Detection of epistasis in genome wide association studies with machine learning methods for therapeutic target identification

Lotfi Slim

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à MINES ParisTech

Detection of Epistasis in Genome Wide Association Studies with Machine Learning Methods for Therapeutic Target Identification

Détection d'épistasie dans les études d'association pangénomiques avec des techniques d'apprentissage pour l'identification de cibles thérapeutiques

Soutenue par
Lotfi Slim
Le 11 Juin 2020

École doctorale n°621
Ingénierie des Systèmes, Matériaux, Mécanique, Energétique

Spécialité
Bio-informatique

Composition du jury :

Chloé-Agathe AZENCOTT MINES ParisTech	*Directrice de thèse*
Gilles BLANCHARD Universität Potsdam	*Président du jury* *Rapporteur*
Karsten BORGWARDT ETH Zürich	*Rapporteur*
Clément CHATELAIN SANOFI R&D	*Examinateur*
Pierre NEUVIAL CNRS & Institut de Mathématiques de Toulouse	*Examinateur*
Jean-Philippe VERT MINES ParisTech & Google	*Invité*
Véronique STOVEN MINES ParisTech	*Directrice de thèse*

MINES ParisTech | PSL

Abstract

By offering an unprecedented picture of the human genome, genome-wide association studies (GWAS) have been expected to fully explain the genetic background of complex diseases. So far, the results have been mitigated to say the least. This, among other things, can be partially attributed to the adopted statistical methodology, which does not often take into account interaction between genetic variants, or epistasis. The detection of epistasis through statistical models presents several challenges for which we develop in this thesis a pair of adequate tools. The first tool, epiGWAS, uses causal inference to detect epistatic interactions between a target SNP and the rest of the genome. The second tool, kernelPSI, instead uses kernel methods to model epistasis between nearby single-nucleotide polymorphisms (SNPs). It also leverages post-selection inference to jointly perform SNP-level selection and gene-level significance testing. The developed tools are – to the best of our knowledge – the first to extend powerful statistical learning frameworks such as causal inference and nonlinear post-selection inference to GWAS. In addition to the methodological contributions, a special emphasis was placed on biological interpretation to validate our findings in multiple sclerosis and body-mass index variations.

Résumé

En offrant une image sans précédent du génome humain, les études d'association pangénomiques (GWAS) expliqueraient pleinement le contexte génétique des maladies complexes. A ce jour, les résultats ont été pour le moins mitigés. Cela peut être partiellement attribué à la méthodologie statistique adoptée, qui ne prend pas souvent en compte l'interaction entre les variants génétiques, ou l'épistasie. La détection d'épistasie à travers des modèles statistiques présente plusieurs défis pour lesquels nous développons dans cette thèse une paire d'outils adéquats. Le premier outil, epiGWAS, utilise l'inférence causale pour détecter les interactions épistatiques entre un SNP cible et le reste du génome. Le deuxième outil, kernelPSI, utilise à la place des méthodes à noyaux pour modéliser l'épistasie entre plusieurs polymorphismes mononucléotidiques (SNPs) voisins. Il tire également partie de l'inférence post-sélection pour effectuer conjointement une sélection au niveau des SNPs et des tests de signification au niveau des gènes. Les outils développés sont - au meilleur de nos connaissances - les premiers à étendre au domains des GWAS des outils puissants d'apprentissage statistique tels que l'inférence causale et l'inférence post-sélection nonlinéaire. En plus des contributions méthodologiques, un accent particulier a été mis sur l'interprétation biologique pour valider nos résultats dans la sclérose en plaques et les variations d'indice de masse corporelle.

Acknowledgments

I would like to start by thanking Chloé-Agathe Azencott for being a supportive advisor and an enthusiastic researcher. Her unique expertise in GWAS helped me navigate the intricacies of the field. Her upbeat attitude has always made our discussions cheerful, even during difficult times. We couldn't help but joke sometimes about the future of GWAS, but other times, we concluded it was the way to go for bioinformatics and biomedical data science.

The second acknowledgement goes naturally for my second advisor, Clément Chatelain who has taught me a lot about biology and bioinformatics. His input was always valuable and was key to taking our work to the next level. I would also like to thank him for putting up with my whims and caprices.

My thesis director during the first half of my PhD was Jean-Philippe Vert. He is a careful listener and a demanding researcher. Both traits are key for success in academia. My third acknowledgement goes to him for helping me pushing the envelope.

Last but not least among my supervisors is Véronique Stoven. Not only was she incredibly helpful, but also a pleasant and agreeable person to speak to. I have always admired her energy and general culture.

I had the pleasure to collaborate with Hélène de Foucauld from Sanofi. She is a passionate biologist. By contrast, I consider myself a passionate statistician. Combining our respective fields was instrumental in making our work exhaustive and unique.

A PhD can not be successful without supportive lab mates. In that respect, I was incredibly lucky by being a member of the CBIO family. Our research topics are often diametrically opposed, but we always acted like a single and tight-knit group. I will miss our lunches at the Curie canteen and our CBIO meetings. I shared wonderful moments with the PhD candidates of my cohort, Hector and Joe.

As a CIFRE PhD candidate, I also had the pleasure of belonging to a second family, the SANOFI family. They taught me a lot about the corporate life and many topics I was completely ignorant of. I will always remember our nerdiest jokes. It was a strong dose of dark humor that only the strongest can withstand.

I am also indebted to all my friends in particular Ayoub, Olivier, Rafik, Mike and François who were curious and attentive to my research. PhDs are a difficult journey that can become pleasant in the presence of such sympathetic and encouraging friends.

My last and most important acknowledgment goes to my family, my father Aissa, my mother Beya and my sister Lobna. My studies have always been the first priority of my parents. They have made many sacrifices to guarantee

me the best conditions for academic success. I hope that I have made their sacrifices worthwhile. This thesis is dedicated to you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Abstract: *Genome-wide association studies have become an ubiquitous approach to unravel the genetic background of complex diseases. Nonetheless, this background remains largely unexplained. Several hypotheses have already been advanced to explain this missing heritability. One of them is the interaction between distinct loci, or epistasis. Intragenic epistasis and intergenic epistasis are the two major types of epistasis. The detection of both types is subject to several statistical challenges due to linkage disequilibrium, high dimensionality and population structure, among others. To tackle them, we propose a pair of novel approaches. They help bridge the gap with statistical learning frameworks such as causal inference and nonlinear post-selection inference to improve the detection of epistatic interactions. These tools are further applied to comprehensive use cases to bridge another gap, namely the gap with biology. Specifically, we focus on intragenic epistasis in body mass index and its variations, and on intergenic epistasis in multiple sclerosis. Bridging the two gaps provides an end-to-end pipeline for the study of epistasis. This is often a major shortcoming of epistasis studies, which makes the work conducted in this thesis a significant contribution to the field.*

Résumé : *Les études d'association à l'échelle du génome sont devenues une approche essentielle pour démêler le fond génétique des maladies complexes. Néanmoins, ce fond génétique reste largement inexpliqué. Plusieurs hypothèses ont déjà été avancées pour expliquer cette héritabilité manquante. L'un d'eux est l'interaction entre des loci distincts, ou épistasie. L'épistasie intragénique et l'épistase intergénique sont les deux principaux types d'épistasie. La détection des deux types est soumise à plusieurs défis statistiques en raison du déséquilibre de liaison, de la haute dimensionnalité et de la structure de la population, entre autres. Pour y faire face, nous proposons deux nouvelles approches. Ils aident à combler l'écart avec les cadres d'apprentissage statistique tels que l'inférence causale et l'inférence post-sélection nonlinéaire pour améliorer la détection des interactions épistatiques. Ces outils sont en outre appliqués à des cas*

d'utilisation complets pour combler un autre fossé, à savoir le fossé avec la biologie. Plus précisément, nous nous concentrons sur l'épistase intragénique dans l'indice de masse corporelle et ses variations, et sur l'épistase intergénique dans la sclérose en plaques. Combler les deux lacunes fournit un outil de bout en bout pour l'étude de l'épistasie. Il s'agit souvent d'une lacune majeure des études sur l'épistasie, ce qui fait des travaux menés dans cette thèse une contribution significative au domaine.

1.1 Genome-wide association studies

The human genome project (Risch & Merikangas, 1996) was hailed as a turning point for humanity. It was the first effort to successfully construct a reference genome. Nonetheless, other equally important goals such as determining the bases of genetic diseases remained unattainable. The first steps in this direction were made thanks to Genome-Wide Association Studies, or GWAS (Visscher et al., 2012). These studies rely on datasets comprising the genotypes of numerous participants and their phenotypic measurements *e.g.* a disease status or a quantitative trait. The statistical association between all genotyped variants and the phenotype is then evaluated. The main rationale is that the discovery of causal variants will further our understanding of biological questions, and hopefully help develop better therapies (Nelson et al., 2015).

Single-nucleotide polymorphisms (SNPs) are the genetic variants of choice in GWAS. They correspond to the substitution of a single nucleotide, the elementary building block of chromosomes. More precisely, SNPs refer to single-nucleotide variants with a frequency larger than 1%. This threshold is owed to the focus of GWAS on common diseases. Behind this lies the hypothesis that common diseases are caused by a large set of interacting variants with small effect sizes. This hypothesis is commonly known as the common disease-common variant (CD-CV) hypothesis (see Figure 1.1). The other category of single-nucleotide variants – those with a frequency lower than 1% – are referred to as rare variants. They are also the subject of genetic studies, in particular for Mendelian diseases (Pritchard, 2002). Genetic studies can additionally include other types of variants such as copy number variations (CNVs) (Marshall et al., 2016).

SNPs approximately occur at a rate of one in every 300 base pairs (Nelson, 2004). 90% of SNPs are located in non-coding regions. The remaining 10% are located in coding regions and can be split into two categories: synonymous (silent) SNPs and nonsynonymous SNPs. Silent SNPs do not alter the amino acid composition of the protein. On the other hand, nonsynonymous SNPs can alter the composition of the protein product in two different ways. If the coding SNP is missense, a complete protein with a different amino acid composition is obtained. Conversely, nonsense coding SNPs often result in incomplete and nonfunctional proteins.

SNPs located in non-coding regions can have an impact in several ways. For instance, they may influence promoter activity (gene expression), messenger RNA (mRNA), conformation (stability), and translational efficiency (Shastry, 2009).

In GWAS, genotypes are typically encoded as the number of allelic mutations at every measured SNP. For biallelic SNPs, this is equivalent to an encoding in $\{0, 1, 2\}$. The positions of the measured SNPs depend on the genotyping technology. In GWAS, the most common technology are SNP arrays thanks to their low cost and high accuracy. Probe-based arrays can now genotype an individual with a $> 99\%$ accuracy (LaFramboise, 2009) for less than 250 dollars [1]. We give an illustration

[1] list prices for the GeneChip Human Mapping 500K Array (source: Affymetrix documentation)

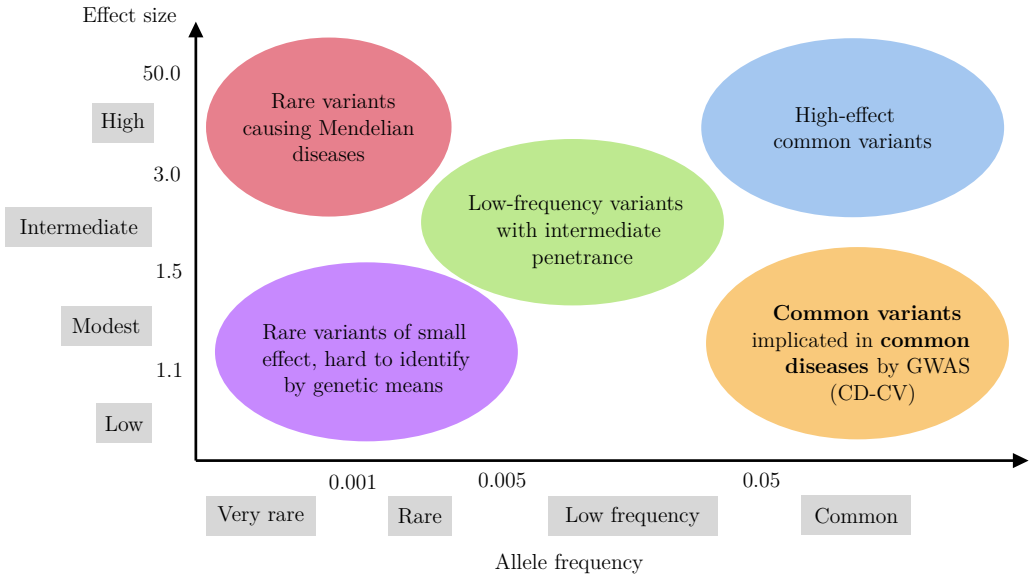


Figure 1.1: Impact of variants by risk allele frequency and effect size. In particular, GWAS focus on common diseases caused by a large set of common variants (bottom-right).

of an Affymetrix SNP array in Figure 1.2. In a standard array, the number of SNPs ranges from $200,000$ to $2,000,000$. The SNP positions are optimized to offer genome-wide coverage and to represent the local linkage disequilibrium (LD) structure. LD corresponds to the non-random association of neighboring alleles (see Section 1.4.4). Thanks to this association, the rest of the genome can then be accurately inferred or imputed.

In GWAS results, significant SNPs are referred to as *lead* or *index* SNPs. Even for "true positives", the lead SNPs are not necessarily causal, but in LD with the true causal variants. This LD relationship is one of many factors that impact the results of GWAS. Other factors include the effect sizes of the causal variants (Zaykin & Zhivotovsky, 2005) and the minor-allele frequencies (MAFs) of both lead and causal variants (Visscher et al., 2017). In all circumstances, the identification of causal variants from lead SNPs must be handled with caution (Schaid et al., 2018).

Despite their inherent difficulties, GWAS have been rather successful at deepening our knowledge of common diseases in the last ten years (Visscher et al., 2017). For instance, GWAS have identified more than one hundred *loci* in type II diabetes (Xue et al., 2018), schizophrenia (Ripke et al., 2014), and outside of the major histocompatibility complex in multiple sclerosis (Oksenberg, 2013). The impact of GWAS goes beyond biological discoveries to support the development of new therapies. Indeed, the odds to reach phase III trials or commercialization are several times larger if the target is backed by genetic evidence (Nelson et al., 2015).

The breadth of conducted GWAS contrasts with the relative simplicity of the implemented statistical methodology. Despite the general awareness within the

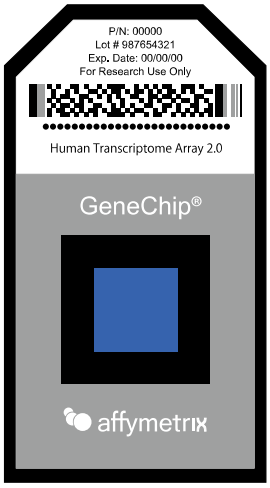


Figure 1.2: Illustration of a GeneChip Human Mapping 500K Array manufactured by Affymetrix. The array interrogates SNPs located on amplicons that range in size from 200 bp to 1,000 bp (Komura et al., 2006).

community of the complex genotype-phenotype relationships, univariate and linear statistical tests of association are still the norm. Their popularity can be explained by their robustness against model misspecification, interpretability and linear complexity in the number of samples. Moreover, the massive leaps of progress in robust statistics and machine learning in recent years have not yet been fully translated to life science disciplines. p-values still remain a universal metric to assess the significance of any reported discovery. Several critics have voiced their concerns against this excessive emphasis on p-values. Ioannidis (2005) used simulations to justify that most research claims are likely to be false. Because of this, the confidence in any reported GWAS association is more and more contingent on its replication across several datasets (Kraft et al., 2009).

Many methodological contributions have improved on the standard techniques in GWAS. Cantor et al. (2010) provides an exhaustive review of recent methods. One approach that is gaining in popularity is meta-analysis and consists in combining the results of several GWAS datasets, even when the original genotypes are unavailable. A second approach is hypothesis-driven GWAS which incorporates prior biological information to narrow the scope around relevant pathways and networks (Kitsios & Zintzaras, 2009). The use of biological information can also be useful a posteriori by mapping the results on pathways, in combination with graph computational tools and pathway databases. Moreover, the a posteriori use can facilitate interpretability and identification of causal SNPs. A third axis of improvement is the modeling of interactions between distinct *loci*, or epistasis (see Section 1.3), to get closer to the underlying biology, and the recovery of missing heritability.

1.2 Missing heritability and epistasis

The GWAS catalog (MacArthur et al., 2016) references more than 11,912 strong associations sourced from 1,751 curated publications (Welter et al., 2013). Despite the scale of such an output, GWAS are frequently criticized for their inability to fully explain the heritability of common diseases and traits. Recovering the full genetic architecture remains a key prerequisite to understanding disease etiology and developing efficient treatments tackling the origins of diseases, and not just their symptoms.

Heritability can be intuitively understood as the genetic contribution to the phenotype (Zuk et al., 2012). This type of heritability is referred to by geneticists as broad-sense heritability, and can be quantified as the proportion of total phenotypic variance that is explained by the genotype:

$$H^2 = \frac{\mathrm{Var}(Y) - \mathrm{Var}(Y|X)}{\mathrm{Var}(Y)},$$

where X is a diploid genotype, Y the phenotype, and $\mathrm{Var}(Y|X)$ the phenotypic variance between genetically-identical individuals.

Broad-sense heritability H^2 constitutes an upper-bound to predictors' capacity to predict phenotype from genotype. On the other hand, narrow-sense heritability h^2 measures the *additive* contribution of a subset of SNPs $P \subset X$ to phenotypic variance. Under linkage equilibrium (independence between SNPs), h^2 is the variance of Y explained by P under a linear regression model:

$$h^2 = 2 \sum_{X_i \in P} \beta_{X_i}^2 f_{X_i}(1 - f_{X_i}),$$

where f_{X_i} is the minor allele frequency (MAF) of SNP X_i and β_{X_i} its corresponding effect size.

To estimate *additive* missing heritability, it would be natural to compare h^2 to h^2_{all}, the *additive* phenotypic variance of all SNPs that affect the response Y. However, the SNPs that affect Y are not exhaustively identified. For this reason, h^2_{all} is approximated thanks to twin studies:

$$h^2_{\text{all}} \approx 2(r_{MZ} - r_{DZ}), \tag{1.1}$$

where r_{MZ} and r_{DZ} are respectively the phenotypic correlations within monozygotic twins and within dizygotic twins. We can finally derive an estimate of *additive* missing heritability in the following way:

$$\pi_{\text{missing}} = 1 - \frac{h^2}{h^2_{all}} \tag{1.2}$$

If the SNPs in P fully explain Y in additive fashion, then $\pi_{\text{missing}} = 0$. This is far from being the typical result in GWAS. Moreover, the estimation of missing heritability in Eq. (1.2) relies on the approximation of narrow-sense heritability in

Eq. (1.1). This approximation makes the underlying assumption that no epistatic interactions are involved, which is inconsistent with the observed biology (Zuk et al., 2012).

It is worth noting that epistasis is not the only hypothesis behind missing heritability. Rare variants, which are often either excluded or poorly detected also contribute. Other types of variants such as copy number variants (CNVs, insertions and deletions) and copy neutral variants (inversions and translocations) are another factor behind missing heritability. This is in addition to a lack of statistical power because of the small sample sizes (Spencer et al., 2009). The last, but not least important factor is the environment through epigenetics and shared environment among relatives (Manolio et al., 2009).

Table 1.1: Estimation of missing heritability for several complex diseases

Disease	Number of loci	Proportion of heritability explained	Heritability measure
Age-related macular degeneration	5	50%	Sibling recurrence risk
Crohn's disease	32	20%	Genetic risk (liability)
Systemic lupus erythematosus	6	15%	Sibling recurrence risk
Type 2 diabetes	18	6%	Sibling recurrence risk
HDL cholesterol	7	5.2%	Residual phenotypic variance
Height	40	5%	Phenotypic variance
Early onset myocardial infarction	9	2.8%	Phenotypic variance
Fasting glucose	4	1.5%	Phenotypic variance

For Crohn's disease, the proportion of explained heritability stands at 20% with 71 identified *loci* (Franke et al., 2010). Zuk et al. (2012) explains that, if interactions among three pathways were included in the estimation of heritability, the explained proportion can be increased to 84%. In schizophrenia, Zuk et al. (2012) completely managed to eliminate missing heritability. The last two examples stress the importance of epistasis modeling in chasing missing heritability, which can be large depending on the disease. In Table 1.1, we give an estimate of explained heritability for several complex diseases. The listed statistics are reproduced from Manolio et al. (2009), and have most likely increased, though moderately (Nolte et al., 2017).

1.3 Understanding the biology of epistasis

Epistasis is considered a prevalent phenomenon that is central to the structure and function of biological pathways (Phillips, 2008). Yet, there is a fair amount of confusion pertaining to its definition, and several reviews have been dedicated to this topic (Phillips, 2008; Cordell, 2002; Örjan Carlborg & Haley, 2004). The major distinction to be made is between *biological epistasis* and *statistical epistasis*. In this section, we review the mechanisms that define biological epistasis. In Section 1.4.2, we characterize epistasis from a statistical perspective.

Epistasis occurs when the phenotypic impact of a genetic variant depends on other variants. For example, the dependency can consist in completely offsetting its impact or modulating its amplitude *e.g.* increasing or decreasing disease propensity. The interacting variants can be located on either distinct genes (intergenic epistasis) or the same genes (intragenic epistasis). The latter form of epistasis is often overlooked despite its importance. For example, (Poon & Chao, 2005) estimates that compensatory mutations in the φX174 bacteriophage are equally split between intergenic and intragenic. Epistatic interactions within non-coding regions exist, too. In particular, epistatic interactions in cis-regulatory regions have recently drawn significant attention (Fish et al., 2016; Lagator et al., 2015, 2017).

1.3.1 Intragenic epistasis

Genetic variants within a gene can have minor individual effects, but their combination can result in a significant impact on protein activity (Bershtein et al., 2006). Intragenic epistatic interactions can additionally impact protein stability. Witt (2008) demonstrates that, in a disulfide bridge, the co-presence of two cysteine aminoacids creates a chemical bond that enhances the stability of the protein. Besides structural and functional properties, intragenic epistasis influences selection. It helps preserve protein function despite continual changes in protein sequence (Weinreich, 2006).

Interestingly, most intragenic interactions are negative. The purpose of the synergistic interaction is to compensate for the change in protein sequence in order to preserve the integrity of the protein. Gonzalez & Ostermeier (2019) studied over $8,000$ mutation pairs in TEM-1 β-Lactamase, and found that negative epistasis occurred 7.6 times as frequently as positive epistasis. Another work from Bank et al. (2014) came to similar conclusions by studying more than $1,000$ pairs in the Hsp90 region in yeast.

Intragenic epistasis encompasses several mechanisms of action. A first type is stability threshold, where both mutations are required to trigger an effect. A second type of mechanism corresponds to suppressor mutations, which neutralize/mask the negative stability effects of other variants. We also mention conformational epistasis: a conformation change due to one mutation is needed so that the beneficial functional effect of another mutation materializes. For a more exhaustive review, we refer the reader to Lehner (2011).

1.3.2 Intergenic epistasis

Intergenic epistasis is the best known and most pervasive form of epistasis. The simplest scenario to consider is the affinity of physical interaction between two proteins. The interaction is deemed epistatic if the affinity depends on the protein SNPs in a non-additive fashion. As in intragenic epistasis, we can also witness neutralizing mechanisms, where the deleterious effect of one SNP on a first protein is conditional on a second SNP located on another protein. Such mechanisms can take place when the second SNP modulates the contact interface with the first protein.

The detrimental effect of a few epistatic pairs is already established in the literature. For instance, Combarros et al. (2009) found 27 gene-gene interactions that were significantly associated with Alzheimer's disease. In systemic lupus erythematosus (SLE), Hughes et al. (2012) provided evidence for 4 epistatic interactions, among which three include SNPs in the human leukocyte antigen (HLA) region. The latter has already been shown to have a deleterious effect in several auto-immune diseases (Simmonds & Gough, 2007). Other diseases with validated epsitatic synergies include tuberculosis (Daya et al., 2015), Crohn's disease (McGovern et al., 2009) and bipolar disorder (Judy, 2013).

Intergenic epistasis can manifest itself in several forms (Lehner, 2011). We note again compensatory mechanisms where two proteins perform the same function, and are a substitute to each other. Another form are sequential interactions along a linear pathway to produce a metabolite. Feedback and cooperation regulatory mechanisms are other forms of intergenic epistasis. A last example of recurrent intergenic epistasis is the non-additive effect of a pair of SNPs which together regulate a physical or chemical property. The complexity of the above interactions demonstrate the difficulty of epistasis detection directly from biology. Hence, the need for powerful statistical tools.

1.4 Challenges of statistical epistasis

The first characterization of epistasis from a statistical perspective dates back to Fisher (1919) who initially coined a similar term "epistacy". It has been gradually substituted with "epistasis" which has resulted in a great deal of confusion among geneticists. Originally, epistasis (Bateson & Mendel, 1909) referred to the blocking effect of some SNPs which occlude the phenotypic effects of other SNPs. On the other hand, Fisher (1919) used epistacy to describe departure from additivity of effects in a quantitative phenotype. Departure from additivity covers Bateson's original definition, and is still the common definition of epistasis.

Aside from the epistemological questions of the definition of epistasis, translating the results of statistical epistasis into plausible scenarios for biological epistasis is the key objective here. It remains a bottleneck because of various challenges that we detail in this section.

1.4.1 Relation to biological epistasis

In type I and type II diabetes, the study of statistical epistasis has successfully led to the discovery of biological interactions (Cordell & Todd, 1995; Cordell et al., 1995; Cox et al., 1999). Nonetheless, it failed in other cases to generate valid interactions (Cordell et al., 2001). Cordell et al. (2001) argues that our capacity to infer biological epistasis from statistical epistasis is limited. This raises the question of the control of false positives in the results of statistical epistasis. For this reason, a more complete picture combining genetic, proteomic and metabolic information is needed (Kim et al., 2016). In addition to detecting biological epistasis, determining the exact type of interaction (see Section 1.3) can also benefit from more information.

1.4.2 The definition of interaction

Statistical interaction was originally defined as the departure from an additive genotype-phenotype model (Fisher, 1919). The easiest way to test this hypothesis are linear models endowed with an interaction term. For a dichotomous phenotype Y and a pair of SNPs (X_1, X_2), we may consider the following logistic regression model:

$$\begin{aligned}\text{logit}\left(P(Y=1|X_1,X_2)\right) = &\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \\ &\alpha_{12} X_1 X_2.\end{aligned} \tag{1.3}$$

If the logistic model in Equation (1.3) is the true model, absence of interaction can be characterized by $\alpha_{12} = 0$. However, in fitted models, drawing similar conclusions directly from the estimated coefficient $\widehat{\alpha}_{12}$ is mistaken, and hypothesis testing is needed to conclude about the true amplitude of the interaction term. Likelihood ratio tests (King, 1998) which compare the goodness-of-fit of two models can be useful in this regard. The two models compared for epistasis are a null model with main effects only ($\alpha_{12} = 0$) and a saturated model with both main and interaction effects' terms.

So far, we have not specified the encoding of the two SNPs (X_1, X_2). The usual encoding is $\{0, 1, 2\}$, which indicates the number of minor alleles in bi-allelic SNPs. However, if we consider binarized SNPs with values in $\{0, 1\}$, interesting equivalences with odd ratios can be easily shown. In fact, SNP binarization can encode for either recessive or dominant mechanisms depending on the binarization rule. Extending the equivalences to bi-allelic SNPS is possible, yet more difficult (VanderWeele & Knol, 2014). First, we define a risk ratio as:

$$R_{ij} = P(Y = 1|X_1 = i, X_2 = j). \tag{1.4}$$

Absence of interaction can be defined as the non-dependency of relative risk ratios w.r.t one SNP on the other SNP. Mathematically speaking, we have,

$$\begin{cases} \frac{R_{11}}{R_{01}} = \frac{R_{10}}{R_{00}}, \\ \frac{R_{11}}{R_{10}} = \frac{R_{01}}{R_{00}}. \end{cases} \tag{1.5}$$

It is straightforward to show that the two conditions in Eq. (1.5) are equivalent. Additionally, we can rewrite them to define absence of statistical interaction in terms of multiplicativity of risk ratios:

$$\frac{R_{11}}{R_{00}} = \frac{R_{10}}{R_{00}} \cdot \frac{R_{01}}{R_{00}}. \tag{1.6}$$

Another common and related way to define statistical interaction are odd ratios, which we define as follows for a reference genotype $(X_1, X_2) = (0, 0)$:

$$OR_{ij} = \frac{R_{ij}/(1 - R_{ij})}{R_{00}/(1 - R_{00})}. \tag{1.7}$$

Similarly to risk ratios, absence of interaction corresponds to multiplicativity of odd ratios:

$$OR_{11} = OR_{10} \cdot OR_{01}. \tag{1.8}$$

Risk and odd ratios are numerically close when the event $\{Y = 1\}$ is rare for all genotypes i.e. $1 - R_{ij} \approx 1$ for all i, j. In this case, the definitions of statistical interaction in Eq. (1.6) and Eq. (1.8) are equivalent.

We used binarized SNPs in this section to demonstrate the interesting mapping between the coefficients of the logit model in Eq. (1.3) and odd ratios. In fact, we always have the following:

$$\begin{cases} \exp(\alpha_0) = R_{00}/(1 - R_{00}), \\ \exp(\alpha_1) = OR_{10}, \\ \exp(\alpha_2) = OR_{01}, \\ \exp(\alpha_{12}) = OR_{11}/(OR_{10} \cdot OR_{01}). \end{cases} \tag{1.9}$$

We can then deduce the equivalence of the two interaction conditions:

$$\alpha_{12} = 0 \Leftrightarrow OR_{11} = OR_{10} \cdot OR_{01}. \tag{1.10}$$

The equivalence in Eq. (1.10) defines interaction on a multiplicative scale. The literature (VanderWeele & Knol, 2014) cites an additive scale given by Eq.(1.11) as well. The two scales are not equivalent.

$$R_{11} - R_{01} - R_{10} + R_{00} = 0 \tag{1.11}$$

Other stronger formulations of statistical interaction in terms of conditional independence and mutual information have also been proposed by statisticians (Whittaker, 2009; Dobrushin, 1959). The multiplicity, intricacy and lack of equivalences between the different formulations of statistical interaction prove the difficulty of constructing a single framework for statistical interaction.

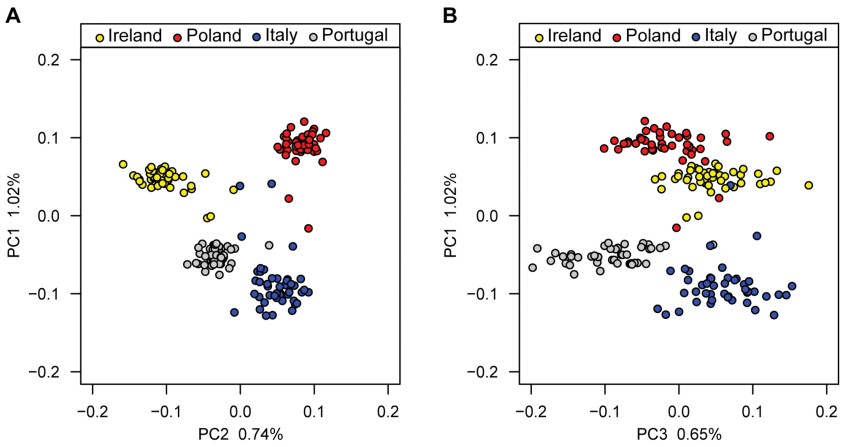


Figure 1.3: Example of PCA results showing how GWAS participants can cluster by country of origin. PC1 is related to the position along the north-south axis, while PC2 to the position along the east-west axis. The figure is sourced from Candille et al. (2012) under a Creative Commons Attribution 2.5 Generic license.

1.4.3 Population structure

Population structure consists in the presence of different subpopulations within the GWAS cohort. It can be formalized as the discrepancy in MAFs owed to the unequal representation of the different subpopulations between cases and controls. The main reason for it is genetic drift (Masel, 2011) which drives variation in MAFs across several generations. A common metric in GWAS to detect population structure is the genomic inflation factor (GIF). It compares the empirical median of the Armitage's trend test statistics for a number of unlinked *loci* to the median of the χ^2 distribution with one degree of freedom (Devlin & Roeder, 1999). Under the null hypothesis of no population structure, the Armitage test statistics asymptotically follow a χ^2 distribution. From a practical standpoint, a GIF value larger than 1.05 indicates presence of population structure.

The classical procedure in GWAS to avoid spurious associations due to population stratification is through principal component analysis (PCA) (Price et al., 2006). The PC analysis not only makes it possible to detect population structure, but also to correct for it by including the top components as covariates in a regression model. However, not all statistical methods can accommodate PC-like correction.

In comparison to univariate GWAS, the problem of population structure in epistasis is more severe. The lower signal-to-noise ratio can result in a higher rate of false discoveries. However, most epistasis detection algorithms do not correct for population structure. Furthermore, the GWAS focusing on epistasis seldom account for it (Wei et al., 2014). Combarros et al. (2009) reviewed more than 100 publications studying Alzheimer's disease, and pointed out the lack of adjustment for population structure among other confounding factors.

1.4.4 Linkage disequilibrium

The non-random association of alleles along chromosomes in a general population is called linkage disequilibrium. Because of it, nearby SNPs are strongly correlated, and this correlation can span hundreds of thousands of base-pairs (bp). The standard way to measure LD between two SNPs is through their squared correlation coefficient, which is usually denoted by r^2.

LD is a double-edged sword. The lead SNPs in GWAS results are often in strong LD with the true causal SNPs. Their identification is possible by fine-mapping the surrounding regions of lead SNPs (Schaid et al., 2018). However, the complex patterns of LD and the large genomic windows it spans can make the task of fine-mapping daunting. Additionally, the presence in the array of SNPs in strong LD with the causal ones is uncertain. Hopefully, with the development of whole-genome sequencing (WGS), the problem of coverage will subside. Yet, with the increased number of SNPs in WGS, other statistical problems are to arise because of the higher LD.

Wei et al. (2014) provide a mathematical explanation to the influence of LD in the univariate setting on statistical power. In the additive case, the explained variance between the measured SNP and the phenotype is a linear function of r^2 and the variance between the causal SNP and the phenotype. In the bivariate additive case, the relationship becomes r^4. In bivariate dominance settings, it even increases to r^8. Under all circumstances, the explained variance in the bivariate case is lower ($r^8 < r^4 < r^2 < 1$), which makes the identification of the causal variants more difficult.

1.4.5 High dimensionality

High dimensionality is one of the major problems in computational biology, and in particular in GWAS. It is often the case that the number of covariates is several times larger than the number of samples. In commercial arrays, the number of SNPs ranges between $200,000$ and $2,000,000$ (Visscher et al., 2017). By contrast, the Wellcome Trust Case Control Consortium (WTCCC) dataset comprises $14,000$ cases for 7 common diseases and $3,000$ controls (Burton et al., 2007). It was launched in 2007, but still remains a gold standard in common diseases. The WTCCC used the Affymetrix 500K with $500,000$ measured SNPs. On average, the SNP-to-sample ratio in a WTCCC case-control dataset is 100. In the machine learning community, the problems created by such large ratios are referred to as the "curse of dimensionality". Despite the rich representations provided by the large number of covariates, the generalization capacity of fitted models is hampered by problems of estimation instability, model overfitting and local convergence (Clarke et al., 2008). Further assumptions *e.g.* sparsity are often added to ensure a better generalization performance (Johnstone & Titterington, 2009). Nonetheless, even with additional assumptions, the ultra-high dimensionality of GWAS datasets sets a limit to their capacity to detect relevant associations.

Direct prediction of phenotype from genotype and biomarker selection are the two main missions of GWAS. In comparison to phenotypic prediction, biomarker selection as a model selection task is more difficult. It is complicated by high-dimensionality and strong correlations between neighboring SNPs, or LD. In epistasis, the problem of high-dimensionality is more acute. For p SNPs, there are $p(p-1)/2$ unique pairs to select from, with high correlation between the pairs since a given SNP is present in $p-1$ pairs.

The problem of high-dimensionality in GWAS is not only statistical, but also computational because of memory requirements and execution time. If one-hot encoding is used for the SNPs, circa 3Gb are needed just to store the WTCCC dataset in RAM memory. If the usual integer encoding $\{0, 1, 2\}$ is used instead, the memory requirements are multiplied by a factor of 10 for 32-bit integers. On top of this, additional memory may be needed for analyzing the dataset.

Beside the computational limitations, several geneticists argue that the problem of statistical power can be overcome with the genotyping of more samples thanks to the rapid decrease in sequencing cost. However, even in a country with a population of 10 million, genotyping all cases for a disease with an incidence rate of 2.0% is not sufficient to reach the setting of $n = p$ for a SNP array with $500,000$ SNPs. A threshold of 2.0% surpasses the prevalence of multiple sclerosis (MS), Crohn's disease and rheumatoid arthritis (RA). Furthermore, constructing a GWAS cohort is a tedious task in practice because of logistics, diagnostics and participants' consent for data sharing.

1.4.6 Nonlinearity

The difficulty of modeling nonlinear effects is another limitation of current approaches in GWAS. For example, the modeling of dominance effects is not directly possible in linear models. Richer classes of models are therefore needed. Nonetheless, linear models and other derivatives remain an attractive and ubiquitous option thanks to their robustness and interpretability. For epistasis detection, a product term between a pair of SNPs can be included to model statistical interaction (Wan et al., 2010). As for linear models in the univariate setting, one can also question the pertinence of such a modeling for statistical epistasis.

To better improve the modeling of nonlinearities, we can include higher-order interactions (tripartite interactions and higher). Indeed, biological interactions can involve more than two entities. The trade-off here is a dramatic increase in complexity and loss of statistical power, which can make them impossible to implement.

Additivity of effects, the original definition proposed by Fisher (1919) for absence of interaction can be easily extended to the nonlinear case. For a continuous phenotype y and two SNPs x_1 and x_2, it can be defined as the existence of two functions $f, g \in \mathbb{R}^{\mathcal{X}}$ such that $y = f(x_1) + g(x_2)$. The definition is intuitive, but fitting the two functions f, g is only possible through additional assumptions (Lim & Hastie, 2015). Moreover, the added assumptions can limit the capacity of the models to detect epistatic interactions.

In Section 1.3, we highlighted several epistatic mechanisms. Each one of them would potentially require a different modeling. However, our knowledge of biological networks and the types of interactions within is still limited. Even if the type of interaction was fully understood, translating it into an adequate nonlinear model is not straightforward because of the mismatch between biological epistasis and statistical epistasis (see Section 1.4.3).

1.4.7 Hypothesis testing

As repeatedly stated, SNP-wise hypothesis testing is the classical strategy in GWAS. More precisely, a chi-squared test to assess odd ratios' significance is used in case-control studies. On the other hand, likelihood-ratio tests and Wald tests are used for continuous traits (Purcell et al., 2007). The output of the tests is a single p-value for each SNP. The computation of genome-wide p-values is followed by their visualization on a Manhattan plot. We provide an illustration of a Manhattan plot in Figure 1.4. The horizontal axis corresponds to genomic coordinates and the vertical axis to p-values. Manhattan plots provide a concise and exhaustive way to appraise the results of a GWA study. Additionally, they indirectly help control for false positives thanks to LD. Neighboring SNPs tend to have similar p-values. Therefore, significant SNPs are usually located near to each other, because all of them are in strong LD with the true causal SNP. An isolated significant p-value can simply be a statistical outlier.

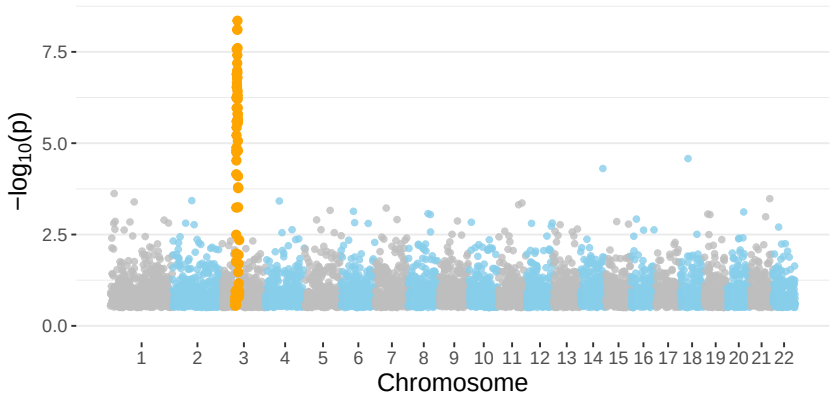


Figure 1.4: Illustration of a Manhattan plot with one significant *locus*.

A statistician would immediately recognize here the setting of multiple hypothesis testing. Most methods correct for it through either family-wise error rate (FWER) or family-discovery rate (FDR) control. The Bonferroni correction is a standard FWER procedure. It is statistically valid, but is more conservative in comparison to FDR procedures such as the Benjamini-Hochberg (BH) procedure. This can lead to a significant loss of statistical power, especially in epistasis. Nevertheless, the application of more recent and less stringent procedures is still uncommon in

GWAS. Aside from this issue, Boyle et al. (2017) propose that regulatory networks are sufficiently interconnected that all expressed genes in these networks impact the core disease pathway. Additionally, the authors attribute most heritability to genes outside of the core pathway. This proposal implies that most genes – and associated SNPs – are "causal". This challenges the very relevance of hypothesis testing, since most expressed genes are positive by indirectly impacting the phenotype.

A major source of criticism toward p-values stems from their manipulation. For example, summary statistics such as p-values must be aggregated in a valid way (Heard & Rubin-Delanchy, 2018). However, this key aspect is often overlooked despite its recurrence. The most important problem with p-values remains the general misunderstanding of them. We mention for example the arbitrary application of 0.05 threshold and the misconception that a p-value is "the probability that the studied hypothesis is true". In light of this, the American statistical association recently issued a lengthy statement (Wasserstein & Lazar, 2016) to clarify a number of misconceptions.

1.5 Bridging the gap with statistics

In the previous sections, we extensively reviewed the challenges of epistasis detection. Armed with this information, we develop in this thesis a pair of novel approaches addressing a number of them. A second and equally-important contribution of this thesis is the extension of a range of statistical frameworks to GWAS.

The first tool we propose is epiGWAS (Slim et al., 2018). To the best of our knowledge, it is the first tool to apply causal inference (Pearl, 2009) to epistasis. Here, we infer the interactions between a predetermined SNP and the rest of the genome. This makes epiGWAS appropriate for the detection of intergenic epistasis. It incorporates several ideas to improve robustness and statistical power. More generally, epiGWAS can be applied to other interaction problems such as clinical trials and social/economic studies.

We also propose a second tool called kernelPSI (Slim et al., 2019). As its name suggests, we use kernel methods (Hofmann et al., 2008) to generalize post-selection inference (PSI) (Lee et al., 2016) to the nonlinear setting. We believe that kernelPSI is the first work not only to develop a general and flexible framework for nonlinear PSI, but also to to jointly apply PSI and kernel methods to GWAS. By contrast with epiGWAS, the main purpose of kernelPSI is the study of intragenic epistasis. Within a given gene, we select blocks of putative SNPs and test their joint association with the phenotype.

To spur the adoption of epiGWAS and kernelPSI by the GWAS community, both tools are provided as R packages downloadable from the CRAN repository. Open-source and user-friendly software can only narrow the gap between GWAS and statistical learning, and hopefully bridge it in the future. During the course of this thesis, it has become obvious to us that bridging this gap is necessary to move the fields of GWAS and epistasis forward.

1.5.1 epiGWAS

Causal inference has swiftly become one of the trendiest topics in machine learning (ML). In particular, extensive research efforts are being dedicated to the investigation of the connections between causal inference and reinforcement learning (Peters et al., 2017). It seeks to determine the response effects of an intervention on the covariates. Rubin (2005) developed the framework of potential outcomes to estimate these effects. A class of methods within this framework rely on propensity scores. The scores were developed for (nonrandomized) observational studies in which they correspond to the probability of treatment assignment conditionally on a set of observed baseline covariates. They reduce the effects of confounding, and the covariates' distributions in cases and controls are similar conditionally on them (Austin, 2011).

Nonrandomized clinical trials study the interactions between a treatment and a set of clinical covariates. They were the main inspiration for epiGWAS, where we analogously study the interactions between a predetermined SNP target and the rest of the genome. The SNP targets can be drawn from the literature, univariate GWAS or results of in *vitro* experiments. Narrowing the scope around such *loci* provides increased statistical power and better interpretability.

In genomic data, propensity scores model the LD structure between the target and the rest of the genome. We include them in several penalized regression models to detect epistatic effects. The key difference between the different models is the normalization of the propensity scores. The goal of the normalization is to correct for the estimation errors of the scores. The latter are estimated using the fastPHASE model which consists of a hidden Markov model (HMM) (Scheet & Stephens, 2006) representation of the chromosomes. The theoretical underpinnings of epiGWAS are detailed in Chapter 2.

EpiGWAS tackles some of the challenges of statistical epistasis highlighted in Section 1.4. It correctly models LD in order to focus on synergistic effects, uses penalized regression and stability selection for high-dimensional feature selection, and finally, completely forgoes hypothesis testing in favor of a more robust scoring procedure.

1.5.2 kernelPSI

The motivation behind kernelPSI is the complete dichotomy between SNP-based and gene-based approaches. Both categories are certainly relevant, yet they answer distinct biological questions. In Section 1.1, we listed SNP effects on protein properties such as expression and stability. On the other hand, interpretation at the gene level offers a functional perspective by analyzing the involved pathways and mechanisms of action. Because of intragenic epistasis (see Section 1.3.1), inference at the gene level is sensible, too: the deleterious effect of one gene can depend on the co-occurrence of multiple mutations.

If all SNPs mapping to a particular gene are used for inference at the gene-level,

statistical performance can suffer. The vast majority of these SNPs are unrelated to the disease, and can even bias the results because of LD and discrepancy in MAFs among other factors. To address this drawback, the most associated SNPs can be selected in a first step, before subsequently testing their joint effect on disease propensity. Mathematically speaking, this can be described as feature selection followed by statistical inference. If the same samples are used in both steps, inference becomes biased. Up until recently, statisticians therefore used different samples for each step for valid inference. This may result in a lack of precision in selection and a lack of statistical power in inference particularly in low sample size settings. By correctly taking into account the feature selection event, PSI allows to use all samples in both steps. The first significant development in this area is owed to Lee et al. (2016) who modeled feature selection in LASSO as a set of linear constraints in the outcome y. In hypothesis testing, the authors tested for the significance of the coefficients in the support. This was achieved by determining the distribution of the test statistics conditionally on the selection constraints. The techniques developed in their work inspired others such Tibshirani et al. (2016), Reid et al. (2017) and Heller et al. (2018).

All of the above contributions are limited to linear models. Going beyond the linear case in genomics is particularly appealing (see Section 1.4.6). In statistical learning, practitioners often resort to kernel methods to model nonlinearity. Classical algorithms e.g. ridge regression, principal component analysis and support vector machines have been adapted for kernels. Put simplistically, kernels can be considered as "generalized dot products". This is achieved by mapping the original features to a reproducing kernel Hilbert space (RKHS), which offers richer descriptions and allows the modeling of nonlinear associations. For the user, the association/similarity between two samples can still be measured using the original features without access to the RKHS. Moreover, the computations with kernels remain linear despite the added complexity. This key aspect probably best explains their success in computational biology (Schölkopf et al., 2004). One of the kernel metrics that allow to measure nonlinear associations between two groups of features is the Hilbert Schmidt Independence Criterion (HSIC). It was originally proposed by Gretton et al. (2005a) who defined HSIC as the squared norm of the cross-covariance operator.

In Chapter 3, we show how HSIC is an example of what we called quadratic kernel association scores. They are quadratic forms of the response y. We use them for nonlinear feature selection through the selection of the corresponding kernels. In subsequent inference, we correctly measure the effect of the selected kernels on the outcome by modeling the selection event as a conjunction of quadratic constraints. Our approach outperformed competing methods relying on either linear PSI or non-selective kernel association scores.

1.6 Bridging the gap with biology

The contributions we detailed in Section 1.5 help bridge the gap with statistical learning. However, the key contributions in GWAS and in genetics are made by providing new insights into the etiology of diseases. This is the reason why large efforts in this thesis have been made to bridge the gap with biology. Simulations and statistical performance measures are essential for validation and benchmarking. Complementing them with new biological discoveries and further interpretation make these tools more valuable. In the case of epiGWAS, we developed a gene-level extension to perform a systemic study of epistasis in Multiple Sclerosis. As for kernelPSI, we studied body mass index (BMI) and its variations ΔBMI to validate the hypothesis of different genetic mechanisms governing the two phenotypes. The MS and BMI studies are respectively detailed in Chapters 4 and 5.

1.6.1 Multiple sclerosis

Multiple sclerosis (MS) is an autoimmune disease that targets the central nervous system (CNS). It can severely hamper the lives of affected people by limiting their movement and their vision. Despite all efforts, its origins are still unknown. We nonetheless have gained valuable knowledge thanks to several GWAS (Baranzini & Oksenberg, 2017). A natural follow-up step would be to study epistasis in MS. Interestingly, the literature already references at least three cases of biological epistasis in MS (Galarza-Muñoz et al., 2017; Harty et al., 2019; Lincoln et al., 2009).

A thorough investigation of all pairwise interactions in a GWA study with epiGWAS is impossible. Therefore, we focused on the interactions between the genes within 15 MS disease maps from the MetaCore pathway database (Ekins et al., 2006). In this study, we developed an extension of epiGWAS at the gene level. It consists in a rank-based aggregation of SNP-SNP scores to derive gene-gene scores. Our study yielded 4 gene pairs involving missense variants and 117 gene pairs with epistasis mediated by eQTLs.

Some of the obtained pairs are already known to be involved in MS. More specifically, GLI-I and SUFU are in direct binding interaction in oligodendrocyteprecursor cell differentiation, and NF-κB regulates the transcription of IP-10. Retrieving such interactions validates the capacity of epiGWAS to reveal novel epistatic interactions in complex disease maps.

1.6.2 Variations of BMI

Some recent studies suggest that BMI and ΔBMI might be influenced by distinct sets of SNPs. This hypothesis can help explain why certain individuals gain weight at a rapid pace even after drastic weight loss (Fothergill et al., 2016).

To study this hypothesis, we used the UK BioBank (Bycroft et al., 2018), which is a large biobank of $500,000$ British individuals with thousands of phenotypes. Similarly to the MS study, the method we developed was not directly applicable in

practice. As a result, we introduced a number of modifications to make kernelPSI scalable for large-sample GWAS. The additions include the use of specific kernels which include MAFs to measure genotypic similarity, mapping the kernels to contiguous LD blocks and transferring some of the computations to graphics processing units (GPUs). We applied kernelPSI to all genes associated to BMI in the GWAS catalog (MacArthur et al., 2016). The pipeline implemented in this study can be transposed to other GWAS with little modification.

Our study demonstrated a weak association between BMI and ΔBMI, in addition to providing a number of putative *loci* for both of them. We also included in this study other gene-level baselines that were outperformed by kernelPSI.

1.7 Contributions

This thesis was simultaneously pursued at the Centre for Computational Biology (CBIO) at Mines ParisTech, and the Bioinformatics group at SANOFI R&D. The collaboration benefited from academic supervision and methodological input at CBIO, and from biological knowledge, application-driven suggestions and logistical support at SANOFI. This led to a comprehensive study of epistasis that covers both statistical and biological aspects.

The work described in this thesis has resulted in a number of publications and preprints, in addition to open-source software. Chapter 2 explains the theoretical framework of epiGWAS. This work has already been submitted to PLOS ONE, and is currently undergoing major revisions (Slim et al., 2018). We also published an eponymous R package directly available from CRAN that facilitates its use by practitioners. The second tool, kernelPSI, is explained in Chapter 3. It was published in the Proceedings of the 36th International Conference on Machine Learning (Slim et al., 2019), and an accompanying R package is also available from CRAN. The multiple sclerosis and body mass index use cases are respectively detailed in Chapter 4 and Chapter 5. Both corresponding manuscripts are still in preparation. Finally, the GPU variant of kernelPSI that we specifically developed for the body mass index study is downloadable from GitHub [2].

[2]The code source of the GPU implementation is on the 'development' branch of the GitHub repository: `https://github.com/EpiSlim/kernelPSI.git`. It automatically detects the supported GPU Nvidia architectures

Chapter 2

EpiGWAS: Novel Methods for Epistasis Detection in Genome-Wide Association Studies

---

Publication and Dissemination: *The work in this chapter has already been submitted to PLOS ONE (Slim et al., 2018), and is currently undergoing major revisions. It was also presented as a poster at the ICML 2019 Workshop on Computational Biology.*

Abstract: *More and more genome-wide association studies are being designed to uncover the full genetic basis of common diseases. Nonetheless, the resulting loci are often insufficient to fully recover the observed heritability. Epistasis, or gene-gene interaction, is one of many hypotheses put forward to explain this missing heritability. In this chapter, we propose epiGWAS, a new approach for epistasis detection that identifies interactions between a target SNP and the rest of the genome. This contrasts with the classical strategy of epistasis detection through exhaustive pairwise SNP testing. We draw inspiration from causal inference in randomized clinical trials, which allows us to take into account linkage disequilibrium. EpiGWAS encompasses several methods, which we compare to state-of-the-art techniques for epistasis detection on simulated and real data, and demonstrate its benefits to identify pairwise interactions.*

Résumé : *De plus en plus d'études d'associations à l'échelle du génome sont conçues pour découvrir la base génétique complète des maladies courantes. Néanmoins, les loci résultants sont souvent insuffisants pour récupérer complètement l'héritabilité observée. L'épistasie, ou interaction gène-gène, est l'une des nombreuses hypothèses avancées pour expliquer cette héritabilité manquante. Dans ce chapitre, nous proposons epiGWAS, une nouvelle approche pour la détection d'épistasie qui identifie les interactions entre un SNP cible et le reste du génome. Cela contraste avec la stratégie*

classique de détection d'épistase grâce à un test statistique par paires de SNPs exhaustif. Nous nous inspirons de l'inférence causale dans les essais cliniques randomisés, ce qui nous permet de prendre en compte le déséquilibre de liaison. EpiGWAS englobe plusieurs méthodes, que nous comparons aux techniques de pointe pour la détection d'épistasie sur des données simulées et réelles, et démontrons ses avantages pour identifier les interactions par paires.

2.1 Introduction

Decrease in sequencing cost has widened the scope of genome-wide association studies (GWAS). Larger cohorts are now built for an ever growing number of diseases. In common ones, the disease risk is dependent on a large number of genes connected through complex interaction networks. The classical approach and still widespread methodology in GWAS is to implement univariate association tests between each single nucleotide polymorphism (SNP) and the phenotype of interest. Such an approach is limited for common diseases, where the interactions between distant genes, or epistasis, need to be taken into account. For instance, several epistatic mechanisms have been highlighted in the onset of Alzheimer's disease (Combarros et al., 2009). Most notably, the interaction between the two genes BACE1 and APOE4 was found to be significant on four distinct datasets. Moreover, at least two epistatic interactions were also reported for multiple sclerosis (Harty et al., 2019; Galarza-Muñoz et al., 2017).

Several strategies (Cordell, 2009; Niel et al., 2015) have been developed for the detection of statistical epistasis. Many of them consist in exhaustive SNP-SNP interaction testing, followed by corrections for multiple hypothesis testing using procedures such as Bonferroni correction (Cabin & Mitchell, 2000) or the Benjamini-Hochberg (Benjamini & Hochberg, 1995) (BH) procedure. For all procedures, the correction comes at the cost of poor statistical power (Nakagawa, 2004). For high-order interactions, the loss in statistical power is aggravated by the large number of SNP tuples to consider. Moreover, exhaustive testing for high-order interactions is also accompanied by an increase in computational complexity. For increased speed, the current state-of-the-art BOOST (Wan et al., 2010) and its GPU-derivative (Yung et al., 2011) add a preliminary screening to filter non-significant interactions. Another fast interaction search algorithm in the high-dimensional setting is the *xyz*-algorithm (Thanei et al., 2018).

By contrast, instead of constructing exhaustive models, we focus on the interactions with a given variant, that we refer to as the target in what follows. The target is a formerly identified SNP that can be extracted from top hits in previous GWAS, causal genes, or experiments. The main rationale behind our approach is to leverage the established dependency between the target and the phenotype for a better detection of epistatic phenomena: a lower number of interactions has to be studied with the additional guarantee that the target affects the phenotype in question. In addition, focusing on interactions with a single variant allows us to model the interaction of this variant with all other SNPs in the genome at once, rather than pair of SNPs by pair of SNPs.

For the purpose of epistasis detection, the pure synergistic effects of the target with other variants must be decoupled from the marginal effects of the target and the other variants. A failure to address this issue can alter the results. One way to do so is to use an ℓ_1-penalized regression model (Tibshirani, 1996) with both marginal effect and quadratic interaction terms. If only one target SNP is investigated, generating as many quadratic interaction terms as remaining SNPs in the genome,

the number of coefficients in this regression is doubled compared to a linear model with only marginal effects, rather than squared if all pairwise interaction terms were to be considered. However, this is still too many in a high-dimensional context such as GWAS. To improve the inference of the interaction coefficients, Bien et al. (2013) introduced hierNET, a LASSO with hierarchy constraints between marginal and interactions terms. However, this approach does not scale to more than a hundred variables and is therefore inapplicable to GWAS data.

We turn instead towards methods developed in the context of randomized controlled trials, which aim at detecting synergies between a treatment (rather than a target SNP) and a set of covariates (rather than other SNPs) towards an outcome (rather than a phenotype). We draw on this analogy to propose two families of methods for epistasis detection. First, *modified outcome* approaches are inspired by the work of Tian et al. (2014). Here we construct a modified phenotype from the phenotype and all SNPs, in such a way that the SNPs in epistasis with the target form the support of a sparse linear regression between this modified phenotype and the non-target SNPs. Second, *outcome weighted learning* approaches are inspired by the work of Zhao et al. (2012). Here the SNPs in epistasis with the target form the support of a weighted sparse linear regression between the phenotype and the non-target SNPs, with samples weighted according to the phenotype and the target SNP.

A major difference between our setting and that of these randomized controlled trial approaches is that, where they assume that the treatment is independent from the covariates, we cannot assume independence between the target SNP and the rest of the genome. Indeed, although recombination can be expected to break down non-random associations between alleles at several loci, such associations exist, and are referred to as linkage disequilibrium (Slatkin, 2008). To account for this dependence, we borrow from the literature on causal inference in observational data and introduce propensity scores. They correspond here to the probability of the target conditionally on all non-target SNPs. In addition, the high dimensionality of the data leads us to use stability selection (Meinshausen & Bühlmann, 2010; Beinrucker et al., 2012) to select the regularization parameter of the ℓ_1-penalized regressions.

In this chapter, we develop a new framework to study epistasis by solely focusing on the synergies with a predetermined target. Most of our methods improve the recovery of interacting SNPs compared to standard methods like GBOOST or a LASSO with interaction terms. We demonstrate the performance of our methods against both of them for several types of disease models. We also conduct a case study on a real GWAS dataset of type II diabetes to demonstrate the scalability of our methods.

2.2 Material and Methods

2.2.1 Setting and notations

We jointly model genotypes and phenotypes as a triplet of random variables (X, A, Y), where Y is a discrete (e.g. in case-control studies) or continuous phenotype, $X = (X_1, \cdots, X_p) \in \{0, 1, 2\}^p$ represents a genotype with p SNPs, and A is the $(p+1)$-th target SNP of interest. The reason why we split the $p + 1$ SNPs into X and A is that our goal is to detect interactions involving A and other SNPs in X. Several selection strategies are possible for the target A: eQTL SNPs for genes with proven effect on the phenotype Y, deleterious splicing variants, or among significant SNPs in previous GWAS. In classical GWAS, the SNPs are identified on the basis of the significance of their main effects. A SNP with interaction effects only can then be overlooked. To detect such SNPs, we can use association measures such as distance correlation (Székely et al., 2007) and mutual information (Cover & Thomas, 2005) which can better capture second-order interaction effects. For the genotype X, we can choose the rest of the genome (the whole genome except the target A) or a given set of SNPs. The SNP set may correspond to a genomic region of interest e.g. gene, promoter region, or a pathway.

We restrict ourselves to a binary encoding of A in $\{-1, +1\}$, which allows us to study both recessive and dominant phenotypes, depending on how we binarize the SNP represented in A. For instance, to model dominant effects, we respectively map $\{0\}$ and $\{1, 2\}$ to $\{-1\}$ and $\{+1\}$. We also introduce a second binarized version of the target SNP A taking values in $\{0, 1\}$ by letting $\tilde{A} = (A+1)/2$. SNP binarization is a common procedure in GWAS in particular for the study of epistasis. Prabhu & Pe'er (2012) and Llinares-López et al. (2018) implement binarized genotypes, while Achlioptas et al. (2011) use locality-sensitive hashing (LSH) to transform the original genotypes into binary vectors. The question is moot in doubled haploid organisms, where the SNPs are homozygous only.

The target SNP A being sign-symmetric and binary, it is always possible to decompose the genotype and phenotype relationship as:

$$Y = \mu(X) + \delta(X) \cdot A + \varepsilon, \tag{2.1}$$

where ε is a zero mean random variable and,

$$\begin{cases} \mu(X) = \dfrac{1}{2}\left[\mathbb{E}(Y|A = +1, X) + \mathbb{E}(Y|A = -1, X)\right] , \\ \delta(X) = \dfrac{1}{2}\left[\mathbb{E}(Y|A = +1, X) - \mathbb{E}(Y|A = -1, X)\right] . \end{cases} \tag{2.2}$$

The term $\delta(X) \cdot A$ in Eq. (2.1) represents the synergistic effects between A and all SNPs in X. In the context of genomic data, we can interpret these synergies as pure epistatic effects: the main effects are accounted for by $\mu(X)$. Furthermore, if $\delta(X)$ is sparse, meaning that it only depends on a subset of elements of X, referred to as the *support* of $\delta(X)$, then the SNPs in this support are the ones interacting

with A. In other words, searching for epistatic interactions between A and SNPs in X amounts to searching for the support of δ.

To estimate this support from GWAS data, we propose several models based on sparse regressions. The common thread between them is the use of propensity scores to estimate $\delta(X)$ and its support without estimating $\mu(X)$. In causal inference, the propensity score $\pi(A|X)$ is defined as the conditional probability of A given X. The propensity score is used to compensate the differences in covariates between the two groups in observational studies, where, by contrast with randomized controlled trials, investigators have no control over the treatment assignment. In our case, this score allows us to model linkage disequilibrium (LD) between A and other nearby SNPs within X. The first family of methods we propose falls under the modified outcome banner (Tian et al., 2014). In these models, an outcome that combines the phenotype Y with the target SNP A and the propensity score $\pi(A|X)$ is fit linearly to the genomic covariates X. We propose several variants of this approach, which differ in their control of estimation errors. Our second proposal is a case-only method based on the framework of outcome weighted learning (Zhao et al., 2012). In this model, which is a weighted linear regression, the outcome is the target SNP A, and the covariates are the rest of the genotype X. The phenotype and the propensity score $\pi(A|X)$ are incorporated in the sample weights $Y/\pi(A|X)$.

Propensity-score approaches require the conditional independence of A and the potential outcomes $\{Y^{(0)}, Y^{(1)}\}$, with respect to X. This assumption still holds for genotypic data. The values of the target A only depend on the genetic background of the individual. In other words, the values of A are not "optimized" to obtain a desired outcome, unlike in non-randomized clinical trials.

The following subsections (Sections 2.2.2 and 2.2.3) elaborate on those methods. Section 2.2.4 details our approach for the estimate of the propensity score $\pi(A|X)$. Finally, Section 2.2.5 explains how we perform model selection through stability selection.

If not stated otherwise, the full data pipeline is written in the **R** language. The methods presented in this work are implemented in the **R** package **epiGWAS**, which is directly available via CRAN. The source code can also be downloaded from the GitHub repository https://github.com/EpiSlim/epiGWAS.

2.2.2 Modified outcome regression

Depending on the underlying target value and the binarization rule, only one of the two possibilities $A = +1$ or $A = -1$ is observed for a given sample. In other words, as in randomized controlled trials where, for each sample, either the treatment is applied or it is not, here, for any given sample, we do not observe the phenotype associated with the same genotype except in A which takes the other value. Hence $\delta(X)$ cannot be estimated directly from GWAS data using Eq. (2.2). The propensity score $\pi(A|X)$ comes into play to circumvent this problem. By considering the new

binarized variable $\widetilde{A} = (A+1)/2 \in \{0,1\}$, we can indeed rewrite Eq. (2.2) as:

$$\delta(X) = \frac{1}{2}\mathbb{E}\left[Y\left(\frac{\widetilde{A}}{\pi(\widetilde{A}=1|X)} - \frac{1-\widetilde{A}}{\pi(\widetilde{A}=0|X)}\right)\middle|X\right].$$

Given an estimate of $\pi(\widetilde{A}|X)$, we define the modified outcome $\widetilde{Y}$ of an observation (X, A, Y) as:

$$\widetilde{Y} = Y\left(\frac{\widetilde{A}}{\pi(\widetilde{A}=1|X)} - \frac{1-\widetilde{A}}{\pi(\widetilde{A}=0|X)}\right), \tag{2.3}$$

and re-express $\delta(X)$ simply as:

$$\delta(X) = \frac{1}{2}\mathbb{E}\left[\widetilde{Y}|X\right]. \tag{2.4}$$

Our definition of modified outcome in Eq. (2.3) generalizes that of Tian et al. (2014), where it is defined as $\widetilde{Y} = Y\widetilde{A}$; both definitions are equivalent in the specific situation considered by Tian et al. (2014) where A and X are independent, *i.e.*, $\pi(\widetilde{A}=1|X) = \pi(\widetilde{A}=1)$, and furthermore $\pi(\widetilde{A}=1) = 1/2$. Our definition (Eq. (2.3)) remains valid even when A and X are not independent. This can accommodate the diversity of the LD landscape and of the broad range of minor allele frequencies.

Given Eq. (2.4), we can estimate the support of δ from GWAS data by first transforming them into genotype-modified outcome pairs $(X_i, \widetilde{Y}_i)_{i=1,\dots,n}$, and then applying a sparse regression model for support recovery. For this purpose, we use an elastic net logistic or linear regression, combined with a stability selection procedure for model selection, as detailed in Section 2.2.5.

The inverse of the propensity score weighting in Eq. (2.3) can create numerical instability. If the conditional probabilities $\hat{\pi}(A_i=0|X_i)$ or $\hat{\pi}(A_i=1|X_i)$ are small, the weight attributed to the sample (i) can be disproportionately large relatively to other samples. Therefore, we propose several alternative definitions of $\widetilde{Y}$, which improve numerical stability and large-sample variance by controlling the inverse of the propensity score $\pi(A|X)$. A first alternative, which we call *shifted modified outcome*, simply consists in the addition of a small term $\xi = 0.1$ to obtain an upper-bound $1/\xi$ on the inverses of propensity scores:

$$\widetilde{Y}_i = Y_i\left(\frac{\widetilde{A}_i}{\pi(\widetilde{A}_i=1|X_i)+\xi} - \frac{1-\widetilde{A}_i}{\pi(\widetilde{A}_i=0|X_i)+\xi}\right).$$

In causal inference, other improvements (Austin, 2011) to the modified outcome in Eq. (2.3) have already been proposed to estimate the average treatment effects Δ given in Eq. (2.5). The transition between the second and third lines in Eq. (2.5) is made possible by the independence of $\widetilde{A}$ and the potential outcomes $\{Y^{(0)}, Y^{(1)}\}$, with respect to X.

$$\begin{aligned}\Delta &= \mathbb{E}\left[Y^{(1)}\right] - \mathbb{E}\left[Y^{(0)}\right] \\ &= \mathbb{E}\left[\mathbb{E}[Y^{(1)}|X]\right] - \mathbb{E}\left[\mathbb{E}[Y^{(0)}|X]\right] \\ &= \mathbb{E}\left[\frac{Y\widetilde{A}}{\pi(\widetilde{A}=1|X)}\right] - \mathbb{E}\left[\frac{Y(1-\widetilde{A})}{\pi(\widetilde{A}=0|X)}\right] \\ &= \mathbb{E}\left[\mathbb{E}[\widetilde{Y}|X]\right].\end{aligned} \tag{2.5}$$

It is clear from the above equation that the modified outcome $\widetilde{Y}$ can be estimated from $\mu_1 = \mathbb{E}[Y^{(1)}|X]$ and $\mu_0 = \mathbb{E}[Y^{(0)}|X]$. Lunceford & Davidian (2004) consider the following family of consistent estimators of μ_0 and μ_1 parameterized by (η_0, η_1):

$$\begin{cases} \frac{\hat{\mu}_1}{n} &= \left(\sum_{j=1}^n \frac{\widetilde{A}_j}{\pi(\widetilde{A}_j=1|X_j)}\right)^{-1} \frac{\widetilde{A}_i Y_i + \eta_1(\widetilde{A}_i - \pi(\widetilde{A}_i=1|X))}{\pi(\widetilde{A}_i=1|X)} \\ \frac{\hat{\mu}_0}{n} &= \left(\sum_{j=1}^n \frac{1-\widetilde{A}_j}{1-\pi(\widetilde{A}_j=1|X_j)}\right)^{-1} \frac{(1-\widetilde{A}_i)Y_i - \eta_0(\widetilde{A}_i - \pi(\widetilde{A}_i=1|X))}{1-\pi(\widetilde{A}_i=1|X)} \end{cases},$$

The case $(\eta_0, \eta_1) = (0, 0)$ yields the second estimator, *normalized modified outcome*, which was found in empirical studies to have a lower variance than the former estimator in Eq. (2.3) :

$$\frac{\widetilde{Y}_i}{n} = \left(\sum_{j=1}^n \frac{\widetilde{A}_j}{\pi(\widetilde{A}_j=1|X_j)}\right)^{-1} \frac{Y_i\widetilde{A}_i}{\pi(\widetilde{A}_i=1|X_i)} - \left(\sum_{j=1}^n \frac{1-\widetilde{A}_j}{\pi(\widetilde{A}_j=0|X_j)}\right)^{-1} \frac{Y_i(1-\widetilde{A}_i)}{\pi(\widetilde{A}_i=0|X_i)}.$$

A second estimator within that family is *robust modified outcome*, which is the estimator with the smallest large-sample variance. We can derive its expression by using empirical estimates of η_0^* and η_1^*, the minimizers of the variance of $\hat{\mu}_0$ and $\hat{\mu}_1$, respectively. We thus obtain:

$$\begin{aligned}\frac{\widetilde{Y}_i}{n} = &\left[\sum_{j=1}^n \frac{\tilde{A}_j}{\pi(\widetilde{A}_j=1|X_j)}\left(1 - \frac{C_1}{\pi(\widetilde{A}_j=1|X_j)}\right)\right]^{-1} \left(1 - \frac{C_1}{\pi(\widetilde{A}_i=1|X_i)}\right) \frac{\widetilde{A}_i Y_i}{\pi(\widetilde{A}_i=1|X_i)} \\ &- \left[\sum_{j=1}^n \frac{1-\widetilde{A}_j}{\pi(\widetilde{A}_j=0|X_j)}\left(1 - \frac{C_0}{\pi(\widetilde{A}_j=0|X_j)}\right)\right]^{-1} \left(1 - \frac{C_0}{\pi(\widetilde{A}_i=0|X_i)}\right) \frac{(1-\widetilde{A}_i)Y_i}{\pi(\widetilde{A}_i=0|X_i)},\end{aligned}$$

where,

$$\begin{cases} C_1 = & \frac{\sum_{j=1}^n \left((\widetilde{A}_j - \pi(\widetilde{A}_j=1|X_j))/\pi(\widetilde{A}_j=1|X_j)\right)}{\sum_{j=1}^n \left((\widetilde{A}_j - \pi(\widetilde{A}_j=1|X_j))/\pi(\widetilde{A}_j=1|X_j)\right)^2} \\ C_0 = - & \frac{\sum_{j=1}^n \left((\widetilde{A}_j - \pi(\widetilde{A}_j=1|X_j))/\pi(\widetilde{A}_j=0|X_j)\right)}{\sum_{j=1}^n \left((\widetilde{A}_j - \pi(\widetilde{A}_j=1|X_j))/\pi(\widetilde{A}_j=0|X_j)\right)^2} \end{cases}.$$

For more details about modified outcome approaches, we refer the reader to Lunceford & Davidian (2004).

2.2.3 Outcome weighted learning

Inspired by the Outcome Weighted Learning (OWL) model of Zhao et al. (2012), developed in the context of randomized clinical trials, we now propose an alternative to the modified outcome approach to estimate $\delta(X)$ and its support using a weighted binary classification formulation. As with OWL, this formulation mathematically amounts to predicting A from X, where prediction errors are weighted according to Y in the fitting process. In the original OWL proposal, the goal is to determine an optimal individual treatment rule d^* that predicts treatment A from prognostic variables X so as to maximize the clinical outcome Y. In our context, this translates to determining an optimal predictor d^* that predicts target SNP A from genotype X, so as to maximize Y (which is larger for cases than controls). We expect such a predictor to rely on the SNPs that interact with A towards predicting the phenotype Y. We assume in this section that Y only takes nonnegative values, e.g., $Y \in \{0,1\}$ for a case-control study. To take into account the dependency between A and X, we replace $\pi(A)$ with $\pi(A|Y)$ in the original OWL definition (Zhao et al., 2012) and look for the following decision rule:

$$d^* \in \underset{d:\{0,1,2\}^p \to \mathbb{R}}{\operatorname{argmin}} \mathbb{E}\left[\frac{Y}{\pi(A|X)}\phi\left(Ad(X)\right)\right], \tag{2.6}$$

where ϕ is a non-increasing loss function such as the logistic loss:

$$\forall u \in \mathbb{R}, \quad \phi(u) = \log\left(1 + e^{-u}\right). \tag{2.7}$$

The reason to consider this formulation is that:

Lemma 2.1. *The solution d^* to ((2.6))-((2.7)) is:*

$$\forall x \in \{0,1,2\}^p, \quad d^*(x) = \ln \frac{\mathbb{E}\left[Y|A=+1, X=x\right]}{\mathbb{E}\left[Y|A=-1, X=x\right]}.$$

Proof. For any $x \in \{0,1,2\}^p$, we see from Eq. (2.6) that $d^*(x)$ must minimize the function $l : \mathbb{R} \to \mathbb{R}$ defined by

$$\begin{aligned}\forall u \in \mathbb{R}, \quad l(u) &= \mathbb{E}\left[\frac{Y}{\pi(A|X=x)}\phi\left(Au\right)\middle| X=x\right] \\ &= \phi(u)\mathbb{E}\left[Y\,|A=1, X=x\right] + \phi(-u)\mathbb{E}\left[Y\,|A=-1, X=x\right].\end{aligned}$$

This function is minimized when $l'(u) = 0$, that is, when $\phi'(u)\mathbb{E}\left[Y\,|A=1, X=x\right] = \phi'(-u)\mathbb{E}\left[Y\,|A=-1, X=x\right]$, which is equivalent to:

$$\frac{\mathbb{E}\left[Y\,|A=1, X=x\right]}{\mathbb{E}\left[Y\,|A=-1, X=x\right]} = e^u.$$

□

Lemma 2.1 clarifies how d^* is related to δ as defined in Eq. (2.2): while δ is half the difference between the expected phenotype conditioned on the two alternative

values of A, d^* is the log-ratio of the same two quantities. In particular, both functions have the same sign for any genotype X. Hence we propose to estimate d^* and its support, as an approximation and alternative to estimating δ and its support, in order to capture SNPs in epistasis with A.

For any given (X, A, Y), if we define the weight $W = Y/\pi(A|X)$, we can interpret d^* in Eq (2.6) as a logistic regression classifier that predicts A from X, with errors weighted by W. Hence d^* and its support can be estimated from GWAS data by standard tools for weighted logistic regression and support estimation. As with modified outcome approaches, we use an elastic net logistic or linear regression, combined with a stability selection procedure for model selection, detailed in Section 2.2.5.

In the case of qualitative GWAS studies, we encode Y as 0 for controls and 1 for cases. The sample weights W of controls thus become 0, resulting in a case-only approach for epistasis detection. Tools such as PLINK (Purcell et al., 2007) and INTERSNP (Herold et al., 2009) similarly implement case-only analyses, which can be more powerful in practice than a joint case-control analysis (Cordell, 2009; Gatto, 2004; Piegorsch et al., 1994; Yang et al., 1999). In the case of PLINK and INTERSNP, additional hypotheses such as the independence of SNP–SNP frequencies are nonetheless needed to ensure the validity of the statistical test. In our case, the family of weights $\{W_i = 1/\pi(A_i|X_i)\}_{i=1,\cdots,n}$ accounts for the dependency between the target A and the genotype X. We can therefore forego such hypotheses on the data. We may even argue that the controls are indirectly included in the regression model through $\pi(A|X)$. It represents the dependency pattern within the general population, which consists of both cases and controls.

2.2.4 Estimate of the propensity score

In causal inference, the estimation of propensity scores $\pi(A|X)$ is often achieved thanks to parametric models such as a logistic regression between A and X. Because of the risk of overfitting in such an ultra high-dimensional setting, we turn instead towards Hidden Markov Models, which are commonly used in genetics to model linkage disequilibrium and were initially developed for imputation (Scheet & Stephens, 2006). In this model, the hidden states represent contiguous clusters of phased haplotypes. The emission states correspond to SNPs.

Since the structural dependence is chromosome-wise, we only retain the SNPs located on the same chromosome as the SNP A – which we denote here by X_A – for the estimate of $\pi(A|X)$. Mathematically, this is equivalent to the independence of the SNPs A and X_A from the SNPs of other chromosomes.

The pathological cases $\pi(A|X_A) \approx 1$ and $\pi(A|X_A) \approx 0$ can be avoided by the removal of all SNPs within a certain distance of A. In our implementation, we first performed an adjacency-constrained hierarchical clustering of the SNPs located on the chromosome of the target A. We fixed the maximum correlation threshold at 0.5. To alleviate strong linkage disequilibrium, we then discarded all neighboring SNPs within a three-cluster window of SNP A. Such filtering is sensible since we

are looking for biological interactions between functionally-distinct regions. The neighboring SNPs are not only removed for the estimation of the propensity score, but also in the regression models searching for interactions.

After the filtering and the fitting of the unphased genotype model using fastPHASE, the last remaining step is the application of the forward algorithm (Rabiner, 1989) to obtain an estimate of the two potential observations $(A = 1, X_A)$ and $(A = -1, X_A)$. The Bayes theorem yields the desired propensity scores $\pi(A|X) = \pi(A|X_A) = \pi(A, X_A)/(\pi(A = +1, X_A) + \pi(A = -1, X_A))$.

2.2.5 Support estimation

In order to estimate the support of δ in the case of modified outcome regression ((2.4)), and of d^* in the case of OWL ((2.6)), we model both functions as linear models and estimate non-zero coefficients by elastic net regression (Zou & Hastie, 2005) combined with stability selection (Haury et al., 2012).

More precisely, given a GWAS cohort $(X_i, A_i, Y_i)_{i=1,\dots,n}$, we first define empirical risks for a candidate linear model $x \mapsto \gamma^\top x$ for δ and d^* as respectively

$$R_1(\gamma) = \frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{Y}_i - \gamma^\top X_i\right)^2, \quad R_2(\gamma) = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{\pi(A_i|X_i)}\phi(A_i\gamma^\top X_i)\,.$$

For a given regularization parameter $\lambda > 0$ and empirical risk $R = R_1$ or $R = R_2$, we then define the elastic net estimator:

$$\widehat{\gamma_\lambda} \in \underset{\gamma}{\operatorname{argmin}}\; R(\gamma) + \lambda\left[(1-s)||\gamma||_1 + \frac{1}{2}s||\gamma||_2^2\right],$$

where we fix $s = 10^{-6}$ to give greater importance to the L_1-penalization. Over a grid of values Λ for the penalization parameter λ, we subsample $N = 50$ times without replacement over the whole cohort. The size of the generated subsamples $I_1, \cdots, I_N$ is $\lfloor n/2 \rfloor$. Each subsample I provides a different support for $\widehat{\gamma}_\lambda$, which we denote $\widehat{S}^\lambda(I)$. For $\lambda \in \Lambda$, the empirical frequency of the variable X_k entering the support is then given by:

$$\widehat{\omega}_k^\lambda = \frac{1}{N}\sum_{j=1}^{N}\mathbb{1}(k \in \widehat{S}^\lambda(I_j)).$$

In the original stability selection procedure (Meinshausen & Bühlmann, 2010), the decision rule for including the variable k in the final model is $\max_{\lambda\in\Lambda}\widehat{\omega}_k^\lambda \geq t$. The parameter t is a predefined threshold. For noisy high-dimensional data, the maximal empirical frequency along the stability path $\max_{\lambda\in\Lambda}\widehat{\omega}_k^\lambda$ may not be sufficiently robust because of its reliance on a single noisy measure of $\widehat{\omega}_k^\lambda$ to derive the maximum. Instead, we used the area under the stability path, $\int_\lambda \widehat{\omega}_k^\lambda \,\mathrm{d}\lambda$, as proposed by Haury et al. (2012). The main intuition behind the better performance is the early entry of causal variables into the LASSO path.

Finally, to determine the grid Λ, we use the **R** package **glmnet** (Friedman et al., 2010). We generate a log-scaled grid of 200 values $(\lambda_l)_{l=1,\cdots,200}$ between $\lambda_1 = \lambda_{max}$ and $\lambda_{200} = \lambda_{max}/100$, where $\lambda_{\max}$ is the maximum λ leading to a non-zero model. To improve inference, we only retain the first half of the path comprised between λ_1 and λ_{100} (see Figure 2.1). The benefit of a thresholded regularization path is to discard a large number of irrelevant covariates that enter the support for low values of λ.

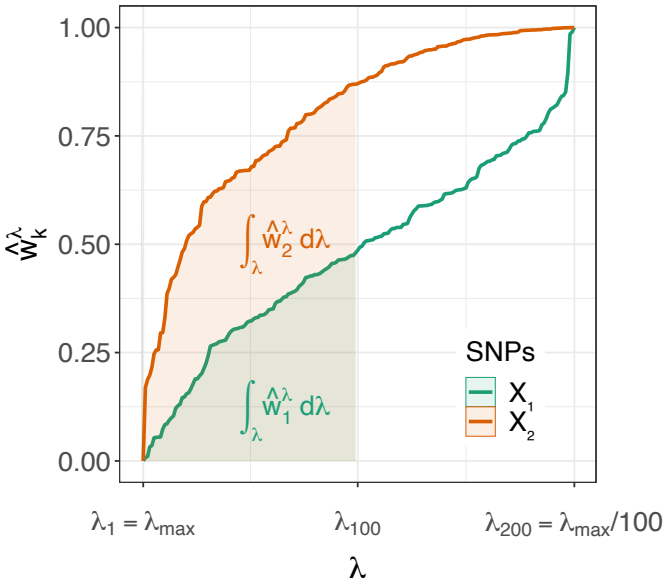


Figure 2.1: Scoring of two SNPs X_1 and X_2. The scores are the areas under the first half of their stability paths comprised between λ_1 and λ_{100}.

2.3 Results

2.3.1 Simulations

Disease model

We simulate phenotypes using a logit model with the following structure:

$$\text{logit}(P(Y = 1|\widetilde{A} = i, X)) = \beta_{i,V}^T X_V + \beta_W^T X_W + X_{Z_1}^T \,\text{diag}\,(\beta_{Z_1,Z_2})\, X_{Z_2},$$

where V, W, Z_1 and Z_2 are random subsets of $\{1, \cdots, p\}$. The variables within the vector X_V interact with A. The variables in X_W corresponds to marginal effects, while X_{Z_1} and X_{Z_2} correspond to pairs of quadratic effects between SNPs that

exclude A. The effect sizes $\beta_{0,V}, \beta_{1,V}, \beta_W$ and β_{Z1,Z_2} are sampled from $\mathcal{N}(0,1)$. Given the symmetry around 0 of the effect size distributions, the simulated cohorts are approximately equally balanced between cases and controls.

To account for the diversity of effect types in disease models, we simulate four scenarios with different overlap configurations between X_V and (X_W, X_{Z_1}):

- Synergistic only effects, $|V \cap W| = 0, |V \cap Z_1| = 0, |V| = |W| = |Z_1| = |Z_2| = 8$;
- Partial overlap between synergistic and marginal effects, $|V \cap W| = 4, |V \cap Z_1| = 0, |V| = |W| = |Z_1| = |Z_2| = 8$;
- Partial overlap between synergistic and quadratic effects, $|V \cap W| = 0, |V \cap Z_1| = 4, |V| = |W| = |Z_1| = |Z_2| = 8$;
- Partial overlap between synergistic and quadratic/marginal effects, $|V \cap W| = 2, |V \cap Z_1| = 2, |V| = |W| = |Z_1| = |Z_2| = 8$.

For each of the above scenarios, we conducted 125 simulations: 5 sets of causal SNPs $\{A, V, W, Z_1, Z_2\} \times 5$ sets of size effects $\{\beta_{0,V}, \beta_{1,V}, \beta_W, \beta_{Z1,Z_2}\} \times 5$ replicates. Within each scenario, we considered multiple SNP sets to model the range of MAFs and LD which can exist between A and X.

Because of the filtering window around the SNP A, the causal SNPs (X_V, X_W, Z_1, Z_2) were sampled outside of that window. The second constraint on the causal SNPs is a lower bound on the minor allele frequencies (MAF). We fixed that bound at 0.2. The goal is to obtain well-balanced marginal distributions for the different variants. For rare variants, it is difficult to untangle the statistical power of any method from the inherent difficulty in detecting them. The lower bound is also coherent with the common disease-common variant hypothesis (Schork et al., 2009): the main drivers of complex/common diseases are common SNPs.

Genotype simulations

For the sake of coherence, we simulated genotypes using the second release of HAPGEN (Su et al., 2011). The underlying model for HAPGEN is the same hidden Markov model used in fastPHASE. The starting point of the simulations is a reference set of population haplotypes. The accompanying haplotypes dataset is the 1000 Genomes phase 3 reference haplotypes (Auton et al., 2015). In our simulations, we only use the European population samples. The second input to HAPGEN is a fine scale recombination map. Consequently, the simulated haplotypes/genotypes exhibit the same linkage disequilibrium structure as the original reference data.

In comparison to the HAPGEN-generated haplotypes, the markers density for SNP arrays is significantly lower. For example, the sequencing technology for the WTCCC case-control consortium (Burton et al., 2007) is the Affymetrix 500K. As its name suggests, "only" five hundred thousand positions are genotyped. As most GWAS are based on SNP array data, we only extract from the simulated genotypes the markers of the Affymetrix 500K. In the subsequent QC step, we only retain

common bi-allelic SNPs defined by a MAF > 0.01. We also remove SNPs that are not in a Hardy-Weinberg equilibrium $(p < 10^{-6})$. We do not conduct any additional LD pruning for the SNPs in X. For univariate GWAS, LD pruning reduces dimensionality while approximately maintaining the same association patterns between genotype and phenotype. For second order interaction effects, the loss of information can be more dramatic, as the retained SNP pairs can be insufficient to represent the complex association of corresponding genomic regions with the phenotype.

For iterative simulations, HAPGEN can be time-consuming, notably for large cohorts consisting of thousands of samples. We instead proceed in the following way: we generate once and for all a large dataset of 20 thousand samples on chromosome 22. To benchmark for varying sample sizes $n \in \{500, 1000, 2000, 5\,000\}$, we iteratively sample uniformly and without replacement n-times the population of 20 000 individuals to create 125 case-control cohorts. On chromosome 22, we then select $p = 5\,000$ SNPs located between the nucleotide positions 16 061 016 and 49 449 618. We do not conduct any posterior pruning to avoid filtering out the true causal SNPs.

Evaluation

We benchmark our new methods against two baselines. The first method is GBOOST (Wan et al., 2010), a state-of-the-art method for epistasis detection. For each SNP pair, it implements the log-likelihood ratio statistic to compare the goodness of fit of two models: the full logistic regression model with both main effect and interaction terms, and the logistic regression model with main effects only. The preliminary sure screening step in GBOOST to discard a number of SNPs from exhaustive pairwise testing was omitted, since we are only interested in the ratio statistic for all pairs of the form (A, X_k), where X_k is the k-th SNP in X. The second method, which we refer to as product LASSO, originates from the machine learning community. It was developed by Tian et al. (2014) to estimate interactions between a treatment and a large number of covariates. It fits an L_1-penalized logistic regression model with $A \times X$ as covariates. The variable of interest A is symmetrically encoded as $\{-1, +1\}$. Under general assumptions, Tian et al. (2014) show how this model works as a good approximation to the optimal decision rule d^* (see Section 2.2.3).

We visualize the support estimation performance in terms of receiver-operating characteristic (ROC) curves and precision-recall (PR) curves. For a particular method in a given scenario, a single ROC (resp. PR) curve allows to visualize the ability of the algorithm to recover causal SNPs. For each SNP, the prediction score is the area under its corresponding stability path. The ground truth label is 1 for the SNPs interacting with the target A, and 0 otherwise. In the high-dimensional setting of GWAS, the use of raw scores instead of p-values lends more robustness to our methods, by avoiding finite-sample approximations of the score distributions and multiple hypothesis corrections.

The covariates and the outcome differ between our methods. That implies a

different regularization path for each method and as a result, incomparable stability paths. For better interpretability and comparability between the methods, we use the position l on the stability path grid $\Lambda = (\lambda_l)$ s.t. $\lambda_l > \lambda_{l+1}$ instead of the value of λ_l for computing the area under the curve.

In Figure 2.2, we provide the ROC and PR curves for the fourth scenario which corresponds to a partial overlap between synergistic and quadratic/marginal effects and for a sample size $n = 500$. Because of space constraints, all ROC/PR figures and corresponding AUC tables are listed in Appendix A.2. The figures represent the average ROC and PR curves of the 125 simulations in each of the four scenarios. To generate those figures, we used the **R** package **precrec** (Saito & Rehmsmeier, 2016). It performs nonlinear interpolation in the PR space. The AUCs were computed with same package.

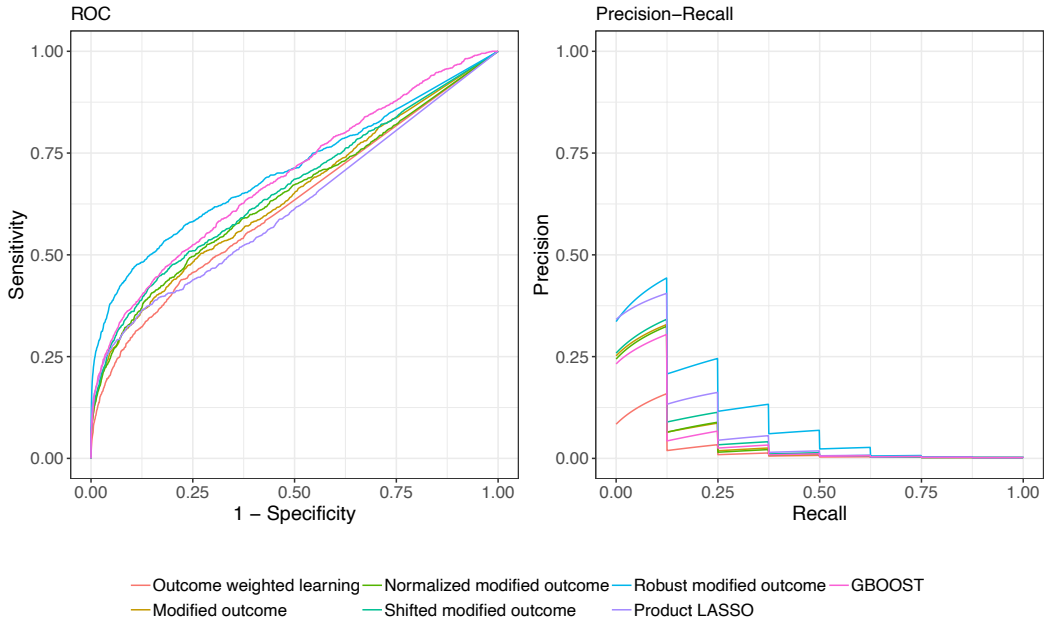


Figure 2.2: Average ROC (left) and PR (right) curves for the fourth scenario and $n = 500$

Regardless of the scenario and the sample size, the areas under all ROC curves are higher than 0.5. That confirms that all of them perform better than random, yet with varying degrees of success. By contrast, the overall areas under the precision-recall curves are low. The maximum area under the precision-recall curve is 0.41, attained by modified outcome with shifted weights for $n = p$. This can be attributed to the imbalanced nature of the problem: 8 synergistic SNPs out of 5 000. We also check that the AUCs increase with the cohort size for both ROC and PR domains.

The best performing methods are robust modified outcome and GBOOST. Robust modified outcome has a slight lead in terms of ROC AUCs, notably for low sample sizes. The latter setup is the closest to our intended application in genome-wide association studies. Of special interest to us in the ROC space is the bottom-

left area. It reflects the performance of highly-ranked instances. For all scenarios, we witness a better start for robust modified outcome. The other methods within the modified outcome family behave similarly. Such a result was expected because of their theoretical similarities. Despite the model misspecification, product LASSO performs rather well. On average, it comes third to GBOOST and robust modified outcome. The outcome weighted learning approach which is an approximation to estimating the sign of δ has consistently been the worst performer in the ROC space.

In PR space, the results are more mixed. For low sample sizes, robust modified outcome is still the best performing method. As the sample size increases, we observe that other methods within the modified outcome family, notably shifted modified outcome, surpass the robust modified outcome approach. Surprisingly, the good performance of GBOOST in ROC space was not reproduced in PR space. This might be explained by the highly imbalanced nature of the problem and the lower performance of GBOOST, compared to robust modified outcome in the high specificity region of the ROC curves (lower left). By contrast, product LASSO is always trailing the best performer of the modified outcome family. As for ROC curves, we are also interested in the beginning of the PR curves. For a recall rate of 0.125, the highest precision rate is near 0.5 for the first, third and fourth scenarios. That implies that we detect on average one causal SNP in the first two SNPs. For the second scenario, the highest precision rate is even higher at approximately 0.68. The area under the stability path is then a robust score for model selection in the high dimensional setting.

It is worth noting the homogeneous behavior of the different methods across the four scenarios. For a given sample size, and for a given method, the ROC and PR AUCs are similar. This suggests they all successfully filtered out the common effects term $\mu(X)$ even in presence of an overlap between the causal SNPs within $\mu(X)$ and $\delta(X)$.

2.3.2 Case study : type II diabetes dataset of the WTCCC

As a case study, we selected the type II diabetes dataset of the WTCCC (Burton et al., 2007) to illustrate the scalability of our methods to real datasets. To the best of our knowledge, no confirmed epistatic interactions exist for type II diabetes. We instead propose to study the synergies with a particular target: $rs41475248$ on chromosome 8. The first criterion to our choice is the presence of a significant epistatic effect. With GBOOST, the SNP $rs41475248$ is involved in 3 epistatic interactions, when controlling for a false discovery rate of 0.05. The second criterion is being a common variant. The MAF of the selected target is 0.45.

Before running our methods on the WTCCC dataset, we applied the same QC procedures with the following thresholds: 0.01 for minor-allele frequencies and $p > 10^{-6}$ for the Hardy-Weinberg equilibrium. No additional pruning is performed. The number of remaining variants is 354 439 SNPs. The number of samples is 4 897, split between 1 953 cases and 2 944 controls.

To solve the different L_1-penalized regressions, we abandoned **glmnet** in favor of another solver, **biglasso** (Zeng & Breheny, 2017). **glmnet** does not accept as input such ultra-high dimensional design matrices. On the other hand, **biglasso** was specifically developed for similar settings thanks to its multi-threaded implementation and utilization of memory-mapped files. Because **biglasso** does not implement sample weighting, it cannot be used to run outcome weighted learning. Moreover, this approach performed worse than the modified outcome approaches on simulated data, and we therefore excluded it from this case study.

The main difficulty for the evaluation of GWAS methods is the biological validation of the study results. We often lack evidence to correctly label each SNP as being involved or not in an epistatic interaction. Evaluating the real model selection performance of the different methods on real datasets is then impossible. However, we can study the concordance between them. A common way to proceed is Kendall's tau which is a measure of rank correlation. In Table 2.1, we give the correlation matrix of our methods and the two baselines of Section 2.3.1. All elements are positive which indicates a relative agreement between the methods. While methods using different mathematical definitions of epistasis cannot be expected to return the same results, those with similar or identical underlying models should capture similar genetic architectures and return more similar results. Modified outcome, normalized modified outcome and shifted modified outcome have the highest correlation coefficients. Such a result was expected because of their theoretical similarities. We also note that the lowest score is for robust modified outcome and GBOOST. In the previous section, these two methods were the best performing. This suggests those two methods can make different true discoveries.

In any follow-up work, we will only exploit the highly-ranked variants. A weighted tau statistic that assigns a higher weight to the first instances is therefore more relevant. Weighted nonnegative tau statistics better assess the relative level of concordance between different pairs of methods, while the sign in Kendall's tau shows if two methods rather agree or disagree. In Table 2.2, we list Kendall's tau coefficients with multiplicative hyperbolic weighting. Similarly, we notice that robust modified outcome is least correlated with GBOOST and most correlated with product LASSO.

Aside from rank correlation, another option to appraise the results is to measure the association between the top SNPs for each method and the phenotype. Table 2.3 lists the Cochran-Armitage test p-values for the top 25 SNPs for each method in an increasing order. Despite being synthetic univariate measures, the Cochran-Armitage statistics give us an indication of the true ranking performance. Robust modified outcome is clearly the method with the lowest p-values. For instance, the top 14 SNPs have a p-value lower than 0.001. That confirms the result of our simulations that robust modified outcome is the best performer for capturing causal SNPs. The p-values associated to product LASSO and GBOOST are also relatively low, with respectively 5 and 4 p-values lower than 0.001. However, we note the overall difficulty in drawing clear conclusions for all methods. Without multiple testing correction, most of the p-values for each method already exceed

	GBOOST	Modified outcome	Normalized modified outcome	Shifted modified outcome	Robust modified outcome	Product LASSO
GBOOST	1.000	0.200	0.203	0.202	0.070	0.152
Modified outcome	0.200	1.000	0.411	0.405	0.150	0.283
Normalized modified outcome	0.203	0.411	1.000	0.406	0.153	0.284
Shifted modified outcome	0.202	0.405	0.406	1.000	0.179	0.301
Robust modified outcome	0.070	0.150	0.153	0.179	1.000	0.257
Product LASSO	0.152	0.283	0.284	0.301	0.257	1.000

Table 2.1: Concordance between methods used to determine SNPs synergistic to rs41475248 in type II diabetes, measured by Kendall's tau.

	GBOOST	Modified outcome	Normalized modified outcome	Shifted modified outcome	Robust modified outcome	Product LASSO
GBOOST	1.000	0.483	0.481	0.517	0.423	0.501
Modified outcome	0.483	1.000	0.851	0.857	0.462	0.586
Normalized modified outcome	0.481	0.851	1.000	0.860	0.467	0.594
Shifted modified outcome	0.517	0.857	0.860	1.000	0.504	0.603
Robust modified outcome	0.423	0.462	0.467	0.504	1.000	0.596
Product LASSO	0.501	0.586	0.594	0.603	0.596	1.000

Table 2.2: Concordance between methods used to determine SNPs synergistic to rs41475248 in type II diabetes, measured by Kendall's tau with multiplicative weights.

classical significance levels *e.g.* 0.05. For 3 out of 6 methods, the p-values of the 25^{th} SNP are greater than 0.90. Nonetheless, the existence of such high p-values further demonstrates the capacity of our methods in discovering novel associations undetected by univariate methods.

GBOOST	Modified outcome	Normalized modified outcome	Shifted modified outcome	Robust modified outcome	Product LASSO
0.0000047	**0.0000000**	**0.0000000**	**0.0000000**	**0.0000000**	**0.0000047**
0.0002632	**0.0000015**	**0.0000015**	**0.0000015**	**0.0000000**	**0.0000075**
0.0002667	**0.0002667**	**0.0002667**	**0.0002667**	**0.0000001**	**0.0000172**
0.0006166	0.0027308	0.0027308	0.0027308	**0.0000012**	**0.0002667**
0.0015069	0.0093734	0.0093734	0.0093734	**0.0000049**	**0.0005286**
0.0028872	0.0633055	0.0633055	0.0633055	**0.0000059**	0.0110392
0.0031533	0.0724198	0.0724198	0.0724198	**0.0000075**	0.0122543
0.0034323	0.0925877	0.0925877	0.0771170	**0.0000172**	0.0152912
0.0081128	0.1126164	0.1043632	0.0925877	**0.0002030**	0.0346055
0.0093734	0.1272777	0.1126164	0.1126164	**0.0002667**	0.0347964
0.0142695	0.2552284	0.1567974	0.1272777	**0.0003047**	0.0396448
0.0633055	0.2926915	0.2971396	0.1639805	**0.0004643**	0.0396932
0.0771170	0.3436741	0.3529366	0.2971396	**0.0005286**	0.0527104
0.1616393	0.3529366	0.5012038	0.3529366	**0.0005841**	0.0633055
0.2089538	0.5871432	0.5506690	0.5012038	0.0015214	0.0763114
0.2114803	0.5985624	0.5985624	0.5707955	0.0016353	0.1126164
0.2256368	0.6016953	0.7183847	0.5985624	0.0025709	0.1185275
0.2586186	0.6361937	0.7199328	0.7000506	0.0064196	0.1796624
0.2654530	0.7183847	0.7342897	0.7183847	0.0080405	0.2552284
0.4105146	0.7342897	0.7656055	0.7342897	0.0110392	0.3308890
0.4323674	0.7979653	0.7706524	0.7979653	0.0122543	0.3867409
0.4376669	0.8683271	0.7979653	0.7993838	0.0124442	0.5045073
0.4796214	0.8820292	0.7993838	0.8683271	0.0136452	0.5985624
0.5871432	0.9188037	0.8820292	0.8821872	0.0346055	0.6238335
0.9479547	0.9903334	0.8821872	0.9188037	0.0396932	0.8821872

Table 2.3: Cochran-Armitage test p-values for the top 25 SNPs for each method

2.4 Discussion

In this chapter, we have proposed several methods, inspired from the clinical trials literature, to select SNPs having synergystic effects with a particular target SNP towards a phenotype. The consistency of our results across the four disease models show that the proposed methods are rather successful. Indeed, their per-

formance is not strongly impacted by the presence/absence of other marginal and epistatic effects. Among the methods we propose, robust modified outcome is the most suited to real GWAS applications. Its superior performance is partially due to its robustness against propensity score misspecification. The AUCs for robust modified outcome are overall the highest in addition to its retrieval performance for highly-ranked instances. More importantly, robust modified outcome outperforms GBOOST and other regression-based methods. This is particularly true for small number of samples ($n = 500$), which is the closest setup to real GWAS datasets. However, the low PR AUCs show that there is still room for improvement. The highest observed PR AUC is 0.17. Interestingly, we note that several of our methods clearly outperform GBOOST across all scenarios and all sample sizes in the PR space. Nonetheless, GBOOST behaves similarly to our methods in the ROC space. Such differences between ROC and PR curves are common for highly-imbalanced datasets where PR curves are more informative and discriminative (Davis & Goadrich, 2006).

In our simulations, ROC and PR AUCs were relatively close between all methods. On the other hand, according to two rank correlation measures (Kendall's tau and weighted Kendall's tau), the results do not strongly overlap between the different methods (values far from 1). For instance, GBOOST least agrees with robust modified outcome. However, the two methods are the best performing in our simulations. Different approaches seem to discover different types of interactions (Bessonov et al., 2015). We conclude that a consensus method combining GBOOST and robust modified outcome could better improve the recovery of interacting SNPs.

The carried simulations prove that the highly-ranked SNPs include false positives. This is accentuated by the imbalanced nature of our problem: a handful of causal SNPs for thousands of referenced SNPs. Hopefully, the continual decrease in genotyping costs will result in a dramatic increase in sample sizes and, in consequence, statistical power. For instance, the UK Biobank (Bycroft et al., 2018) comprises full genome-wide data for five hundred thousand individuals.

The case study that we carried for type II diabetes demonstrates the scalability of our methods to real GWAS. To reduce runtime, one can reduce the number of subsamples used for stability selection; however this may come at the expense of performance. The development of new and faster LASSO solvers (Le Morvan & Vert, 2018; Massias et al., 2018) for large scale problems will further help broaden the adoption of our methods by end-users without compromising statistical performance.

The main contribution of our work is extending the causal inference framework to epistasis detection by developing a new family of methods. They rely on propensity scores to detect interactions with specific SNP targets. Given our partial understanding of common diseases and the overall lack of statistical power of existing tools, such refocused models can be more useful to further our understanding of disease etiologies. Hundreds of genes have already been associated with several diseases via univariate GWAS. The next step is to leverage such findings to detect additional synergies between these genes and the rest of the genome. Beyond a

better understanding of disease mechanisms through new biomarker discovery, we see the development of combination drug therapies as an additional application of our work.

A first area of future improvement for our methods is propensity score estimation, which can benefit from a large number of recent methods (Athey et al., 2018). A second area is incorporating multiple covariates (whether clinical covariates, variables encoding population structure or other genetic variants) to account for, among other things, higher-order interactions and population structure. A straightforward solution is to include additional variables in X, which encode for the other covariates. However, this will impact the consistency and interpretability of the propensity scores. A second potential solution is the use of modified targets which combine the original target with the other covariates e.g. target $\times$ gender. We think that such outcomes have not been explored because of the insufficiency of the representation by a single binary variable. To address this issue we can, for example, borrow some of the ideas in VanderWeele & Hernan (2013) to construct richer representations.

Acknowledgements

This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113, 085475 and 090355.

Chapter 3

kernelPSI: a Post-Selection Inference Framework for Nonlinear Variable Selection

---

Publication and Dissemination: *The work in this chapter has been published as joint work with Chloé-Agathe Azencott, Clément Chatelain and Jean-Philippe Vert in Slim et al. (2019), and orally presented at ICML 2019.*

Abstract: *Model selection is an essential task for many applications in scientific discovery. The most common approaches rely on univariate linear measures of association between each feature and the outcome. Such classical selection procedures fail to take into account nonlinear effects and interactions between features. Kernel-based selection procedures have been proposed as a solution. However, current strategies for kernel selection fail to measure the significance of a joint model constructed through the combination of the basis kernels. In this chapter, we exploit recent advances in post-selection inference to propose a valid statistical test for the association of a joint model of the selected kernels with the outcome. The kernels are selected via a step-wise procedure which we model as a succession of quadratic constraints in the outcome variable.*

Résumé : *La sélection de modèles est une tâche essentielle pour de nombreuses applications scientifiques. Les approches les plus courantes reposent sur des mesures linéaires univariées d'association entre chaque variable et la sortie. De telles procédures de sélection classiques ne prennent pas en compte les effets nonlinéaires et les interactions entre variables. Des procédures de sélection basées sur les noyaux ont été proposées comme solution. Cependant, les stratégies actuelles de sélection des noyaux ne parviennent pas à mesurer l'importance d'un modèle commun construit par la combinaison de plusieurs noyaux de base. Dans ce chapitre, nous exploitons les avancées récentes de l'inférence post-sélection pour proposer un test statistique valide*

pour l'association d'un modèle commun des noyaux sélectionnés avec la sortie. Les noyaux sont sélectionnés via une procédure par étapes que nous modélisons comme une succession de contraintes quadratiques dans la variable de sortie.

3.1 Introduction

Variable selection is an important preliminary step in many data analysis tasks, both to reduce the computational complexity of dealing with high-dimensional data and to discard nuisance variables that may hurt the performance of subsequent regression or classification tasks. Statistical inference about the selected variables, such as testing their association with an outcome of interest, is also relevant for many applications, such as identifying genes associated with a phenotype in genome-wide association studies. If the variables are initially selected using the outcome, then standard statistical tests must be adapted to correct for the fact that the variables tested after selection are likely to exhibit strong association with the outcome, because they were selected for that purpose.

This problem of *post-selection inference* (PSI) can be solved by standard data splitting strategies, where we use different samples for variable selection and statistical inference (Cox, 1975). Splitting data is however not optimal when the total number of samples is limited, and alternative approaches have recently been proposed to perform proper statistical inference after variable selection (Taylor & Tibshirani, 2015). In particular, in the *conditional coverage* setting of Berk et al. (2013), statistical inference is performed conditionally to the selection of the model. For linear models with Gaussian additive noise, Lee et al. (2016); Tibshirani et al. (2016) show that proper statistical inference is possible and computationally efficient in this setting for features selected by lasso, forward stepwise or least angle regression. In these cases it is indeed possible to characterize the distribution of the outcome under a standard null hypothesis model conditionally to the selection of a given set of features. This distribution is a Gaussian distribution truncated to a particular polyhedron. Similar PSI schemes were derived when features are selected not individually but in groups (Loftus & Taylor, 2015; Yang et al., 2016a; Reid et al., 2017).

Most PSI approaches have been limited to linear models so far. In many applications, it is however necessary to account for nonlinear effects or interactions, which requires nonlinear feature selection. This requires generalizing PSI techniques beyond linear procedures. Recently, Yamada et al. (2018) took a first step in that direction by proposing a PSI procedure to follow kernel selection, where kernels are used to generalize linear models to the nonlinear setting. However, their approach is limited to a single way of selecting kernels, namely, marginal estimation of the Hilbert-Schmidt Independent Criterion (HSIC) independence measure (Song et al., 2007). In addition, it only allows to derive post-selection statistical guarantees for one specific question, that of the association of a selected kernel with the outcome.

In this chapter we go one step further and propose a general framework for kernel selection, that leads to valid PSI procedures for a variety of statistical inference questions. Our main contribution is to propose a large family of statistics that estimate the association between a given kernel and an outcome of interest, that can be formulated as a quadratic function of the outcome. This family includes in particular the HSIC criterion used by Yamada et al. (2018), as well as a generalization

to the nonlinear setting (a "kernelization") of the criterion used by Loftus & Taylor (2015); Yang et al. (2016a) to select a group of features in the linear setting. When these statistics are used to select a set of kernels, by marginal filtering or by forward or backward stepwise selection, we can characterize the set of outcomes that lead to the selection of a particular subset as a conjunction of quadratic inequalities. This paves the way to various PSI questions by sampling-based procedures.

3.2 Settings and Notations

Given a data set of n pairs $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, where for each $i \in [1, n]$ the data $x_i \in \mathcal{X}$ for some set $\mathcal{X}$ and the outcome $y_i \in \mathbb{R}$, our goal is to understand the relationship between the data and the outcome. We denote by $Y \in \mathbb{R}^n$ the vector of outcomes ($Y_i = y_i$ for $i \in [1, n]$). We further consider a set of S positive definite kernels $\mathcal{K} = \{k_1, \ldots, k_S\}$ defined over $\mathcal{X}$, and denote $K_1, \ldots, K_S$ the corresponding $n \times n$ Gram matrices (i.e., for any $t \in [1, S], i, j \in [1, n], [K_t]_{ij} = k_t(x_i, x_j)$). We refer to the kernels $k \in \mathcal{K}$ as *local* or *basis* kernels. Our goal is to select a subset of S' local kernels $\{k_{i_1}, \cdots, k_{i_{S'}}\} \subset \mathcal{K}$ that are most associated with the outcome Y, and then to measure the significance of their association with Y.

The choice of basis kernels $\mathcal{K}$ allows us to model a wide range of settings for the underlying data. For example, if $\mathcal{X} = \mathbb{R}^d$, then a basis kernel can only depend on a single coordinate, or on a group of coordinates, in which case selecting kernels leads to variable selection (individually or by groups). Another useful scenario is to consider nonlinear kernels with different hyperparameters, such as a Gaussian kernel with different bandwidth, in which case kernel selection leads to hyperparameter selection.

3.3 Kernel Association Score

Our kernel selection procedure is based on the following general family of association scores between a kernel and the outcome:

Definition 3.1. A quadratic kernel association score is a function $s : \mathbb{R}^{n\times n} \times \mathbb{R}^n \to \mathbb{R}$ of the form

$$s(K, Y) = Y^\top Q(K) Y \,, \tag{3.1}$$

for some function $Q : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$.

If $s(K, Y)$ is a positive definite quadratic form in Y (i.e., if $Q(K)$ is positive semi-definite), we can rewrite it as:

$$s(K, Y) = \|\widehat{Y}_K\|^2 \,, \tag{3.2}$$

where $\widehat{Y}_K = H(K)Y$ is called a *prototype* for a "hat" function $H : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ (take for example $H = Q^{1/2}$). We borrow the term "prototype" from Reid et al. (2017), who use it to design statistical tests of linear association between the outcome and a group of features.

One reason to consider quadratic kernel association scores is that they cover and generalize several measures used for kernel or feature selection. Consider for example $H_{\text{proj}}(K) = KK^{+}$, where K^{+} is the Moore-Penrose inverse of K. The score proposed by Loftus & Taylor (2015) for a group of d features encoded as $X_g \in \mathbb{R}^{n\times d}$ is a special case of H_{proj} with $K = X_g X_g^{\top}$. In this case, the prototype $\widehat{Y}$ is the projection of Y onto the space spanned by the features.

If $K = \sum_{i=1}^{r} \lambda_i u_i u_i^{\top}$ is the singular value decomposition of K, with $\lambda_1 \geq \ldots \geq \lambda_r > 0$, H_{proj} can be rewritten as

$$H_{\text{proj}}(K) = \sum_{i=1}^{r} u_i u_i^{\top} . \tag{3.3}$$

For a general kernel K, which may have large rank r, we propose to consider two regularized versions of Eq. ((3.3)) to reduce the impact of small eigenvalues. The first one is the *kernel principal component regression (KPCR) prototype*, where $\widehat{Y}$ is the projection of Y onto the first $k \leq r$ principal components of the kernel:

$$H_{\text{KPCR}}(K) = \sum_{i=1}^{k} u_i u_i^{\top} .$$

The second one is the *kernel ridge regression (KRR) prototype*, where $\widehat{Y}$ is an estimate of Y by kernel ridge regression with parameter $\lambda \geq 0$:

$$H_{\text{KRR}}(K) = K\left(K + \lambda I\right)^{-1} = \sum_{i=1}^{k} \frac{\lambda_i}{\lambda_i + \lambda} u_i u_i^{\top} .$$

The ridge regression prototype was proposed by Reid et al. (2017) in the linear setting to capture the association between a group of features and an outcome; here we generalize it to the more general kernel setting.

In addition to these prototypes inspired by those used in the linear setting to analyze groups of features, we now show that empirical estimates of the HSIC criterion (Gretton et al., 2005b), widely used to assess the association between a kernel and an outcome (Yamada et al., 2018), is also a quadratic kernel association score. More precisely, given two $n \times n$ kernel matrices K and L, Gretton et al. (2005b) propose the following measure:

$$\widehat{\text{HSIC}}_{\text{biased}}(K, L) = \frac{1}{(n-1)^2} \operatorname{trace}(K\, \Pi_n\, L\, \Pi_n) , \tag{3.4}$$

where $\Pi_n = I_{n\times n} - \frac{1}{n} 1_n 1_n{}^{\top}$. $\widehat{\text{HSIC}}_{\text{biased}}$ is a biased estimator which converges to the population HSIC measure when n increases.

A second, unbiased empirical estimator, which exhibits a convergence speed in $\frac{1}{\sqrt{n}}$, better than that of $\widehat{\text{HSIC}}_{\text{biased}}$, was developed by Song et al. (2007):

$$\begin{aligned} \widehat{\text{HSIC}}_{\text{unbiased}}(X, Y) = \frac{1}{n(n-3)} \Bigg[& \operatorname{trace}(\underline{K}\, \underline{L}) \\ & + \frac{1_n^T \underline{K} 1_n\, 1_n^T \underline{L} 1_n}{(n-1)(n-2)} - \frac{2}{n-2} 1_n^T \underline{K}\, \underline{L} 1_n \Bigg] , \end{aligned} \tag{3.5}$$

where $\underline{K} = K - \mathrm{diag}(K)$ and $\underline{L} = L - \mathrm{diag}(L)$.

Both empirical HSIC estimators fit in our general family of association scores:

Lemma 3.1. *The function*

$$s(K, Y) = \widehat{\mathrm{HSIC}}(K, YY^\top),$$

where $\widehat{\mathrm{HSIC}}$ *is either the biased estimator ((3.4)) or the unbiased one ((3.5)), is a quadratic kernel association score. In addition, the biased estimator is a positive definite quadratic form on* Y *for any kernel* K*.*

Proof. For the biased estimator ((3.4)), we simply rewrite it as

$$\widehat{\mathrm{HSIC}}_{\mathrm{biased}}(K, YY^\top) = \frac{1}{(n-1)^2} Y^\top \Pi_n K \Pi_n Y,$$

which is a positive quadratic form in Y, corresponding to the hat matrix $K^{1/2}\Pi_n/(n-1)$. For the unbiased estimate, the derivation is also simple but a bit tedious, and is postponed to Appendix B.1. □

We highlight that this result is fundamentally different from the results of Yamada et al. (2018), who show that, asymptotically, the empirical block estimator of HSIC (Zhang et al., 2018) has a Gaussian distribution. Here we do not focus on the value of the empirical HSIC estimator itself, but on its dependence on Y, which will be helpful later to derive PSI schemes. We also note that Lemma 3.1 explicitly requires that the kernel L used to model outcomes be the linear kernel, while the approach of Yamada et al. (2018) that leads to a more specific PSI schemes is applicable to any kernel L.

3.4 Kernel Selection

Given any quadratic kernel association score, we now detail different strategies to select a subset of $S' \leq S$ of kernels among the initial set $\mathcal{K}$. We consider three standards strategies, assuming S' is given:

- *Filtering*: we compute the scores $s(K, Y)$ for all candidate kernels $K \in \mathcal{K}$, and select among them the top S' with the highest scores.
- *Forward stepwise selection*: we start from an empty list of kernels, and iteratively add new kernels one by one in the list by picking the one that leads to the largest increase in association score when combined with the kernels already in the list. This is formalized in Algorithm 3.1.
- *Backward stepwise selection*: we start from the full list of kernels, and iteratively remove the one that leads to the smallest decrease in association score, as formalized in Algorithm 3.2.

In addition, we consider *adaptive* variants of these selection methods, where the number S' of selected kernels is not fixed beforehand but automatically selected in a data-driven way. In adaptive estimation of S', we maximize over S' the association score computed at each step, potentially regularized by a penalty function that does not depend on Y. For example, for group selection in the linear regression case, Loftus & Taylor (2015) maximize the association score penalized by an AIC penalty.

Algorithm 3.1 Forward stepwise kernel selection

Input: set of kernels $\mathcal{K} = \{K_1, \ldots, K_S\}$; outcome $Y \in \mathbb{R}^n$; quadratic kernel association score $s(.,.)$; number of kernels to select $S' \leq S$.
Output: a subset of S' selected kernels.
Init: $\mathcal{I} \leftarrow \mathcal{K}$, $\mathcal{J} \leftarrow \emptyset$.
for $i = 1$ to S' **do**
  $\widetilde{K} \leftarrow \underset{K \in \mathcal{I}}{\operatorname{argmax}} \quad s\left(K + \sum_{K' \in \mathcal{J}} K', Y\right)$
  $\mathcal{I} \leftarrow \mathcal{I} \setminus \{\widetilde{K}\}$
  $\mathcal{J} \leftarrow \mathcal{J} \cup \{\widetilde{K}\}$
return $\mathcal{J}$

Algorithm 3.2 Backward stepwise kernel selection

Input: set of kernels $\mathcal{K} = \{K_1, \ldots, K_S\}$; outcome $Y \in \mathbb{R}^n$; quadratic kernel association score $s(.,.)$; number of kernels to select $S' \leq S$.
Output: a subset of S' selected kernels.
Init: $\mathcal{J} \leftarrow \mathcal{K}$.
for $i = 1$ to $S - S'$ **do**
  $\widetilde{K} \leftarrow \underset{K \in \mathcal{J}}{\operatorname{argmax}} \quad s\left(\sum_{K' \in \mathcal{J} \setminus \{K\}} K', Y\right)$
  $\mathcal{J} \leftarrow \mathcal{J} \setminus \{\widetilde{K}\}$
return $\mathcal{J}$

The following result generalizes to the kernel selection problem a result that was proven by Loftus & Taylor (2015) in the feature group selection problem with linear methods.

Theorem 3.1. *Given a set of kernels $\mathcal{K} = \{K_1, \ldots, K_S\}$, a quadratic kernel association score s, and a method for kernel selection discussed above (filtering, forward or backward stepwise selection, adaptive or not), let $\widehat{M}(Y) \subseteq \mathcal{K}$ be the subset of kernels selected given a vector of outcomes $Y \in \mathbb{R}^n$. For any $M \subseteq \mathcal{K}$, there exists $i_M \in \mathbb{N}$, and $(Q_{M,1}, b_{M,1}), \ldots, (Q_{M,i_M}, b_{M,i_M}) \in \mathbb{R}^{n \times n} \times \mathbb{R}$ such that*

$$\{Y \,:\, \widehat{M}(Y) = M\} = \bigcap_{i=1}^{i_M} \{Y \,:\, Y^\top Q_{M,i} Y + b_{M,i} \geq 0\}.$$

Again, the proof is simple but tedious, and is postponed to Appendix B.2. Theorem 3.1 shows that, for a large class of selection methods, we can characterize the set of outcomes Y that lead to the selection of any particular subset of kernels as conjunction of quadratic inequalities. This paves the way to a variety of PSI schemes by conditioning of the event $\widehat{M}(Y) = M$, as explored for example by Loftus & Taylor (2015); Yang et al. (2016a) in the case of group selection.

It is worth noting that Theorem 3.1 is valid in particular when an empirical HSIC estimator is used to select kernels, thanks to Lemma 3.1. In our setting, the kernel selection procedure proposed by Yamada et al. (2018) corresponds precisely to the filtering selection strategy combined with an empirical HSIC estimator. Hence Theorem 3.1 allows to derive an exact characterization of the event $\widehat{M}(Y) = M$ in terms of Y, which in turns allows to derive various PSI procedure involving Y, as detailed below. In contrast, Yamada et al. (2018) provide a characterization of the event $\widehat{M}(Y) = M$ not in terms of Y, but in terms of the vector of values $(s(K_i, Y))_{i=1,\dots,S}$. Combined with the approximation that this vector is asymptotically Gaussian when n tends to infinity, this allows Yamada et al. (2018) to derive PSI schemes to assess the values $s(K_i, Y)$ of the selected kernel. Theorem 3.1 therefore provides a result which is valid non-asymptotically, and which allows to test other types of hypotheses, such as the association of one particular kernel with the outcome, given other selected kernels.

3.5 Statistical Inference

Let us consider the general model

$$Y = \mu + \sigma^2 \varepsilon \,, \tag{3.6}$$

where $\varepsilon \sim \mathcal{N}(0, I_n)$ and $\mu \in \mathbb{R}^n$. Characterizing the set $E = \{Y : \widehat{M}(Y) = M\}$ allows to answer a variety of statistical inference questions about the true signal μ and its association with the different kernels, conditional to the fact that a given set of kernels M has been selected.

For example, testing whether $s(K, \mu) = 0$ for a given kernel $K \in M$, or for the combination of kernels $K = \sum_{K' \in M} K'$, is a way to assess whether K captures information about μ. This is the test carried out by Yamada et al. (2018) to test each individual kernel after selection by marginal HSIC screening. Alternatively, to test whether a given kernel $K \in M$ has information about μ not redundant with the other selected kernels in $M \setminus \{K\}$, one may test whether the prototype of μ built from all kernels in M is significantly better that the prototype built without K. This can translate into testing whether

$$s\left(\sum_{K' \in M} K', \mu\right) = s\left(\sum_{K' \in M, K' \neq K} K', \mu\right) .$$

Such a test is performed by Loftus & Taylor (2015); Yang et al. (2016a) to assess the significance of groups of features in the linear setting, using the projection prototype.

In general, testing a null hypothesis of the form $s(K, \mu) = 0$ for a positive quadratic form s can be done by forming the statistics $V = \|H(K)Y\|^2$, where H is the hat matrix associated with s, and studying its distribution conditionally on the event $Y \in E$. The fact that E is an intersection of subsets defined by quadratic constraints can be exploited to derive computationally efficient procedures to estimate p-values and confidence intervals when, for example, $H(K)$ is a projection onto a subspace (Loftus & Taylor, 2015; Yang et al., 2016a). We can directly borrow these techniques in our setting, for example for the KPCR prototype, where $H(K)$ is a projection matrix. For more general $H(K)$ matrices, the techniques of Loftus & Taylor (2015); Yang et al. (2016a) need to be adapted; another way to proceed is to estimate the distribution of V by Monte-Carlo sampling, as explained in the next section.

Alternatively, Reid et al. (2017) propose to test the significance of groups of features through prototypes, which they argue uses fewer degrees of freedom than statistics based on the norms of prototypes, which can increase statistical power. We adapt this idea to the case of kernels and show here how to test the association of a single kernel (whether one of the selected kernels, or their aggregation) with the outcome. We refer the reader to Reid et al. (2017) for extensions to several groups, that can be easily adapted to several kernels. Given a prototype $\widehat{Y} = H(K)Y$, Reid et al. (2017) propose to test the null hypothesis $H_0 : \theta = 0$ in the following univariate model:

$$Y = \mu + \theta\widehat{Y} + \sigma^2\varepsilon,$$

where again $\varepsilon \sim \mathcal{N}(0, I_n)$, μ is fixed, and θ is the parameter of interest. One easily derives the log-likelihood:

$$\ell_Y(\theta) = \log|I - \theta H(K)| - \frac{1}{2\sigma^2}\|Y - \mu - \theta H(K)Y\|^2,$$

which is a concave function of θ that can be maximized by Newton-Raphson iterations to obtain the maximum likelihood estimator $\widehat{\theta} \in \operatorname{argmax}_\theta \ell_Y(\theta)$. We can then form the likelihood ratio statistics

$$R(Y) = 2\left(\ell_Y(\widehat{\theta}) - \ell_Y(0)\right), \tag{3.7}$$

and study the distribution of $R(Y)$ under H_0 to perform a statistical test and derive a p-value. While $R(Y)$ asymptotically follows a χ^2_1 distribution under H_0 when we do not condition on Y (Reid et al., 2017), its distribution conditioned on the event $\widehat{M}(Y) = M$ is different and must be determined for valid PSI. As this conditional distribution is unlikely to be tractable, we propose to approximate it thanks to empirical sampling. This allows us to derive valid empirical PSI p-values as the fraction of samples Y_t for which $R(Y_t)$ is larger than the $R(Y)$ computed from the data.

3.6 Constrained Sampling

We now discuss how to sample T replicates $Y_1, \ldots, Y_T$ according to the Gaussian model ((3.6)) conditional to the event $\widehat{M}(Y) = M$. As explained in the previous section, this is needed to derive p-values for various statistical tests.

By Theorem 3.1, all replicates must be sampled within the acceptance region defined by a series of quadratic constraints on Y. Several strategies can be deployed to this end. The most straightforward one is rejection sampling, which consists in sampling independently Y_t from $\mathcal{N}(\mu, \sigma^2 I_n)$, and only retaining samples for which all quadratic constraints are satisfied, i.e., $Y_t^T Q_{M,i} Y_t + b_{M,i} \geq 0$, for $i \in \{1, \cdots, i_M\}$. Such a strategy can be time-consuming, especially if the volume of the acceptance region is small, leading to a high number of rejections. Alternatively, one could use the the Hamiltonian Monte Carlo algorithm of Pakman & Paninski (2014). In practice, we found that for large values of n, it does not scale well enough to generate a sufficient number of replicates T. Therefore, we propose a new hit-and run sampler below.

Our proposed sampler is based on the Hypersphere Directions (HD) algorithm, first proposed by Berbee et al. (1987) to detect nonredundant constraints in a system of linear inequalities. The main assumption in the HD algorithm is that the acceptance region is open and bounded. In our case, the boundedness assumption does not necessarily hold. For example, if $b_{M,i} = 0$ for all $i = 1, \ldots, i_M$, then the acceptance region is clearly an unbounded cone, that is, if $Y \in E$ then $\lambda Y \in E$ for any $\lambda \geq 0$. To use the HD algorithm nevertheless, we apply the reparametrization $Z = F(Y)$, where $F : \mathbb{R}^n \to]0,1[^n$ is given by $F(Y)_i = F_{\mu_i,\sigma^2}(Y_i)$ for $i = 1, \ldots, n$. Here $F_{\mu_i,\sigma^2}(Y_i)$ denotes the cumulative distribution function (c.d.f.) of the normal distribution $\mathcal{N}(\mu_i, \sigma^2)$. Without conditioning, Z is uniformly distributed over $]0,1[^n$, and when we condition on $Y \in E$, Z is uniformly distributed on the truncated space region $\mathcal{M}$ given by the quadratic constraints:

$$F^{-1}(Z) Q_{M,i} F^{-1}(Z) + b_{M,i} > 0, \ \forall i \in \{1, \cdots, i_M\}.$$

We use strict inequalities so that $\mathcal{M}$ is both open and bounded; this does not affect the probabilities we estimate.

Algorithm 3.3 presents our hit-and-run sampler (Bélisle et al., 1993), based on iteratively sampling in the hypercube. In the HD algorithm, the unidimensional parameter λ_t is sampled according to the p.d. $f_t^{\lambda}(\lambda_t | Z_{t-1}, \theta_t) \propto f(Z_{t-1} + \lambda_t \theta_t)$, where f is the p.d. of $Z = F(Y)$. Given that Z is uniformly distributed on $\mathcal{M}' =]0,1[^n \cap \mathcal{M}$, λ_t is then uniformly distributed on the region $\Lambda = \{\lambda \text{ s.t. } Z_{t-1} + \lambda\theta_t \in \mathcal{M}'\}$. To sample λ_t, we first start by uniformly sampling on the interval $[a_t, b_t]$ to ensure that $Z_{t-1} + \lambda_t \theta_t \in]0,1[^n$. The sample λ_t is accepted if $Z_{t-1} + \lambda\theta_t \in \mathcal{M}$.

Though our sampling of λ_t is also a rejection sampling, the resulting hit-and-run sampler is faster than a mere rejection sampling of Y_t. Indeed, λ_t is unidimensional

[1] A classical technique to uniformly sample from the n-dimensional sphere is to first sample θ_t from $\mathcal{N}(0,1)$ and normalize, $\theta_t \leftarrow \theta_t / ||\theta_t||_2$

Algorithm 3.3 Hypersphere Directions hit-and-run sampler

Input: Y an admissible point, T the total number of replicates and B the number of burn-in iterations.
Output: a sample of T replicates sampled according to the conditional distribution.
Init: $Z_0 \leftarrow F^{-1}(Y)$, $t \leftarrow 0$
repeat
  $t \leftarrow t + 1$
  Sample uniformly θ_t from $\{\theta \in \mathbb{R}^n,\ ||\theta|| = 1\}$ [1]
  $a_t \leftarrow \max\left\{ \max_{\theta_t^{(i)}>0} -\frac{Z_{t-1}}{\theta_t};\ \max_{\theta_t^{(i)}<0} \frac{1-Z_{t-1}}{\theta_t} \right\}$
  $b_t \leftarrow \min\left\{ \min_{\theta_t^{(i)}<0} -\frac{Z_{t-1}}{\theta_t};\ \min_{\theta_t^{(i)}>0} \frac{1-Z_{t-1}}{\theta_t} \right\}$
  repeat
    Sample uniformly λ_t from $]a_t, b_t[$
    $Z_t \leftarrow Z_{t-1} + \lambda_t \theta_t$
    $Y_t \leftarrow F^{-1}(Z_t)$
  until $Z_t \in \mathcal{M}$
until $t = B + T$
return $\{Y_{B+1}, \cdots, Y_{B+T}\}$

while each replicate Y_t is an n-dimensional normal variable. Moreover, the initial sampling on the interval $]a_t, b_t[$ reduces the total number of rejections.For a proof of the convergence of the HD sampler, we refer the reader to Smith (1984).

In hit-and-run samplers, to generate valid p-values, a large number of burn-in iterations and of replicates are needed. The burn-in period reduces the dependence on the original sample Y, while the large number of replicates addresses the correlation between consecutive replicates.

3.7 Experiments

In our experiments, we focus on the case where each kernel corresponds to a predefined group of features, and where we test the association of the sum of the selected kernels with the outcome. We use $\widehat{\text{HSIC}}_{\text{unbiased}}$ as a quadratic kernel association score for kernel selection in all our experiments.

3.7.1 Statistical Validity

We first demonstrate the statistical validity of our PSI procedure, which we refer to as kernelPSI. We simulate a design matrix X of $n = 100$ samples and $p = 50$ features, partitioned in $S = 10$ disjoint and mutually-independent subgroups of $p' = 5$ features, drawn from a normal distribution centered at 0 and with a covariance

matrix $V_{ij} = \rho^{|i-j|}, \quad i, j \in \{1, \cdots p'\}$. We set the correlation parameter ρ to 0.6. To each group corresponds a *local* Gaussian kernel K_i, of variance $\sigma^2 = 5$.

The outcome Y is drawn as $Y = \theta K_{1:3} U_1 + \varepsilon$, where $K_{1:3} = K_1 + K_2 + K_3$, U_1 is the eigenvector corresponding to the largest eigenvalue of $K_{1:3}$, and ε is Gaussian noise centered at 0. We vary the effect size of θ across the range $\theta \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$, and resample Y 1 000 times to create 1 000 simulations.

In this particular setting where the *local* kernels are additively combined, the three kernel selection strategies in Section 3.4 are equivalent. Along with the *adaptive* variant, we consider 3 variants with a predetermined number of kernels, $S' \in \{1, 3, 5\}$. For inference, we compute the likelihood ratio statistics for KPCR or KRR prototypes, or directly use $\widehat{\text{HSIC}}_{\text{unbiased}}$ as a test statistic (see Section 3.5). Finally, we used our hit-and-run sampler to provide empirical p-values (see Section 3.6), fixing the number of replicates at $T = 5 \times 10^4$ and the number of burn-in iterations at 10^4.

Figure 3.1 shows the Q-Q plot comparing the distribution of the p-values provided by kernelPSI with the uniform distribution, under the null hypothesis ($\theta = 0.0$). All variants give data points aligned with the first diagonal, confirming that the empirical distributions of the statistics are uniform under the null.

Figure 3.2 shows the Q-Q plot comparing the distribution of the p-values provided by kernelPSI with the uniform distribution, under the alternative hypothesis where $\theta = 0.3$. We now expect the p-values to deviate from the uniform. We observe that all kernelPSI variants have statistical power, reflected by low p-values and data points located towards the bottom right of the Q-Q plot. The three strategies (KPCR, KRR and HSIC) enjoy greater statistical power for smaller number of selected kernels. Because of the selection of irrelevant kernels, statistical power decreases when S' increases. The same remark holds for the adaptive variants, which performs similarly to the fixed variant with $S' = 5$. In fact, the average support size for the adaptive kernel selection procedure is $\overline{S'} = 5.05$. We also observe that HSIC has more statistical power than the KRR or KPCR variants, possibly because we used an HSIC estimator for kernel selection, making the inference step closer to the selection one.

3.7.2 Benchmarking

We now evaluate the performance of the kernelPSI procedure against a number of alternatives:

- *protoLasso:* the original, linear prototype method for post-selection inference with L_1-penalized regression (Reid et al., 2017);
- *protoOLS:* a selection-free alternative, where the prototype is obtained from an ordinary least-squares regression, and all variables are retained;
- *protoF:* a classical goodness-of-fit F-test. Here the prototype is constructed similarly as in protoOLS, but the test statistic is an F-statistic rather that a likelihood ratio;

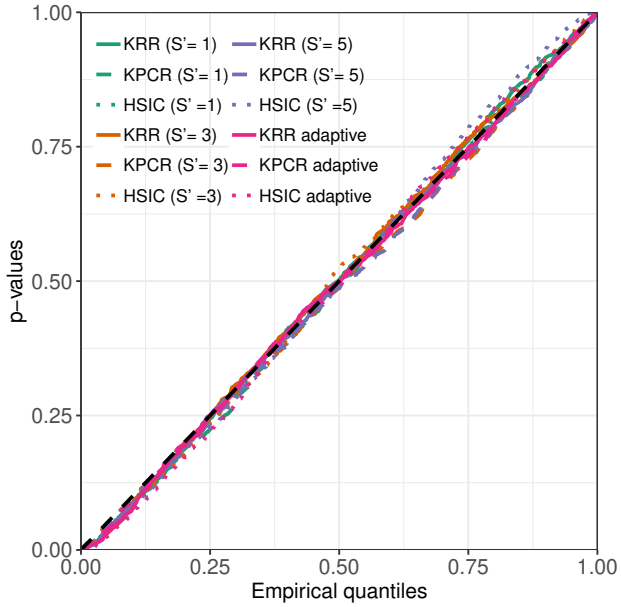


Figure 3.1: Q-Q plot comparing the empirical kernelPSI p-values distributions under the null hypothesis ($\theta = 0.0$) to the uniform distribution.

- *KPCR, KRR, and HSIC:* the non-selective alternatives to our kernelPSI procedure. KPCR and KRR are obtained by constructing a prototype over the sum of all kernels, without the selection step. HSIC is the independence test proposed by Gretton et al. (2008);

- *SKAT:* The Sequence Kernel Association Test (Wu et al., 2011) tests for the significance of the joint effect of all kernels in a non-selective manner, using a quadratic form of the residuals of the null model.

We consider the same setting as in Section 3.7.1, but now add benchmark methods and additionally consider linear kernels over binary features, a setting motivated by the application to genome-wide association studies, where the features are discrete. In this last setting, we vary the effect size θ over the range $\{0.01, 0.02, 0.03, 0.05, 0.07, 0.1\}$. We relegate to Appendix B.3.4.2 an experiment with Gaussian kernels over Swiss roll data.

Figures 3.3 and 3.4 show the evolution of the statistical power as a function of the effect size θ in, respectively, the Gaussian and the linear data setups. These figures confirm that kernel-based methods, particularly selective HSIC and SKAT, are superior to linear ones such as protoLASSO. We observe once more that the selective HSIC variants have more statistical power than their KRR or KPCR counterparts, that methods selecting fewer kernels enjoy more statistical power, and that adaptive methods tend to select too many kernels (closer to $S' = 5$ than to the true $S' = 3$). We also observe that the selective kernelPSI methods ($S' = 1, 3, 5$ or adaptive) have more statistical power than their non-selective counterparts.

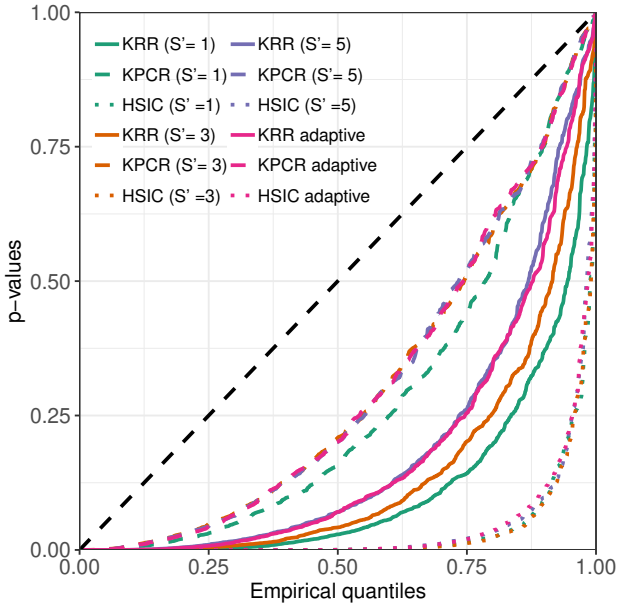


Figure 3.2: Q-Q plot comparing the empirical kernelPSI p-values distributions under the alternative hypothesis ($\theta = 0.3$) to the uniform distribution.

Finally, we note that, in the linear setting, the KRR and KPCR variants perform similarly. We encounter a similar behavior in simulations (not shown) using a Wishart kernel. Depending on the eigenvalues of K, the spectrum of the transfer matrix $H_{\text{KRR}} = K(K + \lambda I_{n\times n})^{-1}$ can be concentrated around 0 and 1. H_{KRR} becomes akin to a projector matrix, and KRR behaves similarly to KPCR.

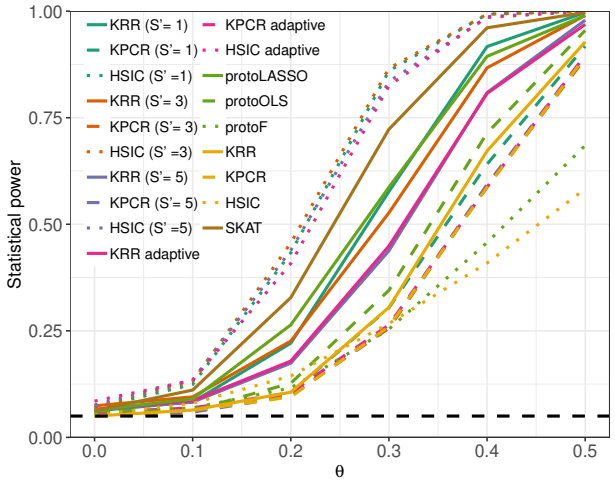


Figure 3.3: Statistical power of kernelPSI variants and benchmark methods, using Gaussian kernels for simulated Gaussian data.

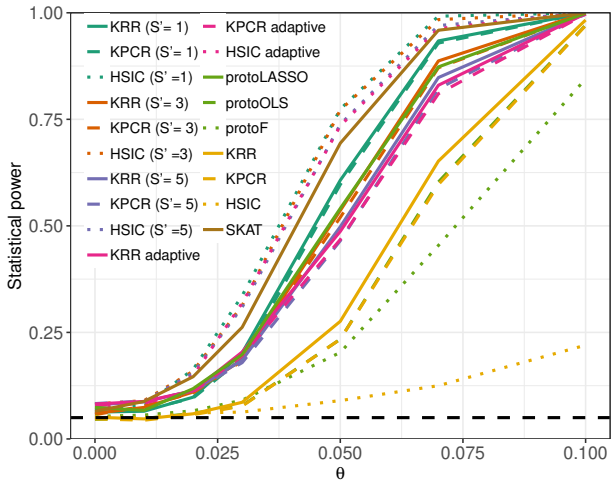


Figure 3.4: Statistical power of kernelPSI variants and benchmark methods, using linear kernels for simulated binary data.

In addition, we evaluate the ability of our kernel selection procedure to recover the three true causal kernels used to simulate the data. Table 3.1 reports the evolution of the precision and recall of our procedures, in terms of selected kernels, for increasing effect sizes in the Gaussian kernels and data setting. Note that when S' is fixed, a random selection method is expected to have a precision of 3/10 (the proportion of kernels that are causal), and a recall of $S'/10$, which corresponds to the values we obtain when there is no signal ($\theta = 0$). As the effect size θ increases, both precision and recall increase.

When S' increases, the precision increases and the recall decreases, which is consistent with our previous observations that increasing S' increases the likelihood to include irrelevant kernels in the selection. Once again, the performance of the adaptive kernelPSI is close to that of the setting where the number of kernels to select is fixed to 5, indicating that the adaptive version tends to select too many kernels.

3.7.3 Case Study: Selecting Genes in a Genome-Wide Association Study

In this section, we illustrate the application of kernelPSI on genome-wide association study (GWAS) data. Here we study the flowering time phenotype "FT_GH" of the *Arabidopsis thaliana* dataset of Atwell et al. (2010). We are interested in using the 166 available samples to test the association of each of 174 candidate genes to this phenotype. Each gene is represented by the single-nucleotide polymorphisms (SNPs) located within ± 20-kilobases. We use hierarchical clustering to create groups of SNPs within each gene; these clusters are expected to correspond to linkage disequilibrium blocks. As is common for GWAS applications, we

Table 3.1: Ability of the kernel selection procedure to recover the true causal kernels, using Gaussian kernels over simulated Gaussian data.

	$\boldsymbol{\theta}$	$S' = 1$	$S' = 3$	$S' = 5$	Adaptive
Recall	0.0	0.102	0.302	0.505	0.435
	0.1	0.150	0.380	0.569	0.523
	0.2	0.263	0.528	0.690	0.678
	0.3	0.324	0.630	0.770	0.768
	0.4	0.332	0.691	0.830	0.822
	0.5	0.333	0.733	0.862	0.855
Precision	0.0	0.306	0.302	0.303	0.305
	0.1	0.450	0.380	0.341	0.352
	0.2	0.791	0.528	0.414	0.437
	0.3	0.974	0.630	0.462	0.485
	0.4	0.997	0.691	0.498	0.518
	0.5	1.000	0.733	0.517	0.548

use the identical-by-state (IBS) kernel (Kwee et al., 2008) to create one kernel by group. We then apply our kernelPSI variants as well as the baseline algorithms used in Section 3.7.2. Further details about our experimental protocol are available in Appendix B.3.6.

We first compare the p-values obtained by the different methods using Kendall's tau coefficient τ to measure the rank correlation between each pair of methods (see Appendix B.3.7). All coefficients are positive, suggesting a relative agreement between the methods. We also resort to non-metric multi-dimensional scaling (NMDS) to visualize the concordance between the methods (see Appendix B.3.9). Altogether, we observe that related methods are located nearby (e.g. KRR near KPCR, protoLASSO near protoOLS, etc.), while selective methods are far away from non-selective ones.

Our first observation is that none of the non-selective methods finds any gene significantly associated with the phenotype ($p < 0.05$ after Bonferroni correction), while our proposed selective methods do. A full list of genes detected by each method is available in Appendix B.3.8. None of those genes have been associated to this phenotype by traditional GWAS (Atwell et al., 2010). We expect the most conservative methods ($S' = 1$) to yield the fewest false positive, and hence focus on those. KRR, KPCR and HSIC find, respectively, 2, 2, and 1 significant genes. One of those, AT5G57360, is detected by all three methods. It is interesting to note that this gene has been previously associated with a very related phenotype, FT10, differing from ours only in the greenhouse temperature (10°C vs 16°C). This is also the case of the other gene detected by KRR, AT5G65060. Finally, the second gene detected by KPCR, AT4G00650, is the well-known FRI gene, which codes for the FRIGIDA protein, required for the regulation of flowering time in late-flowering

phenotypes. All in all, these results indicate that our proposed kernelPSI methods have the power to detect relevant genes in GWAS and are complementary to existing approaches.

3.8 Conclusion

We have proposed kernelPSI, a general framework for post-selection inference with kernels. Our framework rests upon quadratic kernel association scores to measure the association between a given kernel and the outcome. The flexibility in the choice of the kernel allows us to accommodate a broad range of statistics. Conditionally on the kernel selection event, the significance of the association with the outcome of a single kernel, or of a combination of kernels, can be tested. We demonstrated the merits of our approach on both synthetic and real data. In addition to its ability to select causal kernels, kernelPSI enjoys greater statistical power than state-of-the-art techniques. A future direction of our work is to scale kernelPSI to larger datasets, in particular with applications to full GWAS data sets in mind, for example by using the block HSIC estimator (Zhang et al., 2018) to reduce the complexity in the number of samples. Another direction would be to explore whether our framework can also incorporate Multiple Kernel Learning (Bach, 2008). This would allow us to complement our filtering and wrapper kernel selection strategies with an embedded strategy, and to construct an aggregated kernel prototype in a more directly data-driven fashion.

Chapter 4

A systematic analysis of gene-gene epistasis in multiple sclerosis pathways

---

Publication and Dissemination: *in preparation. This is joint work with Clément Chatelain (SANOFI R&D), Hélène de Foucauld (SANOFI R&D) and Chloé-Agathe Azencott (Mines ParisTech).*

Abstract: *Multiple sclerosis is a complex autoimmune disease which genetic basis has been extensively investigated through genome wide association studies. So far, the conducted studies have detected a number of loci independently associated with the disease but few have investigated the interaction between distant loci, or epistasis. In this chapter, we perform a gene level epistasis analysis of multiple sclerosis GWAS from the Wellcome Trust Case Control Consortium 2. We systematically study the epistatic interactions between all pairs of genes within* 19 *multiple sclerosis disease maps from the MetaCore pathway database. We report 4 gene pairs with epistasis involving missense variants, and 117 gene pairs with epistasis mediated by eQTLs. Our epistasis analysis is able to retrieve known interactions linked to multiple sclerosis: direct binding interaction between GLI-I and SUFU, involved in oligodendrocyte precursor cells differentiation, and regulation of IP10 transcription by NF-κB, thus validating the potential of epistasis analysis to reveal biological interaction with relevance in a disease specific context.*

Résumé : *La sclérose en plaques est une maladie auto-immune complexe dont la base génétique a été largement étudiée par des études d'association à l'échelle du génome. Jusqu'à présent, les études menées ont détecté un certain nombre de loci associés indépendamment à la maladie, mais peu ont étudié l'interaction entre des loci distants, ou épistasie. Dans ce chapitre, nous effectuons une analyse de l'épistasie au niveau des gènes sur le GWAS de la sclérose en plaques du Wellcome Trust Case Control Consortium 2. Nous étudions systématiquement les interactions*

épistatiques entre toutes les paires de gènes dans les 19 *cartes de la sclérose en plaques de la base de données MetaCore. Nous rapportons 4 paires de gènes avec une épistasie impliquant des variants faux-sens, et 117 paires de gènes avec une épistasie médiée par des eQTLs. Notre analyse d'épistasie est capable de retrouver des interactions connues liées à la sclérose en plaques: interaction de liaison directe entre GLI-I et SUFU, impliquée dans la différenciation des cellules précurseurs d'oligodendrocytes, et régulation de la transcription IP10 par NF-κB, validant ainsi le potentiel de l'étude d'épistasie pour révéler l'interaction biologique avec pertinence dans un contexte spécifique à la maladie.*

4.1 Introduction

Extensive efforts have been deployed to tackle multiple sclerosis, a chronic disease damaging the central nervous system (Goldenberg, 2012). A number of marketed drugs (Dargahi et al., 2017) attenuate the symptoms of the disease. However, an efficient drug targeting its root causes is still elusive. This is partially due to our limited understanding of the mechanisms governing multiple sclerosis. Several studies demonstrated that heritability is a major component in multiple sclerosis (Dyment, 2006; Dean et al., 2007). The development of GWAS has allowed to explore the genetic causes of this heritability. In GWAS, large cohorts of cases and controls are jointly studied in order to discover new biomarkers and causal loci. In the context of multiple sclerosis, at least fourteen studies (Sawcer et al., 2014) have been put in place in order to develop new hypotheses. So far, hundreds of loci (Baranzini & Oksenberg, 2017; Cotsapas & Mitrovic, 2018) have already been statistically associated with multiple sclerosis. The biology behind some of them (Gregory et al., 2007; Jager et al., 2009; Couturier et al., 2011) has been clarified while for the majority of retained loci, it remains unexplained (Sawcer et al., 2014).

At least two gene-gene interactions have been discovered in multiple sclerosis: high levels of c-Jun may cause enhanced myelinating potential in Fbxw7 Harty et al. (2019) and DDX39B is both a potent activator of IL7R exon6 splicing and a repressor of sIL7R Galarza-Muñoz et al. (2017). An additional tripartite genic interaction has also been reported Lincoln et al. (2009): epistasis between HLA-DRB1, HLA-DQA1, and HLA-DQB1 loci increases multiple sclerosis susceptibility. This further cements the need to study epistasis to understand the genetic basis of multiple sclerosis.

We perform here a selective gene-level analysis of epistasis in multiple sclerosis. The study of epistasis at the gene-level is important because the statistical association at the SNP level might not be strong enough to establish a link between the corresponding genes and the studied disease. We systematically study interactions between pairs of genes contained in 19 multiple sclerosis disease maps from the MetaCore (Ekins et al., 2006) dataset. For this purpose, we apply epiGWAS (see Chapter 2) on the multiple sclerosis GWAS from the Wellcome Trust Case Control Consortium 2 (Sawcer et al., 2011). EpiGWAS was originally developed for SNP-level detection, but we extended here to the gene-level. Our analysis yielded 4 gene pairs with epistasis involving missense variants, and 117 gene pairs with epistasis mediated by eQTLs. Among them, two pairs are already known: direct binding interaction between GLI-I and SUFU, involved in oligodendrocyte precursor cells differentiation, and regulation of IP10 transcription by NF-κB. This confirms the capacity of the statistical study of epistasis to detect biological interactions that further our understanding of disease mechanisms.

4.2 epiGWAS: from the SNP level to the gene level

4.2.1 Detecting SNP-SNP synergies with epiGWAS

In Chapter 2, we have developed epiGWAS, a new framework for targeted epistasis to detect interactions between a given SNP A, which we refer to as the target, and a set of SNPs $X = \{X_1, \cdots, X_p\}$, which can cover either the whole genome or a predetermined region e.g. a gene or a coding region. The output of epiGWAS is a set of interaction scores $\{a_1, \cdots, a_p\}$ between each SNP in the set $X = \{X_1, \cdots, X_p\}$ and the target A. EpiGWAS proposes a family of methods to compute the interaction scores. Among them, we only use the *robust modified outcome* method. In Chapter 2, we have demonstrated its superior performance in comparison with other epistasis detection baselines and the other methods of the modified outcome family.

4.2.2 Gene-level epiGWAS

EpiGWAS can be ran in an exhaustive fashion for each target X_i against the rest of the SNPs $\{X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_p\}$. This procedure generates a list of interaction score vectors. The interpretability and usability of such an output is limited because of the large number of interactions and the different covariates for each target which makes the comparison of the associated scores difficult. For instance, different regularization grids yield different stability curves, and thus, different areas under the curve. Furthermore, despite their robustness, the biological significance of the scores is limited. A first step to improve interpretability is to use rankings. From a practical point of view, rankings are a sensible choice because only the highest-ranking SNPs are used. Rankings also improve comparability between different targets because of the similarity of scale and insensitivity to the underlying parameterization. For a target i, we denote $r_{ij} \in \{1, \cdots, p-1\}$ the rank in a decreasing order of the score of SNP j.

Another immediate benefit of the use of rankings is the possibility of combining different rankings. For example, for two SNPs i and j, we can define the following epistasis interaction score:

$$\text{inter}(i, j) = \frac{1}{r_{ij} + r_{ji}}. \tag{4.1}$$

The interaction score in Eq. (4.1) has the advantages of symmetry and boundedness. The scores take their values in $]0, 1/2]$. Additionally, the combination of two pairwise scores r_{ij} and r_{ji} can help control the estimation errors for one of the targets. For example, if two SNPs i and j are in interaction and the result r_{ij} is not sufficiently high to reflect that, a good ranking of r_{ji} can help compensate that.

We can further aggregate the rankings to detect interactions between genes. More generally, the rankings can be combined to detect interactions between any disjoint sets of SNPs *e.g.* biological pathways, regulatory regions, etc. Let p' be the total number of genes and $\{G_1, \cdots, G_{p'}\}$ the corresponding sets of SNPs such that

$\bigcup_{i=1}^{p'} G_i = [1..p]$. The easiest way to devise an interaction score between two genes i' and j' is to compute the average of all pairwise scores:

$$\text{inter}(G_{i'}, G_{j'}) = \frac{1}{|G_{i'}||G_{j'}|} \sum_{i \in G_{i'}} \sum_{j \in G_{j'}} \frac{1}{r_{ij} + r_{ji}}. \tag{4.2}$$

Thanks to the symmetry of SNP-SNP scores in Eq. (4.1), the gene-gene scores in Eq. (4.2) are symmetric, too. Moreover, the averaging reduces the impact of the size of the genes. In addition to the mean, we can also use the median or the minimum/maximum of all pairwise scores. However, only a single value will be taken into account with the latter strategies. Depending on the implemented regression method, with respect to a target i, the scores, and hence the rankings, of two nearby variants j and j' can be similar because of linkage disequilibrium. This can make the gene-gene scores more robust through the averaging of high nearby rankings. On the other hand, the averaging strategy can be partially biased by the marginal effects of some targets inflating by consequence the interaction scores. Nevertheless, the combination of two rankings in $1/(r_{ij} + r_{ji})$ helps compensate for a low value of either r_{ij} or r_{ji} due to marginal effects.

4.3 Data and experiments

In this section, we describe the data we integrate to perform our systematic gene-gene interaction analysis for MS. For genotypic data, we select the MS dataset from the second release of the Wellcome Trust Case Control Consortium (WTCCC2) et al. (Sawcer et al., 2011). In order to improve statistical power and the downstream biological interpretation, we subset the marker SNPs related to the genes referenced in the MetaCore (Ekins et al., 2006) disease maps for multiple sclerosis. Each gene pair within a disease map is tested for interaction. Within the same disease map, the included genes affect the same MS-related mechanism. Therefore we can use this prior knowledge to evaluate if our method can retrieve known interactions and identify new ones. The SNPs can be mapped to the genes in two different ways:

- Physical mapping: we select all the marker SNPs which positions are within the boundaries of a gene. In this case, we take into account SNPs with an effect on the structure and function of the corresponding protein.
- eQTL-SNP mapping: with the selection of known eQTL SNPs, we study epistasis through the variation in expression of the associated genes in relevant tissues.

4.3.1 Genotypic data

The WTCCC2 study includes 9 772 MS cases and 17 376 controls hailing from 15 different countries. The presence of population structure (see Section 1.4.3), confirmed by a genomic inflation factor (GIF) of 3.72, is poised to lead to inference

issues. To avoid this problem, we only use Caucasian British samples in both cases and controls. The resulting dataset consists of 2048 cases and 5733 controls with a GIF of 1.06 which proves the homogeneity of the dataset. The selected controls come from two distinct cohorts from the UK Blood Services (NBS) and the 1958 British Birth Cohort (58C). The careful reader may notice the important imbalance between the total number of cases and controls which may distort the results. To equalize the field, we randomly subsample controls to obtain a number of controls equal to the number of cases. We also note that we discarded the samples singled out for quality control by the WTCCC.

4.3.2 Variant selection

We give in Table 4.1 the full list of MS disease maps. For ease of reproducibility, we also give the internal ID of the disease maps, as indicated in MetaCore. The number of genes within each map greatly varies. It ranges from 13 genes for DM 3305 to 100 genes (DM 4593). Even for the larger maps, the total number of genes is still low enough to perform exhaustive pairwise analysis for all SNPs mapped to the selected genes. Similarly to sample-wise QC, we first discarded all low quality SNPs designated by the WTCCC2. We then selected SNPs according to the following mappings:

- Physical mapping: corresponds to retrieving all marker SNPs located on a given gene. We use the accompanying R package metabaser (Ishkin, 2019) to first define the boundaries of a given gene, and then subset all SNPs according to their positions, as referenced in dbSNP version 144 (Pagès, 2017).
- eQTL mapping: we use the cis-eQTL dataset from the eQTLGen consortium (Võsa et al., 2018), which provides for each gene a list of significant eQTL-SNPs. The dataset combines 31 684 whole blood samples from 37 cohorts. The reason for this choice is that whole blood composition is affected by MS (Keshari et al., 2016).

For our present study, we chose cis-eQTLs instead of trans-eQTLs because of their higher degree of association to gene expression. The higher association can be attributed to the proximity of the SNPs to the genes: cis-eQTLs are located within 1 Mb from a gene and they often closely map to either the transcription start site or the transcription end site of a gene. The application of a false discovery rate (FDR) of 0.05 resulted in the identification of eQTL-SNPs for 16 989 genes, or approximately 88.3% of all autosomal genes expressed in blood and tested in the cis-eQTL analysis. We restricted ourselves to the genes present in the metaCore disease maps. We observed that the obtained eQTL-mapping datasets were larger than the physical mapping datasets in terms of number of SNPs: the median number of SNPs per disease map is 392 for the physical mapping analysis and 999 for the eQTL-mapping analysis. In Appendix C.1, we give the exact number of SNPs per disease map for each type of mapping. We also included the average number of SNPs per gene for each disease map and for both mappings.

Table 4.1: Titles and internal IDs of MetaCore disease maps related to MS.

internal ID	Title
3302	Notch signaling in oligodendrocyte precursor cell differentiation in multiple sclerosis
3305	SHH signaling in oligodendrocyte precursor cells differentiation in multiple sclerosis
3306	Inhibition of oligodendrocyte precursor cells differentiation by Wnt signaling in multiple sclerosis
4455	Inhibition of remyelination in multiple sclerosis: regulation of cytoskeleton proteins
4593	Axonal degeneration in multiple sclerosis
4693	Role of Thyroid hormone in regulation of oligodendrocyte differentiation in multiple sclerosis
4703	Demyelination in multiple sclerosis
4791	Role of CNTF and LIF in regulation of oligodendrocyte development in multiple sclerosis
4794	Retinoic acid regulation of oligodendrocyte differentiation in multiple sclerosis
4843	Growth factors in regulation of oligodendrocyte precursor cells proliferation in multiple sclerosis
4846	Growth factors in regulation of oligodendrocyte precursor cells survival in multiple sclerosis
4901	Inhibition of remyelination in multiple sclerosis: role of cell-cell and ECM-cell interactions
5199	Cooperative action of IFN-γ and TNF-α on astrocytes in multiple sclerosis
5288	Impaired inhibition of Th17 cell differentiation by IFN-β in multiple sclerosis
5378	Role of IFN-β in the improvement of blood-brain barrier integrity in multiple sclerosis
5398	Role of IFN-β in activation of T cell apoptosis in multiple sclerosis
5518	Role of IFN-β in inhibition of Th1 cell differentiation in multiple sclerosis
5601	IL-2 as a growth factor for T cells in multiple sclerosis
5611	Role of IL-2 in the enhancement of NK cell cytotoxicity in multiple sclerosis

Even though the two analyses are unrelated and use different sets of SNPs, some concordance for the top-scoring genes is to be expected. In fact, for the eQTLGen consortium, Võsa et al. (2018) show that out of 15 317 trait-associated SNPs, 15.2% were in high LD with the lead eQTL SNP showing the strongest association for a cis-eQTL gene. Although the mentioned association is far from perfect, it demonstrates the often-overlooked link between the two analyses.

4.4 Results

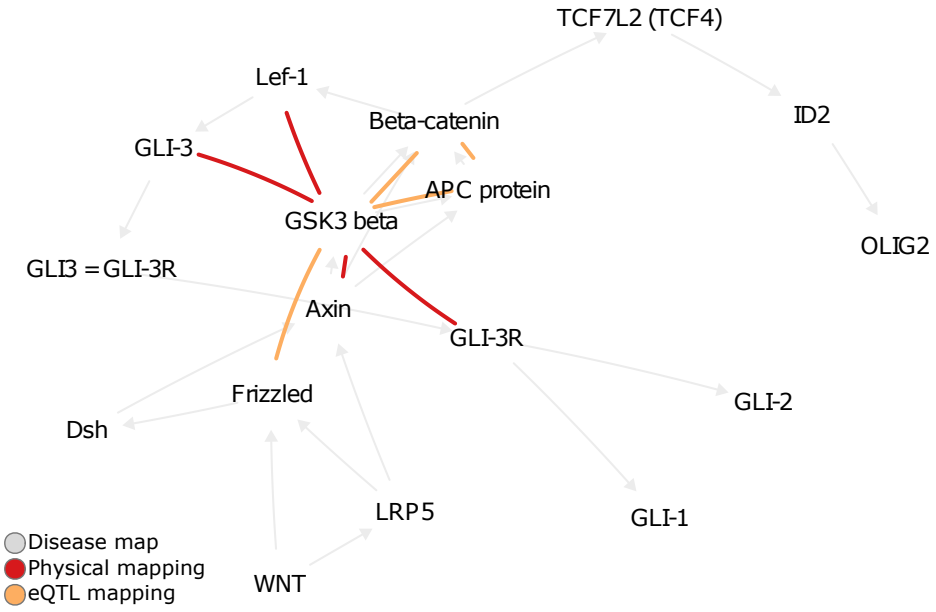


Figure 4.1: The 2% top-scoring pairs in DM 3306 for eQTL and physical mappings.

We exhaustively apply our gene-gene interaction scores in Eq. (4.2) to obtain $p'(p'-1)/2$ interaction scores per disease map, where p' is the number of genes. Given the size of the maps (see Appendix C.1), the interpretation of the full results is rather difficult. We instead focused on the 2% top-scoring pairs for the two analyses. The 2% threshold was manually set with respect to the obtained result. We remark that the top-scoring edges often constitute connected sub-components. We also remark that the obtained sub-components for the eQTL and physical mappings are often interlinked. We further comment on these two remarks in the following paragraphs. We give an illustration of the results in Figure 4.1, in which we plot the obtained subnetworks in addition to the original edges for DM 3306. We relegate the results of the other disease maps to Appendix C.2.

We notice a general consistency of the results between the different disease

maps, which can be formulated through the characteristics below. We also conduct an enrichment analysis, from which we derive empirical p-values to measure the statistical significance of the observed characteristics (see Appendix C.3 for the full results).

- Connectedness: the obtention of connected components for both mappings is the most important aspect of the results. With the exception of DM 3305, 3306 and 4794 consisting of 1 or 2 edges, all disease maps have a p-value lower than 0.05. The p-values were obtained by considering connected components in simulated networks of the same size. Of particular interest are large components because of their significance. In many cases, we obtained an empirical p-value of 0 despite using 10^4 simulations. The discovery of these novel subnetworks can help the understanding of multiple sclerosis by unraveling new disease mechanisms.

- Complementarity: with the exception of DM 4593, the subnetworks of the two mappings are connected *i.e.* they share at least one common node. In fact, they are often connected through multiple nodes without a significant overlap between the edges of the two networks. For instance, they share 5 vertices in DM 4901. In Appendix C.3, we quantify the significance of having 1, 2 or 3 genes in common. The significance values were similarly obtained by considering simulated subnetworks of the same size. We particularly note that 3 edges are in common in DM 3302 for a p-value of 0.038. Therefore, the two types of mappings recover distinct, though connected, interactions, which suggests the complementarity of the two mappings. We can then consider the union of the two subnetworks for further study.

- Centrality: we observed a high degree of connectivity for certain nodes. For example, we mention FAK in DM 4901 ($p_{\text{FAK}} = 0$), SHP-2 in DM 4843 ($p_{\text{SHP-2}} = 0.014$) and TRADD in DM 4843 ($p_{\text{TRADD}} = 0.052$). We attribute this centrality to the existence of important marginal effects that were not completely filtered out. Interestingly, the role of these genes in MS has already been established (Sun et al., 2010; Ahrendsen et al., 2017; Reuss et al., 2014).

- Commonality: despite using the top 2% of all $p'(p'-1)/2$ possible edges for each disease map, some of the retained edges were already present in the original disease maps. In at least 9 out of 19 disease maps, a single edge already exists in the original disease map, and in at least four of them two edges. In DM 3306, we even recover three edges ($p = 0.099$). Nonetheless, drawing conclusions about the underlying biology is challenging given the potential mismatch between biological epistasis and statistical epistasis (Moore & Williams, 2005).

4.4.1 Enrichment analysis for obtained subnetworks

Beyond the validation with existing edges, the main goal of the systematic analysis we conduct here is to discover novel gene-gene interactions in multiple sclerosis. Their biological validation requires laboratory experiments to confirm the observed statistical synergy. As we do not have access to such facilities, we use the enrichment of the recovered networks in terms of existing therapeutic targets as a validation metric. The chosen metric can be criticized in two ways: it is biased in the sense that therapeutic targets only reflect our current understanding of the disease and the existence of effective molecules for the targets. In addition, the targets were often selected on an univariate basis, while the subject of the current study are epistatic interactions. However, an enrichment analysis in terms of therapeutic targets has the advantages of being a trustworthy background thanks to the proven effect of the included genes and its relevance in terms of development of future therapies. For instance, combination therapies if an existing therapeutic target is shown to be interacting with another gene within the recovered subnetworks. Moreover, in light of the new FDA guidance for the co-development of two or more drugs [1], our study pipeline can be of special interest because of its focus on synergistic effects instead of separate additive effects.

In our case, we use OpenTargets (Carvalho-Silva et al., 2018a) as a dataset for therapeutic targets. The dataset is a collaborative effort to create an up-to-date and comprehensive repository to link genomic information of drug targets to a disease of interest. The enrichment analysis studies the overpresence of OpenTargets targets in the obtained networks in comparison with the original disease maps. We use for this matter a classical hypergeometric test (Rivals et al., 2006) to determine the statistical significance of their overpresence. We give the resulting p-values in Appendix C.4. For twelve disease maps, we found at least one common gene between our subnetworks and OpenTargets. Given a significance threshold of 0.05, we found two significant disease maps DM 4593 and DM 5378 with respective p-values of 0.008 and 0.02. The enriched subnetworks require further investigation, especially to study the links within the known targets and between the known targets and the rest of the subnetwork.

4.4.2 Directionality of the synergy

As shown before, our gene-level pipeline with epiGWAS robustly detects the presence of epistatic synergies between two genes. However, the obtained interaction scores do not allow to determine the directionality of the synergy. The synergy can be either positive or negative by respectively increasing or decreasing the disease risk probability. We can nonetheless get a partial answer by studying the nature of interaction between the top-scoring SNPs for each gene pair. We only selected the top-scoring pair because of its disproportionate impact on the corresponding gene-gene score. For example, we can consider the extreme case where for a pair of SNPs

[1] available for download from https://www.fda.gov/media/80100/download

(i, j), we have $r_{ij} = r_{ji} = 1$. The next possible best scoring pair is $r_{i'j'} = r_{j'i'} = 2$ and it further decreases in a hyperbolic manner for the lower rank pairs. So, in the best cases, the top pair will be at least twice as important as the following one.

The direction of the synergy between two uni-dimensional variables can be studied in various ways (VanderWeele & Knol, 2014). In particular, for a binary outcome Y and two variables X_1 and X_2, we can study the sign of the interaction coefficient α_{12} in the following logistic model: $\text{logit}\, P(Y|X_1, X_2) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_{12} X_1 X_2$. Logistic models are widely used for the study of epistasis. For the physical mapping strategy, we conduct a similar analysis. As for the eQTL mapping strategy, the methodology we use for physical mapping can be refined to amount to the desired gene-level interactions. The effect of a SNP i on the expression level e_i of the corresponding gene G_i can be examined through a model of the form $e_i = \gamma_i + \beta_i X_i$. The directionality of the synergy can be deduced from the sign of the following ratio:

$$\text{dir}(G_1, G_2) = \text{sign}\, \frac{\alpha_{12}}{\beta_1 \cdot \beta_2} \tag{4.3}$$

To get a better grasp of the meaning of the score in Eq. (4.3), it suffices to replace the two linear expression models directly in the interaction logistic model. Precisely, we obtain:

$$\begin{aligned} \text{logit}\, P(Y|X_1, X_2) = \alpha_0 &+ \alpha_1 \frac{e_1 - \gamma_1}{\beta_1} + \alpha_2 \frac{e_2 - \gamma_2}{\beta_2} \\ &+ \frac{\alpha_{12}}{\beta_1 \cdot \beta_2} (e_1 - \gamma_1)(e_2 - \gamma_2) \end{aligned} \tag{4.4}$$

The synergy of the two gene expressions is given by the coefficient $\alpha_{12}/(\beta_1 \cdot \beta_2)$ which sign determines the directionality of the epistatic interactions between the two genes. To the best of our knowledge, this is the first study which studies epistasis from such a perspective by including eQTL scores in this way and by moving back and forth between SNP-level and gene-level epistasis. Furthermore, the synergy score in Eq. (4.3) can also be interpreted as an extension of Mendelian randomization (Davies et al., 2018) to second-order interaction effects.

The eQTLGen consortium (Võsa et al., 2018) does not directly supply the effect sizes β_1 and β_2 in the linear expression models. For each SNP, the effect size β is derived from the corresponding Z-score using the following relationship:

$$\beta = \frac{Z}{\sqrt{2\, q(1-q)\, (m + Z^2)}}, \tag{4.5}$$

where q is the MAF of the SNP of interest, as reported in the 1kG v1p3 ALL reference panel and m is the cohort size.

For the significant interactions, we provide a csv file containing the list of coefficients α_{12} in addition to (m_1, q_1, Z_1), (m_2, q_2, Z_2) and the directionality of the synergy $\text{dir}(G_1, G_2) \in \{-1, +1\}$ for the eQTL strategy. One possible approach to appraise the results is to consider a number of summary statistics to get an overview

of the kind of synergies occurring within biological pathways. Interestingly, for all SNP pairs, the interaction coefficient α_{12} is positive in 47% of all cases and the directionality of the synergy $\text{dir}(G_1, G_2)$ is equally split between positive and negative. For the eQTL strategy, we found that α_{12} and $\text{dir}(G_1, G_2)$ agree approximately half of the time (48%). This gives further credence to our gene-gene approach by showing that a different type of information can be obtained by considering more biologically-relevant gene-level interactions.

For each SNP, we also include its PolyPhen (Adzhubei et al., 2013) and SIFT (Ng, 2003) scores reported in BioMart (Kinsella et al., 2011) to better understand its potential deleterious impact on MS. If available, both scores are comprised between 0 and 1, but with opposite interpretations. For SIFT, 0 denotes a deleterious amino-acid substitution, while for PolyPhen, 1 denotes an benign substitution. In total, we obtained 5 variants which were predicted as deleterious by at least one of the two methods.

4.4.3 Biological interpretation

In addition to the preceding statistical analysis, we also conduct a biological analysis of the results for both mappings. Our analysis is built upon existing information in MetaCore disease maps in conjunction with relevant literature.

4.4.3.1 Physical mapping

In total, we obtained 136 epistatic interactions in the 19 disease maps. As an exhaustive analysis of all interactions is out of reach, an a posteriori filtering is needed. In physical mapping, an epistatic interaction between two genes corresponds to a change of their protein structure. We therefore retain an interaction if at least of one the SNPs in the top-scoring pair can lead to a loss of function at the protein level. For that matter, the SNPs are selected according to the following criteria:

- Frameshift variant or incomplete terminal codon variant or missense variant or start loss variant,
- Stop-gained, stop-lost or stop-retained variant,
- Terminal codon variant.

The filtering process yielded 4 gene pairs where one of the the genes presents a missense variant (Appendix C.7). For each of these gene pairs, the impact on the MS phenotype is given as specified (activation or inhibition) or unspecified (unknown), as depicted in Figure 4.2. Among the obtained 4 pairs, GLI-1 and SUFU appear to be particularly interesting, since both genes are in direct binding interaction in DM 3305, which illustrates the SHH (Sonic Hedgehog) signaling in oligodendrocyte precursor cells differentiation in MS (Appendix C.5.1).

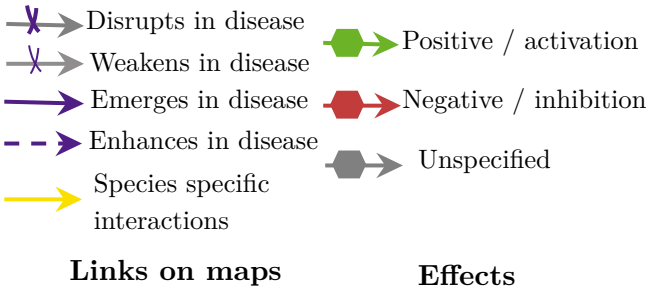


Figure 4.2: The different types of links between proteins/proteins or proteins-phenotypes in MetaCore maps

4.4.3.2 eQTL mapping

In eQTL mapping, an epistatic interaction consists of a gene pair, the simultaneous up/down-regulation of which induces a synergistic effect which lowers or increases the risk of MS. To better understand the impact of simultaneous gene up-regulation on disease propensity, we rewrite Equation (4.4):

$$\text{logit}\, P(Y|X_1, X_2) = \underline{\alpha_0} + \underline{\beta_1} e_1 + \underline{\beta_2} e_2 + \beta_{\text{syner}} e_1 e_2, \tag{4.6}$$

where $\beta_{\text{syner}} = \alpha_{12}/(\beta_1 \cdot \beta_2)$ and the constants α_0, $\underline{\beta_1}$ and $\underline{\beta_2}$ are functions of $(\alpha_0, \alpha_1, \alpha_2, \alpha_{12})$, (γ_1, γ_2) and (β_1, β_2).

The impact of gene up-regulation can be assessed through the signs of $(\beta_1, \beta_2, \beta_{\text{syner}})$. For instance, if β_1, β_2 and β_{syner} are positive, an increase in the expression of either genes leads to a higher disease risk. Hence, a joint inhibition of the two genes reduces the risk. In Table 4.2, we similarly study all possible sign combinations of $(\beta_1, \beta_2, \beta_{\text{syner}})$ to devise a number of recommendations for the application of epistasis to the development of combination therapy.

A total of 117 gene pairs in 19 disease maps were obtained with the eQTL

Table 4.2: Analysis of the impact of genes up-regulation on the risk for humans to develop MS, for each gene individually (signs of β_1 and β_2), and for the pair of genes synergistically (sign of β_{syner}) which is epistasis.

β_1	β_2	β_{syner}	Impact of β_1 and β_2 on MS	Recommendation for combination therapy
> 0	> 0	> 0	detrimental	inhibition of the two genes reduces the risk for MS
> 0	> 0	< 0	beneficial	genes must not be inhibited
< 0	< 0	< 0	beneficial	genes could be activated at the same time
< 0	< 0	> 0	detrimental	genes must not be activated
> 0	< 0	NC	NC	NC

mapping strategy. As in physical mapping, an additional filtering is needed. We selected the gene pairs in which the coefficients ($\beta_1, \beta_2, \beta_{\text{syner}}$) share the same sign (all positive or negative). If positive, the inhibition of both genes reduces the risk for MS. By contrast, if negative, the two genes should be jointly activated to reduce MS risk. This filtering led to 25 gene pairs of interest across 13 maps. Since a thorough study of all 25 pairs is possible, we implemented an additional filtering criterion: existence of a specified effect on MS-related phenotypes e.g. demyelination, remyelination failure, oligodendrocyte death, damage of neural axons, etc. The effect nature is given by the arrow types (see Figure 4.2). This final filter led to 9 gene pairs to consider (see Appendix C.6).

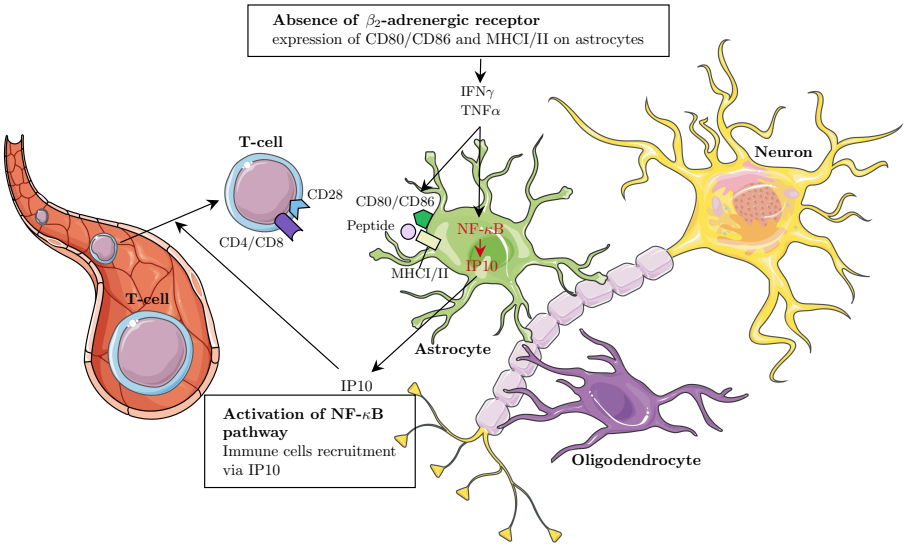
Confident in the single gene pair where both genes have a specified impact on the phenotype, NF-κB and IP10 (see Appendix C.8), we have investigated in further details their role in MS in the aim of assessing their synergistic effect on MS physiopathology. Our analysis is focused on DM 5199 (see Appendix C.5.3) where both genes belong to essential pathways.

Role of IP10 in MS: recruitment of T cell in the CNS IP10 (or IP-10 / CXCL10 (C-X-C motif chemokine ligand 10) / Interferon-Inducible Cytokine IP-10) is an antimicrobial gene which encodes a chemokine of the CXC subfamily, and is a ligand for the receptor CXCR3. This pro-inflammatory cytokine is involved in a wide variety of processes such as chemotaxis, differentiation, and activation of peripheral immune cells, like monocytes, natural killer, T-cell migration, and modulation of adhesion molecule expression (Romagnani et al., 2001; Antonia et al., 2019; Tokunaga et al., 2018).

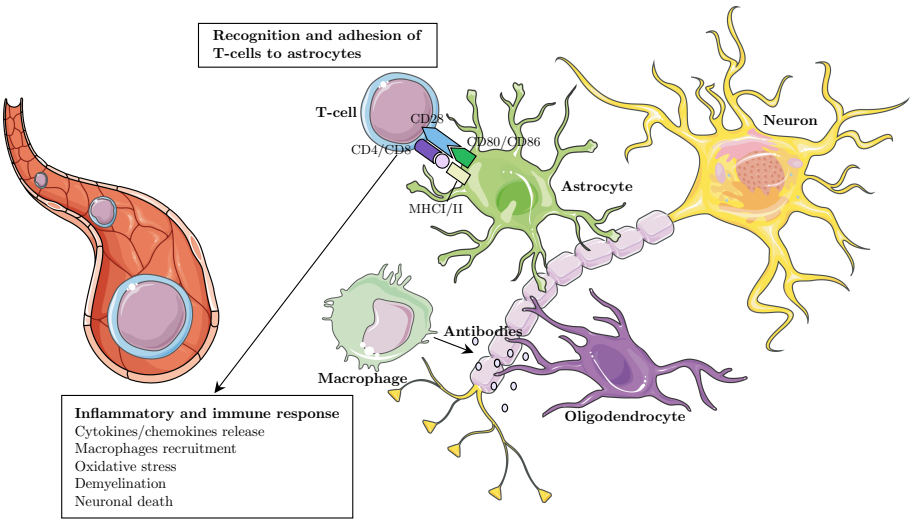
IP-10 is strongly induced by IFN-γ as well as by IFN-α/β (Qian et al., 2006). In vitro, CXCL10 can also be induced by NF-κB, and has been shown to have an early role in hypoxia-induced inflammation (Schmid et al., 2006; Xia et al., 2016). Indeed, in the disease map, the activation of IP10 by NF-κB is clearly indicated by an activation arrow (green arrow). Thus, the two genes are in direct interaction, where NF-κB regulates the transcription of IP10.

DM 5199, which contains IP10 and NF-κB, is focused on the impact of beta-2 adrenergic receptors, which are lacking in astrocytes in MS. This lack enables IFN-γ and TNF-α to trigger the expression of several key pro-inflammatory genes (Keyser et al., 2004, 2010). Whereas human astrocytes are only partially competent antigen presenting cells, the upregulation of MHC-II by IFN-γ alone or in combination with TNF-α enables astrocytes to present myelin as an auto-antigen, and triggers the production of the co-stimulatory molecules C80 and CD86 at their surface. Experimentally, the expression of MHC-class I and MHC-class II, together with the co-stimulatory molecules CD80 and CD86, is detectable in astrocytes in MS plaques (Traugott & Lebon, 1988).

After the transformation of astrocytes in immuno-competent cells, IP10 plays a major role by activating the recruitment of Th1 cells into the CNS (Figure 4.3a). Indeed, in MS, activated CXCR3+ T-cells (IP10 is the ligand for the receptor



(a) Transformation of astrocytes in immuno-competent celles and T-cells recruitment following the NF-κB/IP10 axis activation in MS.



(b) After recruitment of T-cells, adhesion of T-cell/astrocyte leads to inflammatory and immune response inducing neuron damage.

Figure 4.3: Schematic representation of the role played by the gene pairs NF-κB/IP10 in the development of demyelination in MS.

CXCR3) enter the CNS, and can be located in the cerebrospinal fluid or in the brain parenchyma (Lassmann & Ransohoff, 2004). This transport is made possible due to the blood Brain Barrier disruption in MS (Minagar & Alexander, 2003).

Arriving in the CNS, T lymphocytes recognize astrocytes via their MHC-II,

and anchor them via their CD28 which binds to CD80 and CD86 on astrocytes. This intercellular contact between T cells and astrocytes presenting myelin antigens induces the reactivation of T cells in the CNS (Cornet et al., 2000). T cells then secrete pro-inflammatory cytokines; demyelination occurs and macrophages are activated. This further damages myelin and releases cytokines - but also phagocytosing myelin debris - which leads to the damage of neural axons (Williams et al., 2007) (see Figure 4.3b).

Role of NF-κB in MS: transcription regulation Astrocyte reactivity is regulated by key canonical signaling cascades, among which the NF-κB pathway is qualified as pivotal for establishing neuroinflammation (Ponath et al., 2018). TNF-α binds to TNF-R1, which is constitutively expressed in astrocytes, and activates NF-κB signaling pathway (Liang et al., 2004). In cytoplasm, NF-κB is inhibited by I-kB proteins. Phosphorylation of I-kB by IKK (cat) kinase complex marks I-kB for destruction via the ubiquitination pathway, thereby allowing activation of NF-κB complex (Liang et al., 2004). The activated NF-κB translocates into the nucleus and upregulates transcription of target genes including IP10 (Majumder et al., 1998).

Status of IP10 and NF-κB as potential targets in MS treatment assays Human IP10 is a secreted protein, and is mainly located in the extracellular space, but also in the plasma membrane, and to a lesser extent in the cytosol and nucleus (Source: UniProtKB/Swiss-Prot). Today, the ChEMBL database indicates that two antibodies of IP10 are studied in clinical trials: NI-0801 (Phase I completed for allergic contact dermatitis, Phase II terminated for primary biliary cirrhosis) and ELDELUMAB (phase II mainly for rheumatoid arthritis, ulcerative colitis and Crohn's disease; source: Open Targets (Carvalho-Silva et al., 2018b)). The fact that, except for allergic contact dermatitis, all of these diseases belong to the autoimmune diseases family like MS, suggests that IP10 can be a valuable target for MS.

NF-κB is extensively present in the cytosol and the nucleus, to a lesser extent in the extracellular space, but not in the plasma membrane (Source: UniProtKB/Swiss-Prot). No small molecule or antibody is currently under clinical study for a direct blockade of NF-κB, since it is inhibited by IκB proteins in cytoplasm.

Clinical assays trying to inhibit NF-κB have so far focused on its upstream regulators. The phosphorylation of I-κB by the IKK (cat) kinase complex marks I-κB for destruction via the ubiquitination pathway, thereby allowing the activation of the NF-κB complex (Iwai, 2012). Different research groups tried to inhibit undesired NF-κB activity at several regulatory levels (Calzado et al., 2007). For example, inhibitors of IKKB-beta (or IKBKB: Inhibitor Of Nuclear Factor Kappa B Kinase Subunit Beta) aim at blocking the kinase which phosphorylates inhibitors of NF-kappa-B on two critical serine residues. Several small molecules antagonists targeting IKBKB are in phase I, II and III clinical trials for several diseases (source:

Open Target (Carvalho-Silva et al., 2018b)).

Downstream of NF-κB, glucocorticoids receptors (GR) also constitute an interesting research direction. Ligand-bound GR is able to antagonize the activity of immunogenic transcription factors such as nuclear factor-κB (NF-κB)3, AP-14,5, and T-bet6; resulting in a potent attenuation of inflammation (Hudson et al., 2018).

Altogether, these clinical assays for IP10 and NF-κB pathway inhibitors strengthen the potential of the pair as MS targets, where their simultaneous inhibition lowers the risk for MS.

4.5 Conclusion

We study gene-gene interactions for a number of disease maps related to multiple sclerosis. Nonetheless, the pipeline we describe here can be generalized to other diseases. It is based on epiGWAS, a SNP-level epistasis detection tool that we extend to the study of gene-level epistasis. Within each disease map, we obtained a number of significant interactions that formed novel subnetworks. Notably, we have shown complementarity between two different SNP-to-gene mappings: eQTL mapping and physical mapping. We identified 4 gene interactions mediated by potential function modifying variants. Among these interactions we retrieve one known direct binding interaction between GLI-I and SUFU, involved in oligodendrocyte precursor cells differentiation in MS. We also identified 25 gene interactions mediated by eQTLs, in particular a IP10-NF-κB interaction where each gene separately has a known impact on MS. We show that the epistasis mechanism probably pass through the known regulation of IP10 transcription by NF-κB. These observations validate that epistasis analysis can reveal biological interactions and endorse the use of this methodology to predict new biology. To the best of our knowledge, our work is the first application of an epistasis detection tool to a specific disease which is followed by an in-depth statistical analysis and biological interpretation of the results. Nonetheless, more biological and experimental validation is needed to confirm the discovered interactions.

Acknowledgements

This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113, 085475 and 090355.

Chapter 5

Nonlinear post-selection inference for genome-wide association studies

---

Publication and Dissemination: *in preparation. This is joint work with Clément Chatelain (SANOFI R&D) and Chloé-Agathe Azencott (Mines ParisTech).*

Abstract: *Association testing in genome-wide association studies (GWAS) is often performed at either the SNP level or the gene level. The two levels can bring different insights into disease mechanisms. In this chapter, we provide a novel approach based on nonlinear post-selection inference to bridge the gap between them. Our approach selects, within a gene, the SNPs or LD blocks most associated with the phenotype, before testing their combined effect. Both the selection and the association testing are conducted nonlinearly. We apply our tool to the study of BMI and its variation in the UK BioBank. In this study, our approach outperformed other gene-level association testing tools, with the unique benefit of pinpointing the causal SNPs.*

Résumé : *Les tests d'association dans les études d'association à l'échelle du génome (GWAS) sont souvent effectués au niveau du SNP ou au niveau du gène. Les deux niveaux peuvent apporter des informations différentes sur les mécanismes de la maladie. Dans ce chapitre, nous proposons une nouvelle approche basée sur l'inférence post-sélection nonlinéaire pour combler l'écart entre eux. Notre approche sélectionne, au sein d'un gène, les SNPs ou blocs LD les plus associés au phénotype, avant de tester leur effet combiné. Les tests de sélection et d'association sont effectués de manière nonlinéaire. Nous appliquons notre outil à l'étude de l'IMC et de ses variations dans la UK BioBank. Dans cette étude, notre approche a surpassé les autres outils de test d'association au niveau des gènes, avec l'avantage unique de localiser les SNP causaux.*

5.1 Introduction

Lack of statistical power is a major limitation in GWAS. If the analysis is performed at the SNP level, lack of statistical power may stem from small effect sizes and linkage disequilibrium, among others. By modeling the overall association signal, gene level analysis can address this limitation. Being the functional entity, genes have the potential to shed light on yet undiscovered biological and functional mechanisms. However, the incorporation of all mapped SNPs, including non-causal ones, can mask the association signal. An alternative strategy would be to select the SNPs most associated with the phenotype within a given gene, and then test their joint effect. If we do not account for the fact that these SNPs were selected in a first step based on the same data, their overall joint effect is likely to be overestimated. Post-selection inference (PSI) (Lee et al., 2016) was specifically developed to correct for this selection bias, and has already been applied in the context of GWAS (Mieth et al., 2016). In addition, such a framework would also benefit from the incorporation of nonlinearities to model epistatic interactions between neighboring SNPs.

In Chapter 3, we described the theoretical foundations of kernelPSI, a post-selection inference (PSI) framework for nonlinear variable selection. Here, we extend kernelPSI to the demanding setting of GWAS, characterized by its high-dimensionality in both directions: number of samples in large biobanks, and number of SNPs. In kernelPSI, we condition for the selection bias by performing a constrained sampling of replicates of the response vector. We then compare the statistics of the response to those of the replicates to obtain the desired p-values.

The extension of kernelPSI to GWAS required several modifications to improve scalability. Most importantly, we developed a GPU version of the constrained sampling algorithm to speed up linear algebra operations. The rest of the code was also accelerated thanks to a more efficient C++ backend. In particular, we implemented a rapid estimator of the HSIC criterion (Gretton et al., 2005a) based on quadratic-time rank-1 matrix multiplications. HSIC is an example of quadratic kernel association scores (see Chapter 3). The latter are quadratic forms of the response vector, which can model nonlinear effects and epistatic interactions among neighboring SNPs. This extension also generalizes kernelPSI to any non-normally distributed continuous phenotypic outcome.

To illustrate the use of kernelPSI on real GWAS datasets, we study BMI and its fluctuations (ΔBMI) in the UK BioBank. The UK BioBank (Bycroft et al., 2018) is one of the largest available sources of data for the investigation of the contribution of genetic predisposition to a variety of physiological and disease phenotypes. We study both BMI and ΔBMI because of the suspicion that different genetic mechanisms might be governing the two phenotypes (Sandholt et al., 2013). Our study yielded a number of putative genes for BMI and ΔBMI, along with a list of causal *loci* within. Our use case has also shown the better statistical performance of kernelPSI in comparison to other gene-level association tools, with the unique benefit of pinpointing the causal *loci*.

We propose an eponymous R package that implements the full pipeline of kernelPSI. The CPU-only version is directly available from CRAN. The enhanced GPU-version can be downloaded from the development branch of the GitHub repository https://github.com/EpiSlim/kernelPSI.git.

5.2 KernelPSI: post-selection inference for big genomic data

Before covering the modifications we implemented to extend kernelPSI to GWAS data, we start with a brief overview of the framework in the context of GWAS. For further details, we refer the reader to Chapter 3.

We model a GWAS dataset as a set of n pairs $\{(x_1, y_1), \cdots, (x_n, y_n)\}$. For each sample $i \in [\![1, n]\!]$, $y_i \in \mathbb{R}$ represents the phenotype and $x_i \in \mathcal{X}^p$ the genotype, with p the number of SNPs considered. In this study, we defined x_i as a set of p SNPs mapped to a gene (see Section 5.3.1.3), and $\mathcal{X} = \{0, 1, 2\}$ following the dosage encoding of SNPs. We denote by $Y \in \mathbb{R}^n$ the vector of phenotypes, where $Y_i = y_i$ for $i \in [\![1, n]\!]$. We further consider a partition of the genotype in a set of S contiguous SNP clusters $\{\mathcal{S}_1, \cdots, \mathcal{S}_S\}$ (see Section 5.2.2). For each $t \in [\![1, S]\!]$, we define a kernel $\mathcal{K}_t : \{0, 1, 2\}^{|\mathcal{S}_t|} \times \{0, 1, 2\}^{|\mathcal{S}_t|} \to \mathbb{R}$ and the corresponding Gram matrix K_t (see Section 5.2.3 for examples of such kernels). For any $i, j \in [\![1, n]\!]$, $[K_t]_{ij} = \mathcal{K}_t(x_{i,\mathcal{S}_t}, x_{j,\mathcal{S}_t})$, where $x_{i,\mathcal{S}_t}$ contains the values of the SNPs in $\mathcal{S}_t$ for sample i, that is to say, x_i restricted to its entries in $\mathcal{S}_t$.

The goal is to select the SNP clusters that is, the kernels within $\{\mathcal{K}_1, \cdots, \mathcal{K}_S\}$, most associated with the phenotype, and then, to measure their overall association with the phenotype Y. In other words, we perform model selection and measure afterwards the significance of the constructed model.

In both selection and inference stages, a measure of association between a kernel K and a phenotype Y is needed. For this purpose, we define *quadratic kernel association scores* which are quadratic forms in Y:

$$\begin{aligned} s \colon \mathbb{R}^{n\times n} \times \mathbb{R}^n &\to \mathbb{R} \\ (K, Y) &\mapsto Y^\top Q(K) Y, \end{aligned} \tag{5.1}$$

for some mapping $Q : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$.

Quadratic kernel association scores encompass a wide gamut of scores. For instance, empirical estimators of the HSIC criterion. In this chapter, we restrict ourselves to the unbiased empirical HSIC estimator, first proposed by Song et al. (2007):

$$\begin{aligned} \widehat{\mathrm{HSIC}}_{\mathrm{unbiased}}(X, Y) = \frac{1}{n(n-3)} \Bigg[&\mathrm{trace}(\underline{K}\ \underline{L}) \\ &+ \frac{1_n^\top \underline{K} 1_n\ 1_n^\top \underline{L} 1_n}{(n-1)(n-2)} - \frac{2}{n-2} 1_n^\top \underline{K}\ \underline{L} 1_n \Bigg], \end{aligned} \tag{5.2}$$

where $\underline{K} = K - \mathrm{diag}(K)$ and $\underline{L} = L - \mathrm{diag}(L)$.

A multitude of kernel selection strategies can be deployed (see Section 3.4). The kernels can be selected in a *forward* or *backward* stepwise fashion. The number of selected kernels can be either fixed, or adaptively determined. Here, we opt for an adaptive forward strategy, where the number of selected kernels S' is determined according to the maximum of $\widehat{\text{HSIC}}_{\text{unbiased}}$ attained by iteratively adding the kernels.

Regardless of the kernel selection strategy, the selection of a subset of kernels $M \subseteq \mathcal{K}$ can be modeled as a conjunction of quadratic constraints: there exists $i_M \in \mathbb{N}$, and $(Q_{M,1}, b_{M,1}), \cdots, (Q_{M,i_M}, b_{M,i_M}) \in \mathbb{R}^{n\times n} \times \mathbb{R}$ such that

$$\{Y \,:\, \widehat{M}(Y) = M\} = \bigcap_{i=1}^{i_M} \{Y \,:\, Y^\top Q_{M,i} Y + b_{M,i} \geq 0\}. \tag{5.3}$$

Testing the association between the kernels in M and Y needs to account for the statistical bias introduced by the selection event. For valid inference, we need to correct for the fact that the kernels were selected on the basis of their strong association with the outcome Y. As determining the exact distribution of $\text{HSIC}_{\text{unbiased}}$ conditionally to the event $\{Y \,:\, \widehat{M}(Y) = M\}$ was impossible, we developed instead an efficient sampling algorithm to derive empirical p-values. Replicates of the outcome Y which satisfy the quadratic constraints in (5.3) are sampled. The values of their test statistics (in this case, $\widehat{\text{HSIC}}_{\text{unbiased}}$) are then compared to the value of the statistic of the original outcome Y to obtain the desired p-values.

5.2.1 Outcome normalization

Our proposal in Chapter 3 is limited to normally-distributed outcomes. To expand kernelPSI to other continuous outcomes, one needs to transform any continuous outcome Y into a vector of independent normally-distributed variables. A well-known transformation is the Van der Waerden (1952) quantile transformation given by:

$$g(y) = F_{0,1}^{-1}\left(\frac{\text{rank}(y) - 1/2}{n+1}\right), \tag{5.4}$$

where $y \in \mathbb{R}$, $\text{rank}(y)$ is the ranking of y in descending order with respect to $y_1, \cdots, y_n$, and $F_{0,1}$ is the c.d.f of the standard normal distribution.

The accuracy of the transformation in Equation (5.4) depends on the accuracy of the estimation of the regularized quantile $(\text{rank}(y) - 1/2)/(n+1)$, and thus on the number of participants n. Thankfully, many recent GWAS, in particular for physiological measurements, boast a large number of participants.

Other outcome normalization methods have been proposed, such as the Lambert $W \times F$ (Goerg, 2011), or Box-Cox (Box & Cox, 1964) and Yeo-Johnson (Yeo, 2000) transformations. In practice, we found the Van der Waerden transformation in Equation (5.4) to be the most consistent approach across different types of outcome distributions. All the above transformations are implemented in the R package bestNormalize (Peterson & Cavanaugh, 2019).

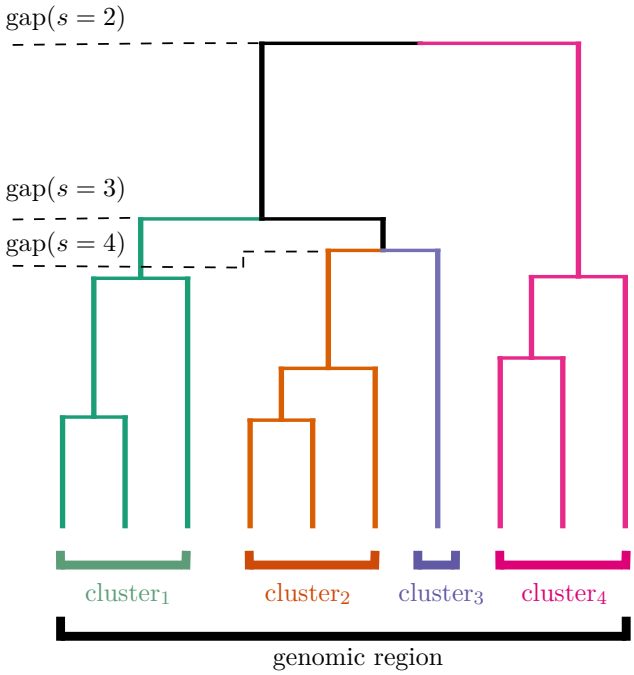


Figure 5.1: Clustering methodology: adjacent hierarchical clustering coupled with the gap statistic to determine the appropriate number of clusters.

5.2.2 Contiguous hierarchical clustering for genomic regions

In GWAS, the true causal SNPs are often unmeasured, but exhibit a strong linkage disequilibrium (LD) with the lead SNPs. The classical strategy to approach this problem is fine mapping (Schaid et al., 2018), where we study the genomic region surrounding the lead SNPs to identify the causal SNPs. A better strategy would then be to directly select regions of strong LD patterns. This amounts to selecting clusters of strongly-correlated SNPs. Such a strategy also has the advantage of reducing the number of clusters/kernels to choose from, while simultaneously modeling the combined cluster effects on the outcome. More statistical power is to be expected.

To define these clusters, we use the R package BALD (Dehman et al., 2015) which implements adjacent hierarchical clustering (AHC) in conjunction with the gap statistic (Tibshirani et al., 2001). Following AHC, the optimal number of clusters S is estimated using the gap statistic.

Here, we apply adjacent hierarchical clustering coupled with the gap statistic to split genomic regions into contiguous groups of SNPs. This approach is illustrated in Figure 5.1, and is readily available from the R package BALD Dehman et al. (2015).

5.2.3 The IBS-kernels and nonlinear SNP selection

It is obviously possible to use linear kernels to define $\{\mathcal{K}_1, \cdots, \mathcal{K}_S\}$. However, such a representation does not take into account MAFs and epistatic interactions between SNPs. To address this limitation, Wu et al. (2010) proposed identical-by-state (IBS) kernels, which measure the number of identical alleles between two individuals i and j. For a cluster t and two genotypes x_i, x_j, IBS kernels are given by:

$$\mathcal{K}_t(x_{i,\mathcal{S}_t}, x_{j,\mathcal{S}_t}) = \sum_{q=1}^{|\mathcal{S}_t|} w_q(2 - \left|[x_{i,\mathcal{S}_t}]_q - [x_{j,\mathcal{S}_t}]_q\right|), \tag{5.5}$$

where the weights $(w_q)_{1:|\mathcal{S}_t|}$ are a function of their respective MAFs $(m_q)_{1:|\mathcal{S}_t|}$:

$$\sqrt{w_q} = \text{Beta}\left(m_q, \alpha_q, \beta_q\right), \tag{5.6}$$

where Beta is the density function of the Beta distribution.

The parameterization $(\alpha_q, \beta_q)_{1:|\mathcal{S}_t|}$ is chosen according to the scope of the GWAS study. For common variants, Ionita-Laza et al. (2013) recommend setting $(\alpha_q, \beta_q) = (0.5, 0.5)$. Such a parameterization still assigns higher weights to rare variants, but the difference is more moderate. For instance, for $(\alpha_q, \beta_q) = (0.5, 0.5)$, we have $w_q^2 = 0.63$ for $m_q = .5$ and for $m_q = 0.01$, $w_q^2 = 3.2$. To get a better understanding of these choices, we compare in Figure 5.2 the Beta densities for different values of (α_q, β_q).

5.2.4 Efficient nonlinear post-selection inference for high-dimensional data

In this section, we detail a number of modifications we included in order to improve the scalability of kernelPSI to the large sample sizes.

5.2.4.1 Rapid estimation of the HSIC criterion

We first recall the unbiased HSIC estimator in Equation (5.2):

$$\begin{aligned}\widehat{\text{HSIC}}_{\text{unbiased}}(X, Y) = \frac{1}{n(n-3)}\Bigg[&\text{trace}(\underline{K}\ \underline{L}) \\ &+ \frac{1_n^\top \underline{K} 1_n\ 1_n^\top \underline{L} 1_n}{(n-1)(n-2)} - \frac{2}{n-2} 1_n^\top \underline{K}\ \underline{L} 1_n\Bigg].\end{aligned} \tag{5.7}$$

The computation of $1_n^\top \underline{K} 1_n$ and $1_n^\top \underline{L} 1_n$ can be performed in quadratic time $\mathcal{O}(n^2)$. However, for $\text{trace}(\underline{K}\ \underline{L})$ and $1_n^\top \underline{K}\ \underline{L} 1_n$, a $\mathcal{O}(n^3)$ complexity can ensue because of the matrix-matrix multiplication of $\underline{K}$ and $\underline{L}$. To avoid that, we decompose $\text{trace}(\underline{K}\ \underline{L})$ as $\sum_{i,j=1}^n [\underline{K}]_{ij} [\underline{L}]_{ji}$, which results in a better $\mathcal{O}(n^2)$ complexity. The same complexity can be achieved for the quadratic form $1_n^\top \underline{K}\ \underline{L} 1_n$ by starting with the matrix-vector multiplication of either $\underline{K} 1_n$ or $\underline{L} 1_n$. Overall, we achieve a $\mathcal{O}(n^2)$ complexity, for which the HSIC criterion can be computed on a single CPU for

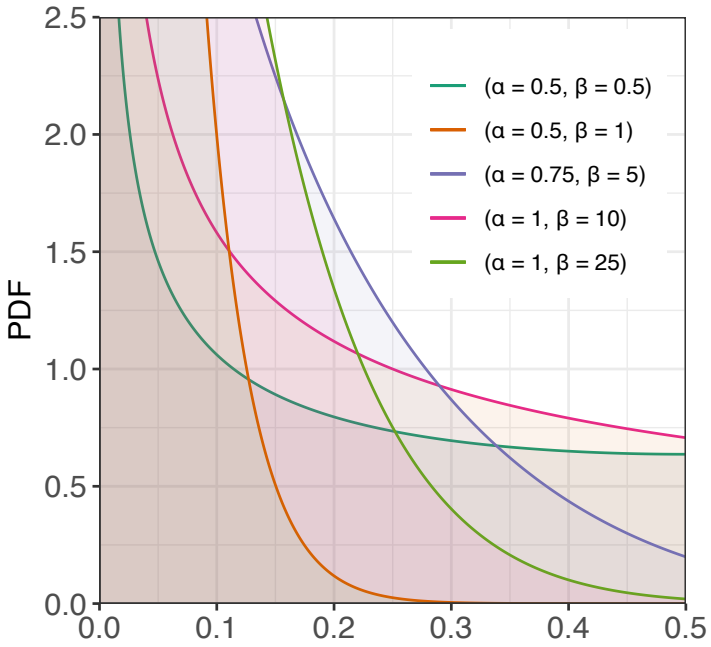


Figure 5.2: Comparison of the Beta densities for different values of the shape parameters (α, β).

thousands of samples in relatively little time. As an illustration, we performed 100 repetitive evaluations of the HSIC criterion for two matrices of size $5,000 \times 5,000$. On a 2.7 GHz intel core i5 processor, the average running time was 1.08s.

5.2.4.2 Accelerated replicates sampling

The gains achieved in Section 5.2.4.1 turned out to be insufficient because of the heavy computational workload involved in replicates' sampling. Our sampling algorithm in Chapter 3 is partly a rejection sampling algorithm. At every iteration, we verify that the candidate replicate satisfies the constraints $Y : Y^\top Q_{M,i} Y + b_{M,i} \geq 0$ for $i \in [\![1, i_M]\!]$. For a large i_M, we observed a significant slow-down due to the overhead between the successive evaluations of the constraints. A single combined evaluation would then eliminate this overhead. We achieve this by encoding all computations in a matrix form, as illustrated in Figure 5.3.

For linear algebra operations, GPUs can dramatically speed up computations (Krüger & Westermann, 2003). We used them here to accelerate the multiplications detailed in Figure 5.3. More specifically, we used the ViennaCL library (Rupp et al., 2016) which enables a simple, high-level access to GPU resources.

A major drawback in hybrid CPU-GPU calculations is the transfer time between the main memory and the GPU memory. With most Nvidia GPUs, the theoretical bandwidth limit is 8 Gb/s. For $i_M = 12$ and $n = 10^4$, the approximate memory size

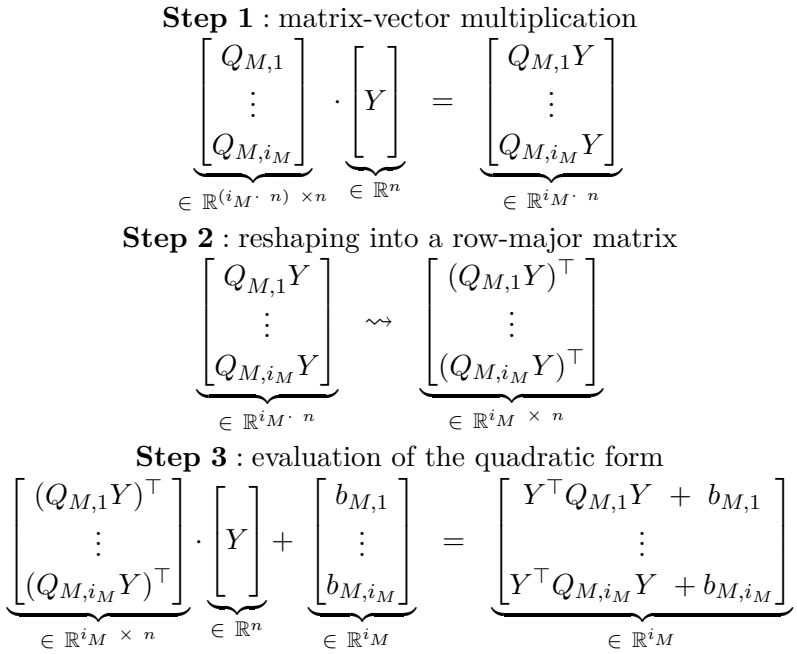


Figure 5.3: A GPU-accelerated pipeline for the evaluation of quadratic constraints.

of all $\mathcal{Q}_M = \{Q_{M,1}, \cdots, Q_{M,i_M}\}$ matrices is 8.94 Gb in a double representation. If we sample $5 \cdot 10^4$ replicates, a repeated data transfer of $\mathcal{Q}_M$ would in the best case last 15 hours and 30 minutes. Such transfer times are prohibitive. To circumvent this problem, we transfer the matrices in $\mathcal{Q}_M$ to GPU memory once and for all before the sampling. However, because of memory size limitations, this imposes an upper limit on the number of matrices i_M in $\mathcal{Q}_M$, and consequently on the number of clusters $S = i_M/2 + 1$.

Finally, we give a rough estimation of the complexity of our sampling algorithm. If we denote by $N_{\text{replicates}}$ the number of replicates, the overall complexity can be approximated as $\mathcal{O}(N_{\text{replicates}} i_M n^2/\tau(n))$. $\tau(n)$ is a decreasing function of n which corresponds to the probability of sampling a replicate in the acceptance region. The average number of iterations to obtain a valid replicate is then $1/\tau(n)$ (mean of a geometric distribution). We are currently unable to propose a closed form for $\tau(n)$.

5.3 A study of BMI and its variation in the UK BioBank

The study of physiological phenotypes in GWAS has so far focused on basic anthropometric measures such as height, weight, and BMI. Their longitudinal fluctuations received little attention, mainly because of the lack of such data. To the best of our knowledge, the fluctuations of BMI have not been the subject of any specific GWA study. In fact, some studies (Sandholt et al., 2013) suggested that BMI and ΔBMI might be influenced by distinct sets of SNPs. Only rare variants impacting

weight loss through gene-diet interaction are referenced in the literature (Qi, 2014). Recent biobanks such as HUNT (Holmen m.fl, 2011), ALSPAC (Fraser et al., 2012) and BiB (Raynor & Group, 2008) are finally making such data available. Another notable biobank is the UK BioBank (Bycroft et al., 2018) which provides extensive phenotypic and health-related information for over $500,000$ British participants. We apply kernelPSI on the UK BioBank dataset to separately study BMI and variations of BMI (ΔBMI).

5.3.1 Data and experiments

5.3.1.1 Quality control

Preprocessing in any GWA study is a manadatory step. Our preprocessing pipeline for the UK BioBank dataset is closely similar to the pipeline of the Neale lab [1] who provides exhaustive summary statistics for over $2,000$ phenotypes in the UK BioBank. We detail below the sample quality-control we conducted.

- Heterozygosity and missing rates: Discarding outliers for both criteria.
- Sex chromosome aneuploidy: only individuals with sex chromosome configurations XX or XY are retained.
- Prior use in phasing: whether sample was selected as input for the phasing of autosomal chromosomes
- Kinship to other participants: we only select participants with no identified relatives in the dataset.
- Ethnic grouping: we subset samples identified as 'white British' to avoid any potential population structure effects.
- Prior use in principal components analysis (PCA): we discard all samples not included in the PCA. The analysis is used for population stratification (see next step).
- Homogeneity: additional population structure artifacts are detected used genomic dispersion (GD). We approximated it through the normalized squared distance of the first six principal components (PCs):

$$\mathrm{GD}(X) = \frac{1}{6}\sum_{i=1}^{6}\langle \mathrm{PC}_i, X\rangle^2$$

The application of the above pipeline yielded $n = 266,679$ final samples. In contrast, Neale lab obtained $337,000$ samples by implementing less stringent thresholds.

We directly extracted the SNPs of the UK BioBank Axiom array from the imputed genotypes provided by the UK BioBank consortium. As for SNP quality

[1] More details are provided on their website https://www.nealelab.is/uk-biobank

control, we focused on bi-allelic SNPs located on autosomal chromosomes. Moreover, we filtered out the SNPs with a MAF < 0.01 or not in a Hardy-Weinberg equilibrium ($p < 1\text{e}{-10}$). Out of caution, we also incorporated two additional filters: we shed the SNPs with an internal UK BioBank information score < 0.8 and a missing proportion rate $> 1/n$. The QC pipeline resulted in $577,811$ SNPs.

5.3.1.2 Phenotypes

ΔBMI is not directly available. We computed it from the participants who attended both the initial assessment visit and the first repeat assessment visit. In total, we obtained $11,992$ samples for ΔBMI. More precisely, we use the average yearly variations $\Delta\text{BMI}/\Delta t_{\text{years}}$. The reason for this is that the time span between the visits is not the same for all participants. For simplicity of notation, we use ΔBMI to denote yearly variations in the rest of this chapter.

As for BMI, we use the measurements of the initial visit. The joint analysis of results stemming from datasets with different samples is not straightforward and requires the utmost attention. We mention two-sample problems as an example of tools providing statistically principled methods to tackle this problem. Our case is even more delicate: BMI and ΔBMI share a number of samples with a huge discrepancy in the total number of samples. To avoid this issue, we restrict ourselves in both phenotypes to the samples for which the ΔBMI measurement is available.

As explained in Section 5.2.1, we apply the Van der Waerden transformation. We illustrate the accuracy of the transformation in Figure 5.4, by visually comparing the empirical c.d.fs of BMI and ΔBMI to the c.d.f of a standard normal distribution. We notice a complete overlap between the c.d.fs. We attribute this good performance to the total number of samples and the low number of ties ($11,933$ unique values).

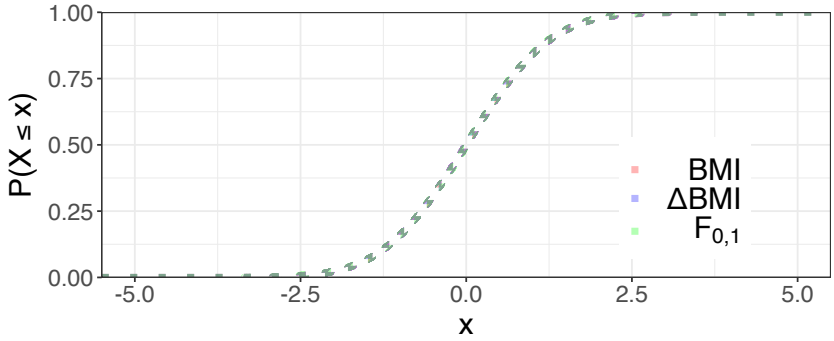


Figure 5.4: Comparison of the empirical c.d.fs of BMI and ΔBMI to the c.d.f of a standard normal distribution.

5.3.1.3 Gene selection

Because of the computation time and resource requirements, a full genome-wide study is impossible. We instead restricted ourselves to the genes already associated with BMI in the GWAS catalog (MacArthur et al., 2016). The scope of the narrower study is then gene prioritization. This is particularly interesting given the large number of genes associated with BMI (1811 genes).

Genes are included in the GWAS catalog, if they contain at least one significant SNP, which can possibly result in a high number of false positive genes. A major strength of kernelPSI is its dual SNP-gene perspective. The gene-level association testing in kernelPSI can assess whether the SNP-level association translates into a gene-level association.

To define genic boundaries, we used the biomaRt tool (Durinck et al., 2009), which provided a genomic interval for 1774 genes. Moreover, the intervals were converted from the GRCh38 coordinate system to the GRCh37 one, since the SNP positions in the UK BioBank are given in the GRCh37 system. We point out that the conversion can result in several noncontiguous intervals (see Hinrichs (2006); Lawrence et al. (2009) for further explanation).

An immediate use of the resulting intervals led to a number of genes without any SNPs within. As a result, we added a downstream/upstream 50kb buffer to cover more SNPs. The same buffer size was also opted for by several other authors (Nakka et al., 2016; Shah et al., 2018).

5.3.1.4 Hierarchical clustering

Despite the extended 50kb buffer, several genes still contained a handful of SNPs. 1215 genes contained at most 3 SNPs. In particular, if only one SNP only is mapped to a given gene, the kernel selection step becomes irrelevant. Nonetheless, we still perform hypothesis testing by directly using the HSIC statistic in Eq. (5.2) to measure the association between the gene and the phenotype. If 2 or 3 SNPs mapped to a gene, we associate a distinct cluster/kernel to each one of them. This allows for a more accurate SNP selection without entailing a dramatic increase in computational complexity.

For all other genes (with more than 4 SNPs), we applied AHC, as explained in Section 5.2.2. The optimal number of clusters S is determined by the gap statistic. In the adaptive kernel selection strategy we use here, this leads to $i_M = 2\,(S-1)$ constraints. To avoid the issues encountered for a large i_M (see Section 5.2.4.2), we set the maximum number of clusters to 5. This leads to a $\sim$ 9.2Gb maximum GPU memory occupancy for the matrices $\mathcal{Q}_M$.

5.3.2 Results

KernelPSI presents the unique benefit of jointly performing SNP-level selection and gene-level significance testing. In this section, we evaluate the performance of kernelPSI in both steps.

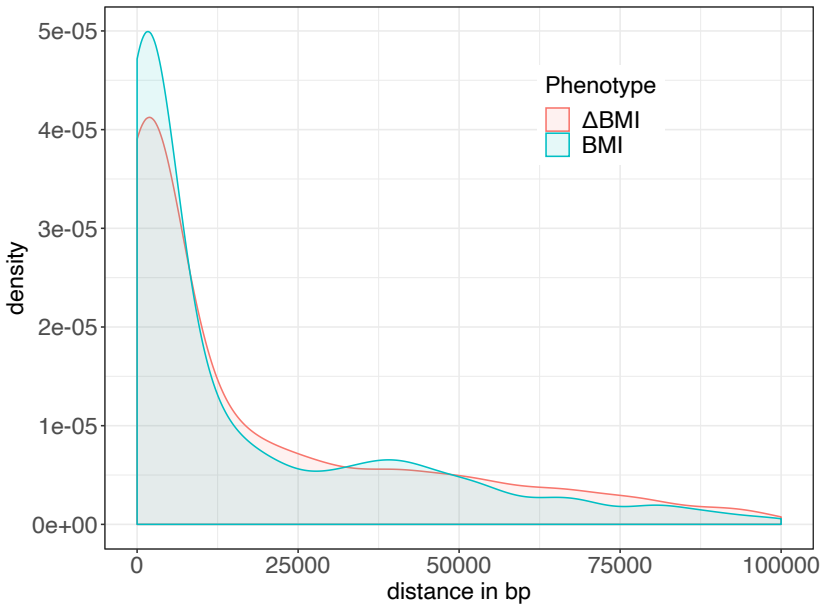


Figure 5.5: Distance between the SNPs of the GWAS Catalog and their closest neighbor among the SNPs in the clusters selected by kernelPSI.

5.3.2.1 Kernel selection

Because of the lack of a ground truth for all genes, validating the results of statistical tools in GWAS has always been difficult. For our study, the validation task is easier, though potentially biased. The genes were retrieved on the basis of their SNP-level association to BMI in the GWAS catalog. We can then compare the distance between the significant SNPs in each gene to their closest SNP neighbor in the clusters selected by kernelPSI. We provide in Figure 5.5 a histogram for the latter distances. The histogram is heavily skewed toward small distances. In other words, the GWAS catalog SNPs are often located near SNPs selected by kernelPSI. This confirms the capacity of kernelPSI to retrieve relevant genomic regions. Moreover, the selected clusters also surround significant SNPs. For BMI and ΔBMI, the selected clusters respectively included at least one significant SNP in 62.5% and 40.6% of genes.

If kernelPSI turned out to be selecting all clusters, the above results would be irrelevant. The clusters would always contain significant SNPs, and a few selected SNPs would also be located near the significant ones. In our application, kernelPSI conservatively selected the number of associated clusters S' (see Table 5.1). For BMI, kernelPSI selected one cluster in $75,9\%$ of the genes for which $S = 3$ and at most 2 clusters in $73,6\%$ of the genes for which $S = 5$. Similar results were obtained for ΔBMI.

Overall, the conservative kernel selection combined with the proximity of the selected kernels to the GWAS catalog SNPs demonstrate the selection performance of kernelPSI.

Table 5.1: Distribution of the number of selected clusters S' depending on the total number of clusters S and the phenotype.

	BMI				
	1	2	3	4	5
1	100.0%				
2	82.9%	17.1%			
3	75.9%	18.5%	5.6%		
4	55.9%	30.9%	10.3%	2.9%	
5	43.9%	29.7%	17.6%	7.2%	1.5%
	ΔBMI				
	1	2	3	4	5
1	100.0%				
2	92.7%	7.3%			
3	70.4%	20.4%	9.3%		
4	50.0%	27.9%	19.1%	2.9%	
5	40.3%	28.4%	21.4%	7.6%	2.3%

Table 5.2: Concordance between BMI and ΔBMI by method, according to three Kendall rank correlation measures (standard, multiplicative, additive).

	Standard		Multiplicative		Additive	
	BMI	ΔBMI	BMI	ΔBMI	BMI	ΔBMI
kernelPSI	1.000	0.015	1.000	0.093	1.000	0.072
	0.015	1.000	0.093	1.000	0.072	1.000
SKAT	1.000	0.020	1.000	0.008	1.000	0.028
	0.020	1.000	0.008	1.000	0.028	1.000
MAGMA	1.000	0.036	1.000	0.058	1.000	0.083
	0.036	1.000	0.058	1.000	0.083	1.000

5.3.2.2 Hypothesis testing

For association testing, we benchmark kernelPSI against two state-of-the-art gene-level baselines. The first one is SKAT (Wu et al., 2011), and can described as a non-selective variant of kernelPSI. Furthermore, it is a quadratic kernel association score which can be incorporated into the framework of kernelPSI. The SKAT score is a variance-component score (Lin, 1997) given by $s_{\text{SKAT}}(K, Y) = Y^\top KY$, for a centered phenotype Y. The second baseline is MAGMA (de Leeuw et al., 2015) which implements the principal components regression gene analysis model. More specifically, it implements an F-test in which the null hypothesis corresponds to absence of effects of all genotype PCs.

A central hypothesis for our study is the different mechanisms involved in BMI and ΔBMI. The low rank correlations of the p -values between the two phenotypes (see Table 5.2) lend further credence to this hypothesis. Interestingly, we observed a similar range of values for kernelPSI and the two benchmarks SKAT and MAGMA. For all metrics and methods, the rank correlations are lower than 0.1.

Despite the low rank correlations between BMI and ΔBMI, we obtained 7 common significant genes [2] out of 64 significant genes for ΔBMI and 40 for BMI. The latter were determined after the application of the Benjamini-Hochberg procedure with an FDR threshold of 0.05. The existence of a number of separate mechanisms does not preclude the existence of common ones simultaneously regulating BMI and ΔBMI.

To compute the empirical p-values in kernelPSI, we sampled $40,000$ replicates in addition to $10,000$ burn-in replicates. The comparison of the distributions of the resulting p-values to those of SKAT and MAGMA shows that kernelPSI clearly enjoys more statistical power than the two baselines for both phenotypes (Figure 5.6). The p-values were altogether significantly lower. Thanks to the large number of replicates, we attribute this performance, not to the lack of accuracy of the empirical p-values, but to the discarding of non-causal clusters in the selection stage.

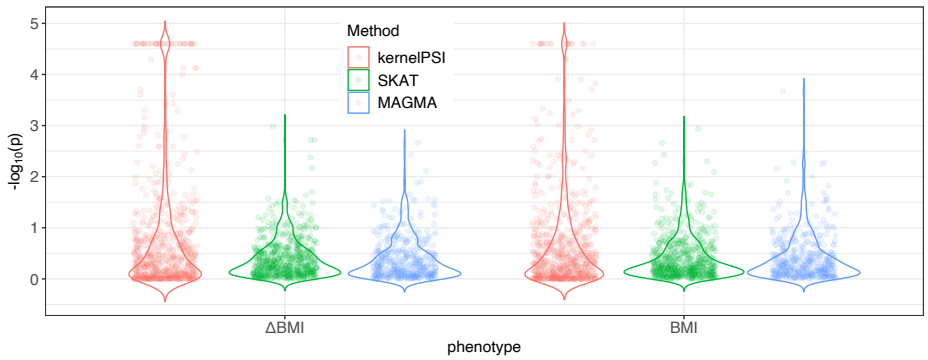


Figure 5.6: A violin plot comparing the p-values of kernelPSI for BMI and ΔBMI to two benchmarks.

[2] The common genes are: CKB, EIF2S2, KSR2, MIR100HG, NRXN3, PDILT and RAB27B

5.4 Conclusion

Most GWAS restricted themselves to SNP-level association testing. In this chapter, we presented a tool that still enables SNP selection, but ascends to the gene-level to perform association testing. The combination of the SNP and gene levels was possible through the use of post-selection inference which properly accounts for the SNP selection bias to perform valid gene inference. A major novelty in our work is the use of kernel methods which can model nonlinear effects and interactions among SNPs. The broad GWAS community can benefit from tools like kernelPSI which combine statistical performance with interpretability. In the future, we look forward to developing exact variants of kernelPSI which forego the sampling step to directly determine the associated p-value. This can dramatically reduce computational times. Another area of major interest to us is the application of kernelPSI to other SNP sets such as cis-regulatory regions or whole pathways to obtain the significance of association of the pathway and the involved genes.

Chapter 6

Conclusion and Perspectives

Put simplistically, epistasis is the interaction between distinct *loci*. The various forms in which epistasis occurs and the overlap between statistical epistasis and biological epistasis are often subject to debate. However, we can unquestionably distinguish between intragenic epistasis and intergenic epistasis. In this thesis, I proposed novel approaches to improve the detection of each of these types of epistasis. I provide below a brief overview of these methods. Not only did our methods consistently outperform state-of-the art baselines, but they also extend powerful fields of statistical learning to GWAS. They only represent a first step towards fully leveraging causal inference and nonlinear post-selection inference in GWAS.

1. **epiGWAS**: is the name of our proposal for intergenic epistasis (see Chapter 2). More specifically, we detect interactions between a predetermined SNP target and the rest of the genome. This approach falls within the framework of causal inference. A major application of this framework is the estimation of interactions between a treatment assignment and a set of clinical covariates. This is analogous to the detection of interactions between a target SNP and a set of SNPs located across the rest of the genome. Based on this analogy, we adapted robust tools for interaction detection in clinical trials to epistasis. The output of epiWGAS are a list of interaction scores between the target SNP and every other SNP. Interestingly, epiGWAS and the baselines that we compared against retrieved different interactions suggesting their combination as a means to further improve statistical power.

2. **kernelPSI**: is the approach we propose for intragenic epistasis (see Chapter 3). The biggest benefit of this approach is its dual functionality: gene-level hypothesis testing in combination with SNP-level selection. This duality is achieved thanks to post-selection inference, which develops a set of techniques to perform feature selection followed by hypothesis testing. So far, the proposed approaches have been limited to linear models. Linear modeling is insufficient for epistasis given its complex nonlinear patterns. We therefore incorporated kernel methods to develop nonlinear post-selection inference. This enables us to model nonlinear effects and interactions among neighboring SNPs. Here, the association between covariates and outcome is measured according to what we call quadratic kernel association scores. They are a quadratic form of the outcome, and are generic in the sense that

they encompass other known forms of association scores. All kernel selection schemas in kernelPSI (foward/backward, with a fixed/adaptive number of kernels) result in a selection event represented by a set of quadratic constraints. We use this representation in hypothesis testing in order to generate *valid* empirical p-values. The result of the initial selection step is to discard irrelevant kernels/covariates to gain in statistical power in hypothesis testing. We demonstrate this by comparing kernelPSI to nonselective techniques.

The above methods were partially driven by the shortcomings we observed in current epistasis detection methods. This makes them noteworthy contributions to the methodological aspect of epistasis. EpiGWAS uses causal inference to model interactions between distant loci in case-control cohorts. On the other hand, to improve biomarker discovery for continuous phenotypes, kernelPSI leverages kernel methods to model interactions within contiguous genomic regions. Despite this, both of them fall short of answering the ultimate goal of GWAS, which is biomarker selection and drug target discovery. We developed two use cases that go the last mile to illustrate our methods on real data. We applied EpiGWAS to the study of epistasis in multiple sclerosis, while we applied kernelPSI to compare the genetic mechanisms governing BMI and variation of ΔBMI. These two cases can be considered as templates to be adapted by end users for their own GWAS datasets.

1. **Multiple sclerosis**: is a neurodegenerative and inflammatory immune disease with severe health consequences. Thankfully, it is a rare disease with several easy-to-access GWAS. Among them, the WTCCC dataset contains 1500 cases and 2000 controls (Burton et al., 2007). We used this dataset to perform a systemic study of epistasis in MS (see Chapter 4). We evaluated all potential interactions in 15 MetaCore disease maps. For this study, we had to extend the original version of epiGWAS to gene-level epistasis detection through the aggregation of SNP-level scores. Our filtering of lead SNP pairs yielded several interactions to be analyzed. Among them, we found two known epistatic interactions. This proves the capacity of gene-level epiGWAS to detect relevant interactions, while simultaneously facilitating the biological analysis thanks to the availability of SNP annotations.

2. **Body-mass index**: our primary goal was to investigate that BMI and ΔBMI have different genetic roots (see Chapter 5). For this matter, we used the recently-released UK BioBank data (Bycroft et al., 2018). We proceeded by applying kernelPSI to both phenotypes on all genes related to BMI in the GWAS catalog (MacArthur et al., 2016). We have shown how the gene p-values in BMI and ΔBMI are not strongly correlated. Our study additionally enabled gene prioritization by dramatically lowering the initial number of related genes. For the sake of completeness, we included two baselines, against which kernelPSI favorably compared.

The burgeoning field of GWAS has been eager for novel statistical approaches, better bioinformatics tools and, most importantly, more data. Our two methods,

epiGWAS and kernelPSI, improved on existing methods and provided future directions of research. Nonetheless, we are far from solving the core problems facing GWAS and its future as a privileged destination for biological answers. We raise below a few issues that we think will become crucial in the next decade.

1. **Can GWAS results be replicated across populations?**
   The decrease in genotyping costs is making GWAS more inclusive. More multi-ethnic and non-Caucasian GWAS are currently being designed (Medina-Gomez et al., 2015). Population-specific discoveries are then possible. On the other hand, results replication will become even more complicated (Gonzalez et al., 2016). In particular, the comparison of results stemming from different populations poses several difficulties. If the same *loci* were found, this increases the confidence in the results. If the *loci* were different, it would be hard to discern false positives from population-specific SNPs. Concluding with high confidence will require more statistical power, and consequently, more participants from each population. For multi-ethnic GWAS, the answer lies in a more balanced representation between populations and better methods to infer the true causal *loci* by rigorously modeling population structure (Sul et al., 2018).

2. **Is chasing epistasis worthwhile?**
   This question runs against the very purpose of this thesis. Nonetheless, it is a question worth asking because of the inherent difficulty in detecting epistasis. As we have shown, it is challenging from both statistical and computational points of view. Validation is even harder. However, being one of the major hypotheses behind missing heritability (Zuk et al., 2012) will always make it an attractive endeavor. Some geneticists admit to the ubiquity of epistasis as a biological phenomenon, but argue that most genetic variation for quantitative traits is additive (Mackay, 2013). This would make the statistical analysis of epistasis essentially superfluous. We think that such conclusions do not reflect our limited knowledge of the underlying biology. Common diseases with their complex (and mostly unknown) architecture, and low effect sizes are hard to analyze in all settings. Moreover, the niche field of epistasis is still in its infancy, and is methodologically a difficult question to tackle. This explains why epistasis detection methods largely failed to deliver significant discoveries that would drum up enthusiasm for them.

3. **Can the original promise of GWAS be kept?**
   This is the second question that challenges the pertinence of GWAS. It is also a common interrogation among many biologists. The original promise of GWAS was to fully unravel the genetic background of complex diseases. The initial enthusiasm has been substituted with more skepticism, as the awareness of the difficulties of GWAS has increased. Thankfully, the picture is not entirely dim (Stranger et al., 2010). The number of *loci* discovered by GWAS is increasing at an ever quicker pace (Visscher et al., 2017). Moreover,

we have now more technologies that complement GWAS: high-throughput technologies, imaging in addition to a wide spectrum of easily-accessible information from network data, clinical trials, etc. This opens the door for new methodological developments to exploit heterogeneous sources of information in tandem. Some of these technologies are even used to validate the results of GWAS. Nonetheless, the original promise of GWAS is still far from being fulfilled.

I frequently mentioned the importance of the number of participants in GWAS (Spencer et al., 2009). The number of SNPs is also important and can lead to better results thanks to fine-mapping (Schaid et al., 2018) and the development of whole-genome sequencing (Gilly et al., 2018). Nonetheless, other statistical issues may arise because of linkage disequilibrium and high dimensionality.

4. **Are multi-omics approaches necessary?**
Multi-omics is one of the approaches that can enhance GWAS (Hasin et al., 2017). It corresponds to the combination of genome, proteome, transcriptome, epigenome, and microbiome. It can provide a broader picture and a more integrated approach across the different layers of biology. Mathematically, the combination of different input sources with non-redundant information for the same task can only benefit classification/regression performance. This probably explains the recent surge in the number of tools tackling multi-omics. Multi-omics is definitely the next frontier in GWAS (Wang et al., 2019). In particular, we look forward to developing a multi-omics variant of kernelPSI. Different kernels can be associated to different types of information *e.g.* genomic sequence data, gene expression, methylation data, etc. However, the success of such tools is heavily dependent on the availability and quality of data (Conesa & Beck, 2019).

5. **Can epistasis and GWAS systematically deliver drug targets?**
The objectives of GWAS are twofold: biomarker selection and therapeutic target discovery. As we explained above, GWAS have delivered a long list of biomarkers. For drug development, the results are more mixed (Visscher et al., 2017), as the transition from target discovery to market clearance is not that straightforward. It is a lengthy process that can take up to fifteen years (Morgan et al., 2011). Target discovery is followed by drug molecule design and activity testing. The initial research steps are then followed by clinical trials, the outcome of which is completely uncertain due to toxicity, side effects, and difficulties in patient recruitment. However, the contribution of genetic data to increasing the odds of success is undisputed (Nelson et al., 2015). Depending on the initial timepoint, the odds are increased by several folds. In the future, this is poised to further increase with the development of systems biology and virtual clinical trials (Smalley, 2018). As for epistasis, the development of combination therapies can only cement the need for its

study (Rochlani et al., 2017). For cancer in particular (Mokhtari et al., 2017), the development of combination therapies brought new hope for many patients.

The pharmaceutical world is currently suffering a productivity crisis because of the inefficiency of proposed targets (Bunnage, 2011). The combination of existing drugs is a first solution to improve patient outcomes in this case. A second solution is to complement an existing drug with a new one. The latter has often an effect only in presence of the former drug. Finally, the therapeutic potential of combination therapy, and as a consequence epistasis, is obvious, but its detection is still far from being solved.

6. **How far can statistical learning benefit healthcare?**
The primary focus of this thesis is epistasis, which is the instigator of our methodological contributions. Yet, the scope of their application is much larger. For instance, post-selection inference can be applied in pathway analysis to jointly perform gene cluster selection and pathway significance testing. Our approach kernelPSI can be easily extended to this setting by implementing graph kernels (Vishwanathan et al., 2010). Another promising application of kernelPSI are clinical trials where kernels can allow the integration of heterogeneous types of data. Beside statistical power, the interpretability of obtained models is of utmost importance. By recovering the key biomarkers, clinicians can better stratify patients, and reposition the treatment to new indications.

   In addition to post-selection inference, health sciences can also benefit from the emerging field of causal inference. For example, it has been used in pharmacovigilance for the identification of adverse drug effects (Agbabiaka et al., 2008), in cancer to discover distinct disease mechanisms underlying cancer subtypes (Xue et al., 2019), and in epidemiology to investigate the ecological drivers of disease emergence (Plowright et al., 2008). Currently, the major bottleneck in causal inference is its limitation to two-level treatment assignments. New approaches are being proposed for continuous (Fong et al., 2018) and multi-level (Yang et al., 2016b) treatments. Similarly, it would be interesting to extend epiGWAS to three-level bi-allelic SNPs.

Appendix A

EpiGWAS supplementary material

A.1 Genotypic hidden Markov model

Several authors (Scheet & Stephens, 2006; Sun et al., 2007; Rastas et al., 2005; Kimmel & Shamir, 2005) consider hidden Markov models more flexible for modeling linkage disequilibrium than block representations based on patterns of high LD. We also chose this model because regression models were severly overfitting because of the high dimensionality of the data, which was heavily skewing estimated propensity scores towards 0 and 1.

The hidden Markov model representation of the genome was developed to perform imputation, and has essentially remained confined to that application. For example, the fastPHASE software (Scheet & Stephens, 2006) based on this model leads to near-perfect imputation results, with error rates typically lower than 0.01. Among other applications, this representation has been used to construct knockoff copies of SNPs (Barber & Candès, 2015) to control the false discovery rate in GWAS (Sesia et al., 2018). The estimate of the propensity scores $\pi(A|X)$ is a new application of this representation in the context of GWAS.

In this Appendix, we explicit the transition and emission probabilities for the genotypic hidden Markov model. For that purpose, we start by considering a pair of ordered haplotypes $H^a = \left(H_1^a, \cdots, H_p^a\right) \in \{0,1\}^p$ and $H^b = \left(H_1^b, \cdots, H_p^b\right) \in \{0,1\}^p$. We recall that the two haplotypes correspond to the same positions. The hidden variables $Z^a = \left(Z_1^a, \cdots, Z_p^a,\right)$ and $Z^b = \left(Z_1^b, \cdots, Z_p^b,\right)$ represent cluster memberships. They take discrete values in $\{1, \cdots, K\}^p$. Scheet and Stephens Scheet & Stephens (2006) define the clusters as a "(common) combination of alleles at tightly linked SNPs". The underlying hidden Markov models for the two alleles have identical forms. We then focus on the first allele a. We follow the notations of Sesia et al. (2018).

The marginal distribution of the first hidden state can be written as:

$$q_1^{hap}(k) = \alpha_{1,k}, \quad k \in \{1, \cdots, K\}.$$

For $j \in \{2, \cdots, p\}$, the transition matrix Q_j^{hap} is given by:

$$Q_j^{hap}(k'|k) = P(H_j = k'|H_{j-1} = k') = \begin{cases} e^{-r_j} + (1 - e^{-r_j})\, \alpha_{j,k'}, & k' = k \\ (1 - e^{-r_j})\, \alpha_{j,k'}, & k' \neq k \end{cases}.$$

The parameter $r = (r_2, \cdots, r_p)$ can be assimilated to the recombination rate between loci $j-1$ and j, although Scheet and Stephens Scheet & Stephens (2006) point out the general mismatch between the observed recombination rates and the estimate of r. The parameter $\alpha = (\alpha_{j,k})_{(j,k)\in\{1,\cdots p\}\times\{1,\cdots,K\}}$ is the relative frequency of the cluster k in locus j.

Conditionally on the latent state $Z_j^{hap} = z_j$, the allele H_j is a Bernoulli random variable, $H_j|Z_j \sim \mathcal{B}(\theta_{j,z_j})$. θ_{j,z_j} is the frequency of allele 1 in cluster z_j at the position j:

$$f_j^{hap} = (h_j; z_j, \theta) = \begin{cases} 1-\theta_{j,z_j}, & h_j = 0 \\ \theta_{j,z_j}, & h_j = 1 \end{cases}.$$

Under the Hardy-Weinberg equilibrium (HWE), a third hidden Markov model for the unphased genotype can be derived by combining the HMMs of the two alleles a and b. The emission states $X = (X_1, \cdots, X_p) \in \{0,1,2\}^p$ are given by the sum of the emission states, $H^a + H^b = \left(H_1^a + H_1^b, \cdots, H_p^a + H_p^b\right)$. Because of the phase indetermination, the latent states are unordered pairs of haplotype latent states, $Z = (\{Z_1^a, Z_1^b\}, \cdots, \{Z_p^a, Z_p^b\})$. Thus, the dimensionality of the latent variable space is $K(K+1)/2$. The different probabilities of the genotype model are computed by considering the two cases: $Z_j^a = Z_j^b$ and $Z_j^a \neq Z_j^b$.

The initial latent state distribution is given by:

$$q_1^{gen}(\{k^a, k^b\}) = \begin{cases} (\alpha_{1,k^a})^2, & k^a = k^b \\ 2\alpha_{1,k^a}\alpha_{1,k^b} & k^a \neq k^b \end{cases},$$

In a similar fashion, the transition probabilities:

$$Q_j^{gen}(\{\underline{k}^a, \underline{k}^b\}|\{k^a, k^b\}) = \begin{cases} Q_j^{hap}(\underline{k}^a|k^a)Q_j^{hap}(\underline{k}^b|k^b) + Q_j^{hap}(\underline{k}^b|k^a)Q_j^{hap}(\underline{k}^a|k^b), & \underline{k}^a \neq \underline{k}^b \\ Q_j^{hap}(\underline{k}^a|k^a)Q_j^{hap}(\underline{k}^b|k^b), & \text{otherwise} \end{cases},$$

and, the emission probabilities are

$$f_j(x_j; \{k^a, k^b\}, \theta) = \begin{cases} (1-\theta_{j,k^a})(1-\theta_{j,k^b}), & x_j = 0 \\ \theta_{j,k^a}(1-\theta_{j,k^b}) + (1-\theta_{j,k^a})\theta_{j,k^b}, & x_j = 1 \\ \theta_{j,k^a}\theta_{j,k^b}, & x_j = 2 \end{cases}.$$

For the estimate of the parameters $\nu = (\alpha, r, \theta)$, we use the imputation software fastPHASE (Scheet & Stephens, 2006) which fits the hidden Markov model using an expectation-maximization (EM) algorithm (Dempster et al., 1977). Its computational complexity is $\mathcal{O}\left(npK^2\right)$. The complexity scales linearly for both p and n, rendering fastPHASE well-suited for real case-control datasets where the number of SNPs is typically in the hundreds of thousands and the number of samples in the thousands. In practice, as a trade-off between a rich representation of the clusters and the ensuing quadratic complexity, we chose $K = 12$.

A.2 Additional simulation results for epiGWAS

A.2.1 First scenario: synergistic only effects

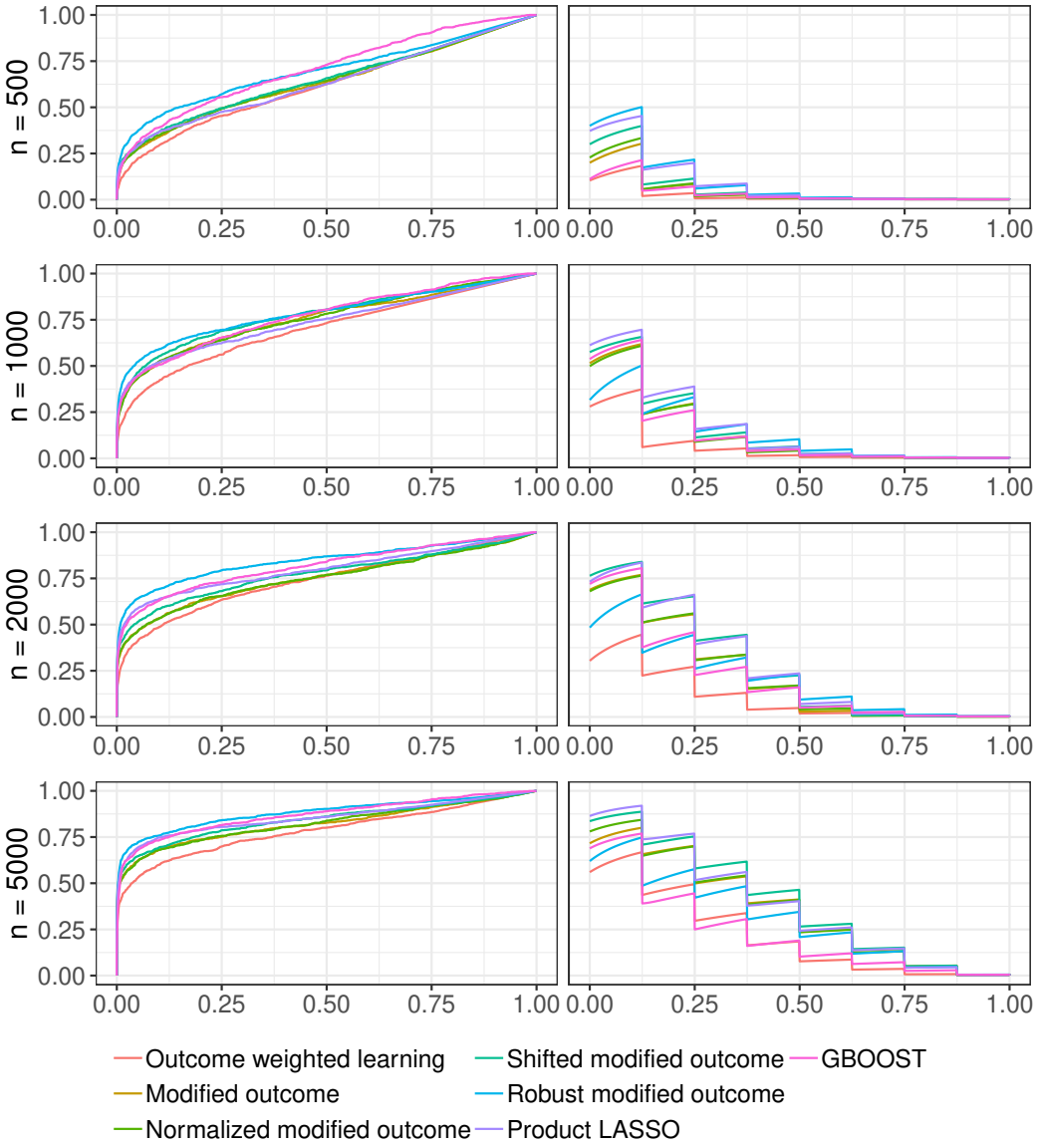


Figure A.1: Average ROC (left column) and PR (right column) curves for the first scenario

Table A.1: Average ROC and PR AUCs for the first scenario

Method	PR	ROC
n =500		
GBOOST	0.0362	0.7075
Modified outcome	0.0468	0.6747
Robust modified outcome	0.0973	0.7414
Normalized modified outcome	0.0512	0.6754
Shifted modified outcome	0.0644	0.6794
Outcome weighted learning	0.0254	0.6282
Product LASSO	0.0895	0.6514
n =1000		
GBOOST	0.1270	0.7688
Modified outcome	0.1284	0.7131
Robust modified outcome	0.1302	0.7434
Normalized modified outcome	0.1255	0.7120
Shifted modified outcome	0.1470	0.7224
Outcome weighted learning	0.0613	0.6764
Product LASSO	0.1619	0.7032
n =2000		
GBOOST	0.2103	0.8169
Modified outcome	0.2252	0.7512
Robust modified outcome	0.2070	0.8449
Normalized modified outcome	0.2266	0.7501
Shifted modified outcome	0.2704	0.7753
Outcome weighted learning	0.1045	0.7394
Product LASSO	0.2711	0.7989
n =5000		
GBOOST	0.2276	0.8697
Modified outcome	0.3512	0.8218
Robust modified outcome	0.3011	0.8818
Normalized modified outcome	0.3548	0.8248
Shifted modified outcome	0.3907	0.8423
Outcome weighted learning	0.2139	0.7847
Product LASSO	0.3779	0.8546

A.2.2 Second scenario: partial overlap between marginal and synergistic effects

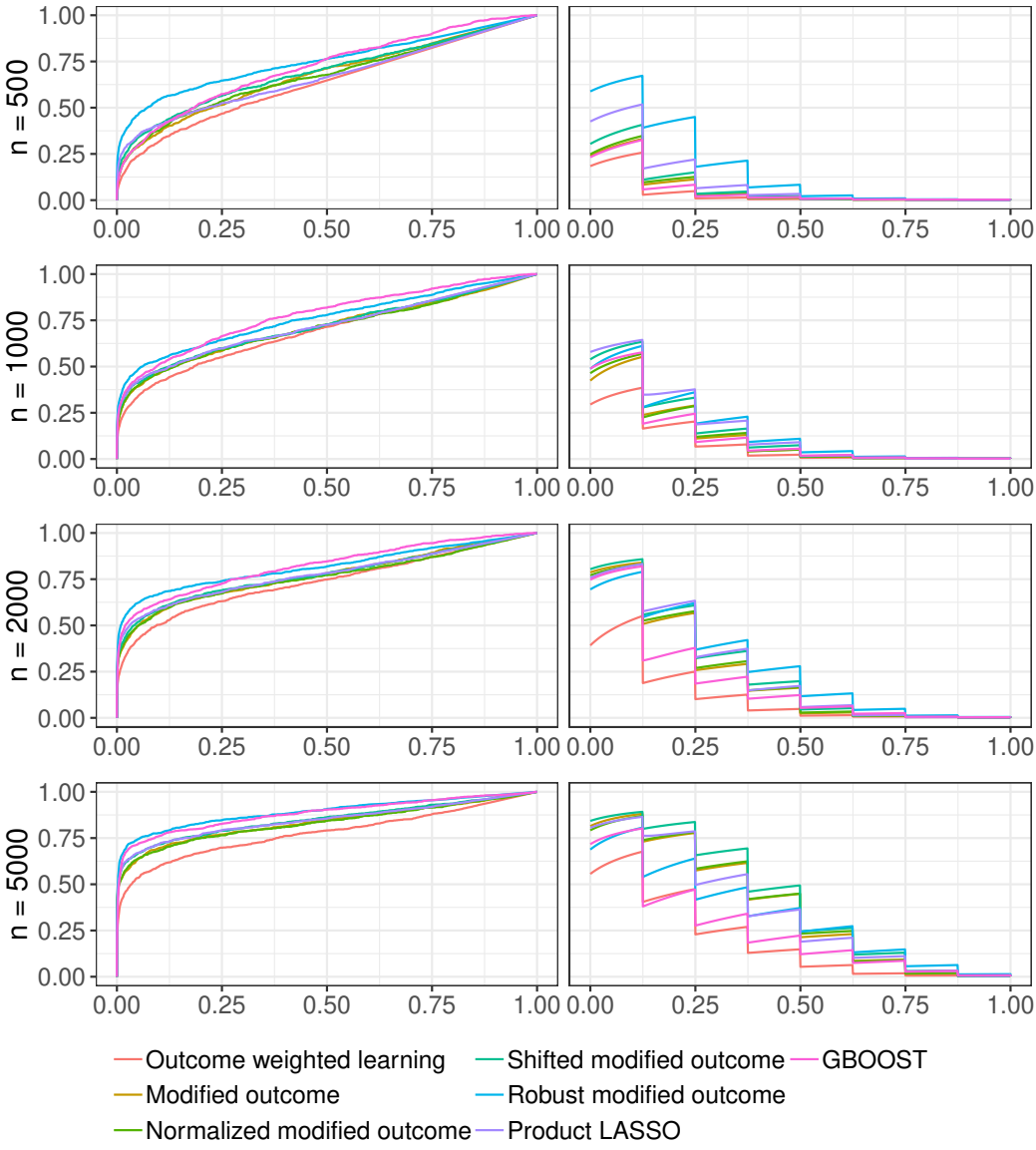


Figure A.2: Average ROC (left column) and PR (right column) curves for the second scenario

Table A.2: Average ROC and PR AUCs for the second scenario

Method	PR	ROC
n =500		
GBOOST	0.0516	0.7186
Modified outcome	0.0563	0.6750
Robust modified outcome	0.1716	0.7502
Normalized modified outcome	0.0590	0.6713
Shifted modified outcome	0.0712	0.6918
Outcome weighted learning	0.0367	0.6345
Product LASSO	0.0994	0.6659
n =1000		
GBOOST	0.1190	0.7773
Modified outcome	0.1195	0.7092
Robust modified outcome	0.1574	0.7601
Normalized modified outcome	0.1233	0.7080
Shifted modified outcome	0.1443	0.7160
Outcome weighted learning	0.0805	0.6923
Product LASSO	0.1609	0.7170
n =2000		
GBOOST	0.1933	0.8226
Modified outcome	0.2294	0.7708
Robust modified outcome	0.2732	0.8183
Normalized modified outcome	0.2321	0.7623
Shifted modified outcome	0.2532	0.7753
Outcome weighted learning	0.1114	0.7360
Product LASSO	0.2507	0.7762
n =5000		
GBOOST	02454	0.8821
Modified outcome	0.3718	0.8344
Robust modified outcome	0.3286	0.8916
Normalized modified outcome	0.3739	0.8309
Shifted modified outcome	0.4079	0.8487
Outcome weighted learning	0.1930	0.7769
Product LASSO	0.3537	0.8467

A.2.3 Third scenario: partial overlap between quadratic and synergistic effects

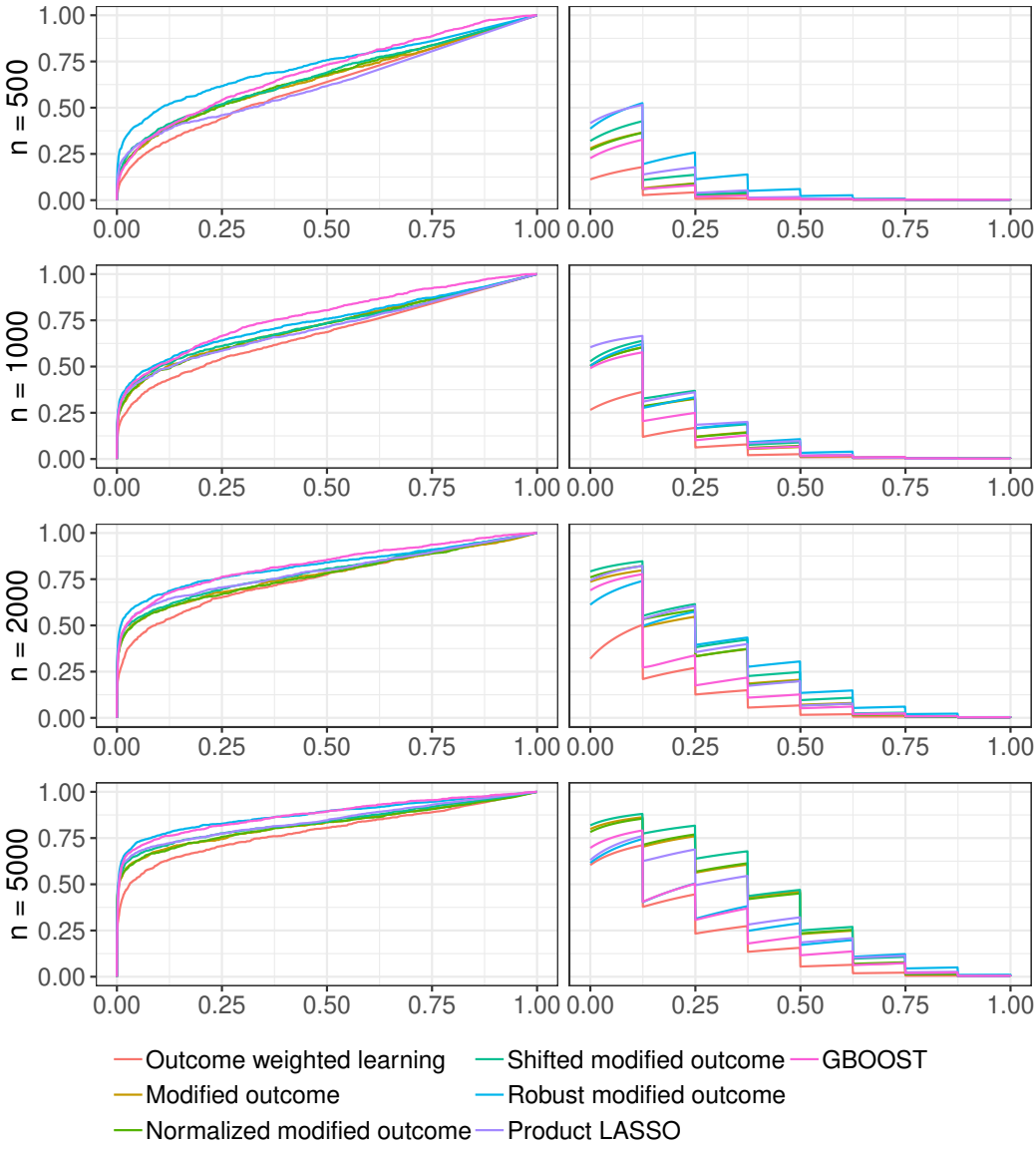


Figure A.3: Average ROC (left column) and PR (right column) curves for the third scenario

Table A.3: Average ROC and PR AUCs for the third scenario

Method	PR	ROC
n =500		
GBOOST	0.050	0.6970
Modified outcome	0.0570	0.6559
Robust modified outcome	0.1148	0.7296
Normalized modified outcome	0.0569	0.6627
Shifted modified outcome	0.0714	0.6703
Outcome weighted learning	0.0260	0.6233
Product LASSO	0.0889	0.6282
n =1000		
GBOOST	0.1228	0.7746
Modified outcome	0.1362	0.7181
Robust modified outcome	0.1513	0.7444
Normalized modified outcome	0.1373	0.7175
Shifted modified outcome	0.1546	0.7226
Outcome weighted learning	0.0728	0.6778
Product LASSO	0.1620	0.7100
n =2000		
GBOOST	0.1814	0.8307
Modified outcome	0.2430	0.7733
Robust modified outcome	0.2697	0.8235
Normalized modified outcome	0.2496	0.7724
Shifted modified outcome	0.2737	0.7886
Outcome weighted learning	0.1129	0.7535
Product LASSO	0.2543	0.7921
n =5000		
GBOOST	0.2467	0.8767
Modified outcome	0.3663	0.8241
Robust modified outcome	0.2660	0.8790
Normalized modified outcome	0.3669	0.8236
Shifted modified outcome	0.3944	0.8376
Outcome weighted learning	0.1965	0.7893
Product LASSO	0.3158	0.8439

A.2.4 Fourth scenario: partial overlap between quadratic/marginal and synergistic effects

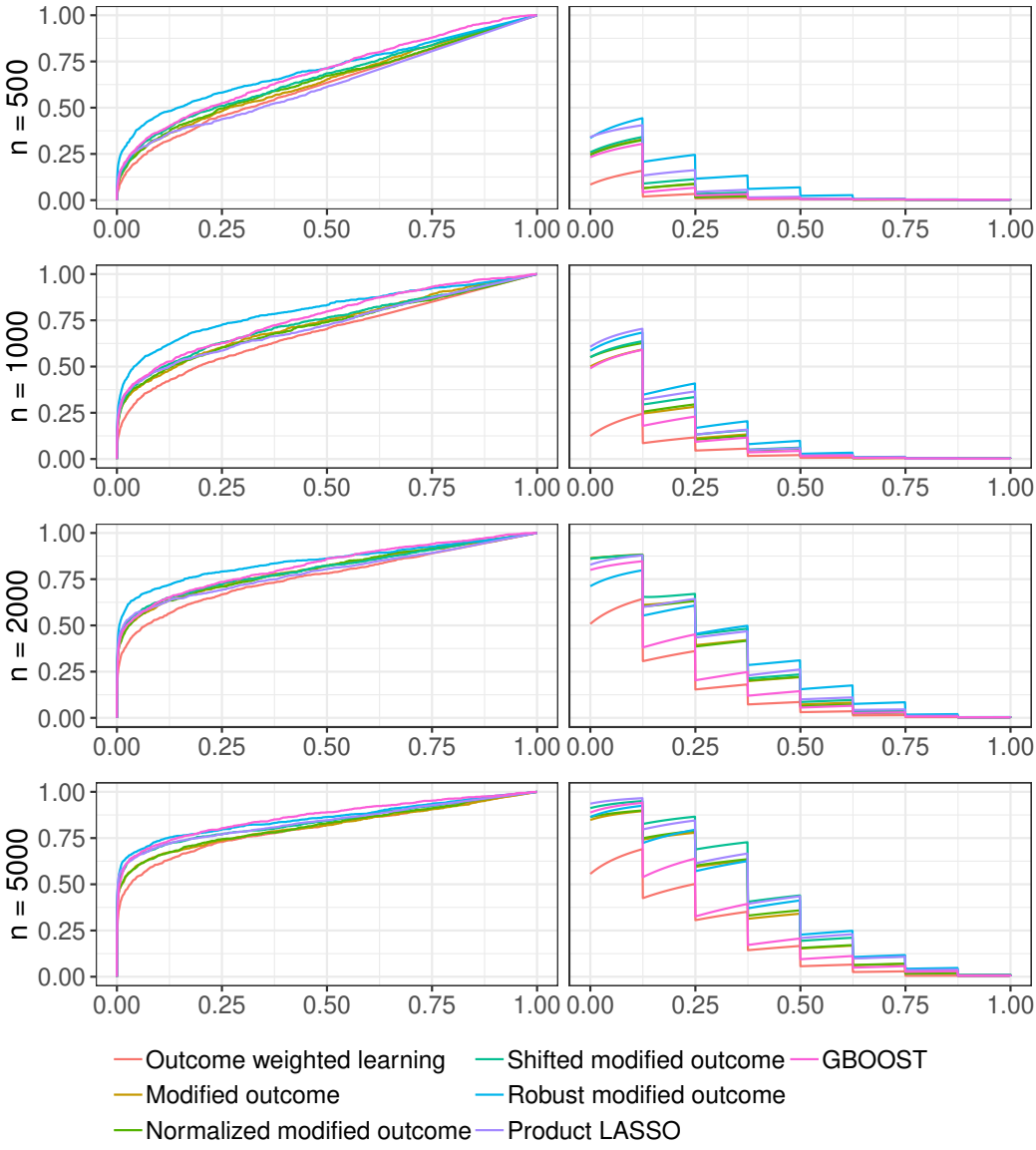


Figure A.4: Average ROC (left column) and PR (right column) curves for the fourth scenario

Table A.4: Average ROC and PR AUCs for the fourth scenario

Method	PR	ROC
n =500		
GBOOST	0.0479	0.6900
Modified outcome	0.0521	0.6427
Robust modified outcome	0.1066	0.7065
Normalized modified outcome	0.0513	0.6460
Shifted modified outcome	0.0591	0.6623
Outcome weighted learning	0.0227	0.6218
Product LASSO	0.0762	0.6174
n =1000		
GBOOST	0.1163	0.7647
Modified outcome	0.1283	0.7288
Robust modified outcome	0.1687	0.8049
Normalized modified outcome	0.1338	0.7200
Shifted modified outcome	0.1438	0.7388
Outcome weighted learning	0.0479	0.6838
Product LASSO	0.1554	0.7206
n =2000		
GBOOST	0.2129	0.8237
Modified outcome	0.2794	0.8007
Robust modified outcome	0.2986	0.8478
Normalized modified outcome	0.2763	0.8032
Shifted modified outcome	0.2960	0.8050
Outcome weighted learning	0.1530	0.7641
Product LASSO	0.2927	0.7899
n =5000		
GBOOST	0.2823	0.8656
Modified outcome	0.3541	0.8127
Robust modified outcome	0.3823	0.8568
Normalized modified outcome	0.3597	0.8175
Shifted modified outcome	0.4091	0.8388
Outcome weighted learning	0.2106	0.8031
Product LASSO	0.4000	0.8399

Appendix B

KernelPSI supplementary material

B.1 Proof of Lemma 3.1

In this Appendix, for a linear kernel of the outcome $L = YY^T$, we detail the necessary steps to transform the empirical HSIC estimators into a quadratic form. For the biased estimator, the result is straightforward:

$$\begin{aligned}\widehat{\text{HSIC}}_{\text{biased}} &= \frac{1}{(n-1)^2}\ \text{trace}(K\Pi_n L\Pi_n) \\ &= \frac{1}{(n-1)^2} Y^T(\Pi_n K\Pi_n)Y \\ &= Y^T Q_{\text{biased}} Y,\end{aligned}$$

where $\Pi_n = I_{n\times n} - \frac{1}{n}1_n {1_n}^T$.

For the unbiased estimator, the calculations are more tedious:

$$\widehat{\text{HSIC}}_{\text{unbiased}}(K,L) = \frac{1}{n(n-3)}\left[\text{trace}(\underline{K}\ \underline{L}) + \frac{1^T\underline{K}1\ 1^T\underline{L}1}{(n-1)(n-2)} - \frac{2}{n-2}1^T\underline{K}\ \underline{L}1\right],$$

where $\underline{K} = K - \text{diag}(K)$ and $\underline{L} = L - \text{diag}(L)$. The diagonal matrices $\text{diag}(K)$ and $\text{diag}(L)$ can be respectively rewritten as : $\text{diag}(K) = \sum_{i=1}^n P^{(i)}KP^{(i)}$ and $\text{diag}(L) = \sum_{i=1}^n P^{(i)}LP^{(i)}$. $P^{(i)}$ is the projection on the i^{th} coordinate. We remark that $P^{(i)}P^{(j)} = \delta_{ij}P^{(i)}$.

We now develop each term of the previous equation.

$$
\begin{aligned}
\operatorname{trace}(\underline{K}\ \underline{L}) &= \operatorname{trace}\left(KL - \sum_{i=1}^{n} KP^{(i)}LP^{(i)} - \sum_{i=1}^{n} P^{(i)}KP^{(i)}L + \sum_{i,j=1}^{n} P^{(j)}KP^{(j)}P^{(i)}LP^{(i)}\right) \\
&= \operatorname{trace}(KL) - \operatorname{trace}\left(\sum_{i=1}^{n} P^{(i)}KP^{(i)}L\right) - \operatorname{trace}\left(\sum_{i=1}^{n} P^{(i)}KP^{(i)}L\right) \\
&+ \operatorname{trace}\left(\sum_{i,j=1}^{n} P^{(j)}KP^{(j)}P^{(i)}LP^{(i)}\right) \\
&= \operatorname{trace}(KL) - 2\ \operatorname{trace}\left(\sum_{i=1}^{n} P^{(i)}KP^{(i)}L\right) + \operatorname{trace}\left(\sum_{i}^{n} P^{(i)}KP^{(i)}LP^{(i)}\right) \\
&= \operatorname{trace}(KL) - \operatorname{trace}\left(\sum_{i=1}^{n} P^{(i)}KP^{(i)}L\right) \\
&= Y^T KY - Y^T\left(\sum_{i=1}^{n} P^{(i)}KP^{(i)}\right)Y \\
&= Y^T(K - K_P)Y \qquad \text{with} \qquad K_P = \sum_{i=1}^{n} P^{(i)}KP^{(i)} \\
\operatorname{trace}(\underline{K}\ \underline{L}) &= Y^T K_1 K
\end{aligned}
$$

Similarly, we obtain:

$$
\begin{aligned}
1^T\underline{K}1 &= 1^T\left(K - K_P\right)1 = c_X \\
1^T\underline{L}1 &= 1^T\left(L - \sum_{i=1}^{n} P^{(i)}LP^{(i)}\right)1 \\
&= Y^T 11^T Y\ - \operatorname{trace}\left(11^T \sum_{i=1}^{n} P^{(i)}LP^{(i)}\right) \\
&= Y^T\left(11^T - \sum_{i=1}^{n} P^{(i)}11^T P^{(i)}\right)Y \\
1^T\underline{L}1 &= Y^T K_2 Y
\end{aligned}
$$

As for the last term:

$$
\begin{aligned}
1^T \underline{K}\ \underline{L} 1 &= \operatorname{trace}(11^T \underline{K}\ \underline{L}) \\
&= \operatorname{trace}\left(11^T \left(K - \sum_{i=1}^{n} P^{(i)} K P^{(i)}\right)\ \left(L - \sum_{j=1}^{n} P^{(j)} L P^{(j)}\right)\right) \\
&= \operatorname{trace}(11^T K\ L) - \operatorname{trace}\left(11^T K \sum_{i=1}^{n} P^{(i)} L P^{(i)}\right) - \operatorname{trace}\left(11^T \sum_{i=1}^{n} P^{(i)} K P^{(i)} L\right) \\
&+ \operatorname{trace}\left(11^T \sum_{i=1}^{n} P^{(i)} K P^{(i)} \sum_{j=1}^{n} P^{(j)} L P^{(j)}\right) \\
&= \operatorname{trace}(11^T K\ L) - \operatorname{trace}\left(\sum_{i=1}^{n} P^{(i)} 11^T K P^{(i)} L\right) - \operatorname{trace}\left(\sum_{i=1}^{n} 11^T P^{(i)} K P^{(i)} L\right) \\
&+ \operatorname{trace}\left(11^T \sum_{i=1}^{n} P^{(i)} K P^{(i)} L P^{(i)}\right) \\
&= Y^T (11^T K) Y - Y^T \left(\sum_{i=1}^{n} P^{(i)} 11^T K P^{(i)}\right) Y - Y^T \left(\sum_{i=1}^{n} 11^T P^{(i)} K P^{(i)}\right) Y \\
&+ Y^T \left(\sum_{i=1}^{n} P^{(i)} 11^T P^{(i)} K P^{(i)}\right) Y \\
1^T \underline{K}\ \underline{L} 1 &= Y^T K_3 Y
\end{aligned}
$$

That yields the following quadratic form:

$$
\begin{aligned}
\widehat{\mathrm{HSIC}}_{\mathrm{unbiased}}(X, Y) &= \frac{1}{n(n-3)} \left[Y^T K_1 Y + c_X \frac{Y^T K_2 Y}{(n-1)(n-3)} - \frac{2}{n-2} Y^T K_3 Y \right] \\
&= Y^T Q_{\mathrm{unbiased}} Y
\end{aligned}
$$

B.2 Proof of Theorem 3.1

For a quadratic kernel association score $s(K,Y) = Y^T Q(K) Y$, we represent the three kernel selection strategies as an intersection of quadratic constraints.

For marginal screening, we can write the selection event of the top S' kernels $i_1, \cdots, i_{S'}$ in the following way :

$$E_{S'}^{\text{screening}} = \bigcap_{l=1}^{S'-1} \left\{ Y^T Q(K_{i_l}) Y \geq Y^T Q(K_{i_{l+1}}) Y \right\} \cap \bigcap_{l' \notin \{i_1, \cdots, i_{S'}\}} \left\{ Y^T Q(K_{i_{S'}}) Y \geq Y^T Q(K_{l'}) Y \right\}$$

In Yamada et al. Yamada et al. (2018), the authors obtain $S'(S-S')$ constraints by comparing the association score of each selected kernel to the association scores of all discarded kernels. Here, by conditioning on the order of selection of the kernels, we only obtain $S-1$ constraints in $E_{S'}^{\text{screening}}$.

For forward stepwise selection (Algorithm 3.1), we first start by modeling an intermediate step s. The selection of the kernel K_{i_s} is equivalent to the following selection event:

$$\bigcap_{\substack{i \notin \mathcal{J}^{(s-1)} \\ i \neq i_s}} \left\{ Y \text{ s.t. } Y^T Q(K_{\mathcal{J}^{(s-1)} \cup \{i_s\}}) Y \geq Y^T Q(K_{\mathcal{J}^{(s-1)} \cup \{i\}}) Y \right\},$$

where $\mathcal{J}^{(m)}$ represents the set of selected kernels at step m and $K_{\mathcal{A}} = \sum_{p \in \mathcal{A}} K_p$ for a subset $\mathcal{A}$ of $\{1, \cdots, S\}$.

We can then recursively define the event E_s, representing the selection of $s \leq S'$ groups:

$$E_s^{\text{forward}} = E_{s-1}^{\text{forward}} \cap \bigcap_{\substack{i \notin \mathcal{J}^{(s-1)} \\ i \neq i_s}} \left\{ Y \text{ s.t. } Y^T Q(K_{\mathcal{J}^{(s-1)} \cup \{i_s\}}) Y \geq Y^T Q(K_{\mathcal{J}^{(s-1)} \cup \{i\}}) Y \right\}$$

For $s = S'$, we then obtain a conjunction of quadratic constraints.

For backward selection (Algorithm 3.2), we can derive a similar set of recursive constraints to model the elimination of the kernels $\mathcal{I}^{(s)} = \{i_1, \cdots, i_s\}$:

$$E_s^{\text{backward}} = E_{s-1}^{\text{backward}} \cap \bigcap_{\substack{i \notin \mathcal{I}^{(s-1)} \\ i \neq i_s}} \left\{ Y \text{ s.t. } Y^T Q(K^-_{\mathcal{I}^{(s-1)} \cup \{i_s\}}) Y \geq Y^T Q(K^-_{\mathcal{I}^{(s-1)} \cup \{i\}}) Y \right\},$$

where $K^-_{\mathcal{A}} = \sum_{p \in \mathcal{A}^c} K_p$. The set $\mathcal{A}^c$ is the complement of $\mathcal{A}$ in $\{1, \cdots, S\}$.

To model the a posteriori choice of S' in the adaptive variants, an additional set of constraints must be introduced in the selection event. In Equation (B.1), we model the selection event $E_{\text{adaptive}}^{\text{forward}}$ corresponding to the adaptive extension of

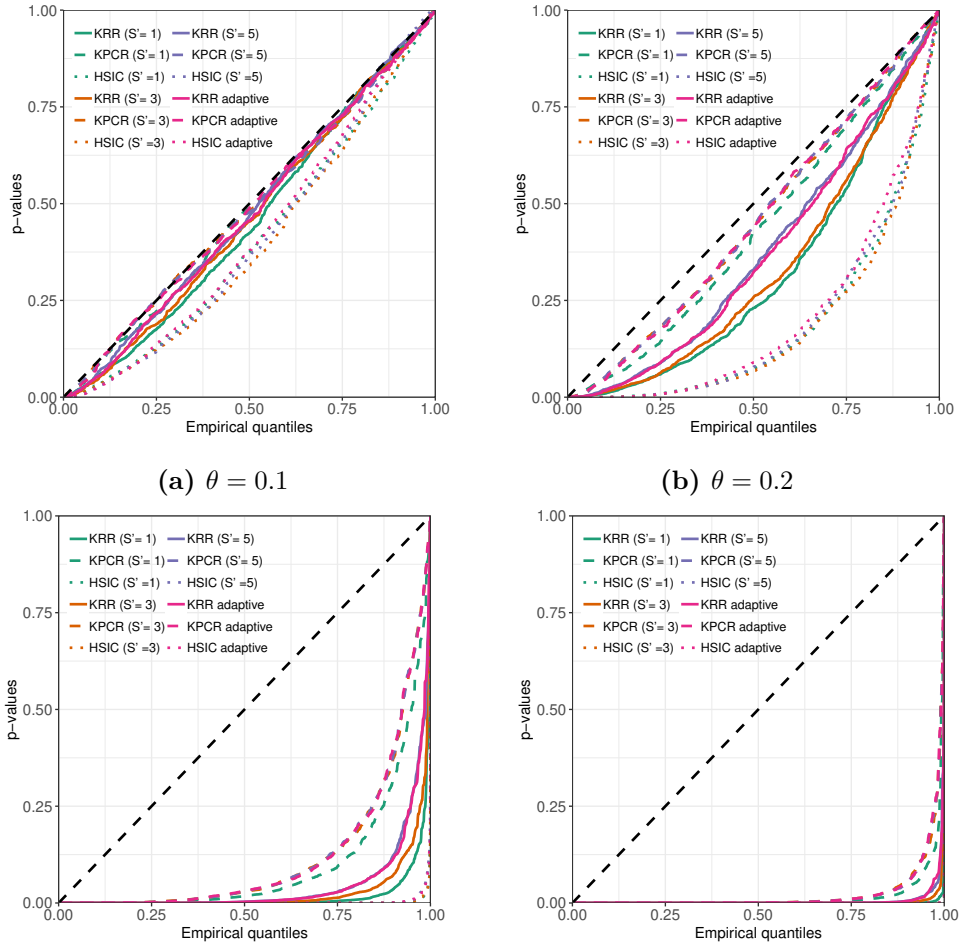
forward stepwise selection. The quadratic set of constraints in E_S^{forward} represents the order of selection of the kernels $\mathcal{J}^{(S)} = \{i_1, \cdots, i_S\}$, while the intersection of the other constraints represents the selection of the number of kernels S'. The backward version $E_{\text{adaptive}}^{\text{backward}}$ can be easily deduced in a similar fashion.

$$E_{\text{adaptive}}^{\text{forward}} = E_S^{\text{forward}} \cap \bigcap_{\substack{m=1 \\ m \neq S'}}^{S} \left\{ Y \text{ s.t. } Y^T Q(K_{\mathcal{J}^{(S')}}) Y \geq Y^T Q(K_{\mathcal{J}^{(m)}}) Y \right\} \tag{B.1}$$

The result in Theorem 3.1 is more general by adding to the quadratic form a constant, which can be used as a form of penalization. The above proof can be easily extended to the setting of Theorem 3.1.

B.3 Additional experiments on kernelPSI

B.3.1 Statistical validity: Statistical power of kernelPSI for different effect sizes, on simulated data

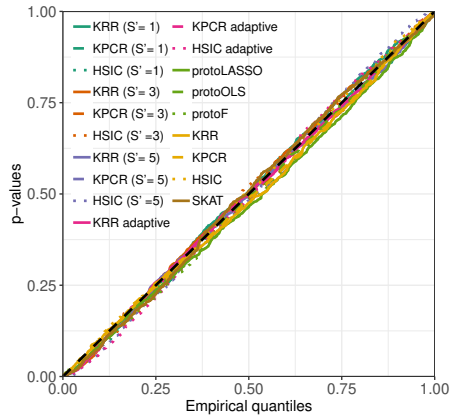


(a) $\theta = 0.1$ **(b)** $\theta = 0.2$

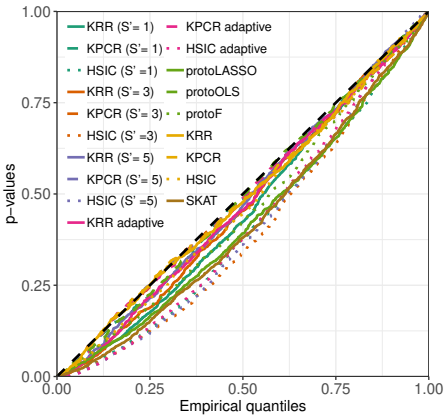
(c) $\theta = 0.4$ **(d)** $\theta = 0.5$

Figure B.1: Q-Q plots comparing the empirical kernelPSI p-values distributions under the alternative hypothesis to the uniform distribution, for different effect sizes θ. The data is generated as described in Section 3.7.1.

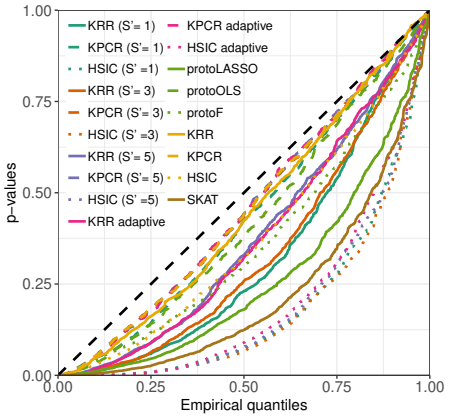
B.3.2 Benchmarking for the first configuration: using Gaussian kernels over simulated Gaussian data



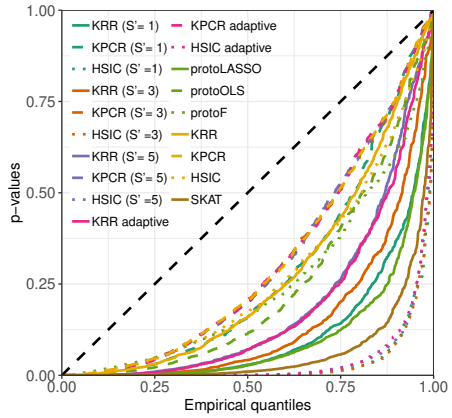
(a) $\theta = 0.0$



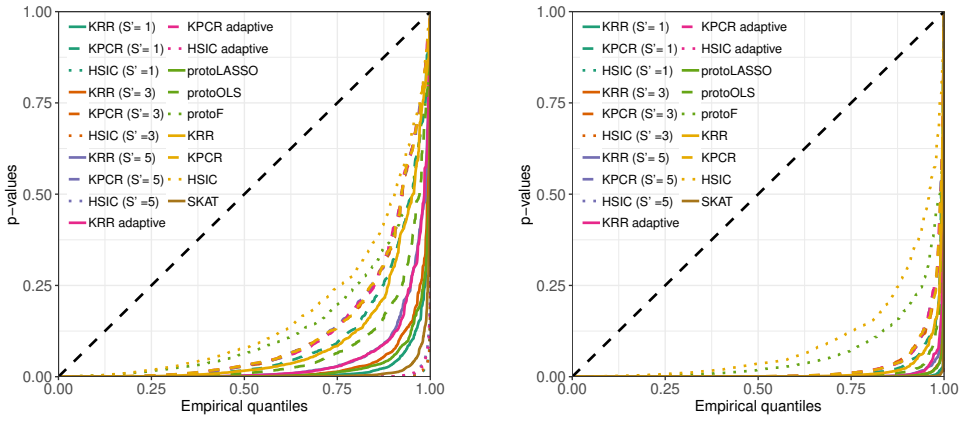
(b) $\theta = 0.1$



(c) $\theta = 0.2$



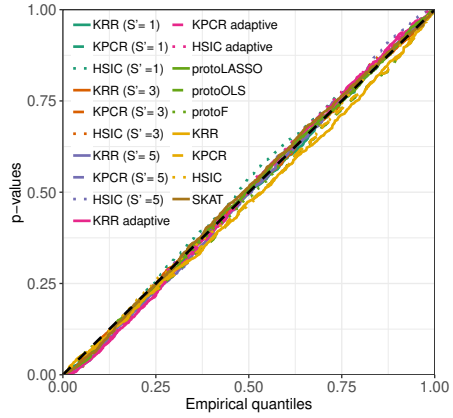
(d) $\theta = 0.3$



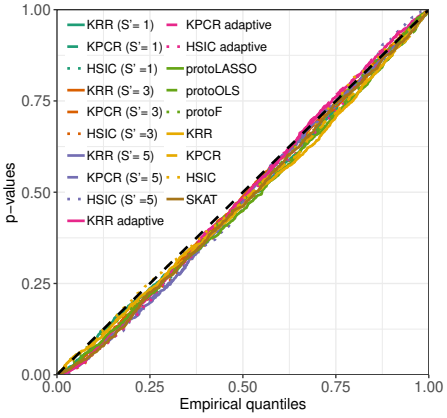
(e) $\theta = 0.4$ **(f)** $\theta = 0.5$

Figure B.2: Q-Q plots comparing the empirical kernelPSI and benchmarking p-values distributions under the null ($\theta = 0$) or alternative hypothesis ($\theta > 0$) to the uniform distribution, for different effect sizes θ, using Gaussian kernels for simulated Gaussian data. The data generation and benchmarked methods are described in Section 3.7.2.

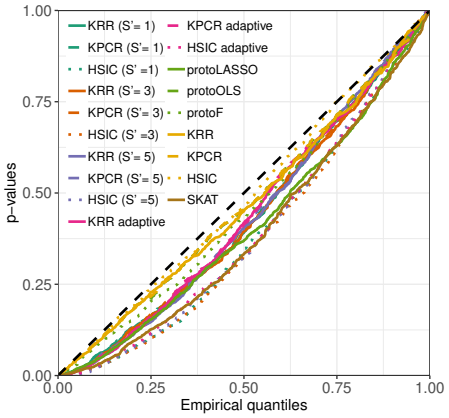
B.3.3 Benchmarking for the second configuration: using linear kernels over simulated binary data



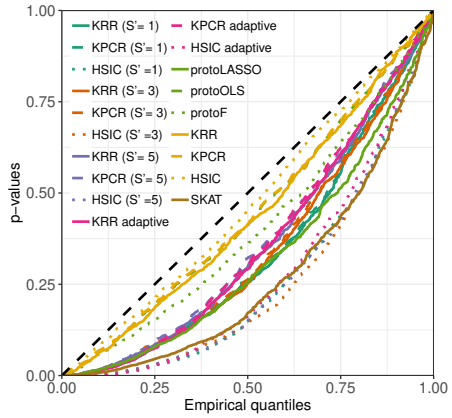
(a) $\theta = 0.0$



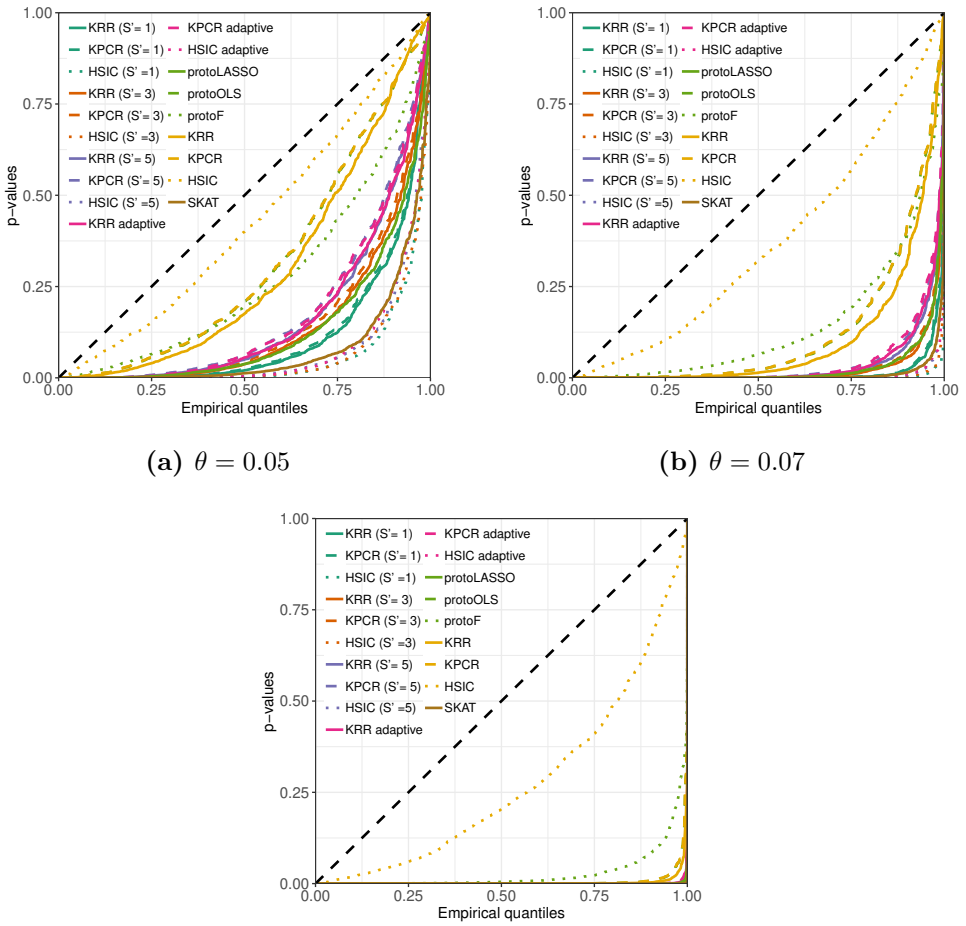
(b) $\theta = 0.01$



(c) $\theta = 0.02$



(d) $\theta = 0.03$



(a) $\theta = 0.05$

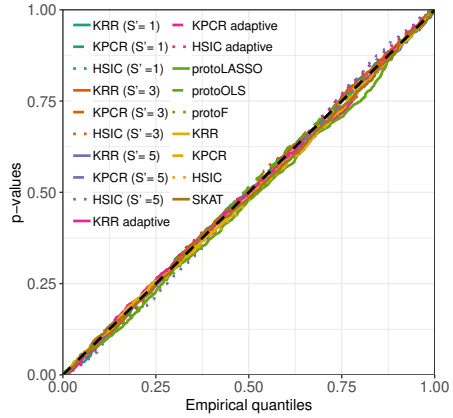
(b) $\theta = 0.07$

(c) $\theta = 0.10$

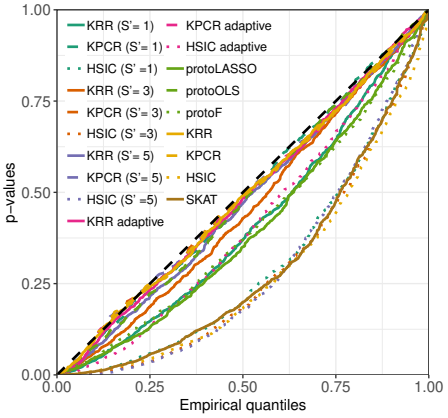
Figure B.4: Q-Q plots comparing the empirical kernelPSI and benchmarking p-values distributions under the null ($\theta = 0$) or alternative hypothesis ($\theta > 0$) to the uniform distribution, for different effect sizes θ, using linear kernels for simulated binary data. The data generation and benchmarked methods are described in Section 3.7.2.

B.3.4 Benchmarking for the third configuration: using Gaussian kernels over simulated Swiss roll data

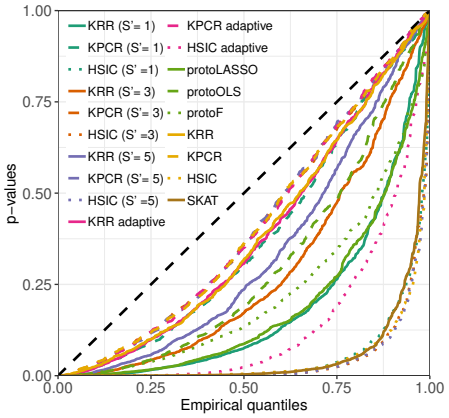
B.3.4.1 Statistical validity: Q-Q plots for various effect sizes



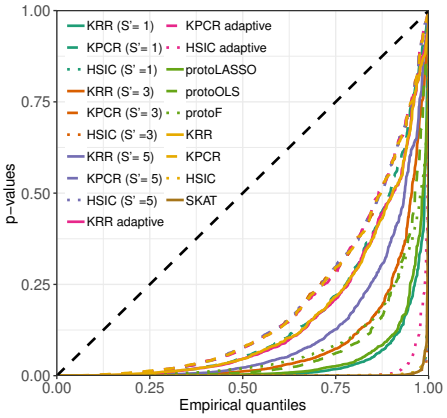
(a) $\theta = 0.0$



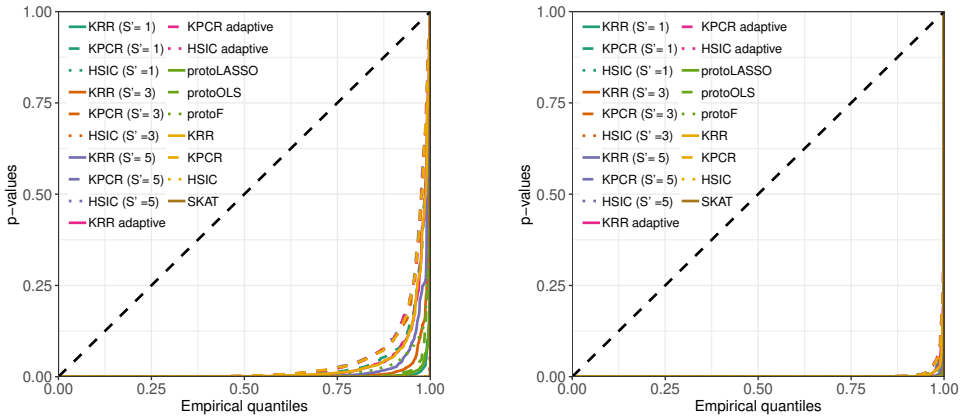
(b) $\theta = 0.1$



(c) $\theta = 0.2$



(d) $\theta = 0.3$



(e) $\theta = 0.4$ **(f)** $\theta = 0.5$

Figure B.5: Q-Q plots comparing the empirical kernelPSI and benchmarking p-values distributions under the null ($\theta = 0$) or alternative hypothesis ($\theta > 0$) to the uniform distribution, for different effect sizes θ, using Gaussian kernels for simulated Swiss roll data. The data generation and benchmarked methods are described in Section 3.7.2.

B.3.4.2 Evolution of the statistical power as a function of the effect size

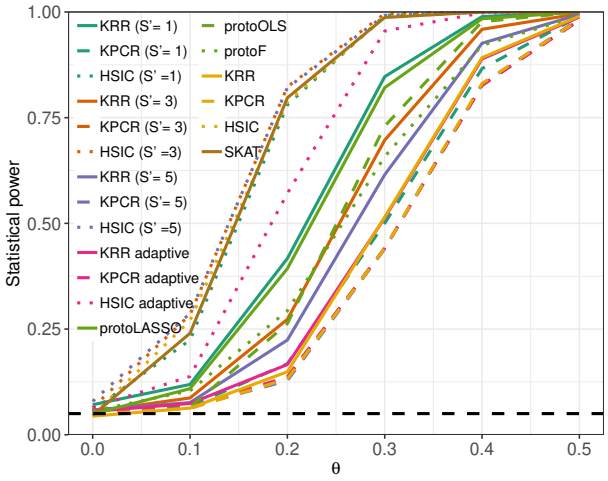


Figure B.6: Statistical power of kernelPSI variants and benchmark methods, using Gaussian kernels for simulated Swiss roll data.

B.3.5 Kernel selection performance

Table B.1: Ability of the kernel selection procedure to recover the true causal kernels, using linear kernels over with binary data.

	$\boldsymbol{\theta}$	$S' = 1$	$S' = 3$	$S' = 5$	Adaptive
Recall	0.0	0.102	0.316	0.529	0.390
	0.01	0.117	0.341	0.553	0.414
	0.02	0.163	0.388	0.590	0.482
	0.03	0.224	0.445	0.632	0.559
	0.05	0.311	0.512	0.683	0.639
	0.07	0.332	0.542	0.713	0.686
	0.1	0.333	0.581	0.759	0.750
Precision	0.0	0.308	0.316	0.317	0.320
	0.01	0.351	0.341	0.331	0.335
	0.02	0.489	0.388	0.354	0.377
	0.03	0.673	0.445	0.379	0.422
	0.05	0.935	0.512	0.409	0.463
	0.07	0.996	0.542	0.428	0.473
	0.1	1.000	0.581	0.455	0.484

Table B.2: Ability of the kernel selection procedure to recover the true causal kernels, using Gaussian kernels over simulated Swiss roll data.

	$\boldsymbol{\theta}$	$S' = 1$	$S' = 3$	$S' = 5$	Adaptive
Recall	0.0	0.112	0.306	0.505	0.488
	0.1	0.125	0.331	0.541	0.715
	0.2	0.157	0.404	0.579	0.963
	0.3	0.196	0.462	0.621	0.999
	0.4	0.239	0.507	0.645	1.000
	0.5	0.275	0.537	0.655	1.000
Precision	0.0	0.337	0.306	0.303	0.328
	0.1	0.377	0.331	0.325	0.318
	0.2	0.471	0.404	0.347	0.303
	0.3	0.588	0.462	0.372	0.300
	0.4	0.717	0.507	0.387	0.300
	0.5	0.825	0.537	0.393	0.300

B.3.6 *A. thaliana* case study of kernelPSI: data description and pre-processing

For this dataset, we are interested in the effect of each gene on the outcome Y, which corresponds to the flowering time in green house, corrected for population structure. We follow the same correction procedure as in Azencott et al. Azencott et al. (2013). The total number of samples is $n = 166$. The features are 9 938 binary SNPs located within a ± 20-kilobase window of 174 pre-selected genes. These genes, known as candidate genes, have been selected by experts as most likely to be involved in flowering time traits. The full list of genes with additional functional information is available from the following URL: https://www.mpipz.mpg.de/14637/Arabidopsis_flowering_genes.

We start with applying hierarchical clustering algorithm to define clusters within each gene. For a given cluster, the associated SNPs are expected to be in linkage disequilibrium. The genes are clustered differently depending on the sample size. Genes with a number of SNPs lower than the gene median size (58 SNPs) are split into 6 clusters. We apply the fixed version of kernelPSI for the three parameterizations $S' \in \{1, 2, 4\}$. For genes larger than the median size, we split them into 12 clusters and consider a number of selected clusters $S' \in \{1, 3, 6\}$.

We use the identical-by-state (IBS) kernel Kwee et al. (2008) for the clusters. This kernel is commonly used in GWAS. For two samples i and j, the IBS kernel corresponds to the fraction of identical SNPs between the two samples:

$$K_{ij} = \frac{|X_i| - ||X_i - X_j||}{|X_i|},$$

where $|X_i|$ is the length of X_i.

B.3.7 *A. thaliana* case study: rank concordance between the methods

Table B.3: Concordance between kernelPSI and benchmark methods, measured by the Kendall's tau coefficient between the p-values returned for the 50% smallest genes.

	KRR (S'= 1)	KPCR (S'= 1)	HSIC (S'= 1)	KRR (S'= 2)	KPCR (S'= 2)	HSIC (S'= 2)	KRR (S'= 4)	KPCR (S'= 4)	HSIC (S'= 4)	KRR adaptive	KPCR adaptive	HSIC adaptive	protoLASSO	protoOLS	KRR	KPCR	HSIC	SKAT
KRR (S'= 1)	1.000	0.795	0.664	0.592	0.501	0.593	0.359	0.204	0.443	0.556	0.412	0.410	0.497	0.198	0.263	0.174	0.421	0.424
KPCR (S'= 1)	0.795	1.000	0.579	0.517	0.566	0.495	0.306	0.258	0.327	0.512	0.486	0.368	0.380	0.230	0.273	0.218	0.362	0.369
HSIC (S'= 1)	0.664	0.579	1.000	0.467	0.361	0.655	0.217	0.071	0.552	0.393	0.251	0.529	0.481	0.060	0.150	0.055	0.575	0.577
KRR (S'= 2)	0.592	0.517	0.467	1.000	0.724	0.553	0.491	0.293	0.453	0.580	0.434	0.367	0.495	0.280	0.363	0.255	0.418	0.424
KPCR (S'= 2)	0.501	0.566	0.361	0.724	1.000	0.396	0.410	0.326	0.273	0.533	0.507	0.284	0.336	0.250	0.325	0.229	0.311	0.315
HSIC (S'= 2)	0.593	0.495	0.655	0.553	0.396	1.000	0.255	0.159	0.664	0.441	0.323	0.561	0.504	0.195	0.214	0.181	0.544	0.549
KRR (S'= 4)	0.359	0.306	0.217	0.491	0.410	0.255	1.000	0.621	0.272	0.486	0.394	0.182	0.324	0.498	0.666	0.494	0.181	0.179
KPCR (S'= 4)	0.204	0.258	0.071	0.293	0.326	0.159	0.621	1.000	0.115	0.346	0.460	0.081	0.155	0.581	0.578	0.620	0.070	0.069
HSIC (S'= 4)	0.443	0.327	0.552	0.453	0.273	0.664	0.272	0.115	1.000	0.360	0.214	0.475	0.471	0.169	0.183	0.124	0.569	0.571
KRR adaptive	0.556	0.512	0.393	0.580	0.533	0.441	0.486	0.346	0.360	1.000	0.685	0.403	0.429	0.330	0.438	0.322	0.307	0.307
KPCR adaptive	0.412	0.486	0.251	0.434	0.507	0.323	0.394	0.460	0.214	0.685	1.000	0.241	0.300	0.415	0.419	0.424	0.169	0.176
HSIC adaptive	0.410	0.368	0.529	0.367	0.284	0.561	0.182	0.081	0.475	0.403	0.241	1.000	0.310	0.132	0.144	0.114	0.381	0.385
protoLASSO	0.497	0.380	0.481	0.495	0.336	0.504	0.324	0.155	0.471	0.429	0.300	0.310	1.000	0.218	0.274	0.189	0.483	0.486
protoOLS	0.198	0.230	0.060	0.280	0.250	0.195	0.498	0.581	0.169	0.330	0.415	0.132	0.218	1.000	0.622	0.856	0.107	0.106
KRR	0.263	0.273	0.150	0.363	0.325	0.214	0.666	0.578	0.183	0.438	0.419	0.144	0.274	0.622	1.000	0.641	0.168	0.164
KPCR	0.174	0.218	0.055	0.255	0.229	0.181	0.494	0.620	0.124	0.322	0.424	0.114	0.189	0.856	0.641	1.000	0.092	0.088
HSIC	0.421	0.362	0.575	0.418	0.311	0.544	0.181	0.070	0.569	0.307	0.169	0.381	0.483	0.107	0.168	0.092	1.000	0.972
SKAT	0.424	0.369	0.577	0.424	0.315	0.549	0.179	0.069	0.571	0.307	0.176	0.385	0.486	0.106	0.164	0.088	0.972	1.000

Table B.4: Concordance between kernelPSI and benchmark methods, measured by the Kendall's tau coefficient between the p-values returned for the 50% largest genes.

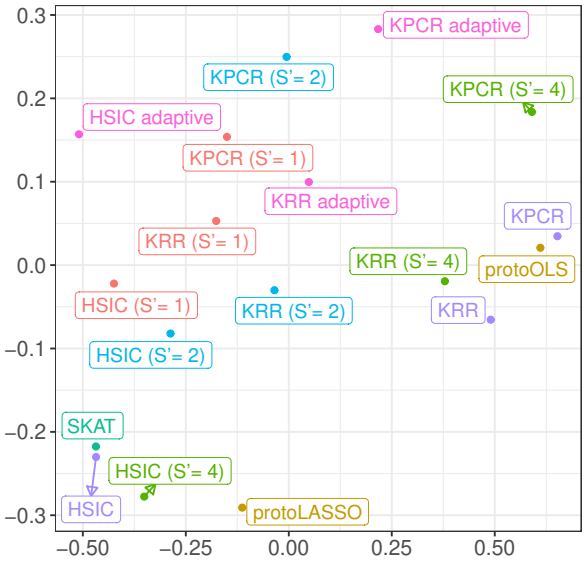
	KRR (S'= 1)	KPCR (S'= 1)	HSIC (S'= 1)	KRR (S'= 3)	KPCR (S'= 3)	HSIC (S'= 3)	KRR (S'= 6)	KPCR (S'= 6)	HSIC (S'= 6)	KRR adaptive	KPCR adaptive	HSIC adaptive	protoLASSO	protoOLS	KRR	KPCR	HSIC	SKAT
KRR (S'= 1)	1.000	0.936	0.778	0.454	0.385	0.450	0.328	0.121	0.385	0.301	0.187	0.332	0.357	-0.032	0.220	0.033	0.522	0.520
KPCR (S'= 1)	0.936	1.000	0.724	0.439	0.396	0.414	0.308	0.129	0.357	0.303	0.189	0.310	0.324	-0.029	0.219	0.034	0.488	0.490
HSIC (S'= 1)	0.778	0.724	1.000	0.421	0.345	0.527	0.366	0.128	0.437	0.299	0.204	0.338	0.397	-0.004	0.202	0.063	0.538	0.536
KRR (S'= 3)	0.454	0.439	0.421	1.000	0.779	0.550	0.495	0.247	0.439	0.403	0.270	0.293	0.380	0.039	0.274	0.105	0.475	0.476
KPCR (S'= 3)	0.385	0.396	0.345	0.779	1.000	0.444	0.449	0.293	0.380	0.381	0.283	0.235	0.282	0.062	0.289	0.149	0.382	0.383
HSIC (S'= 3)	0.450	0.414	0.527	0.550	0.444	1.000	0.327	0.004	0.492	0.273	0.150	0.288	0.354	-0.093	0.133	-0.024	0.516	0.517
KRR (S'= 6)	0.328	0.308	0.366	0.495	0.449	0.327	1.000	0.422	0.411	0.394	0.283	0.270	0.342	0.180	0.442	0.250	0.339	0.336
KPCR (S'= 6)	0.121	0.129	0.128	0.247	0.293	0.004	0.422	1.000	0.127	0.265	0.415	0.069	0.115	0.416	0.350	0.468	0.093	0.094
HSIC (S'= 6)	0.385	0.357	0.437	0.439	0.380	0.492	0.411	0.127	1.000	0.299	0.177	0.350	0.290	0.043	0.205	0.114	0.411	0.405
KRR adaptive	0.301	0.303	0.299	0.403	0.381	0.273	0.394	0.265	0.299	1.000	0.603	0.392	0.345	0.081	0.270	0.156	0.223	0.228
KPCR adaptive	0.187	0.189	0.204	0.270	0.283	0.150	0.283	0.415	0.177	0.603	1.000	0.286	0.248	0.243	0.236	0.327	0.125	0.126
HSIC adaptive	0.332	0.310	0.338	0.293	0.235	0.288	0.270	0.069	0.350	0.392	0.286	1.000	0.250	-0.044	0.163	0.036	0.322	0.321
protoLASSO	0.357	0.324	0.397	0.380	0.282	0.354	0.342	0.115	0.290	0.345	0.248	0.250	1.000	0.084	0.222	0.057	0.345	0.347
protoOLS	-0.032	-0.029	-0.004	0.039	0.062	-0.093	0.180	0.416	0.043	0.081	0.243	-0.044	0.084	1.000	0.332	0.594	-0.037	-0.038
KRR	0.220	0.219	0.202	0.274	0.289	0.133	0.442	0.350	0.205	0.270	0.236	0.163	0.222	0.332	1.000	0.387	0.204	0.205
KPCR	0.033	0.034	0.063	0.105	0.149	-0.024	0.250	0.468	0.114	0.156	0.327	0.036	0.057	0.594	0.387	1.000	-0.001	-0.007
HSIC	0.522	0.488	0.538	0.475	0.382	0.516	0.339	0.093	0.411	0.223	0.125	0.322	0.345	-0.037	0.204	-0.001	1.000	0.984
SKAT	0.520	0.490	0.536	0.476	0.383	0.517	0.336	0.094	0.405	0.228	0.126	0.321	0.347	-0.038	0.205	-0.007	0.984	1.000

B.3.8 *A. thaliana* case study: list of significant genes

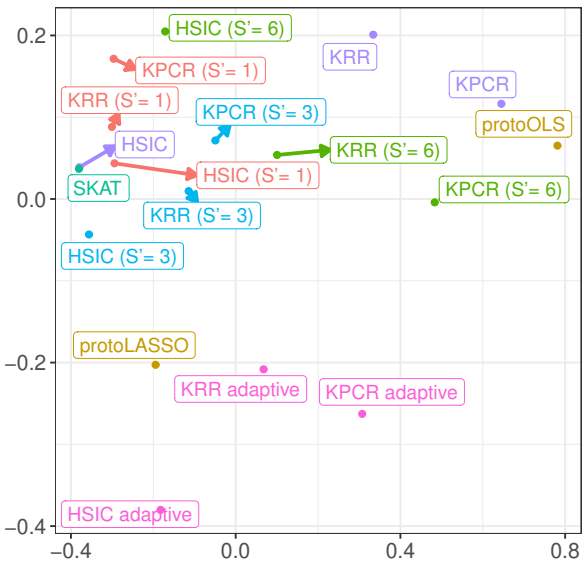
Table B.5: Genes detected as significantly associated to the FT GH phenotype, by method.

Method	Significant genes
HSIC (S'= 4)	AT1G53090, AT1G53160, AT1G56170, AT3G60250
KPCR (S'= 1)	AT5G57360, AT4G00650
protoLASSO	–
KPCR (S'= 3)	AT1G80340, AT4G00650
KPCR	–
HSIC (S'= 3)	–
HSIC (S'= 1)	AT5G57360]
KPCR adaptive	AT4G35900, AT5G47640, AT5G55835
HSIC	–
HSIC adaptive	AT1G53160, AT2G18790, AT4G08920, AT5G46210, AT5G47640, AT5G55835
KPCR (S'= 4)	AT2G22540, AT4G35900, AT5G60100
KRR	–
KPCR (S'= 6)	AT1G69120, AT5G10945
KPCR (S'= 2)	AT1G56170
KRR adaptive	AT2G18790, AT2G25930, AT4G00650, AT5G47640, AT5G55835
SKAT	–
KRR (S'= 6)	AT1G69120, AT2G21070
KRR (S'= 4)	AT1G68840, AT4G35900, AT5G60100, AT5G65050, AT5G65070
KRR (S'= 3)	AT1G80340, AT2G21070
KRR (S'= 2)	AT1G56170, AT2G38880, AT5G65060
HSIC (S'= 6)	AT4G08920, AT5G26147, AT5G47640
KRR (S'= 1)	AT5G57360, AT5G65060
HSIC (S'= 2)	AT1G53090, AT1G56170, AT2G27990
protoOLS	–

B.3.9 *A. thaliana* case study: non-metric multi-dimensional scaling of the results.



(a) NMDS results for the 50% smallest genes



(b) NMDS results for the 50% largest genes

Figure B.7: Non-metric multi-dimensional scaling (NMDS) of the p-values obtained by the kernelPSI and benchmark methods on *Arabidopsis thaliana* data, using $1 - \tau$ as a distance.

Appendix C

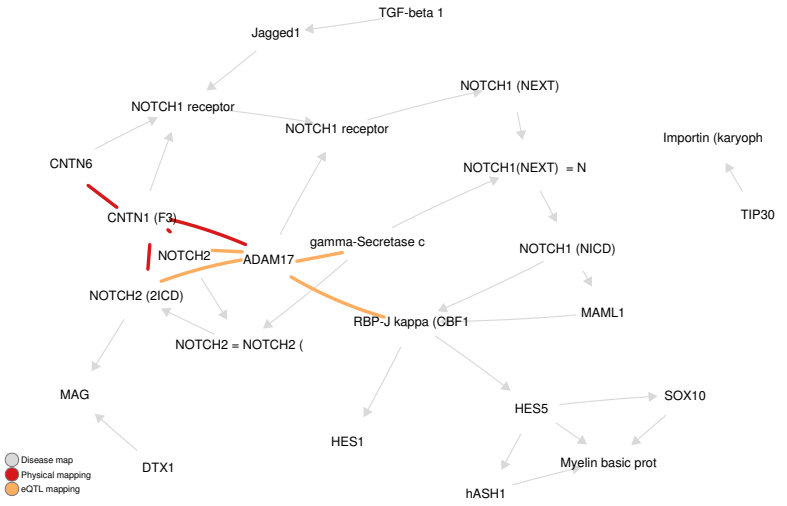
EpiGWAS on multiple sclerosis: supplementary materials

C.1 Distribution of SNPs in MS disease maps

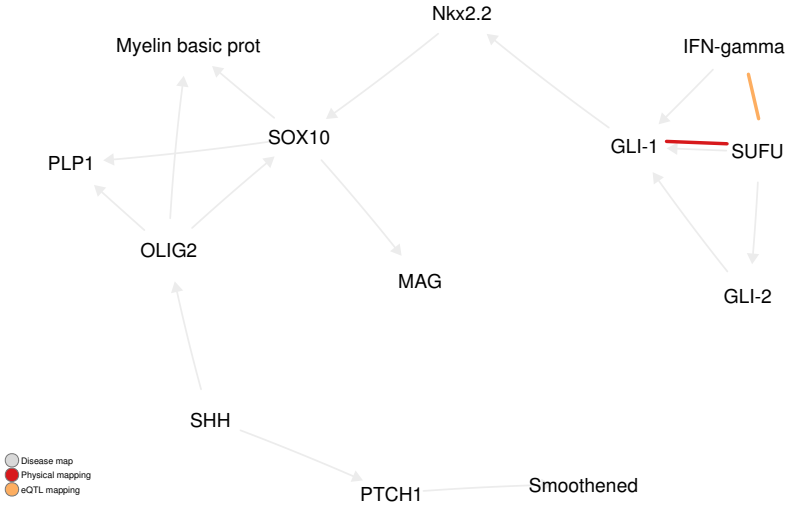
Table C.1: SNP and gene distributions in each disease map for eQTL and physical mappings

internal ID	Physical mapping			eQTL mapping		
	#SNPs	#genes	average #SNPs per gene	#SNPs	#genes	average #SNPs per gene
3302	416	21	19.81	833	19	43.84
3305	70	10	7.00	238	8	29.75
3306	383	21	18.24	869	19	45.74
4455	755	38	19.87	1813	36	50.36
4593	1295	24	53.96	1647	17	96.88
4693	544	34	16.00	912	27	33.78
4703	331	28	11.82	999	27	37.00
4791	252	24	10.50	1264	23	54.96
4794	84	15	5.60	331	12	27.58
4843	984	32	30.75	1401	29	48.31
4846	1318	36	36.61	1555	32	48.59
4901	1173	35	33.51	1209	24	50.38
5199	656	28	23.43	1320	32	41.25
5288	515	27	19.07	724	22	32.91
5378	257	22	11.68	907	22	41.23
5398	141	21	6.71	1050	24	43.75
5518	392	29	13.52	1474	27	54.59
5601	348	28	12.43	742	25	29.68
5611	224	22	10.18	906	24	37.75

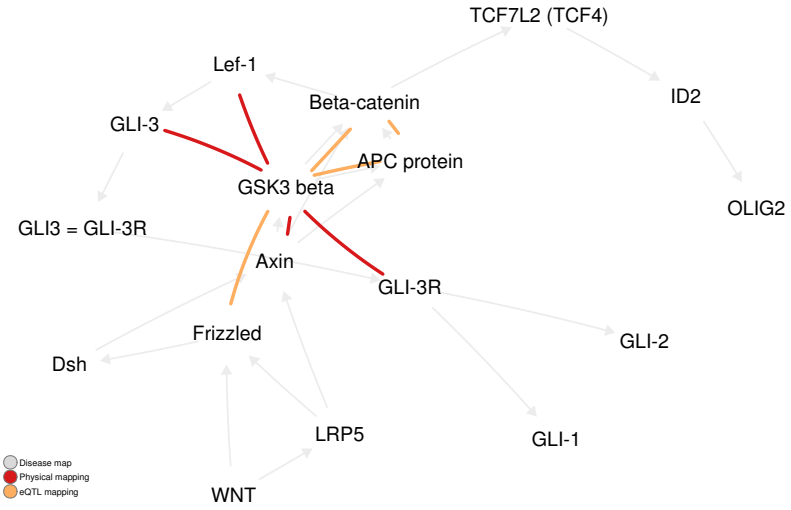
C.2 Visualization of epiGWAS results on MetaCore disease maps for multiple sclerosis



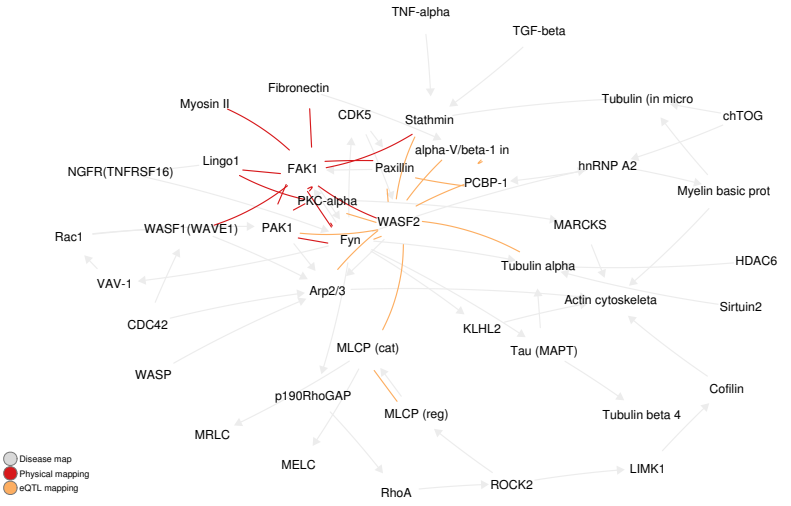
(a) DM 3302: Notch signaling in oligodendrocyte precursor cell differentiation in multiple sclerosis



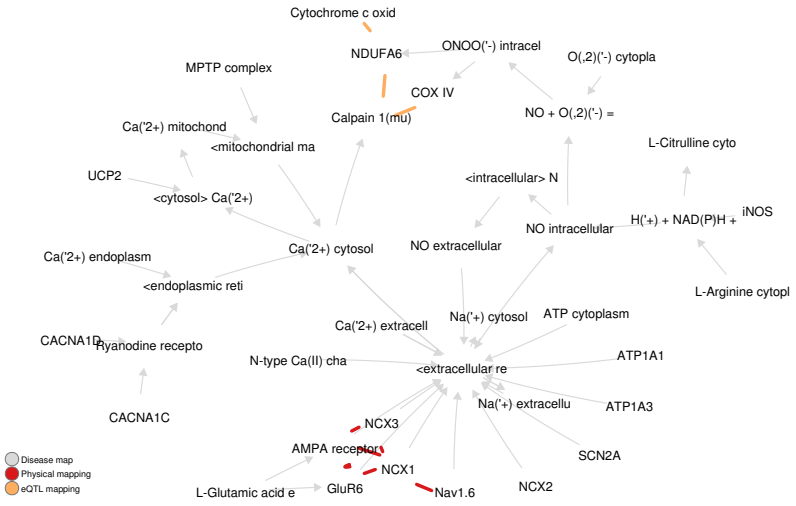
(b) DM 3305: SHH signaling in oligodendrocyte precursor cells differentiation in multiple sclerosis



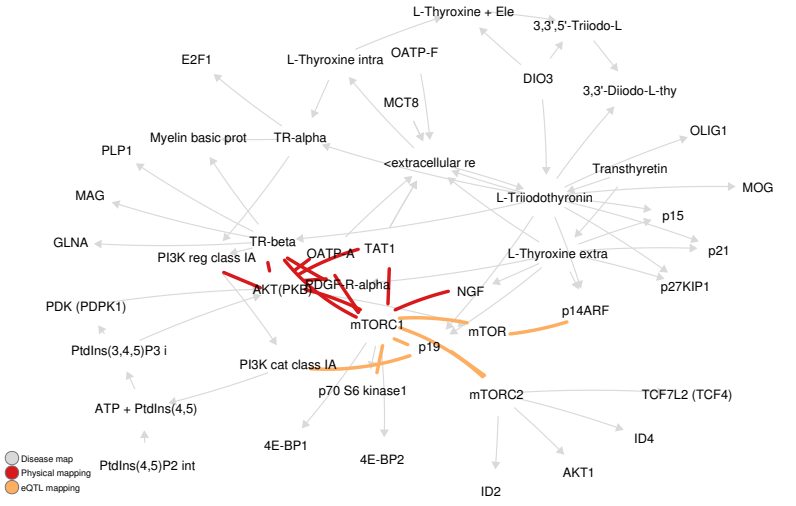
(c) DM 3306: Inhibition of oligodendrocyte precursor cells differentiation by Wnt signaling in multiple sclerosis



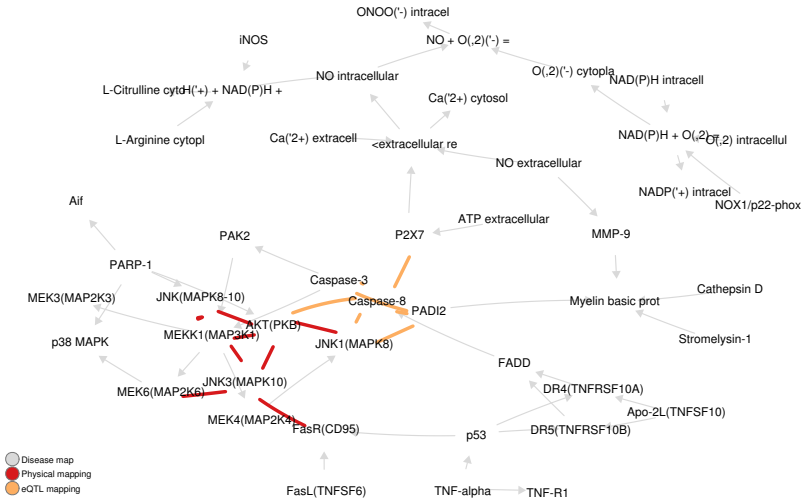
(d) DM 4455: Inhibition of remyelination in multiple sclerosis: regulation of cytoskeleton proteins



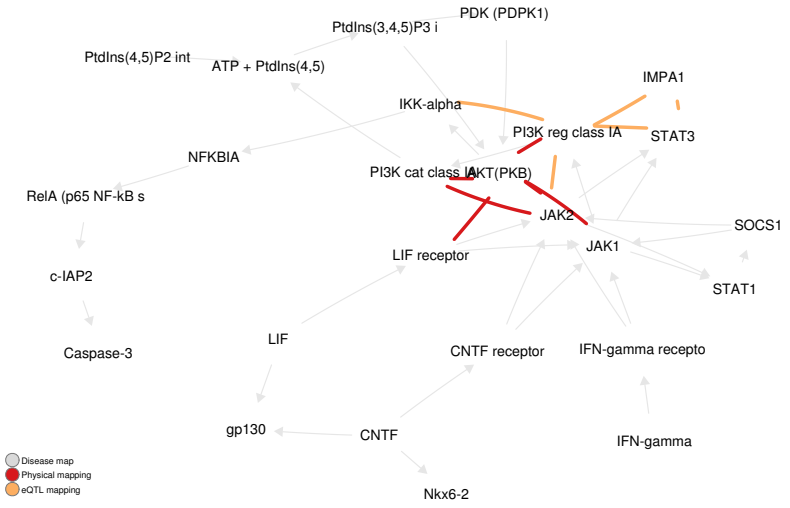
(e) DM 4593: Axonal degeneration in multiple sclerosis



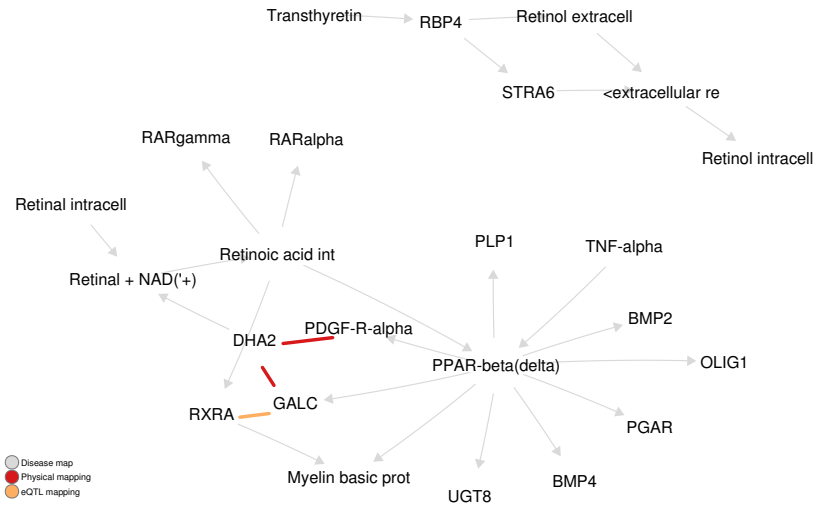
(f) DM 4693: Role of Thyroid hormone in regulation of oligodendrocyte differentiation in multiple sclerosis



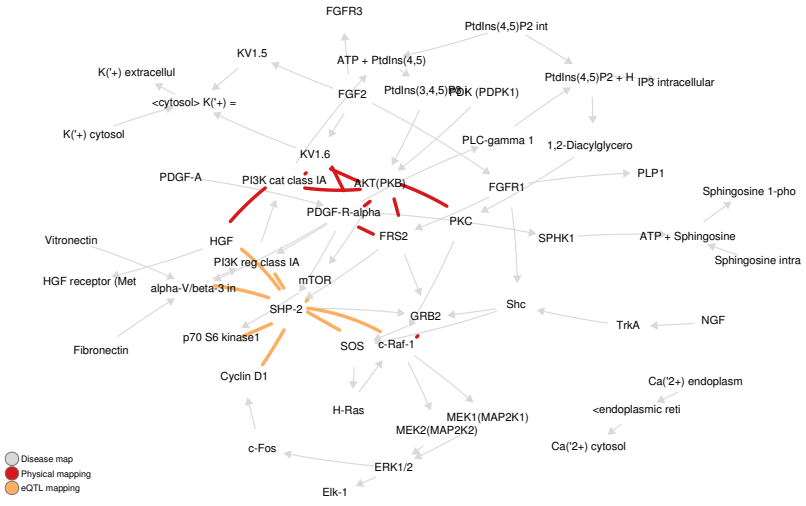
(g) DM 4703: Demyelination in multiple sclerosis



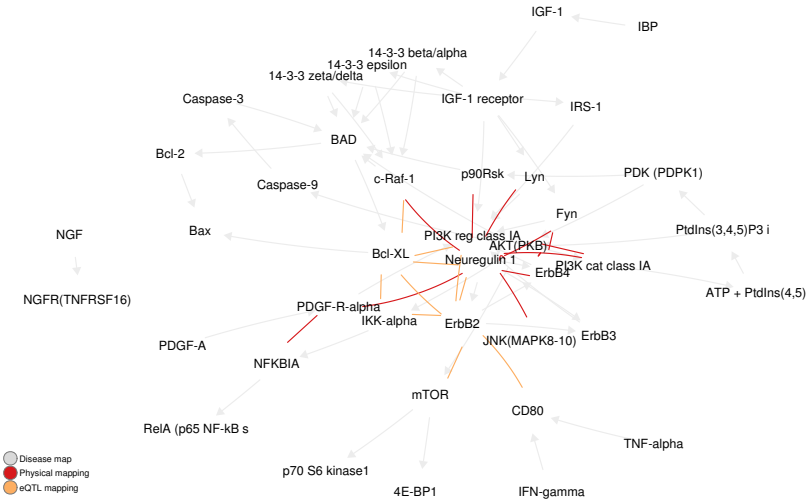
(h) DM 4791: Role of CNTF and LIF in regulation of oligodendrocyte development in multiple sclerosis



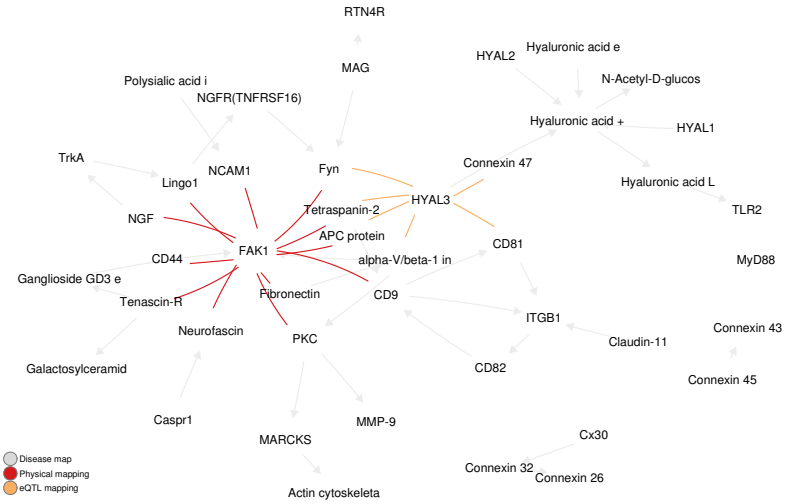
(i) DM 4794: Retinoic acid regulation of oligodendrocyte differentiation in multiple sclerosis



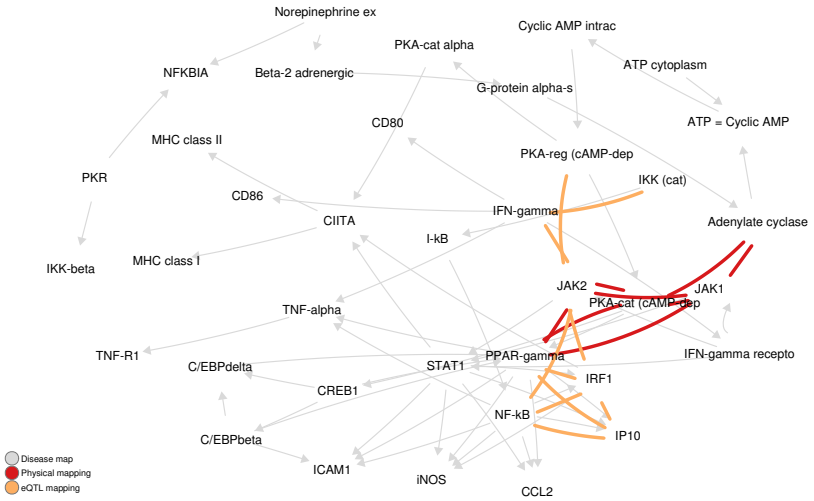
(j) DM 4843: Growth factors in regulation of oligodendrocyte precursor cells proliferation in multiple sclerosis



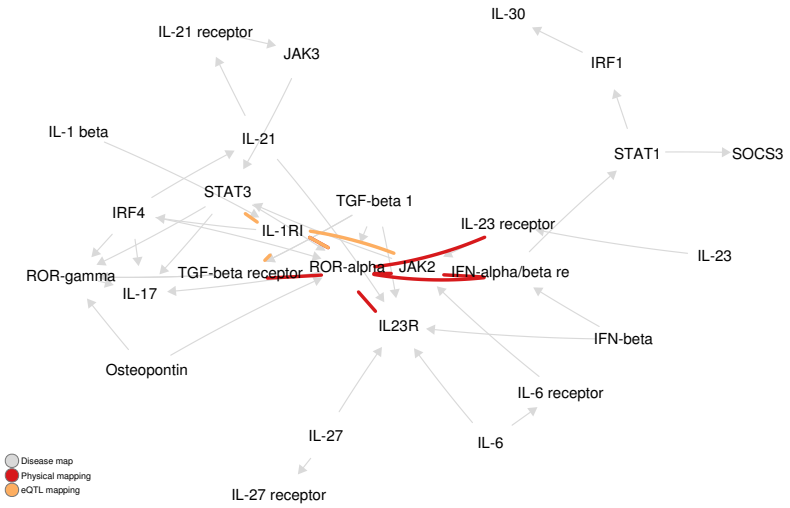
(k) DM 4846: Growth factors in regulation of oligodendrocyte precursor cells survival in multiple sclerosis



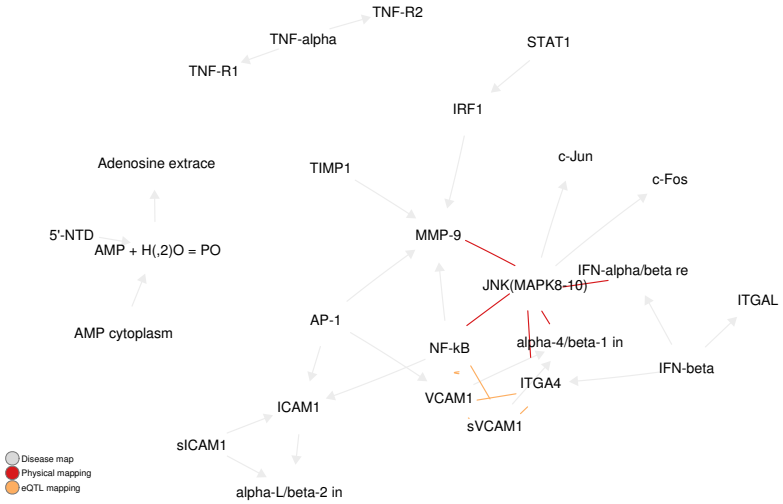
(l) DM 4901: Inhibition of remyelination in multiple sclerosis: role of cell-cell and ECM-cell interactions



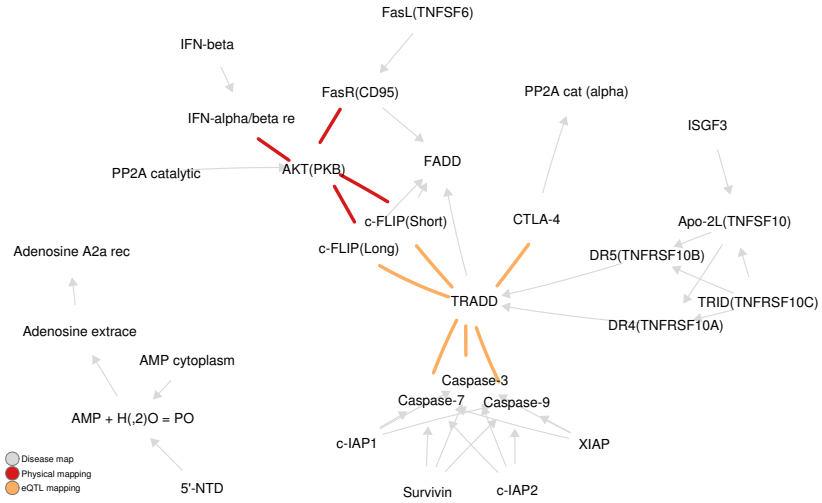
(m) DM 5199: Cooperative action of IFN-γ and TNF-α on astrocytes in multiple sclerosis



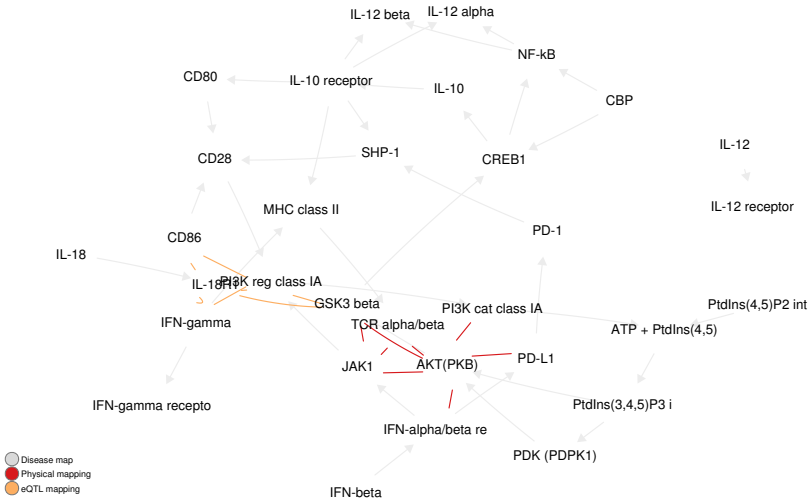
(n) DM 5288: Impaired inhibition of Th17 cell differentiation by IFN-β in multiple sclerosis



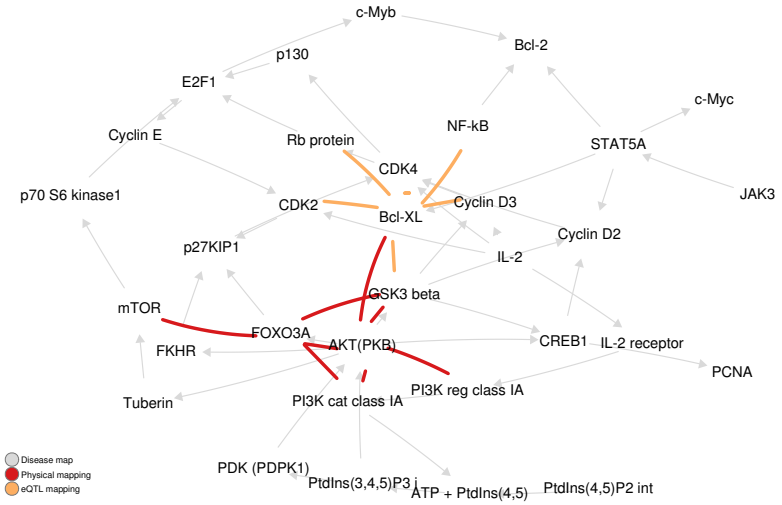
(o) DM 5378: Role of IFN-β in the improvement of blood-brain barrier integrity in multiple sclerosis



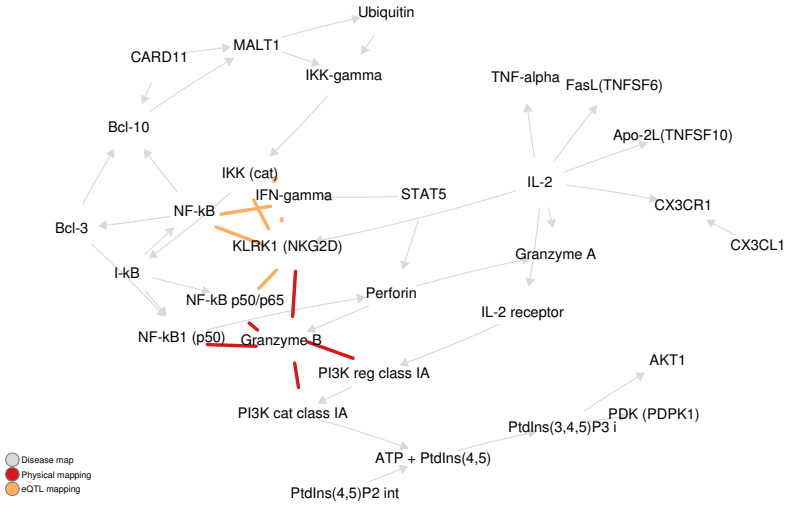
(p) DM 5398: Role of IFN-β in activation of T cell apoptosis in multiple sclerosis



(q) DM 5518: Role of IFN-β in inhibition of Th1 cell differentiation in multiple sclerosis



(r) DM 5601: IL-2 as a growth factor for T cells in multiple sclerosis



(s) DM 5611: Role of IL-2 in the enhancement of NK cell cytotoxicity in multiple sclerosis

Figure C.1: figure
Representation of the 2% top-scoring interactions for physical and eQTL mappings on the original disease maps.

C.3 Statistical significance of the observed network characteristics

Table C.2: Enrichment analysis results for four network characteristics: connectedness, complementarity, centrality and commonality.

internal ID	Connectedness		Complementarity			Centrality	Commonality		
	physical mapping	eQTL mapping	1 vertex	2 vertices	3 vertices	maximal degree	1 edge	2 edges	3 edges
3302	0.019	0.021	0.705	0.254	0.038	0.147	0.542	0.149	0.022
3305	1.000	1.000	0.294	0.013	0.000	1.000	0.354	0.035	0.000
3306	0.061	0.060	0.792	0.335	0.054	0.001	0.761	0.357	0.099
4455	0.000	0.000	0.988	0.905	0.679	0.000	0.822	0.499	0.224
4593	0.001	0.014	0.392	0.056	0.004	0.724	0.368	0.072	0.007
4693	0.000	0.000	0.728	0.314	0.072	0.000	0.619	0.246	0.069
4703	0.000	0.000	0.649	0.208	0.033	0.750	0.479	0.126	0.020
4791	0.008	0.011	0.778	0.340	0.070	0.407	0.728	0.333	0.096
4794	0.161	1.000	0.241	0.011	0.000	1.000	0.233	0.028	0.001
4843	0.000	0.000	0.836	0.477	0.171	0.014	0.551	0.179	0.037
4846	0.000	0.000	0.938	0.699	0.361	0.000	0.782	0.447	0.191
4901	0.000	0.002	0.947	0.729	0.391	0.000	0.561	0.187	0.040
5199	0.000	0.000	0.726	0.291	0.057	0.134	0.785	0.439	0.183
5288	0.004	0.014	0.791	0.366	0.082	0.009	0.697	0.307	0.090
5378	0.012	0.011	0.673	0.215	0.026	0.341	0.561	0.172	0.030
5398	0.012	0.004	0.779	0.346	0.074	0.052	0.567	0.178	0.034
5518	0.002	0.002	0.723	0.283	0.050	0.251	0.665	0.275	0.074
5601	0.001	0.002	0.847	0.472	0.146	0.032	0.713	0.338	0.109
5611	0.004	0.004	0.687	0.245	0.037	0.405	0.602	0.210	0.048

C.4 Content of epiGWAS-selected subnetworks in therapeutic targets

Table C.3: Number of drug targets in the resulting subnetworks for each disease map and its statistical significance.

internal ID	Number of included drug targets	p-value
3302	0	1.000
3305	0	1.000
3306	1	0.378
4455	2	0.380
4593	6	0.009
4693	2	0.382
4703	0	1.000
4791	2	0.154
4794	1	0.222
4843	2	0.808
4846	2	0.500
4901	2	0.265
5199	1	0.875
5288	2	0.728
5378	4	0.024
5398	2	0.347
5518	4	0.275
5601	1	0.768
5611	0	1.000

C.5 MetaCore disease maps for multiple sclerosis

C.5.1 Disease map 3305

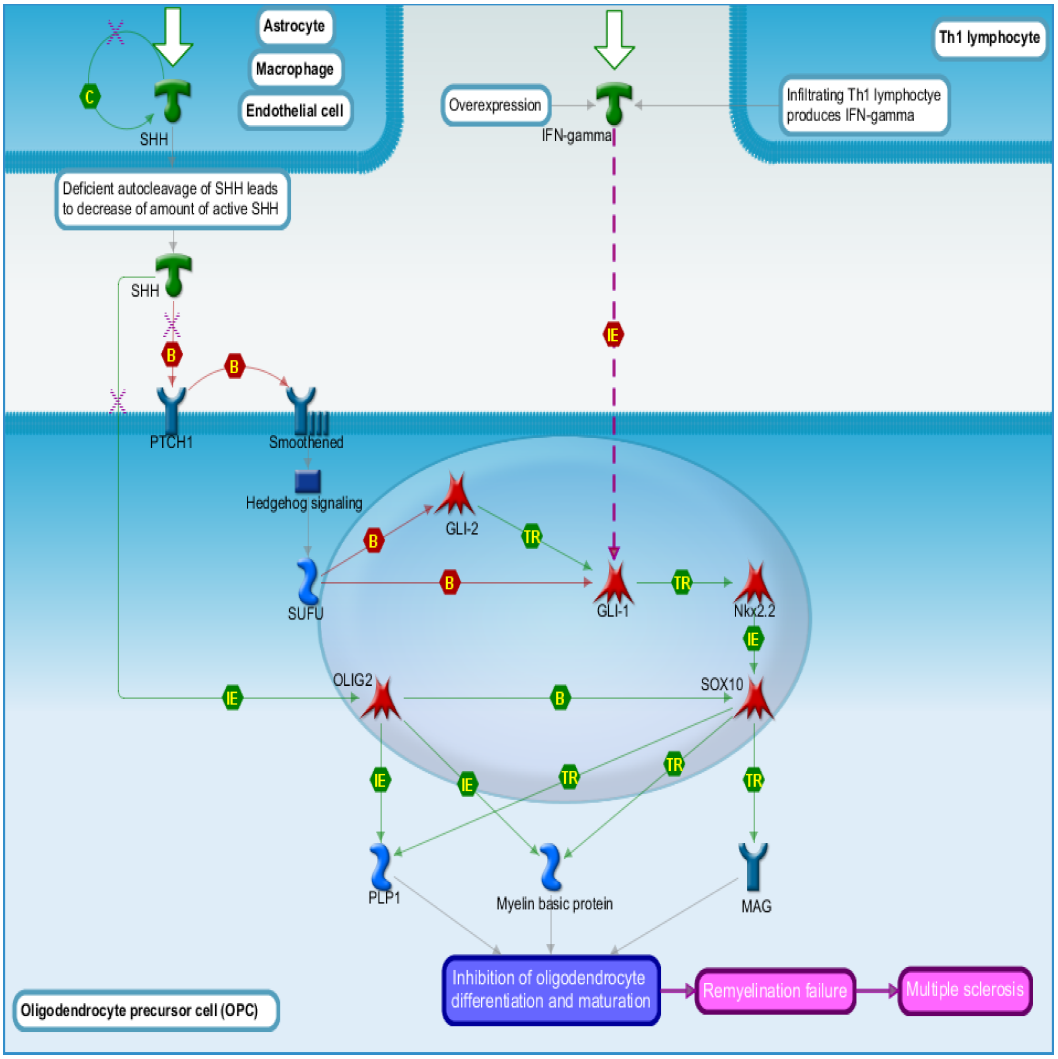


Figure C.2: Sonic Hedgehog signaling in oligodendrocyte precursor cells differentiation in multiple sclerosis (DM 3305).

C.5.2 Disease map 4455

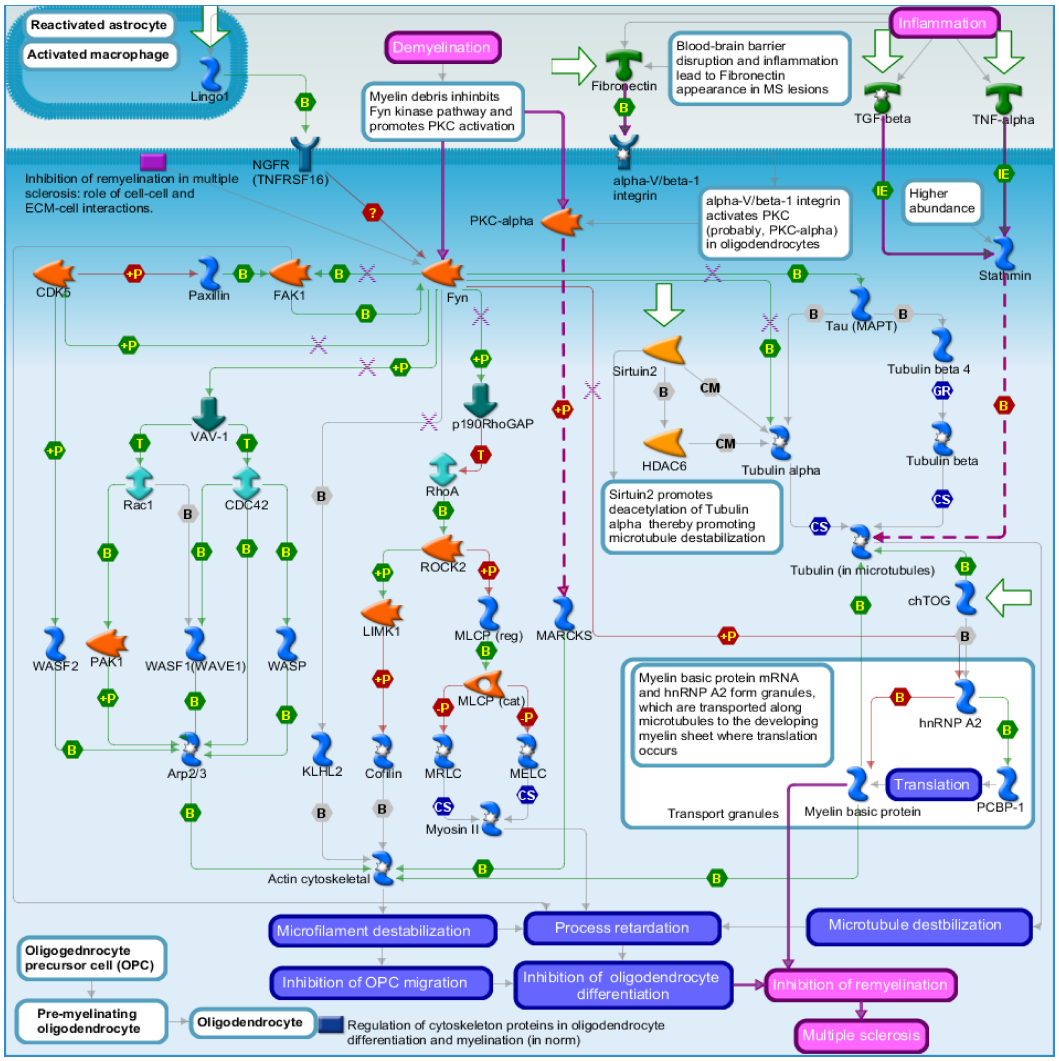


Figure C.3: Inhibition of remyelination in multiple sclerosis: regulation of cytoskeleton proteins (DM 4455).

C.5.3 Disease map 5199

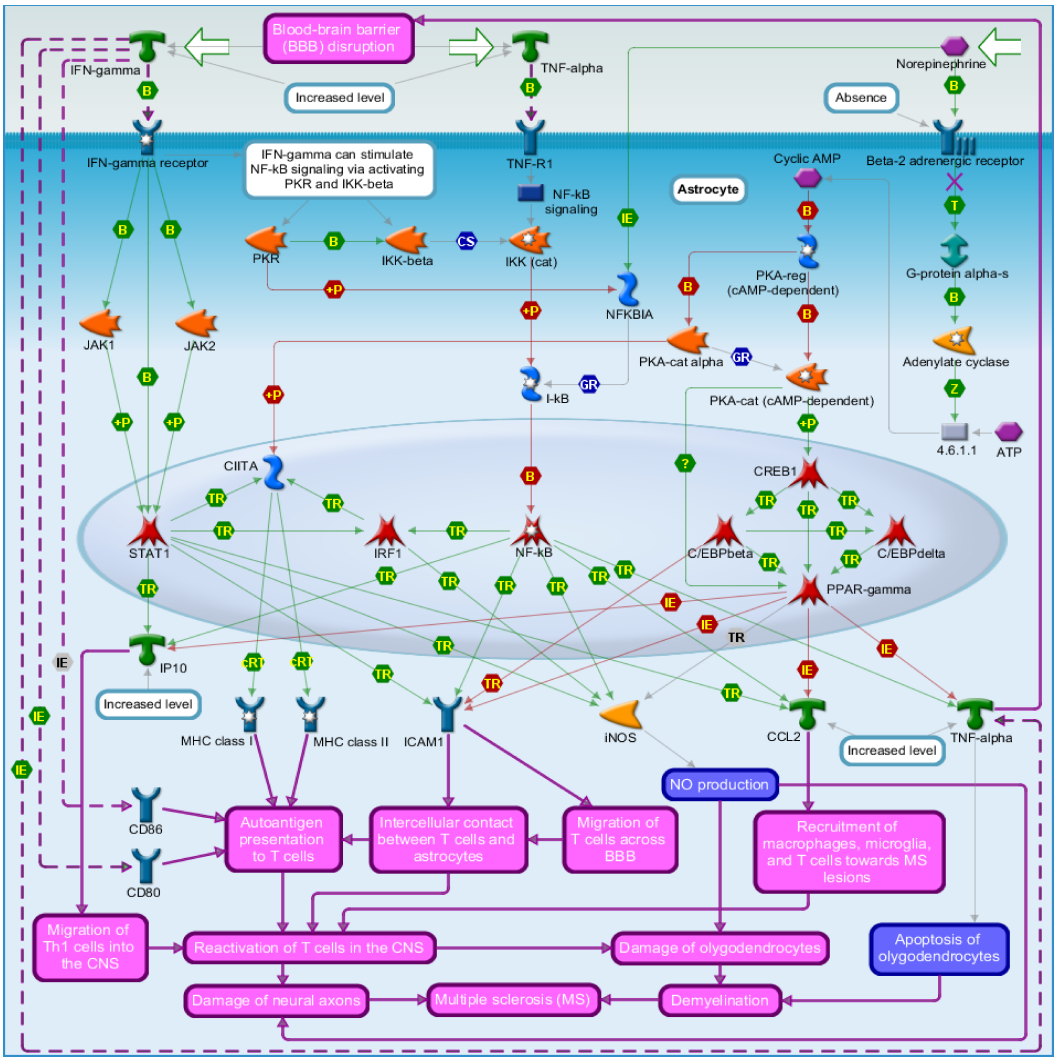


Figure C.4: Cooperative action of IFN-gamma and TNF-alpha on astrocytes in multiple sclerosis (DM 5199).

C.6 Filtering pipeline

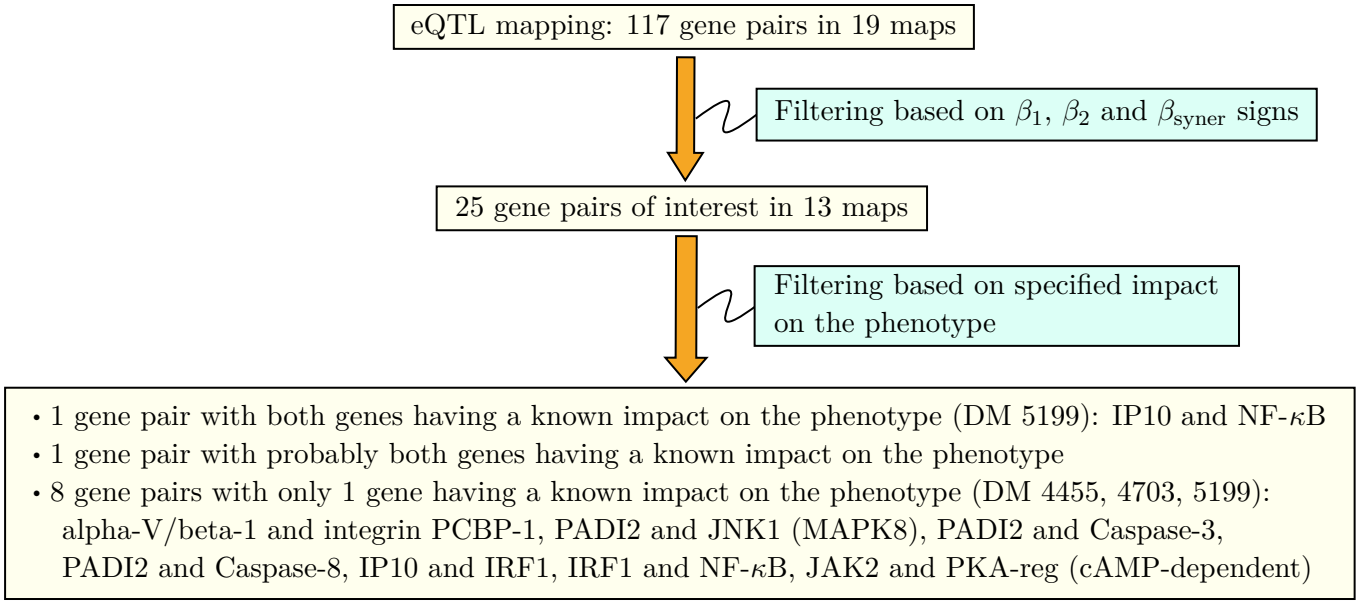


Figure C.5: Filtering process for gene pairs identified by eQTL mapping.

C.7 Physical mapping of SNPs selected by epiGWAS

Table C.4: Pairs of genes identified by physical mapping, and selected on the basis of their SNPs' consequence as a protein dysfunction.

internal ID	Gene pair	Type of interaction
3305	GLI-1 and SUFU	direct interaction between the genes, but unspecified impact on MS
4703	AKT (PKB) and MEKK1 (MAP3K1)	no direct interaction between the genes, but AKT has a specified impact on MS
5611	Granzy me B and KLRK1 (NKG2D)	no direct interaction between the genes, and unspecified impact on MS
	Granzyme B and PI3K cat class IA	no direct interaction between the genes, and unspecified impact on MS

C.8 eQTL mapping

Table C.5: Compiled results of gene pairs identified by epistasis, and filtered according to the scheme in Fig 4.2, with their specified or unknown impact on MS.

internal ID	Title	Interacting gene pair		β_x	β_y	β_{syner}	Specified impact on MS (activation or inhibition)
3302	Notch signaling in oligodendrocyte precursor cell differentiation in multiple sclerosis	RBP-J kappa (CBF1)	ADAM17	1.40	1.37	0.02	no
3305	SHH signaling in oligodendrocyte precursor cells differentiation in multiple sclerosis						
3306	Inhibition of oligodendrocyte precursor cells differentiation by Wnt signaling in multiple sclerosis	Beta-catenin	GSK3 beta	1.27	1.84	0.00	no
4455	Inhibition of remyelination in multiple sclerosis: regulation of cytoskeleton proteins	alpha-V/beta-1 integrin	PCBP-1	1.27	0.96	0.01	probably yes for alpha-V/beta-1 integrin
4593	Axonal degeneration in multiple sclerosis						
4693	Role of Thyroid hormone in regulation of oligodendrocyte differentiation in multiple sclerosis						
4703	Demyelination in multiple sclerosis	PADI2	JNK1(MAPK8)	-1.42	-1.58	-0.02	PADI2 enhances in disease

4703	Demyelination in multiple sclerosis	PADI2	Caspase-3	-1.56	-1.96	-0.01	PADI2 enhances in disease
4703	Demyelination in multiple sclerosis	PADI2	Caspase-8	-1.47	-1.21	-0.03	PADI2 enhances in disease
4703	Demyelination in multiple sclerosis	JNK1(MAPK8)	Caspase-8	-1.58	-1.21	-0.01	no
4791	Role of CNTF and LIF in regulation of oligodendrocyte development in multiple sclerosis	IMPA1	STAT3	1.41	1.10	0.02	no
4791	Role of CNTF and LIF in regulation of oligodendrocyte development in multiple sclerosis	PI3K reg class IA	STAT3	1.40	1.10	0.05	no
4794	Retinoic acid regulation of oligodendrocyte differentiation in multiple sclerosis						
4843	Growth factors in regulation of oligodendrocyte precursor cells proliferation in multiple sclerosis	alpha-V/beta-3 integrin	SHP-2	1.34	1.97	0.07	no
4843	Growth factors in regulation of oligodendrocyte precursor cells proliferation in multiple sclerosis	SHP-2	c-Raf-1	1.63	1.63	0.09	no
4846	Growth factors in regulation of oligodendrocyte precursor cells survival in multiple sclerosis	ErbB2	Neuregulin 1	1.10	1.58	0.11	no
4846	Growth factors in regulation of oligodendrocyte precursor cells survival in multiple sclerosis	Neuregulin 1	Bcl-XL	-1.49	-1.17	-0.02	no

4901	Inhibition of remyelination in multiple sclerosis: role of cell-cell and ECM-cell interactions	Fyn	HYAL3	-1.99	-1.38	-0.07	no
5199	Cooperative action of IFN-γ and TNF-α on astrocytes in multiple sclerosis	IP10	IRF1	1.41	1.12	0.09	yes for IP10
5199	Cooperative action of IFN-γ and TNF-α on astrocytes in multiple sclerosis	IP10	NF-κB	1.39	0.98	0.09	yes for both genes
5199	Cooperative action of IFN-γ and TNF-α on astrocytes in multiple sclerosis	IRF1	NF-κB	1.16	0.88	0.07	yes for NF-κB
5199	Cooperative action of IFN-γ and TNF-α on astrocytes in multiple sclerosis	JAK2	PKA-reg (cAMP-dependent)	1.14	1.25	0.02	yes for JAK2
5288	Impaired inhibition of Th17 cell differentiation by IFN-beta in multiple sclerosis	IL-1RI	ROR-alpha	-1.16	-1.29	-0.09	yes (probable)
5378	Role of IFN-beta in the improvement of blood-brain barrier integrity in multiple sclerosis						
5398	Role of IFN-beta in activation of T cell apoptosis in multiple sclerosis	CTLA-4	TRADD	-1.61	-2.61	-0.04	no
5398	Role of IFN-beta in activation of T cell apoptosis in multiple sclerosis	Caspase-3	TRADD	-1.96	-2.21	-0.07	no
5518	Role of IFN-beta in inhibition of Th1 cell differentiation in multiple sclerosis	IFN-γ	PI3K reg class IA	1.29	1.40	0.07	no
5518	Role of IFN-beta in inhibition of Th1 cell differentiation in multiple sclerosis	GSK3 beta	IL-18R1	1.39	1.36	0.02	no
5518	Role of IFN-beta in inhibition of Th1 cell differentiation in multiple sclerosis	PI3K reg class IA	CD86	-0.96	-1.13	-0.18	no

5601	IL-2 as a growth factor for T cells in multiple sclerosis	GSK3 beta	Bcl-XL	-0.85	-1.17	-0.03	no
5611	Role of IL-2 in the enhancement of NK cell cytotoxicity in multiple sclerosis						

Bibliography

Achlioptas, P., Schölkopf, B. and Borgwardt, K. Two-locus association mapping in subquadratic time. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11*. ACM Press, 2011. doi: 10.1145/2020408.2020521. (Cited on page 25.)

Adzhubei, I., Jordan, D.M. and Sunyaev, S.R. Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics*, 76 (1):7.20.1–7.20.41, January 2013. doi: 10.1002/0471142905.hg0720s76. (Cited on page 72.)

Agbabiaka, T.B., Savović, J. and Ernst, E. Methods for causality assessment of adverse drug reactions. *Drug Safety*, 31(1):21–37, 2008. doi: 10.2165/00002018-200831010-00003. (Cited on page 99.)

Ahrendsen, J.T., Harlow, D.E., Finseth, L.T., Bourne, J.N., Hickey, S.P., Gould, E.A., Culp, C.M. and Macklin, W.B. The protein tyrosine phosphatase shp2 regulates oligodendrocyte differentiation and early myelination and contributes to timely remyelination. *The Journal of Neuroscience*, 38(4):787–802, December 2017. doi: 10.1523/jneurosci.2864-16.2017. (Cited on page 69.)

Antonia, A.L., Gibbs, K.D., Trahair, E.D., Pittman, K.J., Martin, A.T. et al. Pathogen evasion of chemokine response through suppression of CXCL10. *Frontiers in Cellular and Infection Microbiology*, 9, August 2019. doi: 10.3389/fcimb.2019.00280. (Cited on page 74.)

Athey, S., Imbens, G.W. and Wager, S. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, feb 2018. doi: 10.1111/rssb.12268. (Cited on page 41.)

Atwell, S., Huang, Y.S., Vilhjálmsson, B.J., Willems, G., Horton, M. et al. Genome-wide association study of 107 phenotypes in arabidopsis thaliana inbred lines. *Nature*, 465(7298):627–631, mar 2010. doi: 10.1038/nature08800. (Cited on pages 57 and 58.)

Austin, P.C. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46 (3):399–424, May 2011. doi: 10.1080/00273171.2011.568786. (Cited on pages 17 and 27.)

Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Abecasis, G.R. et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015. ISSN 1476-4687. doi: 10.1038/nature15393. (Cited on page 33.)

Azencott, C.A., Grimm, D., Sugiyama, M., Kawahara, Y. and Borgwardt, K.M. Efficient network-guided multi-locus association mapping with graph cuts. *Bioinformatics*, 29(13):i171–i179, 2013. (Cited on page 125.)

Bach, F.R. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9:1179–1225, June 2008. ISSN 1532-4435. (Cited on page 59.)

Bank, C., Hietpas, R.T., Jensen, J.D. and Bolon, D.N. A systematic survey of an intragenic epistatic landscape. *Molecular Biology and Evolution*, 32(1):229–238, November 2014. doi: 10.1093/molbev/msu301. (Cited on page 8.)

Baranzini, S.E. and Oksenberg, J.R. The genetics of multiple sclerosis: From 0 to 200 in 50 years. *Trends in Genetics*, 33(12):960–970, December 2017. doi: 10.1016/j.tig.2017.09.004. (Cited on pages 19 and 63.)

Barber, R.F. and Candès, E.J. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, oct 2015. doi: 10.1214/15-aos1337. (Cited on page 101.)

Bateson, W. and Mendel, G. *Mendel's principles of heredity, by W. Bateson.* University Press,, 1909. doi: 10.5962/bhl.title.1057. (Cited on page 9.)

Beinrucker, A., Dogan, U. and Blanchard, G. A simple extension of stability feature selection. In *Lecture Notes in Computer Science*, pp. 256–265. Springer Berlin Heidelberg, 2012. doi: 10.1007/978-3-642-32717-9_26. (Cited on page 24.)

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 00359246. (Cited on page 23.)

Berbee, H.C.P., Boender, C.G.E., Ran, A.H.G.R., Scheffer, C.L., Smith, R.L. and Telgen, J. Hit-and-run algorithms for the identification of nonredundant linear inequalities. *Mathematical Programming*, 37(2):184–207, jun 1987. (Cited on page 52.)

Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. Valid post-selection inference. *Ann. Stat.*, 41(2):802–837, 2013. (Cited on page 45.)

Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N. and Tawfik, D.S. Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature*, 444(7121):929–932, November 2006. doi: 10.1038/nature05385. (Cited on page 8.)

Bessonov, K., Gusareva, E.S. and Steen, K.V. A cautionary note on the impact of protocol changes for genome-wide association SNP × SNP interaction studies: an example on ankylosing spondylitis. *Human Genetics*, 134(7):761–773, May 2015. doi: 10.1007/s00439-015-1560-7. (Cited on page 40.)

Bien, J., Taylor, J. and Tibshirani, R. A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111–1141, June 2013. doi: 10.1214/13-aos1096. (Cited on page 24.)

Box, G.E.P. and Cox, D.R. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243, July 1964. doi: 10.1111/j.2517-6161.1964.tb00553.x. (Cited on page 82.)

Boyle, E.A., Li, Y.I. and Pritchard, J.K. An expanded view of complex traits: From polygenic to omnigenic. *Cell*, 169(7):1177–1186, June 2017. doi: 10.1016/j.cell.2017.05.038. (Cited on page 16.)

Bunnage, M.E. Getting pharmaceutical r&d back on target. *Nature Chemical Biology*, 7(6):335–339, May 2011. doi: 10.1038/nchembio.581. (Cited on page 99.)

Burton, P.R., Clayton, D.G., Cardon, L.R., Craddock, N., Deloukas, P. 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. (Cited on pages 13, 33, 36 and 96.)

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T. et al. The UK biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, October 2018. doi: 10.1038/s41586-018-0579-z. (Cited on pages 19, 40, 80, 87 and 96.)

Bélisle, C.J.P., Romeijn, H.E. and Smith, R.L. Hit-and-run algorithms for generating multivariate distributions. *Mathematics of Operations Research*, 18(2):255–266, 1993. (Cited on page 52.)

Cabin, R.J. and Mitchell, R.J. To bonferroni or not to bonferroni: when and how are the questions. *Bulletin of the Ecological Society of America*, 81(3):246–248, 2000. (Cited on page 23.)

Calzado, M., Bacher, S. and Schmitz, M.L. NF-κb inhibitors for the treatment of inflammatory diseases and cancer. *Current Medicinal Chemistry*, 14(3):367–376, February 2007. doi: 10.2174/092986707779941113. (Cited on page 76.)

Candille, S.I., Absher, D.M., Beleza, S., Bauchet, M., McEvoy, B. et al. Genome-wide association studies of quantitatively measured skin, hair, and eye pigmentation in four european populations. *PLoS ONE*, 7(10):e48294, October 2012. doi: 10.1371/journal.pone.0048294. (Cited on pages xi and 12.)

Cantor, R.M., Lange, K. and Sinsheimer, J.S. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *The American Journal of Human Genetics*, 86(1):6–22, January 2010. doi: 10.1016/j.ajhg.2009.11.017. (Cited on page 5.)

Carvalho-Silva, D., Pierleoni, A., Pignatelli, M., Ong, C., Fumis, L. et al. Open Targets Platform: new developments and updates two years on. *Nucleic Acids Research*, 47(D1):D1056–D1065, 11 2018a. ISSN 0305-1048. doi: 10.1093/nar/gky1133. (Cited on page 70.)

Carvalho-Silva, D., Pierleoni, A., Pignatelli, M., Ong, C., Fumis, L. et al. Open targets platform: new developments and updates two years on. *Nucleic Acids Research*, 47(D1):D1056–D1065, November 2018b. doi: 10.1093/nar/gky1133. (Cited on pages 76 and 77.)

Clarke, R., Ressom, H.W., Wang, A., Xuan, J., Liu, M.C., Gehan, E.A. and Wang, Y. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Reviews Cancer*, 8(1):37–49, January 2008. doi: 10.1038/nrc2294. (Cited on page 13.)

Combarros, O., Cortina-Borja, M., Smith, A.D. and Lehmann, D.J. Epistasis in sporadic alzheimer's disease. *Neurobiology of Aging*, 30(9):1333–1349, September 2009. doi: 10.1016/j.neurobiolaging.2007.11.027. (Cited on pages 9, 12 and 23.)

Conesa, A. and Beck, S. Making multi-omics data accessible to researchers. *Scientific Data*, 6(1), October 2019. doi: 10.1038/s41597-019-0258-4. (Cited on page 98.)

Cordell, H.J. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20):2463–2468, October 2002. doi: 10.1093/hmg/11.20.2463. (Cited on page 8.)

Cordell, H.J. Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6):392–404, June 2009. doi: 10.1038/nrg2579. (Cited on pages 23 and 30.)

Cordell, H.J. and Todd, J.A. Multifactorial inheritance in type 1 diabetes. *Trends in Genetics*, 11(12):499–504, December 1995. doi: 10.1016/s0168-9525(00)89160-x. (Cited on page 10.)

Cordell, H.J., Todd, J.A., Bennett, S.T., Kawaguchi, Y. and Farrall, M. Two-locus maximum lod score analysis of a multifactorial trait: joint consideration of IDDM2 and IDDM4 with IDDM1 in type 1 diabetes. *American journal of human genetics*, 57(4):920–934, oct 1995. ISSN 0002-9297. (Cited on page 10.)

Cordell, H.J., Todd, J.A., Hill, N.J., Lord, C.J., Lyons, P.A., Peterson, L.B., Wicker, L.S. and Clayton, D.G. Statistical modeling of interlocus interactions in a complex disease: Rejection of the multiplicative model of epistasis in type 1 diabetes. *Genetics*, 158(1):357–367, 2001. ISSN 0016-6731. (Cited on page 10.)

Cornet, A., Bettelli, E., Oukka, M., Cambouris, C., Avellana-Adalid, V., Kosmatopoulos, K. and Liblau, R.S. Role of astrocytes in antigen presentation and

naive t-cell activation. *Journal of Neuroimmunology*, 106(1-2):69–77, July 2000. doi: 10.1016/s0165-5728(99)00215-5. (Cited on page 76.)

Cotsapas, C. and Mitrovic, M. Genome-wide association studies of multiple sclerosis. *Clinical & Translational Immunology*, 7(6):e1018, 2018. doi: 10.1002/cti2.1018. (Cited on page 63.)

Couturier, N., Bucciarelli, F., Nurtdinov, R.N., Debouverie, M., Lebrun-Frenay, C. et al. Tyrosine kinase 2 variant influences t lymphocyte polarization and multiple sclerosis susceptibility. *Brain*, 134(3):693–703, February 2011. doi: 10.1093/brain/awr010. (Cited on page 63.)

Cover, T.M. and Thomas, J.A. *Elements of Information Theory*. John Wiley & Sons, Inc., April 2005. doi: 10.1002/047174882x. (Cited on page 25.)

Cox, D.R. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441–444, 1975. (Cited on page 45.)

Cox, N.J., Frigge, M., Nicolae, D.L., Concannon, P., Hanis, C.L., Bell, G.I. and Kong, A. Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in mexican americans. *Nature Genetics*, 21(2):213–215, February 1999. doi: 10.1038/6002. (Cited on page 10.)

Dargahi, N., Katsara, M., Tselios, T., Androutsou, M.E., de Courten, M., Matsoukas, J. and Apostolopoulos, V. Multiple sclerosis: Immunopathology and treatment update. *Brain Sciences*, 7(12):78, July 2017. (Cited on page 63.)

Davies, N.M., Holmes, M.V. and Smith, G.D. Reading mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ*, pp. k601, July 2018. doi: 10.1136/bmj.k601. (Cited on page 71.)

Davis, J. and Goadrich, M. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning - ICML '06*. ACM Press, 2006. doi: 10.1145/1143844.1143874. (Cited on page 40.)

Daya, M., van der Merwe, L., van Helden, P.D., Möller, M. and Hoal, E.G. Investigating the role of gene-gene interactions in TB susceptibility. *PLOS ONE*, 10(4):e0123970, April 2015. doi: 10.1371/journal.pone.0123970. (Cited on page 9.)

de Leeuw, C.A., Mooij, J.M., Heskes, T. and Posthuma, D. MAGMA: Generalized gene-set analysis of GWAS data. *PLOS Computational Biology*, 11(4):e1004219, April 2015. doi: 10.1371/journal.pcbi.1004219. (Cited on page 92.)

Dean, G., Yeo, T.W., Goris, A., Taylor, C.J., Goodman, R.S. et al. HLA-DRB1 and multiple sclerosis in malta. *Neurology*, 70(2):101–105, December 2007. doi: 10.1212/01.wnl.0000284598.98525.d7. (Cited on page 63.)

Dehman, A., Ambroise, C. and Neuvial, P. Performance of a blockwise approach in variable selection using linkage disequilibrium information. *BMC Bioinformatics*, 16(1), May 2015. doi: 10.1186/s12859-015-0556-6. (Cited on page 83.)

Dempster, A.P., Laird, N.M. and Rubin, D.B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. ISSN 00359246. (Cited on page 102.)

Devlin, B. and Roeder, K. Genomic control for association studies. *Biometrics*, 55 (4):997–1004, December 1999. doi: 10.1111/j.0006-341x.1999.00997.x. (Cited on page 12.)

Dobrushin, R.L. A general formulation of the fundamental theorem of shannon in the theory of information. *Uspekhi Matematicheskikh Nauk*, 14(6):3–104, 1959. (Cited on page 11.)

Durinck, S., Spellman, P.T., Birney, E. and Huber, W. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature Protocols*, 4:1184–1191, 2009. (Cited on page 89.)

Dyment, D.A. Multiple sclerosis in stepsiblings: recurrence risk and ascertainment. *Journal of Neurology, Neurosurgery & Psychiatry*, 77(2):258–259, February 2006. doi: 10.1136/jnnp.2005.063008. (Cited on page 63.)

Ekins, S., Nikolsky, Y., Bugrim, A., Kirillov, E. and Nikolskaya, T. Pathway mapping tools for analysis of high content data. In Taylor, D.L., Haskins, J.R. and Giuliano, K.A. (eds.), *High Content Screening: A Powerful Approach to Systems Cell Biology and Drug Discovery*, pp. 319–350. Humana Press, Totowa, NJ, 2006. ISBN 978-1-59745-217-5. doi: 10.1385/1-59745-217-3:319. (Cited on pages 19, 63 and 65.)

Fish, A.E., Capra, J.A. and Bush, W.S. Are interactions between cis -regulatory variants evidence for biological epistasis or statistical artifacts? *The American Journal of Human Genetics*, 99(4):817–830, October 2016. doi: 10.1016/j.ajhg.2016.07.022. (Cited on page 8.)

Fisher, R.A. XV.—the correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1919. doi: 10.1017/s0080456800012163. (Cited on pages 9, 10 and 14.)

Fong, C., Hazlett, C. and Imai, K. Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics*, 12(1):156–177, March 2018. doi: 10.1214/17-aoas1101. (Cited on page 99.)

Fothergill, E., Guo, J., Howard, L., Kerns, J.C., Knuth, N.D. et al. Persistent metabolic adaptation 6 years after "the biggest loser" competition. *Obesity*, 24 (8):1612–1619, May 2016. doi: 10.1002/oby.21538. (Cited on page 19.)

Franke, A., McGovern, D.P.B., Barrett, J.C., Wang, K., Radford-Smith, G.L. et al. Genome-wide meta-analysis increases to 71 the number of confirmed crohn's disease susceptibility loci. *Nature Genetics*, 42(12):1118–1125, November 2010. doi: 10.1038/ng.717. (Cited on page 7.)

Fraser, A., Macdonald-Wallis, C., Tilling, K., Boyd, A., Golding, J. et al. Cohort profile: The avon longitudinal study of parents and children: ALSPAC mothers cohort. *International Journal of Epidemiology*, 42(1):97–110, April 2012. doi: 10.1093/ije/dys066. (Cited on page 87.)

Friedman, J., Hastie, T. and Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 2010. doi: 10.18637/jss.v033.i01. (Cited on page 32.)

Galarza-Muñoz, G., Briggs, F.B.S., Evsyukova, I., Schott-Lerner, G., Kennedy, E.M. et al. Human Epistatic Interaction Controls IL7R Splicing and Increases Multiple Sclerosis Risk. *Cell*, 169(1):72–84.e13, mar 2017. ISSN 1097-4172. doi: 10.1016/j.cell.2017.03.007. (Cited on pages 19, 23 and 63.)

Gatto, N.M. Further development of the case-only design for assessing gene-environment interaction: evaluation of and adjustment for bias. *International Journal of Epidemiology*, 33(5):1014–1024, May 2004. doi: 10.1093/ije/dyh306. (Cited on page 30.)

Gilly, A., Southam, L., Suveges, D., Kuchenbaecker, K., Moore, R. et al. Very low-depth whole-genome sequencing in complex trait association studies. *Bioinformatics*, 35(15):2555–2561, December 2018. doi: 10.1093/bioinformatics/bty1032. (Cited on page 98.)

Goerg, G.M. Lambert w random variables—a new family of generalized skewed distributions with applications to risk estimation. *The Annals of Applied Statistics*, 5(3):2197–2230, September 2011. doi: 10.1214/11-aoas457. (Cited on page 82.)

Goldenberg, M.M. Multiple sclerosis review. *P & T : a peer-reviewed journal for formulary management*, 37(3):175–184, March 2012. ISSN 1052-1372. (Cited on page 63.)

Gonzalez, C.E. and Ostermeier, M. Pervasive pairwise intragenic epistasis among sequential mutations in TEM-1 β-lactamase. *Journal of Molecular Biology*, 431 (10):1981–1992, May 2019. doi: 10.1016/j.jmb.2019.03.020. (Cited on page 8.)

Gonzalez, S., Gupta, J., Villa, E., Mallawaarachchi, I., Rodriguez, M. et al. Replication of genome-wide association study (GWAS) susceptibility loci in a latino bipolar disorder cohort. *Bipolar Disorders*, 18(6):520–527, September 2016. doi: 10.1111/bdi.12438. (Cited on page 97.)

Gregory, S.G., Schmidt, S., Seth, P., Oksenberg, J.R., Hart, J. et al. Interleukin 7 receptor alpha chain (IL7R) shows allelic and functional association with multiple

sclerosis. *Nature genetics*, 39(9):1083–91, sep 2007. ISSN 1061-4036. doi: 10.1038/ng2103. (Cited on page 63.)

Gretton, A., Bousquet, O., Smola, A. and Schölkopf, B. Measuring statistical dependence with hilbert-schmidt norms. In Jain, S., Simon, H.U. and Tomita, E. (eds.), *Algorithmic Learning Theory*, pp. 63–77, Berlin, Heidelberg, 2005a. Springer Berlin Heidelberg. ISBN 978-3-540-31696-1. (Cited on pages 18 and 80.)

Gretton, A., Smola, A., Bousquet, O., Herbrich, R., Belitski, A. et al. Kernel constrained covariance for dependence measurement. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pp. 1–8, January 2005b. (Cited on page 47.)

Gretton, A., Fukumizu, K., Teo, C.H., Song, L., Schölkopf, B. and Smola, A.J. A Kernel Statistical Test of Independence. In Platt, J.C., Koller, D., Singer, Y. and Roweis, S.T. (eds.), *Advances in Neural Information Processing Systems 20*, pp. 585–592. Curran Associates, Inc., 2008. (Cited on page 55.)

Harty, B.L., Coelho, F., Pease-Raissi, S.E., Mogha, A., Ackerman, S.D. et al. Myelinating Schwann cells ensheath multiple axons in the absence of E3 ligase component Fbxw7. *Nature Communications*, 10(1):2976, 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-10881-y. (Cited on pages 19, 23 and 63.)

Hasin, Y., Seldin, M. and Lusis, A. Multi-omics approaches to disease. *Genome Biology*, 18(1), May 2017. doi: 10.1186/s13059-017-1215-1. (Cited on page 98.)

Haury, A.C., Mordelet, F., Vera-Licona, P. and Vert, J.P. TIGRESS: Trustful inference of gene REgulation using stability selection. *BMC Systems Biology*, 6 (1):145, 2012. doi: 10.1186/1752-0509-6-145. (Cited on page 31.)

Heard, N.A. and Rubin-Delanchy, P. Choosing between methods of combining p-values. *Biometrika*, 105(1):239–246, January 2018. doi: 10.1093/biomet/asx076. (Cited on page 16.)

Heller, R., Chatterjee, N., Krieger, A. and Shi, J. Post-selection inference following aggregate level hypothesis testing in large-scale genomic data. *Journal of the American Statistical Association*, 113(524):1770–1783, June 2018. doi: 10.1080/01621459.2017.1375933. (Cited on page 18.)

Herold, C., Steffens, M., Brockschmidt, F.F., Baur, M.P. and Becker, T. INTERSNP: genome-wide interaction analysis guided by a priori information. *Bioinformatics*, 25(24):3275–3281, October 2009. doi: 10.1093/bioinformatics/btp596. (Cited on page 30.)

Hinrichs, A.S. The UCSC genome browser database: update 2006. *Nucleic Acids Research*, 34(90001):D590–D598, January 2006. doi: 10.1093/nar/gkj144. (Cited on page 89.)

Hofmann, T., Schölkopf, B. and Smola, A.J. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171–1220, June 2008. doi: 10.1214/009053607000000677. (Cited on page 16.)

Holmen m.fl, J. The nord-trøndelag health study 1995-97 (HUNT 2). *Norsk Epidemiologi*, 13(1), October 2011. doi: 10.5324/nje.v13i1.305. (Cited on page 87.)

Hudson, W.H., de Vera, I.M.S., Nwachukwu, J.C., Weikum, E.R., Herbst, A.G. et al. Cryptic glucocorticoid receptor-binding sites pervade genomic NF-κb response elements. *Nature Communications*, 9(1), April 2018. doi: 10.1038/s41467-018-03780-1. (Cited on page 77.)

Hughes, T., Adler, A., Kelly, J.A., Kaufman, K.M., Williams, A.H. et al. Evidence for gene-gene epistatic interactions among susceptibility loci for systemic lupus erythematosus. *Arthritis & Rheumatism*, 64(2):485–492, January 2012. doi: 10.1002/art.33354. (Cited on page 9.)

Ioannidis, J.P.A. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, August 2005. doi: 10.1371/journal.pmed.0020124. (Cited on page 5.)

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J.D. and Lin, X. Sequence kernel association tests for the combined effect of rare and common variants. *The American Journal of Human Genetics*, 92(6):841–853, June 2013. doi: 10.1016/j.ajhg.2013.04.015. (Cited on page 84.)

Ishkin, A. *metabaser: Library of functions to work with Clarivate Analytics' MetaBase*, 2019. R package version 4.4.0. (Cited on page 66.)

Iwai, K. Diverse ubiquitin signaling in NF-κb activation. *Trends in Cell Biology*, 22(7):355–364, July 2012. doi: 10.1016/j.tcb.2012.04.001. (Cited on page 76.)

Jager, P.L.D., Baecher-Allan, C., Maier, L.M., Arthur, A.T., Ottoboni, L. et al. The role of the CD58 locus in multiple sclerosis. *Proceedings of the National Academy of Sciences*, 106(13):5264–5269, February 2009. doi: 10.1073/pnas.0813310106. (Cited on page 63.)

Johnstone, I.M. and Titterington, D.M. Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4237–4253, November 2009. doi: 10.1098/rsta.2009.0159. (Cited on page 13.)

Judy, J. Converging evidence for epistasis between ANK3 and potassium channel gene KCNQ2 in bipolar disorder. *Frontiers in Genetics*, 4, 2013. doi: 10.3389/fgene.2013.00087. (Cited on page 9.)

Keshari, P.K., Harbo, H.F., Myhr, K.M., Aarseth, J.H., Bos, S.D. and Berge, T. Allelic imbalance of multiple sclerosis susceptibility genes IKZF3 and IQGAP1 in human peripheral blood. *BMC Genetics*, 17(1), April 2016. doi: 10.1186/s12863-016-0367-4. (Cited on page 66.)

Keyser, J.D., Zeinstra, E. and Wilczak, N. Astrocytic β2-adrenergic receptors and multiple sclerosis. *Neurobiology of Disease*, 15(2):331–339, March 2004. doi: 10.1016/j.nbd.2003.10.012. (Cited on page 74.)

Keyser, J.D., Laureys, G., Demol, F., Wilczak, N., Mostert, J. and Clinckers, R. Astrocytes as potential targets to suppress inflammatory demyelinating lesions in multiple sclerosis. *Neurochemistry International*, 57(4):446–450, November 2010. doi: 10.1016/j.neuint.2010.02.012. (Cited on page 74.)

Kim, D., Li, R., Lucas, A., Verma, S.S., Dudek, S.M. and Ritchie, M.D. Using knowledge-driven genomic interactions for multi-omics data analysis: metadimensional models for predicting clinical outcomes in ovarian carcinoma. *Journal of the American Medical Informatics Association*, pp. ocw165, December 2016. doi: 10.1093/jamia/ocw165. (Cited on page 10.)

Kimmel, G. and Shamir, R. A block-free hidden markov model for genotypes and its application to disease association. *Journal of Computational Biology*, 12(10): 1243–1260, dec 2005. doi: 10.1089/cmb.2005.12.1243. (Cited on page 101.)

King, G. *Unifying Political Methodology*. University of Michigan Press, 1998. doi: 10.3998/mpub.23784. (Cited on page 10.)

Kinsella, R.J., Kahari, A., Haider, S., Zamora, J., Proctor, G. et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database*, 2011(0): bar030–bar030, July 2011. doi: 10.1093/database/bar030. (Cited on page 72.)

Kitsios, G.D. and Zintzaras, E. Genome-wide association studies: hypothesis-"free" or "engaged"? *Translational Research*, 154(4):161–164, October 2009. doi: 10.1016/j.trsl.2009.07.001. (Cited on page 5.)

Komura, D., Shen, F., Ishikawa, S., Fitch, K.R., Chen, W. et al. Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. *Genome Research*, 16(12):1575–1584, October 2006. doi: 10.1101/gr.5629106. (Cited on pages xi and 5.)

Kraft, P., Zeggini, E. and Ioannidis, J.P.A. Replication in genome-wide association studies. *Statistical Science*, 24(4):561–573, November 2009. doi: 10.1214/09-sts290. (Cited on page 5.)

Krüger, J. and Westermann, R. Linear algebra operators for GPU implementation of numerical algorithms. *ACM Transactions on Graphics*, 22(3):908, July 2003. doi: 10.1145/882262.882363. (Cited on page 85.)

Kwee, L.C., Liu, D., Lin, X., Ghosh, D. and Epstein, M.P. A powerful and flexible multilocus association test for quantitative traits. *The American Journal of Human Genetics*, 82(2):386–397, feb 2008. doi: 10.1016/j.ajhg.2007.10.010. (Cited on pages 58 and 125.)

LaFramboise, T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Research*, 37(13):4181–4193, July 2009. doi: 10.1093/nar/gkp552. (Cited on page 3.)

Lagator, M., Igler, C., Moreno, A.B., Guet, C.C. and Bollback, J.P. Epistatic interactions in the ArabinoseCis-regulatory element. *Molecular Biology and Evolution*, 33(3):761–769, November 2015. doi: 10.1093/molbev/msv269. (Cited on page 8.)

Lagator, M., Paixão, T., Barton, N.H., Bollback, J.P. and Guet, C.C. On the mechanistic nature of epistasis in a canonical cis-regulatory element. *eLife*, 6, May 2017. doi: 10.7554/elife.25192. (Cited on page 8.)

Lassmann, H. and Ransohoff, R.M. The CD4–th1 model for multiple sclerosis: a crucial re-appraisal. *Trends in Immunology*, 25(3):132–137, March 2004. doi: 10.1016/j.it.2004.01.007. (Cited on page 75.)

Lawrence, M., Gentleman, R. and Carey, V. rtracklayer: an r package for interfacing with genome browsers. *Bioinformatics*, 25:1841–1842, 2009. doi: 10.1093/bioinformatics/btp328. (Cited on page 89.)

Le Morvan, M. and Vert, J. WHInter: A working set algorithm for high-dimensional sparse second order interaction models. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pp. 3632–3641, 2018. (Cited on page 40.)

Lee, J.D., Sun, D.L., Sun, Y. and Taylor, J.E. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, June 2016. doi: 10.1214/15-aos1371. (Cited on pages 16, 18, 45 and 80.)

Lehner, B. Molecular mechanisms of epistasis within and between genes. *Trends in Genetics*, 27(8):323–331, August 2011. doi: 10.1016/j.tig.2011.05.007. (Cited on pages 8 and 9.)

Liang, Y., Zhou, Y. and Shen, P. NF-kappaB and its regulation on the immune system. *Cellular & molecular immunology*, 1(5):343–50, oct 2004. ISSN 1672-7681. (Cited on page 76.)

Lim, M. and Hastie, T. Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3):627–654, July 2015. doi: 10.1080/10618600.2014.938812. (Cited on page 14.)

Lin, X. Variance component testing in generalised linear models with random effects. *Biometrika*, 84(2):309–326, June 1997. doi: 10.1093/biomet/84.2.309. (Cited on page 92.)

Lincoln, M.R., Ramagopalan, S.V., Chao, M.J., Herrera, B.M., DeLuca, G.C. et al. Epistasis among HLA-DRB1, HLA-DQA1, and HLA-DQB1 loci determines multiple sclerosis susceptibility. *Proceedings of the National Academy of Sciences*, 106

(18):7542–7547, April 2009. doi: 10.1073/pnas.0812664106. (Cited on pages 19 and 63.)

Llinares-López, F., Papaxanthos, L., Roqueiro, D., Bodenham, D. and Borgwardt, K. CASMAP: detection of statistically significant combinations of SNPs in association mapping. *Bioinformatics*, 35(15):2680–2682, December 2018. doi: 10.1093/bioinformatics/bty1020. (Cited on page 25.)

Loftus, J.R. and Taylor, J.E. Selective inference in regression models with groups of variables. *arXiv preprint arXiv:1511.01478*, 2015. (Cited on pages 45, 46, 47, 49, 50 and 51.)

Lunceford, J.K. and Davidian, M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23(19):2937–2960, August 2004. doi: 10.1002/sim.1903. (Cited on page 28.)

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P. et al. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). *Nucleic Acids Research*, 45(D1):D896–D901, November 2016. doi: 10.1093/nar/gkw1133. (Cited on pages 6, 20, 89 and 96.)

Mackay, T.F.C. Epistasis and quantitative traits: using model organisms to study gene–gene interactions. *Nature Reviews Genetics*, 15(1):22–33, December 2013. doi: 10.1038/nrg3627. (Cited on page 97.)

Majumder, S., Zhou, L., Chaturvedi, P., Babcock, G., Aras, S. and Ransohoff, R. Regulation of human IP-10 gene expression in astrocytoma cells by inflammatory cytokines. *Journal of Neuroscience Research*, 54(2):169–180, October 1998. doi: 10.1002/(sici)1097-4547(19981015)54:2<169::aid-jnr5>3.0.co;2-c. (Cited on page 76.)

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A. et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, October 2009. doi: 10.1038/nature08494. (Cited on page 7.)

Marshall, C.R., , Howrigan, D.P., Merico, D., Thiruvahindrapuram, B. et al. Contribution of copy number variants to schizophrenia from a genome-wide study of 41, 321 subjects. *Nature Genetics*, 49(1):27–35, November 2016. doi: 10.1038/ng.3725. (Cited on page 3.)

Masel, J. Genetic drift. *Current Biology*, 21(20):R837–R838, October 2011. doi: 10.1016/j.cub.2011.08.007. (Cited on page 12.)

Massias, M., Gramfort, A. and Salmon, J. Celer: a Fast Solver for the Lasso with Dual Extrapolation. In *ICML 2018 - 35th International Conference on Machine Learning*, volume 80 of *PMLR*, pp. 3321–3330, Stockholm, Sweden, July 2018. (Cited on page 40.)

McGovern, D.P., Rotter, J.I., Mei, L., Haritunians, T., Landers, C. et al. Genetic epistasis of IL23/IL17 pathway genes in crohn's disease. *Inflammatory Bowel Diseases*, 15(6):883–889, June 2009. doi: 10.1002/ibd.20855. (Cited on page 9.)

Medina-Gomez, C., Felix, J.F., Estrada, K., Peters, M.J., Herrera, L. et al. Challenges in conducting genome-wide association studies in highly admixed multiethnic populations: the generation r study. *European Journal of Epidemiology*, 30(4):317–330, March 2015. doi: 10.1007/s10654-015-9998-4. (Cited on page 97.)

Meinshausen, N. and Bühlmann, P. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, July 2010. doi: 10.1111/j.1467-9868.2010.00740.x. (Cited on pages 24 and 31.)

Mieth, B., Kloft, M., Rodríguez, J.A., Sonnenburg, S., Vobruba, R. et al. Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies. *Scientific Reports*, 6(1), November 2016. doi: 10.1038/srep36671. (Cited on page 80.)

Minagar, A. and Alexander, J.S. Blood-brain barrier disruption in multiple sclerosis. *Multiple Sclerosis Journal*, 9(6):540–549, December 2003. doi: 10.1191/1352458503ms965oa. (Cited on page 75.)

Mokhtari, R.B., Homayouni, T.S., Baluch, N., Morgatskaya, E., Kumar, S., Das, B. and Yeger, H. Combination therapy in combating cancer. *Oncotarget*, 8(23), March 2017. doi: 10.18632/oncotarget.16723. (Cited on page 99.)

Moore, J.H. and Williams, S.M. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *BioEssays*, 27(6):637–646, 2005. doi: 10.1002/bies.20236. (Cited on page 69.)

Morgan, S., Grootendorst, P., Lexchin, J., Cunningham, C. and Greyson, D. The cost of drug development: A systematic review. *Health Policy*, 100(1):4–17, April 2011. doi: 10.1016/j.healthpol.2010.12.002. (Cited on page 98.)

Nakagawa, S. A farewell to bonferroni: the problems of low statistical power and publication bias. *Behavioral Ecology*, 15(6):1044–1045, nov 2004. doi: 10.1093/beheco/arh107. (Cited on page 23.)

Nakka, P., Raphael, B.J. and Ramachandran, S. Gene and network analysis of common variants reveals novel associations in multiple complex diseases. *Genetics*, 204(2):783–798, August 2016. doi: 10.1534/genetics.116.188391. (Cited on page 89.)

Nelson, M.R. Large-scale validation of single nucleotide polymorphisms in gene regions. *Genome Research*, 14(8):1664–1668, August 2004. doi: 10.1101/gr.2421604. (Cited on page 3.)

Nelson, M.R., Tipney, H., Painter, J.L., Shen, J., Nicoletti, P. et al. The support of human genetic evidence for approved drug indications. *Nature Genetics*, 47(8): 856–860, June 2015. doi: 10.1038/ng.3314. (Cited on pages 3, 4 and 98.)

Ng, P.C. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814, July 2003. doi: 10.1093/nar/gkg509. (Cited on page 72.)

Niel, C., Sinoquet, C., Dina, C. and Rocheleau, G. A survey about methods dedicated to epistasis detection. *Frontiers in Genetics*, 6(SEP), 2015. ISSN 16648021. doi: 10.3389/fgene.2015.00285. (Cited on page 23.)

Nolte, I.M., van der Most, P.J., Alizadeh, B.Z., de Bakker, P.I., Boezen, H.M. et al. Missing heritability: is the gap closing? an analysis of 32 complex traits in the lifelines cohort study. *European Journal of Human Genetics*, 25(7):877–885, April 2017. doi: 10.1038/ejhg.2017.50. (Cited on page 7.)

Oksenberg, J.R. Decoding multiple sclerosis: an update on genomics and future directions. *Expert Review of Neurotherapeutics*, 13(sup2):11–19, November 2013. doi: 10.1586/14737175.2013.865867. (Cited on page 4.)

Örjan Carlborg and Haley, C.S. Epistasis: too often neglected in complex trait studies? *Nature Reviews Genetics*, 5(8):618–625, August 2004. doi: 10.1038/nrg1407. (Cited on page 8.)

Pagès, H. *SNPlocs.Hsapiens.dbSNP144.GRCh37: SNP locations for Homo sapiens (dbSNP Build 144)*, 2017. R package version 0.99.20. (Cited on page 66.)

Pakman, A. and Paninski, L. Exact hamiltonian monte carlo for truncated multivariate gaussians. *Journal of Computational and Graphical Statistics*, 23(2): 518–542, apr 2014. doi: 10.1080/10618600.2013.788448. (Cited on page 52.)

Pearl, J. Causal inference in statistics: An overview. *Statistics Surveys*, 3(0):96–146, 2009. doi: 10.1214/09-ss057. (Cited on page 16.)

Peters, J., Janzing, D. and Schlkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN 0262037319, 9780262037310. (Cited on page 17.)

Peterson, R.A. and Cavanaugh, J.E. Ordered quantile normalization: a semiparametric transformation built for the cross-validation era. *Journal of Applied Statistics*, pp. 1–16, 2019. doi: 10.1080/02664763.2019.1630372. (Cited on page 82.)

Phillips, P.C. Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9(11):855–867, November 2008. doi: 10.1038/nrg2452. (Cited on page 8.)

Piegorsch, W.W., Weinberg, C.R. and Taylor, J.A. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in Medicine*, 13(2):153–162, January 1994. doi: 10.1002/sim.4780130206. (Cited on page 30.)

Plowright, R.K., Sokolow, S.H., Gorman, M.E., Daszak, P. and Foley, J.E. Causal inference in disease ecology: investigating ecological drivers of disease emergence. *Frontiers in Ecology and the Environment*, 6(8):420–429, October 2008. doi: 10.1890/070086. (Cited on page 99.)

Ponath, G., Park, C. and Pitt, D. The role of astrocytes in multiple sclerosis. *Frontiers in Immunology*, 9, February 2018. doi: 10.3389/fimmu.2018.00217. (Cited on page 76.)

Poon, A. and Chao, L. The rate of compensatory mutation in the DNA bacteriophage φx174. *Genetics*, 170(3):989–999, May 2005. doi: 10.1534/genetics.104.039438. (Cited on page 8.)

Prabhu, S. and Pe'er, I. Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease. *Genome Research*, 22(11):2230–2240, July 2012. doi: 10.1101/gr.137885.112. (Cited on page 25.)

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, July 2006. doi: 10.1038/ng1847. (Cited on page 12.)

Pritchard, J.K. The allelic architecture of human disease genes: common disease-common variant... or not? *Human Molecular Genetics*, 11(20):2417–2423, October 2002. doi: 10.1093/hmg/11.20.2417. (Cited on page 3.)

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A. et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, September 2007. doi: 10.1086/519795. (Cited on pages 15 and 30.)

Qi, L. Gene–diet interaction and weight loss. *Current Opinion in Lipidology*, 25(1):27–34, February 2014. doi: 10.1097/mol.0000000000000037. (Cited on page 87.)

Qian, C., An, H., Yu, Y., Liu, S. and Cao, X. TLR agonists induce regulatory dendritic cells to recruit th1 cells via preferential IP-10 secretion and inhibit th1 proliferation. *Blood*, 109(8):3308–3315, December 2006. doi: 10.1182/blood-2006-08-040337. (Cited on page 74.)

Rabiner, L.R. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, Feb 1989. ISSN 0018-9219. doi: 10.1109/5.18626. (Cited on page 31.)

Rastas, P., Koivisto, M., Mannila, H. and Ukkonen, E. A hidden markov technique for haplotype reconstruction. In *Lecture Notes in Computer Science*, pp. 140–151. Springer Berlin Heidelberg, 2005. doi: 10.1007/11557067_12. (Cited on page 101.)

Raynor, P. and Group, B.i.B.C. Born in Bradford, a cohort study of babies born in Bradford, and their parents: Protocol for the recruitment phase. *BMC Public Health*, 8(1):327, 2008. ISSN 1471-2458. doi: 10.1186/1471-2458-8-327. (Cited on page 87.)

Reid, S., Taylor, J. and Tibshirani, R. A general framework for estimation and inference from clusters of features. *Journal of the American Statistical Association*, 113(521):280–293, September 2017. doi: 10.1080/01621459.2016.1246368. (Cited on pages 18, 45, 46, 47, 51 and 54.)

Reuss, R., Mistarz, M., Mirau, A., Kraus, J., Bödeker, R.H. and Oschmann, P. FADD is upregulated in relapsing remitting multiple sclerosis. *Neuroimmunomodulation*, 21(5):221–225, 2014. doi: 10.1159/000356522. (Cited on page 69.)

Ripke, S., Neale, B.M., Corvin, A., Walters, J.T.R., Farh, K.H. et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421–427, July 2014. doi: 10.1038/nature13595. (Cited on page 4.)

Risch, N. and Merikangas, K. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517, September 1996. doi: 10.1126/science.273.5281.1516. (Cited on page 3.)

Rivals, I., Personnaz, L., Taing, L. and Potier, M.C. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, 23(4):401–407, December 2006. doi: 10.1093/bioinformatics/btl633. (Cited on page 70.)

Rochlani, Y., Khan, M.H., Banach, M. and Aronow, W.S. Are two drugs better than one? a review of combination therapies for hypertension. *Expert Opinion on Pharmacotherapy*, 18(4):377–386, February 2017. doi: 10.1080/14656566.2017.1288719. (Cited on page 99.)

Romagnani, P., Annunziato, F., Lazzeri, E., Cosmi, L., Beltrame, C. et al. Interferon-inducible protein 10, monokine induced by interferon gamma, and interferon-inducible t-cell alpha chemoattractant are produced by thymic epithelial cells and attract t-cell receptor (TCR) $\alpha\beta$+CD8+ single-positive t cells, TCR$\gamma\delta$+ t cells, and natural killer–type cells in human thymus. *Blood*, 97(3):601–607, February 2001. doi: 10.1182/blood.v97.3.601. (Cited on page 74.)

Rubin, D.B. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331, March 2005. doi: 10.1198/016214504000001880. (Cited on page 17.)

Rupp, K., Tillet, P., Rudolf, F., Weinbub, J., Morhammer, A., Grasser, T., Jüngel, A. and Selberherr, S. ViennaCL—linear algebra library for multi- and many-core architectures. *SIAM Journal on Scientific Computing*, 38(5):S412–S439, January 2016. doi: 10.1137/15m1026419. (Cited on page 85.)

Saito, T. and Rehmsmeier, M. Precrec: fast and accurate precision–recall and ROC curve calculations in r. *Bioinformatics*, 33(1):145–147, sep 2016. doi: 10.1093/bioinformatics/btw570. (Cited on page 35.)

Sandholt, C.H., Allin, K.H., Toft, U., Borglykke, A., Ribel-Madsen, R. et al. The effect of GWAS identified BMI loci on changes in body weight among middle-aged danes during a five-year period. *Obesity*, 22(3):901–908, August 2013. doi: 10.1002/oby.20540. (Cited on pages 80 and 86.)

Sawcer, S., Hellenthal, G., Pirinen, M., Spencer, C.C.A., Patsopoulos, N.A. et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, 476(7359):214–219, 2011. doi: 10.1038/nature10251. (Cited on pages 63 and 65.)

Sawcer, S., Franklin, R.J.M. and Ban, M. Multiple sclerosis genetics. *The Lancet Neurology*, 13(7):700–709, July 2014. doi: 10.1016/s1474-4422(14)70041-9. (Cited on page 63.)

Schaid, D.J., Chen, W. and Larson, N.B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, 19 (8):491–504, May 2018. doi: 10.1038/s41576-018-0016-z. (Cited on pages 4, 13, 83 and 98.)

Scheet, P. and Stephens, M. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, 78(4):629–644, April 2006. doi: 10.1086/502802. (Cited on pages 17, 30, 101 and 102.)

Schmid, H., Boucherot, A., Yasuda, Y., Henger, A., Brunner, B. et al. Modular activation of nuclear factor- b transcriptional programs in human diabetic nephropathy. *Diabetes*, 55(11):2993–3003, October 2006. doi: 10.2337/db06-0477. (Cited on page 74.)

Schölkopf, B., Scholkopf, M., Tsuda, K., zur Förderung der Wissenschaften, M.P.G., Vert, J. et al. *Kernel Methods in Computational Biology*. A Bradford book. Bradford Bks, 2004. ISBN 9780262195096. (Cited on page 18.)

Schork, N.J., Murray, S.S., Frazer, K.A. and Topol, E.J. Common vs. rare allele hypotheses for complex diseases. *Current Opinion in Genetics & Development*, 19(3):212–219, June 2009. doi: 10.1016/j.gde.2009.04.010. (Cited on page 33.)

Sesia, M., Sabatti, C. and Candès, E.J. Gene hunting with hidden markov model knockoffs. *Biometrika*, aug 2018. doi: 10.1093/biomet/asy033. (Cited on page 101.)

Shah, R.L., Li, Q., Zhao, W., Tedja, M.S., Tideman, J.W.L. et al. A genome-wide association study of corneal astigmatism: The CREAM Consortium. *Molecular vision*, 24:127–142, feb 2018. (Cited on page 89.)

Shastry, B.S. SNPs: Impact on gene function and phenotype. In *Methods in Molecular Biology*, pp. 3–22. Humana Press, 2009. doi: 10.1007/978-1-60327-411-1_1. (Cited on page 3.)

Simmonds, M. and Gough, S. The HLA region and autoimmune disease: Associations and mechanisms of action. *Current Genomics*, 8(7):453–465, November 2007. doi: 10.2174/138920207783591690. (Cited on page 9.)

Slatkin, M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–485, June 2008. doi: 10.1038/nrg2361. (Cited on page 24.)

Slim, L., Chatelain, C., Azencott, C.A. and Vert, J.P. Novel methods for epistasis detection in genome-wide association studies. October 2018. doi: 10.1101/442749. (Cited on pages 16, 20 and 21.)

Slim, L., Chatelain, C., Azencott, C.A. and Vert, J.P. kernelPSI: a post-selection inference framework for nonlinear variable selection. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5857–5865, Long Beach, California, USA, 09–15 Jun 2019. PMLR. (Cited on pages 16, 20 and 43.)

Smalley, E. Clinical trials go virtual, big pharma dives in. *Nature Biotechnology*, 36(7):561–562, July 2018. doi: 10.1038/nbt0718-561. (Cited on page 98.)

Smith, R.L. Efficient Monte Carlo Procedures for Generating Points Uniformly Distributed over Bounded Regions. *Operations Research*, 32(6):1296–1308, 1984. (Cited on page 53.)

Song, L., Smola, A., Gretton, A., Borgwardt, K.M. and Bedo, J. Supervised feature selection via dependence estimation. In *Proceedings of the 24th international conference on Machine learning - ICML '07*. ACM Press, 2007. doi: 10.1145/1273496.1273600. (Cited on pages 45, 47 and 81.)

Spencer, C.C.A., Su, Z., Donnelly, P. and Marchini, J. Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. *PLoS Genetics*, 5(5):e1000477, May 2009. doi: 10.1371/journal.pgen.1000477. (Cited on pages 7 and 98.)

Stranger, B.E., Stahl, E.A. and Raj, T. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, 187(2):367–383, November 2010. doi: 10.1534/genetics.110.120907. (Cited on page 97.)

Su, Z., Marchini, J. and Donnelly, P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics*, 27(16):2304–2305, jun 2011. doi: 10.1093/bioinformatics/btr341. (Cited on page 33.)

Sul, J.H., Martin, L.S. and Eskin, E. Population structure in genetic studies: Confounding factors and mixed models. *PLOS Genetics*, 14(12):e1007309, December 2018. doi: 10.1371/journal.pgen.1007309. (Cited on page 97.)

Sun, S., Greenwood, C.M. and Neal, R.M. Haplotype inference using a bayesian hidden markov model. *Genetic Epidemiology*, 31(8):937–948, dec 2007. doi: 10.1002/gepi.20253. (Cited on page 101.)

Sun, X., Wang, X., Chen, T., Li, T., Cao, K. et al. Myelin activates FAK/akt/NF-κb pathways and provokes CR3-dependent inflammatory response in murine system. *PLoS ONE*, 5(2):e9380, February 2010. doi: 10.1371/journal.pone.0009380. (Cited on page 69.)

Székely, G.J., Rizzo, M.L. and Bakirov, N.K. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, dec 2007. doi: 10.1214/009053607000000505. (Cited on page 25.)

Taylor, J. and Tibshirani, R.J. Statistical learning and selective inference. *Proc. Natl. Acad. Sci. U.S.A.*, 112:7629–7634, June 2015. (Cited on page 45.)

Thanei, G.A., Meinshausen, N. and Shah, R.D. The xyz algorithm for fast interaction search in high-dimensional data. *Journal of Machine Learning Research*, 19 (37):1–42, 2018. (Cited on page 23.)

Tian, L., Alizadeh, A.A., Gentles, A.J. and Tibshirani, R. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532, October 2014. doi: 10.1080/01621459.2014.951443. (Cited on pages 24, 26, 27 and 34.)

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, January 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x. (Cited on page 23.)

Tibshirani, R., Walther, G. and Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, May 2001. doi: 10.1111/1467-9868.00293. (Cited on page 83.)

Tibshirani, R.J., Taylor, J., Lockhart, R. and Tibshirani, R. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical*

Association, 111(514):600–620, April 2016. doi: 10.1080/01621459.2015.1108848. (Cited on pages 18 and 45.)

Tokunaga, R., Zhang, W., Naseem, M., Puccini, A., Berger, M.D. et al. CXCL9, CXCL10, CXCL11/CXCR3 axis for immune activation – a target for novel cancer therapy. *Cancer Treatment Reviews*, 63:40–47, February 2018. doi: 10.1016/j.ctrv.2017.11.007. (Cited on page 74.)

Traugott, U. and Lebon, P. Demonstration of α, β, and γ interferon in active chronic multiple sclerosis lesions. *Annals of the New York Academy of Sciences*, 540(1 Advances in N):309–311, November 1988. doi: 10.1111/j.1749-6632.1988.tb27083.x. (Cited on page 74.)

Van der Waerden, B. Order tests for the two-sample problem and their power. In *Indagationes Mathematicae (Proceedings)*, volume 55, pp. 453–458. Elsevier, 1952. (Cited on page 82.)

VanderWeele, T.J. and Hernan, M.A. Causal inference under multiple versions of treatment. *Journal of Causal Inference*, 1(1):1–20, January 2013. doi: 10.1515/jci-2012-0002. (Cited on page 41.)

VanderWeele, T.J. and Knol, M.J. A tutorial on interaction. *Epidemiologic Methods*, 3(1), January 2014. doi: 10.1515/em-2013-0005. (Cited on pages 10, 11 and 71.)

Vishwanathan, S.V.N., Schraudolph, N.N., Kondor, R. and Borgwardt, K.M. Graph kernels. *Journal of Machine Learning Research*, 11(Apr):1201–1242, 2010. (Cited on page 99.)

Visscher, P.M., Brown, M.A., McCarthy, M.I. and Yang, J. Five years of GWAS discovery. *The American Journal of Human Genetics*, 90(1):7–24, January 2012. doi: 10.1016/j.ajhg.2011.11.029. (Cited on page 3.)

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A. and Yang, J. 10 years of GWAS discovery: Biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, July 2017. doi: 10.1016/j.ajhg.2017.06.005. (Cited on pages 4, 13, 97 and 98.)

Võsa, U., Claringbould, A., Westra, H.J., Bonder, M.J., Deelen, P. et al. Unraveling the polygenic architecture of complex traits using blood eqtl metaanalysis. *bioRxiv*, 2018. doi: 10.1101/447367. (Cited on pages 66, 68 and 71.)

Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N.L. and Yu, W. BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *The American Journal of Human Genetics*, 87(3):325–340, September 2010. doi: 10.1016/j.ajhg.2010.07.021. (Cited on pages 14, 23 and 34.)

Wang, Q., Chen, R., Cheng, F., Wei, Q., Ji, Y. et al. A bayesian framework that integrates multi-omics data and gene networks predicts risk genes from

schizophrenia GWAS data. *Nature Neuroscience*, 22(5):691–699, April 2019. doi: 10.1038/s41593-019-0382-7. (Cited on page 98.)

Wasserstein, R.L. and Lazar, N.A. The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133, April 2016. doi: 10.1080/00031305.2016.1154108. (Cited on page 16.)

Wei, W.H., Hemani, G. and Haley, C.S. Detecting epistasis in human complex traits. *Nature Reviews Genetics*, 15(11):722–733, September 2014. doi: 10.1038/nrg3747. (Cited on pages 12 and 13.)

Weinreich, D.M. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science*, 312(5770):111–114, April 2006. doi: 10.1126/science.1123539. (Cited on page 8.)

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P. et al. The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(D1):D1001–D1006, December 2013. doi: 10.1093/nar/gkt1229. (Cited on page 6.)

Whittaker, J. *Graphical models in applied multivariate statistics*. Wiley Publishing, 2009. (Cited on page 11.)

Williams, A., Piaton, G. and Lubetzki, C. Astrocytes-friends or foes in multiple sclerosis? *Glia*, 55(13):1300–1312, July 2007. doi: 10.1002/glia.20546. (Cited on page 76.)

Witt, D. Recent developments in disulfide bond formation. *Synthesis*, 2008(16):2491–2509, August 2008. doi: 10.1055/s-2008-1067188. (Cited on page 8.)

Wu, M.C., Kraft, P., Epstein, M.P., Taylor, D.M., Chanock, S.J., Hunter, D.J. and Lin, X. Powerful SNP-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, 86(6):929–942, jun 2010. doi: 10.1016/j.ajhg.2010.05.002. (Cited on page 84.)

Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M. and Lin, X. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, July 2011. doi: 10.1016/j.ajhg.2011.05.029. (Cited on pages 55 and 92.)

Xia, J.B., Liu, G.H., Chen, Z.Y., Mao, C.Z., Zhou, D.C. et al. Hypoxia/ischemia promotes CXCL10 expression in cardiac microvascular endothelial cells by NFkB activation. *Cytokine*, 81:63–70, May 2016. doi: 10.1016/j.cyto.2016.02.007. (Cited on page 74.)

Xue, A., , Wu, Y., Zhu, Z., Zhang, F. et al. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nature Communications*, 9(1), July 2018. doi: 10.1038/s41467-018-04951-w. (Cited on page 4.)

Xue, Y., Cooper, G., Cai, C., Lu, S., Hu, B., Ma, X. and Lu, X. Tumour-specific causal inference discovers distinct disease mechanisms underlying cancer subtypes. *Scientific Reports*, 9(1), September 2019. doi: 10.1038/s41598-019-48318-7. (Cited on page 99.)

Yamada, M., Umezu, Y., Fukumizu, K. and Takeuchi, I. Post selection inference with kernels. In Storkey, A. and Perez-Cruz, F. (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 152–160, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR. (Cited on pages 45, 47, 48, 50 and 114.)

Yang, F., Barber, R.F., Jain, P. and Lafferty, J. Selective inference for group-sparse linear models. In *Advances in Neural Information Processing Systems*, pp. 2469–2477, 2016a. (Cited on pages 45, 46, 50 and 51.)

Yang, Q., Khoury, M.J., Sun, F. and Flanders, W.D. Case-only design to measure gene-gene interaction. *Epidemiology (Cambridge, Mass.)*, 10(2):167–70, mar 1999. ISSN 1044-3983. doi: 10.1002/sim.4780130206. (Cited on page 30.)

Yang, S., Imbens, G.W., Cui, Z., Faries, D.E. and Kadziola, Z. Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics*, 72(4):1055–1065, March 2016b. doi: 10.1111/biom.12505. (Cited on page 99.)

Yeo, I.K. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959, December 2000. doi: 10.1093/biomet/87.4.954. (Cited on page 82.)

Yung, L.S., Yang, C., Wan, X. and Yu, W. GBOOST: a GPU-based tool for detecting gene–gene interactions in genome–wide case control studies. *Bioinformatics*, 27(9):1309–1310, March 2011. doi: 10.1093/bioinformatics/btr114. (Cited on page 23.)

Zaykin, D.V. and Zhivotovsky, L.A. Ranks of genuine associations in whole-genome scans. *Genetics*, 171(2):813–823, July 2005. doi: 10.1534/genetics.105.044206. (Cited on page 4.)

Zeng, Y. and Breheny, P. The biglasso package: A memory- and computation-efficient solver for lasso model fitting with big data in r. *ArXiv e-prints*, 2017. (Cited on page 37.)

Zhang, Q., Filippi, S., Gretton, A. and Sejdinovic, D. Large-scale kernel methods for independence testing. *Statistics and Computing*, 28(1):113–130, 2018. ISSN 1573-1375. doi: 10.1007/s11222-016-9721-7. (Cited on pages 48 and 59.)

Zhao, Y., Zeng, D., Rush, A.J. and Kosorok, M.R. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical*

Association, 107(499):1106–1118, June 2012. doi: 10.1080/01621459.2012.695674. (Cited on pages 24, 26 and 29.)

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2): 301–320, April 2005. doi: 10.1111/j.1467-9868.2005.00503.x. (Cited on page 31.)

Zuk, O., Hechter, E., Sunyaev, S.R. and Lander, E.S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198, January 2012. doi: 10.1073/pnas.1119675109. (Cited on pages 6, 7 and 97.)

RÉSUMÉ

En offrant une image sans précédent du génome humain, les études d'association pangénomiques (GWAS) expliqueraient pleinement le contexte génétique des maladies complexes. A ce jour, les résultats ont été pour le moins mitigés. Cela peut être partiellement attribué à la méthodologie statistique adoptée, qui ne prend pas souvent en compte l'interaction entre les variants génétiques, ou l'épistasie. La détection d'épistasie à travers des modèles statistiques présente plusieurs défis pour lesquels nous développons dans cette thèse une paire d'outils adéquats. Le premier outil, epiGWAS, utilise l'inférence causale pour détecter les interactions épistatiques entre un SNP cible et le reste du génome. Le deuxième outil, kernelPSI, utilise à la place des méthodes à noyaux pour modéliser l'épistasie entre plusieurs polymorphismes mononucléotidiques (SNPs) voisins. Il tire également partie de l'inférence post-sélection pour effectuer conjointement une sélection au niveau des SNPs et des tests de signification au niveau des gènes. Les outils développés sont - au meilleur de nos connaissances - les premiers à étendre au domains des GWAS des outils puissants d'apprentissage statistique tels que l'inférence causale et l'inférence post-sélection nonlinéaire. En plus des contributions méthodologiques, un accent particulier a été mis sur l'interprétation biologique pour valider nos résultats dans la sclérose en plaques et les variations d'indice de masse corporelle.

MOTS CLÉS

Apprentissage automatique, statistique en grande dimension, GWAS, épistasie, génomique

ABSTRACT

By offering an unprecedented picture of the human genome, genome-wide association studies (GWAS) have been expected to fully explain the genetic background of complex diseases. So far, the results have been mitigated to say the least. This, among other things, can be partially attributed to the adopted statistical methodology, which does not often take into account interaction between genetic variants, or epistasis. The detection of epistasis through statistical models presents several challenges for which we develop in this thesis a pair of adequate tools. The first tool, epiGWAS, uses causal inference to detect epistatic interactions between a target SNP and the rest of the genome. The second tool, kernelPSI, instead uses kernel methods to model epistasis between nearby single-nucleotide polymorphisms (SNPs). It also leverages post-selection inference to jointly perform SNP-level selection and gene-level significance testing. The developed tools are – to the best of our knowledge – the first to extend powerful statistical learning frameworks such as causal inference and nonlinear post-selection inference to GWAS. In addition to the methodological contributions, a special emphasis was placed on biological interpretation to validate our findings in multiple sclerosis and body-mass index variations.

KEYWORDS

Machine learning, high-dimensional statistics, GWAS, epistasis, genomics