental disorders". In: *World Psychiatry* 14.1 (2015), p. 27.

[190] Paolo Fusar-Poli et al. "Transdiagnostic psychiatry: a systematic review". In: *World Psychiatry* 18.2 (2019), pp. 192–207.

[191] Buhm Han and Eleazar Eskin. "Interpreting meta-analyses of genome-wide association studies". In: *PLoS genetics* 8.3 (2012), e1002555.

[192] Cue Hyunkyu Lee et al. "PLEIO: a method to map and interpret pleiotropic loci with GWAS summary statistics". In: *The American Journal of Human Genetics* 108.1 (2021), pp. 36–48.

[193] Patrick Turley et al. "Multi-trait analysis of genome-wide association summary statistics using MTAG". In: *Nature genetics* 50.2 (2018), pp. 229–237.

[194] Wouter J Peyrot and Alkes L Price. "Identifying loci with different allele frequencies among cases of eight psychiatric disorders using CC-GWAS". In: *Nature Genetics* (2021), pp. 1–10.

[195] Christopher N Foley et al. "A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits". In: *Nature communications* 12.1 (2021), pp. 1–18.

[196] Claudia Giambartolomei et al. "Bayesian test for colocalisation between pairs of genetic association studies using summary statistics". In: *PLoS genetics* 10.5 (2014), e1004383.

[197] Claudia Giambartolomei et al. “A Bayesian framework for multiple trait colocalization from summary association statistics”. In: *Bioinformatics* 34.15 (2018), pp. 2538–2545.

[198] Daniel Sik Wai Ho et al. “Machine learning SNP based prediction for precision medicine”. In: *Frontiers in genetics* 10 (2019), p. 267.

[199] Naomi R Wray, Michael E Goddard, and Peter M Visscher. “Prediction of individual genetic risk to disease from genome-wide association studies”. In: *Genome research* 17.10 (2007), pp. 1520–1528.

[200] Florian Privé et al. “Making the most of clumping and thresholding for polygenic scores”. In: *The American Journal of Human Genetics* 105.6 (2019), pp. 1213–1221.

[201] Timothy Shin Heng Mak et al. “Polygenic scores via penalized regression on summary statistics”. In: *Genetic epidemiology* 41.6 (2017), pp. 469–480.

[202] Tian Ge et al. “Polygenic prediction via Bayesian regression and continuous shrinkage priors”. In: *Nature communications* 10.1 (2019), pp. 1–10.

[203] Luke R Lloyd-Jones et al. “Improved polygenic prediction by Bayesian multiple regression on summary statistics”. In: *Nature communications* 10.1 (2019), pp. 1–11.

[204] Bjarni J Vilhjálmsson et al. “Modeling linkage disequilibrium increases accuracy of polygenic risk scores”. In: *The american journal of human genetics* 97.4 (2015), pp. 576–592.

[205] Naomi R Wray et al. “Pitfalls of predicting complex traits from SNPs”. In: *Nature Reviews Genetics* 14.7 (2013), pp. 507–515.

[206] Ulrike Von Luxburg and Bernhard Schölkopf. “Statistical learning theory: Models, concepts, and results”. In: *Handbook of the History of Logic*. Vol. 10. Elsevier, 2011, pp. 651–706.

[207] Laramie Duncan et al. “Analysis of polygenic risk score usage and performance in diverse human populations”. In: *Nature communications* 10.1 (2019), pp. 1–9.

[208] Tiffany Amariuta et al. “Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements”. In: *Nature genetics* 52.12 (2020), pp. 1346–1354.

[209] Carla Márquez-Luna et al. “Multiethnic polygenic risk scores improve risk prediction in diverse populations”. In: *Genetic epidemiology* 41.8 (2017), pp. 811–823.

[210] Marc A Coram et al. “Leveraging multi-ethnic evidence for risk assessment of quantitative traits in minority populations”. In: *The American Journal of Human Genetics* 101.2 (2017), pp. 218–226.

[211] Omer Weissbrod et al. “Leveraging fine-mapping and non-European training data to improve trans-ethnic polygenic risk scores”. In: *medRxiv* (2021).

[212] Malgorzata Maciukiewicz et al. “GWAS-based machine learning approach to predict duloxetine response in major depressive disorder”. In: *Journal of psychiatric research* 99 (2018), pp. 62–68.

[213] Vincent Botta et al. “Exploiting SNP correlations within random forest for genome-wide association studies”. In: *PloS one* 9.4 (2014), e93379.

[214] Urko M Marigorta et al. “Transcriptional risk scores link GWAS to eQTLs and predict complications in Crohn’s disease”. In: *Nature genetics* 49.10 (2017), pp. 1517–1521.

[215] Robert M Maier et al. “Improving genetic prediction by leveraging genetic correlations among human diseases and traits”. In: *Nature communications* 9.1 (2018), pp. 1–17.

[216] Danielle M Dick et al. "Post-GWAS in psychiatric genetics: a developmental perspective on the "other" next steps". In: *Genes, Brain and Behavior* 17.3 (2018), e12447.

[217] Laramie E Duncan, Michael Ostacher, and Jacob Ballon. "How genome-wide association studies (GWAS) made traditional candidate gene studies obsolete". In: *Neuropsychopharmacology* 44.9 (2019), pp. 1518–1523.

[218] Eva C Verbeek et al. "A fine-mapping study of 7 top scoring genes from a GWAS for major depressive disorder". In: *PLoS One* 7.5 (2012), e37384.

[219] Hywel J Williams et al. "Fine mapping of ZNF804A and genome-wide significant evidence for its involvement in schizophrenia and bipolar disorder". In: *Molecular psychiatry* 16.4 (2011), pp. 429–441.

[220] JP Lepine et al. "Prevalence and comorbidity of psychiatric disorders in the French general population". In: *L'encephale* 31.2 (2005), pp. 182–194.

[221] Matthew E Hirschtritt et al. "Lifetime prevalence, age of risk, and genetic relationships of comorbid psychiatric disorders in Tourette syndrome". In: *JAMA psychiatry* 72.4 (2015), pp. 325–333.

[222] Steven R Pliszka. "Comorbid psychiatric disorders in children with ADHD." In: *Attention-deficit hyperactivity disorder: A handbook for diagnosis and treatment* (2015), pp. 140–168.

[223] Joanne L Doherty and Michael J Owen. "Genomic insights into the overlap between psychiatric disorders: implications for research and clinical practice". In: *Genome medicine* 6.4 (2014), pp. 1–13.

[224] Jie Huang et al. "Cross-disorder genomewide analysis of schizophrenia, bipolar disorder, and depression". In: *American Journal of Psychiatry* 167.10 (2010), pp. 1254–1263.

[225] Nanda NJ Rommelse et al. "Shared heritability of attention-deficit/hyperactivity disorder and autism spectrum disorder". In: *European child & adolescent psychiatry* 19.3 (2010), pp. 281–295.

[226] Marco A Grados. "The genetics of obsessive-compulsive disorder and Tourette syndrome: an epidemiological and pathway-based approach for gene discovery". In: *Journal of the American Academy of Child & Adolescent Psychiatry* 49.8 (2010), pp. 810–819.

[227] Cross-Disorder Group of the Psychiatric Genomics Consortium et al. "Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis". In: *The Lancet* 381.9875 (2013), pp. 1371–1379.

[228] Phil H Lee et al. "Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders". In: *Cell* 179.7 (2019), pp. 1469–1482.

[229] International Schizophrenia Consortium. "Common polygenic variation contributes to risk of schizophrenia that overlaps with bipolar disorder". In: *Nature* 460.7256 (2009), p. 748.

[230] Rachel L Kember et al. "Polygenic risk of psychiatric disorders exhibits cross-trait associations in electronic health record data from european ancestry individuals". In: *Biological Psychiatry* 89.3 (2021), pp. 236–245.

[231] Eva Krapohl et al. "Phenome-wide analysis of genome-wide polygenic scores". In: *Molecular psychiatry* 21.9 (2016), pp. 1188–1193.

[232] Beate Leppert et al. "A cross-disorder PRS-pheWAS of 5 major psychiatric disorders in UK Biobank". In: *PLoS genetics* 16.5 (2020), e1008185.

[233] Fernando Pires Hartwig et al. “Body mass index and psychiatric disorders: a Mendelian randomization study”. In: *Scientific reports* 6.1 (2016), pp. 1–11.

[234] Xue Gao et al. “The bidirectional causal relationships of insomnia with five major psychiatric disorders: a Mendelian randomization study”. In: *European Psychiatry* 60 (2019), pp. 79–85.

[235] Enda M Byrne et al. “Conditional GWAS analysis to identify disorder-specific SNPs for psychiatric disorders”. In: *Molecular psychiatry* (2020), pp. 1–12.

[236] Douglas M Ruderfer et al. “Genomic dissection of bipolar disorder and schizophrenia, including 28 subphenotypes”. In: *Cell* 173.7 (2018), pp. 1705–1715.

[237] Iordanis Karagiannidis et al. “The genetics of Gilles de la Tourette syndrome: a common aetiological basis with comorbid disorders?” In: *Current Behavioral Neuroscience Reports* 3.3 (2016), pp. 218–231.

[238] Elena Cravedi et al. “Tourette syndrome and other neurodevelopmental disorders: a comprehensive review”. In: *Child and adolescent psychiatry and mental health* 11.1 (2017), pp. 1–12.

[239] Peristera Paschou et al. “Genetic susceptibility and neurotransmitters in Tourette syndrome”. In: *International review of neurobiology* 112 (2013), pp. 155–177.

[240] Heidi A Browne et al. “Genetics of obsessive-compulsive disorder and related disorders”. In: *Psychiatric Clinics* 37.3 (2014), pp. 319–335.

[241] Jacob AS Vorstman et al. “Autism genetics: opportunities and challenges for clinical translation”. In: *Nature Reviews Genetics* 18.6 (2017), pp. 362–376.

[242] Paul D Arnold et al. “Revealing the complex genetic architecture of obsessive-compulsive disorder using meta-analysis”. In: *Molecular psychiatry* 23.5 (2018), pp. 1181–1181.

[243] Verneri Anttila et al. “Analysis of shared heritability in common disorders of the brain”. In: *Science* 360.6395 (2018).

[244] Mark J Taylor et al. “Association of genetic risk factors for psychiatric disorders and traits of these disorders in a Swedish population twin sample”. In: *JAMA psychiatry* 76.3 (2019), pp. 280–289.

[245] Mohamed Abdulkadir et al. “Investigation of previously implicated genetic variants in chronic tic disorders: a transmission disequilibrium test approach”. In: *European archives of psychiatry and clinical neuroscience* 268.3 (2018), pp. 301–316.

[246] S Hong Lee et al. “Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs”. In: *Nature genetics* 45.9 (2013), pp. 984–995.

[247] S Evelyn Stewart et al. “Genome-wide association study of obsessive-compulsive disorder”. In: *Molecular psychiatry* 18.7 (2013), pp. 788–798.

[248] Manuel Mattheisen et al. “Genome-wide association study in obsessive-compulsive disorder: results from the OCGAS”. In: *Molecular psychiatry* 20.3 (2015), pp. 337–344.

[249] Samsiddhi Bhattacharjee et al. “A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits”. In: *The American Journal of Human Genetics* 90.5 (2012), pp. 821–835.

[250] Dan-Yu Lin and Patrick F Sullivan. “Meta-analysis of genome-wide association studies with overlapping subjects”. In: *The American Journal of Human Genetics* 85.6 (2009), pp. 862–872.

[251] Christiaan A de Leeuw et al. "MAGMA: generalized gene-set analysis of GWAS data". In: *PLoS computational biology* 11.4 (2015), e1004219.

[252] Kyoko Watanabe et al. "Functional mapping and annotation of genetic associations with FUMA". In: *Nature communications* 8.1 (2017), pp. 1–11.

[253] Kai Wang, Mingyao Li, and Hakon Hakonarson. "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data". In: *Nucleic acids research* 38.16 (2010), e164–e164.

[254] Prateek Kumar, Steven Henikoff, and Pauline C Ng. "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm". In: *Nature protocols* 4.7 (2009), pp. 1073–1081.

[255] Ivan Adzhubei, Daniel M Jordan, and Shamil R Sunyaev. "Predicting functional effect of human missense mutations using PolyPhen-2". In: *Current protocols in human genetics* 76.1 (2013), pp. 7–20.

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

[257] Irene Papatheodorou et al. "Expression Atlas: gene and protein expression across multiple studies and organisms". In: *Nucleic acids research* 46.D1 (2018), pp. D246–D251.

[258] Guangchuang Yu et al. "clusterProfiler: an R package for comparing biological themes among gene clusters". In: *Omics: a journal of integrative biology* 16.5 (2012), pp. 284–287.

[259] Marc Carlson. *org.Hs.eg.db*. 2021. URL: `https://bioconductor.org/packages/release/data/annotation/manuals/org.Hs.eg.db/man/org.Hs.eg.db.pdf`.

[260] Elizabeth I Boyle et al. "GO:: TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes". In: *Bioinformatics* 20.18 (2004), pp. 3710–3715.

[261] Ting Qi et al. "Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood". In: *Nature communications* 9.1 (2018), pp. 1–12.

[262] Alexis Battle et al. "Genetic effects on gene expression across human tissues." In: *Nature* 550.7675 (2017), pp. 204–213.

[263] Bernard Ng et al. "An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome". In: *Nature neuroscience* 20.10 (2017), pp. 1418–1426.

[264] Lifeng Dong et al. "LINC00461 promotes cell migration and invasion in breast cancer through miR-30a-5p/integrin β3 axis". In: *Journal of cellular biochemistry* 120.4 (2019), pp. 4851–4862.

[265] Pietro Laneve et al. "A minicircuitry involving REST and CREB controls miR-9-2 expression during human neuronal differentiation". In: *Nucleic acids research* 38.20 (2010), pp. 6895–6905.

[266] Lorenzo F Sempere et al. "Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation". In: *Genome biology* 5.3 (2004), pp. 1–11.

[267] Sha Liu et al. "Identifying common genome-wide risk genes for major psychiatric traits". In: *Human genetics* 139.2 (2020), pp. 185–198.

[268] Thomas E Willnow, Claus M Petersen, and Anders Nykjaer. “VPS10P-domain receptors—regulators of neuronal viability and function”. In: *Nature Reviews Neuroscience* 9.12 (2008), pp. 899–909.

[269] Scott A Small and Gregory A Petsko. “Retromer in Alzheimer disease, Parkinson disease and other neurological disorders”. In: *Nature Reviews Neuroscience* 16.3 (2015), pp. 126–132.

[270] Scott A Small et al. “Model-guided microarray implicates the retromer complex in Alzheimer’s disease”. In: *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society* 58.6 (2005), pp. 909–919.

[271] Tilman Breiderhoff et al. “Sortilin-related receptor SORCS3 is a postsynaptic modulator of synaptic depression and fear extinction”. In: *PloS one* 8.9 (2013), e75006.

[272] Gitte B Christiansen et al. “The sorting receptor SorCS3 is a stronger regulator of glutamate receptor functions compared to GABAergic mechanisms in the hippocampus”. In: *Hippocampus* 27.3 (2017), pp. 235–248.

[273] Kaushik Chakrabarty et al. “Glutamatergic dysfunction in OCD”. In: *Neuropsychopharmacology* 30.9 (2005), pp. 1735–1740.

[274] Harvey S Singer, Christina Morris, and Marco Grados. “Glutamatergic modulatory therapy for Tourette syndrome”. In: *Medical hypotheses* 74.5 (2010), pp. 862–867.

[275] Paromita Roy Choudhury, Sanjukta Lahiri, and Usha Rajamma. “Glutamate mediated signaling in the pathophysiology of autism spectrum disorders”. In: *Pharmacology Biochemistry and Behavior* 100.4 (2012), pp. 841–849.

[276] Stefanos Maltezos et al. “Glutamate/glutamine and neuronal integrity in adults with ADHD: a proton MRS study”. In: *Translational psychiatry* 4.3 (2014), e373–e373.

[277] Danielle S Rudd et al. “A genome-wide CNV analysis of schizophrenia reveals a potential role for a multiple-hit model”. In: *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 165.8 (2014), pp. 619–626.

[278] Sara Calderoni et al. “Basal ganglia and restricted and repetitive behaviours in Autism Spectrum Disorders: current status and future perspectives”. In: *Epidemiology and psychiatric sciences* 23.3 (2014), pp. 235–238.

[279] Philip Shaw et al. “Mapping the development of the basal ganglia in children with attention-deficit/hyperactivity disorder”. In: *Journal of the American Academy of Child & Adolescent Psychiatry* 53.7 (2014), pp. 780–789.

[280] Ricardo Oliveira Horta Maciel et al. “Executive dysfunction, obsessive–compulsive symptoms, and attention deficit and hyperactivity disorder in Systemic Lupus Erythematosus: Evidence for basal ganglia dysfunction?” In: *Journal of the neurological sciences* 360 (2016), pp. 94–97.

[281] Daniele Caligiore et al. “Dysfunctions of the basal ganglia-cerebellar-thalamo-cortical system produce motor tics in Tourette syndrome”. In: *PLoS computational biology* 13.3 (2017), e1005395.

[282] Stephanie HM Van Goozen et al. “Hypothalamic-pituitary-adrenal axis and autonomic nervous system activity in disruptive children and matched controls”. In: *Journal of the American Academy of Child & Adolescent Psychiatry* 39.11 (2000), pp. 1438–1445.

[283] Nestor L Lopez-Duran, Maria Kovacs, and Charles J George. “Hypothalamic–pituitary–adrenal axis dysregulation in depressed children and adolescents: A meta-analysis”. In: *Psychoneuroendocrinology* 34.9 (2009), pp. 1272–1283.

[284] Benjamin L Hankin et al. “Hypothalamic–pituitary–adrenal axis dysregulation in dysphoric children and adolescents: Cortisol reactivity to psychosocial stress from preschool through middle adolescence”. In: *Biological psychiatry* 68.5 (2010), pp. 484–490.

[285] Davide Martino, Antonella Macerollo, and James F Leckman. “Neuroendocrine aspects of Tourette syndrome”. In: *International review of neurobiology* 112 (2013), pp. 239–279.

[286] Corinna Reichl et al. “Hypothalamic-pituitary-adrenal axis, childhood adversity and adolescent nonsuicidal self-injury”. In: *Psychoneuroendocrinology* 74 (2016), pp. 203–211.

[287] Constantine Tsigos and George P Chrousos. “Hypothalamic–pituitary–adrenal axis, neuroendocrine factors and stress”. In: *Journal of psychosomatic research* 53.4 (2002), pp. 865–871.

[288] Dirk Van West, Stephan Claes, and Dirk Deboutte. “Differences in hypothalamic–pituitary–adrenal axis functioning among children with ADHD predominantly inattentive and combined types”. In: *European child & adolescent psychiatry* 18.9 (2009), pp. 543–553.

[289] BA Corbett et al. “Examining cortisol rhythmicity and responsivity to stress in children with Tourette syndrome”. In: *Psychoneuroendocrinology* 33.6 (2008), pp. 810–820.

[290] David W Craig et al. “Assessing and managing risk when sharing aggregate genetic variant data”. In: *Nature Reviews Genetics* 12.10 (2011), pp. 730–736.

[291] Bogdan Pasaniuc and Alkes L Price. “Dissecting the genetics of complex traits using summary association statistics”. In: *Nature Reviews Genetics* 18.2 (2017), p. 117.

[292] Ju-Hyun Park et al. “Estimation of effect size distribution from genome-wide association studies and implications for future discoveries”. In: *Nature genetics* 42.7 (2010), pp. 570–575.

[293] Yan Zhang et al. “Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits”. In: *Nature genetics* 50.9 (2018), pp. 1318–1326.

[294] Zhiyu Yang et al. “Investigating shared genetic basis across Tourette Syndrome and comorbid neurodevelopmental disorders along the impulsivity-compulsivity spectrum”. In: *Biological Psychiatry* (2021).

[295] Fotis Tsetsos et al. “Meta-analysis of Tourette syndrome and attention deficit hyperactivity disorder provides support for a shared genetic basis”. In: *Frontiers in neuroscience* 10 (2016), p. 340.

[296] Christian Benner et al. “FINEMAP: efficient variable selection using summary data from genome-wide association studies”. In: *Bioinformatics* 32.10 (2016), pp. 1493–1501.

[297] Bogdan Pasaniuc et al. “Fast and accurate imputation of summary statistics enhances evidence of functional enrichment”. In: *Bioinformatics* 30.20 (2014), pp. 2906–2914.

[298] Sina Rüeger, Aaron McDaid, and Zoltán Kutalik. “Evaluation and application of summary statistic imputation to discover new height-associated loci”. In: *PLoS genetics* 14.5 (2018), e1007371.

[299] Brendan K Bulik-Sullivan et al. “LD Score regression distinguishes confounding from polygenicity in genome-wide association studies”. In: *Nature genetics* 47.3 (2015), pp. 291–295.

[300] Brielin C Brown et al. “Transethnic genetic-correlation estimates from summary statistics”. In: *The American Journal of Human Genetics* 99.1 (2016), pp. 76–88.

[301] Jie Zheng et al. “LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis”. In: *Bioinformatics* 33.2 (2017), pp. 272–279.

[302] Robert A Power et al. “Polygenic risk scores for schizophrenia and bipolar disorder predict creativity”. In: *Nature neuroscience* 18.7 (2015), pp. 953–955.

[303] Ali Torkamani, Nathan E Wineinger, and Eric J Topol. “The personal and clinical utility of polygenic risk scores”. In: *Nature Reviews Genetics* 19.9 (2018), pp. 581–590.

[304] Frank Dudbridge. “Power and predictive accuracy of polygenic risk scores”. In: *PLoS Genet* 9.3 (2013), e1003348.

[305] Luigi Palla and Frank Dudbridge. “A fast method that uses polygenic scores to estimate the variance explained by genome-wide marker panels and the proportion of variants affecting a trait”. In: *The American Journal of Human Genetics* 97.2 (2015), pp. 250–259.

[306] Michael P LaValley. “Logistic regression”. In: *Circulation* 117.18 (2008), pp. 2395–2399.

[307] David J Balding and Richard A Nichols. “A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity”. In: *Genetica* 96.1-2 (1995), pp. 3–12.

[308] Sebanti Sengupta. *METAL, unpublished Material and Methods*. https://genome.sph.umich.edu/w/images/7/7b/METAL_sample_overlap_method_2017-11-15.pdf. 2017.

[309] Cristen J Willer, Yun Li, and Gonçalo R Abecasis. “METAL: fast and efficient meta-analysis of genomewide association scans”. In: *Bioinformatics* 26.17 (2010), pp. 2190–2191.

[310] Samsiddhi Bhattacharjee et al. “A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits”. In: *The American Journal of Human Genetics* 90.5 (2012), pp. 821–835.

[311] Shing Wan Choi and Paul F O’Reilly. “PRSice-2: Polygenic Risk Score software for biobank-scale data”. In: *Gigascience* 8.7 (2019), giz082.

[312] Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium et al. “A mega-analysis of genome-wide association studies for major depressive disorder”. In: *Molecular psychiatry* 18.4 (2013).

[313] Eli A Stahl et al. “Genome-wide association study identifies 30 loci associated with bipolar disorder”. In: *Nature genetics* 51.5 (2019), pp. 793–803.

[314] Stephan Ripke et al. “Biological insights from 108 schizophrenia-associated genetic loci”. In: *Nature* 511.7510 (2014), p. 421.

[315] Peter M Visscher and William G Hill. “The limits of individual identification from sample allele frequencies: theory and statistical analysis”. In: *PLoS Genet* 5.10 (2009), e1000628.

[316] W Scott Watkins et al. “Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms”. In: *Genome research* 13.7 (2003), pp. 1607–1618.

[317] Guiyan Ni et al. “A comparison of ten polygenic score methods for psychiatric disorders applied across multiple cohorts”. In: *Biological Psychiatry* (2021).

[318] Florian Privé, Julyan Arbel, and Bjarni J Vilhjálmsson. “LDpred2: better, faster, stronger”. In: *BioRxiv* (2020).

[319] L Duncan et al. "Analysis of polygenic risk score usage and performance in diverse human populations". In: *Nature communications* 10.1 (2019), pp. 1–9.

[320] Luke R Lloyd-Jones et al. "Transformation of summary statistics from linear mixed model association on all-or-none traits to odds ratio". In: *Genetics* 208.4 (2018), pp. 1397–1408.

[321] Luigi Freda. *Logistic Regression*. `http://nlp.chonbuk.ac.kr/BML/slides_freda/lec7.pdf`. Accessed: 2020-04-13. 2016.

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

[323] David Firth. "Bias reduction of maximum likelihood estimates". In: *Biometrika* (1993), pp. 27–38.

[324] Georg Heinze and Michael Schemper. "A solution to the problem of separation in logistic regression". In: *Statistics in medicine* 21.16 (2002), pp. 2409–2419.

[325] Clement Ma et al. "Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants". In: *Genetic epidemiology* 37.6 (2013), pp. 539–550.

[326] Pedro RD Bom and Heiko Rachinger. "A generalized-weights solution to sample overlap in meta-analysis". In: *Research Synthesis Methods* (2020).

[327] Alkes L Price et al. "Principal components analysis corrects for stratification in genome-wide association studies". In: *Nature genetics* 38.8 (2006), pp. 904–909.

[328] Aritra Bose et al. "TeraPCA: a fast and scalable software package to study genetic variation in tera-scale genotypes". In: *Bioinformatics* 35.19 (2019), pp. 3679–3683.

[329] Maximilian Haeussler et al. "The UCSC genome browser database: 2019 update". In: *Nucleic acids research* 47.D1 (2019), pp. D853–D858.

[330] Laramie Duncan et al. "Significant locus and metabolic genetic correlations revealed in genome-wide association study of anorexia nervosa". In: *American journal of psychiatry* 174.9 (2017), pp. 850–858.

[331] Eli A Stahl et al. "Genome-wide association study identifies 30 loci associated with bipolar disorder". In: *Nature genetics* 51.5 (2019), pp. 793–803.

VITA

EDUCATION

Ph.D. candidate in Biological Sciences Anticipated 2021 Fall
Major Concentration: Integrative Neuroscience-PULSe
Purdue University West Lafayette, IN, USA
Research Advisor: Prof. Peristera Paschou
Current Graduate School GPA: 3.78/4.0

M.S. in Statistics 2020 - 2021
Major Concentration: Statistics and Computer Science
Purdue University West Lafayette, IN, USA

B.S. in Biotechnology 2011 - 2015
University of Science and Technology of China Hefei, Anhui, China
Undergraduate GPA: 3.39/4.3

RESEARCH EXPERIENCE

Graduate Research Assistant 08/2017-present
Paschou lab, Department of Biological Sciences, Purdue University

Research Advisors: Prof. Peristera Paschou (Dept. of Biological Sciences)
Prof. Petros Drineas (Dept. of Computer Science)

- Identifying a shared regulatory background for neurodevelopmental disorders through meta-analysis of genome-wide association studies (GWAS)
- Developing novel methodologies to effectively extract and analyze information from summary statistics of genome-wide association studies
- Familiar with GWAS, rare variants analyses and downstream analyses using individual level data or summary statistics

Undergraduate Research Assistant 04/2014 – 06/2016
Cognitive Neuropsychology Laboratory, School of Life Science, University of Science and Technology of China

Research Advisors: Prof. Xiaochu Zhang
Dr. Lizhuang Yang

- Evaluating individual differences in the lateralization of social cognitive function in temporo-parietal junction (TPJ)
- Investigating impacts of electrical stimulation over bilateral occipito-temporal regions on subjects' electroencephalogram (EEG) and the composite face effect

- Experience in collecting and analyzing magnetic resonance imaging (MRI), EEG and behavior data of human subjects

TEACHING EXPERIENCE

Teaching Assistant 2019 Summer
BIOL 22100 Microbiology Laboratory, Purdue University
Course Instructor: Dr. Ashwana Fricker

Teaching Assistant 2018 Spring
BIOL 22100 Microbiology Laboratory, Purdue University
Course Instructor: Dr. Kiryl Datsenka

HONORS & AWARDS

Purdue University Biological Sciences Department Travel Award
Purdue University 2019

Purdue Institute for Integrative Neuroscience (PIIN) Travel Grant
Purdue University 2018

Outstanding Student Scholarship of USTC
University of Science and Technology of China 2014

Outstanding Student Scholarship of USTC
University of Science and Technology of China 2013

Outstanding Student Scholarship of USTC
University of Science and Technology of China 2012

Freshman Scholarship of USTC
University of Science and Technology of China 2011

PRESENTATIONS

Z. Yang, P. Paschou, P. Drineas, *Reconstructing SNP Allele and Genotype Frequencies from GWAS Summary Statistics.* Virtual poster presentation. Presented at American Society of Human Genetics 2021 Virtual Meeting, October 18th, 2021, Online

Z. Yang, PGC Cross-disorder Working group, P. Paschou, *Investigating shared genetic basis across Tourette Syndrome and comorbid neurodevelopmental disorders along the impulsivity-compulsivity spectrum.* Virtual poster presentation. Presented at Greater Indiana Society for Neuroscience 2021 Annual Meeting, April 8th, 2021, Online

Z. Yang, PGC Cross-disorder Working group, P. Paschou, *Shared biological mechanisms across*

ADHD,ASD,OCD and TS – from Genetics to Neurobiology. Oral presentation. Presented at Purdue Biological Retreat, November 15-17th, 2019, Plymouth, Indiana

Z. Yang, H. Wu, P. Lee, F.Tsetsos, L. Davis, D. Yu, S. Lee, S. Dalsgaard, J. Haavik, C. Barta, T. Zayats, V. Eapen, N. Wray, B. Devlin, M. Daly, B. Neale, A. Børglum, J. Crowley, J. Scharf, C. Mathews, S. Faraone, B. Franke, M. Mattheisen, J. Smoller, P. Paschou, *Cross-disorder meta-analysis of genomewide association studies sheds light into potentially shared neurobiology across Attention Deficit Hyperactivity Disorder, Autism Spectrum Disorders, Obsessive Compulsive Disorder, and Tourette Syndrome.* Oral presentation. Presented at 2019 World Congress of Psychiatric Genetics, October 30, 2019, Anaheim, California

Z. Yang, Tsetsos, P. Paschou, *Uncovering the Shared Genetic Basis across Childhood-onset Neuropsychiatric Disorders: ADHD, ASD, OCD and TS.* Poster presentation. Presented at Health and Disease: Science, Technology, Culture and Policy poster session, February 28th, 2019, West Lafayette, Indiana

Z. Yang, Tsetsos, P. Paschou, *Uncovering the Shared Genetic Basis across Childhood-onset Neuropsychiatric Disorders: ADHD, ASD, OCD and TS.* Poster presentation. Presented at Sigma Xi 2019 poster night, February 26th, 2019, West Lafayette, Indiana

Z. Yang, A. Bose, P. Drineas, P. Paschou, *Application of Denoising Techniques on SNP-based Risk Prediction for Complex Neuropsychiatric Disorders.* Poster presentation (PgmNr 1376). Presented at the 68th Annual Meeting of The American Society of Human Genetics, October 19th, 2018, San Diego, California

Z. Yang, A. Bose, P. Drineas, P. Paschou, *Application of Denoising Techniques on Genome-wide Association Studies.* Poster presentation. Presented at Health and Disease: Science, Technology, Culture and Policy poster session, March 1st, 2018, West Lafayette, Indiana

Z. Yang, A. Bose, TS/OCD PGC Working group, P. Drineas, P. Paschou, *Application of machine learning approaches on genotype based disease prediction.* Poster presentation. Presented at Purdue Biological Retreat, November 10-12th, 2017, Bloomington, Indiana

Z. Yang, T. Tsetsos,P. Paschou, *Identifying a shared regulatory background for neuro-developmental disorders through meta-analysis of genomewide association studies.* Poster presentation (PgmNr 2142). Presented at the 67th Annual Meeting of The American Society of Human Genetics, October 19th, 2017, Orlando, Florida

PUBLICATIONS

Yang, Z., Paschou, P., Drineas, P. (2021). Reconstructing SNP Allele and Genotype Frequencies from GWAS Summary Statistics. *bioRxiv*

Topaloudi, A., Zagoriti, Z., Flint, A. C., Martinez, M. B.,**Yang, Z.**, Tsetsos, F., ... & Paschou, P. (2021). Myasthenia gravis genome-wide association study implicates AGRN as a risk locus. *Journal of Medical Genetics.*

Yang, Z., Wu, H., Lee, P. H., Tsetsos, F., Davis, L. K., Yu, D., ... & Paschou, P. (2021). Investigating shared genetic basis across Tourette Syndrome and comorbid neurodevelopmental disorders along the impulsivity-compulsivity spectrum. *Biological Psychiatry.*

Yang L. Z., Zhang W., Wang, W., **Yang, Z.**, Wang, H., ... & Zhang, X. (2020). Neural and psychological predictors of cognitive enhancement and impairment from neurostimulation. *Advanced Science.* No. advs.201902863.

Wang, S., Mandell JD., Kumar, Y., Sun, N., Morris, MT., Arbelaez, J., Nasello, C., Dong, S., Duhn, C., Zhao, X., **Yang, Z.**, . . . & State MW. (2018). De Novo Sequence and Copy Number Variants Are Strongly Associated with Tourette Disorder and Implicate Cell Polarity in Pathogenesis. *Cell reports.* Sep 25;24(13):3441-54.

Yang, L. Z., **Yang, Z.**, & Zhang, X. (2016). Non-invasive brain stimulation for the treatment of nicotine addiction: potential and challenges.*Neuroscience bulletin.* 32(6), 550-556.

Ndasauka, Y., Hou, J., Wang, Y., Yang, L., **Yang, Z.**, Ye, Z., ... & Zhang, X. (2016). Excessive use of Twitter among college students in the UK: Validation of the Microblog Excessive Use Scale and relationship to social interaction and loneliness.*Computers in Human Behavior.* 55, 963-971.

Yang, L. Z., Zhang, W., Shi, B., **Yang, Z.**, Wei, Z., Gu, F., ... & Zhang, X. (2014). Electrical stimulation over bilateral occipito-temporal regions reduces N170 in the right hemisphere and the composite face effect. *PloS one.* 9(12), e115772.