Fakultät für Medizin
Technische Universität München

Exploring Genetic Interactions: from Tools Development with Massive Parallelization on GPGPU to Multi-Phenotype Studies on Dyslexia

Beibei Jiang

Vollständiger Abdruck der von der Fakultät für Medizin der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:
Prof. Dr. Thomas Korn

Prüfende der Dissertation:
1. apl. Prof. Dr. Bertram Müller-Myhsok
2. Prof. Dr. Burkhard Rost

Die Dissertation wurde am 02. 05. 2019 bei der Technischen Universität München eingereicht und durch die Fakultät für Medizin am 08.10.2019 angenommen.

Abstract

Over a decade, genome-wide association studies (GWASs) have provided insightful information into the genetic architecture of complex traits. However, the variants found by GWASs explain just a small portion of heritability. Meanwhile, as large scale GWASs and meta-analyses of multiple phenotypes are becoming increasingly common, there is a need to develop computationally efficient models/tools for multi-locus studies and multi-phenotype studies. Thus, we were motivated to focus on the development of tools serving for epistatic studies and to seek for analysis strategy jointly analyzed multiple phenotypes.

By exploiting the technical and methodological progress, we developed three R packages. **SimPhe** was built based on the Cockerham epistasis model to simulate (multiple correlated) phenotype(s) with epistatic effects. Another two packages, **episcan** and **gpuEpiScan**, simplified the calculation of `EPIBALSTER` and `epiHSIC` and were implemented with high performance, especially the package based on Graphics Processing Unit (GPU). The two packages can be employed by epistasis detection in both case-control studies and quantitative trait studies. Our packages might help drive down costs of computation and increase innovation in epistatic studies.

Moreover, we explored the gene-gene interactions on developmental dyslexia, which is mainly characterized by reading problems in children. Multivariate meta-analysis was performed on genome-wide interaction study (GWIS) for reading-related phenotypes in the dyslexia dataset, which contains nine cohorts from different locations. We identified one genome-wide significant epistasis, rs1442415 and rs8013684, associated with word reading, as well as suggestive genetic interactions which might affect reading abilities. Except for rs1442415, which has been reported to influence educational attainment, the genetic variants involved in the suggestive interactions have shown associations with psychiatric disorders in previous GWASs, particularly with bipolar disorder. Our findings suggest

making efforts to investigate not just the genetic interactions but also multiple correlated psychiatric disorders.

Zusammenfassung

Über ein Jahrzehnt haben GWASs aufschlussreiche Informationen zur genetischen Architektur komplexer Merkmale geliefert. Die von GWASs gefundenen Varianten erklären jedoch nur einen kleinen Teil der Erblichkeit. In der Zwischenzeit, da sich GWASs im großen Maßstab und Meta-Analysen für mehrere Phänotypen immer mehr durchsetzen, besteht die Notwendigkeit, rechenintensive Modelle/Werkzeuge für Multi-Locus-Studien und Multi-Phänotyp-Studien zu entwickeln. Daher haben wir Programme und Analysestrategien für epistatische Studien entwickelt, die mehrere Phänotypen gemeinsam behandeln.

Wir nutzten den technischen und methodischen Fortschritt aus und entwickelten drei R-Pakete. **SimPhe** basiert auf dem Epistasemodell von Cockerham, um (mehrfach korrelierte) Phänotyp(en) mit epistatischen Effekten zu simulieren. Die beiden anderen Pakete, **episcan** und **gpuEpiScan**, vereinfachen die Berechnung von `EPIBLASTER` und `epiHSIC` und wurden mit hoher Rechenleistung implementiert, insbesondere das auf GPU basierende Paket. Die beiden Pakete können zum Epistasisnachweis sowohl in Fall-Kontroll-Studien als auch in quantitativen Merkmalsstudien eingesetzt werden. Unsere Pakete könnten dazu beitragen, die Rechenkosten zu senken und die Innovation bei epistatischen Studien zu steigern.

Darüber hinaus untersuchten wir die Gen-Gen-Wechselwirkungen bei Entwicklungsstörungen, die hauptsächlich durch Leseprobleme bei Kindern gekennzeichnet sind. Multivariate Meta-Analysen wurden an GWIS für lesebezogene Phänotypen in den Dyslexie-Daten durchgeführt, die neun Kohorten aus verschiedenen Orten enthält. Wir identifizierten eine genomweit signifikante Epistase, rs1442415 und rs8013684, die mit dem Lesen von Wörtern assoziiert sind, sowie suggestive genetische Interaktionen, die die Lesefähigkeiten beeinflussen könnten. Mit Ausnahme von rs1442415, von dem berichtet wurde, dass es den Bildungserfolg beeinflusst, haben die genetischen Varianten, die an den suggestiven Interaktionen beteiligt sind, Assoziationen mit psychiatrischen Störungen in früheren GWASs gezeigt, insbesondere mit bipolaren Störungen. Unsere Ergeb-

nisse legen nahe, Anstrengungen zu unternehmen, um nicht nur die genetischen Interaktionen zu untersuchen, sondern auch die Beziehungen zwischen verschiedenen psychiatriSchen Störungen.

Acknowledgements

Life goes with the gains from the current and the past. To a future me, please remember the people who have raised you, supervised you, mentored you, helped you, accompanied you, and loved you.

My deepest gratitude goes to my supervisors. My doctoral advisor, Bertram Müller-Myhsok, always believes in me and supports me during my Ph.D. Thanks for giving me the opportunity to study in the Statistical Genetics group and creating such a comfortable working atmosphere. Without your enlightening instruction and impeccable kindness, I could not have completed my doctorate and dissertation. I would also like to give many thanks to my second advisor, Burkhard Rost, who provided me with valuable guidance in every stage of my doctor life and offered me impressive wisdom and outlook.

Moreover, I would like to express my great gratitude to my doctoral mentors. Many thanks to Benno Pütz for guiding me into the wonderful High Performance Computing world and encouraging me to challenge myself in unknown areas. Also to my second mentor, Nazanin Mirza-Schreiber, thanks for introducing me to human genetics and helping me shaping an efficient analysis. I am so happy that you are my mentors and you make my Ph.D. life much easier.

In addition, I would like to thank all my teachers during every stage of my studentship, especially in bachelor and master, who has helped me to develop the fundamental and essential academic competence.

My great thanks to all members of the Statistical Genetics group at the Max Planck Institute of Psychiatry in the past three and a half years. Thank you for all the fruitful discussions and the "girl's nights". Particularly, special thanks to the previous members, Meiwen Jia, Ilaria Bonavita, Alessandro Gialluisi, and Till Andlauer, who accompanied me through the difficult first year in Germany.

To my family and friends, as well as to the person who loves/loved me, I thank you for all the incredible support. You light up my life and bring me so much

happiness. My life is full of the beauty you have created. The last special thanks to my parents and grandparents for raising me and teaching me to work hard and be kind. As you always say, amazing things will come after efforts.

Beibei Jiang
Max Planck Institute of Psychiatry
Munich, March 18, 2019

Contents

List of Figures

List of Tables

Acronyms

API	Application Programming Interface
ARAN	alphanumeric rapid automatized naming
BD	bipolar disorder
BLAS	Basic Linear Algebra Subroutines
CAD	coronary artery disease
CCA	Canonical Correlation Analysis
CD	Crohn disease
CPU	Central Processing Unit
CRAN	Comprehensive R Archive Network
DD	Developmental Dyslexia
DNA	deoxyribonucleic acid
DRAN	digits rapid automatized naming
DS	digit span
eQTL	expression quantitative trait locus
eSNP	expression single nucleotide polymorphism
FDR	false discovery rate
GPC	Graphics Processing Cluster
GPGPU	General Purpose Graphics Processing Unit
GPU	Graphics Processing Unit
GUI	Graphical User Interface
GWAS	genome-wide association study
GWIS	genome-wide interaction study

HPC	High Performance Computing
HSIC	Hilbert-Schmidt Independence Criterion
HWE	Hardy-Weinberg equilibrium
IBD	identity-by-distance
IBS	identity-by-state
IC	inbreeding coefficient
KB	Kilobyte
Kb	kilo base pairs
LD	linkage disequilibrium
LD/ST	load/store unit
LRAN	letters rapid automatized naming
MAF	minor allele frequency
Mb	mega base pairs
MDS	Multidimensional Scaling
ML	Maximum Likelihood
MPI	Message Passing Interface
NCBI	National Center for Biotechnology Information
NWR	non-word reading
OpenMP	Open Multi-Processing
PCA	Principal Component Analysis
PD	phoneme deletion
PRAN	pictures rapid automatized naming
PTSD	post-traumatic stress disorder
Q-Q	Quantile-Quantile
QC	quality control
QTL	quantitative trait locus
REML	Restricted Maximum Likelihood

RNA	ribonucleic acid
SFU	Special Function Units
SM	Streaming Multiprocessor
SNP	single nucleotide polymorphism
SP	Streaming Processor
T1D	type 1 diabetes
T2D	type 2 diabetes
TPC	Texture Processing Cluster
UK	United Kingdom
US	United States
WNWR	word and non-word reading
WR	word reading
WS	word spelling
WTCCC	Wellcome Trust Case Control Consortium

1 Introduction

High-throughput technologies, producing dense SNPs information across the whole genome, and current automated phenotyping technologies have enabled the study of the relationship between genotype and phenotype (using GWAS) at an unprecedented level of detail. GWASs have detected a large number of genetic variants associated with various complex traits. However, they fail to explain much of the phenotypic variation and meanwhile epistasis has been suggested to be considered for better understanding the genetic architecture of complex traits [1, 2, 3].

1.1 Motivation for epistasis related tools development

Epistasis analysis on the whole-genome level is facing many challenges. It acquires not only the appropriate statistical methods but also the efficient implementation to accelerate the calculation process. Due to the technological and computational advances in the past years, there are several statistical methods as well as related software (tools) which have been developed to identify epistasis. Some of them have taken advantage of the General Purpose Graphics Processing Units (GPGPUs), especially with the **CUDA** parallel model, which succeed in reducing the computational burden causing by the exhaustive search [4, 5, 6, 7, 8, 9]. Most of them have limitations in the phenotypic type, only case-control or quantitative phenotype studies, or in the accessibility owing to the requirement of command line knowledge for their deployment.

Motivated by the importance of epistasis studies, the development of computing architectures, the advantages of GPU utilization, and providing tools easily exploited by domain experts with the knowledge to interpret the epistasis results, we were aiming to develop R packages with high performance for epistasis detection in both case-control and quantitative phenotype studies.

Besides, these methods (tools) are developed in diverse mathematical ways indicating that the performance of their results is difficult to compare. A controllable simulation tool can be used to evaluate type I error rates or to perform power comparisons for the different statistical tests [10]. Nevertheless, the current simulation tools rarely take into account epistatic effects when generating quantitative phenotypes. Thus, we were also motivated to develop a tool which can simulate single or multiple (correlated) quantitative phenotypes based on genotypes with additive, dominance, and epistatic effects.

1.2 Motivation in multi-phenotype studies

With the availability of cheaper and accurate assays to quantify multiple phenotypes in large population cohorts, large scale GWASs, as well as meta-analyses, of multiple phenotypes are becoming increasingly common and many efforts have been made on developing methods for multi-phenotype studies [11, 12, 13, 14, 15, 16, 17, 18, 19]. There is an increasing need to develop models and computationally efficient algorithms for joint analysis of multi-SNP and multi-phenotype data [20]. Moreover, Webber et al. have discussed the problem in epistasis analysis from phenotype definition and suggested to study the similar behavioral traits or endophenotypes, which might be more likely to detect epistatic effects influencing the disease [21]. Therefore, we were motivated to find an effective analytical strategy to detect epistasis to conduct multi-phenotype studies, with expectations to gain a deeper understanding of the genetic architecture of complex traits.

Dyslexia, one of the most common neurodevelopmental disorders affecting children across languages, writing systems, and educational approaches, is usually diagnosed by collecting several psychological and psychometric measures which are statistically or functionally correlated [22]. Previous linkage and association analysis only partly reveal the genetic architecture for dyslexia [23, 24, 25]. By considering the interplay of genetic factors on the endophenotypes determining dyslexia, we aimed to find novel genetic variants interacting in an epistatic way influencing the reading abilities in children.

1.3 Overview of the thesis structure

Apart from the chapter of a real case application on dyslexia utilizing the tool described in the method development part, the chapters of this thesis are self-contained which allows reading each of them without having to know the details in the previous one(s).

Being an interdisciplinary work combining human genetics, statistics, and computer science, a brief introduction of the research interests is presented as Chapter 1 (Introduction). The background related to the general issues in the human genome, the computational burden from the whole-genome analysis, and the current research trends in solving the related problems is given in Chapter 2 (State of the Art). Specifically, it covers the discussion about the GWAS limits, the reason to conduct GWIS, the analytic and computational challenges for epistatic analysis, and the methods to be considered for epistasis detection in multi-phenotype studies.

Chapter 3 (Methods and Tools Development) describes the algorithms and implementations of the three R packages, **SimPhe**, **episcan**, and **gpuEpiScan**, developed according to our interests (see motivations in Section 1.1). Each R package was presented with examples. The efficiency of the key function in GPU based R package, **gpuEpiScan**, is discussed.

Chapter 4 (A Real Dataset Application: Dyslexia) provides an application on reading abilities with our analytical strategy in multiple phenotypes study, which is related to the motivations in Section 1.2. The strategy is provided and evaluated before being applied on entire dyslexic data. The details and novel findings of the epistasis analysis are given. Functional meanings of the results are explored and discussed to gain more valuable knowledge on the general correlations among psychiatric disorders.

In Chapter 5 (Conclusion & Outlook), we conclude the findings, discuss the contributions and limitations, and provide perspectives for future developments, of which some attempts are already shown. Supplementary information on the algorithms, the datasets, and the detailed results which have not presented in previous chapters are given in the Appendix A and B.

2 State of the Art

2.1 Genome-wide association study and epistasis

2.1.1 Genome-wide association study

GWASs are aimed at detecting variants at genomic loci associated with human diseases or other complex traits in the population and, in particular, at detecting associations between common SNPs and phenotypes.

2.1.1.1 Genetic variants

The term of DNA refers to a double-stranded molecule mainly located within the cell nucleus. A genetic variation, in which a nucleotide on a certain genomic position in the DNA is exchanged, is called single nucleotide polymorphism (SNP, Figure 2.1). Chemically (Figure 2.2) and most frequently, the exchange happens between adenine (A) and guanine (G) or between cytosine (C) and thymine (T). SNPs are the most common type of genetic variation on humans. SNPs may fall within coding sequences, non-coding regions, or in the intergenic regions of genes (regions between genes). Generally, SNPs are more frequent in non-coding regions than in more conserved coding-regions [27]. SNPs within a coding sequence do not necessarily change the amino acid sequence of the protein that is produced, due to the degeneracy of the genetic code. SNPs in non-coding regions may still affect gene splicing, transcription factor binding, messenger ribonucleic acid (RNA) degradation, or the sequence of noncoding RNA. A SNP upstream or downstream of a gene could affect the gene expression, which is referred to as an expression single nucleotide polymorphism (eSNP).

Since 2005 [28], SNPs are considered as the biological markers in GWASs to map the association between genetic factors and measurable phenotypes, for example complex diseases. With the progress of modern technologies, especially

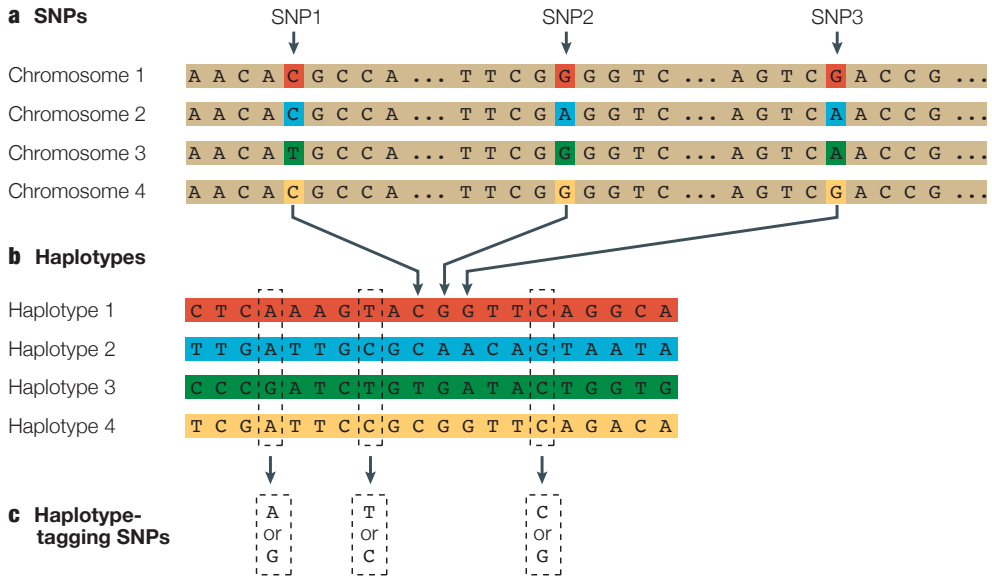

Figure 2.1: SNPs, haplotypes and haplotype-tagging SNPs [26].
a. SNPs are shown in a short stretch of deoxyribonucleic acid (DNA) in four versions of the same chromosomal region taken from different individuals. Most of the DNA sequence is identical in these chromosomes, but variation is shown to occur at three bases. Each SNP has two possible alleles; the first SNP (SNP1) has the alleles C and T.
b. A haplotype consists of a particular combination of alleles at nearby SNPs. Shown here are the observed genotypes for 20 SNPs that extend across 6 000 bases of DNA. Only the variable bases are shown, including the three SNPs that are shown in panel a. For this region, most of the chromosomes in a population survey have haplotypes 1–4.
c. Genotyping of just the 3 haplotype-tagging SNPs out of the 20 SNPs is sufficient to uniquely identify these 4 haplotypes. For example, if a particular chromosome has the sequence A-T-C at these three haplotype-tagging SNPs, this sequence matches the pattern determined for haplotype 1. Note that many chromosomes carry the common haplotypes in the population.

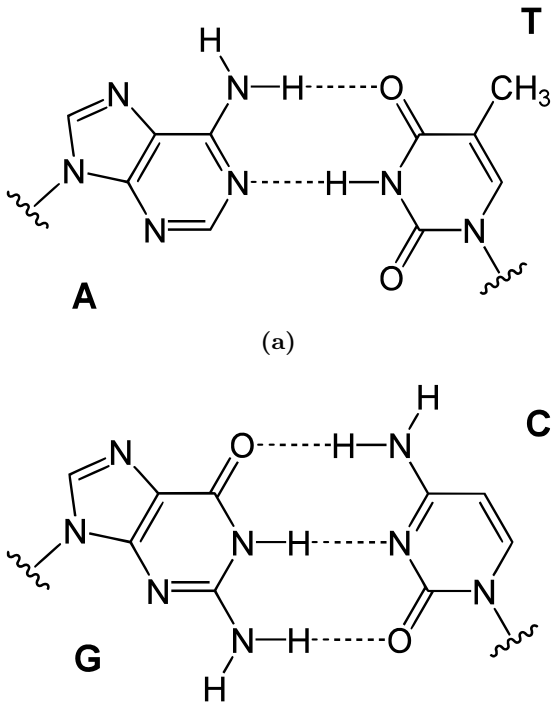

Figure 2.2: Base pairs (from `https://en.wikipedia.org/wiki/Base_pair`).
(a) An **A.T** base pair with two hydrogen bonds.
(b) A **G.C** base pair with three hydrogen bonds. Non-covalent hydrogen bonds between the bases are shown as dashed lines. The wiggly lines stand for the connection to the pentose sugar and point in the direction of the minor groove.

the development of microarray platforms and sequencing strategies, several hundred thousand to more than a million SNPs are usually assayed in thousands of individuals in a typical GWAS.

Though SNPs are the most common genetic variants utilized in GWAS, there are also other variants which are nowadays included in the analysis, e.g., indels, the genetic variations referring to the insertions or deletions of bases in the genome, which are the second most common type of genetic variation on humans. The public genetic variation database, **dbSNP**, from the National Center for Biotechnology Information (NCBI) contains a large number of SNPs and indels. In 2015, the 1000 Genomes Project Consortium has contributed or validated 80 million of

the 100 million variants in **dbSNP**[1]. They reported that over 99.9 % of variants consist of SNPs and short indels and the number of different sites between a typical genome and the reference human genome ranges from 4.1 to 5.0 million. The range comes from the different total number of observed non-reference sites among populations (Figure 2.3) [29].

SNPs and indels provide researchers a powerful way to study the genetic root of the differences that are apparent across the human race.

2.1.1.2 Linkage and association analysis

Considering the experimental design and the coverage of the genome, the first well-established GWAS, was reported in 2007 by Wellcome Trust Case Control Consortium (WTCCC) [30] while a precursor study was conducted in 2005 [28]. Before that, linkage analysis was the commonly used tool to map genomic loci, which have effects on complex traits in the past decades. Gene mapping by linkage relies on the cosegregation of causal variants, causing the observed association signal, with marker alleles within pedigrees [31]. It succeeded in mapping genes and gene variants affecting Mendelian traits (e.g., single-gene disorders) [32] while failed to reliably identify complex-trait loci in human pedigrees. GWAS, based upon the principle of LD (details in Section 2.2.3) at the population level, performs the unbiased scan of the genome compared with linkage analysis in which only a few genetic markers per chromosome were used to tag a causal variant since the number of recombination events per meiosis is relatively small [31]. GWAS represents an important step beyond family-based linkage studies and a powerful tool for investigating the genetic architecture of complex traits.

2.1.1.3 Missing heritability

The continuous success is achieving by GWASs in the over ten years (staged reviewed in [31, 34], Figure 2.4). However, most variants identified by GWASs so far explain only a small proportion (typically $< 50\,\%$) of estimated heritability — the proportion of variation in a particular trait that is attributable to genetic

[1] By the early of this year (2019), there are over 113 million validated of the 660 million recorded variants in **dbSNP** (version 151, `https://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi`).

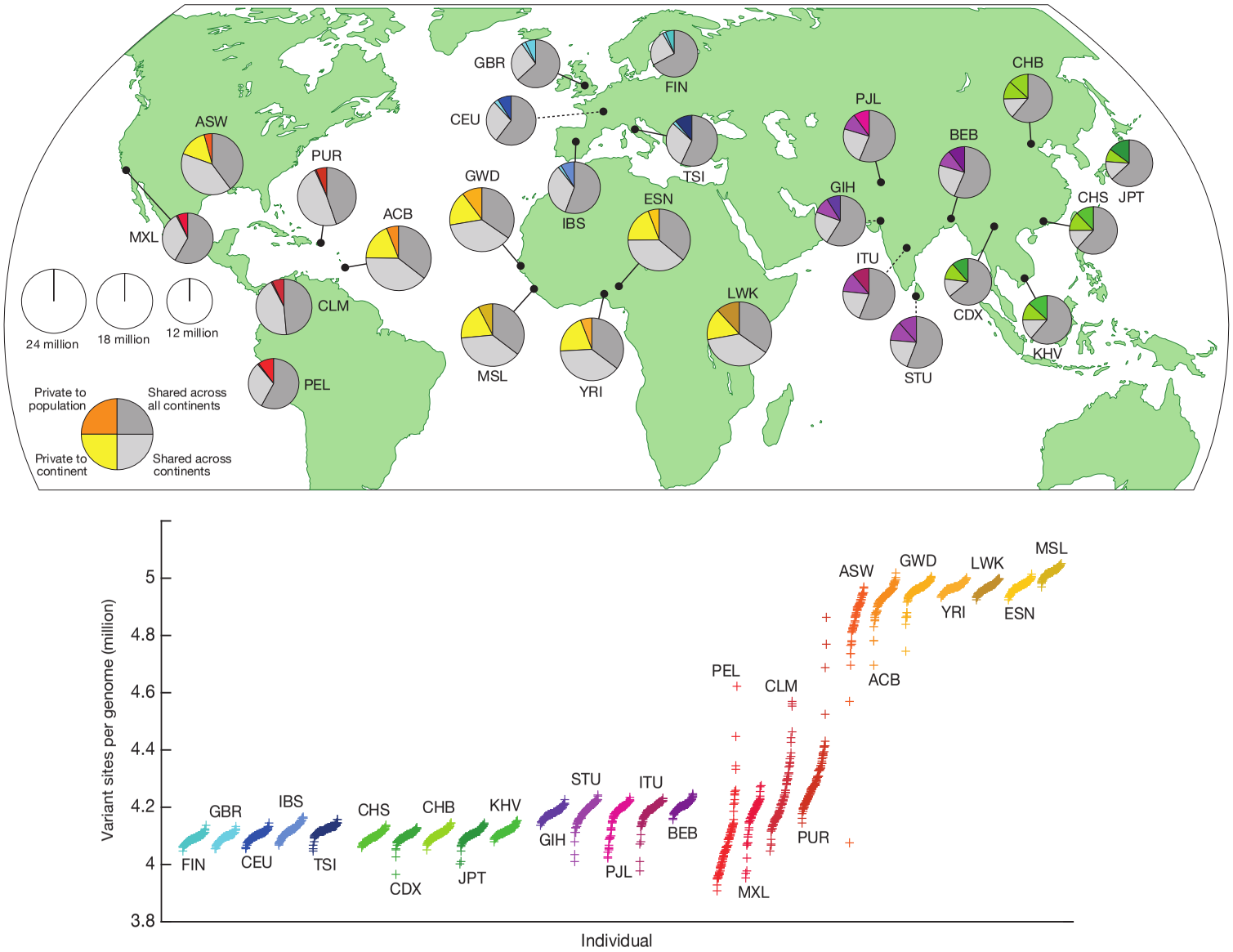

Figure 2.3: The number of variant sites per genome. Individuals from African ancestry populations harbour the greatest numbers of variant sites and individuals from recently admixed populations show great variability in the number of variants, roughly proportional to the degree of recent African ancestry in their genomes [29].

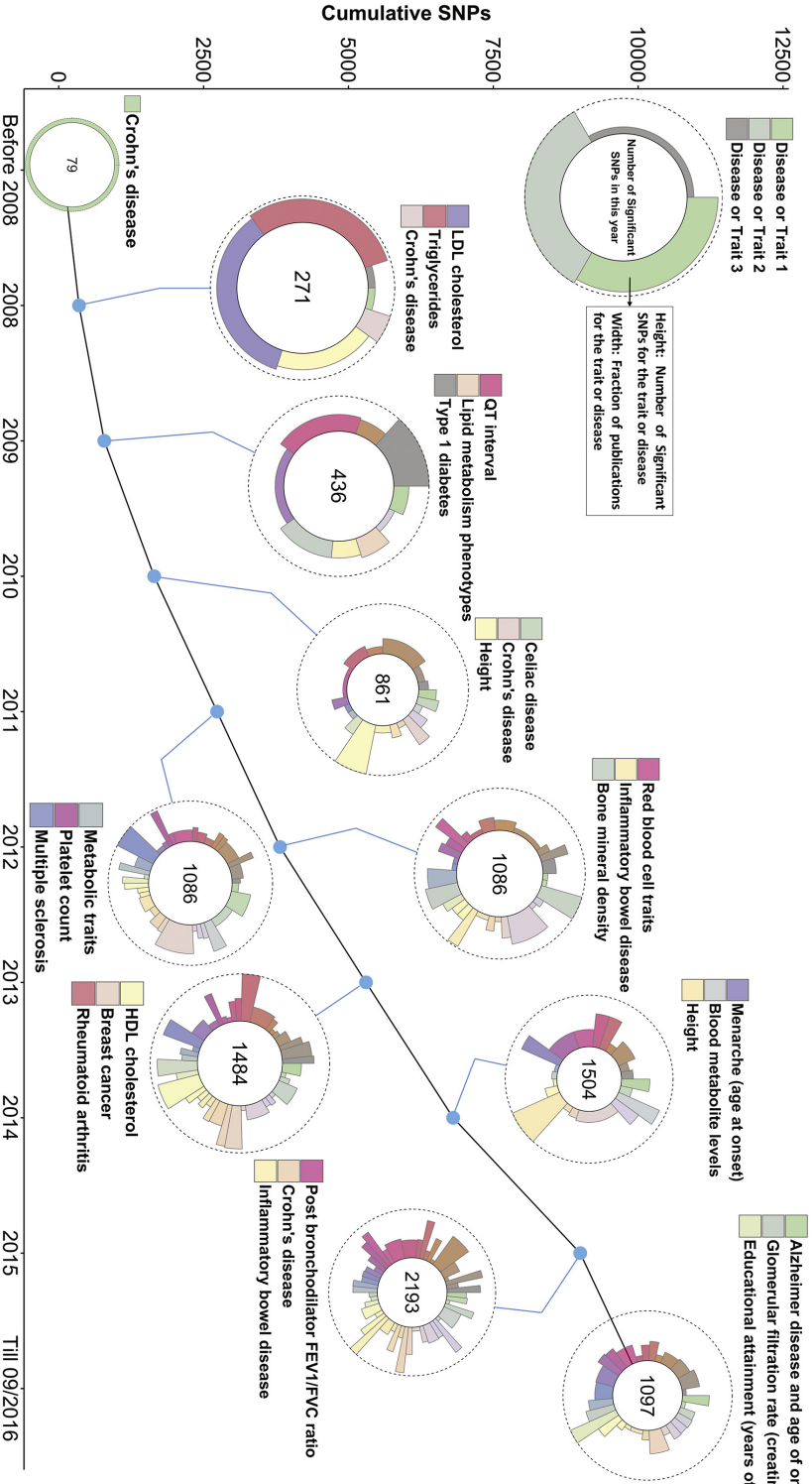

Figure 2.4: GWAS SNP-trait discovery timeline. Data used for generating the graph were taken from the GWAS Catalog [33] SNPs and traits were selected according to the following filters. SNPs were selected with a p-value $< 5 \times 10^{-8}$. For each trait with two or more selected SNPs, SNPs were removed if they had an LD $r^2 > 0.5$ (calculated from 1000 Genomes phase 3 data) with another selected SNPs and their p-value was larger. For each year of discovery, only the top three traits and diseases with the largest number of SNPs are labeled in the circle [34].

factors [2, 3]. For example, the 697 genetic variants with genome-wide significance on adult height explain only 20 % of the phenotypic variance [35]. The unexplained heritability is so-called missing heritability. Many explanations for the missing heritability have been suggested, including much larger numbers of variants with smaller effect yet to be found; rarer variants (possibly with larger effects) which are poorly detected by available genotyping arrays that focus on variants present in 5 % or more of the population; structural variants poorly captured by existing arrays; low power to detect gene-gene interactions; and inadequate accounting for shared environment among relatives (summarized in [1]).

2.1.2 Epistasis

Due to the fact that gene-gene interaction could be an explanation of the missing heritability (discussed in Section 2.1.1.3), epistasis, defined generally as the interaction between different genes, is a hot topic of discussion in complex trait genetics in recent years [36, 37, 38, 39, 3, 21, 40].

2.1.2.1 Genotypic contexts of epistasis

The term "epistasis" has many different meanings. Over one hundred year ago, Bateson and Punnett coined the term ("epistatic") to describe a masking effect whereby a variant or allele at one locus (denoted at that time as an "allelomorphic pair") prevents the variant at another locus from manifesting its effect on a phenotype [41]. In this sense, the variants must be interacting with one another, at least in the loose sense that they exist within pathways that both influence the same phenotype. Later, Fisher [42] used a derivative of the term "epistacy" to average any statistical deviation from the additive combination of two loci in their effects on a phenotype, which has been rapidly adopted as "epistasis" by population geneticists.

Since then, the term "epistasis" has been expanded to describe nearly any set of complex interactions among genetic loci, which was condensed into three main categories by Phillips [37]:

functional epistasis addresses the molecular interactions that any genetic elements have with one another, e.g., the interaction consist of proteins that

operate within the same pathway or of proteins that directly complex with one another [43].

compositional epistasis describes the ways a specific genotype is composed and the influence that this specific genetic background has on the effects of a given set of alleles. It can be expanded to include genetic interactions beyond those that are exposed in the double mutant homozygote, which is the definition of epistasis used in modern systems biology. Since enumerating all possible genetic interactions for any real population is not possible, a compositional epistasis approach can not be formally applied to natural populations.

statistical epistasis attributed to Fisher is the average deviation of combinations of alleles at different loci estimated over all other genotypes present within a population. It does not simply mean that phenotypes are measured quantitatively, but that they are sampled from a population as opposed to being intentionally constructed, as is described by compositional epistasis [37].

Functional epistasis and compositional epistasis are jointly considered by Sackton as physiological epistasis [44], in a broad sense, referring to any situation in which the genotype at one locus modifies the phenotypic expression of the genotype at another locus [3]. Statistical epistasis, according to Cordell, mostly relies on the concept of a linear model that describes the relationship between an outcome variable and a predictor variable or variables [45].

Ultimately, the variety of the meaning for epistasis may be the reason that there are various uses or applications of epistasis.

2.1.2.2 Epistasis as a tool

Since the first application conducted by Bateson and Punnett revealed a pattern of epistasis influencing flower color in sweet peas [41], epistasis analysis has utility in the human genetic analysis of biosynthetic pathways, developmental pathways, and other genetic networks. For instance, with WTCCC data, Wan et al. have found many gene-gene interactions in the MHC region for type 1 diabetes (T1D) and one SNP pair for Crohn disease (CD) [46]. While Lippert et al. performed an exhaustive epistatic search on expanded WTCCC data and did not detect any genetic interaction for CD. Instead, they identified several genetic interactions

associated with coronary artery disease (CAD) and one epistatic interaction for bipolar disorder (BD) [47]. The difference between the results from Wan et al. and Lippert et al. might come from that the different amount of data and different calculation methods they used [47]. The failure of replication may be the reason for the debate on the role of epistasis on human genetics.

However, the unreplicable situation was broken in 2014. Hemani et al. identified 30 pairs of SNPs that interacted to affect the expression of 19 different gene transcripts [38]. These interactions were robust to adjustment for multiple testing and were successfully replicated across two independent studies. Most of the replicated apparently interacting SNP pairs were associated with gene expression in *cis* and were located close to each other on the same chromosome (all $<$ 520 kilo base pairs (Kb)). Wood et al. replicated 11 of the *cis–cis* pairs in the InCHIANTI dataset [48]. The phenomenon of *cis–cis* interaction was also observed in a recent large-scale statistical interaction analysis, in which one SNP pair at the LPA locus have identified that epistatically affects CAD susceptibility [49].

Apart from the lack of replication across different datasets, the epistasis concept is controversial in human genetics also due to the lack of well-characterized examples. Nevertheless, recently, some researchers from New Zealand and Australia have shown the clinic evidence that two mutations of TNFRSF13B/TACI as well as TCF3 genes interact in an epistatic way causing severe immunodeficiency and autoimmunity in the digenic proband [50, 51]. Though epistasis has not been reported to be widespread, the clinical findings encourage researchers to put more efforts into epistasis studies.

2.2 Analysis challenge

The studies using epistasis to analyze the structure of genetic pathways have utilized a small set of genes that had previously been identified to affect the trait with single-mutant analysis. Nevertheless, the entire premise of epistasis is that genetic interaction can influence phenotypes when found in combination with one another. It is better to conduct a large-scale systematic study of the possible pairwise interactions between all genes [37]. The datasets investigated with GWAS can be appropriate sources to search for pairwise epistasis.

However, searching for epistatic effects on a genome-wide level is confronted with many challenges, including the same statistical issues as in GWAS resulting from the genetic part, e.g., false positive caused by population heterogeneity, and the computational expense due to the increasing sample sizes and scale of genotyping.

2.2.1 Genotype call rate and allele frequencies

Genotype call rate, calculated by $\frac{number\ of\ genotyped\ individuals}{total\ number\ of\ individuals}$, is a good indicator of marker quality. Large variations exist in DNA sample quality which can have substantial effects on genotype call rate and genotype accuracy. The below-average call rates and accuracy usually happen in samples of low DNA quality or concentration. Thus, genotype failure rate, also called missing genotype rate, can be one measure of DNA sample quality. Typically, SNPs (or variants, depending on the studies) with lower call rate than, approximately, 98 % to 99 % are recommended to be removed from analysis while the threshold may vary from study to study based on the sample size.

Allele frequency, defined as the relative occurence of every single allele in the population, can be calculated by $\frac{number\ of\ individuals\ with\ specific\ allele}{2 \times total\ number\ of\ individuals}$. The minor allele frequency (MAF) refers to the frequency of the allele in a locus appearing with a low rate in the population. It is important to filter SNPs based on MAF since the statistical power is extremely low for rare SNPs. Besides, the SNPs with extremely low MAF also have the potential to lead to spurious associations due to either genotyping errors or population stratification [52]. The clustering algorithms can be challenged for making genotype calls when SNP with low MAF are involved, usually resulting in poor clusters. Fake associations, caused by potential stratification issues rather than disease association, happen if specific alleles are present only in certain ancestral populations, with low MAF in those populations.

In order to avoid misclassification, bias, high false positive or negative rates, classically, a MAF threshold of 1 % to 5 % is applied to remove all SNPs with very low MAF. Studies with small sample sizes may require a higher threshold [53].

2.2.2 Hardy-Weinberg equilibrium

Hardy-Weinberg equilibrium (HWE) is a principle stating that allele and genotype frequencies in a population will remain constant from one generation to the next in the absence of disturbing factors. In a natural population, the mating is random (without disturbation) which means the frequencies, theoretically, are constant. However, HWE can be influenced by many factors, for example, genetic drift, mate choice, mutation, gene flow, population bottleneck, and founder effect. The Hardy-Weinberg equation states as

$$p^2 + 2pq + q^2 = 1 \tag{2.1}$$

where assumes the frequency of major allele (A) is p and the frequency of minor allele (a) is q. p^2 is the predicted frequency of homozygous people with major alleles (AA) in a population, $2pq$ is the predicted frequency of heterozygous (Aa) people, and q^2 is the predicted frequency of homozygous people with minor alleles (aa).

For each genetic variant in genome-wide dataset, a χ^2 goodness-of-fit test is commonly used to examine whether observed genotypes conform to Hardy-Weinberg expectations. Wigginton proposed exact tests to control type I error inflating in χ^2 test [54]. The test statistics can show the deviations from HWE which may suggest problems with genotyping or population structure or imputation inconsistencies or, in samples of affected individuals, an association between the genetic variants and disease susceptibility. The p-value threshold for declaring genetic variants to be in HWE has varied significantly between studies. The reason is that although departure from HWE can indicate potential genotyping error, disequilibrium can also result from a true association. It has been consistently noted that many more SNPs are out of HWE at any given significance threshold than would be expected by chance [52]. SNPs severely out of HWE should therefore not be eliminated from the analysis while flagged for further analysis after the association analyses are performed while in practice, the variants with HWE p-value $< 5 \times 10^{-5}$ are typically removed from the study.

2.2.3 Linkage disequilibrium

LD, the non-random association of alleles at different loci in a given population, can also be affcted by the factors influencing HWE. The LD concept is related to chromosomal linkage, where two variants on a chromosome remain physically joined on a chromosome through generations of a family. As Figure 2.2.3 from Bush described, chromosomal segments are broken apart by recombination events within a family from generation to generation. Through generations, the effect is amplified. Until all alleles in the population are in linkage equilibrium or are independent, repeated random recombination events keep breaking apart segments of a contiguous chromosome (containing linked alleles) [55]. Namely, the rate of LD decay decreases according to the increasing number of generations (Figure 2.2.3). Since more recombination events have been accumulated in African-descent populations, the most ancestral, they have smaller regions of LD than European-descent and Asian-descent populations, created by founder events (a sampling of chromosomes from the African population).

The level of LD between two loci with allele A (minor allele: a) and allele B (minor allele: b) is commonly measured by D' and r^2. If $f(A)$, $f(B)$, $f(a)$, and $f(b)$ represent the frequency of allele A, B, a, and b, respectively. $f(AB)$, $f(Ab)$, $f(aB)$, and $f(ab)$ represent the different combinations among alleles. Then D' is defined as

$$D' = \begin{cases} \frac{f(AB)f(ab)-f(Ab)f(aB)}{\min(f(A)f(b),f(a)f(B))} & \text{when} \quad f(AB)f(ab) - f(Ab)f(aB) > 0 \\ \\ \frac{f(AB)f(ab)-f(Ab)f(aB)}{\max(-f(A)f(B),-f(a)f(b))} & \text{when} \quad f(AB)f(ab) - f(Ab)f(aB) < 0 \end{cases} \tag{2.2}$$

and

$$r^2 = \frac{(f(AB)f(ab) - f(Ab)f(aB))^2}{f(A)f(B)f(a)f(b)} \tag{2.3}$$

Both values, D' and r^2, range from 0 to 1, where 0 implies frequent recombination between the two variants and statistical independence under principles of

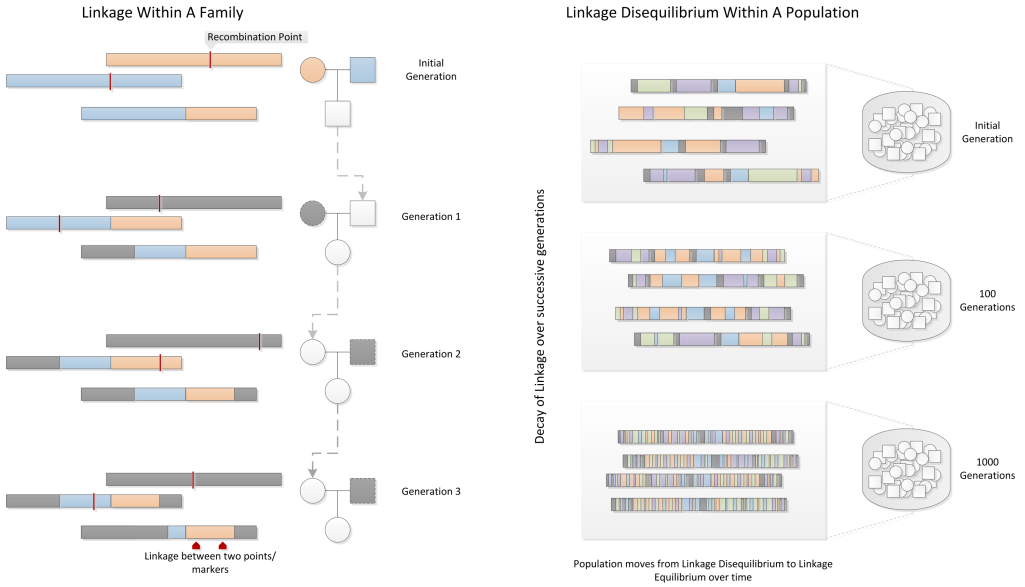

Figure 2.5: Linkage and LD. Within a family, linkage occurs when two genetic markers (points on a chromosome) remain linked on a chromosome rather than being broken apart by recombination events during meiosis, shown as red lines. In a population, contiguous stretches of founder chromosomes from the initial generation are sequentially reduced in size by recombination events. Over time, a pair of markers or points on a chromosome in the population move from linkage disequilibrium to linkage equilibrium, as recombination events eventually occur between every possible point on the chromosome [55].

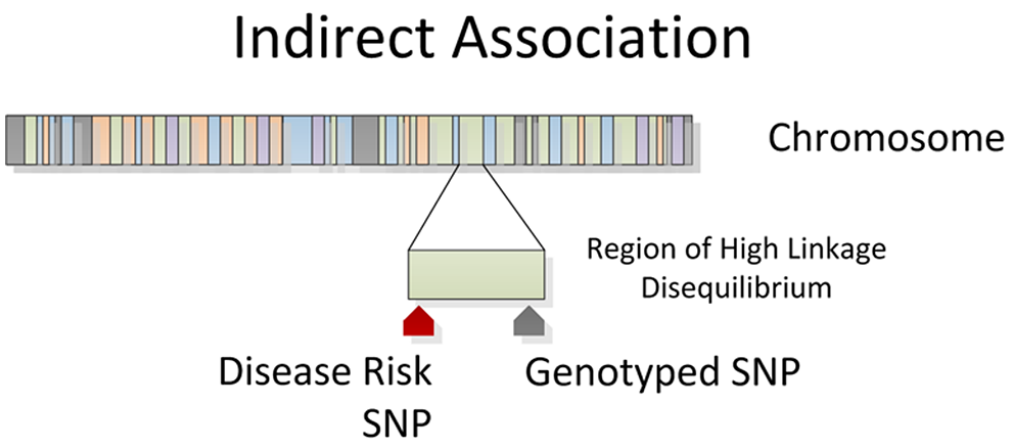

Figure 2.6: Indirect association . Genotyped SNPs often lie in a region of high linkage disequilibrium with an influential allele. The genotyped SNP will be statistically associated with disease as a surrogate for the disease SNP through an indirect association [55].

HWE and 1 means a complete LD, indicating no recombination between the two variants within the population.

The existence of LD creates the different outcomes from a genetic association study. In GWAS, tag SNPs refer to the SNPs selected specifically to capture the variation at nearby sites in the genome. If a tag SNP is associated with a specific trait, it can be the exact signal (direct association) or the indirect association indicating within the LD region of the tag SNP, there is a risk SNP (Figure 2.6) [55]. Thus, the detected SNP might not be the causual variant due to the presence of indirect association. For the interaction analysis, strong LD between the tested SNP pair, according to Cordell [45], might cause the fake interaction suggesting that reducing the number of variants in the variant set might be required, not only for computational reasons, but also to ensure that the variants are in low LD.

2.2.4 Population stratification

The above has explained several factors which influence the genetic studies, in which "population" has been mentioned frequently. Figure 2.3 shows only the difference in the number of genetic variant sites among different populations. Besides, the allele frequencies and LD structures can also differ through populations. In modern genome-wide studies, large sample size and admixture cohorts

have been involved. The allele frequency differences between different cohorts or the hidden groups within the cohort can lead to false positive associations driven for instance by unexpected relatedness of individuals. The population stratification, allele frequency differences between different groups due to systematic ancestry differences, is also an issue for genetic studies.

Efforts have been made to remove or reduce the effect of population stratification through the removal of individuals of divergent ancestry. Price and his colleagues have reported the limitations of using genomic control [56] and structured association [57, 58] for population stratification correction [59]. The uniform adjustment in genomic control may lead it to be insufficient for markers having unusually strong differentiation across ancestral populations and be superfluous for markers devoid of such differentiation, resulting in a loss in power. Structure association is limited by the computational cost on large datasets and the predefined number of clusters. They proposed correcting for stratification with Principal Component Analysis (PCA), known as EIGENSTRAT, which identifies continuous axes with large genetic variation estimated by several top eigenvectors of the pairwise correlation matrix for subjects, and then treats those axes as continuous covariates in the association analyses. The linear projections are assumed to suffice to correct for the effect of stratification while Kimmel presented evidence from a simulation study that EIGENSTRAT correction will often be inaccurate (details in [60]). Meanwhile, Patterson recommended filtering out loci with high LD in large data to minimize distortion in PCA due to high correlation in adjacent SNPs based on multivariate regression analysis for each marker in the SNP matrix [61]. Later, a method using classical Multidimensional Scaling (MDS) corroborates the use of independent SNPs [62]. The open-source tool PLINK [63] has provided a MDS-based approach to population stratification which uses whole genome SNP data to measure the pairwise identity-by-state (IBS) distance matrix of each individual then conduct MDS. It turns out that the approaches based on PCA and MDS become widely applied on genome-wide genetic studies for population stratification and the PCA-components or MDS-components are utilized as covariates in the statistical analysis.

2.2.5 Haplotype and imputation

A haplotype is a group of alleles on the same chromosome that are inherited together from a single parent (Figure 2.1). In genetic study of populations, it is

common to estimate haplotypes since the precise sequence (phase) of alleles on each homologous copy of a chromosome is not directly observed by genotyping and must be inferred by statistical methods [64]. Estimated haplotypes can be further used for genotype imputation to increase the power of association studies by inferring missing genotypes, harmonizing data sets for meta-analyses, and increasing the overall number of markers available for association testing [65]. The process of haplotype estimation before genotype imputation is called pre-phasing [66]. Figure 2.7 explains how genotype imputaion is processed.

Several programs have been developed to perform genotype imputation, such as, PLINK, Beagle [68], and IMPUTE2 [69]. Among them, IMPUTE2 is computationally intensive but provides a better estimate of missing genotypes compared with other methods because it takes into account all available markers when imputing each missing genotype [65]. The phasing process in IMPUTE2 has been reported to be less accurate than the phasing method called SHAPEIT2 [66, 70]. Thus, it is recommended to utilize SHAPEIT2 to estimate the haploptypes from genotype or sequencing data then impute the genotypes with IMPUTE2.

2.3 Computational challenge

Despite the challenges on the genetic issues, due to the impressively fast progress in high-throughput sequencing, the massive data production confronts genetic researchers with challenges in the areas of data mining, memory on device, and computational efficiency. Performing a genome-wide pair-wise epistasis study for n SNPs requires $\frac{n\times(n-1)}{2}$ tests, indicating GWIS faces a more serious computational challenge compared to GWAS.

2.3.1 High performance computing with central processing unit(s)

The development of programming models and standards have facilitated the deployment of parallel applications. For example, the widely used techniques: Message Passing Interface (MPI) [71] for distributed memory systems and Open Multi-Processing (OpenMP) [72] for shared memory systems. Upton and his colleagues compared these two models on the advantages and disadvantages from the programming flexibility and scalability to the performance (see [73]). Hybrid MPI+OpenMP can be applied on modern clusters that connect multiple shared

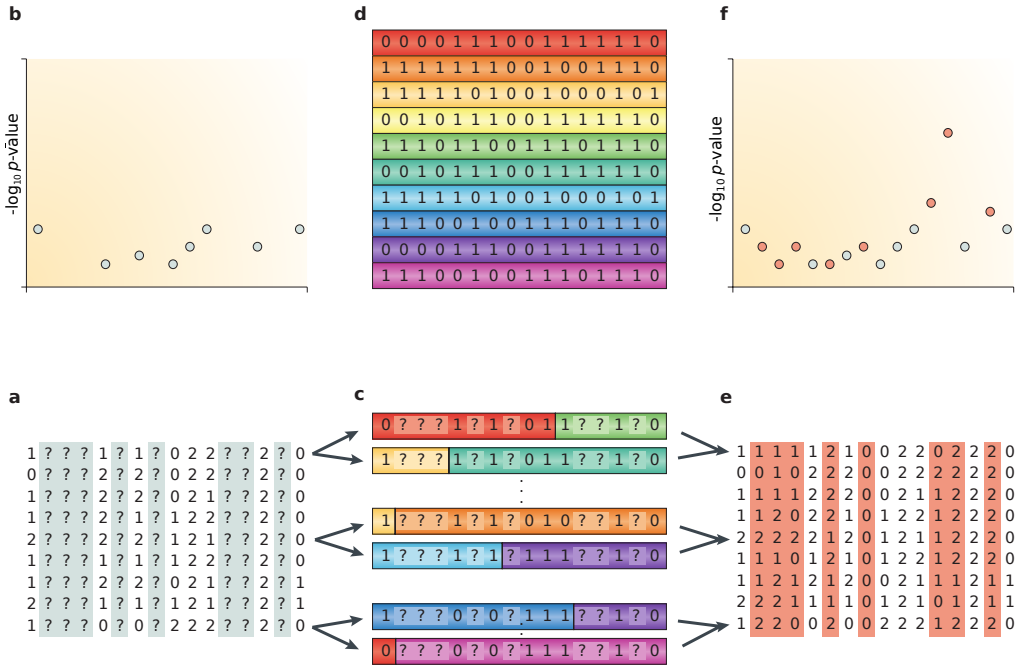

Figure 2.7: How genotype imputation works.

a. Genotype data with missing data at untyped SNPs (grey question marks). The raw data consist of a set of genotyped SNPs that has a large number of SNPs without any genotype data. Genotyped SNPs often lie in a region of high linkage disequilibrium with an influential allele.

b. Testing association at typed SNPs may not lead to a clear signal. Testing for association at just these SNPs may not lead to a significant association.

c. Each sample is phased and the haplotypes are modelled as a mosaic of those in the haplotype reference panel. The figure highlights three phased individuals.

d. Reference set of haplotypes, for example, HapMap. These haplotypes are compared to the dense haplotypes in the reference panel.

e. The reference haplotypes are used to impute alleles into the samples to create imputed genotypes (orange).

f. Testing association at imputed SNPs may boost the signal [67].

memory nodes with multi-core CPUs. MPI enables communications between the nodes on the cluster and OpenMP feeds the cores on each node. The use of this hybrid programming model can bring improved reduced memory consumption and reduced communication needs [73].

Performing epistatic analysis on clusters can reduce the runtime from years to days due to the benefits from parallel computing environments. A resource manager, such as SLURM [74], provides the cluster users opportunity to define resources, from which epistatic analysis can be deployed across multiple cores in parallel on a cluster. Linear or near-linear speed-ups in execution time can be achieved depending on the algorithm, implementation, and optimization. One example shown by Upton [73] is that a 600-days-running task would be completed in just over 2 days when using 300 cores.

Accessing High Performance Computing (HPC) facilities via a cluster to parallelize the task of epistatic analysis across multiple cores is one of the solutions to manage the computational burden.

2.3.2 Graphics processing unit and CUDA

An alternative approach is to harness the power of modern consumer graphics cards. The GPGPU[2] is the form of stream processor (or a vector processor), running compute kernels. GPGPU turns the massive computational power of a modern graphics accelerator's shader pipeline into general-purpose computing power, as opposed to being hard wired solely to do graphical operations. The programmable GPU, driven by the insatiable market demand, has evolved into a highly parallel, multithreaded, manycore processor.

Several companies are making efforts on the development of the GPU. Among them, NVIDIA currently has the most successfuls GPU products. The NVIDIA Tesla-class GPUs are composed of an array of Graphics Processing Clusters (GPCs), Texture Processing Clusters (TPCs), Streaming Multiprocessors (SMs), and memory controllers. For example, a full Tesla P100 consists of six GPCs, 60 Pascal SMs, 30 TPCs (each including two SMs), and eight 512-bit memory controllers (4096 bits total). Each SM of Tesla P100 is partitioned into two processing blocks, each having 32 single-precision **CUDA** Cores (also called

[2] The GPU term in this thesis is used in the sense of the GPGPU, going to general computations beyond the originally intended graphics applications.

Streaming Processors (SPs)), an instruction buffer, a warp scheduler, and two dispatch units (Figure 2.8).

CUDA is the hardware and software architecture that enables NVIDIA GPUs to execute programs written with C, C++, and other languages. A **CUDA** program calls parallel kernels, which execute in parallel across a set of parallel threads. The GPU instantiates a kernel program on a grid of parallel (pre-defined number of) thread blocks. Each thread within a thread block executes an instance of the kernel and has a thread ID within its thread block, program counter, registers, per-thread private memory, inputs, and output results (Figure 2.9). The **CUDA** parallel model distributes efficiently the data from the global memory into shared memory on the device (GPU). Besides, similar to the idea using a CPU cluster, it distributes the task into a number of independent subtasks.

The development of GPU and **CUDA** parallel programming model enable researchers to evolve the tools for epistatic analysis. A number of epistasis analysis tools, taking advantage of **CUDA**, have been proposed, such as, SHEsisEpi [4], GBOOST [5], cuGWAM [6], GLIDE [7], SNPsyn [8], and SingleMI [9].

Except **CUDA**, **OpenCL** is an alternative open-standard Application Programming Interface (API) for GPUs, not only NVIDIA GPUs but also for the others, which was originally geared toward a heterogeneous mix of target devices such as cable set top boxes, smart phones, and desktop CPUs. Hemani developed epiGPU employing **OpenCL** for exhaustive pairwise epistasis scans [76]. While **OpenCL** code is more portable, developing in **OpenCL** is more labor-intensive than in **CUDA** [77].

2.4 Multi-phenotype studies

Apart from the analysis of binary phenotype or single quantitative phenotype, researchers are paying close attention to multiple phenotypes studies. Jointly analyzing multiple traits may boost power to detect novel associations [13], measure heritable covariance between traits [14], and has the potential to make causal inference between traits [78].

[3] GPU clock is the GPU clock speed, measured in megahertz (MHz), also called engine clock.

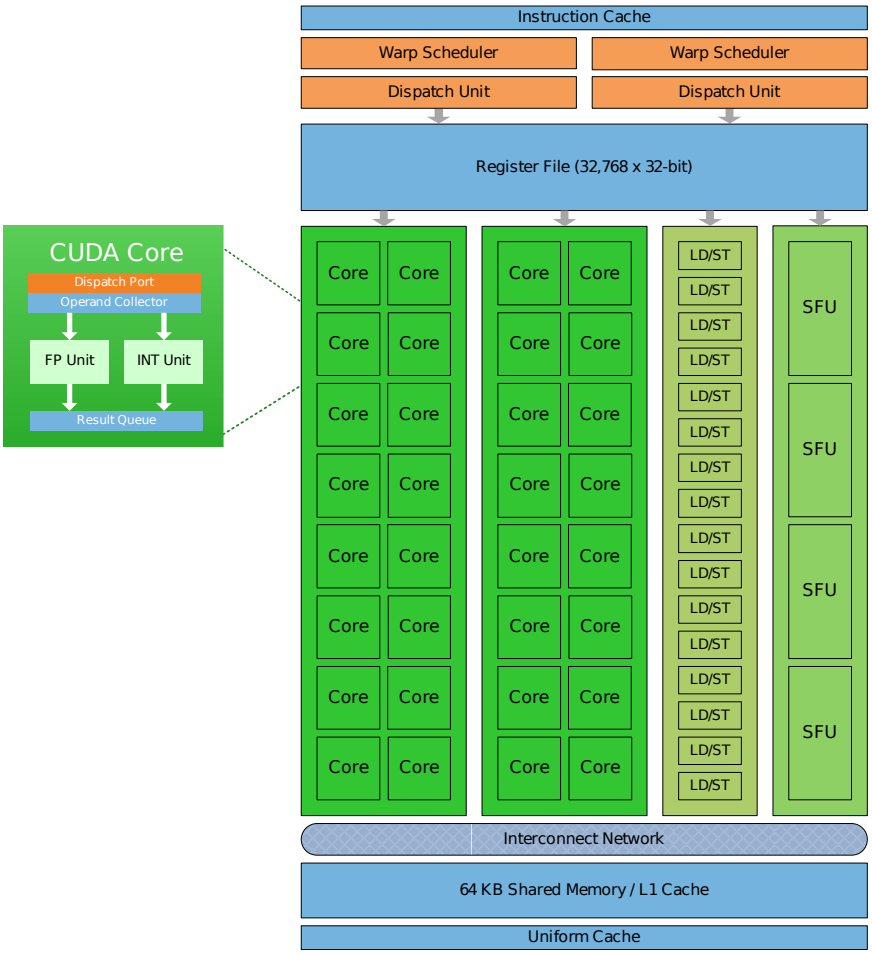

Figure 2.8: NVIDIA Fermi architecture. Eash SM features 32 single-precision **CUDA** cores, 16 load/store units (LD/STs), four Special Function Unitss (SFUs), a 64 Kilobyte (KB) block of high speed on-chip shared memory (L1+Shared Memory subsection) and an interface to the L2 cache. LD/ST allows source and destination addresses to be calculated for 16 threads per clock[3]. It loads and stores the data from/to cache or DRAM. Each SFU executes one instruction per thread, per clock. The SM executes threads in groups of 32 threads called a warp, which conducts over eight clocks. The SFU pipeline is decoupled from the dispatch unit, allowing the dispatch unit to issue to other execution units while the SFU is occupied. Each core has both floating-point (FP) and integer (INT) execution units. Convention in figures: orange for scheduling and dispatch, green for execution, and light blue for registers and caches [75].

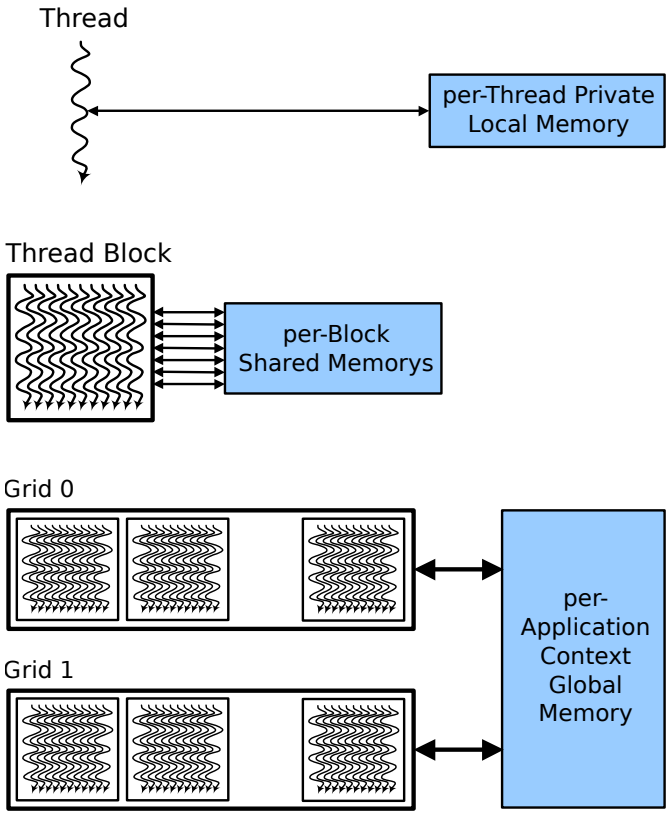

Figure 2.9: CUDA Hierarchy of threads, blocks, and grids, with corresponding per-thread private, per-block shared, and per-application global memory spaces [75].

2.4.1 Multiple correlated phenotypes

A phenotype is defined from physical or chemical measure(s) to represent the specific biological function(s), e.g., human height. For complex disease, the definition of phenotype (case or control) remains somewhat ad-hoc [13]. Usually, the diagnosis relies on a complex range of overlapping clinical characteristics, especially for neuropsychiatric disorders. These clinical characteristics can be also called endophenotypes which are typically correlated.

Furthermore, family (including twin and adoption) studies suggest correlated familial-genetic liabilities to psychiatric disorders. Phenotypic and genetic overlap has also been observed, for example, Lee and co-workers have provided molecular evidence for the sharing of genetic risk factors across key psychiatric disorders [15] and estimated a significant positive genetic correlation between risk of type 2 diabetes (T2D) and hypertension [14].

2.4.2 Multivariate analysis of multiple phenotypes

Correlated phenomena among endophenotypes and multiple disorders of interest lead the attention to multi-phenotype studies, from developing suitable methods for joint-analysis to applications on diverse complex traits. For example, O'Reilly and Bottolo have developed **MultiPhen** and **R2GUESS** to provide approaches for multi-phenotype analysis [13, 20]. They also reported more genetic factors detected from multi-phenotype studies which are associated with several blood lipid traits and have not been identified by single phenotype studies.

Webber has discussed the problem in epistasis analysis from phenotype definition and suggested to study the similar behavioural traits or endophenotypes, instead of treating disease status in the simplest manner as a dichotomous trait (either present or absent), which might be more likely to detect epistatic effects influencing the disease [21]. However, to our knowledge, the above packages, **MultiPhen** and **R2GUESS**, and other methods for multi-phenotype analysis fail to address the genetic interactions. Though **R2GUESS** can analyze multiple genetic variants together, the result shows only the assciation between a set of variants and the phenotypes of interest, which can not explain whether the genetic variants interacted with each other or not.

To deal with multiple phenotypic dimensions in epistasis analysis, there are several solutions from GWASs that can be taken advantage of. For instance, performing dimensionality reduction on the phenotype space then apply more standard methods for epistasis detection. Some attempts in association studies have been reported which use PCA to find directions in the phenotypes that explain most variability [79, 80]. Apart from performing dimensionality reduction on the phenotypes, there are also methods with abilities to handle multiple correlated outcomes directly that have been applied to genetic studies, such as multivariate regression [81, 13, 82], Canonical Correlation Analysis (CCA) (or CCA-based methods like semiparametric CCA) [17], and Fisher's product method for testing multiple phenotypes with summary statistics [83]. Although O'Reilly has proved that the standard CCA and MANOVA (implemented in **R2GUESS** [20]) can have higher inflated type I error rates when testing case-control or non-normal continuous phenotypes compared to **MultiPhen**, there is no comparison on the performance of the adjusted CCA, PCA, multiple regression, and Fisher's product method.

2.4.3 Multivariate meta-analysis

Since genome-wide genetic studies usually contain many different cohorts, meta-analysis is commonly used to combine data and identify the common effect or the reason for the variation of different effects. Currently, the statistical methodologies for performing meta-analysis of GWASs are oriented towards the "one gene, one trait" approach. As discussed above, multi-phenotype studies would be a promising research area. The advantages of multivariate meta-analysis have also been pointed out by researchers. Dimou and his coworkers hold the view that the multivariate meta-analysis jointly synthesizes the estimates arising from each study [84, 85, 86], which provides estimates for all the effect sizes and makes easier the comparison of the results, avoiding multiple comparisons and consequently, the inflation of type I error rate [87]. Furthermore, the potential correlation between the estimates is considered which yields more robust and precise estimates while the univariate approach ignores such correlation structure [88].

Developing or investigating methods suitable for multivariate meta-analysis is on demand. Many efforts have been made on multivariate meta-analysis of genetic association studies. Van Houwelingen and collaborators used a formulation as following for multivariate meta-analysis

$$\mathbf{y}_i \sim \mathbf{MVN}(\beta, \mathbf{\Sigma} + \mathbf{C}_i) \tag{2.4}$$

where $\mathbf{y}_i$ is the vector containing the p different estimates and by β, the vector of the overall means. $\mathbf{C}_i$ represents a within-studies covariance matrix, the diagonal elements of which are the study-specific estimates of the variance and are assumed known [84]. The off-diagonal elements of $\mathbf{C}_i$ correspond to the pairwise within-studies covariances. $\mathbf{\Sigma}$ represents a between-studies covariance matrix, whereas the off-diagonal elements correspond to the between studies covariances that are estimated during the fitting procedure. When the number of the freely estimated parameters increases, a structured specification for $\mathbf{\Sigma}$ can be imposed [87].

Illustrative code was provided in SAS `PROC MIXED` [84]. Bagos has shown the same for `gllamm` in Stata [89], whereas R users can utilize the **metafor** package [90]. Besides, the Stata command `mvmeta` can perform not only inferences based on either Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML)

by direct maximization of the approximate likelihood using a Newton–Raphson algorithm [91], but also the DerSimonian and Laird's method of moments [88, 92]. The same functionality has been implemented in R package **mvmeta** while the implementation is more powerful since it contains a regression function to fit the models, as well as a series of auxiliary functions and data used in examples, then to summarize the results, extract predictions, residuals, and statistics, perform statistical tests [93].

Genome-wide interaction analysis in multi-phenotypes studies can take into accout the overhead methods for multivariate meta-analysis to maximize the power of epistasis detection and to acchive robust and precise estimates.

3 Methods and Tools Development

3.1 SimPhe

3.1.1 Introduction

For complex traits, GWAS have been widely used to uncover the genetic basis in the past fifteen years [94, 95, 96]. However, the variants found through GWAS only explain a small portion of the heritability of complex traits [2]. This may be due to complicating factors such as an increased number of contributing loci and susceptibility alleles, incomplete penetrance, and contributing environmental effects. Epistasis, generally defined as the interaction between different genes, has been a hot topic in quantitative genetics for a long time [97, 98, 99]. There still is a controversy about the role of epistasis because the majority of researchers only concentrate on additive effects and most genetic variation is currently assumed additive [100, 101, 36, 102, 39]. Even for the search of epistasis, only few tools can take dominance effects into consideration. Recently, the detection of dominance or the interactions it is involved in have been reported [38, 103]. We believe it is essential to consider all the genetic effects when conducting GWIS. In addition, the performance of GWIS methods needs to be analyzed before using them in real applications.

A controllable simulation tool can be used to evaluate type I error rates for new statistical tests or power comparisons between the new tests and other existing tests [10]. In the past decade, many simulation programs have been developed aiming to generate genetic or phenotypic data. Each of them has its own pros and cons. For example, genomeSIMLA can simulate large scale genomic data in population-based case-control or family-based samples [104] but is not appropriate to test methods for detecting significant associations between genetic and phenotypic variation. SITDEM uses the observed parameters in GWAS to simulate disease/endpoint models in three different approaches (Bayes's theorem,

odds ratio, and relative risk) [105] but is limited to single locus studies. Similar limitations exist in SeqSIMLA, which uses GENOME [106] as default to simulate sequence data in unrelated case-control or family samples with user-specified disease or quantitative trait models [10]. To our knowledge, phenosim [107], epiSIM [108], and the function `trio.sim` implemented in R [109] package **trio** [110] on Bioconductor [111] are the only structured and published tools that can model epistatic interactions based on SNP genotypes without using HapMap project data. Among them, epiSIM works for generating realistic case-control samples and `trio.sim` in **trio** is developed for case-parent trios. The Python-based [112] phenosim is a nice tool to add a phenotype to genotypes simulated by coalescent-based simulators. It supports multiple output formats, as well as the input taken from the output of different coalescent simulation tools. However, it is not easy to use realistic genotype data, e. g., stored in PLINK [63] format (bed, bim, and fam files) which has been widely adopted in genetic analysis. To embed the dominance interactions with other genetic items, users have to modify the initial code in phenosim which further complicates the simulation process.

Here, we present the R package **SimPhe** [113], to simulate single phenotype or multiple (correlated) quantitative phenotypes based on genotypes with additive, dominance, and epistatic effects. Through parameters to the different functions, users can easily specify the number of quantitative trait loci (QTL), genetic effect size, the number of quantitative traits, and proportions of variance explained by the QTLs. It is convenient for GWAS or GWIS tool developers to test their methods, and for users who would like to compare the performance of such methods, especially the biological data analyst that wants to apply methods in real datasets and is interested in knowing which method performs better for a certain trait or traits with correlation.

3.1.2 Epistasis model

There are different models for modeling epistasis [42, 114, 115, 116]. Among them, Cockerham's model [114] has been reported as more appropriate than the other models for the study of epistasis between genes [117]. In this article, we use multi-locus-two-allele (G2A), representing the Cockerham model, to simulate the

phenotype(s) [118]: for a single phenotype, y, the relation to the genetic effects can be expressed as[1]

$$y = \sum_{p=1}^{n} G_{pij} + \varepsilon \tag{3.1}$$

where $p \in \{1, \ldots, n\}$ is the index of the epistatic pair (SNP pair). The minor allele counts at locus A and B are given by i and j, respectively (i. e., $i, j \in \{0, 1, 2\}$). For the examples presented below we will assume SNP A with genotype Aa at locus A and SNP B with genotype Bb at locus B (the minor alleles would be a and b, respectively). The random effect is represented by ε. Under the assumption of Hardy-Weinberg and linkage equilibrium, for a single interactive SNP pair the genetic value G_{ij} is given by

$$G_{ij} = \beta_0 + \sum_{t=1}^{8} \beta_{G_{w_t}} w_{tij} \tag{3.2}$$

where $\beta_{G_{w_t}}$ are the regression coefficients with $t \in \{1 \ldots 8\}$. The w_{tij} are the scale components of genotype ij for the t-th contrast. Four are the linear and quadratic orthogonal contrasts for locus A (locus B), the statistical linear and quadratic terms correspond to the genetic additive and dominance terms, respectively. The remaining four represent the interaction scales:

w_1	linear	additive of locus A
w_2	quadratic	dominance of locus A
w_3	linear	additive of locus B
w_4	quadratic	dominance of locus B
$w_5 = w_1 \times w_3$	linear × linear	additive × additive of locus A and B
$w_6 = w_1 \times w_4$	linear × quadratic	additive × dominance of locus A and B
$w_7 = w_2 \times w_3$	quadratic × linear	dominance × additive of locus A and B
$w_8 = w_2 \times w_4$	quadratic × quadratic	dominance × dominance of locus A and B

[1] The equations, tables, and cooresponding descriptions below in Section 3.1.2 are based on the two publications [117] and [118].

Due to their orthogonal design, the variables in this model are mutually independent to each other and are defined as

$$w_1 = \begin{cases} 1 & \text{for AA} \\ 0 & \text{for Aa} \\ -1 & \text{for aa} \end{cases} \qquad w_2 = \begin{cases} -\frac{1}{2} & \text{for AA} \\ \frac{1}{2} & \text{for Aa} \\ -\frac{1}{2} & \text{for aa} \end{cases}$$

$$w_3 = \begin{cases} 1 & \text{for BB} \\ 0 & \text{for Bb} \\ -1 & \text{for bb} \end{cases} \qquad w_4 = \begin{cases} -\frac{1}{2} & \text{for BB} \\ \frac{1}{2} & \text{for Bb} \\ -\frac{1}{2} & \text{for bb} \end{cases}$$

$$w_5 = w_1 \times w_3, \quad w_6 = w_1 \times w_4, \quad w_7 = w_2 \times w_3, \quad w_8 = w_2 \times w_4$$

The vectors of orthogonal contrasts for the nine possible genotype combinations are given in Table 3.1.

Table 3.1: Eight orthogonal scales based on nine genotypes

Scale	AABB	AABb	AAbb	AaBB	AaBb	Aabb	aaBB	aaBb	aabb
G	G_{00}	G_{01}	G_{02}	G_{10}	G_{11}	G_{12}	G_{20}	G_{21}	G_{22}
w_1	1	1	1	0	0	0	-1	-1	-1
w_2	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$
w_3	1	0	-1	1	0	-1	1	0	-1
w_4	$-\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$
w_5	1	0	-1	0	0	0	-1	0	1
w_6	$-\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	0	0	0	$\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$
w_7	$-\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$	$-\frac{1}{2}$	0	$\frac{1}{2}$
w_8	$\frac{1}{4}$	$-\frac{1}{4}$	$\frac{1}{4}$	$-\frac{1}{4}$	$\frac{1}{4}$	$-\frac{1}{4}$	$\frac{1}{4}$	$-\frac{1}{4}$	$\frac{1}{4}$

3.1.3 Variation and correlation

Due to the changes in allele frequency and LD, the relative strengths of the genetic effects will be different. Then, the total genetic variance contains the covariance between different genetic effects through linkage [117]. But while sim-

ulating, we suppose the two SNPs which have epistatic effects are unlinked. Thus, the total genetic variance for one pair of interactive SNPs can be partitioned into eight independent components without covariance as

$$V_G = \sum_{t=1}^{8} V_{w_t} \tag{3.3}$$

Each variance V_{w_t} of genetic effects is contributed by its own genetic parameter without any covariance with other effects. Based on Equation 3.3, the heritability of the phenotype, h^2, contributed by an epistatic event can be written as

$$h^2 = \frac{V_G}{V_G + V_\varepsilon} \tag{3.4}$$

where V_G is the total variance of genetic effects and V_ε is the variance of the random effect. According to Cockerham and Kao [119, 117], the variance components contributed by one pair of SNPs in a population are shown in Appendix A.1 (Equation A.1, adjusted from [117]).

With the known heritability and given regression coefficients, according to Equation 3.4 and the detailed genetic variance for each type of epistatic effects in Equation A.1 (Appendix A.1), it is not difficult to get the variance of the random effect. In other words, to simulate a phenotype with specific heritability, for example, a phenotype with 40 % heritability, only the regression coefficients and heritability equal to 0.4 need to be given. There is no need to define the variation of the random effects. By this way, the simulated phenotype can be similar with the trait under analysis with certain heritability.

For multiple phenotype studies, it is quite common that the phenotypes are related. To establish the correlation among simulated phenotypes, there are two ways: one is to set shared QTLs for each phenotype and the other is to build the correlation matrix and then convert to the independent phenotypes. For the latter, the correlated phenotypes $\mathbf{Y}_{\text{new}}$ can be generated by converting the correlation matrix on independent simulated phenotypes $\mathbf{Y}$ like

$$\mathbf{Y}_{\text{new}} = \mathbf{L}\mathbf{Y}^{'} \tag{3.5}$$

where $\mathbf{Y}^{'}$ is normalized (centered and divided by standard deviation). $\mathbf{L}$ is a lower triangular matrix with real and positive diagonal entries in a Cholesky decomposition

$$\mathbf{C} = \mathbf{L}\mathbf{L}^{\top} \tag{3.6}$$

where $\mathbf{C}$ is a covariance matrix and $\mathbf{L}^{\top}$ is the transpose of $\mathbf{L}$.

Chosing which way to build the correlation for the simulated phenotypes depends on the purpose of the simulation. We highly recommend to set shared QTLs since the major objective for most of the simulations in multiple phenotype studies is to test whether the used method can find the common QTLs among phenotypes. However, since the genotypes of the set QTLs are varying, the correlation coefficient cannot be controlled. It might differ from the expected value which may be not desirable for some relatedness-oriented simulation.

3.1.4 Contents of SimPhe

In this section we describe the functions included in **SimPhe** and provide some notes on their implementation. We start by describing the main function of the package, `sim.phe`. In the following, we offer details of some of the secondary functions that could be useful for users. The key input files or parameters for the main function are also crucial for some secondary functions which are also explained.

3.1.4.1 Main function

The main function of **SimPhe** is `sim.phe`. This function contains steps that how the simulation works in this package. Among the arguments listed below, the `sim.pars` and `fgeno` are most important as they are the basis of the simulation.

A flowchart is provided in Figure 3.1 to guide the reader through the specification of the arguments for `sim.phe`.

`sim.pars:` name of the file or name of the prepared list object with genetic parameter settings.

`fgeno:` string specifying the filename of genotype information or pre-read data with genotype information.

`ftype:` genotype file format for `fgeno`.

`fwrite:` whether to write simulated data file.

`fphename:` name of the phenotype(s) file.

`fusepar:` name of output file saving the settings for the simulation.

`seed:` integer seed for the random number generator.

`Dskim:` the coefficient of linkage disequilibrium.

`noise.var:` variance of random noise.

`pattern:` ignore pattern for detecting the phenotype index from the parameter names.

`plink.path:` path of the PLINK executable.

`genetic.model:` a string specifying the genetic model to use for the simulation.

Simulation parameters. All of the information about the parameter settings related to a simulation should be given either in the prepared list or in the file named in `sim.pars`. The file format is shown in the example below:

```
[P1mean]
mean
10

[P1main]
SNP additive dominance
SNP01 4 2
SNP02 2.5 1.5
SNP03 5.5 1.8
SNP04 -2.5 1
SNP05 8.5 9.8
SNP06 8.7 8.4
```

```
[P1epistasis]
SNPA SNPB additive_additive additive_dominance dominance_additive dominance_dominance
SNP01 SNP02 10 2 2.8 1.8
SNP03 SNP04 8 3 3.5 2.5
SNP05 SNP06 20.5 20.6 20.4 10.2

[P1heritability]
heritability
0.6

[P2mean]
mean
20

[P2main]
SNP additive dominance
SNP01 8 2
SNP02 4.5 1.8
SNP03 6.5 1.5
SNP04 -3.5 -1.2
SNP07 10.4 8.6
SNP08 12.3 11.9

[P2epistasis]
SNPA SNPB additive_additive additive_dominance dominance_additive dominance_dominance
SNP01 SNP02 15 4 3 2
SNP03 SNP04 12 5 3.5 2.5
SNP07 SNP08 21.5 20.9 20.5 20.3
```

The general structure is blocks with a bracketed headline, each followed by a plain-text table with one header line followed by a variable number of data lines up to either an empty line, the next block header, or the end of the file.

The blocks starting with "[P1 . . .]" refer to phenotype 1, those with "[P2 . . .]" to phenotype 2. There should be the following four blocks for each phenotype with the columns described here (for "P1"):

[P1mean]

mean β_0: coefficient parameter of "basic" genetic effects in $G_{ij} = \beta_0 + \sum_{t=1}^{8} \beta_{G_{w_t}} w_{tij}$

[P1main]

SNP SNP name

additive coefficient of additive effect

dominance coefficient of dominance effect

[P1epistasis]

SNPA first SNP

SNPB second SNP

additive_additive coefficient for additive_additive interaction

additive_dominance coefficient for additive_dominance interaction

dominance_additive coefficient for dominance_additive interaction

dominance_dominance coefficient for dominance_dominance interaction

[P1heritability]

heritability expected heritability

Similar meanings apply to "[P2mean]", "[P2main]", ... for phenotype 2. Users can simulate multiple phenotypes by adding more information. The important and necessary items for each phenotype are "main" and "epistasis" which determine the effect sizes of genetic effects. With the optional block "heritability", the simulated phenotype can be designed with a specific propotion of genetic variance. If "heritability" is missing, there is another way to specify the heritability by assigning `noise.var`. Otherwise, using the default `noise.var` (equal to 1), the heritability of the simulated phenotype could be extremely high. We also provide some helpful functions associated with the calculation of heritability. More details can be found in Section 3.1.4.2.

Genotype data. The arguments `fgeno` and `ftype` provide information about the genotypes. Here the genotype file (`fgeno`) only needs to contain all the information of SNPs dedicating to genetic effects. In other words, only the SNPs mentioned in the parameter file for the simulation are sufficient. `fgeno` should be a string specifying the filename to get genotyope data from or a dataframe with the genotype information. If given as character string, the format of the genotype file must be specified (`ftype`) to avoid problems when reading the file. We provide three options for `ftype`:

"plink" : genotype file is written in PLINK format. For this option (default), `fgeno` needs to be given without suffix and `plink.path` may need to be assigned by the user because PLINK will be run from it within **SimPhe**. We highly recommend users to check the system settings with R and PLINK when using genotype with PLINK format. There is always some uncontrollable problem with different devices, we can not guarantee that our general commands have no trouble to get access to the PLINK from R.

"ind.head" : columns are the indviduals and rows are SNPs.

"snp.head" : columns are SNPs and rows are indviduals.

For the last two options ("ind.head" and "snp.head"), `fgeno` must be the full name (with suffix and path if necessary) of the genotype file. Of course, this does not apply if `fgeno` is provided as a dataframe, in that case SNPs should be in columns, individuals in rows.

Outputs. After simulation, `fwrite` determines whether write out the simulated phenotype(s) into a file and `fphename` can specify the filename. To achieve clearly understanding and make future check, all the information in the parameter file and additional information about the simulation will be automatically recorded into a file with filename containing in `fusepar`, for example, allele frequencies and heritability for each phenotype simulation.

Linkage disequilibrium. The definitions of LD are various but it shows the nonrandom association of alleles at different loci. More details can be found in Slatkin's review [120]. For epistasis simulation, it is essential to know the relationship of the alleles at the two loci involved in an interaction. The argument `Dskim` can specify LD. However, we believe that setting `Dskim` equal to 0 is enough because there are other kinds of genetic variants one needs to pay more attention to in research, e. g., microsatellites, insertions, deletions, and inversions. Also, assuming no LD is an easy and safe way for simulation.

Other useful arguments. In the parameter file for the simulation, the user does not need to give the number of phenotypes. This will be detected automatically by matching character string with `pattern`. To capture the same simulation results every time, `seed` is an option to use. In the current version, the only option to `genetic.model` is "epistasis". Users can set unrequired genetic item(s) to zero to exclude unexpected genetic effect(s).

The detail of pipeline to simulate phenotype(s) in main function `sim.phe` is shown in Figure 3.1. Each step involves one secondary function.

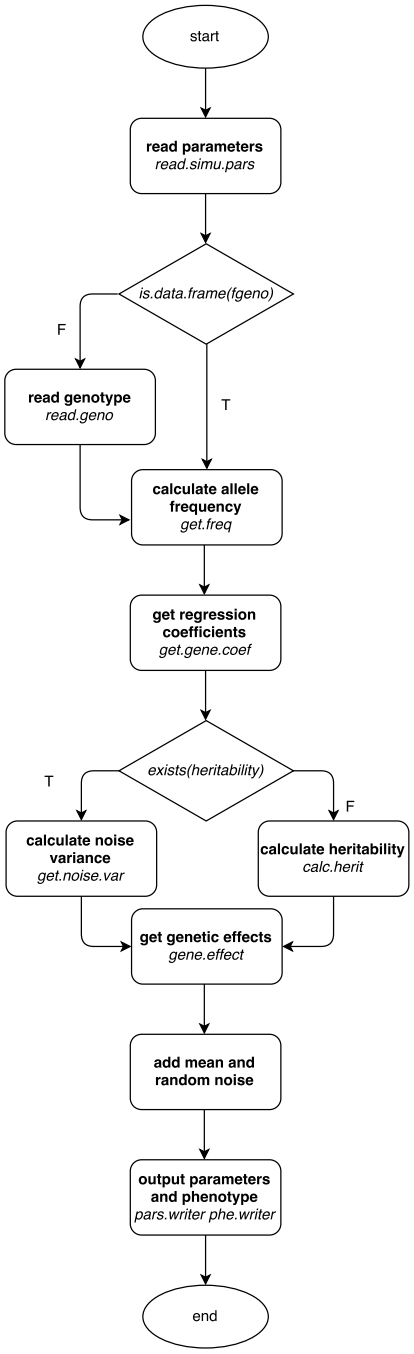

Figure 3.1: Flowchart of main function: `sim.phe` including secondary functions.

3.1.4.2 Secondary functions

In real-world datasets, there is always various correlation among different phenotypes.To simulate this kind of phenotypes, we provide the function `build.cor.phe` to build relationships between independent phenotypes. As mentioned before, there are two ways to specify correlation: either set shared QTLs or use a correlation matrix to convert data. `build.cor.phe` only works for the latter.

`pheno:` data to build correlation for.

`corMtr:` a correlation matrix, e. g., $\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$ for two variables, r is the correlation coefficient for variable 1 and variable 2.

`sdMtr:` a matrix with standard deviations, e. g., $\begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}$ for two variables, σ_1 and σ_2 are the standard deviations for variable 1 and variable 2.

`margin:` a vector giving the subscripts which the function will be applied over.

Among the 4 arguments, `sdMtr` is optional. If missing, `build.sd.matrix` will compute the standard deviation matrix based on the given data.

Below we list some additional functions which are part of the core of `sim.phe`. Figure 3.1 shows where these secondary functions are used.

Table 3.2: List of core functions in **SimPhe**

Generic name	Function
`calc.gene.var`	Calculate total genetic variance based on regression coefficients
`calc.herit`	Calculate heritability if it is not given
`get.noise.var`	Get noise variance according to the expected heritability
`gene.effect`	Get genetic effect for each individual based on the genotype
`read.geno`	Read genotype data
`read.simu.pars`	Read file specifying the simulation parameters

Even though the main purpose, simulation, can be achieved by simply running `sim.phe`, we believe secondary functions still need to be paid attention to. These functions are called from `sim.phe` but they can also be used independently.

3.1.4.3 Necessary Inputs

SimPhe requires some essential information to simulate phenotypes. They are the basis for the main function and some secondary functions.

First there are the simulation parameters defined in a file with the special format described above. The name of this file is specified as a parameter to function `sim.phe`. We provide a demo file in the package (for details, see Section 3.1.5). Users can copy this file and modify the values (numbers) to set their own simulation parameters.

The other is a file with the genotyping information. This can be the whole genome genotyping data or just the subset including the genotyping information of the QTLs referenced in the parameters file, but the names of these QTLs must exist in both files. It means there is no need not preprocess a big dataset prior to a simulation. Also, a preloaded dataframe with GWAS genotypes can be used directly from memory (even if not loaded through **SimPhe**, if it has SNPs in columns). QTLs not referenced in the simulation parameter file will be ignored. As shown in Figure 3.1, users can reference either a variable with the genotype data or give the filename of the genotype data as well as the file format to the package (for details about the supported types in this package, see Section 3.1.4.1). Genotype needs to be coded as one of $\{0, 1, 2\}$. Therefore, either a real genotype dataset from sequencing or simulated genotypic data from some simulators is applicable for **SimPhe** if it is in the right format.

3.1.5 Sample implementation

Here we demonstrate the use of **SimPhe** by applying it to a real-world genotype dataset. The simulation parameters, as shown in the last section, are randomly set without any special meaning. First, we show how easy it is to get phenotype(s) by using `sim.phe`. Before running `sim.phe`, we need to specify the parameter file and genotype file for simulation.

If not already available on your system, **SimPhe** can be installed from Comprehensive R Archive Network (CRAN) using

```
install.packages("SimPhe")
```

After installing **SimPhe**, two toy files including the information of simulation parameters and genotype exist in the package folder:

```
library("SimPhe")
fpar.path <- system.file("extdata", "simupars.txt", package="SimPhe")
fgeno.path <- system.file("extdata", "10SNP.txt", package="SimPhe")
```

Then simulate the phenotypes as designed in the parameter file after loading the package:

```
phe <- sim.phe(sim.pars = fpar.path, fgeno = fgeno.path,
               ftype = "snp.head", seed = 123, fwrite = FALSE)
```

In the parameter file, we describe two phenotypes contributed to by two common SNP pairs with epistatic effects and one independent SNP pair with epistatic effects, the simulation parameters are taken from the toy example file: simupars.txt. Users can inspect them by looking at the variable `genepars`.

```
genepars

## $P1mean
##   mean
## 1   10
##
## $P1main
##     SNP additive dominance
## 1 SNP01      4.0       2.0
## 2 SNP02      2.5       1.5
## 3 SNP03      5.5       1.8
## 4 SNP04     -2.5       1.0
## 5 SNP05      8.5       9.8
## 6 SNP06      8.7       8.4
##
## $P1epistasis
##    SNPA  SNPB additive_additive additive_dominance dominance_additive
## 1 SNP01 SNP02               10                  2                2.8
## 2 SNP03 SNP04                8                  3                3.5
## 3 SNP05 SNP06               20                 21               20.4
##   dominance_dominance
## 1                 1.8
## 2                 2.5
## 3                10.2
##
## $P1heritability
##   heritability
## 1          0.6
##
## $P2mean
##   mean
```

```
## 1   20
##
## $P2main
##     SNP additive dominance
## 1 SNP01      8.0       2.0
## 2 SNP02      4.5       1.8
## 3 SNP03      6.5       1.5
## 4 SNP04     -3.5      -1.2
## 5 SNP07     10.4       8.6
## 6 SNP08     12.3      11.9
##
## $P2epistasis
##    SNPA  SNPB additive_additive additive_dominance dominance_additive
## 1 SNP01 SNP02               15                  4               3.0
## 2 SNP03 SNP04               12                  5               3.5
## 3 SNP07 SNP08               22                 21              20.5
##   dominance_dominance
## 1                 2.0
## 2                 2.5
## 3                20.3
```

Phenotype 1 has been set with a certain heritability (`$P1heritability`) but phenotype 2 has not. With the following steps we will check whether the heritability of the simulated phenotype 1 is the same as the set one. **SimPhe** includes the coefficients and the allele frequencies for simulating phenotype 1: `gene.coef` and `allele.freq`, which is extracted from the simulation parameters.

```
gene.coefficients
```

```
## $epi.par1
##    SNPA  SNPB additiveA dominanceA additiveB dominanceB additive_additive
## 1 SNP01 SNP02         4          2       2.5        1.5                10
##   additive_dominance dominance_additive dominance_dominance
## 1                  2                2.8                 1.8
##
## $epi.par2
##    SNPA  SNPB additiveA dominanceA additiveB dominanceB additive_additive
## 1 SNP03 SNP04       5.5        1.8      -2.5          1                 8
##   additive_dominance dominance_additive dominance_dominance
## 1                  3                3.5                 2.5
##
## $epi.par3
##    SNPA  SNPB additiveA dominanceA additiveB dominanceB additive_additive
## 1 SNP05 SNP06       8.5        9.8       8.7        8.4                20
##   additive_dominance dominance_additive dominance_dominance
## 1                 21                 20                  10
```

```
allele.freq
```

```
##     SNP major.frequency minor.frequency
## 1 SNP05            0.64            0.36
## 2 SNP01            0.72            0.28
## 3 SNP03            0.74            0.26
## 4 SNP06            0.73            0.27
## 5 SNP02            0.64            0.36
## 6 SNP04            0.66            0.34
```

```
genevar <- calc.gene.var(gene.coefficients, allele.freq)
phe1var <- var(phe[, "p1"])
simuht <- genevar / phe1var
simuht
```

```
## [1] 0.63
```

The result is not the exactly 0.6 due to the (pseudo) random numbers generated in R. To get phenotype 2 with a specific heritability, for example, 0.45, we could proceed as:

```
genecoef <- get.gene.coef(
    main.pars = specify.pars(genetic.pars = genepars,
                             effect.type = "main", phe.index = 2),
    epi.pars = specify.pars(genetic.pars = genepars,
                            effect.type = "epistasis",
                            phe.index = 2))
genotype <- read.geno(fname = fgeno.path, ftype = "snp.head")
freq2 <- get.freq(geno = genotype,
                  epi.pars = specify.pars(genetic.pars = genepars,
                                          effect.type = "epistasis",
                                          phe.index = 2))
exp.noise.var <- get.noise.var(gene.coef = genecoef,
                               freq = freq2,
                               heritability = 0.45)
```

Then, when simulating a phenotype, just give this value as argument `noise.var` to function `sim.phe`. It will generate a phenotype which has a heritability close to 0.45.

As mentioned earlier, building the correlation by setting the shared interactive SNP pairs cannot be controlled. We can take a look at the correlation between the simulated phenotype 1 and phenotype 2:

```
cor.test(phe[, "p1"], phe[, "p2"])
```

```
##
##  Pearson's product-moment correlation
```

```
##
## data:  phe[, "p1"] and phe[, "p2"]
## t = 2, df = 100, p-value = 0.03
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.023 0.397
## sample estimates:
##  cor
## 0.22
```

According to the result of the correlation test, the two simulated phenotypes are significantly correlated but the correlation coefficient is small and its value cannot easily be predicted. To get a certain value, we can impose correlation by applying the correlation matrix to two independent variables. For two phenotypes, if we set different SNP pairs for each, we assume these two phenotypes are independent. Here we specify another parameter file (for `sim.pars`) and use another genotype file with more samples (for `fgeno`), compared to the genotype file within the package **SimPhe**, for `sim.phe`:

```
fpar.path <- system.file("extdata", "sep_simupars.txt", package="SimPhe")
fgeno <- 'data/geno.txt'
indphe <- sim.phe(sim.pars = fpar.path, fgeno = fgeno,
                  ftype = "snp.head", seed = 123, fwrite = FALSE)
```

We can test the correlation between the initial phenotypes with separated SNP pairs settings:

```
cor.test(indphe[, "p1"], indphe[, "p2"])

##
##  Pearson's product-moment correlation
##
## data:  indphe[, "p1"] and indphe[, "p2"]
## t = 1, df = 5000, p-value = 0.2
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.010  0.047
## sample estimates:
##   cor
## 0.018
```

Apparently, these two phenotypes are not related. To make them correlated, first specify a correlation matrix:

```
corm <- matrix(c(1, 0.6, 0.6, 1), ncol = 2)
corm

##      [,1] [,2]
```

```
## [1,]  1.0  0.6
## [2,]  0.6  1.0
```

Before applying the correlation matrix to simulated phenotypes, we would like to know what the data look like:

```
apply(indphe, 2, mean)

## p1 p2
## 20 33

apply(indphe, 2, sd)

## p1 p2
## 21 18
```

Then we can build correlation between the two initial phenotypes:

```
corphe <- build.cor.phe(indphe, corMtr = corm)
```

To test the correlation between the two new phenotypes and to see if there is any difference:

```
apply(corphe, 2, mean)

## p1 p2
## 20 33

apply(corphe, 2, sd)

## p1 p2
## 21 18

cor.test(corphe[, "p1"], corphe[, "p2"])

##
##  Pearson's product-moment correlation
##
## data:  corphe[, "p1"] and corphe[, "p2"]
## t = 50, df = 5000, p-value <2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.59 0.63
## sample estimates:
##  cor
## 0.61
```

While there is no significant difference with regard to mean and standard deviation between the initial and the new phenotypes, the correlation within the latter is now close to the value specified in `corm`. This can be visualized in scatter plots of the independent and correlated simulated phenotypes shown in Figure 3.2.

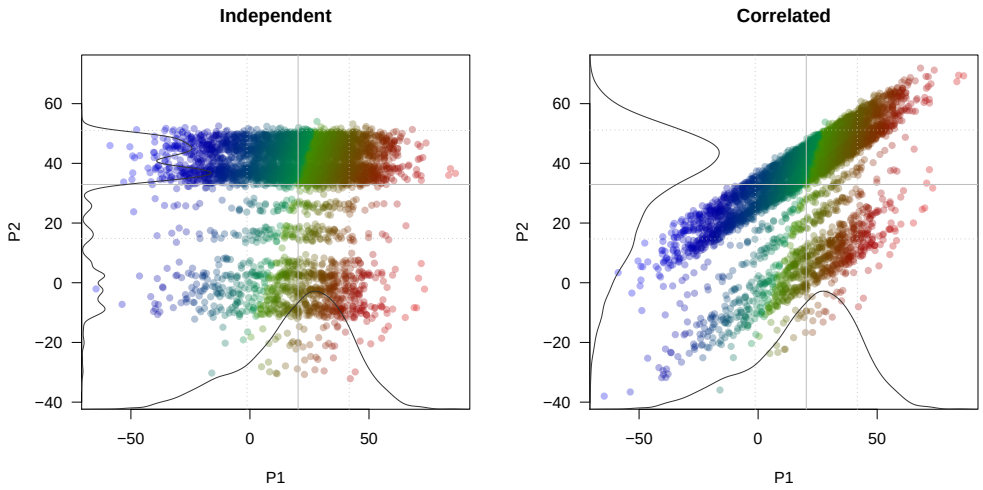

Figure 3.2: Scatter plots of the independent and correlated simulated phenotypes. P1 is not affected while P2 is sheared to reach the desired correlation. The lines indicate mean ± standard deviation and the respective distributions. The colors indicate the shift in P2, blue meaning downwards, red meaning upwards.

3.1.6 Discussion

We introduced the R package **SimPhe**, which implements the methods of the Cockerham model [114, 117, 118] to simulate phenotypes with epistasis based on genotyping data. The main features of **SimPhe** have been explained and illustrated using the examples of parameter setting through either `genepars` or the files in the installation folder which are available with the package. The simulation of quantitative trait(s) with epistatic interactions based on realistic genotypes has great potential for the assessment and improvement of GWAS and GWIS methods because the QTL(s) setting and the simulated data can provide clear information about whether designed or chosen algorithm(s) are capable of detecting causal factors associated with complex trait(s). Additionally, **SimPhe** might be a valuable addition and a nice complement to the simulators oriented to generate genotype data. It is a practical tool to build the relationship(s) between genotypes and phenotype(s). Implemented by R, **SimPhe** does not require high-level programming or advanced mathematical skills. It is easy to understand and use.

While the current release is stable, we plan enhancements of the package in the following areas: First, although the Cockerham model is suitable for epistasis studies, as a simulation tool for the phenotype(s), **SimPhe** may need to consider other models as well. Second, the interactions between loci and environment are not applied as separate items in the current version. This is an active area of research and we hope to implement a full genetic model in the future. Third, due to the fact that three locus or even higher order interactions are more complicated, currently only two locus epistatic interactions are allowed in **SimPhe** but we would like to investigate the possibility to model high-order interactions. Furthermore, in order to make complex models more easily available to researchers without the need to write R code, developing a Graphical User Interface (GUI) using **shiny** [121] is in progress.

3.2 episcan **and** gpuEpiScan

3.2.1 Introduction

Epistasis, the interaction between genes, causes hidden quantitative genetic variation in natural populations and could be responsible for the small effects, missing heritability, and the lack of replication that are typically observed for human complex traits [39]. It should be considered in genetic analysis to better understand the genotype-phenotype map. By allowing for epistatic interactions between potential genetic loci, researchers may succeed in identifying genetic variants which might otherwise have remained undetected [99]. However, mapping epistatic interactions is computationally challenging in a big dataset, especially nowadays, due to the increasing number of experimental designs, the decreasing costs of sequencing individual genomes and the prospects for high throughput [39]. Besides, improving the power of detection for exhaustive pairwise epistasis through meta-analyses on multiple existing GWASs datasets (the data for which are already readily available) is also demanding [122].

Remarkable activity in the development of methods and tools for detecting epistasis has been seen in the past few years. The developed methods, detecting whether the joint effect of two or more loci differs from that predicted by their individual effects, range from conventional regression-based methods to nature-inspired algorithms (reviewed in [122]). The available tools take advantage of modern computing facilities to reduce the runtime of the exhaustive search on epistasis. Different parallel computing models, architectures and devices have been deployed and exploited, such as **MPI** and **OpenMP** for CPU-only machines, **CUDA** and **OpenCL** for GPU machines (reviewed in [73]). However, even though these valuable methods/tools rooted in HPC make epistasis study much easier in genome-wide level, the usability is still limited since they are only avaliable through the command line [73], as well as the potential conflicts caused by the different operating systems, for example during installation.

Here, we present two R packages, **episcan** [123] and **gpuEpiScan** implemented as described in Kam-Thong's papers [124, 125], to provide efficient ways to detect pairwise genetic interactions. Both of the packages support the epistasis analysis not just in case-control study (binary phenotype) but also in quantitative trait study (continuous phenotype). To enable the selection to use and avoid the redundant installation on the different computing facilities, **episcan** built on CPU

and **gpuEpiScan** built on GPU have been developed as two independent packages, which makes users be able to select according to the avaliable computing resources.

3.2.2 Implementation methodology

3.2.2.1 Epistasis search for case-control study and quantitative trait

Both **episcan** and **gpuEpiScan** contain methods for scanning pair-wise epistasis associated with diseases and quantitative traits.

The method for epistasis detection in case-control studies is a filtering strategy simplified from the method introduced in Kam-Thong's paper (called `EPIBLASTER`) [124], which uses the difference of Pearson correlation coefficients between cases and controls. The correlation difference $\Delta\rho$ between these two groups for each pair of variants is calculated as

$$\Delta\rho(X_{(A,B)}, Y) = \frac{1}{n_1 - 1} \sum_{i:y_i=1} \tilde{x}_{A_i}\tilde{x}_{B_i} - \frac{1}{n_0 - 1} \sum_{i:y_i=0} \tilde{x}_{A_i}\tilde{x}_{B_i} \tag{3.7}$$

where $\tilde{x}_{A_i}$ and $\tilde{x}_{B_i}$ represent the two variants A and B, which are Z-score normalized within each group (case: $y_i = 1$ and control: $y_i = 0$) indicating all variants have mean 0 and variance 1. The number of subjects in each group are n_1 for cases and n_0 for controls.

The method for epistasis analysis in quantitative trait studies is a fast approach derived from Hilbert-Schmidt Independence Criterion (HSIC) [126] to test whether the two variants are independent of the phenotype. It was originally applied on epistasis detection by Kam-Thong [125] who also showed the close relationship between HSIC and linear regression by examining the derivation of estimates using the least squares regression method. Given a finite number of observations m and for each $x \in \mathcal{X}, y \in \mathcal{Y}$, empirical HSIC as reported by Gretton [126] is[2]

[2] According to Gretton [126], the empirical **HSIC** is recovered by replacing the population expectations with their empirical counterparts, and some additional manipulations.

$$\texttt{HSIC}(Z, \mathcal{F}, \mathcal{G}) \coloneqq (m-1)^{-2}\mathbf{tr}(\mathbf{KHLH}) \tag{3.8}$$

where $\mathcal{F}$ and $\mathcal{G}$ are reproducing kernel Hilbert spaces on $\mathcal{X}$ and $\mathcal{Y}$ with associated kernels $k\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ and $l\colon \mathcal{Y}\times\mathcal{Y}\to\mathbb{R}$. $Z \coloneqq \{(x_1, y_1), \ldots, (x_m, y_m)\} \subseteq \mathcal{X}\times\mathcal{Y}$. $\mathbf{H}, \mathbf{K}, \mathbf{L} \in \mathbb{R}^{m\times m}$, $\mathbf{K}_{ij} \coloneqq k(x_i, x_j)$, $\mathbf{L}_{ij} \coloneqq l(y_i, y_j)$ and $\mathbf{H}_{ij} \coloneqq \delta_{i,j} - m^{-1}$ ($\delta_{i,j} = 1$, if $i = j$; $\delta_{i,j} = 0$, otherwise)[3]. Large HSIC values indicate high correlation of the tested random variables. HSIC is zero if and only if the random variables are independent [126].

To measure the dependence between the pair of variants and the phenotype, the kernels for the phenotype and the potential epistatic variants are defined via

$$k(x_i, x_i) = \phi(x_i)\phi(x_i) = \tilde{x}_{A_i}\tilde{x}_{B_i} \tag{3.9}$$

$$l(y_i, y_i) = \varphi(y_i)\varphi(y_i) = \tilde{y}_i \tag{3.10}$$

where $\tilde{x}_{A_i}$, $\tilde{x}_{B_i}$ and $\tilde{y}_i$ represent variant A, variant B, and the phenotype, which are all Z-score normalized. Then the empirical HSIC for epistasis, called `epiHSIC` (or $\texttt{epiHSIC}_{\texttt{empirical}}$), is described as

$$\texttt{epiHSIC}((X, Y), \mathcal{F}, \mathcal{G}) \propto \sum_{i=1}^{m} \tilde{x}_{A_i}\tilde{x}_{B_i}\tilde{y}_i \tag{3.11}$$

where X and Y represent the epistatic effect and phenotypic score, respectively.

In either case, a Z-test is further performed on the correlation differences and `epiHSIC` values to obtain the significance for each pair of variants.

[3] Here, i and j are the indices in the matrix.

3.2.2.2 Matrix manipulation

The essential process in `EPIBLASTER` (Equation 3.7) is the calculation of the correlation coefficients for the two groups. From Equation 3.7, the correlation matrix for all pairs of variants represented by $\mathbf{R}$ is

$$\mathbf{R} = \frac{1}{m-1}\widetilde{\mathbf{A}}^{\top}\widetilde{\mathbf{B}} \tag{3.12}$$

where m is the total number of subjects. $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}$ are the matrices obtained from $\mathbf{A}$ and $\mathbf{B}$ by column-wise Z-score normalization, of which the rows are the individual information and the columns are the variant information. If $\mathbf{A} = \mathbf{B}$, then $\mathbf{R}$ is a symmetric matrix meaning only the values in the upper or lower triangle are necessary for further steps.

The key computation of the empirical HSIC is the trace of the underlying product of gram matrices, which is simplified in empirical `epiHSIC` (Equation 3.11). If $\mathbf{C}$ represents the `epiHSIC` matrix for all pairs of variants, then

$$\mathbf{C} = \widetilde{\mathbf{A}}^{\top}(\widetilde{\mathbf{Y}} \circ \widetilde{\mathbf{B}}) \tag{3.13}$$

where $\widetilde{\mathbf{Y}}$ is the phenotype matrix with rows for individuals and one column for one Z-score normalized phenotype. The operator ($\circ$) between $\widetilde{\mathbf{Y}}$ and $\widetilde{\mathbf{B}}$ means Hadamard product (or Schur/entrywise product), which is an element-wise (or pointwise) operation. In R, the above can be easily reached with `vector(`$\widetilde{\mathbf{Y}}$`) * matrix(`$\widetilde{\mathbf{B}}$`)` since R supports vectorized operation.

3.2.2.3 Implementation

Due to the exhaustive search, it is necessary to parallelize the calculation, along with array programming, to achieve high performance. The first idea is to split the big dataset into small subsets then run each of those as a computational unit. For example, if a big dataset is split to 22 small subsets according to the 22

autosomes, we need 253 ($\frac{22\times(22+1)}{2}$) jobs to scan all the interactions[4]. Expressly, if there are n small subsets, the resulting $\frac{n\times(n+1)}{2}$ jobs can be parallelized. **episcan** and **gpuEpiScan** are both capable of dealing with only one or with two (genotype) inputs for the different parallelized jobs (i-th subset vs j-th subset whether $i = j$). A single genotype input can be used for both $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}$ in Equations 3.12 and 3.13 leading to symmetric matrices $\mathbf{R}$ and $\mathbf{C}$ with valuable information in the upper or lower triangle.

Moreover, not all potentical users may have access to a cluster to parallelize jobs and one job can still include a large number of variants which might reach the limit of the computer's physical memory during calculation, particularly on GPUs. To address this, **episcan** and **gpuEpiScan** can split each genotype input into chunks then consider each chunk as one computational unit (Figure 3.3). The chunk size can be specified via an optional argument[5]. The chunking procedure facilitates a fast, robust and consistent calculation, especially for **gpuEpiScan** since the memory on the GPU is much more limited than the main memory accessible by the CPU.

Further, the performance of matrix multiplication on CPU can be boosted by many commonly available Basic Linear Algebra Subroutines (BLAS) libraries. There is also a fast implementation of matrix multiplication on NVIDIA GPU called **cuBLAS** [127].

3.2.3 Contents of episcan and gpuEpiScan

3.2.3.1 Main function

In this section we describe the main function, `episcan` in **episcan** and `gpuEpiScan` in **gpuEpiScan**, and provide some notes on their usage. The main functions contain the steps which show how the analysis works in the packages. Since they are quite similar, we describe them together and detail their differences when is necessary. All the arguments are listed below.

[4] Interactions include the interacting variants from the same chromosome and from two different chromosomes

[5] A chunk size equal to the total number of variants in the genotype input means no chunking.

[6] Since the purpose of the chunking process is to avoid occupying too much memory, no parallelization is implemented for the combination of chunks. Matrix-matrix multiplication is boosted by linking commonly available BLAS libraries or using **CUDA** on GPU.

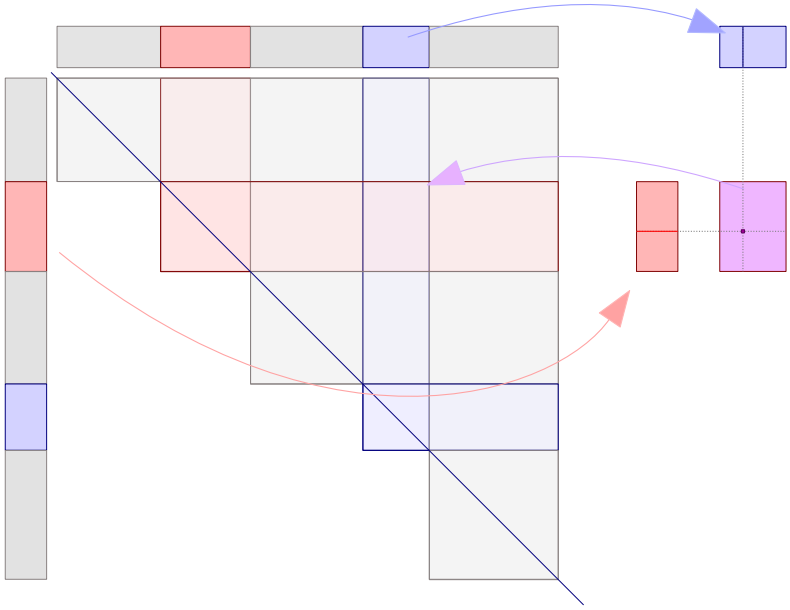

Figure 3.3: Dataset splitting and chunking. The top and left sidebars are matrix representations of the entire dataset, which can be split into several subsets. The top sidebar has the variants as columns and individuals as rows. The left sidebar is the transpose of the top sidebar, which has the variants as rows and individuals as columns. If there is no chunking process, the red and purple rectangles represent two subsets and can be taken as examples to explain how parallelization works. The combinations of subsets can be parallelized by HPC cluster or using R parallel processing packages, e.g., **parallel**. The figure on the right side intuitively explains the matrix-matrix multiplication for parallel computation with **CUDA**. Each thread within one block per grid is responsible for the calculation of (one pair of) the elements, responding to one value in the result matrix indicated by the purple dot. If chunking is requested indicated by the red and blue lines, the red and blue rectangles represent chunks[6]. If the combination containing two same subsets/chunks, only the upper/lower triangle without the diagonal of the result matrix would be recorded.

`geno1:` first genotype input with columns for variants and rows for samples.

`geno2:` optional second genotype input with columns for variants and rows for samples.

`pheno:` phenotype input, either case-control phenotype (0 for control and 1 for case) or quantitative phenotype.

`phetype:` character string to specify the type of phenotype, either "case-control" or "quantitative".

`outfile:` file name for output.

`suffix:` suffix for output file.

`zpthres:` the threshold of significance to select variant pairs for output.

`chunksize:` the number of variants in each chunk. If `chunksize` is larger than the number of given variants (n), it will be reset to be equal to n, which means no chunking procedure.

`scale:` scale the data or not.

`gpuidx:` (only for `gpuEpiScan` in **gpuEpiScan**) index of GPU (device).

The required parameters to be given are `geno1` (as well as `geno2` if needed), `pheno`, and `phetype`. The others have default settings. Figure 3.4 shows the workflow of `episcan` and `gpuEpiScan`, where `phetype` and the number of genotype input determine the type of secondary function to use. As described in Section 3.2.2.2, the matrices need to be normalized before executing multiplication. If any of the necessary inputs are missing or `phetype` is not specified (default is a character string with length equal to 2), the function will exit with an error. The parameter `scale` indicates whether the input data needs to be scaled (Z-score normalized). To limit the amount of the results recorded on the disk, `zpthres` defines the significance threshold for output, namely only the interactions with p-values $\leq$ `zpthres` are saved in the output file.

3.2.3.2 Examples

Here we demonstrate the use of **episcan** and **gpuEpiScan** by applying them to dummy data (pseudo random numbers). If not already available on your system, **episcan** can be installed from CRAN using

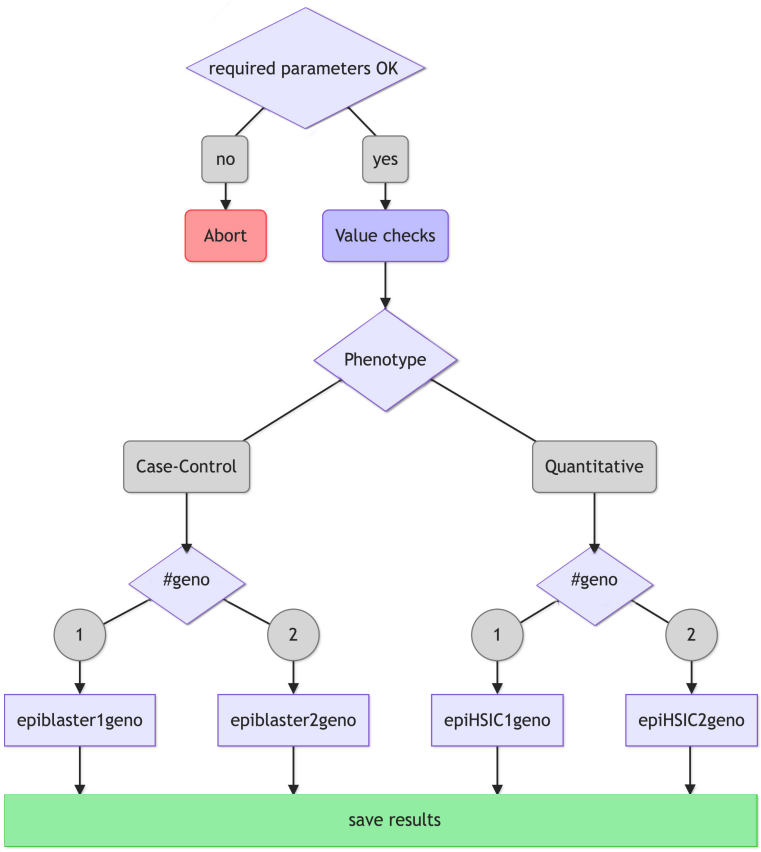

Figure 3.4: Workflow of `episcan` and `gpuEpiScan`. `Quant` means quantitative and `#geno` means the number of genotype inputs. `epiblaster1geno` and `epiblaster1geno` are the secondary functions for case-control studies with one genotype input and two genotype inputs, respectively. `epiHSIC1geno` and `epiHSIC2geno` are the secondary functions for quantitative triat studies with one genotype input and two genotype inputs, separately.

```
install.packages("episcan")
```

Since **gpuEpiScan** is currently only available for Linux system with NVIDIA GPU, the easiest way to install it with the specific **CUDA** path (YOUR_CUDA_PATH, usually the path is `/usr/local/cuda` but it may differ on different **CUDA** installation) is as

```
devtools::install_git('https://github.com/beibeiJ/gpuEpiScan.git',
                      build_opts = '--configure-args=YOUR_CUDA_PATH')
```

or download the source file then install from command line, e.g.,

```
# x.y.z need to be replaced by the real version, e.g., 0.0.1
R CMD INSTALL gpuEpiScan_x.y.z.tar.gz --configure-args=YOUR_CUDA_PATH
```

Afterwards, we load the package

```
library("episcan")
```

or

```
library("gpuEpiScan")
```

To run pair-wise epistasis analysis, first we randomly generate a small genotype dataset (`geno`) with 100 subjects and 100 variants (e.g., SNPs) as well as a case-control phenotype (`p`).

```
set.seed(123)
geno <- matrix(sample(x = 0:2,
                      size = 10000,
                      replace = TRUE,
                      prob = c(0.5, 0.3, 0.2)),
               ncol = 100)
dimnames(geno) <- list(row = paste0("IND", 1:nrow(geno)),
                       col = paste0("rs", 1:ncol(geno)))
p <- sample(rep(x = 0:1, times = c(60,40)))
```

We can take a look at `geno` and `p`

```
geno[1:5, 1:10]
```

```
##       col
## row    rs1 rs2 rs3 rs4 rs5 rs6 rs7 rs8 rs9 rs10
##   IND1   0   1   0   1   2   0   0   2   0    2
##   IND2   1   0   2   0   0   0   1   0   0    1
##   IND3   0   0   1   1   2   0   0   0   0    2
##   IND4   2   2   1   1   1   0   0   2   0    1
##   IND5   2   0   0   1   0   0   0   0   0    1
```

```
p[1:10]
```

```
##  [1] 0 0 1 1 0 0 1 1 1 0
```

Then simply use **episcan** as

```
episcan(geno1 = geno,
        pheno = p,
        phetype = "case-control",
        outfile = "episcan_1geno_cc",
        suffix = ".txt",
        zpthres = 0.5,
        chunksize = 20,
        scale = TRUE)
```

```
## p-value threshold of Z test for output: 0.5
## set chunksize: 20
## [1] "episcan starts:"
## [1] "Thu Jan 31 16:47:55 2019"
## [1] "1 chunk loop: Thu Jan 31 16:47:55 2019"
## [1] "2 chunk loop: Thu Jan 31 16:47:56 2019"
## [1] "3 chunk loop: Thu Jan 31 16:47:56 2019"
## [1] "4 chunk loop: Thu Jan 31 16:47:56 2019"
## [1] "5 chunk loop: Thu Jan 31 16:47:56 2019"
## [1] "epiblaster calculation is over!"
## [1] "Thu Jan 31 16:47:56 2019"
```

The **chunksize** can be any number between 1 and the total number of variants in genotype input. The above command saves the interactions with p-values ≤ 0.9. In a real genome-wide case, it can be more strict, e.g., 5×10^{-8}. We can load the output file to check the result:

```
result <- read.table("episcan_1geno_cc.txt",
                     header = TRUE,
                     stringsAsFactors = FALSE)
print(paste("Total rows:", nrow(result)))
```

```
## [1] "Total rows: 2508"
```

```
head(result)
```

```
##   SNP1 SNP2 Zscore    ZP
## 1  rs1  rs3   0.82 0.412
## 2  rs2  rs3  -0.89 0.373
```

```
## 3  rs1  rs4   0.88 0.379
## 4  rs2  rs4   0.78 0.437
## 5  rs3  rs4  -0.75 0.454
## 6  rs1  rs5  -1.74 0.082
```

The number of combinations for 100 variants is 4950. We have 2508 interactions passing the threshold for output. A `zpthres` value equal to 1 means to record the test result of all pairs. For `gpuEpiScan`, we generate a larger dataset, with 5000 subjects and 100000 variants splitting in two genotype inputs, to show the efficiency when utilizing the GPU. Below is the simulation to get data similar with a real-world quantitative triat study:

```
geno1 <- matrix(sample(0:2, size = 60000 * 5000,
                       replace = TRUE,
                       prob = c(0.5, 0.3, 0.2)),
                ncol = 60000)
geno2 <- matrix(sample(0:2, size = 40000 * 5000,
                       replace = TRUE,
                       prob = c(0.4, 0.3, 0.3)),
                ncol = 40000)
dimnames(geno1) <- list(row = paste0("IND", 1:nrow(geno1)),
                        col = paste0("rs", 1:ncol(geno1)))
dimnames(geno2) <- list(row = paste0("IND", 1:nrow(geno2)),
                        col = paste0("exm", 1:ncol(geno2)))
p <- rnorm(5000)
```

We can then conduct an epistasis search via

```
gpuEpiScan(geno1 = geno1,
           geno2 = geno2,
           pheno = p,
           phetype = "quantitative",
           outfile = "gpuEpiScan_2geno_quant",
           suffix = ".txt",
           zpthres = 1e-5,
           chunksize = 10000,
           scale = FALSE)

## p-value threshold of Z test for output: 1e-05
## set chunksize: 10000
## GPU index: 0
## [1] "episcan starts:"
```

```
## [1] "Thu Jan 31 16:53:09 2019"
## [1] "1 chunk loop: Thu Jan 31 16:53:09 2019"
## [1] "2 chunk loop: Thu Jan 31 16:53:56 2019"
## [1] "3 chunk loop: Thu Jan 31 16:54:48 2019"
## [1] "4 chunk loop: Thu Jan 31 16:55:33 2019"
## [1] "5 chunk loop: Thu Jan 31 16:56:20 2019"
## [1] "6 chunk loop: Thu Jan 31 16:57:05 2019"
## [1] "GPUepiHSIC calculation is over!"
## [1] "Thu Jan 31 16:57:50 2019"
```

The above calculation was conducted on one Tesla P100 GPU and finished one chunk loop (i-th chunk in `geno1` vs all the chunks in `geno2`) within one minute (plus the result saving time). With GPU, the genome-wide epistasis search becomes less time-consuming.

3.2.4 Discussion

The limited amount of heritability explained by the variants identified in GWASs is partly a result of the concentration on single locus association. Epistasis, also be referred to as SNP–SNP interactions in GWIS, have been considered in the contribution of missing heritability. However, search for epistatic interactions is a difficult task, practically, in terms of implementing algorithms for detecting epistasis and adjusting the search space appropriately [73]. High-performance and cloud computing have been widely utilized and facilitated the exhaustive epistasis search. Based on the development of the modern computer architecture, we derived the methods from Kam-Thong's work [124, 125], `EPIBLASTER` and `epiHSIC` (see Section 3.2.2), and developed two packages, **episcan** and **gpuEpiScan**, to provide efficient ways for epistasis detection not only in case-control studies but also in quantitative trait studies.

To show the efficiency particularly with GPU, we compared the performance of the core function, e.g., Pearson correlation, implemented in **gpuEpiScan** against the other implementations (Figure 3.5), from which we have seen the benefit making advantage of GPU. Although **gpuEpiScan** can only be installed on Linux system with NVIDIA GPU so far, the high performance achieved by GPU motivates us to further explore the potential that enables the package with broad adaptability.

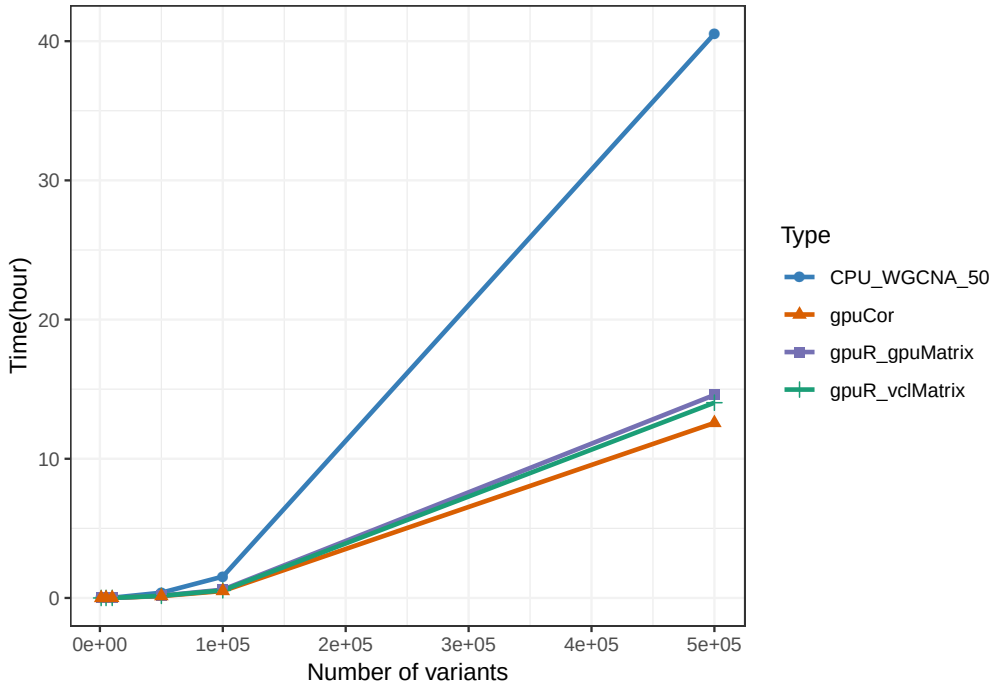

Figure 3.5: Runtime comparison of correlation calculation on GPU and multiple CPUs. Computing Pearson correlation with **WGCNA**::`cor` (CPU_WGCNA_50, where 50 indicates the number of used threads), implementation based on **CUDA** (gpuCor), implementation based on **OpenCL** using **gpuR** [128, 129] object `gpuMatrix` (gpuR_gpuMatrix), and **gpuR** object `vclMatrix` (gpuR_vclMatrix). The comparison was operated on Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz and Tesla P100 (GPU) with 10000 individuals and different number of variants (1000, 5000, 10000, 50000, 100000, 500000).

Though the current releases are stable, we plan enhancements of **episcan** and **gpuEpiScan** in the following areas: First, efforts will make not only on enabling the packages, mainly **gpuEpiScan**, available across different platforms (operating systems) but also on different GPU devices (e.g., NVIDIA GPU, AMD GPU). Second, the methods we implemented are non-parametric methodologies for epistasis detection which are suggested to be used as the scanning strategy then further conduct parametric based analysis (e.g., regression) or biological functional analysis (e.g., enrichment test) with the filtered interactions. Third, due to the biologist and clinician may have limited computer skills, it might be fruitful to have an accessible implementation, such as, web service, a GUI developed with **shiny**.

4 A Real Dataset Application: Dyslexia

4.1 Introduction

Developmental Dyslexia (DD) is a hereditary disorder which affects the visual processing of words and shows a prevalence of 5–12 % among school-aged children, implying life-long learning difficulties for most of the affected individuals [23]. The key symptom of a dyslexic individual is a discrepancy between the impaired acquisition of visual reading skills while showing normal oral and non-verbal abilities [130]. The disabilities usually appear not independently and the proportion of inherited factors involved in these cognitive skills ranges from 40 % to 80 % as reported in family and twin studies [131, 132, 133, 134].

Previously, linkage analysis and candidate gene association studies have been applied to investigate the genetic and neurobiological basis of the underlying cognitive skills, which have identified several robust candidate susceptibility genes, such as *DYX1C1* (*15q21*), *DCDC2* and *KIAA0319* (*6p22.3*), *GCFC2* and *MRPL19* (*2p12*), and *ROBO1* (*3p12.3*) (reviewed in [23, 24, 25]). Moreover, several GWAS have been performed to understand genetic basis of DD but few SNPs have shown genome-wide significant association (p-value $\leq 5 \times 10^{-8}$) [135, 136, 137, 138, 139, 140, 22] and the associations were not replicated in recent association analysis of dyslexia candidate genes on an independent dataset [141]. Despite all of the promising findings above, the candidate susceptibility genes identified and replicated so far, only explain a small portion of the genetic variance of underlying dyslexia and the related cognitive traits. It appears that the genetic architecture of dyslexia is very complicated. Also, the syndrome might be explainable by considering the interplay of genetic factors on the different cognitive traits.

Abundant evidence demonstrates that epistasis influences the genetic evolution of populations and the heritability of complex traits (reviewed in [37, 45, 142, 143, 122, 3]). Detecting genomic interactions still represents a great challenge

that could be met with a better understanding of epistasis from a mechanistic point of view [40]. Hence, it is reasonable to take an epistatic look at the genetic basis of dyslexia.

To diagnose dyslexia, researchers usually need to collect several resulting measures which are statistically or functionally correlated [22]. Such phenotypes, also called endophenotypes or sub-phenotypes, motivate the development of multi-phenotype studies [13, 20, 144, 16, 18, 19]. Since univariate analysis (single phenotype) misses the underlying covariance across two or more correlated traits, it may result in the low sensitivity for detecting shared genetic factors (pleiotropy) [140]. However, analyzing multiple phenotypes together can increase the power to detect novel associations [13], measure the heritable covariance between traits [14], and identify the potential to make causal inference between traits [78]. Additionally, Webber has reviewed the epistasis studies in neuropsychiatric disorders and discussed the problem present in detecting epistatic effects on disease status [21]. He showed the trend that, instead of studying epistasis on a dichotomous trait (present or absent), understanding the epistatic contribution on the behavioral traits or endophenotypes and their underlying neurological circuits and systems becomes significant and fruitful. Thus, jointly analyzing the reading abilities is necessary for epistasis study on dyslexia.

However, conducting a genome-wide epistasis search on thousands of dyslexic samples is extremely time-consuming, especially in several independent studies with multiple phenotypes. To reduce the computational burden of GWIS, as well as the heterogeneity caused by multiple studies, we have derived a simple analysis strategy to capture the significant interacting SNP pairs contributing to the complex disease. Shortly, the analysis follows that first, a multi-phenotype `epiHSIC` was performed on the merged data per phenotype to filter out the relatively unrelated SNP pairs after the strict quality control processing for each cohort, as well as for the merged dataset. Second, a multivariate regression was conducted on each cohort data fitting linear model per selected SNP pairs to assemble the statistics of different studies. Last, a multivariate meta-analysis was applied to the cohort statistics of joint phenotypes to measure the summary statistics. Following this analysis strategy, we searched for pairwise epistasis exhaustively in a dataset containing thousands of dyslexic children and aimed to find genetic interactions significantly associated with reading ability in multiple phenotype studies.

4.2 Subjects and Preprocessing

4.2.1 Datasets

The original datasets have been described in the context of a large quantitative GWAS on reading-related cognitive abilities [22]. Briefly, the datasets include seven cohorts containing unrelated DD cases and controls from seven different European countries and two family-based datasets. The seven European cohorts are from Austria (n=374), Germany (n=1061), Switzerland (n=67), Finland (n=336), France (n=165), Hungary (n=243), and the Netherlands (n=311). The two family-based datasets are from Colorado, United States (US) with children showing a school history of reading difficulties and their siblings (n=563) and the United Kingdom (UK) consisting of subjects with a formal diagnosis of dyslexia and their siblings (n=983) [145, 146, 139].

4.2.2 Phenotypic measures

The original cohorts consist of ten different phenotypic measures, which are word reading (WR), non-word reading (NWR), word spelling (WS), phoneme deletion (PD), digit span (DS), letters rapid automatized naming (LRAN), digits rapid automatized naming (DRAN), pictures rapid automatized naming (PRAN), alphanumeric rapid automatized naming (ARAN), and word and non-word reading (WNWR). The details on statistical elaboration were reported in previous studies [22, 139, 147, 148]. In summary, raw values from psychometric tests were grade-normed (age-adjusted in Colorado) then Z-score normalized to reduce skewness, except for the DS score which was only Z-score normalized within cohort since it was already standardized and normally distributed [147]. Additionally, no phenotypic outliers were detected in any of the analyzed datasets [22].

4.2.3 Data preprocessing

4.2.3.1 Genotype data

Initially, individuals were genotyped using different chips and autosomal genotype data have been subjected to different quality control (QC) procedures utilizing PLINK v1.90 to the degree that previous studies did [136, 22]. Briefly,

within each dataset, SNPs were filtered out if they showed a variant call rate $< 98\%$, a MAF $< 5\%$, or a HWE test p-value $< 10^{-6}$. Furthermore, individuals were removed if they showed a genotyping rate $< 98\%$. Besides, genetic ancestry outliers based on the MDS analysis on the IBS matrix and homogeneous samples with a proportion identity-by-distance (IBD) ≥ 0.0625 (corresponding to the inbreeding coefficient (IC)) were excluded, which means there are no related samples not only in the unrelated cohorts but also in the family-based cohorts (Colorado and UK).

4.2.3.2 Imputation

According to Gialluisi [22], autosomal variants were aligned to the 1000 Genomes phase I v3 reference panel (ALL populations, June 2014 release) [29] and pre-phased using SHAPEIT v2 (r837) [70]. Imputation was performed using IMPUTE2 v2.3.226 in 5 mega base pairs (Mb) chunks with 500 Kb buffers, filtering out variants that were monomorphic in the 1000 Genomes EUR (European) samples. Chunks with < 51 genotyped variants or concordance rates $< 92\%$ were fused with neighboring chunks and re-imputed.

Beyond the above, imputed variants (genotype probabilities) within each cohort were filtered with the criteria of IMPUTE2 INFO metric < 0.8, MAF $< 5\%$, the proportion of missing genotype data (null genotype call probabilities) across all samples for the SNP $< 2\%$ and HWE test p-values $< 10^{-6}$ using QCTOOL v1.5.

4.2.3.3 Cohort and phenotype selection

The German cohort includes sub-datasets from different places, collected in Marburg and Würzburg twice and in Munich once (namely MarWu1, MarWu2, and Munich), which have different experimental designs, measuring strategies and scaling methods to collect and deal with the data. Figure 4.1 shows the difference of word reading values among the datasets (a similar plot for non-word reading can be found in Appendix B.1). To exclude the potential effect of the diversity among various sub-cohorts, especially the difference in disease status (MarWu1 and MarWu2 contain no controls), only the Munich cohort was considered in further analysis.

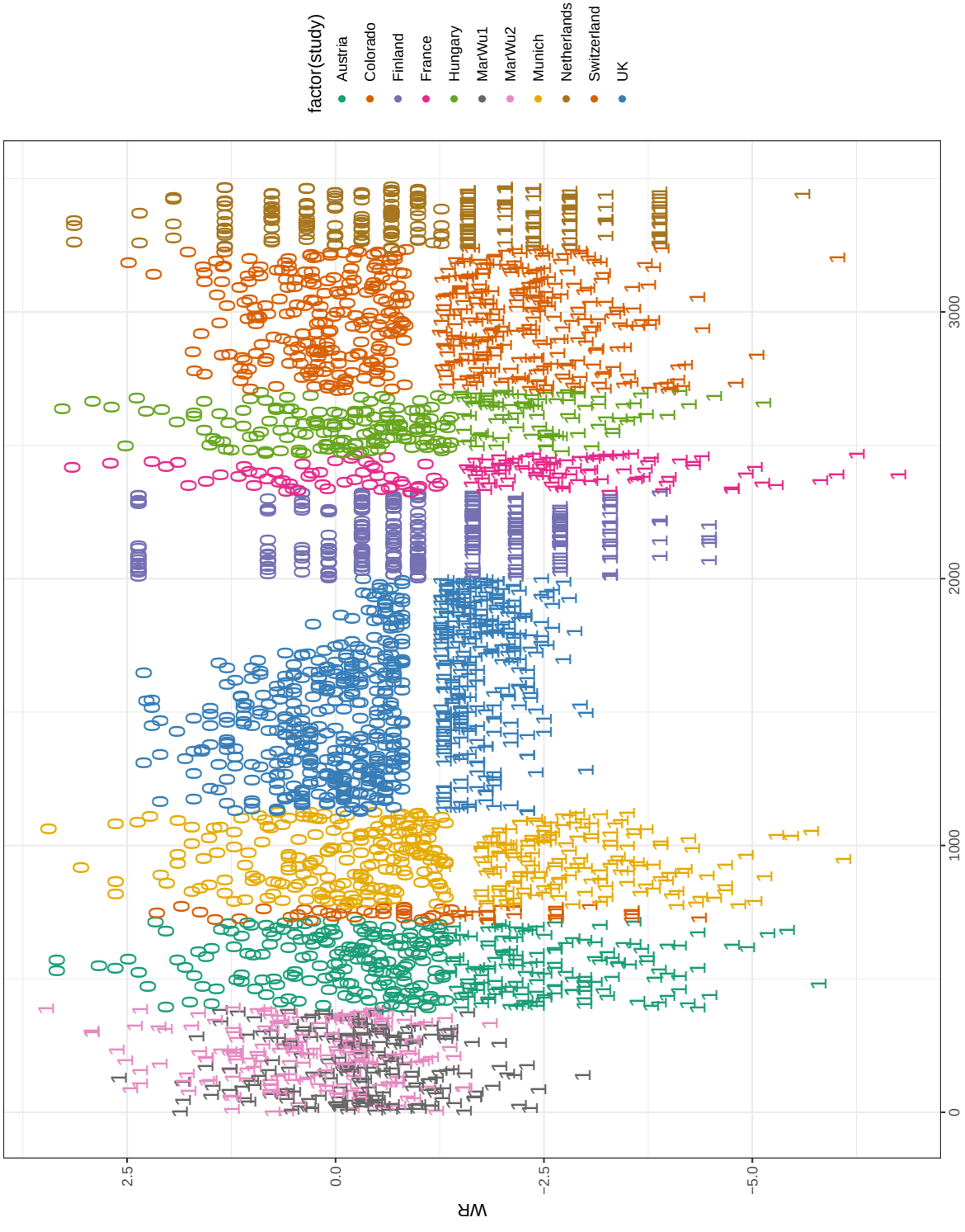

Figure 4.1: Word reading in all datasets. Different colors indicate the different datasets (studies) and the symbols (0 and 1) identify the sample as control or case, respectively.

Table 4.1: Sample size of phenotypes in nine datasets

Study	nPhe	nWR(without NA)	nNWR(without NA)
Austria	328	328	328
Colorado	533	533	529
Finland	324	324	300
France	143	143	120
Hungary	236	236	236
Munich	352	352	351
Netherlands	232	232	230
Switzerland	56	56	56
UK	875	873	868
Total	3079	3077	3018

Note:
[1] nPhe: number of phenotype pairs;
[2] nWR(without NA): number of word reading without missing value(s);
[3] nNWR(without NA): number of non-word reading without missing value(s).

To select the phenotypes suitable to perform multiple (correlated) phenotypes studies, we did (pair-based) correlation tests of the ten phenotypic measures using all selected cohorts (Austria, Colorado, Finland, France, Hungary, Munich, Netherlands, Switzerland, and UK). The skills showed moderate to high cross-trait correlations (see Figure 4.2). Among them, WNWR shows an extremely high correlation with both WR and NWR (Pearson correlation coefficients: $r = 0.97$ and $r = 0.96$, respectively). However, we were not interested in this highly correlated phenomenon because WNWR is a derived measure based on WR and NWR. Besides, since one of the biggest dataset, the UK cohort, only contains phenotypic information for WR, WS, NWR, and PD, we mainly concentrated on the selection of these four skills. The third highest correlation within our area of interest, appears between WR and NWR ($r = 0.86$). Therefore, we focused on two core phenotypes of dyslexia: word reading and non-word reading.

Further, considering the limitation of the maximum missing rate for both phenotypes, only the individuals that have at least one phenotypic information, either WR or NWR, were selected. The number of individuals with WR and NWR for each cohort was shown in Table 4.1 (Appendix B.1 shows the information of the all datasets). The Q-Q plots of WR (Figure 4.3) and NWR (Figure 4.4) per cohort shows the distribution of normality.

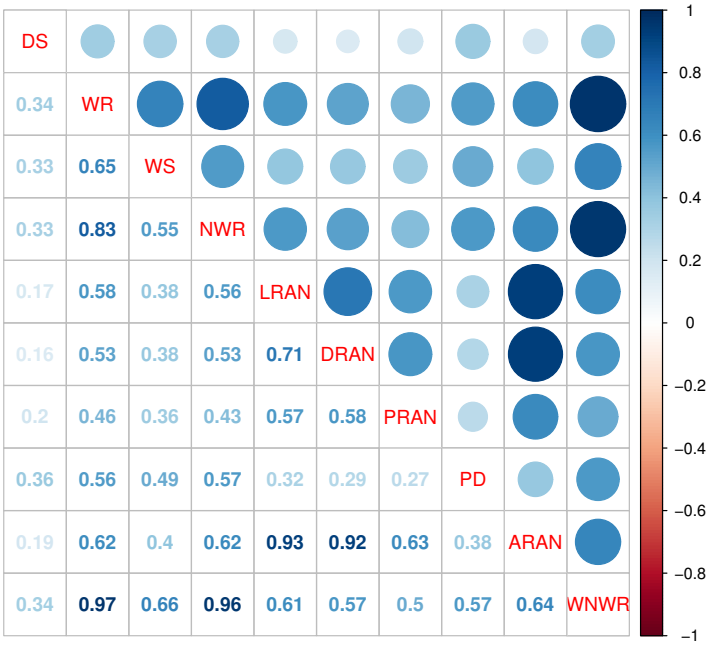

Figure 4.2: Correlations of ten dyslexia phenotypes. [1]DS: digit span; [2]WR: word reading; [3]WS: word spelling; [4]NWR: non-word reading; [5]PD: phoneme deletion; [6]LRAN: letters rapid automatized naming; [7]DRAN: digits rapid automatized naming; [8]PRAN: pictures rapid automatized naming; [9]ARAN: alphanumeric rapid automatized naming; [10]WNWR: word and non-word reading. The sample size for the different pairs of phenotypes differs since the coefficient was computed using all complete pairs of observations on the oriented two phenotypes. Particularly, for the Colorado dataset, there is no information of ARAN and WNWR while for the UK dataset, there is no information of DS, LRAN, DRAN, PRAN, ARAN, and WNWR.

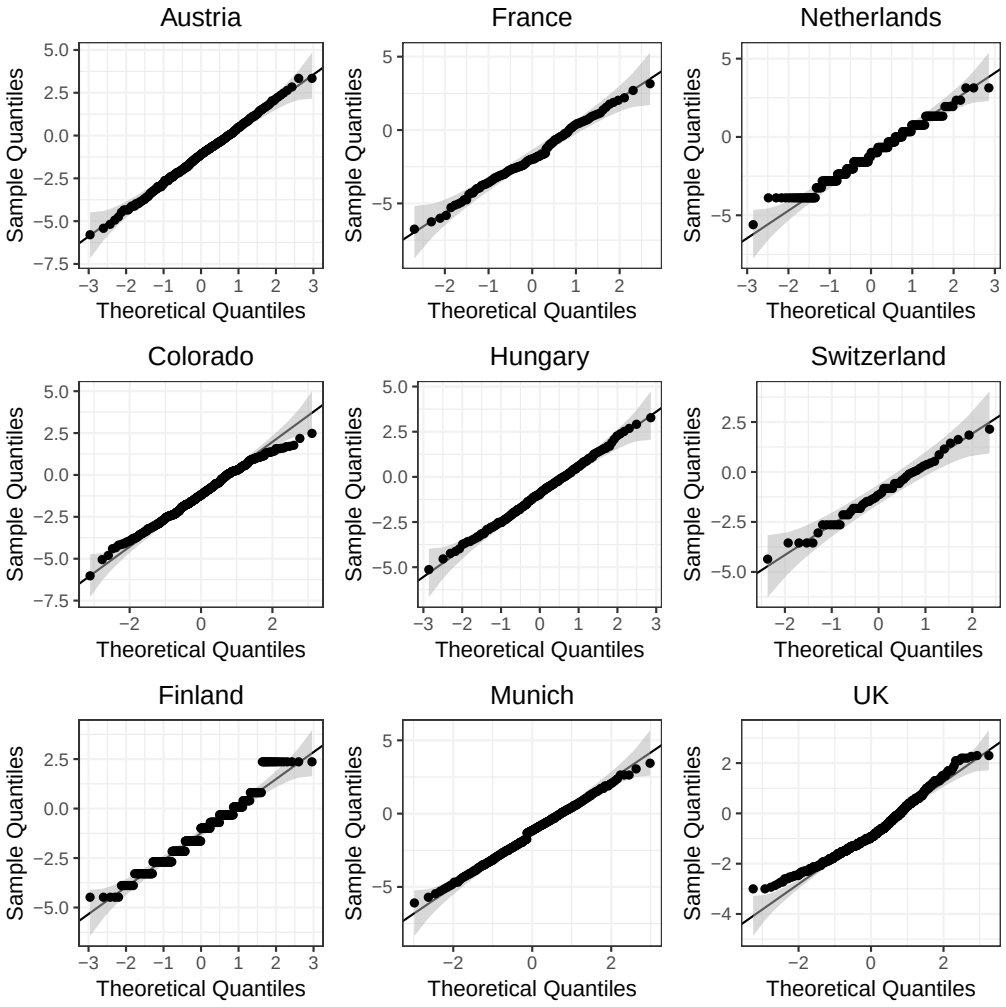

Figure 4.3: Q-Q plots of word reading per cohort. For each plot, the x-axis represents the theoretical normally distributed values and the y-axis represents the observed values.

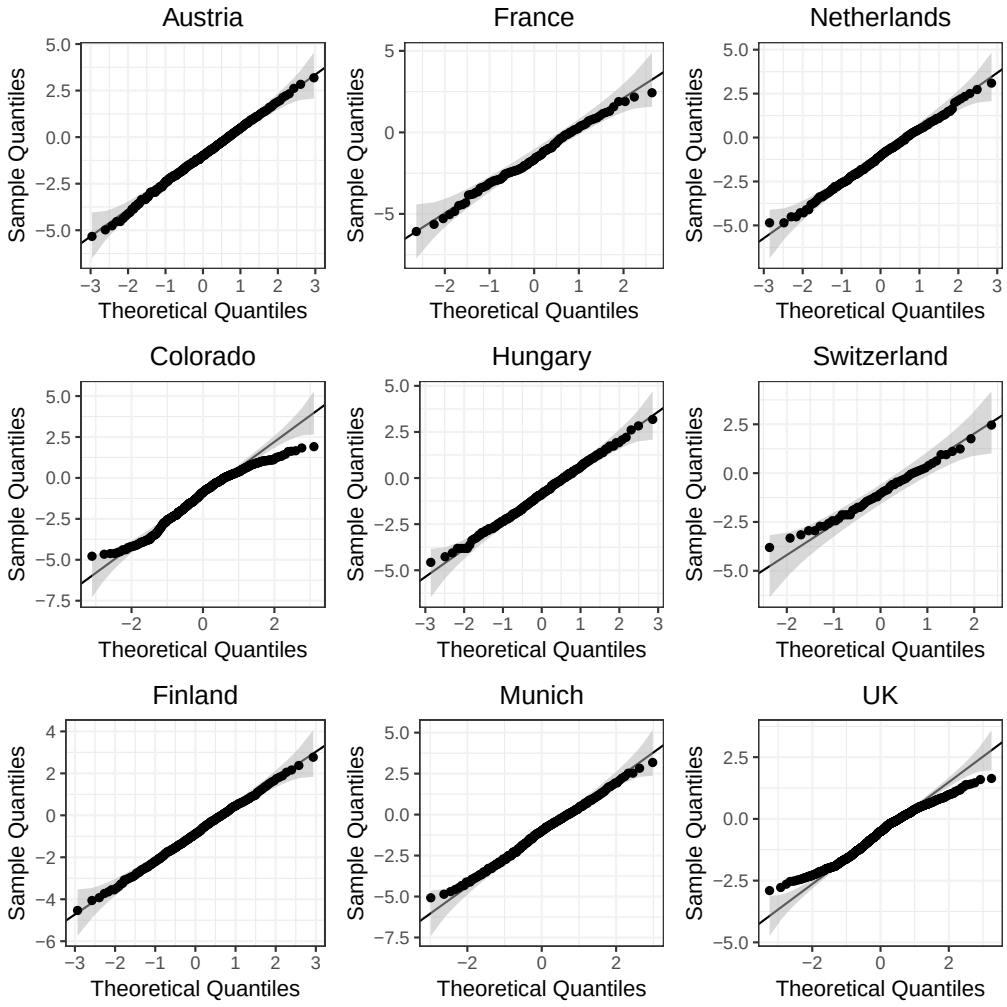

Figure 4.4: Q-Q plots of non-word reading per cohort. For each plot, the x-axis represents the theoretical normally distributed values and the y-axis represents the observed values.

4.2.4 Split samples

We split the nine datasets (Austria, Colorado, Finland, France, Hungary, Munich, Netherlands, Switzerland, and UK) into discovery and replication datasets based on the sample size and distribution of word reading in each cohort (see Figure 4.1 and Figure 4.3). Austria, Colorado, Hungary, Munich, and UK have a larger number of phenotype pairs than France and Switzerland. Also, the difference of numbers (without NA) between WR an NWR in these five cohorts are relatively smaller than for Finland and France. Moreover, unlike the distribution of Finland and Netherlands datasets, there is no jumping point (discrete-like values) in the other datasets. Thus, Austria, Colorado, Hungary, Munich, and UK were analyzed as the discovery datasets and Finland, France, Netherland, and Switzerland datasets were considered as the replication datasets.

4.2.5 Data preprocessing within discovery and replication

4.2.5.1 Discovery datasets

Discovery datasets were extracted from the whole datasets (genotype data, imputed data, and phenotype data) to process the following steps:

a) conducted QC for genotype data per cohort as described in Section 4.2.3.1

b) combined five cohorts (after QC) then operated QC and applied MDS analysis to get the first ten components as covariates for further merged dataset analysis

c) extracted the samples from the merged dataset (after QC) per cohort then calculated the MDS components per cohort (also checked whether MDS plots change or not) and saved them as covariates for further analysis per cohort

d) compared the samples containing WR and NWR information with the samples in genotype data then kept the common samples (more individual information in genotype data than in phenotyping data)

e) merged the imputed data with PLINK and conducted QC

f) pruned out the variants based on LD score ($r^2 > 0.7^{\frac{1}{2}}$)

g) extracted the remaining variants (275561) and samples from merged imputed dataset and performed QC (same like the QC in Section 4.2.3.2)

Table 4.2: Sample size of phenotypes in discovery set

Study	nPhe	nWR(without NA)	nNWR(without NA)
Austria	325	325	325
Colorado	181	181	179
Hungary	234	234	234
Munich	352	352	351
UK	510	509	508
Total	1602	1601	1597

Note:
[1] nPhe: number of phenotype pairs;
[2] nWR(without NA): number of word reading without missing value(s);
[3] nNWR(without NA): number of non-word reading without missing value(s).

h) extracted the remaining variants (275559) and samples (see Figure 4.5) per cohort, performed QC (same like the QC in Section 4.2.3.2)

i) selected the common variants through the five cohorts

j) extracted the common variants (271437) for merged dataset and each cohort

In the end, 271437 SNPs were used for analysis. The information of sample size per cohort in discovery datasets is shown in Table 4.2.

4.2.5.2 Replication datasets

An analogous workflow, as described in Section 4.2.5.1 without the steps related to getting the final variants list, was applied on replication datasets. The reference variants list was the same as the final SNP list (271437) in the discovery dataset[1]. The information of sample size per cohort in replication datasets is shown in Table 4.3.

4.3 Analysis strategy

To reduce runtime and memory costs, we first applied **gpuEpiScan** to quickly scan for epistasis on the whole genome level with merged discovery data (five cohorts) then saved the results of the variant pairs with an arbitrarily determined p-value threshold: 10^{-4}. Later, we tested all the saved variant pairs with linear regression

[1] The number of variants remaining in some replication cohorts may be less than 271437.

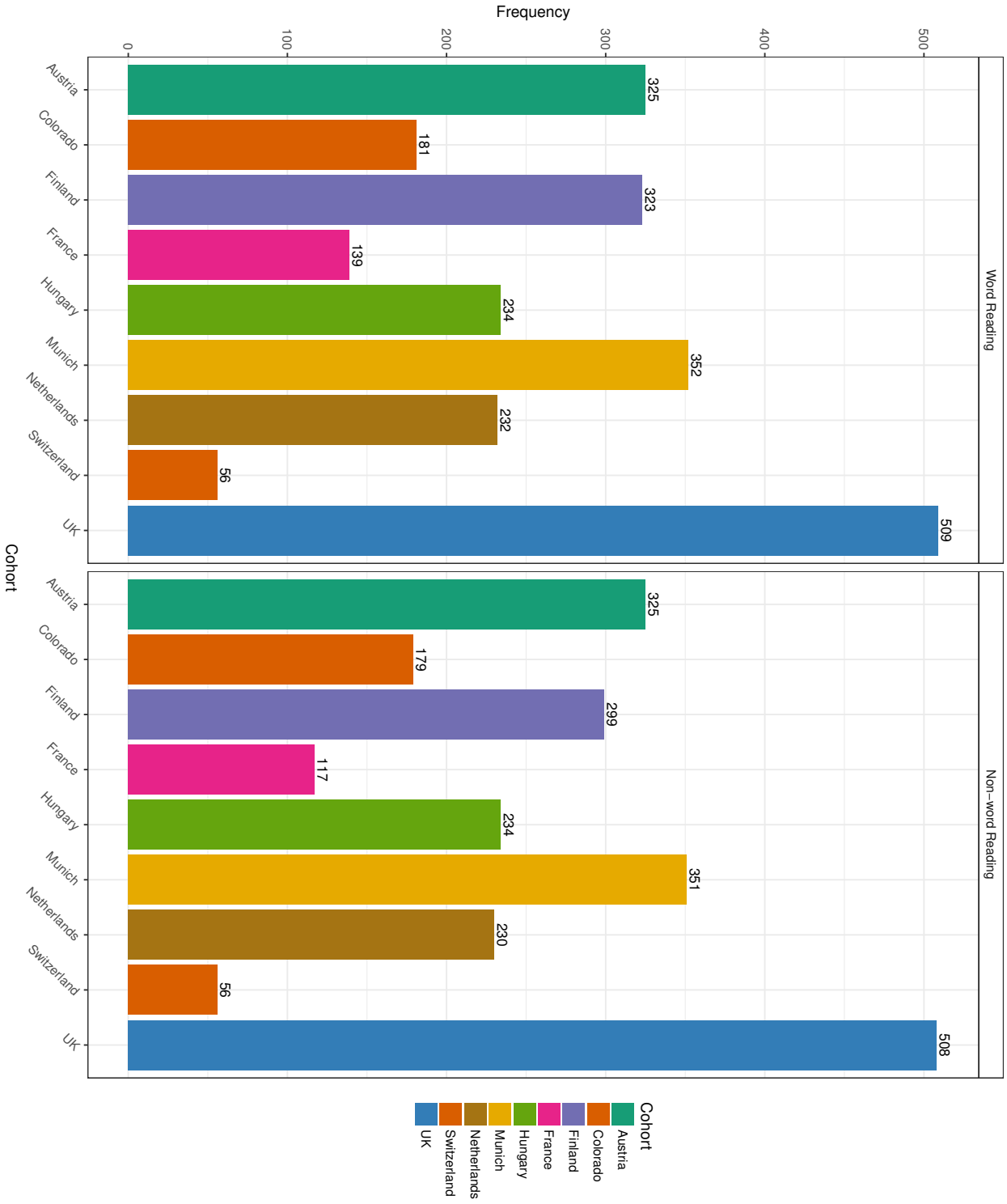

Figure 4.5: Number of phenotypes in final cohorts.

Table 4.3: Sample size of phenotypes in replication set

Study	nPhe	nWR(without NA)	nNWR(without NA)
Finland	323	323	299
France	143	143	120
Netherland	232	232	230
Switzerland	56	56	56
Total	754	754	705

Note:
[1] nPhe: number of phenotype pairs;
[2] nWR(without NA): number of word reading without missing value(s);
[3] nNWR(without NA): number of non-word reading without missing value(s).

per cohort then performed the multivariate meta-analysis. Also, linear regression on the merged data was operated to compare how **gpuEpiScan** results vary with traditional analysis method (linear regression).

4.3.1 Merged data

Scanning for epistasis was conducted using our R package **gpuEpiScan** on the merged data. The input variables (the dosage coding variants) were corrected by the covariates (the first ten MDS components). To get the outcomes, the merged phenotypes (WR and NWR) were scaled per phenotype and corrected by the covariates (the first ten MDS components). Afterwards, a PCA was applied to reduce the dimension and get the main component of the two corrected phenotypes. Later, the corrected variants and the PCA component were used as input for **gpuEpiScan**[2].

4.3.2 Cohort data

Multivariate regression within each cohort used the dosage coding variants as variables, the MDS components as covariates, and the scaled phenotypes as outcomes. The genetic association with WR or NWR is written in matrix notation as

[2] Within **gpuEpiScan**, the first step is to scale the inputs which mean both the input variables and the outcome will be scaled.

$$\begin{pmatrix} y_{1_1} & y_{2_1} \\ y_{1_2} & y_{2_2} \\ \vdots & \vdots \\ y_{1_n} & y_{2_n} \end{pmatrix} = \begin{pmatrix} 1 & x_{A_1} & x_{B_1} & x_{A_1B_1} & c_{1_1} & c_{2_1} & \cdots & c_{10_1} \\ 1 & x_{A_2} & x_{B_2} & x_{A_2B_2} & c_{1_2} & c_{2_2} & \cdots & c_{10_2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{A_n} & x_{B_n} & x_{A_nB_n} & c_{1_n} & c_{2_n} & \cdots & c_{10_n} \end{pmatrix} \cdot \begin{pmatrix} \beta_{1_0} & \beta_{2_0} \\ \beta_{1_1} & \beta_{2_1} \\ \beta_{1_2} & \beta_{2_2} \\ \vdots & \vdots \\ \beta_{1_{13}} & \beta_{2_{13}} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1_1} & \varepsilon_{2_1} \\ \varepsilon_{1_2} & \varepsilon_{2_2} \\ \vdots & \vdots \\ \varepsilon_{2_n} & \varepsilon_{2_n} \end{pmatrix}$$

where $1, 2, ..., n$ indicates the individuals and x_A and x_B are the variant A and variant B (dosage coding). y_1, y_2, ε_1, ε_2, β_1, and β_2 represent standardized scores, random errors, and regression coefficients of regression items (from 0 to 13) for WR and NWR, respectively. The ten MDS components as covariates are denoted by $c_1, c_2, \cdots, c_{10}$.

4.3.3 Meta-analysis

To avoid any influences on the statistical scores resulting from the different populations (studies) and to capture possible heterogeneity, multivariate meta-analysis was performed to summarize the association between the variants of interest and the two correlated phenotypes across multiple studies using the R package **mvmeta** [93]. Also, the inverse variance based meta-analysis with fixed effect[3] for each phenotype was subsequently applied via the R package **rmeta** [149].

4.4 Results

4.4.1 Evidence for the effectiveness of the analysis strategy

Before applying the analysis strategy (described in section 4.3) to the entire datasets, we investigated whether the main principle component from a PCA or the product of original phenotypes (Z-score normalized) can be an effective representative in `epiHSIC` for the multiple phenotypes in epistasis scanning step. The experiment was conducted on artificial phenotypes, simulated by **SimPhe**, which were contributed by the pre-defined quantitative trait loci (QTLs) with marginal

[3] Heterogeneity tests were performed and there was no significant heterogeneity.

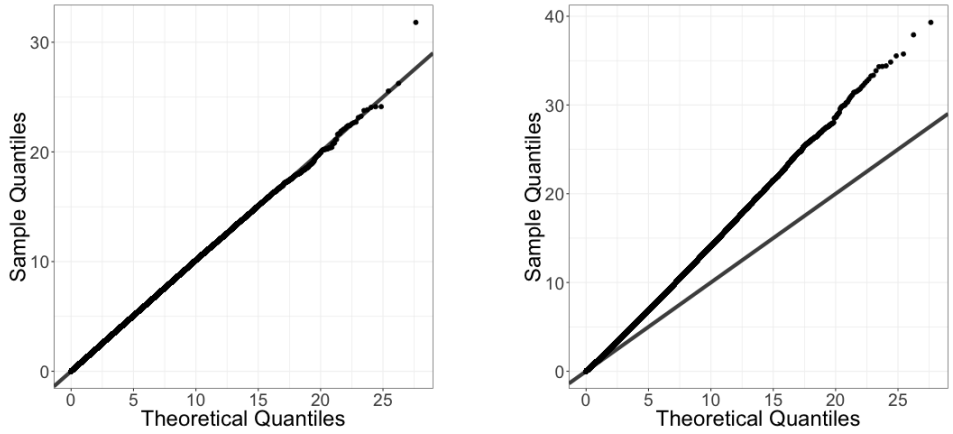

Figure 4.6: Q-Q plots of permutation tests. The left Q-Q plot is for the permutation test of taking the main principle component from PCA into `epiHSIC` and the right Q-Q plot is for the permutation test of taking the product of normalized phenotypes into `epiHSIC`. The x-axis represents the theoretical p-values and the y-axis represents the observed p-values. All p-values were shown as $-\log_{10}(p\text{-values})$.

effects and epistatic effects. The real interacting QTLs have been detected with high significance in either the main principle component from the PCA or the product of normalized phenotypes as phenotypic representative. Then a permutation test (one million times) was conducted to get the distribution of the `epiHSIC` statistics for these two phenotypic representatives. Figure 4.6 gives the evidence that it is better to use the main principle component from PCA rather than the product of normalized phenotypes to represent outcome in `epiHSIC` since the later causes high inflation[4].

Besides, we randomly chose one thousand variants from each discovery cohorts as exploration datasets to test how well our analysis strategy performed when potential heterogeneity of experimental designs exists. We tested all the combinations of one thousand variants and saved the results, not just the results of the Z-test in `epiHSIC` on the merged exploration data but also the results of the linear regression per cohort, as well as the meta-analysis results. Figure 4.7 and Figure 4.8 show the relationship among the p-values of the Z-test in `epiHSIC` for WR and NWR on the merged (exploration) data, of the inverse variance based meta-analysis with fixed effects, and of the multivariate meta-analysis with fixed

[4] Since reading abilities show high linear correlation, we consider to use PCA to reduce the phenotypic dimensions. If potentially non-linear correlations exist, we suppose principal components from kernel PCA can be fitted in `epiHSIC`.

effects (Appendix B.2 with random effects). To help delineate any logarithmic trend, the p-values are shown as the negative logarithmic values. In the low significance part of the plots (lower left), we observe the correlation between the two analytic scores. Both of the figures demonstrate the validity in the approximation of p-values of Z-test in `epiHSIC` for multiple phenotypes to the resulting meta-analysis p-values on the interaction term in a linear regression. Therefore, multi-phenotype `epiHSIC` was considered as the scanning method in the analysis strategy, which was further applied to entire dyslexia dataset to detect the epistasis associated with word reading and non-word reading.

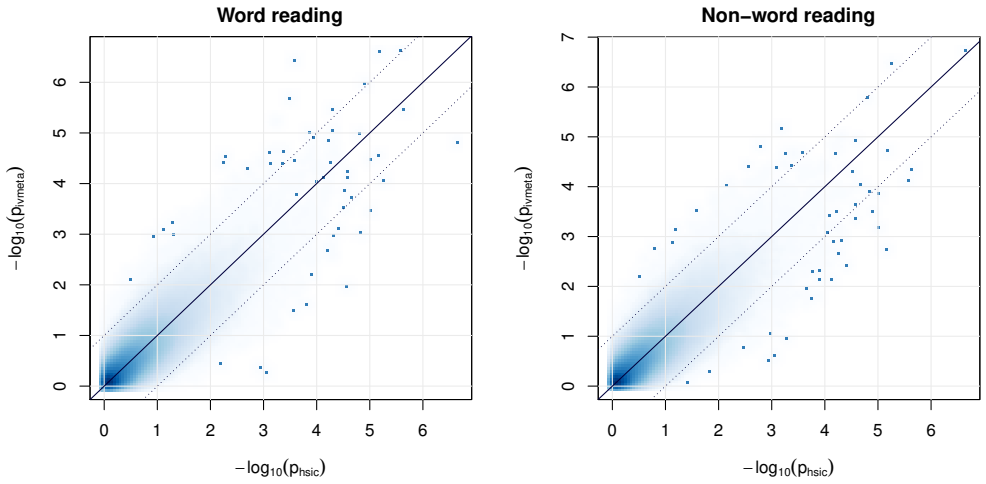

Figure 4.7: Comparison of p-values between `epiHSIC` for multiple phenotypes and inverse variance based meta-analysis. p_{hsic} represents the p-values of Z-test in `epiHSIC` for multiple phenotypes on merged (exploration) data and p_{ivmeta} means the p-values of inverse variance based meta-analysis with fixed effects. The superimposed 50 points indicate the outer border of the distribution. All p-values are shown as $-\log_{10}(p\text{-values})$.

4.4.2 Epistasis on dyslexia

The epistasis scanning process was applied to 271437 variants with $\binom{271437}{2}$ ($\approx$ 36 billion) combinations in merged discovery datasets. The genetic interactions (variant pairs) with Z-test p-values $< 10^{-4}$ were chosen and passed to conduct linear regression on merged discovery datasets and multivariate regression on cohort data.

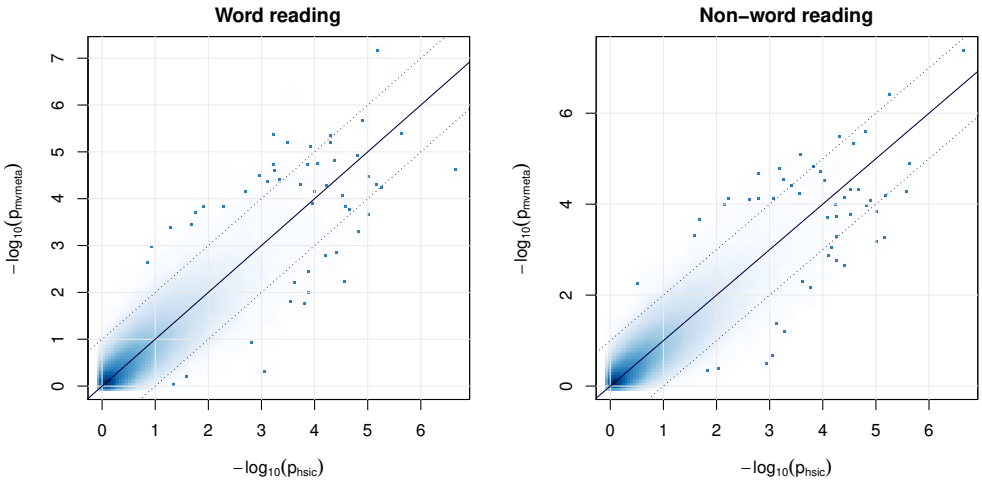

Figure 4.8: Comparison of p-values between `epiHSIC` for multiple phenotypes and multivariate meta-analysis. p_{hsic} represents the p-values of Z-test in `epiHSIC` for multiple phenotypes on merged (exploration) data and p_{mvmeta} means the p-values of multivariate meta-analysis with fixed effects. The superimposed 50 points indicate the outer border of the distribution. All p-values are shown as $-\log_{10}(p\text{-values})$.

There are around 3.8 million SNP pairs passed the filtering process with $p < 10^{-4}$ in multi-phenotype `epiHSIC` study. We then fitted these pairs (one by one) in linear regression for word reading and non-word reading on the merged discovery datasets, separately. Also, the multivariate regression of these pairs was applied within each cohort to get the statistics of multiple studies. Later, different meta-analysis methods, multivariate meta-analysis and inverse variance based meta-analysis, were performed on the cohort statistical scores. However, none of the SNP pairs stood out with strong association (Bonferroni correction threshold: 1.36×10^{-12}) for WR or NWR in discovery dataset (see Appendix B.2).

To check whether increasing the sample size can change the result, we performed multivariate regression of the ($\approx$ 3.8 million) SNP pairs on the replication datasets per cohort for WR and NWR, independently and applied the same meta-analysis methods as for discovery datasets to statistical scores across all nine cohorts (5 cohorts in discovery plus 4 cohorts in replication).

4.4.2.1 Significant interactions

Multivariate meta-analysis results on nine cohorts were presented in Figure 4.9 and 4.10 (only the selected interactions p-values $\leq 5\times10^{-8}$ were plotted[5]). There are 1697 epistatic SNP pairs associated with word reading and 1617 epistatic SNP pairs associated with non-word reading. The circos plots clearly show that most of the interactions appear with $10^{-10} < p \leq 5 \times 10^{-8}$ and one pair of variants passing the Bonferroni corrected threshold of genome-wide significance ($p \leq 1.36 \times 10^{-12}$) for word reading (Figure 4.9) but none of the interaction reached the threshold for non-word reading (Figure 4.10).

The most significant interaction associated with word reading was observed with the $p = 1.26 \times 10^{-12}$ between rs8013684 on chromosome 14 (position: 92365790, minor allele: A) and rs1442415 on chromosome 15 (position: 96216254, minor allele: T). Both of the SNPs have been imputed with high INFO score (Table 4.4). The diversity on marker level across the nine cohorts only shows on minor allele frequency in Finland data which is higher than in other cohorts for both of the SNPs. Nevertherless, no heterogeneity among the cohorts ($p \approx 0.76$ for WR and $p \approx 0.68$ for NWR) were given and most of the cohorts, except Switzerland due to the small sample size, demonstrated negative effect directions for word reading (Figure 4.11).

For below, the significant (SNP or variant) pair (interaction or epistasis) points to the epistasis between rs8013684 and rs1442415 for word reading.

4.4.2.2 Interaction between rs8013684 and rs1442415

To confirm the epistasis occurring between rs8013684 and rs1442415, we plotted the average score of word reading cross different combinations of genotypes (Figure 4.13). From the colored lines and their trends, we concluded that purely epistatic effects exist in the absence of main effects at either locus.

To study the performance of the other SNPs within the LD region of rs8013684 and rs1442415 (some of the SNPs have been removed during pre-processing with the $r^2 > 0.7^{\frac{1}{2}}$), 223 SNPs around rs8013684 and rs1442415 ($\pm$ 500kb) have been selected to perform the meta-analysis. Among them, 115 are on chromosome

[5] The p-value threshold was chosen by no intention. It was just used as filtering to have fewer genetic interactions to easily display.

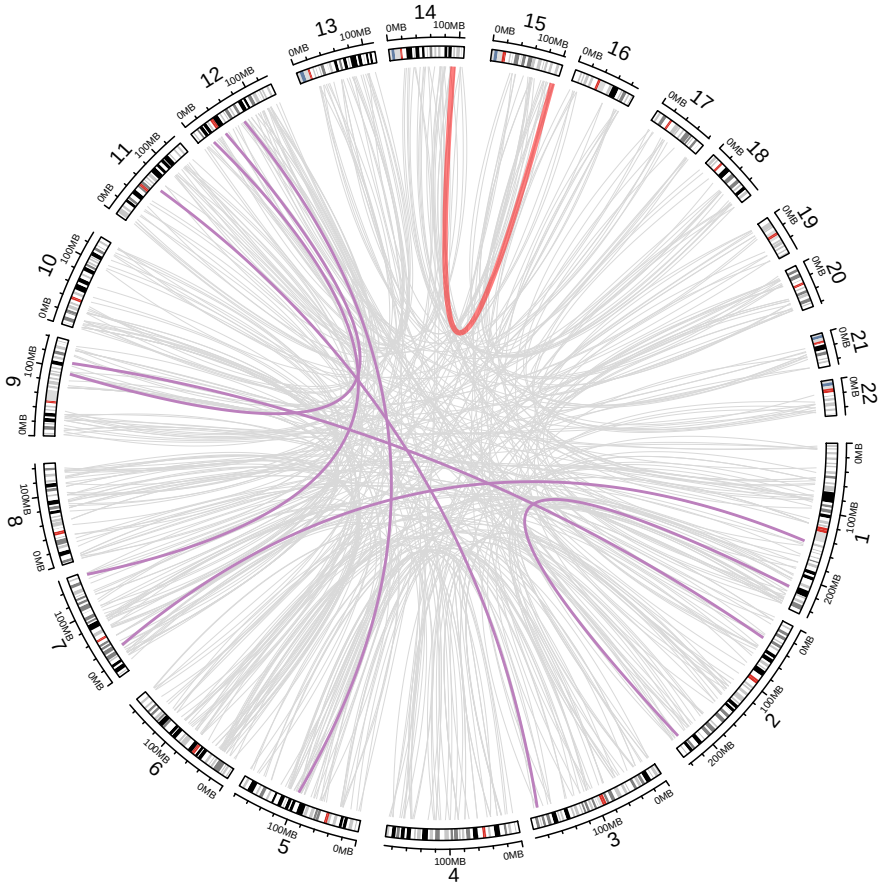

Figure 4.9: Circos diagram showing inter-chromosomal interactions for word reading. The connected lines indicate the genetic interactions and the colors indicate the significance of the interactions, grey meaning $10^{-10} < p \leq 5 \times 10^{-8}$, purple meaning $1.36 \times 10^{-12} < p \leq 10^{-10}$, and red meaning $p \leq 1.36 \times 10^{-12}$.

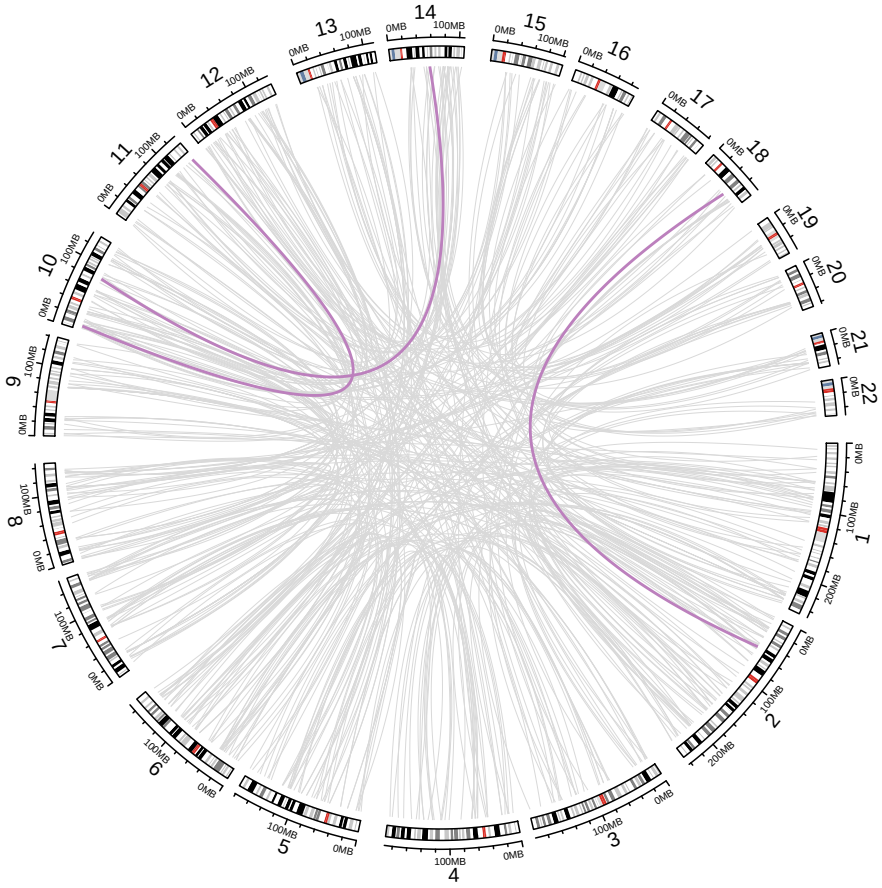

Figure 4.10: Circos diagram showing inter-chromosomal interactions for non-word reading. The connected lines indicate the genetic interactions and the colors indicate the significance of the interactions, grey meaning $10^{-10} < p \leq 5 \times 10^{-8}$ and purple meaning $1.36 \times 10^{-12} < p \leq 10^{-10}$.

Table 4.4: SNP information for the top signal

SNP	Study	MAF	HWE	SNPmiss	INDmiss	Info
rs8013684	Austria	0.36	0.55	0	0.02	0.99
	Munich	0.34	0.20	0	0.01	1.00
	Hungary	0.34	2.27	0	0.02	0.99
	UK	0.37	0.00	0	0.02	0.99
	Colorado	0.34	0.13	0	0.02	0.99
	Finland	0.46	0.00	0	0.02	0.99
	France	0.33	0.46	0	0.01	1.00
	Netherlands	0.33	0.33	0	0.02	0.99
	Switzerland	0.33	0.27	0	0.02	0.99
rs1442415	Austria	0.07	0.19	0	0.02	0.95
	Munich	0.08	0.34	0	0.03	0.94
	Hungary	0.06	0.00	0	0.03	0.95
	UK	0.07	0.00	0	0.02	0.97
	Colorado	0.05	0.00	0	0.01	0.97
	Finland	0.14	0.19	0	0.01	0.99
	France	0.08	0.22	0	0.02	0.96
	Netherlands	0.08	0.00	0	0.01	0.98
	Switzerland	0.07	0.00	0	0.02	0.95

Note:
[1] SNP: SNP name;
[2] MAF: minor allele frequency;
[3] HWE: $-\log_{10}(p\text{-value})$ for Hardy-Weinberg equilibrium;
[4] SNPmiss: the proportion of missing genotype data (null genotype call probabilities) across all samples for the SNP;
[5] INDmiss: the proportion of individuals for which the maximum genotype probability is less than a threshold of 0.9;
[6] Info: IMPUTE's info score measuring how much uncertainty there is in the genotype calls. Equal to zero when the genotype call probabilities are obtained from the allele frequency, to 1 when the calls are all certain.

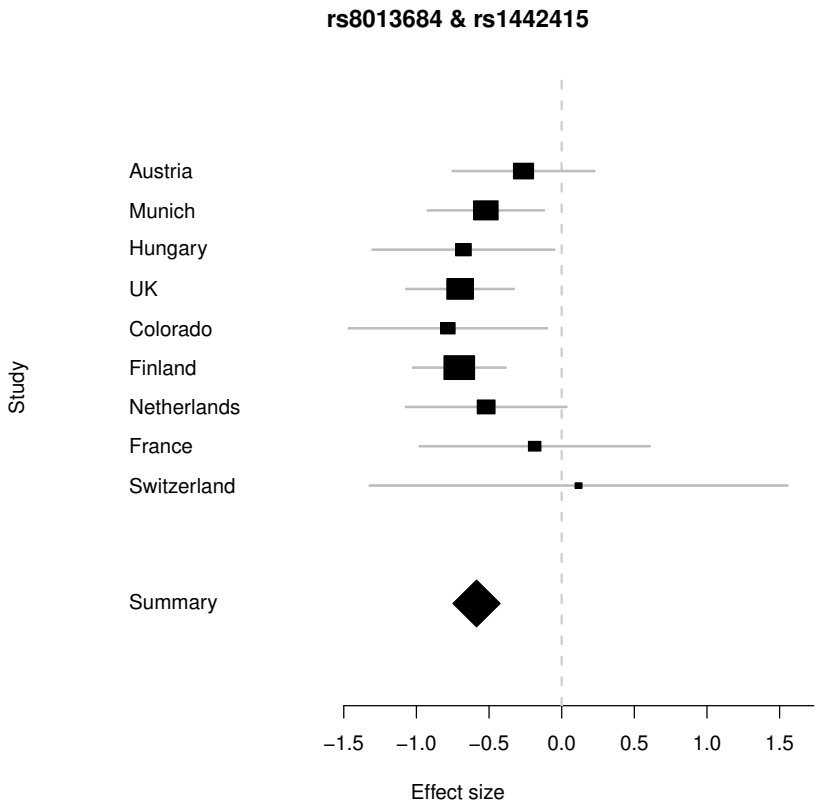

Figure 4.11: Top hit of meta-analysis for word reading. For every single cohort (Austria, Munich, Hungary, UK, Colorado, Finland, Netherlands, France, and Switzerland) the square and horizontal line show the estimated regression coefficient β and 95 % confidence interval, representing the effect of each copy of the reference-allele on reading performance. The size of the square is inversely proportional to the standard error of the estimated effect. Below the individual cohorts, a summary diamond shows the multivariate meta-analysis with fixed effects when analyzing all nine cohorts together. Notably, except in Switzerland cohort, clear negative effect presents in the most single cohort and in the combined sample.

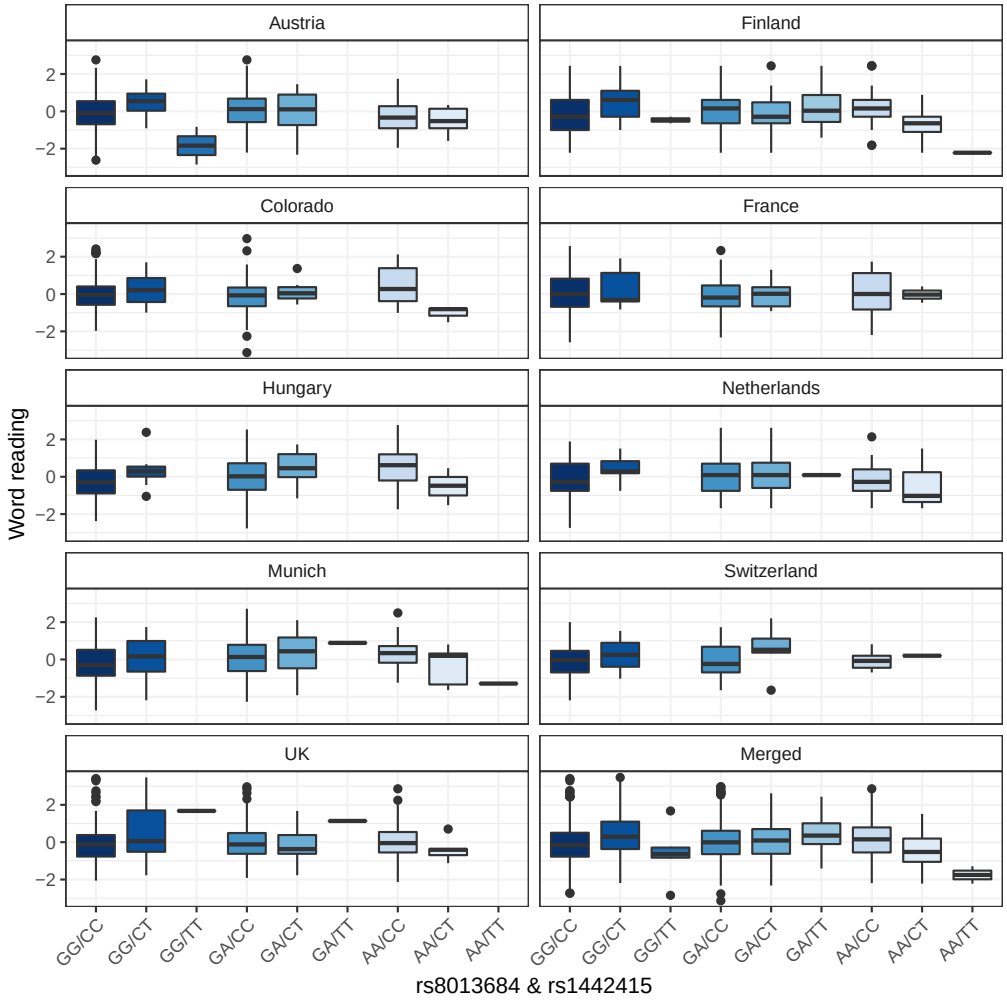

Figure 4.12: Boxplot of word reading through all genotype combinations per cohort. The y-axis defines the range of the measured reading scores normalized with the mean of 0 and the variance of 1. A higher score indicates the weaker reading performance. The x-axis is the according genotypic combination of the SNPs rs8013684 (chromosome 14, minor allele: A) and rs1442415 (chromosome 15, minor allele: T). The box defines the interquartile range, the thick line shows the median, and single dot means possible outlier. Box colors separate classes of genotype combinations.

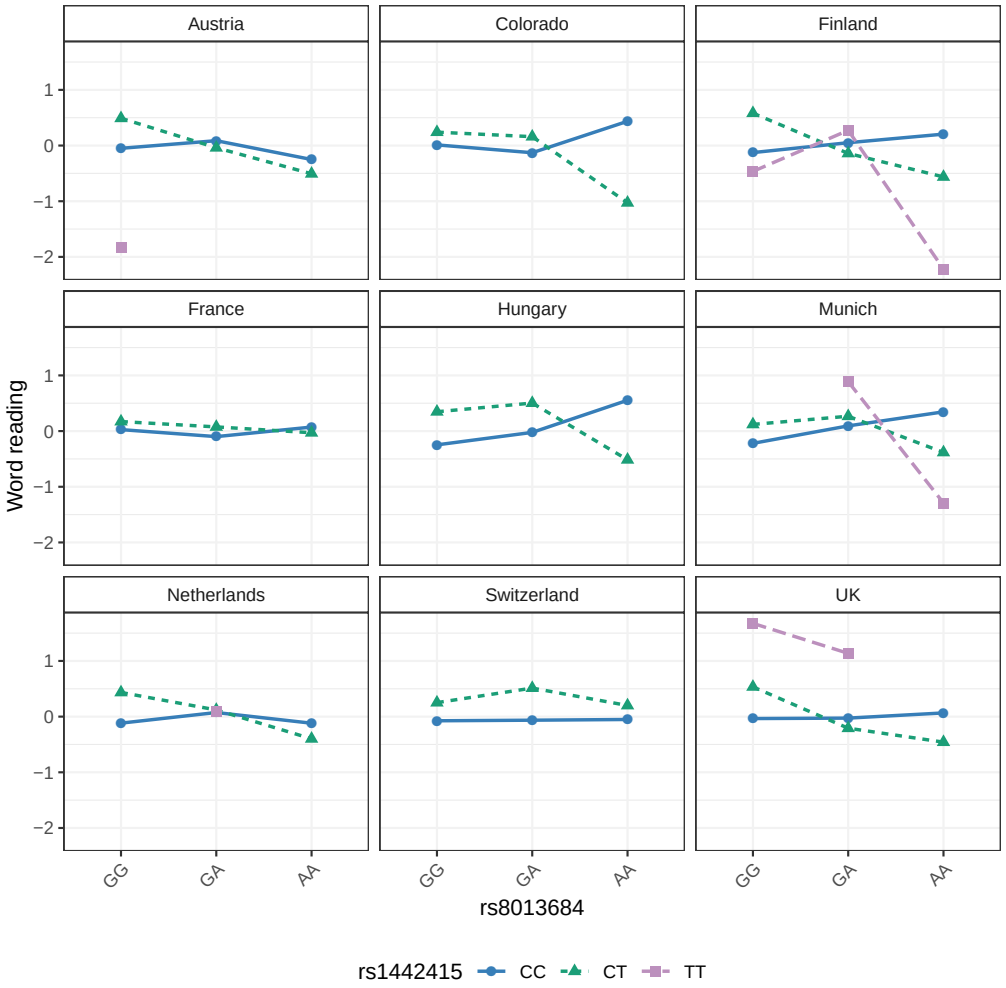

Figure 4.13: Genotypic effects at two locus per cohort. Two-locus model of the top hit illustrates purely epistatic effects in the absence of main effects at either locus. The x-axis represents the three genotypes of rs8013684 (GG (homozygote of major allele), GA (heterozygote), AA (homozygote of minor allele)) while the three lines indicate the different genotypes of rs1442415 with the differently colored lines, a blue solid line for homozygote of major allele (CC), a green dotted line for heterozygote (CT), and a purple dashed line where present for homozygote of minor allele (TT)[6]. The y-axis displays the respective mean of the measured word reading scores.

14 and 110 are on chromosome 15. Figure 4.14 shows the interactions only between rs8013684/rs1442415 and other SNPs. Except for the detected epistasis (rs8013684 vs rs1442415) in the main analysis, LD region analysis also found other SNPs interacted with rs8013684 or rs1442415 are associated with word reading. For example, rs8013684 and rs60109817 have been identified with a strong interaction (p-value $\approx 5.06 \times 10^{-8}$ below the Bonferroni correction threshold for $\binom{225}{2}$ times tests). However, since rs60109817 is within the LD region of rs1442415, it might be detected due to strong linkage.

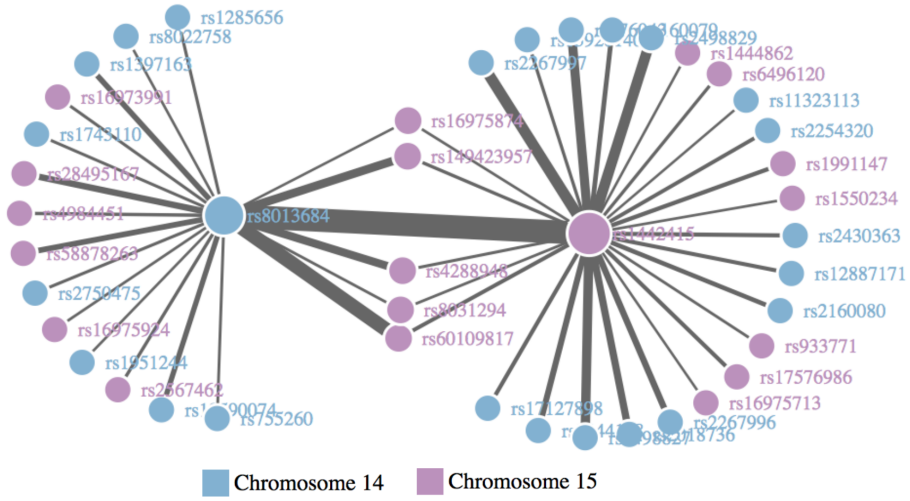

Figure 4.14: Interactions of the SNPs within the LD regions of rs8013684 and rs1442415 for word reading. The lines indicate the genetic interactions and the width of the line represents the significance of the interaction. All the visible lines have $p \leq 0.05$. The color of the nodes represents the chromosome, blue for chromosome 14 and purple for chromosome 15.

4.5 Discussion

A GWIS of around 2500 samples containing nine independent cohorts has been conducted and reading-related traits (*word reading* and *non-word reading*) were investigated jointly to better understand the genetic architecture. To our knowledge, it is the richest epistasis study in terms of the reading ability, as well as the involved countries and languages. Besides, multivariate meta-analysis was performed on the correlated phenotypes and appeared to improve the precision

of the analysis than the standard meta-analysis for single phenotype (Table B.4 and B.5) (discussed in [87]).

In the beginning, there was no SNP pair reaching the Bonferroni corrected significance (1.36×10^{-12}) for the specific phenotypes when only discovery datasets were considered. When including replication datasets into the analysis, we found significant epistasis which illustrates that sample size affects the detection of epistatic effects. Figure 4.15 presents the Pearson correlation coefficients between the predictions of the top epistatic signal (rs8013684 and rs1442415) and phenotypic measures change against the increasing sample size. The correlation coefficient rises almost 45 % for word reading and 70 % for non-word reading when the replication datasets were analyzed together with the discovery datasets. As discussed in experimental design in neuroscience [150] and epistasis [45], our study confirms the well-established fact that statistical detection power increases when more samples have been involved. The large sample size is required both to detect significant interactions and to sample the landscape of possible genetic interactions [39].

Our study did not detect many SNP pairs with the Bonferroni corrected significance for GWIS even though the sample size increased, consistent with the phenomenon of previous GWASs on dyslexia, where few SNPs reached the genome-wide significance: 5×10^{-8} in GWASs [135, 136, 137, 138, 139, 140, 22]. The interaction between rs8013684 and rs1442415 is the only one which shows high significance ($p = 1.26 \times 10^{-12}$). Though none of the two SNPs is in any gene regions, rs1442415 was discovered with $p \approx 9.64 \times 10^{-4}$ in previous GWASs including over 400000 individuals on educational attainment [151]. According to Undheim's clinical work [152], dyslexic individuals showed lower educational attainment compared to non-dyslexic people. Thus, it might be the genetic interactions that rs1442415 participated in that affects the performance of word reading and then the affected group shows a different level of education.

Except the most significant SNP pair associated with word reading, there were more than one thousand interactions with p-value $\leq 5 \times 10^{-8}$ which have been recognized to affect word reading and non-word reading, as shown in Figure 4.9 and Figure 4.10. To further understand the result, we annotated the SNPs involved in the interactions with p-value $\leq 10^{-10}$ and performed the enrichment test to explore whether any of the genes mapped by the selected SNPs are linked to the GWAS catalog [33] to provide insight into previously reported associations

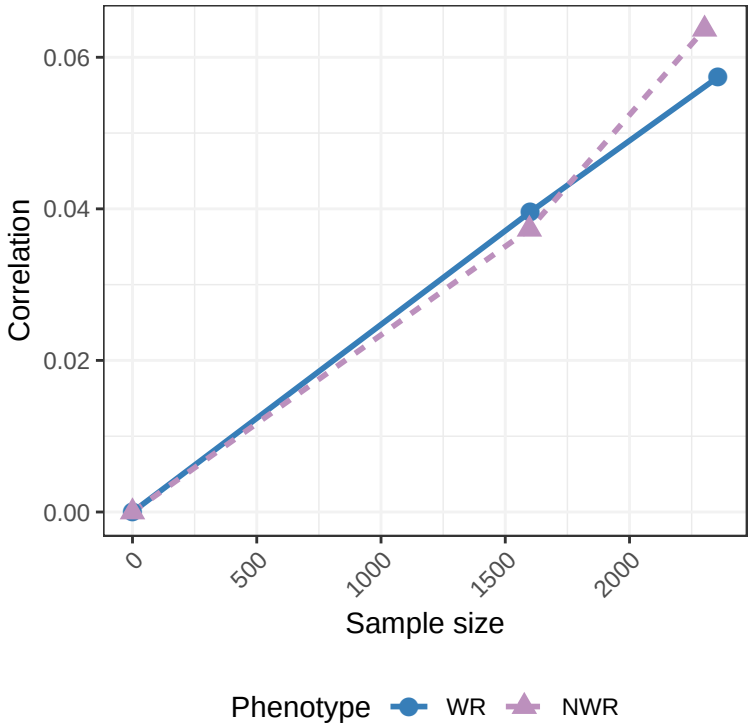

Figure 4.15: Correlations between the predictions of the top epistatic signal and phenotypic measures versus the change of sample sizes. The x-axis shows the number of individuals when different cohort(s) have been included for the analysis, from 0 to the number of total samples in discovery datasets and further to the number of total samples in all nine cohorts (discovery plus replication datasets). The y-axis indicates the Pearson correlation coefficients between the values predicted by the epistatic SNP pair (rs8013684 and rs1442415) and the observed phenotypic scores for the different amount of samples. The colored lines display the trend of the correlation coefficient change for the two phenotypes, a blue solid line for word reading and a purple dotted line for non-word reading.

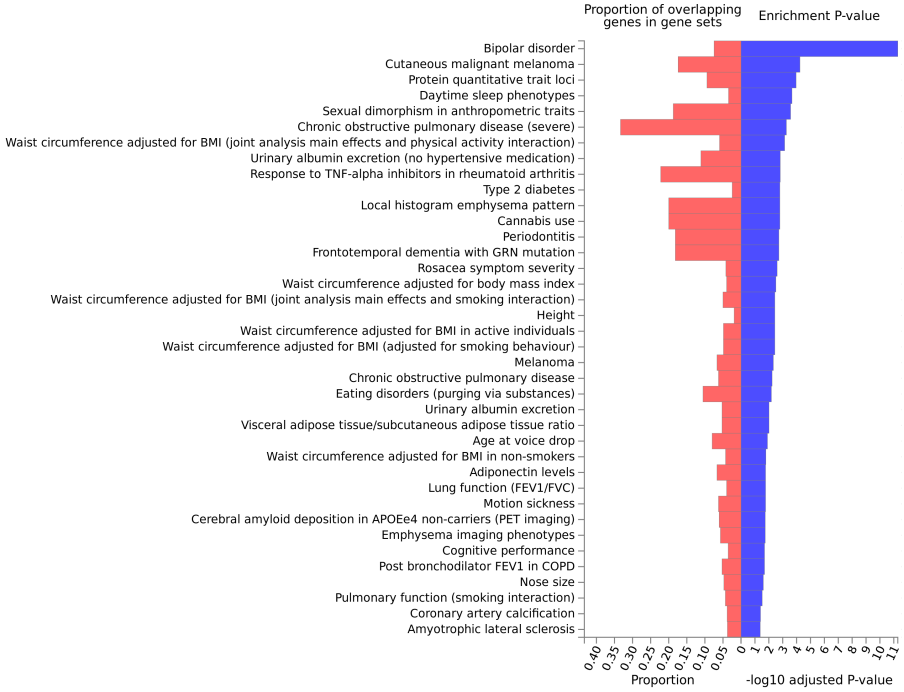

Figure 4.16: Gene enrichment compared with reported GWASs gene for word reading. Red boxes are for the overlapping proportions of each complex trait and blue boxes are for Benjamini-Hochberg (false discovery rate (FDR)) adjusted p-values (shown as $-\log_{10}(p\text{-values})$) of hypergeometric distribution tests. Note: only the complex traits of which the genes were tested against our mapped gene set with adjusted p-values < 0.05 are plotted.

of the SNPs in the risk loci with a variety of phenotypes. The gene mapping, expression quantitative trait locus (eQTL) mapping, and chromatin interaction mapping were done by FUMA [153] building on several related sources or tools[7] [154, 155, 156, 157, 158]. Surprisingly, the selected SNPs (eight pairs referring to 16 independent SNPs) identified in GWIS on word reading are mapped within or near the gene regions which have been investigated in many GWASs on human diseases (see Figure 4.16). Some of them are neurobiological traits, such as bipolar disorder, daytime sleep phenotypes, and cognitive performance. To get an idea how the enrichment depends on SNP pairs, we performed leave-one-out (leave-one-pair-out) stepwise enrichment test and found the SNP pair, rs75222724 (chromosome 3, *EHHADH–AS1*, ncRNA_intronic) and rs1700189 (chromosome 11, *DPP3*, intronic), carrying more information than the others (Figure 4.17). By reducing the number of input genes (two SNPs were mapped to 27 genes), the enrichment test shows the genes of interest mapped by rs75222724 and rs1700189 also overrepresented in the gene set associated with major depressive disorder. Our study might be a support for the idea that genetic factors contributing to risk of one disease could, on average, also be informative of the risk of correlated diseases [15, 159].

Despite cognitive abnormalities have been considered as the fundamental defects in schizophrenia, supported by [160], our enrichment test did not show the correlation between the genes mapped by the interacted SNPs associated with word reading and the reported genes associated with schizophrenia in previous GWASs. However, a significant overlap appears between our gene set and the gene set reported to be associated with bipolar disorder. It suggests that reading abilities might share the genetic factors with bipolar disorder leading the genetic evidence for the finding in recent co-occurrences study which discovered that individuals with reading problems had increased risks of several psychiatric disorders including bipolar disorder but without schizophrenia [161].

In summary, we performed multivariate meta-analysis on GWIS for reading-related phenotypes and identified one genome-wide significant epistasis associated with word reading, as well as suggestive genetic interactions which might affect reading abilities. Except the SNP (rs1442415) significantly interacting with rs8013684 has been reported to influence educational attainment, the SNPs

[7] The details of how the processes are conducted by FUMA can be found in the tutorial page (`http://fuma.ctglab.nl/tutorial`). The version we used is FUMA v1.3.4 released in February 17, 2019, with which the reference gene set of GWAS catalog is e93 2019-01-11.

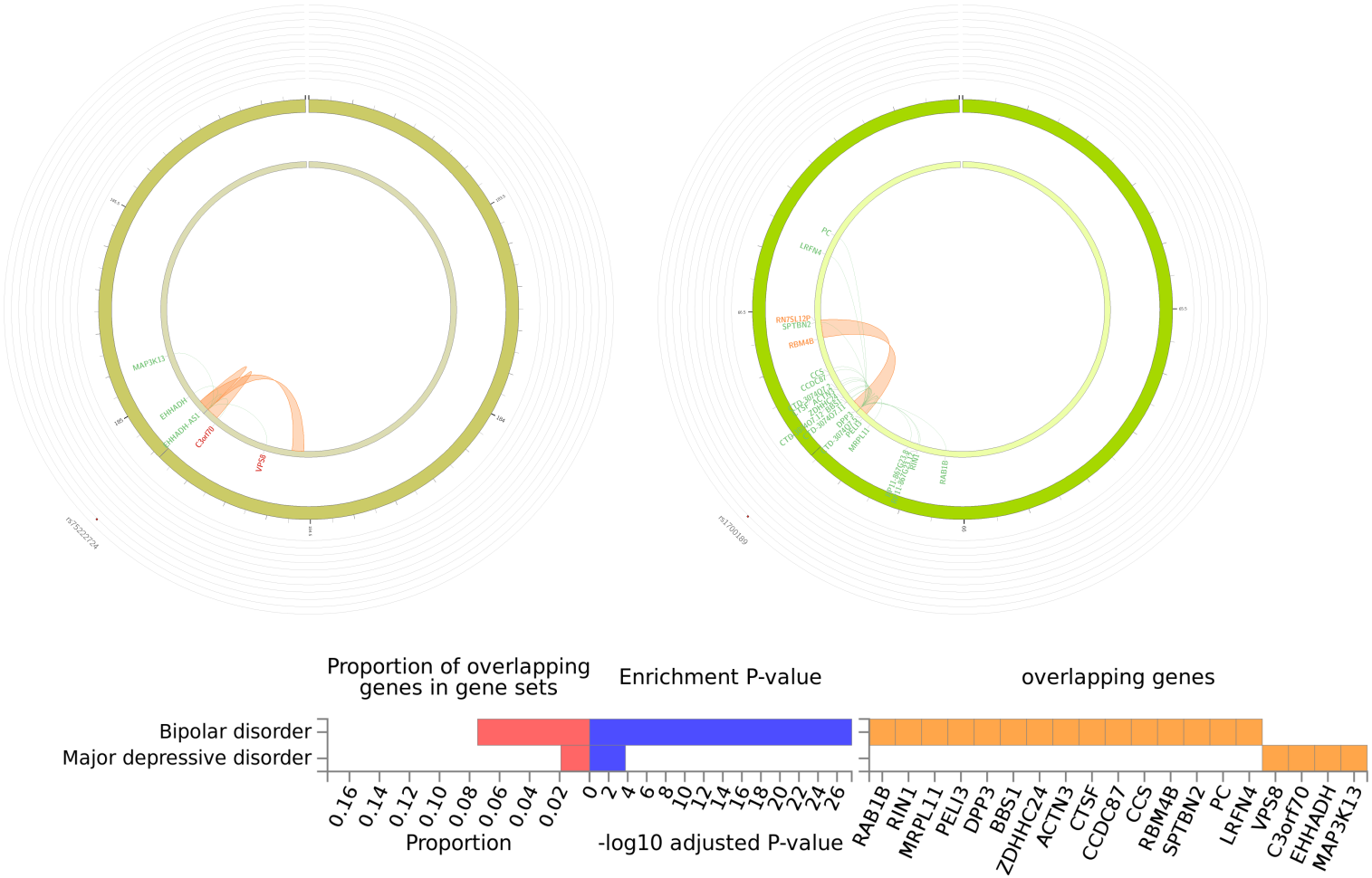

Figure 4.17: Chromatin interactions and enrichment of rs75222724 & rs1700189. The circos plots indicate the interactions between the genes mapped by the SNP and other genomic risk loci. The bottom plot shows the overpresentation information of mapped genes(orange colored). Red boxes are for the overlapping proportions of each complex trait and blue boxes are for Benjamini-Hochberg (FDR) adjusted p-values (shown as $-\log_{10}(p\text{-values})$) of hypergeometric distribution tests.

involved in the suggestive interactions have shown the associations with psychiatric disorders in previous GWASs, especially with bipolar disorder. Our results will likely contribute to efforts to investigate not just on the genetic interactions but also on the multiple correlated phenotypes (skills/disorders).

5 Conclusion & Outlook

The thesis is intended to develop efficient tools, with high accessibility to domain experts, for exhaustive epistasis search based on hypothesis-free methods. Meanwhile, we also aimed to extend our understanding of statistical epistasis and to expose complex synergetic factors underlying susceptibility to developmental dyslexia.

5.1 Tools development

5.1.1 Statement of contributions on pair-wise interactions

From the computational and statistical perspective, the three R packages we have developed can help identify the genetic interplay or gain insights into the diversity of tools for epistatic analysis. The R package **SimPhe** to simulate (multiple correlated) phenotype(s) with epistatic effects, can provide God's perspective for researchers to evaluate the performance of various epistasis methods. The two packages **episcan** and **gpuEpiScan** for detecting epistasis with high performance not only in case-control studies but also in quantitative trait studies, can help address the presence of noise measured in all variables and the existence of nonlinearities in the system describing the effects between the genotypic and the phenotypic output. The detection methods utilized by the two packages are correlation-based methods, not novel but computationally simple, to solve the statistical epistasis search problem [124, 125]. Coupled with matrix manipulation, the implementations, especially on GPU, are more powerful than previous approaches.

5.1.2 Attempt on high-order interactions and other interests

The R packages were developed with only a focus on pair-wise interactions. Future concentration can be on high-order, e.g., three-way interactions. Besides, from the application point of view, the correlation-based methods, especially `HSIC`, can be adapted to not only epistasis studies but also multi-omics studies where researchers could seek for the network (connection) among genomics, transcriptomics, metabolomics, and proteomics. The idea of three-way interactions or network has been attempted on one PTSD dataset (discovered in [162]) containing genotyping, expression, and methylation information. To reduce the computational burden, the toy application only considered the SNPs mapped to *FKBP5*. When each omic variant was considered as one variable, `HSIC` for three-way interaction was formulated as

$$\texttt{HSIC}_{three-way}((X,Y),\mathcal{F},\mathcal{G}) \propto \sum_{i=1}^{m} \tilde{x}_{G_i}\tilde{x}_{E_i}\tilde{x}_{M_i}\tilde{y}_i \tag{5.1}$$

where $\tilde{x}_{G_i}$, $\tilde{x}_{E_i}$, and $\tilde{x}_{M_i}$ represent the genotype, expression, and methylation data, which are all Z-score normalized. The total number of individuals is represented by m. The phenotype indicator, $\tilde{y}_i$, has

$$\tilde{y}_i = \begin{cases} \frac{1}{n_1-1} & \text{when} \quad y_i = 1 \\ -\frac{1}{n_0-1} & \text{when} \quad y_i = 0 \end{cases} \tag{5.2}$$

where n_1 is the total number of cases and n_0 is the total number of controls.

The p-values of three-way `HSIC`, according to the definition and initial application on epistasis [126, 125], supposed to be the approximation of p-values from the three-way interaction term in a linear model like

$$y \sim \beta_0 + \beta_1\tilde{x}_G + \beta_2\tilde{x}_E + \beta_3\tilde{x}_M + \beta_4\tilde{x}_G\tilde{x}_E + \beta_5\tilde{x}_E\tilde{x}_M + \beta_6\tilde{x}_G\tilde{x}_M + \beta_7\tilde{x}_G\tilde{x}_E\tilde{x}_M + \varepsilon \tag{5.3}$$

where $\beta_0, \beta_2, \cdots, \beta_7$ represent the regression coefficients.

Figure 5.1 shows the result of network analysis on PTSD three-omics data. Since the analysis is a toy-like attempt, we did not go further with the deep understanding of the network. However, the links among the genes mapped by three types of markers from different omics data provide the information that the biological architecture of PTSD is complicated. Seeking only a single locus with significant effect, e.g., in GWAS, might not be an efficient way to study the functional mechanism influencing PTSD. It reveals the potential interest of looking at across omics interactions on PTSD, as well as on other complex traits.

Apart from developing methods/tools for high-order interactions, we are interested in continuously contributing to multi-phenotype studies. Not only because the correlations have been observed among endophenotypes but also because the multivariate analysis on dyslexia found potentially shared genetic variants in an interacting way or in a single-locus way associated with psychiatric disorders. Although we have shown evidence that taking the main principal component of a PCA as a representative of multiple phenotypes to `epiHSIC` can be an effective approach for epistasis detection in multi-phenotype studies, it is hard to have deep insight on how precise the latent variable from linear-based dimensionality reduction (PCA) could indicate the true correlations of multiple phenotypes. Further study will be on investigating methods, linear or nonlinear, to have better coverage of the diverse correlations among multiple phenotypes (two or more).

Moreover, building a GUI with R package **shiny** is also under consideration (mentioned in Section 3.1.6 and Section 3.2.4) to make complex models more easily available to broader range researchers.

5.2 Biological perspective

Although substantial genetic interactions associated with reading abilities have been uncovered by our multi-phenotype studies on the dyslexia dataset, we could not draw well-founded conclusions yet whether the identified interacting loci have biologically reasonable functions, especially the functional mechanisms, due to the lack of attention and well-established experimental verification for genetic interactions. As discussed in Chapter 2 and Section 4.5, many studies including ours are trying to understand the role of statistical epistasis in human health and disease by

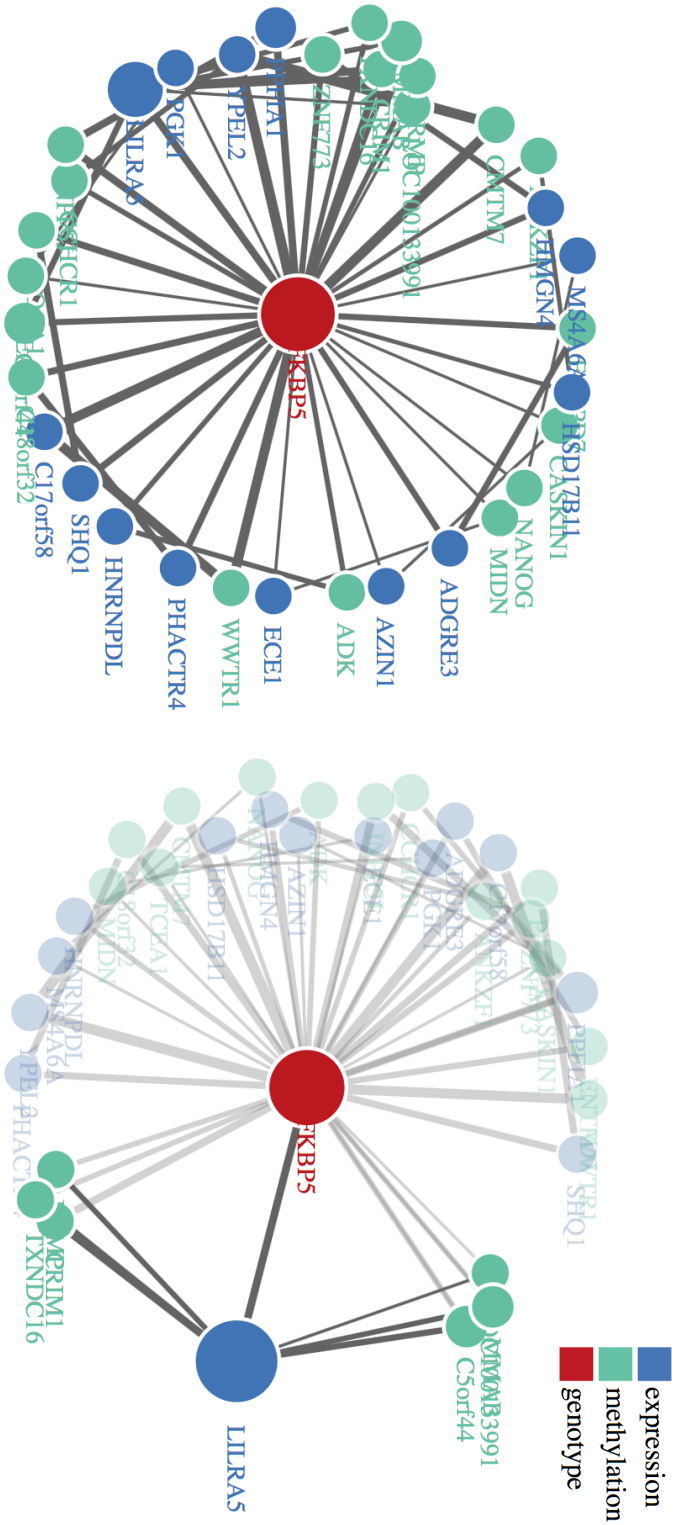

Figure 5.1: Netwrok analysis of genomics, transcriptomics, and metabolites on PTSD.

- looking at the pathway/network the variants involved in
- comparing epistatic effects to the single locus effects on complex traits
- seeking for generalizability of the interaction presence across populations

We believe that it requires time to offer state of the art capabilities in the wide area of interaction computation for any kind of omics data (Section 5.1.2). These could help to improve personalized clinical approaches, answer questions regarding functionality, and shed light upon the often discussed necessity of single-locus penetrance in an interacting complex (marginal effects).

Bibliography

[1] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009. doi:10.1038/nature08494.

[2] B. Maher. Personal genomes: The case of the missing heritability. *Nature*, 456(7218):18–21, 2008. doi:10.1038/456018a.

[3] T. B. Sackton and D. L. Hartl. Genotypic context and epistasis in individuals and populations. *Cell*, 166(2):279–287, 2016. doi:10.1016/j.cell.2016.06.047.

[4] X. Hu, Q. Liu, Z. Zhang, Z. Li, S. Wang, L. He, and Y. Shi. SHEsisEpi, a GPU-enhanced genome-wide SNP-SNP interaction scanning algorithm, efficiently reveals the risk genetic epistasis in bipolar disorder. *Cell Res*, 20(7):854–7, 2010. doi:10.1038/cr.2010.68.

[5] L. S. Yung, C. Yang, X. Wan, and W. C. Yu. Gboost: a GPU-based tool for detecting gene-gene interactions in genome-wide case control studies. *Bioinformatics*, 27(9):1309–1310, 2011. doi:10.1093/bioinformatics/btr114.

[6] M. S. Kwon, K. Kim, S. Lee, and T. Park. cuGWAM: Genome-wide association multifactor dimensionality reduction using **CUDA**-enabled high-performance graphics processing unit. *International Journal of Data Mining and Bioinformatics*, 6(5):471–481, 2012. doi:10.1504/Ijdmb.2012.049301.

[7] T. Kam-Thong, C. A. Azencott, L. Cayton, B. Pütz, A. Altmann, N. Karbalai, P. G. Samann, B. Schölkopf, B. Müller-Myhsok, and K. M. Borgwardt. GLIDE: GPU-based linear regression for detection of epistasis. *Human Heredity*, 73(4):220–236, 2012. doi:10.1159/000341885.

[8] D. Sluga, T. Curk, B. Zupan, and U. Lotric. Heterogeneous computing architecture for fast detection of SNP-SNP interactions. *BMC Bioinformatics*, 15:216, 2014. doi:10.1186/1471-2105-15-216.

[9] D. Junger, C. Hundt, J. G. Dominguez, and B. Schmidt. Speed and accuracy improvement of higher-order epistasis detection on **CUDA**-enabled gpus. *Cluster Computing-the Journal of Networks Software Tools and Applications*, 20(3):1899–1908, 2017. doi:10.1007/s10586-017-0938-9.

[10] R. H. Chung and C. C. Shih. SeqSIMLA: a sequence and phenotype simulation tool for complex disease studies. *BMC Bioinformatics*, 14(1):199, 2013. doi:10.1186/1471-2105-14-199.

[11] C. Sabatti, S. K. Service, A. L. Hartikainen, A. Pouta, S. Ripatti, J. Brodsky, C. G. Jones, N. A. Zaitlen, T. Varilo, M. Kaakinen, U. Sovio, A. Ruokonen, J. Laitinen, E. Jakkula, L. Coin, C. Hoggart, A. Collins, H. Turunen, S. Gabriel, P. Elliot, M. I. McCarthy, M. J. Daly, M. R. Jarvelin, N. B. Freimer, and L. Peltonen. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet*, 41(1):35–46, 2009. doi:10.1038/ng.271.

[12] T. M. Teslovich, K. Musunuru, A. V. Smith, A. C. Edmondson, I. M. Stylianou, M. Koseki, J. P. Pirruccello, S. Ripatti, D. I. Chasman, C. J. Willer, C. T. Johansen, S. W. Fouchier, A. Isaacs, G. M. Peloso, M. Barbalic, S. L. Ricketts, J. C. Bis, Y. S. Aulchenko, G. Thorleifsson, M. F. Feitosa, J. Chambers, M. Orho-Melander, O. Melander, T. Johnson, X. Li, X. Guo, M. Li, Y. Shin Cho, M. Jin Go, Y. Jin Kim, J. Y. Lee, T. Park, K. Kim, X. Sim, R. Twee-Hee Ong, D. C. Croteau-Chonka, L. A. Lange, J. D. Smith, K. Song, J. Hua Zhao, X. Yuan, J. Luan, C. Lamina, A. Ziegler, W. Zhang, R. Y. Zee, A. F. Wright, J. C. Witteman, J. F. Wilson, G. Willemsen, H. E. Wichmann, J. B. Whitfield, D. M. Waterworth, N. J. Wareham, G. Waeber, P. Vollenweider, B. F. Voight, V. Vitart, A. G. Uitterlinden, M. Uda, J. Tuomilehto, J. R. Thompson, T. Tanaka, I. Surakka, H. M. Stringham, T. D. Spector, N. Soranzo, J. H. Smit, J. Sin-

isalo, K. Silander, E. J. Sijbrands, A. Scuteri, J. Scott, D. Schlessinger, S. Sanna, V. Salomaa, J. Saharinen, C. Sabatti, A. Ruokonen, I. Rudan, L. M. Rose, R. Roberts, M. Rieder, B. M. Psaty, P. P. Pramstaller, I. Pichler, M. Perola, B. W. Penninx, N. L. Pedersen, C. Pattaro, A. N. Parker, G. Pare, B. A. Oostra, C. J. O'Donnell, M. S. Nieminen, D. A. Nickerson, G. W. Montgomery, T. Meitinger, R. McPherson, M. I. McCarthy, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707–13, 2010. doi:10.1038/nature09270.

[13] P. F. O'Reilly, C. J. Hoggart, Y. Pomyen, F. C. F. Calboli, P. Elliott, M. R. Jarvelin, and L. J. M. Coin. **MultiPhen**: Joint model of multiple phenotypes can increase discovery in GWAS. *Plos One*, 7(5), 2012. doi:10.1371/journal.pone.0034861.

[14] S. H. Lee, J. Yang, M. E. Goddard, P. M. Visscher, and N. R. Wray. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*, 28(19):2540–2542, 2012. doi:10.1093/bioinformatics/bts474.

[15] S. H. Lee, S. Ripke, B. M. Neale, S. V. Faraone, S. M. Purcell, R. H. Perlis, B. J. Mowry, A. Thapar, M. E. Goddard, J. S. Witte, D. Absher, I. Agartz, H. Akil, F. Amin, O. A. Andreassen, A. Anjorin, R. Anney, V. Anttila, D. E. Arking, P. Asherson, M. H. Azevedo, L. Backlund, J. A. Badner, A. J. Bailey, T. Banaschewski, J. D. Barchas, M. R. Barnes, T. B. Barrett, N. Bass, A. Battaglia, M. Bauer, M. Bayes, F. Bellivier, S. E. Bergen, W. Berrettini, C. Betancur, T. Bettecken, J. Biederman, E. B. Binder, D. W. Black, D. H. R. Blackwood, C. S. Bloss, M. Boehnke, D. I. Boomsma, G. Breen, R. Breuer, R. Bruggeman, P. Cormican, N. G. Buccola, J. K. Buitelaar, W. E. Bunney, J. D. Buxbaum, W. F. Byerley, E. M. Byrne, S. Caesar, W. Cahn, R. M. Cantor, M. Casas, A. Chakravarti, K. Chambert, K. Choudhury, S. Cichon, C. R. Cloninger, D. A. Collier, E. H. Cook, H. Coon, B. Cormand, A. Corvin, W. H. Coryell, D. W. Craig, I. W. Craig, J. Crosbie, M. L. Cuccaro, D. Curtis, D. Czamara, S. Datta, G. Dawson, R. Day, E. J. De Geus, F. Degenhardt, S. Djurovic, G. J. Donohoe, A. E. Doyle, J. B. Duan, F. Dudbridge, E. Duketis, R. P. Ebstein, H. J. Edenberg, J. Elia, S. Ennis, B. Etain, A. Fanous, A. E. Farmer, I. N. Ferrier, M. Flickinger, E. Fombonne, T. Foroud, J. Frank, B. Franke,

C. Fraser, et al. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics*, 45(9):984–+, 2013. doi:10.1038/ng.2711.

[16] A. Dahl, V. Iotchkova, A. Baud, Å. Johansson, U. Gyllensten, N. Soranzo, R. Mott, A. Kranis, and J. Marchini. A multiple-phenotype imputation method for genetic studies. *Nature Genetics*, 48:466, 2016. doi:10.1038/ng.3513.

[17] D. Agniel and T. Cai. Analysis of multiple diverse phenotypes via semiparametric canonical correlation analysis. *Biometrics*, 73(4):1254–1265, 2017. doi:10.1111/biom.12690.

[18] Y. Hu, Q. Lu, W. Liu, Y. Zhang, M. Li, and H. Zhao. Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. *PLOS Genetics*, 13(6):e1006836, 2017. doi:10.1371/journal.pgen.1006836.

[19] P. Turley, R. K. Walters, O. Maghzian, A. Okbay, J. J. Lee, M. A. Fontana, T. A. Nguyen-Viet, R. Wedow, M. Zacher, N. A. Furlotte, P. Magnusson, S. Oskarsson, M. Johannesson, P. M. Visscher, D. Laibson, D. Cesarini, B. M. Neale, D. J. Benjamin, M. Agee, B. Alipanahi, A. Auton, R. K. Bell, K. Bryc, S. L. Elson, P. Fontanillas, N. A. Furlotte, D. A. Hinds, B. S. Hromatka, K. E. Huber, A. Kleinman, N. K. Litterman, M. H. McIntyre, J. L. Mountain, C. A. M. Northover, J. F. Sathirapongsasuti, O. V. Sazonova, J. F. Shelton, S. Shringarpure, C. Tian, J. Y. Tung, V. Vacic, C. H. Wilson, S. J. Pitts, T. andMe Research, and C. Social Science Genetic Association. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature Genetics*, 50(2):229–237, 2018. doi:10.1038/s41588-017-0009-4.

[20] L. Bottolo, M. Chadeau-Hyam, D. I. Hastie, T. Zeller, B. Liquet, P. Newcombe, L. Yengo, P. S. Wild, A. Schillert, A. Ziegler, S. F. Nielsen, A. S. Butterworth, W. K. Ho, R. Castagné, T. Munzel, D. Tregouet, M. Falchi, F. Cambien, B. G. Nordestgaard, F. Fumeron, A. Tybjærg-Hansen, P. Froguel, J. Danesh, E. Petretto, S. Blankenberg, L. Tiret, and S. Richardson. GUESS-ing polygenic associations with multiple phenotypes using a GPU-based evolutionary stochastic search algorithm. *PLOS Genetics*, 9(8):e1003657, 2013. doi:10.1371/journal.pgen.1003657.

[21] C. Webber. Epistasis in neuropsychiatric disorders. *Trends in Genetics*, 33(4):256–265, 2017. doi:10.1016/j.tig.2017.01.009.

[22] A. Gialluisi, T. F. M. Andlauer, N. Mirza-Schreiber, K. Moll, J. Becker, P. Hoffmann, K. U. Ludwig, D. Czamara, B. St Pourcain, W. Brandler, F. Honbolygo, D. Toth, V. Csepe, G. Huguet, A. P. Morris, J. Hulslander, E. G. Willcutt, J. C. DeFries, R. K. Olson, S. D. Smith, B. F. Pennington, A. Vaessen, U. Maurer, H. Lyytinen, M. Peyrard-Janvid, P. H. T. Leppanen, D. Brandeis, M. Bonte, J. F. Stein, J. B. Talcott, F. Fauchereau, A. Wilcke, C. Francks, T. Bourgeron, A. P. Monaco, F. Ramus, K. Landerl, J. Kere, T. S. Scerri, S. Paracchini, S. E. Fisher, J. Schumacher, M. M. Nothen, B. Müller-Myhsok, and G. Schulte-Körne. Genome-wide association scan identifies new variants associated with a cognitive predictor of dyslexia. *Transl Psychiatry*, 9(1):77, 2019. doi:10.1038/s41398-019-0402-0.

[23] S. Mascheretti, A. De Luca, V. Trezzi, D. Peruzzo, A. Nordio, C. Marino, and F. Arrigoni. Neurogenetics of developmental dyslexia: from genes to behavior through brain neuroimaging and cognitive and sensorial mechanisms. *Translational psychiatry*, 7:e987, January 2017. doi:10.1038/tp.2016.240.

[24] A. Carrion-Castillo, B. Franke, and S. E. Fisher. Molecular genetics of dyslexia: An overview. 19(4):214–240, 2013. doi:doi.org/10.1002/dys.1464.

[25] J. Kere. The molecular genetics and neurobiology of developmental dyslexia as model of a complex phenotype. *Biochemical and Biophysical Research Communications*, 452(2):236–243, 2014. doi:https://doi.org/10.1016/j.bbrc.2014.07.102.

[26] D. A. Hafler and P. L. De Jager. Applying a new generation of genetic maps to understand human inflammatory disease. *Nat Rev Immunol*, 5(1):83–91, 2005. doi:10.1038/nri1532.

[27] J. C. Castle. Snps occur in regions with less genomic sequence conservation. *PLoS One*, 6(6):e20660, 2011. doi:10.1371/journal.pone.0020660.

[28] R. J. Klein, C. Zeiss, E. Y. Chew, J. Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, M. B. Bracken, F. L.

Ferris, J. Ott, C. Barnstable, and J. Hoh. Complement factor H polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389, 2005. doi:10.1126/science.1109557.

[29] C. Genomes Project, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, and G. R. Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015. doi:10.1038/nature15393.

[30] P. R. Burton, D. G. Clayton, L. R. Cardon, N. Craddock, P. Deloukas, A. Duncanson, D. P. Kwiatkowski, M. I. McCarthy, W. H. Ouwehand, N. J. Samani, J. A. Todd, P. Donnelly, J. C. Barrett, D. Davison, D. Easton, D. Evans, H. T. Leung, J. L. Marchini, A. P. Morris, C. C. A. Spencer, M. D. Tobin, A. P. Attwood, J. P. Boorman, B. Cant, U. Everson, J. M. Hussey, J. D. Jolley, A. S. Knight, K. Koch, E. Meech, S. Nutland, C. V. Prowse, H. E. Stevens, N. C. Taylor, G. R. Walters, N. M. Walker, N. A. Watkins, T. Winzer, R. W. Jones, W. L. McArdle, S. M. Ring, D. P. Strachan, M. Pembrey, G. Breen, D. St Clair, S. Caesar, K. Gordon-Smith, L. Jones, C. Fraser, E. K. Green, D. Grozeva, M. L. Hamshere, P. A. Holmans, I. R. Jones, G. Kirov, V. Moskvina, I. Nikolov, M. C. O'Donovan, M. J. Owen, D. A. Collier, A. Elkin, A. Farmer, R. Williamson, P. McGuffin, A. H. Young, I. N. Ferrier, S. G. Ball, A. J. Balmforth, J. H. Barrett, D. T. Bishop, M. M. Iles, A. Maqbool, N. Yuldasheva, A. S. Hall, P. S. Braund, R. J. Dixon, M. Mangino, S. Stevens, J. R. Thompson, F. Bredin, M. Tremelling, M. Parkes, H. Drummond, C. W. Lees, E. R. Nimmo, J. Satsangi, S. A. Fisher, A. Forbes, C. M. Lewis, C. M. Onnie, N. J. Prescott, J. Sanderson, C. G. Mathew, J. Barbour, M. K. Mohiuddin, C. E. Todhunter, J. C. Mansfield, T. Ahmad, F. R. Cummings, D. P. Jewell, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007. doi:10.1038/nature05911.

[31] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang. Five years of GWAS discovery. *American Journal of Human Genetics*, 90(1):7–24, 2012. doi:10.1016/j.ajhg.2011.11.029.

[32] D. Botstein and N. Risch. Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nature Genetics*, 33:228–237, 2003. doi:10.1038/ng1090.

[33] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, and H. Parkinson. The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(D1):D1001–D1006, 2014. doi:10.1093/nar/gkt1229.

[34] P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang. 10 years of GWAS discovery: Biology, function, and translation. *Am J Hum Genet*, 101(1):5–22, 2017. doi:10.1016/j.ajhg.2017.06.005.

[35] A. R. Wood, T. Esko, J. Yang, S. Vedantam, T. H. Pers, S. Gustafsson, A. Y. Chu, K. Estrada, J. Luan, Z. Kutalik, N. Amin, M. L. Buchkovich, D. C. Croteau-Chonka, F. R. Day, Y. Duan, T. Fall, R. Fehrmann, T. Ferreira, A. U. Jackson, J. Karjalainen, K. S. Lo, A. E. Locke, R. Magi, E. Mihailov, E. Porcu, J. C. Randall, A. Scherag, A. A. Vinkhuyzen, H. J. Westra, T. W. Winkler, T. Workalemahu, J. H. Zhao, D. Absher, E. Albrecht, D. Anderson, J. Baron, M. Beekman, A. Demirkan, G. B. Ehret, B. Feenstra, M. F. Feitosa, K. Fischer, R. M. Fraser, A. Goel, J. Gong, A. E. Justice, S. Kanoni, M. E. Kleber, K. Kristiansson, U. Lim, V. Lotay, J. C. Lui, M. Mangino, I. Mateo Leach, C. Medina-Gomez, M. A. Nalls, D. R. Nyholt, C. D. Palmer, D. Pasko, S. Pechlivanis, I. Prokopenko, J. S. Ried, S. Ripke, D. Shungin, A. Stancakova, R. J. Strawbridge, Y. J. Sung, T. Tanaka, A. Teumer, S. Trompet, S. W. van der Laan, J. van Setten, J. V. Van Vliet-Ostaptchouk, Z. Wang, L. Yengo, W. Zhang, U. Afzal, J. Arnlov, G. M. Arscott, S. Bandinelli, A. Barrett, C. Bellis, A. J. Bennett, C. Berne, M. Bluher, J. L. Bolton, Y. Bottcher, H. A. Boyd, M. Bruinenberg, B. M. Buckley, S. Buyske, I. H. Caspersen, P. S. Chines, R. Clarke, S. Claudi-Boehm, M. Cooper, E. W. Daw, P. A. De Jong, J. Deelen, G. Delgado, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet*, 46(11):1173–86, 2014. doi:10.1038/ng.3097.

[36] O. Carlborg and C. S. Haley. Epistasis: too often neglected in complex trait studies? *Nat Rev Genet*, 5(8):618–25, 2004. doi:10.1038/nrg1407.

[37] P. C. Phillips. Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9(11):855–867, 2008. doi:10.1038/nrg2452.

[38] G. Hemani, K. Shakhbazov, H. J. Westra, T. Esko, A. K. Henders, A. F. McRae, J. Yang, G. Gibson, N. G. Martin, A. Metspalu, L. Franke, G. W. Montgomery, P. M. Visscher, and J. E. Powell. Detection and replication of epistasis influencing transcription in humans. *Nature*, 508(7495):249–53, 2014. doi:10.1038/nature13005.

[39] T. F. C. Mackay. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nature Reviews Genetics*, 15(1):22–33, 2014. doi:10.1038/nrg3627.

[40] G. Pedruzzi, A. Barlukova, and I. M. Rouzine. Evolutionary footprint of epistasis. *PLOS Computational Biology*, 14(9):e1006426, 2018. doi:10.1371/journal.pcbi.1006426.

[41] W. Bateson. *Mendel's principles of heredity*. University Press, Cambridge, 1909.

[42] R. A. Fisher. xv.— The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1918. doi:10.1017/S0080456800012163.

[43] C. Boone, H. Bussey, and B. J. Andrews. Exploring genetic interactions and networks with yeast. *Nature Reviews Genetics*, 8(6):437–449, 2007. doi:10.1038/nrg2085.

[44] J. M. Cheverud and E. J. Routman. Epistasis and its contribution to genetic variance components. *Genetics*, 139(3):1455–61, 1995.

[45] H. J. Cordell. Detecting gene-gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6):392–404, 2009. doi:10.1038/nrg2579.

[46] X. A. Wan, C. Yang, Q. A. Yang, H. Xue, X. D. Fan, N. L. S. Tang, and W. C. Yu. Boost: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *American Journal of Human Genetics*, 87(3):325–340, 2010. doi:10.1016/j.ajhg.2010.07.021.

[47] C. Lippert, J. Listgarten, R. I. Davidson, J. Baxter, H. Poon, C. M. Kadie, and D. Heckerman. An exhaustive epistatic SNP association analysis on expanded Wellcome Trust data (vol 3, pg 1099, 2013). *Scientific Reports*, 3, 2013. doi:ARTN132110.1038/srep01321.

[48] A. R. Wood, M. A. Tuke, M. A. Nalls, D. G. Hernandez, S. Bandinelli, A. B. Singleton, D. Melzer, L. Ferrucci, T. M. Frayling, and M. N. Weedon. Another explanation for apparent epistasis. *Nature*, 514(7520):E3–E5, 2014. doi:10.1038/nature13691.

[49] L. Zeng, N. Mirza-Schreiber, C. Lamina, S. Coassin, C. P. Nelson, O. Franzen, M. E. Kleber, S. Friedel, T. F. M. Andlauer, B. Jiang, B. Stiller, L. Li, C. Willenborg, M. Munz, T. Kessler, A. Kastrati, K.-L. Laugwitz, J. Erdmann, S. Moebus, M. M. Noethen, A. Peters, K. Strauch, M. Müller-Nurasyid, C. Gieger, T. Meitinger, E. Steinhagen-Thiessen, W. Maerz, J. L. M. Bjorkegren, N. J. Samani, F. Kronenberg, B. Müller-Myhsok, and H. Schunkert. *Cis*-epistasis at the *LPA* locus and risk of coronary artery disease. *bioRxiv*, page 518290, 2019. doi:10.1101/518290.

[50] R. Ameratunga, W. Koopmans, S. T. Woon, E. Leung, K. Lehnert, C. A. Slade, J. C. Tempany, A. Enders, R. Steele, P. Browett, P. D. Hodgkin, and V. L. Bryant. Epistatic interactions between mutations of TACI (*TNFRSF13B*) and *TCF3* result in a severe primary immunodeficiency disorder and systemic lupus erythematosus. *Clinical & Translational Immunology*, 6, 2017. doi:10.1038/cti.2017.41.

[51] R. Ameratunga, S. T. Woon, V. L. Bryant, R. Steele, C. Slade, E. Y. Leung, and K. Lehnert. Clinical implications of digenic inheritance and epistasis in primary immunodeficiency disorders. *Frontiers in Immunology*, 8, 2018. doi:ARTN196510.3389/fimmu.2017.01965.

[52] S. Turner, L. L. Armstrong, Y. Bradford, C. S. Carlson, D. C. Crawford, A. T. Crenshaw, M. de Andrade, K. F. Doheny, J. L. Haines, G. Hayes, G. Jarvik, L. Jiang, I. J. Kullo, R. Li, H. Ling, T. A. Manolio, M. Matsumoto, C. A. McCarty, A. N. McDavid, D. B. Mirel, J. E. Paschall, E. W. Pugh, L. V. Rasmussen, R. A. Wilke, R. L. Zuvich, and M. D. Ritchie. Quality control procedures for genome-wide association studies. *Curr Protoc Hum Genet*, Chapter 1:Unit1 19, 2011. doi:10.1002/0471142905.hg0119s68.

[53] C. A. Anderson, F. H. Pettersson, G. M. Clarke, L. R. Cardon, A. P. Morris, and K. T. Zondervan. Data quality control in genetic case-control association studies. *Nat Protoc*, 5(9):1564–73, 2010. doi:10.1038/nprot.2010.116.

[54] J. E. Wigginton, D. J. Cutler, and G. R. Abecasis. A note on exact tests of Hardy-Weinberg equilibrium. *American Journal of Human Genetics*, 76(5):887–893, 2005. doi:10.1086/429864.

[55] W. S. Bush and J. H. Moore. Chapter 11: Genome-wide association studies. *PLoS Comput Biol*, 8(12):e1002822, 2012. doi:10.1371/journal.pcbi.1002822.

[56] B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, 1999. doi:doi.org/10.1111/j.0006-341X.1999.00997.x.

[57] G. A. Satten, W. D. Flanders, and Q. H. Yang. Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *American Journal of Human Genetics*, 68(2):466–477, 2001. doi:doi.org/10.1086/318195.

[58] J. K. Pritchard, M. Stephens, N. A. Rosenberg, and P. Donnelly. Association mapping in structured populations. *American Journal of Human Genetics*, 67(1):170–181, 2000. doi:10.1086/302959.

[59] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38(8):904–9, 2006. doi:10.1038/ng1847.

[60] G. Kimmel, M. I. Jordan, E. Halperin, R. Shamir, and R. M. Karp. A randomization test for controlling population stratification in whole-genome association studies. *American Journal of Human Genetics*, 81(5):895–905, 2007. doi:10.1086/521372.

[61] N. Patterson, A. L. Price, and D. Reich. Population structure and eigenanalysis. *Plos Genetics*, 2(12):2074–2093, 2006. doi:10.1371/journal.pgen.0020190.

[62] Q. Z. Li and K. Yu. Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. *Genetic Epidemiology*, 32(3):215–226, 2008. doi:10.1002/gepi.20296.

[63] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham. PLINK:

a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3):559–75, 2007. doi:10.1086/519795.

[64] O. Delaneau, B. Howie, A. J. Cox, J. F. Zagury, and J. Marchini. Haplotype estimation using sequencing reads. *American Journal of Human Genetics*, 93(4):687–696, 2013. doi:10.1016/j.ajhg.2013.09.002.

[65] E. Porcu, S. Sanna, C. Fuchsberger, and L. G. Fritsche. Genotype imputation in genome-wide association studies. *Curr Protoc Hum Genet*, Chapter 1:Unit 1 25, 2013. doi:10.1002/0471142905.hg0125s78.

[66] B. Howie, C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet*, 44(8):955–9, 2012. doi:10.1038/ng.2354.

[67] J. Marchini and B. Howie. Genotype imputation for genome-wide association studies. *Nat Rev Genet*, 11(7):499–511, 2010. doi:10.1038/nrg2796.

[68] S. R. Browning and B. L. Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, 81(5):1084–1097, 2007. doi:10.1086/521987.

[69] B. N. Howie, P. Donnelly, and J. Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, 5(6):e1000529, 2009. doi:10.1371/journal.pgen.1000529.

[70] O. Delaneau, J. F. Zagury, and J. Marchini. Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods*, 10(1):5–6, 2013. doi:10.1038/nmeth.2307.

[71] W. Gropp, E. Lusk, N. Doss, and A. Skjellum. A high-performance, portable implementation of the **MPI** message passing interface standard. *Parallel Computing*, 22(6):789–828, 1996. doi:10.1016/0167-8191(96)00024-5.

[72] L. Dagum and R. Menon. Openmp: An industry standard API for shared-memory programming. *IEEE Computational Science & Engineering*, 5(1):46–55, 1998. doi:10.1109/99.660313.

[73] A. Upton, O. Trelles, J. A. Cornejo-Garcia, and J. R. Perkins. Review: High-performance computing to detect epistasis in genome scale data sets. *Briefings in Bioinformatics*, 17(3):368–379, 2016. doi:10.1093/bib/bbv058.

[74] A. B. Yoo, M. A. Jette, and M. Grondona. SLURM: Simple linux utility for resource management. *Job Scheduling Strategies for Parallel Processing*, 2862:44–60, 2003.

[75] NVIDIA. Nvidia's next generation **CUDA** compute architecture: Fermi. v1.1, 2009. doi:citeulike-article-id:11826606.

[76] G. Hemani, A. Theocharidis, W. Wei, and C. Haley. Epigpu: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. *Bioinformatics*, 27(11):1462–5, 2011. doi:10.1093/bioinformatics/btr172.

[77] G. K. Chen and Y. Guo. Discovering epistasis in large scale genetic association studies by exploiting graphics cards. *Front Genet*, 4:266, 2013. doi:10.3389/fgene.2013.00266.

[78] E. E. Schadt, J. Lamb, X. Yang, J. Zhu, S. Edwards, D. GuhaThakurta, S. K. Sieberts, S. Monks, M. Reitman, C. Zhang, P. Y. Lum, A. Leonardson, R. Thieringer, J. M. Metzger, L. Yang, J. Castle, H. Zhu, S. F. Kash, T. A. Drake, A. Sachs, and A. J. Lusis. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, 37:710, 2005. doi:10.1038/ng1589.

[79] H. Aschard, B. J. Vilhjalmsson, N. Greliche, P. E. Morange, D. A. Tregouet, and P. Kraft. Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *Am J Hum Genet*, 94(5):662–76, 2014. doi:10.1016/j.ajhg.2014.03.016.

[80] J. S. Ried, J. M. Jeff, A. Y. Chu, J. L. Bragg-Gresham, J. van Dongen, J. E. Huffman, T. S. Ahluwalia, G. Cadby, N. Eklund, J. Eriksson, T. Esko, M. F. Feitosa, A. Goel, M. Gorski, C. Hayward, N. L. Heard-Costa, A. U. Jackson, E. Jokinen, S. Kanoni, K. Kristiansson, Z. Kutalik, J. Lahti, J. A. Luan, R. Magi, A. Mahajan, M. Mangino, C. Medina-Gomez, K. L. Monda, I. M. Nolte, L. Perusse, I. Prokopenko, L. Qi, L. M. Rose, E. Salvi, M. T. Smith, H. Snieder, A. Stancakova, Y. J. Sung, I. Tachmazidou, A. Teumer,

G. Thorleifsson, P. van der Harst, R. W. Walker, S. R. Wang, S. H. Wild, S. M. Willems, A. Wong, W. H. Zhang, E. Albrecht, A. C. Alves, S. J. L. Bakker, C. Barlassina, T. M. Bartz, J. Beilby, C. Bellis, R. N. Bergman, S. Bergmann, J. Blangero, M. Bluher, E. Boerwinkle, L. L. Bonnycastle, S. R. Bornstein, M. Bruinenberg, H. Campbell, Y. D. I. Chen, C. W. K. Chiang, P. S. Chines, F. S. Collins, F. Cucca, L. A. Cupples, F. D'Avila, E. J. C. de Geus, G. Dedoussis, M. Dimitriou, A. Doring, J. G. Eriksson, A. E. Farmaki, M. Farrall, T. Ferreira, K. Fischer, N. G. Forouhi, N. Friedrich, A. P. Gjesing, N. Glorioso, M. Graff, H. Grallert, N. Grarup, J. Grassler, J. Grewal, A. Hamsten, M. N. Harder, C. A. Hartman, M. Hassinen, N. Hastie, A. T. Hattersley, A. S. Havulinna, M. Heliovaara, H. Hillege, A. Hofman, O. Holmen, et al. A principal component meta-analysis on multiple anthropometric traits identifies novel loci for body shape. *Nature Communications*, 7, 2016. doi:ARTN1335710.1038/ncomms13357.

[81] S. Kim, K. A. Sohn, and E. P. Xing. A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformatics*, 25(12):I204–I212, 2009. doi:10.1093/bioinformatics/btp218.

[82] P. Mitteroecker, J. M. Cheverud, and M. Pavlicev. Multivariate analysis of genotype-phenotype association. *Genetics*, 202(4):1345–63, 2016. doi:10.1534/genetics.115.181339.

[83] Z. Liu and X. Lin. Multiple phenotype association tests using summary statistics in genome-wide association studies. *Biometrics*, 74(1):165–175, 2018. doi:10.1111/biom.12735.

[84] H. C. van Houwelingen, L. R. Arends, and T. Stijnen. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine*, 21(4):589–624, 2002. doi:DOI10.1002/sim.1040.

[85] J. J. Kirkham, R. D. Riley, and P. R. Williamson. A multivariate meta-analysis approach for reducing the impact of outcome reporting bias in systematic reviews. *Statistics in Medicine*, 31(20):2179–95, 2012. doi:10.1002/sim.5356.

[86] D. Mavridis and G. Salanti. A practical introduction to multivariate meta-analysis. *Stat Methods Med Res*, 22(2):133–58, 2013. doi:10.1177/0962280211432219.

[87] N. L. Dimou, K. G. Pantavou, G. G. Braliou, and P. G. Bagos. *Multivariate Methods for Meta-Analysis of Genetic Association Studies*, pages 157–182. Springer New York, New York, NY, 2018. doi:10.1007/978-1-4939-7868-7_11.

[88] D. Jackson, R. Riley, and I. R. White. Multivariate meta-analysis: Potential and promise. *Statistics in Medicine*, 30(20):2481–2498, 2011. doi:10.1002/sim.4172.

[89] P. G. Bagos. Meta-analysis in stata using gllamm. *Res Synth Methods*, 6(4):310–32, 2015. doi:10.1002/jrsm.1157.

[90] W. Viechtbauer. Conducting meta-analyses in R with the **metafor** package. *Journal of Statistical Software*, 36(3):1–48, 2010.

[91] I. R. White. Multivariate random-effects meta-analysis. *Stata Journal*, 9(1):40–56, 2009. doi:10.1177/1536867x0900900103.

[92] D. Jackson, I. R. White, and S. G. Thompson. Extending dersimonian and laird's methodology to perform multivariate random effects meta-analyses. *Statistics in Medicine*, 29(12):1282–97, 2010. doi:10.1002/sim.3602.

[93] A. Gasparrini, B. Armstrong, and M. G. Kenward. Multivariate meta-analysis for non-linear and other multi-parameter associations. *Statistics in Medicine*, 31(29):3821–3839, 2012. doi:10.1002/sim.5471.

[94] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*, 106(23):9362–7, 2009. doi:10.1073/pnas.0903103106.

[95] S. Atwell, Y. S. Huang, B. J. Vilhjálmsson, G. Willems, M. Horton, Y. Li, D. Meng, A. Platt, A. M. Tarone, T. T. Hu, R. Jiang, N. W. Muliyati, X. Zhang, M. A. Amer, I. Baxter, B. Brachi, J. Chory, C. Dean, M. Debieu, J. de Meaux, J. R. Ecker, N. Faure, J. M. Kniskern, J. D. G. Jones, T. Michael, A. Nemri, F. Roux, D. E. Salt, C. Tang, M. Todesco, M. B. Traw, D. Weigel, P. Marjoram, J. O. Borevitz, J. Bergelson, and M. Nordborg. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature*, 465:627–631, June 2010. doi:10.1038/nature08800.

[96] B. E. Stranger, E. A. Stahl, and T. Raj. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, 187(2):367–83, 2011. doi:10.1534/genetics.110.120907.

[97] R. A. Fisher. *The Genetical Theory of Natural Selection*. The Clarendon Press, 1930. doi:10.5962/bhl.title.27468.

[98] S. Wright. Evolution in Mendelian populations. *Genetics*, 16(2):97–159, 1931. doi:10.1007/bf02459575.

[99] H. J. Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20):2463–2468, 2002. doi:10.1093/hmg/11.20.2463.

[100] D. S. Falconer and T. F. Mackay. *Introduction to Quantitative Genetics (4th Edition)*. Longman, 1996.

[101] M. Lynch, B. Walsh, et al. *Genetics and Analysis of Quantitative Traits*, volume 1. Sinauer Associates, 1998.

[102] W. G. Hill, M. E. Goddard, and P. M. Visscher. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet*, 4(2):e1000008, 2008. doi:10.1371/journal.pgen.1000008.

[103] J. L. Sztepanacz and M. W. Blows. Dominance genetic variance for traits under directional selection in *Drosophila serrata*. *Genetics*, 200(1):371–84, 2015. doi:10.1534/genetics.115.175489.

[104] T. L. Edwards, W. S. Bush, S. D. Turner, S. M. Dudek, E. S. Torstenson, M. Schmidt, E. Martin, and M. D. Ritchie. Generating linkage disequilibrium patterns in data simulations using genomesimla, 2008. URL: https://doi.org/10.1007/978-3-540-78757-0_3.

[105] J. H. Oh and J. O. Deasy. Sitdem: a simulation tool for disease/endpoint models of association studies based on single nucleotide polymorphism genotypes. *Comput Biol Med*, 45:136–42, 2014. doi:10.1016/j.compbiomed.2013.11.021.

[106] L. Liang, S. Zollner, and G. R. Abecasis. Genome: a rapid coalescent-based whole genome simulator. *Bioinformatics*, 23(12):1565–7, 2007. doi:10.1093/bioinformatics/btm138.

[107] T. Gunther, I. Gawenda, and K. J. Schmid. phenosim—a software to simulate phenotypes for testing in genome-wide association studies. *BMC Bioinformatics*, 12(1):265, 2011. doi:10.1186/1471-2105-12-265.

[108] J. Shang, J. Zhang, X. Lei, W. Zhao, and Y. Dong. **EpiSIM**: Simulation of multiple epistasis, linkage disequilibrium patterns and haplotype blocks for genome-wide interaction analysis. *Genes & Genomics*, 35(3):305–316, 2013. doi:10.1007/s13258-013-0081-9.

[109] R Core Team. R: A language and environment for statistical computing, 2017. URL: https://www.R-project.org/.

[110] H. Schwender, Q. Li, P. Berger, C. Neumann, M. Taub, and I. Ruczinski. **trio***: Testing of SNPs and SNP Interactions in Case-Parent Trio Studies*, 2015. R package version 3.14.0.

[111] W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K. D. Hansen, R. A. Irizarry, M. Lawrence, M. I. Love, J. MacDonald, V. Obenchain, A. K. Oles, H. Pages, A. Reyes, P. Shannon, G. K. Smyth, D. Tenenbaum, L. Waldron, and M. Morgan. Orchestrating high-throughput genomic analysis with bioconductor. *Nat Methods*, 12(2):115–21, 2015. doi:10.1038/nmeth.3252.

[112] Python Core Team. Python: A dynamic, open source programming language, 2017. URL: https://www.python.org/.

[113] B. Jiang and B. Pütz. **SimPhe***: Tools to Simulate Phenotype(s) with Epistatic Interaction*, 2018. R package version 0.2.0. URL: https://CRAN.R-project.org/package=SimPhe.

[114] C. C. Cockerham. An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics*, 39(6):859–882, 1954.

[115] B. Hayman and K. Mather. The description of genic interactions in continuous variation. *Biometrics*, 11(1):69–82, 1955. doi:10.2307/3001481.

[116] O. Kempthorne. *An Introduction to Genetic Statistics*. John Wiley & Sons, 1957. doi:10.2307/1526810.

[117] C. H. Kao and Z. B. Zeng. Modeling epistasis of quantitative trait loci using Cockerham's model. *Genetics*, 160(3):1243–1261, 2002.

[118] T. Wang and Z. B. Zeng. Contribution of genetic effects to genetic variance components with epistasis and linkage disequilibrium. *BMC Genet*, 10(1):52, 2009. doi:10.1186/1471-2156-10-52.

[119] C. C. Cockerham and B. S. Weir. Quadratic analyses of reciprocal crosses. *Biometrics*, 33(1):187–203, 1977. doi:10.2307/2529312.

[120] M. Slatkin. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nat Rev Genet*, 9(6):477–85, 2008. doi:10.1038/nrg2361.

[121] RStudio, Inc. Easy web applications in R, 2017. URL: http://www.rstudio.com/shiny/.

[122] W.-H. Wei, G. Hemani, and C. S. Haley. Detecting epistasis in human complex traits. *Nature Reviews Genetics*, 15:722, 2014. doi:10.1038/nrg3747.

[123] B. Jiang and B. Pütz. **episcan***: Scan Pairwise Epistasis*, 2018. R package version 0.0.1. URL: https://CRAN.R-project.org/package=episcan.

[124] T. Kam-Thong, D. Czamara, K. Tsuda, K. Borgwardt, C. M. Lewis, A. Erhardt-Lehmann, B. Hemmer, P. Rieckmann, M. Daake, F. Weber, C. Wolf, A. Ziegler, B. Pütz, F. Holsboer, B. Schölkopf, and B. Müller-Myhsok. EPIBLASTER—fast exhaustive two-locus epistasis detection strategy using graphical processing units. *European Journal of Human Genetics*, 19(4):465–471, 2011. doi:10.1038/ejhg.2010.196.

[125] T. Kam-Thong, B. Pütz, N. Karbalai, B. Müller-Myhsok, and K. Borgwardt. Epistasis detection on quantitative phenotypes by exhaustive enumeration using GPUs. *Bioinformatics*, 27(13):i214–21, 2011. doi:10.1093/bioinformatics/btr218.

[126] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In S. Jain, H. U. Simon, and E. Tomita, editors, *Algorithmic Learning Theory*, pages 63–77. Springer Berlin Heidelberg.

[127] **clBLAS**. `https://github.com/clMathLibraries/clBLAS`. Accessed: 2016-05-14.

[128] K. Rupp, P. Tillet, F. Rudolf, J. Weinbub, A. Morhammer, T. Grasser, A. Jüngel, and S. Selberherr. Viennacl—linear algebra library for multi- and many-core architectures. *SIAM Journal on Scientific Computing*, 38(5):S412–S439, 2016. `doi:10.1137/15M1026419`.

[129] C. D. Jr. *gpuR: GPU Functions for R Objects.*, 2017. R package version 2.0.0. URL: `http://github.com/cdeterman/gpuR`.

[130] J. Stein. What is developmental dyslexia? *Brain Sci*, 8(2), 2018. `doi:10.3390/brainsci8020026`.

[131] C. J. Davis, J. Gayán, V. S. Knopik, S. D. Smith, L. R. Cardon, B. F. Pennington, R. K. Olson, and J. C. DeFries. Etiology of reading difficulties and rapid naming: The colorado twin study of reading disability. *Behavior Genetics*, 31(6):625–635, 2001. `doi:10.1023/a:1013305730430`.

[132] J. Gayán and R. K. Olson. Genetic and environmental influences on orthographic and phonological skills in children with reading disabilities. *Developmental neuropsychology*, 20(2):483–507, 2001.

[133] J. Gayán and R. K. Olson. Genetic and environmental influences on individual differences in printed word recognition. *Journal of Experimental Child Psychology*, 84(2):97–123, 2003. `doi:https://doi.org/10.1016/S0022-0965(02)00181-9`.

[134] J. Schumacher, P. Hoffmann, C. Schmäl, G. Schulte-Körne, and M. M. Nöthen. Genetics of dyslexia: the evolving landscape. *Journal of Medical Genetics*, 44:289–297, 2007. `doi:10.1136/jmg.2006.046516`.

[135] E. L. Meaburn, N. Harlaar, I. W. Craig, L. C. Schalkwyk, and R. Plomin. Quantitative trait locus association scan of early reading disability and ability using pooled DNA and 100K SNP microarrays in a sample of 5760 children. *Molecular Psychiatry*, 13:729, 2007. `doi:10.1038/sj.mp.4002063`.

[136] J. D. Eicher, N. R. Powers, L. L. Miller, N. Akshoomoff, D. G. Amaral, C. S. Bloss, O. Libiger, N. J. Schork, B. F. Darst, B. J. Casey, L. Chang, T. Ernst, J. Frazier, W. E. Kaufmann, B. Keating, T. Kenet, D. Kennedy, S. Mostofsky, S. S. Murray, E. R. Sowell, H. Bartsch, J. M. Kuperman, T. T. Brown, D. J. Hagler, A. M. Dale, T. L. Jernigan, B. St Pourcain,

G. D. Smith, S. M. Ring, and J. R. Gruen. Genome-wide association study of shared components of reading disability and language impairment. *Genes Brain and Behavior*, 12(8):792–801, 2013. doi:10.1111/gbb.12085.

[137] M. Luciano, D. M. Evans, N. K. Hansell, S. E. Medland, G. W. Montgomery, N. G. Martin, M. J. Wright, and T. C. Bates. A genome-wide association study for reading and language abilities in two population cohorts. *Genes Brain Behav*, 12(6):645–52, 2013. doi:10.1111/gbb.12053.

[138] L. L. Field, K. Shumansky, J. Ryan, D. Truong, E. Swiergala, and B. J. Kaplan. Dense-map genome scan for dyslexia supports loci at 4q13, 16p12, 17q22; suggests novel locus at 7q36. 12(1):56–69, 2013. doi:10.1111/gbb.12003.

[139] A. Gialluisi, D. F. Newbury, E. G. Wilcutt, R. K. Olson, J. C. DeFries, W. M. Brandler, B. F. Pennington, S. D. Smith, T. S. Scerri, N. H. Simpson, S. L. I. Consortium, M. Luciano, D. M. Evans, T. C. Bates, J. F. Stein, J. B. Talcott, A. P. Monaco, S. Paracchini, C. Francks, and S. E. Fisher. Genome-wide screening for DNA variants associated with reading and language traits. *Genes Brain and Behavior*, 13(7):686–701, 2014. doi:10.1111/gbb.12158.

[140] D. Truong, A. K. Adams, R. Boada, J. C. Frijters, D. E. Hill, M. W. Lovett, M. E. Mahone, E. G. Willcutt, M. Wolf, , J. C. Defries, S. E. Fisher, C. Francks, A. Gialluisi, R. K. Olson, B. Pennington, S. D. Smith, J. Bosson-Heenan, and J. R. Gruen. Multivariate genome-wide association study of rapid automatized naming and rapid alternating stimulus in Hispanic and African American youth. *bioRxiv*, 2017.

[141] A. Carrion-Castillo, B. Maassen, B. Franke, A. Heister, M. Naber, A. van der Leij, C. Francks, and S. E. Fisher. Association analysis of dyslexia candidate genes in a dutch longitudinal sample. *European Journal Of Human Genetics*, 25:452, 2017. doi:10.1038/ejhg.2016.194.

[142] O. Zuk, E. Hechter, S. R. Sunyaev, and E. S. Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A*, 109(4):1193–8, 2012. doi:10.1073/pnas.1119675109.

[143] T. F. C. Mackay and J. H. Moore. Why epistasis is important for tackling complex human disease genetics. *Genome Medicine*, 6, 2014. doi:10.1186/gm561.

[144] X. Zhou and M. Stephens. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, 11:407, 2014. doi:10.1038/nmeth.2848.

[145] E. G. Willcutt, B. F. Pennington, R. K. Olson, N. Chhabildas, and J. Hulslander. Neuropsychological analyses of comorbidity between reading disability and attention deficit hyperactivity disorder: in search of the common deficit. *Dev Neuropsychol*, 27(1):35–78, 2005. doi:10.1207/s15326942dn2701_3.

[146] W. M. Brandler, A. P. Morris, D. M. Evans, T. S. Scerri, J. P. Kemp, N. J. Timpson, B. St Pourcain, G. D. Smith, S. M. Ring, J. Stein, A. P. Monaco, J. B. Talcott, S. E. Fisher, C. Webber, and S. Paracchini. Common variants in left/right asymmetry genes and pathways are associated with relative hand skill. *PLoS Genet*, 9(9):e1003751, 2013. doi:10.1371/journal.pgen.1003751.

[147] K. Moll, F. Ramus, J. Bartling, J. Bruder, S. Kunze, N. Neuhoff, S. Streiftau, H. Lyytinen, P. H. T. Leppanen, K. Lohvansuu, D. Toth, F. Honbolygo, V. Csepe, C. Bogliotti, S. Iannuzzi, J. F. Demonet, E. Longeras, S. Valdois, F. George, I. Soares-Boucaud, M. F. Le Heuzey, C. Billard, M. O'Donovan, G. Hill, J. Williams, D. Brandeis, U. Maurer, E. Schulz, S. van der Mark, B. Müller-Myhsok, G. Schulte-Körne, and K. Landerl. Cognitive mechanisms underlying reading and spelling development in five european orthographies. *Learning and Instruction*, 29:65–77, 2014. doi:10.1016/j.learninstruc.2013.09.003.

[148] K. Landerl, F. Ramus, K. Moll, H. Lyytinen, P. H. T. Leppanen, K. Lohvansuu, M. O'Donovan, J. Williams, J. Bartling, J. Bruder, S. Kunze, N. Neuhoff, D. Toth, F. Honbolygo, V. Csepe, C. Bogliotti, S. Iannuzzi, Y. Chaix, J. F. Demonet, E. Longeras, S. Valdois, C. Chabernaud, F. Delteil-Pinton, C. Billard, F. George, J. C. Ziegler, I. Comte-Gervais, I. Soares-Boucaud, C. L. Gerard, L. Blomert, A. Vaessen, P. Gerretsen, M. Ekkebus, D. Brandeis, U. Maurer, E. Schulz, S. van der Mark, B. Müller-Myhsok, and G. Schulte-Körne. Predictors of developmental dyslexia in eu-

ropean orthographies with varying complexity. *Journal of Child Psychology and Psychiatry*, 54(6):686–694, 2013. doi:10.1111/jcpp.12029.

[149] T. Lumley. **rmeta***: Meta-Analysis*, 2018. R package version 3.0. URL: https://CRAN.R-project.org/package=rmeta.

[150] K. S. Button, J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, and M. R. Munafo. Power failure: why small sample size undermines the reliability of neuroscience (vol 14, pg 365-376, 2013). *Nature Reviews Neuroscience*, 14(6):444–444, 2013. doi:10.1038/nrn3502.

[151] A. Okbay, J. P. Beauchamp, M. A. Fontana, J. J. Lee, T. H. Pers, C. A. Rietveld, P. Turley, G.-B. Chen, V. Emilsson, S. F. W. Meddens, S. Oskarsson, J. K. Pickrell, K. Thom, P. Timshel, R. de Vlaming, A. Abdellaoui, T. S. Ahluwalia, J. Bacelis, C. Baumbach, G. Bjornsdottir, J. H. Brandsma, M. Pina Concas, J. Derringer, N. A. Furlotte, T. E. Galesloot, G. Girotto, R. Gupta, L. M. Hall, S. E. Harris, E. Hofer, M. Horikoshi, J. E. Huffman, K. Kaasik, I. P. Kalafati, R. Karlsson, A. Kong, J. Lahti, S. J. v. d. Lee, C. deLeeuw, P. A. Lind, K.-O. Lindgren, T. Liu, M. Mangino, J. Marten, E. Mihailov, M. B. Miller, P. J. van der Most, C. Oldmeadow, A. Payton, N. Pervjakova, W. J. Peyrot, Y. Qian, O. Raitakari, R. Rueedi, E. Salvi, B. Schmidt, K. E. Schraut, J. Shi, A. V. Smith, R. A. Poot, B. St Pourcain, A. Teumer, G. Thorleifsson, N. Verweij, D. Vuckovic, J. Wellmann, H.-J. Westra, J. Yang, W. Zhao, Z. Zhu, B. Z. Alizadeh, N. Amin, A. Bakshi, S. E. Baumeister, G. Biino, K. Bønnelykke, P. A. Boyle, H. Campbell, F. P. Cappuccio, G. Davies, J.-E. De Neve, P. Deloukas, I. Demuth, J. Ding, P. Eibich, L. Eisele, N. Eklund, D. M. Evans, J. D. Faul, M. F. Feitosa, A. J. Forstner, I. Gandin, B. Gunnarsson, B. V. Halldórsson, T. B. Harris, A. C. Heath, L. J. Hocking, E. G. Holliday, G. Homuth, M. A. Horan, et al. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*, 533:539, 2016. doi:10.1038/nature17671.

[152] A. M. Undheim. Dyslexia and psychosocial factors. a follow-up study of young norwegian adults with a history of dyslexia in childhood. *Nordic Journal of Psychiatry*, 57(3):221–226, 2003. doi:10.1080/08039480310001391.

[153] K. Watanabe, E. Taskesen, A. van Bochoven, and D. Posthuma. Functional mapping and annotation of genetic associations with FUMA. *Nature Communications*, 8, 2017. doi:10.1038/s41467-017-01261-5.

[154] K. Wang, M. Y. Li, and H. Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16), 2010. doi:10.1093/nar/gkq603.

[155] M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46:310, 2014. doi:10.1038/ng.2892.

[156] A. P. Boyle, E. L. Hong, M. Hariharan, Y. Cheng, M. A. Schaub, M. Kasowski, K. J. Karczewski, J. Park, B. C. Hitz, S. Weng, J. M. Cherry, and M. Snyder. Annotation of functional variation in personal genomes using regulomedb. *Genome Research*, 22(9):1790–1797, 2012. doi:10.1101/gr.137323.112.

[157] J. Ernst and M. Kellis. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*, 9(3):215–216, 2012. doi:10.1038/nmeth.1906.

[158] A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y. C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K. H. Farh, S. Feizi, R. Karlic, A. R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. M. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L. H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic,

B. Ren, J. A. Stamatoyannopoulos, T. Wang, M. Kellis, and R. E. Consortium. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015. doi:10.1038/nature14248.

[159] R. Maier, G. Moser, G. B. Chen, S. Ripke, C. Cross-Disorder Working Group of the Psychiatric Genomics, W. Coryell, J. B. Potash, W. A. Scheftner, J. Shi, M. M. Weissman, C. M. Hultman, M. Landen, D. F. Levinson, K. S. Kendler, J. W. Smoller, N. R. Wray, and S. H. Lee. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. *Am J Hum Genet*, 96(2):283–94, 2015. doi:10.1016/j.ajhg.2014.12.006.

[160] H. Stefansson, A. Meyer-Lindenberg, S. Steinberg, B. Magnusdottir, K. Morgen, S. Arnarsdottir, G. Bjornsdottir, G. BragiWalters, G. A. Jonsdottir, O. M. Doyle, H. Tost, O. Grimm, S. Kristjansdottir, H. Snorrason, S. R. Davidsdottir, L. J. Gudmundsson, G. F. Jonsson, B. Stefansdottir, I. Helgadottir, M. Haraldsson, B. Jonsdottir, J. H. Thygesen, A. J. Schwarz, M. Didriksen, T. B. Stensbol, M. Brammer, S. Kapur, J. G. Halldorsson, S. Hreidarsson, E. Saemundsen, E. Sigurdsson, and K. Stefansson. CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature*, 505(7483):361–+, 2014. doi:10.1038/nature12818.

[161] M. Cederlof, B. Maughan, H. Larsson, B. M. D'Onofrio, and R. Plomin. Reading problems and major mental disorders — co-occurrences and familial overlaps in a swedish nationwide cohort. *Journal of Psychiatric Research*, 91:124–129, 2017. doi:10.1016/j.jpsychires.2017.03.014.

[162] D. Mehta, M. Gonik, T. Klengel, M. Rex-Haffner, A. Menke, J. Rubel, K. B. Mercer, B. Pütz, B. Bradley, F. Holsboer, K. J. Ressler, B. Müller-Myhsok, and E. B. Binder. Using polymorphisms in *FKBP5* to define biologically distinct subtypes of posttraumatic stress disorder: evidence from endocrine and gene expression studies. *Arch Gen Psychiatry*, 68(9):901–10, 2011. doi:10.1001/archgenpsychiatry.2011.50.

[163] B. S. Weir. *Genetic Data Analysis II*. Sinauer Associates, 1996. doi:10.2307/2533134.

A Appendix of SimPhe

A.1 Genetic Variance

$$
\begin{aligned}
V_G = {} & 2\pi_A\beta^2_{G_{w_1}} + \pi_A(1+\tau_A^2)\beta^2_{G_{w_2}} + 2\pi_B\beta^2_{G_{w_3}} + \pi_B(1+\tau_B^2)\beta^2_{G_{w_4}} \\
& + 2(\pi_A\tau_B^2 + \pi_B\tau_A^2 + 2\pi_A\pi_B - \tau_A\tau_B D)\beta^2_{G_{w_5}} \\
& + \frac{1}{4}[(1-2\pi_A) - \tau_B^2(\tau_A\tau_B + 4D)^2]\beta^2_{G_{w_6}} \\
& + \frac{1}{4}[(1-2\pi_B) - \tau_A^2(\tau_A\tau_B + 4D)^2]\beta^2_{G_{w_7}} \\
& + \frac{1}{16}[1 - (\tau_A\tau_B + 4D)^4]\beta^2_{G_{w_8}} \\
& + 4\pi_A\tau_A\beta_{G_{w_1}}\beta_{G_{w_2}} + 4D\beta_{G_{w_1}}\beta_{G_{w_3}} + 4\tau_B D\beta_{G_{w_1}}\beta_{G_{w_4}} - 4\tau_B\pi_A\beta_{G_{w_1}}\beta_{G_{w_5}} \\
& - 2(\pi_A\tau_B^2 + 4D^2)\beta_{G_{w_1}}\beta_{G_{w_6}} - 2[2\tau_B\pi_A\tau_A + (1-2\tau_A^2)D]\beta_{G_{w_1}}\beta_{G_{w_7}} \\
& - 2[\tau_A\pi_A\tau_B^2 - (\tau_A^2 - 4\pi_A)\tau_B D - 4\tau_A D^2]\beta_{G_{w_1}}\beta_{G_{w_8}} \\
& + 4\tau_A D\beta_{G_{w_2}}\beta_{G_{w_3}} - 4D(\tau_A\tau_B - 2D)\beta_{G_{w_2}}\beta_{G_{w_4}} - 4\pi_A(\tau_A\tau_B + 2D)\beta_{G_{w_2}}\beta_{G_{w_5}} \\
& - 2\tau_B\pi_A(\tau_A\tau_B + 4D)\beta_{G_{w_2}}\beta_{G_{w_6}} - 2[\tau_B\pi_A(\tau_A^2+1) + \tau_A^3 D]\beta_{G_{w_2}}\beta_{G_{w_7}} \\
& - [\pi_A\tau_B^2(1+\tau_A^2) - 2\tau_A^3\tau_B D - 4\tau_A^2 D^2]\beta_{G_{w_2}}\beta_{G_{w_8}} \\
& + 4\tau_B\pi_B\beta_{G_{w_3}}\beta_{G_{w_4}} - 4\tau_A\pi_B\beta_{G_{w_3}}\beta_{G_{w_5}} \\
& - 2[2\tau_A\pi_B\tau_B + (1-2\tau_B^2)D]\beta_{G_{w_3}}\beta_{G_{w_6}} - 2(\pi_B\tau_A^2 + 4D^2)\beta_{G_{w_3}}\beta_{G_{w_7}} \\
& - 2[2\tau_B\pi_B\tau_A^2 - (\tau_B^2 - 4\pi_B)\tau_A D - 4\tau_B D^2]\beta_{G_{w_3}}\beta_{G_{w_8}} \\
& - 4\pi_B(\tau_A\tau_B + 2D)\beta_{G_{w_4}}\beta_{G_{w_5}} \\
& - 2[\tau_A\pi_B(\tau_B^2+1) + \tau_B^3 D]\beta_{G_{w_4}}\beta_{G_{w_6}} - 2\tau_A\pi_B(\tau_A\tau_B + 4D)\beta_{G_{w_4}}\beta_{G_{w_7}} \\
& - [\pi_B\tau_A^2(1+\tau_B^2) - 2\tau_B^3\tau_A D - 4\tau_B^2 D^2]\beta_{G_{w_4}}\beta_{G_{w_8}} \\
& + 2[\tau_B(\pi_A + 2\pi_B\tau_A^2) + \tau_A(1-3\tau_B^2)D - 4\tau_B D^2]\beta_{G_{w_5}}\beta_{G_{w_6}} \\
& + 2[\tau_A(\pi_B + 2\pi_A\tau_B^2) + \tau_B(1-3\tau_A^2)D - 4\tau_A D^2]\beta_{G_{w_5}}\beta_{G_{w_7}} \\
& + \frac{1}{2}(\tau_A\tau_B + 2D)[1 - (\tau_A\tau_B + 4D)^2]\beta_{G_{w_5}}\beta_{G_{w_8}} \\
& + \frac{1}{2}[\tau_A\tau_B + 2D - \tau_A\tau_B(\tau_A\tau_B + 4D)^2]\beta_{G_{w_6}}\beta_{G_{w_7}} \\
& + \frac{1}{4}[\tau_A - \tau_B(\tau_A\tau_B + 4D)^3]\beta_{G_{w_6}}\beta_{G_{w_8}} + \frac{1}{4}[\tau_B - \tau_A(\tau_A\tau_B + 4D)^3]\beta_{G_{w_7}}\beta_{G_{w_8}}
\end{aligned}
\tag{A.1}
$$

The four terms, f_A, f_a, f_B, and f_b, denote the frequencies of alleles A, a, B, and b in locus A and B (remember f_A and f_B mean the *major* allele frequencies in locus A and B, f_a and f_b are the *minor* allele frequencies) while $\tau_A = f_a - f_A$, $\tau_B = f_b - f_B$, $\pi_A = f_a f_A$, $\pi_B = f_b f_B$ [163]. D is the LD score indicating the association of alleles at locus A and B. If we assume linkage equilibrium (LE), the quantity D is equal to 0.

B Appendix of Dyslexia

B.1 Datasets

Originally, the datasets includes German dataset, containing three sub-datasets (MarWu1, MarWu2, and Munich), accompanied with Austria, Colorado, Finland, France, Hungary, Netherlands, Switzerland, and UK datasets. There are eleven datasets in total (see Table B.1 and Figure B.1).

B.2 Results

B.2.1 Exploration

We also compared the p-values between multi-phenotype `epiHSIC` and multivariate meta-analysis (random effect, see Figure B.2).

Figure B.1: Non-word reading in all datasets. The colors indicate the different datasets (studies) and the symbols (0 and 1) represent whether the sample is control or case, respectively.

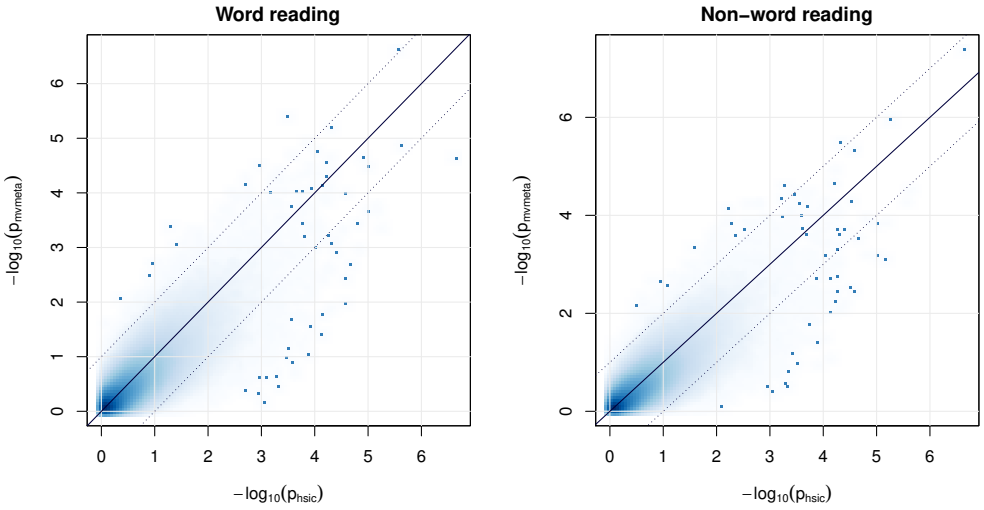

Figure B.2: Comparison of p-values between `epiHSIC` for multiple phenotypes and multivariate meta-analysis (random effect). p_{hsic} represents the p-values of Z-test in `epiHSIC` for multiple phenotypes for merged (exploration) data and p_{mvmeta} means the p-values of multivariate meta-analysis with random effects. The top 50 points were superimposed on the density image. All p-values were shown as $-\log_{10}(p\text{-values})$.

Table B.1: Sample size of two phenotypes in all datasets

Study	nPhe	nWR(without NA)	nNWR(without NA)
Austria	328	328	328
Colorado	533	533	529
Finland	324	324	300
France	143	143	120
Hungary	236	236	236
MarWu1	193	193	193
MarWu2	197	197	197
Munich	352	352	351
Netherlands	232	232	230
Switzerland	56	56	56
UK	875	873	868
Total	3469	3467	3408

Note:
[1] nPhe: number of phenotype pairs;
[2] nWR(without NA): number of word reading without missing value(s);
[3] nNWR(without NA): number of non-word reading without missing value(s).

B.2.2 Discovery

There are no significant pairs which pass the Bonferroni correction threshold (1.36×10^{-12}). The top 20 significant epistasis for word reading and non-word reading are shown in Table B.2 and B.3, respectively.

Table B.2: Top 20 SNP pairs for word reading in the meta-analysis on the discovery datasets

SNP1	SNP2	Chr1	Chr2	Position1	Position2	mvmetaFix	ivmetaFix
rs56235036	rs10978628	2	9	35702379	109529338	5.19e-11	7.54e-11
rs7556628	rs8106506	1	19	26685190	57686211	5.59e-11	5.33e-10
rs659459	rs4030873	1	18	6873932	76673358	7.53e-11	3.91e-11
rs7566682	rs1316439	2	9	228154137	120250392	8.81e-11	1.32e-10
rs10802208	rs75837095	1	9	245462775	129161865	1.06e-10	1.06e-10
rs6446498	rs4236980	4	8	6400680	71035615	1.11e-10	1.88e-10
rs4146566	rs10847264	5	12	56101894	127518105	1.58e-10	1.25e-09
rs2075799	rs3108847	6	8	31778529	110724480	1.65e-10	1.68e-10
rs113094458	rs4758291	6	11	143935188	8125294	2.07e-10	3.37e-10
rs4569449	rs10770155	2	11	10229697	11018431	2.16e-10	8.13e-10
rs10222516	rs184467	3	13	1880191	29622636	2.20e-10	5.39e-10
rs61194909	rs11613110	2	12	99906576	75539335	2.39e-10	4.75e-10
rs2208173	rs72671845	6	8	95257918	91028394	2.49e-10	2.77e-10
rs1807395	rs740442	4	17	7487093	55779494	2.72e-10	6.05e-10
rs6586770	rs178376	8	14	18694360	80176931	2.82e-10	5.14e-10
rs2966625	rs9524522	7	13	12001136	95150154	3.00e-10	4.02e-10
rs17113884	rs61912375	5	11	152393125	133155383	3.17e-10	7.47e-10
rs7537024	rs28711477	1	14	29859210	78783124	3.43e-10	1.46e-09
rs12676286	rs11641691	8	16	81874602	12609267	3.53e-10	4.59e-10
rs12533769	rs569151	7	18	70193190	58215383	3.69e-10	3.96e-09

Note:

[1] SNP1/2: first/second SNP in interacting pair;

[2] Chr1/2: chromosome SNP1/2 lies on;

[3] Position1/2: position of SNP1/2 on its respective chromosome;

[4] mvmetaFix: p-value of multivariate meta-analysis with fixed effect;

[5] ivmetaFix: p-value of inverse variance based meta-analysis with fixed effect.

Table B.3: Top 20 SNP pairs for non-word reading in the meta-analysis on the discovery datasets

SNP1	SNP2	Chr1	Chr2	Position1	Position2	mvmetaFix	ivmetaFix
rs9877231	rs1104305	3	8	132562606	3212189	4.82e-11	2.85e-10
rs72903413	rs2051344	6	18	86073575	74715653	8.30e-11	8.55e-10
rs12464039	rs4409410	2	8	150732821	125941652	1.26e-10	8.78e-10
rs72918500	rs10421891	18	19	51457056	46315809	1.46e-10	1.59e-09
rs244436	rs1157636	5	11	166373903	24404247	1.55e-10	3.71e-10
rs7323715	rs11159295	13	14	91076725	78306035	1.60e-10	2.74e-10
rs61194909	rs11613110	2	12	99906576	75539335	1.68e-10	3.55e-10
rs13028210	rs9298395	2	8	115088054	83784479	1.74e-10	5.40e-11
rs72920305	rs10421891	18	19	51458536	46315809	2.21e-10	2.52e-09
rs11686138	rs12581660	2	12	11579054	83788369	2.26e-10	1.21e-09
rs62287078	rs10813980	3	9	175218378	33437140	2.27e-10	1.31e-09
rs9454693	rs12549115	6	8	69808955	18382444	2.32e-10	2.67e-10
rs2269652	rs37389	1	5	175097357	35085180	2.33e-10	1.99e-10
rs72920305	rs4802279	18	19	51458536	46322830	2.34e-10	7.05e-09
rs6716601	rs6915661	2	6	85596553	155411606	2.43e-10	8.31e-11
rs72918500	rs4802279	18	19	51457056	46322830	2.57e-10	5.03e-09
rs8179786	rs448247	2	10	202143032	30896831	2.59e-10	1.44e-09
rs486661	rs72458511	2	4	141368094	76988600	2.74e-10	1.75e-10
rs17813338	rs12605711	8	18	442755	69252263	2.87e-10	3.27e-10
rs12071420	rs10150544	1	14	167760165	30615135	2.90e-10	1.82e-10

Note:

[1] SNP1/2: first/second SNP in interacting pair;

[2] Chr1/2: chromosome SNP1/2 lies on;

[3] Position1/2: position of SNP1/2 on its respective chromosome;

[4] mvmetaFix: p-value of multivariate meta-analysis with fixed effect;

[5] ivmetaFix: p-value of inverse variance based meta-analysis with fixed effect.

B.2.3 Discovery and replication together

There is only one significant pair which passes the Bonferroni correction threshold (1.36×10^{-12}). The top 20 epistasis pairs for word reading and non-word reading are shown in Table B.4 and B.5, respectively.

Table B.4: Top 20 SNP pairs for word reading in the meta-analysis on the discovery and replication datasets

SNP1	SNP2	Chr1	Chr2	Position1	Position2	mvmetaFix	ivmetaFix
rs8013684	rs1442415	14	15	92365790	96216254	1.26e-12	3.83e-12
rs13182682	rs2580960	5	12	99263712	74063951	1.11e-11	6.83e-12
rs56235036	rs10978628	2	9	35702379	109529338	2.18e-11	4.82e-11
rs2771090	rs1596534	9	12	93080054	41172010	2.73e-11	2.56e-11
rs1418555	rs10205244	1	2	218523650	230208486	4.00e-11	5.38e-11
rs4633327	rs1894909	1	7	146959913	36810390	4.31e-11	6.76e-11
rs7807195	rs11044444	7	12	154443525	19328609	6.51e-11	3.08e-10
rs75222724	rs75546553	3	11	184882243	66258916	9.04e-11	1.51e-10
rs12133344	rs4936711	1	11	179758019	122457726	1.12e-10	1.95e-10
rs7861646	rs11373754	9	12	23732000	86315625	1.21e-10	1.87e-10
rs1480802	rs7278012	8	21	136224211	20231604	1.22e-10	7.21e-11
rs578446	rs604514	11	13	96030606	71331644	1.23e-10	3.04e-10
rs3765953	rs150589259	1	6	119917253	30436380	1.27e-10	8.25e-10
rs9995726	rs4810329	4	20	37904553	40266126	1.80e-10	1.75e-09
rs1489221	rs11655584	8	17	73548109	38139024	1.90e-10	2.31e-10
rs12224885	rs219613	11	21	6846268	27774233	2.37e-10	1.09e-09
rs7826828	rs2077345	8	10	134229883	27899128	2.42e-10	3.27e-10
rs2430047	rs7216082	7	17	122294225	29889952	2.85e-10	2.88e-09
rs7545768	rs2711019	1	7	237524328	25028903	2.89e-10	5.94e-10
rs9917220	rs11614973	2	12	212058193	133723621	3.52e-10	3.68e-10

Note:

[1] SNP1/2: first/second SNP in interacting pair;

[2] Chr1/2: chromosome SNP1/2 lies on;

[3] Position1/2: position of SNP1/2 on its respective chromosome;

[4] mvmetaFix: p-value of multivariate meta-analysis with fixed effect;

[5] ivmetaFix: p-value of inverse variance based meta-analysis with fixed effect.

Table B.5: Top 20 SNP pairs for non-word reading in the meta-analysis the discovery and replication datasets

SNP1	SNP2	Chr1	Chr2	Position1	Position2	mvmetaFix	ivmetaFix
rs41295077	rs2277271	10	11	6118262	132399272	2.71e-11	1.54e-10
rs1558852	rs17417046	2	18	51252137	50984997	3.43e-11	2.32e-10
rs9664198	rs1189827	10	14	81843222	57533464	6.19e-11	1.59e-10
rs11686138	rs12581660	2	12	11579054	83788369	1.05e-10	4.78e-10
rs61194909	rs11613110	2	12	99906576	75539335	1.07e-10	1.07e-10
rs767665	rs6094931	5	20	144327414	46617331	1.18e-10	1.22e-10
rs12995732	rs2241485	2	4	45545872	10100812	1.20e-10	1.99e-10
rs12071420	rs10150544	1	14	167760165	30615135	1.71e-10	1.09e-10
rs62366885	rs7312553	5	12	50622637	103982568	1.95e-10	2.52e-10
rs147692383	rs6475840	2	9	138908598	25230166	2.22e-10	2.70e-10
rs10434537	rs10778374	5	12	62010640	105660904	2.23e-10	1.72e-10
rs193464	rs12605672	7	18	145046518	42343914	2.80e-10	5.67e-10
rs9811939	rs993376	3	12	70368592	17795398	2.83e-10	3.07e-10
rs17170328	rs59738387	7	8	146834604	13200212	2.89e-10	1.33e-09
rs67156578	rs6475840	2	9	138909308	25230166	2.94e-10	8.03e-10
rs7768381	rs137891724	6	6	130037945	151835512	3.45e-10	2.04e-10
rs4846875	rs2271218	1	5	231038020	98110531	3.59e-10	1.20e-09
rs11232302	rs266319	11	15	80434157	67266756	3.72e-10	7.55e-10
rs2614143	rs177038	10	16	14123459	72500483	3.83e-10	4.58e-09
rs11740851	rs2828245	5	21	148306141	24911815	3.95e-10	8.09e-10

Note:

[1] SNP1/2: first/second SNP in interacting pair;

[2] Chr1/2: chromosome SNP1/2 lies on;

[3] Position1/2: position of SNP1/2 on its respective chromosome;

[4] mvmetaFix: p-value of multivariate meta-analysis with fixed effect;

[5] ivmetaFix: p-value of inverse variance based meta-analysis with fixed effect.