# *Association Analysis Using Set-Based Approaches in the Post-GWAS Era*

## *Dissertation*

*zur Erlangung des humanwissenschaftlichen Doktorgrades*

*in der Medizin*

*der Georg-August-Universität Göttingen*

vorgelegt von

Summaira Yasmeen

aus Khushab, Pakistan

Göttingen, 2021

## Thesis Committee Members

| | |
|---|---|
| **Supervisor:** | **Prof. Dr. Heike Bickeböller** |
| | Institut für Genetische Epidemiologie<br>Universitätsmedizin Göttingen<br>Georg-August-Universität Göttingen |
| **Second Member:** | **Prof. Dr. Tim Beißbarth** |
| | Institut für Medizinische Statistik<br>Universitätsmedizin Göttingen<br>Georg-August-Universität Göttingen |
| **Third Member:** | **Prof. Dr. Thomas Kneib** |
| | Professur für Statistik<br>Wirtschaftswissenschaftliche Fakultät<br>Georg-August-Universität Göttingen |

## Date of Disputation: 21st July, 2021

# AFFIDAVIT

I hereby declare that my doctoral thesis entitled ***"Association Analysis Using Set-Based Approaches in the Post-GWAS Era"***, was written independently by myself and without the use of any other sources or aids than quoted.

Summaira Yasmeen
Göttingen- 25th May, 2021

# ACKNOWLEDGEMENTS

It's the most difficult part of thesis to write, as I really don't want to miss any name but the list of amazing humans who contributed, and supported me throughout this doctoral journey is long. I will keep try to keep this section of thesis interesting for readers. Important: I have avoided copying typical phrases of thanks from the internet, the text written here is true-to-heart!

A heartfelt thanks to my PhD supervisor, Prof. Dr. Heike Bickeböller- as in German, 'Doktormutter', for her guidance, continuous support, and kind advice throughout my PhD research studies. Prof. Bickeböller not only ensured her support in my academic journey also she has been very understanding in stressful times during the PhD work. I am also very grateful to the two thesis committee members namely Prof. Dr. Tim Beißbarth and Prof. Dr. Thomas Kneib for listening to my reports on the state of my research sharing their ideas and advice in stimulating discussions.

I would like to pay thanks towards our research collaborators based in Munich from the PsyCourse Study namely Prof. Dr. Peter Falkai, Dr. Sergi Papiol, and Dr. Urs Heilbronner. A whole shout-out for my colleagues at the institute of Genetic Epidemiology for being very friendly and helpful in various situations. Andrew Entwistle- Thanks for proofreading this thesis and all previous published academic texts in my PhD work. Bernadette Wendel- Thanks for being always there- My person to go! I will miss our brain-storming sessions about research ideas. Katharina Stahl- Thanks for retreating me with vegan food. Albert Rosenberger- Thanks for being my next door colleague, I would definitely miss learning German phrases from him. Anja Sapara- Thanks for always greatly handling administrative issues particularly those which required German skills that I lack. Dr. Stefanie Friedrichs- a former colleague, I greatly acknowledge

her support for helping me getting started here smoothly! A funny moment that might make reader laugh, in the beginning of my PhD, around 10:00 pm I called Dr. Friedrichs that I am lost in a street and its name is similar to a street before. She asked what's the name, I responded, 'Einbahnstrasse'- that means one way street. However, she figured out where I was and helped me to reach home safely. ☺

My work presented in this thesis was financially supported by the Deutsche Forschungsgemeinschaft (DFG) in the context of my membership in the research training group "Scaling problems in statistics" (RTG 1644). I thank you all the members of RTG especially the coordinators Barbara Strauss and Dörte Dede.

I am greatly indebted to my parents Mr. Malik Muhammad Ramzan and Mrs. Zaib-un-Nisa. Thanks for believing in me and allowing me to pursue my dreams in a foreign land, gazillions of miles away from the family. I thank you my Dad for being the best guy in my life! Lastly I am thankful to my friends and family, in particular Amna Farooq, Muzammil Adeel, Aisha Shakoor, Aisha Younas, and Abiha Zahra!

It is important for me to write it here that my colleagues including my PhD supervisor ensured that I should not feel alone in my personal life struggles.

With amazing memories,
Summaira aka Sunny

# ABSTRACT

Genotyping arrays have greatly facilitated genetic epidemiological studies into genetic risk factors for numerous complex diseases such as psychiatric disorders. The use of genome-wide association analysis (GWAS) is unequivocally established. More recently, DNA methylation arrays have

enabled genome-wide profiling of the methylome, in addition to contemporary genetic epidemiology study design. An example of one such study is the Genetics of Lipid Lowering Drugs and Diet Network (GOLDN) Lipidomics Study, which identified methylation markers (CpG markers) and single nucleotide polymorphisms (SNPs), associated with the change in triglyceride levels after drug intervention.

Genotyping and methylation arrays assay several hundred thousand markers; however, single-marker association analysis suffers greatly from the burden of multiple testing. Set-based (SNP or CpG set) association approaches offer great flexibility, thus allowing the joint testing of a set of variants. For instance, a polygenic risk score (PRS) is a set-based approach, which, in addition to the strongly associated SNPs identified by large-scale GWAS, recruits SNPs with moderate to weak effects. The genotype information of the SNP set in the PRS is taken from an independent sample (target sample) and is then weighted by individual SNP effects derived from a relevant GWAS performed on a separate sample (discovery sample) into a cumulative score for each individual in the target sample. The resulting score, based on a SNP set or the PRS, is then regressed on the target phenotype. Such a regression model is evaluated by the amount of variance explained ($R^2$) by the PRS in the target phenotype. Another strategy of set-based association analysis is kernel machine regression (KMR): a semi-parametric regression approach, in which the effects of markers within a set (CpG set or SNP set) are modelled via a kernel function and thus evaluated by a single-component variance test. A kernel function computes pairwise genomic similarity between the individuals, that is, the inner product of a set of variants under analysis, maybe comprising a gene or a biological pathway.

For my first article, I performed a simulation study to evaluate the performance of PRS in correlated discovery and target traits by considering various sample sizes of the target sample, namely n=200, 500, and 1000. The PRS for correlated traits can be viewed as a situation of calculating

schizophrenia-PRS for psychosocial endophenotypes such as global assessment functioning (GAF) score or positive and negative syndrome scale (PANSS) score. Considering such a situation, I simulated four correlated target traits that had varying degrees of correlation ($r^2$) with the discovery trait, i.e., $r^2$= 1.00, 0.8, 0.6, and 0.4. The results demonstrated that the average $R^2$ estimates by the PRS roughly decreased by the square of the correlation between the target traits. In addition, the range of estimated $R^2$ is most inflated in the sample size of the target trait n=200. Thus, the simulation findings alert researchers conducting clinical studies with endophenotypes to the fact that they need to pay attention to two important factors: first, the sample size of the target trait and secondly, the shared amount of genetic correlation between the target and discovery traits.

In my second article, I implemented a KMR approach for set-based association testing of a CpG set. KMR has been successfully employed on SNP sets. In preparation of the second article, I used real and simulated datasets (based on a real dataset) provided by the Genetic Analysis Workshop 20 (GAW20) from the GOLDN study. GOLDN is a longitudinal study with individuals recruited from pedigrees. In my analysis, I only used independent individuals, which restricted the sample size in the real and simulated datasets to n<200. CpG sets were devised using the evidence of association reported by the GOLDN study in the real data set. For simulated datasets, true causal CpGs were provided by GAW20. Thus, I formulated candidate genomic regions of varying lengths while keeping the associated CpG(s) inside the region. The results replicated the evidence of association reported by GOLDN in the real data, and in simulated datasets albeit nominally. Moreover, in the simulated data, causal SNPs exert their full effect on the phenoytpes given when the causal CpG loci had no methylation (B-value=0). Thus, I also considered modelling an interaction term along with the main effects. The results yielded significant association.

As part of the discussion, simulation results on the performance of the linear kernel for a CpG set with original (B-values) and logit transformed methylation values (M-values) indicated that logit transformation results in a loss of power. There, I also considered analysing an additive kernel that combines the genotype kernel and the methylation kernel and then tests for association with the phenotype. The initial simulations suggest that an additive kernel with a CpG set including hypo, semi, and hypermethylated sites simultaneously might not improve the model over only including a SNP set. However, it appears fruitful to investigate further the situation in which only one type of methylation state is present in a CpG set.

# CONTENTS

# 1 INTRODUCTION

The completion of the Human Genome Project in 2003 enabled researchers to search extensively for genetic loci responsible for diseases such as type 2 diabetes mellitus (T2DM), coronary artery disease (CAD), cancer, and psychiatric disorders [1]. Since 2003, there have been extensive advances in numerous molecular techniques, from the first-generation DNA sequencing using Sanger's method to massively parallel sequencing technologies enabling rapid sequencing of the whole genome. The application of these new technologies led to a considerable amount of development in statistical genetics methods. Thus, a key focus of genetic research has been to identify the molecular aberrations that make humans more susceptible to disease or a more severe disease course, and to explain the genetic architecture of a disease (or a phenotype).

The molecular aetiology of diseases is complex; in addition to genetics, numerous other factors such as epigenetics and the environment also play some role in the susceptibility, development, and progression of diseases [2]. Various classifications of disease exist; for instance the number of genes causing and/or influencing a disease is one of the common methods to classify diseases, as in the terms monogenic – a single gene; oligogenic – a few genes; and polygenic – several to many genes [3]. Another classification bases on the disease prevalence in populations, i.e., the number of affected individuals in a population for a particular disease thus dichotomised into a rare or common disease [4].

Existing evidence suggests that most rare diseases exhibit a Mendelian pattern of inheritance, an example of which being maturity-onset diabetes of the young (MODY), which displays autosomal dominant inheritance [5]. Mendelian diseases are rare in population, i.e., in a sample of unrelated individuals with European ancestry, the proportion of affected individuals carrying the pathogenic variant(s) is small-this is also defined as penetrance. However, this is otherwise for a sample comprised of related individuals such as those belonging to a pedigree. In a pedigree, the penetrance of Mendelian diseases is high or even complete i.e., all individuals that inherited pathogenic variant(s) exhibit disease. This hints towards searching for chromosomal segment(s) harbouring disease-causing mutation(s) that tend to be localised within a pedigree owing to co-segregation. Thus, mapping of such loci has been very fruitful via linkage analysis, which is a statistical genetics method of locating chromosomal segments that co-segregate with the disease phenotype through families [6].

To date, the loci for almost 6,800 phenotypes have been successfully mapped using linkage analysis [7], for instance *CFTR* for cystic fibrosis, *HNF1a* for maturity-onset diabetes of the young (MODY), and *BRCA1* and *BRCA2* for breast cancer [8]. However, an important fact is that localization of co-segregated genomic regions with disease causing mutation(s) for individuals belonging to a pedigree does not necessarily mean that all such individuals will exhibit the disease, this is also referred to as reduced or incomplete penetrance. Incomplete penetrance has been observed in the pedigree studies for breast cancer i.e., individuals albeit carrying pathogenic mutations in *BRCA1* and *BRCA2* do not have breast cancer [8].

On other hand, the majority of common diseases exhibit a complex polygenic architecture, that is, an intricate interplay between several to many genetic loci spanned across the coding and non-coding parts of the

human genome [6]. The pattern of inheritance of these loci is not yet fully understood. Indeed, they seem to influence disease aetiology by interrupting multiple biological pathways and gene regulatory networks [3, 6]. One well-studied example of such a complex disease aetiology is diabetes mellitus - a highly heterogeneous group of diseases exhibiting different pathophysiology with hypoglycaemia as a common feature [5]. The identification of the genes behind MODY encouraged and accelerated the genetic studies into T2DM, the more prevalent diabetic phenotype [9]. Using pedigree-based study designs, genome-wide linkage analysis revealed an initial set of loci at *PPARG*, *KCNJ11,* and near *TCF7L2* for T2DM [5]. The linkage-based analysis for T2DM was not particularly fruitful; the acceleration in risk-variant discovery for T2DM has been primarily driven by the introduction of genome-wide association studies (GWASs). By definition, a GWAS is a statistical approach that scans genetic variants, genotyped on a commercial array for a number of unrelated individuals sampled from a population, to find statistical evidence of genetic variations associated with a particular disease. The first round of findings from the GWASs for T2DM confirmed evidence of strong association (odds ratio (OR) > 4.0) for previously identified loci through linkage analysis. In addition, it revealed a set of novel loci with modest to weak signals (OR approximately 1.05–1.35) near *CDKAL1*, *HHEX*, *SLC30A8*, *IGF2BP2*, and *CDKN2A* [5]. Nevertheless, the contributions made by pedigree-based study designs are unprecedented in both rare as well as common diseases.

Another well-known example of a common disease is CAD, for which genome-wide linkage analysis also unravelled an initial set of genetic variants, in a similar fashion to T2DM. For instance, a genome-wide linkage analysis conducted on the U.S GeneQuest cohort with 428 nuclear families identified six novel loci on chromosomes 3p25.1, 3p29, 9q22.3, 9p34.11, 17p12, and 21q22.3 [10]. Also for CAD, another genome-wide linkage analysis of 156 sibling pairs revealed two more genetic loci on

chromosomes 2q21.1–22 and Xq23–26 [11]. Thus, genome-wide linkage analysis enabled the discovery of only a handful of variants (with strong effect sizes) for common diseases with a polygenic architecture [12].

With the increasing amount of evidence furnished by the GWASs it became clearer that polygenic diseases are driven by several to many genetic variants with modest-to-weak effect sizes and minor allele frequencies (MAFs) >1% (also called common variants). These common variants do not necessarily cause the disease but rather influence the risk of developing the disease [13].

Along with the polygenic and highly multifactorial nature, the diagnosis of one common disease might also confer genetic predisposition to developing another distinct disease. For instance, patients with T2DM are at a higher risk of developing CAD than are non-T2DM patients - hinting towards a shared set of genetic variants in both distinct disease aetiologies. Results from a large-scale GWAS for T2DM and CAD has also shown a strong evidence of shared genetic correlation in both diseases i.e., genetic variants associated with increased risk of CAD are also associated with increased risk of T2DM [1, 14].

## 1.1 Association Analysis

Association determines whether a particular allele or genotype in a population is associated with the disease more often than expected by chance. Those positions in the DNA sequence displaying an exchange of a single nucleotide are called single nucleotide polymorphisms (SNPs). Let the alleles be denoted by $A$, and $a$, so that the individual genotype at any bi-allelic SNP site is $AA$, $Aa$, or $aa$.

In statistical modelling, in a traditional GWAS, the most common outcome of interest is either a quantitative measure of phenotype such as height, body mass index, or disease status such as diagnosis of T2DM

(yes/no), and the features (or variables) are several hundred thousands or millions of genotyped or imputed SNPs. The standard approach to analysing a GWAS is based on testing each genotyped SNP in the genome individually for statistical significance of its association with the phenotype under investigation. Logistic regression is employed for a binary phenotype and linear regression for a quantitative phenotype [15]. The first GWAS in 2005 on 100K genotyped SNPs with a sample size of 146 individuals gave robust evidence of association for complement factor H (CFH) with age-related macular degeneration with an Odds Ratio (OR) of 4.6 [15]. Later on in 2007, the explosion in GWAS analyses revealed that the majority of common risk alleles conferred effect sizes of < 1.5 OR [15].

It is estimated that a typical human genome differs from the reference human genome at 4.1 million to 5.0 million sites [16]. The latest information available at the 1000 Human Genomic Consortium project website for the phase 3 reports 84.4 million variants in n= 2504 individuals from 26 populations [17]. Nevertheless, SNP arrays genotype far less variants. This can be explained by the phenomenon of linkage disequilibrium (LD). In a sample of individuals from a population with a common genetic ancestry such as European, alleles in two SNPs that are physically close to each other appear together more often than would be expected by chance , thus these two SNPs are said to be in LD with each other [18]. Mathematically, the LD between two genetic variants can be quantified as a correlation between SNPs across a population. Two SNPs that are in strong LD can serve as proxies for one another [19, 6]. That is, if the correlation between the two SNPs is high, genotyping one of these provides almost complete genotype information of another. Therefore, a SNP array that genotypes ~1 million SNPs can effectively assay a larger proportion of the human genome [19]. However, the issue of array coverage also needs to be considered while performing the quality control and imputation of genotypes. Taking advantage of high LD among SNPs,

commercial genotyping arrays have been specifically designed to genotype SNPs that correlate with, or 'tag/represent', a large number of other SNPs in the human genome [19].

Over the last years, GWASs have made major contributions to the efforts of gene mapping by identifying numerous novel genetic associations. However, early studies had small sample sizes [15] and were thus underpowered to detect the small effect sizes expected for the common variants, hinting that these variants require large sample sizes. Hence, meta-analysis of available GWAS data from different studies was soon recognized as an appropriate method in order to achieve adequate sample sizes and the optimum power to discover genetic associations with modest to weak effect sizes [20]. For instance, a meta-analysis GWAS of T2DM, with ~16 million genetic variants in 62,892 cases and 596,424 controls identified 143 SNPs [21]. This approach led to the establishment of disease-specific consortia such as the Psychiatric GWAS Consortium (PGC), the Cognitive Genomics Consortium (COGENT), and the Genetic Investigation of ANthropometric Traits Consortium (GIANT).

Undoubtedly, the contribution of GWASs to a better understanding of complex diseases is unprecedented, but it does suffer several limitations. The following are some of the key limitations of GWAS:

1. High dimensionality – Multiple testing problem

A large sample size for performing a GWAS is necessary because it essentially requires testing hundreds of thousands of SNPs (high dimensionality of the data), resulting in hundreds of thousands of tests (the multiple testing problem). As a result, GWASs are underpowered to detect a major part of the genetic variance in a phenotype that might be explained by SNPs, which do not achieve the required significance level owing to multiple testing correction. The fraction of total variance (V) determined by genetics is often terms as genetic variance ($V_g$) or heritability ($h^2$). In human

genetics, an additive model of heritability is often assumed, simply summing the contributions of all the additive alleles influencing that trait [22].

2. LD hinders pinpointing the causal variants

Local correlation between SNPs in LD facilitates the initial identification of a locus i.e., genomic region but makes it difficult to discern the causal variant(s). Most GWAS-identified association signals so far map to non-coding regions of the genome [22], for which any biological interpretation is inherently challenging.

3. Missing Heritability: A post-GWAS challenge

The variance (V) of a phenotype can be sub-divided as a sum of two components, one part explained by genetics (heritability) and the other explained by environmental or other unknown factors. Given the polygenic architecture of complex diseases and unknown patterns of Mendelian inheritance, we assume an additive model of genetics for polygenic diseases; all genetic factors contribute towards $V_G$ in an additive fashion. Although evidence of a non-additive model of genetics (or heritability) is difficult to assess in humans, model organisms (for example, yeast, worm, fly, or mouse) have established epistasis as a pivotal component of the genetic architecture of complex traits [22, 23]. However, the identification of significant gene–gene interactions has been challenging in GWAS and post-GWAS experiments in humans, owing primarily to a lack of statistical power and to methodological challenges.

## 1.2 The Post GWAS -Era

In order to enhance our existing understanding of molecular underpinnings behind complex diseases, in addition to genetic variations, geneticists have investigated various other -omic or molecular processes such as epigenetic events. An epigenetic event can be defined as the

reversible attachment of a chemical cap that does not change the DNA sequence such as the addition of a functional group (methyl) or a protein (histone) to the DNA sequence [23]. These events control various molecular processes such as regulation of the transcriptional state of a gene, i.e. gene activation or gene silencing. In other words, it controls the production of the functional form of gene(s). Thus, epigenetic processes have also been exploited as biological markers for disease characterization. For example, in cancer, the over-expression of gene(s) is suspected as a function of observed high levels of DNA methylation (DNAm) in the blood sample. DNAm refers to the reversible addition of a methyl group to cytosine-guanine dinucleotide (CpG) sites in the DNA sequence. In parallel to the progress made by molecular techniques, numerous statistical methods have been developed to address the computational limitations of GWAS such as set-based approaches for association analysis.

### 1.2.1 Set-based approaches for association analysis

The GWAS design suffers from the curse of high dimensionality of the data, strong LD between variants, and multiple weakly associated variants, while set-based approaches allow joint testing of a subset of variants from the total set of genotyped variants, greatly reducing dimensions of the data [22].

Several methods have been proposed to combine SNPs in a set. One such method is to use the genomic annotation and/or genomic features of SNPs and then mapping these to a gene or multiple genes(s) involved in biological pathway(s) [13]. Another strategy is to exploit the statistical evidence of association obtained through GWAS analysis of SNPs such as p-value and thus partitioning the genome-wide SNPs into several subsets of SNPs ranked by their *p-values*. The joint analysis of several SNPs together not only yields improved power in settings where SNPs individually have moderate to small effect size but also greatly reduces the burden of multiple

testing (Single marker analysis in the GWAS design). In addition, tag SNPs are in LD with causal loci and thus a set-based approach allows testing the association of a batch of biologically important SNPs with the phenotype [24], instead of an individual SNP. Many GWASs may not release individual-level data owing to logistic challenges or data confidentiality agreements. Instead, it is much more likely that a marginal test statistic for association with the outcome is available for each individual SNP. The individual SNP association estimates for numerous complex phenotypes such as schizophrenia (SZ), major depression disorder (MDD), and autism are publically accessible. These statistical estimates from large consortia have been used as weights to aggregate the genome-wide genotype SNPs for an individual into a single value estimate, named the **polygenic risk score** (PRS). PRS can be viewed as a set-based approach, which initially recruits all genotyped SNPs whose association estimates are available. Let us assume the set with all genotyped SNPs is a superset *M*. In the next steps, several proper subsets of SNPs from the set *M* are formulated, ranked by the *p-values* of single SNPs. $S$ is a proper subset of $M$, if there is at least one element of $M$ that is not an element of $S$. Moreover, all the elements in these proper subsets are necessarily members of the superset *M.* From these several proper subsets of SNPs, several PRSs are then constructed. Each PRS is then associated with the phenotype of interest and through several regressions, the optimal PRS is selected, which is actually constructed on a subset of SNPs (more details in the methods chapter).

Building the additive, albeit weighted sum of additive genetic heritability is the most commonly exploited model, as it allows parametric modelling based on linear regression. However, this model is indeed an oversimplification for polygenic diseases. Therefore, a hybrid regression approach called **kernel machine regression** (KMR) is proposed. In KMR covariates such as age, gender, and smoking status are modelled parametrically, and the joint effect of a set of genetic or epigenetic markers

non-parametrically. More specifically, the non-parametric effect of multiple markers is modelled via a kernel (more details in the methods chapter). The KMR framework has shown to be robust as it allows great flexibility in the functional relationship between the SNPs belonging to a set and the disease or the outcome of interest. Numerous kernels exist, such as the linear kernel, Gaussian kernel, and quadratic kernel. For genetic variants, the linear kernel has been successfully employed in association testing [25].

### 1.2.2 DNA methylation: CpG sites

A big challenge after GWAS is to explain the functions of the identified SNPs, and to illustrate the mechanisms underlying the associations. The current GWAS catalog stats released on 04$^{th}$ April 2021, hosts data for 158,358 SNPs with 255,015 associations from 5002 GWAS publications [26]. Most of these associated SNPs are located in the non-coding region of DNA, which might be the genomic regions harbouring the transcriptional machinery of gene(s) such as promoters, enhancers, or silencers [27].

DNAm is a reversible-dynamic epigenetic event; thus, the degree of methylation at a CpG site or several closely located CpG sites of DNA determines the transcriptional state of nearby gene(s) [28, 29, 30]. For instance, hypo-methylation of a DNA sequence in a promoter triggers gene activation while hyper-methylation signals gene silencing [28, 29]. According to the ENCODE database, the human genome has approximately 28 million CpG sites that exhibit varying methylation patterns [31]. Similar to genotyping arrays, Illumina also provides DNAm arrays, namely the Illumina Human Methylation 450 K (also 850K: K refers to 1000 i.e., 1K=1000 sites) and Infinium Methylation EPIC (EPIC) BeadChips (Illumina Inc, San Diego, CA) [32]. These arrays have limited coverage of the methylome and can only detect up to 870K CpGs across the human epigenome, leaving a large proportion of CpG sites unmeasured. DNAm patterns are specific to tissues and developmental stages, and

change over time [28]. DNAm via arrays is usually profiled in the whole blood samples. Whole blood contains several cells of distinct types in various proportions; this is one of the prominent confounding factors of DNAm data generated from arrays. Analogous to the whole-genome sequencing technique of DNA, DNAm profiling can also be done at single-base-pair resolution using the whole-genome bisulphite technique, but is expensive. In the whole-genome bisulphite technique, DNAm is profiled in the cell lines.

The DNAm level of a CpG site is a beta-distributed continuous value varying from 0 to 1. At each CpG site, methylation is quantified by the beta value, denoted as:

$$B\text{-}value := M \,/\, (M + U + a),$$

where $M > 0$ and $U > 0$ denote the methylated and unmethylated signal intensities [33, 34]. The offset $a \geq 0$ is usually set equal to 100 and is added to $M$+$U$ to stabilize beta values when both $M$ and $U$ are small [34]. The distribution of an individual CpG site across various individuals can be considered as beta-distributed with values bounded between zero and one [33]. If the methylation of a site is zero, it refers to a state of no methylation, a biological indicator of transcriptional activity; and a value of one is maximum methylation, a biological indicator of minimum or no transcriptional activity. In practice, the methylation state of a CpG site depending upon the corresponding measured *B-value* of DNAm belongs to one of the three classes, i.e., hypo-methylated (*B-value* <0.20), semi-methylated (*B-value* >0.20 and <0.70), or hyper-methylated (*B-value* >0.70) [35]. Gaussian regression with the beta-distributed *B-values* of DNAm data is problematic. The variance of *B-values* is usually smaller near the boundaries than the middle of the interval (0 , 1), implying violation of the homoscedasticity assumption required in Gaussian regression [35]. To address this problem, several modelling strategies have been proposed,

including Gaussian regression with logit-transformed *B-values*, called *M-values*, and generalized regression models incorporating *B-values* as responses, e.g. beta regression.

In addition, alike GWASs epigenome-wide association studies (EWAS) analyses have been conducted to study the disrupted genome-wide patterns of DNAm for numerous diseases such as for metabolic syndrome, schizophrenia, and inflammatory or autoimmune disorders [35]. For EWAS analysis, an identical approach to the traditional GWAS has been used, which also suffers from similar limiting factors. Moreover, recent studies have demonstrated evidence for loci harbouring SNPs that influence the methylation state [31]. Such loci have been termed methylation quantitative trait loci (methQTLs). In most methQTL, the correlations with the nearby genotypes (cis-genotype: on same DNA strand) are most pronounced. There is some evidence that SNPs can also influence methylation state(s) of CpG site(s) in trans (located on another strand of DNA), but this does not seem to be as prevalent as *cis*-effects. It is also important to note that in most of these previous studies, the true causative SNP was not identified unequivocally [22]. In some cases, disease-associated epigenetic variation could arise prior to disease onset, but still not be causative for the disease [23]. This type of epi-phenomenon could be a result of confounding, when an environmental factor such as smoking, or a genetic variant, induces both aberrant epigenetic states and disease. These potential relationships between epigenetic variation and complex disease have important implications for the design and analysis of EWAS [30]. There are several EWAS designs being opted, such as monozygotic twins, and longitudinal cohorts [35]. In addition to single marker association analysis, EWAS has also been employed in the elucidation of the drug response by recording pre and post-treatment methylation data.

More recently along with GWAS, researchers have started performing EWAS on the same individuals along with considering the gene expression datasets as well. These datasets have enabled integrated analysis of multiple layers of omics data for the phenotype of interest with the aim of improving our existing understanding of disease. One of the noticeable examples is the integration of the gene expression data from blood ($n$ = 14,115 and 2765) with the GWAS results for T2DM, which identified 33 putative functional genes, three of which were targeted by approved drugs [21]. A further integration of DNAm ($n$ = 1980) and epigenomic annotation data highlighted three genes (*CAMK1D*, *TP53INP1*, and *ATP5G1*) with plausible regulatory mechanisms, whereby a genetic variant exerted an effect on T2DM through epigenetic regulation of gene expression [21].

## 1.3 Objectives and Outline of Thesis

The main objective of the research work done for this thesis is an evaluation of the set-based association approaches. Two statistical approaches are part of this thesis; one is a purely parametric regression approach: the polygenic risk score. The other approach is semi-parametric in nature and is named kernel machine regression. The research towards this thesis aims to enhance the existing understanding of both methods through an extensive simulation study and by using other omics data (epigenetic data: DNA methylation) with and without considering the interaction between genetic and epigenetic data. For the PRS method, I performed extensive simulations with varying sample sizes (small to moderate as is common in clinical studies). In addition, I evaluated the performance of PRS for correlated complex phenotypes instead of only using identical or similar phenotype. The KMR method has been exploited fairly well for SNP sets. In this thesis, my focus was to review the performance of the kernel for CpG markers. In addition, I also considered modelling an interaction term between SNPs and CpGs

# 2 METHODS

## 2.1 Model Notation

Let us suppose we have genotyped *m* SNPs for *n* individuals, for which we measured a quantitative, normally distributed phenotype *y*. A traditional GWAS analysis proceeds by sequentially testing the null hypothesis of no association between the SNP and the phenotype for each single SNP [36, 37]. For a bi-allelic SNP with alleles denoted as *A* and *a*, there are various genotype-coding methods for the three possible genotypes i.e., *AA, Aa,* and *aa*. I will use the count of minor alleles in a genotype. Let us assume allele ***a*** is the minor allele, thus 0 for genotype *AA*, 1 for *Aa*, and 2 for *aa*. In GWAS analysis, an additive model is assumed, which implies that if the risk conferred by the minor allele *a* increases by r-fold for the heterozygous genotype (*Aa*) then it increases 2r-fold for the homozygous genotype (*aa*) [38].

The linear regression model in a typical GWAS for a quantitative outcome for an individual *i* is as follows:

$$y_i = x_i^T \beta + \beta_q g_{iq} + \varepsilon_i \; ; \quad \varepsilon_i \sim^{i.i.d} \mathcal{N}(0, \sigma^2) \qquad \textit{Eq. 2.1}$$

where $g_{iq}$ denotes the minor allele count for q$^{th}$ SNP; $q$ =1,…,$m$ in the i$^{th}$ individual; i =1,…,n , and $\beta_q$ is the regression coefficient for the q$^{th}$ SNP. $x_i^T$ denotes the transposed ($T$) vector of considered covariates, such as age, gender, and/or smoking status including the intercept; $\beta_x$ denotes the vector of corresponding regression coefficients. $\varepsilon_i$ is the vector of error

terms, which are identically independently distributed (i.i.d.) and follow a normal distribution. We consider a test for no association between the genotype $g_{iq}$ and outcome $y_i$ , i.e., $\beta_q = 0$ as given in *Eq. 2.1*. Given *m* genotyped SNPs, *m* linear regression models are performed for a traditional GWAS. Thus, for the q$^{th}$ SNP, the fitted regression model estimates the regression coefficient $\hat{\beta}_q$; this is also called the estimated effect size. In addition, the fitted regression model also provides association statistics for the estimated effect size $\hat{\beta}_q$, which includes the standard error (SE) of the estimated effect $\hat{\beta}_q$, and the *p-value* for the q$^{th}$ SNP with the outcome $y_i$. If the *p-value* for $\hat{\beta}_q$ is less than the defined level of significance *a*, the null hypothesis can be rejected. The commonly used alpha values are 0.01 and 0.05. However, in a GWAS for *m* genotyped SNPs, *m* regression analyses are performed, i.e. simultaneously testing *m* null hypotheses of no association between the genotypes of *m* SNPs and the phenotype $y$, for *n* individuals. Thus, the individual *p-values* for the *m* SNPs need to be corrected for multiple testing. A commonly used method for multiple-comparison correction is the Bonferroni correction; other methods include the Tukey-Kramer and Scheffe method [39, 40]. The Bonferroni correction is a conservative multiple-comparison correction method that resets the alpha value (*a* =0.05) for the *m* regression tests to $\alpha/m$ and thus the signficance level for the estimated *p-value* is adjusted for multiple tests. For GWAS analysis, the proposed genome-wide significance level is $5 \times 10^{-8}$ [39].

The association statistics for the *m* tested SNPs altogether are termed summary statistics of a GWAS analysis. In a typical summary statistics report of a GWAS analysis, each line represents the association statistic $A_q$ for the q$^{th}$ SNP, such that $A_q = \{\hat{\beta}_q, SE_q, p - value_q\}$. Along with the association statistics, other information such as the genomic location of SNPs is also given. These summary statistics reports are publicly available for a number of published GWASs on the GWAS catalogue database [26].

## 2.2 Polygenic Risk Score Analysis

The main goal of GWASs so far has been to identify causal variants that tell us about the biology of the phenotype and propose ways for targeted treatments [22, 36, 41]. GWAS findings have unravelled robust statistical association evidence for quite a number of loci, albeit the number of statistically significant SNPs is small and the effects are far from explaining the missing heritability [6, 22]. Even though many SNPs with weak to moderate effects can be assumed as associated to phenotypes however, a wide majority of loci do not achieve any genome-wide significance level owing to the multiple testing burden [22, 37, 41].

A polygenic risk score (PRS) is an individual-level score of genetic risk, which can be conceptualised as an aggregate measure of allelic counts across a set of genetic variants weighted by effect sizes, derived from an appropriate GWAS result [42, 43]. Initially, PRS computation was restricted to SNPs that reached genome-wide significance [42]. However, with the availability of GWAS results from much larger studies such as those from consortia, PRS also included SNPs that did not reach genome-wide significance [43, 44, 45, 46, 47].

In principle, the computation of PRS requires two main ingredients: First a sample of independent individuals with genotype and phenotype information for a complex trait/disease - the so-called target trait. Secondly, an appropriate GWAS summary statistic estimated in an independent sample with identical or correlated phenotype to that of the target trait is needed. This phenotype is termed the discovery phenotype. Usually the sample size of GWAS for the discovery phenotype is quiet large in comparison to the target trait. Let us suppose a *SNP set R*, which is a proper subset of SNP set *M*. Here SNP set *M* refers to the *m* genotyped SNPs for the target trait. For simplicity, we assume the GWAS summary statistics information is available for the genotyped SNPs in the SNP set *M*. The *SNP*

*set R* has $r$ SNPs. *G* is then an $n \times r$ matrix, which has genotype information for the $r$ SNPs and $n$ individuals in the target sample. Each element of the matrix *G,* denoted as $g_{iq}$ is the genotype information (0, 1, or 2) for the q$^{th}$ SNP, $q=1, \ldots, r$ , in the i$^{th}$ individual $i=1, \ldots, n$. $\hat{\beta}_q$ is the genetic effect size estimated for a single SNP in an external large-scale GWAS of a discovery trait. Thus, PRS for the individual $i$ can be computed as follows:

$$PRS_i = \sum_{q=1}^{r} \hat{\beta}_q G_{iq} \qquad Eq.\ 2.2$$

$PRS_i$ is the cumulative sum of minor allele counts of genotypes across the $r$ SNPs, weighted by the respective estimated effect of the SNPs [42]. The GWAS summary statistics for the discovery trait are usually estimated on a very large sample size, such as several thousand individuals, and are publicly accessible. For instance, at the Psychiatric Genomic Consortium (PGC) for schizophrenia, performed a GWAS analysis on 36,989 cases and 113,075 controls with European ancestry, and published the GWAS summary statistics [44].

PRS analysis can be viewed as a search of the *SNP set R,* such that $R \subset M$. In order to find the *SNP set R,* several subsets of SNP set *M* are formulated. Prior to sub-setting the SNP set *M*, preliminary SNP filtering is recommended to address problems such as multicollinearity. Thus, SNP filtering followed by the sub-setting of SNP set *M* is a two-stepped approach. The first step is clumping, which removes correlated SNPs from the SNP-set *M*; the resulting SNP set is denoted as $M_{clumped}$, followed by the second step of sub-setting SNPs in the SNP set $M_{clumped,}$ using a thresholding-based criterion. This approach is called clumping and thresholding (C+T), sometimes also known as pruning and thresholding (P+T).

In the first step, clumping (C) is performed such that SNPs in set *M,* which have weak correlation with one another, are retained, thus avoiding

multicollinearity among SNPs [45]. Clumping selects the most significant SNPs iteratively by computing the correlation ($r^2$) between an index SNP and its nearby SNPs within a window based on genetic distance between SNPs [43]. This removes the nearby SNPs that demonstrate greater correlation with the index SNP beyond a given threshold [45]. The recommended threshold is $r^2 = 0.8$, which we also used in our analysis [43]. The clumping step prunes redundant correlated effects caused by linkage disequilibrium (LD) between SNPs [43]. For clumping, keeping in consideration the LD pattern in the human genome, a window size between 250 kbp to 500 kbp is recommended [43, 46]. However, this procedure may also remove independently predictive SNP(s) that are in LD with the index SNP.

The second step is thresholding (T). The GWAS summary statistics of SNPs in SNP-set $M_{clumped}$ obtained from the discovery sample are ranked from lowest to highest *p-value* [43]. Let us suppose a vector *S of* length *s* comprises of positive real numbers ranging between 0 and 1. Each element in the vector *S* represents the discrete *p-value* grid points deployed as the *p-value* threshold to create *s* number of proper SNP subsets from the SNP set $M_{clumped}$. Each subsequent SNP set which is subset of SNP set $M_{clumped}$, necessarily incorporates cumulatively SNPs from the previous *p-value* threshold. That is, SNPs at the 0.001 *p-value* threshold are a proper subset of SNPs at the *p-value* threshold of 0.01.

In this way, a PRS that considers genome-wide SNPs can be defined as a weighted sum of allele counts for SNPs meeting a *p-value* threshold, yielding a set of PRSs for a vector of thresholds. In theory, this *p-value* based thresholding approach is also applicable to considering SNPs only within a gene set or a pathway instead of all genome-wide genotyped SNPs on an array.

Combining information for non-significant SNPs, but with lower-ranked *p-values*, in addition to significant SNPs, the PRS analysis can be considered as a search to find an "optimal" *p-value* threshold corresponding to the *SNP set R*, which has the maximum possible predictive power compared to all other selected *p-value* thresholds. Thus in a parametric regression framework, the resulting PRS for each grid point or *p-value* threshold is regressed on the phenotype as target trait.

Let us assume $y_i^{target}$ denotes a normally distributed phenotype of the target trait, the regression model is as follows:

$$y_i^{target} = \beta_{PRS} PRS_i + \varepsilon_i \qquad Eq\ 2.3$$

where $\beta_{PRS}$ is the regression coefficient of PRS. The *s* number of distinct p-value thresholds defined in the vector *S*, also determine the number of computed PRSs and thus the number of regression models in *Eq. 2.3*. The explained variance ($R^2$) and association *p-value* of $y_i^{target}$ with the PRS is compared for those analysed models. The PRS among *s* the PRSs that explains the maximum variance in the phenotype, with significant evidence of association between $y_i^{target}$ and PRS of the model, is referred to as the optimal PRS ($PRS_{opt}$), and the corresponding *p-value* threshold is called the optimal *p-value* threshold ($p_{opt}$*-value)*. Both steps, clumping and thresholding represent a statistical compromise between signal and noise [43]. Clumping aims to ensure the inclusion of truly predictive variants and reducing noise in the score by excluding variants that are highly correlated, while thresholding allows consideration of SNPs beyond the significance level [43, 45]. Thus using genome-wide SNPs, PRS analysis outputs a *SNP set R*, which can be further subjected to gene or pathway-based enrichment analysis.

## 2.3 Kernel Machine Regression

Kernel methods are a machine learning class of algorithms that allow association testing of a SNP set with a phenotype of interest, reducing the dimensionality of tests greatly in comparison to the traditional GWAS [47]. Instead of single SNP analyses as in *Eq. 2.1* we now consider a SNP set *R* with $r$ number of SNPs where $g_i$ denotes the vector of genotypes for the i$^{th}$ individual, such that $g_i = (g_{i1}, g_{i2}, \dots, g_{ir})^T,\ i = 1., .., n.$ Unlike the definition of SNP set *R* in the PRS analysis section, here the SNP set *R* can be the SNPs in the SNP set *M*, mapped to a gene or multiple genes belonging to a pathway.

The regression model for the phenotype $y_i$ is as follows:

$$y_i = x_i^T \beta + h(g_i) + \varepsilon_i \quad \textit{Eq. 2.4}$$

where $h(g_i)$ is an unknown function, which models the genetic information for individual *i* in the model. $x_i^T$ is the transposed vector of covariates that are parametrically modelled by the linear model. The genetic effects can be modelled via the function *h* with great flexibility, e.g., parametrically (such as via a PRS) or non-parametrically via a kernel function. Thus, in the kernel machine regression (KMR) model in *Eq. 2.4*, we test for association between a SNP set modelled via function $h(\cdot)$ and the phenotype, the hypothesis is as follows:

$$H_0 : h(\cdot) = 0 \ \ versus \quad H_1 : h(\cdot) \neq 0$$

### 2.3.1 Kernel function and association testing

Let us assume the unknown function $h(g_i)$ lies within a reproducing kernel Hilbert space $\mathcal{H}_K$ generated by a positive definite kernel function $k(\cdot,\cdot)$. A reproducing kernel Hilbert space $\mathcal{H}_K$ allows the specification of a user-defined feature map $\varphi$ that in turn allows the transformation of data points from their original space into a higher dimensional feature space [48, 49].

The specific properties of the defined feature space $\varphi$ do not require its explicit evaluation [49]. However, as per Mercer's theorem, if a function $k(\cdot,\cdot)$ on the data points $i$ and $j$ satisfy Mercer's constraints, then there exists a function $\varphi(\cdot,\cdot)$ that maps $i$, and $j$ into a higher dimension [50].

Mathematically, a kernel can be represented as follows:

$$k(g_i, g_j) = < \varphi(g_i), \varphi(g_j) >$$

Here $k$ denotes the kernel function, $g_i$ and $g_j$ are the $r$ dimensional inputs from the genotype matrix $G$ for any two individuals $i$ *and* $j$. $\varphi$ is a map from $n \times r$ −dimensional space to $n \times n$ −dimensional space. Thus, the $n \times r$ −dimensional matrix $G$ *is* converted into a $n \times n$ −dimensional kernel matrix $K$, which is a symmetric and semi-definite positive matrix. Any element of the resulting $n \times n$ −dimensional matrix quantifies the similarity between any two individuals $i$ *and* $j$ determined by the specified kernel function $k(g_i, g_j)$. Thus, $K$ is also called the genomic similarity matrix [50].

For genetic data, a simple and popular choice of kernel function is the linear kernel: $k(g_i, g_j) = g_i^T g_j$, which is a dot product between any two data points [47]. This indicates that the overall genetic effect is a linear combination of the individual effects in the *SNP set R*. Other kernels include the polynomial kernel $k(g_i, g_j) = (g_i^T g_j + c)^d$ where $c$ is a constant term and $d$ is the polynomial degree, and the Gaussian kernel $k(g_i, g_j) = exp\left(-\|g_i - g_j\|^2 / \rho\right)$ [47]. To quantify similarity or dissimilarity between any two data points, numerous distance-based methods are available, such as Euclidian distance, Manhattan distance, Cosine similarity, Minkowski similarity, and Jackard Index [51]. For genetic data, methods such as identity–by-state, identity-by-descent, or shared-haplotype-based measures have been frequently opted. The details on these various general

as well as genomic similarity measures, and relationships between them can be found elsewhere [25, 26, 52, 54].

In the KMR framework, once the kernel function is specified, the next step is to test the hypothesis. In *Eq. 2.5*, $H_0 : h(g_i) = 0$, is equivalent to $H_0 : \sigma_K^2 = 0$, where $\sigma_K^2$ is the variance explained by the kernel function. Thus, a high-dimensional test is reduced to testing a single variance component. The regression model presented in *Eq. 2.4* can also be viewed as a linear mixed model (LMM), assuming fixed effects for the design matrix $X$, a nonparametric function $h(G)$ modelling the genetic information as random effect with $h(\cdot) \sim \mathrm{N}(0, \tau^2 K)$, and $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 I)$ [52, 50, 47]. $\tau^2$ is the unknown variance component, which is expressed as a function of the scaling parameter $\lambda$ and the variance $\sigma_K^2$ as follows: $\tau^2 = \lambda^{-1}\sigma_K^2$ [47]. The overall variance $\sigma^2$ can be defined as $\sigma^2 = \sigma_K^2 + \sigma_\varepsilon^2$, and $\sigma_\varepsilon^2$ is the residual variance. The conditional distribution of $y$ given the random effects $h$ is normal: $y|h \sim \mathcal{N}(X\beta + h, \sigma^2 I)$ and marginally (averaged across the individuals) $y \sim \mathcal{N}(X\beta, \tau^2 K + \sigma^2 I)$ [49, 52, 54]. We can estimate $h$ by noticing the fact that the distribution of $y$ and $h$ is jointly normal and their covariance is $\tau^2 K$ [52, 50, 47]. Making use of the conditional multivariate normal distribution, the expectation of $h$ given the observation $y$ can be estimated as $\tau^2 K \sum^{-1}(y - X\widehat{\beta})$ [49, 52, 54]. The estimates $\widehat{\beta}$ and $h$ are obtained by minimizing the penalized likelihood function for the KMR, which are equivalent to the best linear unbiased estimator and the best linear unbiased predictor of the LMM [47]. This connection bridges machine learning and regression statistics, specifically LMM and KMR, and allows for a unified framework of model fitting and statistical inferences [47]. We can estimate the variance component parameters $\tau^2$ and $\sigma^2$ and any unknown parameter in the kernel function by maximizing the likelihood of the LMM. Thus, within the KMR framework, we test the following hypotheses, which intuitively are all the same: $H_0 : h(G_i) = 0$ *or* $H_0 : \sigma_K^2 = 0$ *or* $H_0 : \tau = 0$

The score test statistic can be derived by taking the first derivative of the restricted maximum likelihood (REML) equation with respect to $\sigma_K^2$ and evaluating it under the null hypothesis [47]. The score statistic $Q$ follows a mixture of chi-squared distributions under $H_0$ and takes on the following quadratic form:

$$Q = \frac{1}{2\sigma_\varepsilon^2}\,(y - \hat{y})^T K (y - \hat{y}) \qquad \textit{Eq. 2.5}$$

where $\hat{y}$ is the fitted value of $y$ under the null model, and is easy to fit with standard regression models for fixed effects (e.g. linear regression for quantitative traits, or logistic regression for binary traits). The test statistic $Q$ depends on the true covariance matrix $\Sigma$ of $y_i$ , which is often unknown in practice and requires estimation of a large number of parameters. Although the sample covariance can be used to estimate $\Sigma$, it is not stable when the number of SNPs in the SNP set $R$ is large or moderate and the number of individuals $n$ is small. Some statisticians use $\sigma_\varepsilon^2$ in the denominator to compute the test statistics; others ignore this term and a formal derivation based on derivatives of the log-likelihood function would use $\sigma_\varepsilon^4$ in the denominator. However, these variations are just different scalings of the quadratic form. As long as the scaling is considered, i.e. assuring that Q follows a mixture of chi-squared distributions, the resulting *p-value* will be valid [26, 49, 54]. The Satterthwaite approximation - an anti-conservative approach and the Davies method- an analytical solution, have been used to compute *p-values* in the KMR. Between both computation methods for the *p-value*, the Davies method is more accurate. However, owing to computational constraints, the Satterthwaite approximation can also be used *[47]*.

# 3 SUMMARIES

## 3.1 PRS Approach in Correlated Phenotypes with Moderate to Small Sample Sizes

My first article is a simulation study aimed at investigating the performance of polygenic risk scores (PRSs) in correlated phenotypes with varying sample sizes, typical in the clinical setting; with an application to real data [53]. The aims and questions of this study were motivated by our collaboration with the researchers from the PsyCourse study. As stated above, PsyCourse is a multicentre, transdiagnostic longitudinal study of the affective-to-psychotic continuum that combines longitudinal deep phenotyping and dimensional assessment of psychopathology [54].

It is known that affective and psychotic disorders partially share psychopathological features and are genetically correlated [55]. Affective disorders, also called mood disorders, are mainly characterised by mood episodes and most typically involve bipolar disorders (Bipolar I and II), major depressive disorder, and mania [56]. Psychotic disorders, also referred to as delusion disorders, are mainly characterised by severe mental disorder that causes abnormal thinking and perceptions such as schizophrenia and schizoaffective disorder [56].

At the time I employed the data from the PsyCourse study [54], the sample size was n= 771 individuals. Key questions addressed in this research work are as follows:

I. How much variance in the target traits of various sample sizes i.e., n=200, 500, and 1000 can be explained by the PRS, if both discovery and target trait are identical?
II. In another situation in which the target and discovery trait are not the same but are correlated, how much variance can the PRS still explain in the target trait?
III. How much prediction capability of the PRS is lost for target traits of multiple sample sizes that have a strong to weak degree of correlation with the discovery trait?
IV. Can the results of the simulation analysis performed to address the questions stated in I, II, and III be interpreted with the real data application from the PsyCourse study? If Yes, then how?

Before stating the findings of the simulation study, I would first like to introduce the simulation setup briefly. Keeping in consideration the sample size used in the GWAS summary statistics available for schizophrenia from the Psychiatric Genomic Consortium(PGC) [44], I simulated datasets for a quantitatively distributed discovery phenotype for N=34,000 individuals. Effect sizes for the causal markers were derived from the total heritability assumed for the discovery trait, i.e., 80%. I then performed a GWAS analysis to generate the summary statistics for the discovery sample. Furthermore, I generated datasets for four target traits, namely T1, T2, T3, and T4. The phenotype generation model for T1 was identical to that of the discovery trait. The target traits T2 to T4 were generated keeping a correlation $r^2$ =0.8, 0.6, and 0.4 with T1. For all target traits, we generated datasets with varying sample sizes, namely n=200, 500, and 1000. The number of replicates for the simulation analysis described above was set to 100.

Given an identical discovery and target trait scenario as stated in question I, i.e. 100% common genetic aetiology, the variance ($R^2$) explained by PRS substantially reduced from that of the total simulated trait heritability. The loss in $R^2$ becomes more evident with the decreasing

sample sizes for T1. On average, the PRS based on summary statistics of the discovery trait explained 40% of the true simulated heritability. With the correlation structure between T1 and other target traits (T2-T4), the simulation results demonstrated an interesting agreement with a formula for loss in average explained variance by the PRS for sample size n=1000. On average, the loss in $R^2$ decreased by the square of the correlation between T1 and other target traits. Thus, from T1 (average $R^2 = 0.32$), the average $R^2$ estimates decrease for T2 to $R^2 = 0.82 \times 0.32 = 0.21$, for T3 to $R^2 = 0.62 \times 0.32 = 0.12$, and for T4 to $R^2 = 0.42 \times 0.32 = 0.05$. This gradual decrease in the average $R^2$ estimates by the PRS from T1 to T4 corresponds well with decreasing empirical correlation among target traits. With the decreasing sample size for T2-T4, the loss in $R^2$ by PRS became more evident. Moreover, the difference in the smallest and largest values of $R^2$ across the 100 simulation replicates increased with decreasing sample size, indicating a higher probability of finding false positive results.

I compare these findings of my simulation analysis with results of a dataset from the PsyCourse study. Briefly, after removing missing data from our considered phenotypes for the analysis, our sample size reduced to n=653 individuals. The individuals in the PsyCourse study belonged to two diagnostic groups: affective (n= 266) and psychotic (n=386). Thus, I analysed the total dataset and each diagnostic group separately. In the analysis, I used two well-developed psychosocial functioning scores, namely the Global Assessment Function (GAF) score and the Positive and Negative Syndrome Scale (PANSS) score from the PsyCourse dataset as target phenotypes. I used the GWAS summary statistics for schizophrenia (SZ; as discovery trait) available from the Psychiatric Genomic consortium. Only a small amount of variance was explained by the SZ-based PRS for GAF and PANSS.

## 3.2 Kernel Machine Regression for DNA Methylation Data (CpG) and Modelling the Interaction Term between SNPs and CpG Variants

In the second article, I investigated a very well-known set-based semi-parametric regression approach called kernel machine regression (KMR) now for DNA methylation (DNAm) data [57]. Previous research done by our group [58, 59] investigated the existing kernels and developed new kernels for genetic data association testing with the KMR approach. In the methods section, I have briefly highlighted how KMR works for genetic datasets. However, unlike the SNP or genotype data that are encoded as the count of minor alleles in a SNP (i.e. 0, 1, or 2 respectively), the DNAm data for CpG markers vary continuously between 0 and 1, depicting the methylation state of the respective marker. An individual CpG marker is assumed to have beta distribution, with two parameters $\alpha$ and $\beta$ controlling the shape of distribution. Thus, data from methylation are not necessarily normal.

The research work towards the second article is my contribution to Genetic Analysis Workshop 20 (GAW20). GAW20 provided genome-wide genotype (bi-allelic SNPs) and DNA methylation levels (CpG) from the Genetics of Lipid Lowering Drugs and Diet Network (GOLDN) study [60]. GOLDN is a longitudinal (four data points: visits 1 to 4) family-based study involving 991 participants of European descent. The study goal was to localize novel loci contributing to triglyceride (TG) and very-low-density lipoprotein cholesterol (VLDL-C) response in connection with a lipid-lowering drug. Blood levels of TG and VLDL-C were measured before the diet (visits 1 and 2) and after drug intervention (visits 3 and 4). DNAm levels of CpGs were collected only at two time points, during visits 2 and 4. Visit 1 and 2 were one day apart from each other, as was the case for visits 3 and 4. However, visit 2 and 4 were three weeks apart from each other. The data obtained on DNA methylation during visit 2 and 4 were generated using the 450K Illumina methylation array. EWAS analysis revealed four

CpGs markers, namely cg00574958, cg17058475, cg01082498, and cg09737197 in intron 1 of the carnitine palmitoyltransferase 1A (CPT1A) gene with strong evidence of association with VLDL-C and TG (Irvin et al., 2014).

In addition to the real datasets from GOLDN study, GAW20 also provided simulated datasets generated with a model using the GOLDN study real sample, genotype, and methylation datasets. The answer sheet provided by GAW20 described the simulation model. Post-treatment methylation levels were modelled based on pre-methylation with a higher variation at ten CpG markers than for the remaining CpG markers. Post-TG levels were influenced by five causal SNPs with decreasing heritability and several polygenes. However, the influence of each of the five causal SNPs on the TG levels (pre to post) decreased with increasing degree of methylation of nearby CpG markers to the causal SNPs, such that five out of ten CpG markers were close to the causal SNPs. Using the information from the findings of the GOLDN study and the answer sheet for the simulated data from GAW20, I defined genomic regions harbouring causal and non-causal SNPs and their nearby causal and non-causal CpGs.

I investigated both simulated and real datasets and our research aimed to address the following questions:

I. Can KMR be implemented for the sets of CpG markers that do not follow a normal distribution?
II. How does the kernel work, by increasing the size of genomic region under analysis or by incorporating more CpGs into the kernel?
III. Does a model that considers an interaction term between the pair(s) of nearby SNP and CpG markers improve the overall performance of the model? This corresponds to a genetic by epigenetic interaction term (genome-by-epigenome interactions)?

I restricted my analysis only to independent individuals in both simulated as well as real datasets. This reduced the sample size from n= 991

to n=150 individuals in the real dataset and n=111 individuals in the simulated datasets. Thus, a loss of power in the analysis was expected. A research focus of our group was on kernel methods for genetic datasets, so that SNP-sets were jointly tested for association. I also adapted the set-based approach for CpG markers. Here, two crucial aspects concerning CpG markers need attention: first, the 450K methylation array only gives information for CpGs that are located either inside or close to protein-coding genes. Secondly, the CpG density in the human genome is not uniform. Research has revealed that instead of a single CpG site/variant, few to several CpGs as a CpG-set within a genomic region exhibit a similar pattern of methylation [29, 61]. This CpG-set controls the transcriptional activity of nearby genes. Considering the non-uniform density and the set-based behaviour of CpGs, I defined genomic regions of interest of varying length. In the simulated data, the genomic regions were defined within the boundaries of lying zero kilobase pairs (kbp), 3 kbp, and 15 kbp upstream and downstream of true causal markers. In the real data, the boundaries of genomic regions were defined using the intronic boundaries of CPT1A, as reported evidence by the GOLDN study findings for association of CpG markers with the TG levels are mapped to *CPT1A*.

Overall, the analyses for both the real and simulated data indicated that the use of KMR for CpG markers is feasible. I modelled the set of CpGs in defined genomic regions via a linear kernel in the KMR. In addition, a linear model was used to validate the findings by the linear-kernel-based KMR. In particular, even though I only considered independent individuals for analyses, KMR was able to replicate the association from the original study, albeit nominally. For simulated data, no direct effect of CpG markers was modelled; the KMR approach did not yield any significant findings. My results for KMR were supported with that of linear regression analysis. Most importantly, an interaction regression model for the causal SNP with the nearest CpG marker identified an effect for the SNPs with the three greatest heritabilities simulated. The simulation model assumed full SNP

effect only for unmethylated regions, decreasing to zero in the case of full methylation. Thus, in the context of a clear candidate setting, interaction between epigenetic and genetic data may enhance information, albeit nominally, even with small sample sizes.

# 4 DISCUSSION

Many parametric and non-parametric methods have been developed and tested for the set-based association testing of genetic datasets with complex phenotypes. In this thesis, I present two set-based approaches, polygenic risk scores (PRSs) for genetic data and kernels for genetic and methylation data. PRS is an often-applied strategy that calculates a genetic risk score for a particular phenotype based on GWAS summary statistics obtained from an independent sample.

In my first research article [53], I investigated the use of GWAS results for target traits that are not identical to the discovery trait but might have a strong-to-moderate correlation. Our collaboration with psychiatrists based in Munich working on the Psycourse study [54], motivated this research. Sergi Papiol and his workgroup conducted a longitudinal study in 2019 recruiting schizophrenia patients and healthy controls, in which all study participants went through aerobic endurance training for three months [62]. Magnetic resonance imaging (MRI) scans were collected at baseline and at the end of training. PRSs were calculated using the GWAS summary statistics of schizophrenia available at the PGC [44]. A change in hippocampal volume before and after training was found to be associated with the schizophrenia PRS. Change in hippocampal volume is a psychopathological endophenotype [63]; when used as a target trait, it reflects a situation of modest to strong correlation between the target and discovery traits. In the PsyCourse study, several other psychopathological endophenotypes such as cognitive functions investigating the working memory or scores assessing psychological functioning were recorded.

In molecular psychiatric research, endophenotypes have gained quite a lot in momentum in the last few years [54]. It is speculated that endophenotypes share genetic burden in the affective-to-psychotic continuum of psychiatric disorders, and they tend to appear in both patients and their unaffected relatives [64]. Hence, endophenotypes can potentially decipher the genetic burden of disease better than the transdiagnostic groups. Significant associations have been reported for schizophrenia-based PRS analysis with P300 and the digit-span test, two commonly recorded psychopathological endophenotypes [65].

My simulation study provides an insight into using various endophenotypes where the magnitude of genetic correlation between an endophenotype and the discovery trait is different. The relevant GWAS summary statistics can be used while taking into account the sample size of the target trait under PRS analysis. The findings of my study can also be used for situations of correlated diseases, such as taking T2DM as an endophenotype/risk factor in cardiovascular disease.

Although the applications of PRS have been useful for various complex diseases, this method has many limitations. First, PRS requires appropriate and relevant GWAS summary statistics calculated on a very large sample [43]. Such large-scale GWAS results are only available for a few complex diseases. Moreover, a wide majority of these GWASs are performed on individuals with European ancestry [46]. Therefore, PRS is calculated for the target sample with the same genetic ancestry.

In the PRS studies conducted between 2008 and 2017, 67% of these studies included participants exclusively with European ancestry, 19% included participants with East Asian ancestry and only 3.8 % of studies were performed for African, Hispanic, or other indigenous ancestry [46]. Duncan et al., 2019 performed an interesting PRS analysis with admixture populations using height as discovery and target trait. In the above state of analysis, the GWAS summary statistics for height, calculated for European

individuals, were used from the UK Biobank. However, for the target trait, individuals with African ethnic ancestry were employed. A linear regression model as function of age, gender, as well as European ancestry components with and without PRS revealed that the predictive performance of European ancestry-derived polygenic scores is lower in non-European ancestry samples [46]. Thus, the underrepresentation of non-European GWASs limits the predictive power of PRSs.

More recently, Middle East Asian countries [66, 67] have taken the initiative and established data banks similar to the UK Biobank such as the Saudi Human Genome Project and Qatar Biobank for medical research. Biobanks in East Asian countries such as China (Pan-Asia Population SNP Database; Human Genetic Resources Platform), South Korea (Korean Genomic Variant Database; Korea Biobank Project), Japan (Japanese Genotype-Phenotype Archive; Biobank Japan), and Taiwan (Taiwan Biobank) were established earlier. In the South East Asian countries such as Pakistan and India, no biobanks have been established yet; however, numerous GWASs targeting phenotypes such as T2DM and lung cancer have been carried out there. In PRS analysis, the model is adjusted for population stratification; principal component analysis (PCA) and multidimensional scaling (MDS) are frequently used. However, there is need for development of methods that allow combining data from admixed populations. In addition, there is the need to conduct well-powered studies in non-white populations.

PRS facilitates the evaluation of relative risk in individuals. Hence, it cannot be used to infer the absolute risk of genetic predisposition for a particular disease in a particular individual. Numerous PRS analyses have demonstrated that adding known non-genetic risk factors along with the PRS in the regression model has resulted in an increase in the explained variance of the target trait, thus improving the risk stratification in individuals in the target sample. The performance of risk models can be evaluated with the ROC (receiver operating characteristics) curve and the

area under the curve (AUC) [68]. The clinical utility of models with an AUC < 0.65 is generally not very great. Models with an AUC > 0.8 are generally informative for most patients, and can be used for clearer stratification of study participants into groups of high, intermediate, and low risk.

In the second article, I implemented a kernel method, which allows joint testing of a set of markers with a phenoytpe. Kernel methods have been well exploited for genetic datasets based on SNPs. Previous research work done by our group has demonstrated the use of kernels for dependent and independent individuals in a dataset, as well as developing a kernel that allows the incorporation of pathways with interacting genes. For my research work, I used the linear kernel for methylation datasets (set of CpG markers). CpG markers have often been employed as response variable and associated with gene expression datasets. In my work, I considered a reverse regression model that associates the phenotype with the CpG markers. For the regression model in which CpG markers are employed as response variable, a logit transformation is applied to normalise the CpG values, as these are beta-distributed. However, a CpG value that is continuous from 0 to 1 yields infinity and zero for 0s and 1s in the data after logit transformation. In order to overcome this problem, the original data interval of values is reduced such that 0s and 1s are replaced. This results in a severe loss of information. The untransformed value is referred to as B-value. After logit transformation, these are called M-values. In my second article, I used B-values with a limited sample size of n<200 individuals in a linear kernel.

Subsequently, I performed simulations (results not published) to compare the performance of the linear kernel by using B-values versus M-values. Using a copula-based approach through beta distribution, I generated methylation datasets for a quantitative phenotype *y*. I considered various scenarios based on correlation structure, i.e., from no correlation to 0.10, and 0.25 or block-wise correlation among the simulated CpG markers.

In all these scenarios, I found that using the logit transformed M-values resulted in a loss of power, albeit not significant.

Furthermore (results not published), I considered using an additive kernel that combined genotype and methylation datasets in one kernel, and testing the association with the outcome $y$. The simulation scenarios were the same as stated above, with varying numbers of SNPs and CpGs in each dataset. First, simulation results testing the association with the outcome revealed that the power remained approximately the same for an additive kernel for a SNP set, regardless of whether or not the CpG set was added into the kernel. These results are in good agreement with Zhao et al., 2018 [69]. In these simulations however, I simulated CpG markers that were a mixture from the three distributional shapes for methylation, i.e. hypo/semi/hyper-methylated, and this, along with the SNP set, did not really improve power. Thus, I subsequently simulated CpG markers only from one distributional pattern; the performance of the additive kernel varied then. More simulations are required to test the power of an additive kernel integrating a methylation-state-specific CpG set with a SNP set into one kernel.

The research work completed in preparation of this thesis contributes to the use of the PRS technique for correlated target traits with careful consideration of the sample size, in particular for clinical studies. In addition, the research into implementing kernels for methylation datasets hints towards exploiting the opportunity of investigating other -omics datasets by deploying other kernels. Given the existing understanding of the genetic architecture of complex diseases, an additive model of heritability is used. The importance of interactions is unclear and the additive model seems to capture more variance than the interactions when simple models for the interaction are employed. Kernel methods seem a promising strategy for multi-omics data integration. Moreover, the integration of multi-omics data can be more profitable by taking existing

knowledge such as that available on the ENCODE database or the Gene Expression Omnibus (GEO) into account.

# 5 REFERENCES

[1] R. Pranavchand and B. M. Reddy, "Genomics era and complex disorders: Implications of GWAS with special reference to coronary artery disease, type 2 diabetes mellitus, and cancers," *Journal of Postgradaute Medicine,* vol. 62, no. 3, pp. 188-198, 2016.

[2] E. B. Bookman, K. McAllister, E. Gillanders , K. Wanke, D. Balshaw , J. Rutter, J. Reedy , D. Shaughnessy , T. Agurs-Collins, D. Paltoo, A. Atienza , L. Bierut , M. D. Fallin , F. Perera , E. Turkheimer , J. Boardman, M. L. Marazita , S. M. Rappaport , E. Boerwinkle , . S. J. Suomi , N. E. Caporas, I. Hertz-Picciotto , K. C. Jacobson , W. L. Lowe, L. R. Goldman , P. Duggal , M. R. Gunnar , T. A. Manolio , E. D. Green, D. H. Olster and L. S. Birnbaum, "Gene-environment interplay in common complex diseases: forging an integrative model—recommendations from an NIH workshop," *Genetic Epidemiology,* vol. 35, no. 4, pp. 217-225, 2011.

[3] M. Darrason, "Unifying diseases from a genetic point of view: the example of the genetic theory of infectious diseases," *Theoretical Medicine and Bioethics,* vol. 34, no. 4, p. 327–344, 2013.

[4] B. H. Y. Chung, J. F. T. Chau and G. K.-S. Won, "Rare versus common diseases: a false dichotomy in precision medicine," *Genomic Medicine,* vol. 6, no. 19, 2021.

[5] A. Pal and M. McCarthy, "The genetics of type 2 diabetes and its clinical relevance," *Clinical Genetics,* vol. 83, no. 4, pp. 297-306, 2013.

[6] G. M. Cooper and J. Shendure , "Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data," *Nature Reviews Genetics,* vol. 12, no. 9, p. 628–640, 2011.

[7] Online Inheritance in Man (OMIM), "Online Inheritance in Man," 1960. [Online]. Available: https://www.omim.org/statistics/geneMap. [Accessed 12 May 2021].

[8] S. Shiovitz and L. A. Korde, "Genetics of breast cancer: a topic in evolution," *Annals of oncology : official journal of the European Society for Medical Oncology,* vol. 26, no. 7, pp. 1291-1299, 2015.

[9] M. T. Malecki, "Genetics of type 2 diabetes mellitus," *Diabetes Research and Clinical Practice,* vol. 68, no. 1, pp. 10-21, 2005.

[10] H. Gao, L. Li, S. Rao, G. Shen, Q. Xi, S. Chen, Z. Zhang, K. Wang, S. G. Ellisi, Q. Chen, E. J. Topol and Q. K. Wang, "Genome-Wide Linkage Scan Identifies Two Novel Genetic Loci for Coronary Artery Disease: In GeneQuest Families," *PLoSONE,* vol. 9, no. 12, 2014.

[11] Y. Guo, F. Wang, L. Li, H. Gao, . S. Arckacki, I. Z. Wang, J. Barnard, S. Elis, C. Hubbard, . E. J. Topol, Q. Chen and Q. K. Wang, "Genome-Wide Linkage Analysis of Large Multiple Multigenerational Families Identifies Novel Genetic Loci for Coronary Artery Disease," *Scientific Reports,* vol. 7, no. 1, 2017.

[12] G. Lett and P. L. Auer, "Rare variant association studies: considerations, challenges and opportunities," *Genome medicine,* vol. 7, no. 1, 2015.

[13] R. M. Cantor, K. Lange and J. S. Sinsheimer, "Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application," *The American Journal of Human Genetics,* vol. 1, no. 86, pp. 6-22, 2010.

[14] C. S. Ku, E. Y. Loy, Y. Pawitan and K. S. Chia, "The pursuit of genome-wide association studies: where are we now?," *Journal of Human Genetics,* vol. 55, pp. 195-206, 2010.

[15] J. K. Pritchard and M. Przeworski, "Linkage Disequilibrium in Humans: Models and Data," *American Journal of Human Genetics,* vol. 69, no. 1, pp. 1-14, 2001.

[16] T. LaFramboise, "Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances," *Nucleic Acids Research,* vol. 37, no. 4, p. 4181–4193, 2009.

[17] O. A. Panagiotou, C. J. Wille, J. N. Hirschhorn and J. P. A. Ioannidis, "The Power of Meta-Analysis in Genome Wide Association Studies," *Annual Review of Genomics and Human Genetics,* vol. 65, pp. 441- 465, 2013.

[18] A. Xue, Y. Wu, Z. Zhu, . F. Zhang, K. E. Kemper, Z. Zheng, L. Yengo, L. R. Lloyd-Jones, J. Sidorenko, Y. Wu, A. F. McRae, P. M. Visscher, J. Zeng, J. Yang and eQTLGen Consortium, "Genome-wide association

analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes," *Nature Communications,* vol. 9, 2018.

[19] V. Tam, N. Patel, M. Turcotte, Y. Bossé, G. Paré and D. Meyre, "Benefits and limitations of genome-wide association studies," *Nature Reviews Genetics,* vol. 20, pp. 467- 484, 2019.

[20] M. Civelek and A. J. Lusis, "Systems genetics approaches to understand complex traits," *Nature Reviews Genetics,* vol. 15, pp. 34 - 48, 2014.

[21] A. E. Handel, G. C. Ebers and S. V. Ramagopalan, "Epigenetics: molecular mechanisms and implications for disease," *Trends in Molecular Medicine,* vol. 16, no. 1, pp. 7-16, 2010.

[22] D. J. Schaid, C. M. Rowland, D. E. Tines, R. M. Jacobso and G. A. Poland, "Score tests for association between traits and haplotypes when linkage phase is ambiguous," *American Journal of Human Genetics,* vol. 70, no. 2, pp. 425-434, 2002.

[23] D. J. Schaid, S. K. McDonnell, S. J. Hebbring, J. M. Cunningham and S. N. Thibodeau, "Nonparametric Tests of Association of Multiple Genes with Human Disease," *The American Journal of Human Genetics,* vol. 76, no. 5, pp. 780-793, 2005.

[24] GWAS catalog, "GWAS catalog," 04 April 2021. [Online]. Available: https://www.ebi.ac.uk/gwas/docs/about. [Accessed 30 April 2021].

[25] Y. G. Tak and P. J. Farnham, "Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome," *Epigenetics & Chromatin,* vol. 57, no. 8, 2015.

[26] E. R. Gibney and C. M. Nolan, "Epigenetics and gene expression," *Heredity,* vol. 105, pp. 4-13, 2010.

[27] L. D. Moore, T. Le and G. Fan, "DNA Methylation and Its Basic Function," *Neuropsychopharmacology,* vol. 38, no. 1, pp. 23-38, 2013.

[28] C. Luo, P. Hajkova and J. R. Ecker, "Dynamic DNA methylation: in the right place at the right time," *Science,* vol. 361, no. 6409, pp. 1336-1340, 2019.

[29] The ENCODE Project Consortium, "Expanded encyclopaedias of DNA elements in the human and mouse genomes," *Nature,* vol. 583, pp. 699-710, 2020.

[30] L. H. Chadwick, A. Sawa, I. V. Yang, A. Baccarelli, X. O. Breakefield, H.-W. Deng, D. C. Dolinoy, M. . D. Fallin, N. T. Holland, E. A. Houseman, S. Lomvardas, M. Rao, J. S. Satterlee, F. L. Tyson, P. Vijayanand and J. M. Greally, "New insights and updated guidelines for epigenome-wide association studies," *Neuroepigenetics,* vol. 1, pp. 14-19, 2015.

[31] E. Raineri, M. Dabad and S. Heath, "A Note on Exact Differences between Beta Distributions in Genomic (Methylation) Studies," *PlosONE,* 2014.

[32] L. Weinhold, S. Wahl, S. Pechlivanis, P. Hoffman and M. Schmid, "A statistical model for the analysis of beta values in DNA methylation studies," *BMC Bioinformatics,* vol. 17, 2016.

[33] J. M. Flanagan, "Epigenome-wide association studies (EWAS): past, present, and future," *Methods in Molecular Biology,* pp. 51-63, 2015.

[34] P. Zeng, Y. Zhao, C. Qian, L. Zhang, R. Zhang, J. Gou, J. Liu, L. Liu and F. Chen, "Statistical analysis for genome-wide association study," *The Journal of Biomedical Research,* vol. 29, no. 4, pp. 285-297, 2015.

[35] A. T. Marees, H. d. d. Kluiver, S. Stringer, F. Vorspan, E. Curis, C. Marie-Claire and E. M. Derks, "A tutorial on conducting genome-wide association studies: Quality control and statistical analysis," *International Journal of Methods in Psychological Research,* vol. 27, no. 2, p. e1608, 2018.

[36] M. T. Dorak, "Genome Biology for Genetic Epidemiologists," 07 April 2018. [Online]. Available: http://www.dorak.info/epi/glosge.html. [Accessed 04 January 2021].

[37] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *PNAS,* vol. 100, no. 16, pp. 9440-9445, 2003.

[38] J. Fadista, A. K. Manning, J. C. Florez and L. Groop , "The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants," *European Journal of Human Genetics,* vol. 24, pp. 1202-1205, 2016.

[39] F. Dudbridge, "Power and Predictive Accuracy of Polygenic Risk Scores," *PLoS Genetics,* vol. 9, no. 3, p. e1003348, 2013.

[40] S. W. Choi, T. S.-H. Mak and P. F. O'Reilly, "Tutorial: a guide to performing polygenic risk score analyses," *Nature Protocols,* vol. 15, pp. 2759-2722, 2020.

[41] Psychiatric GWAS Consortium Steering Committee, "A framework for interpreting genome-wide association studies of psychiatric disorders," *Molecular Psychiatry,* vol. 14, no. 1, 2009.

[42] L. Duncan, H. Shen, B. Gelaye, J. Meijsen, K. Ressler, M. Feldman, R. Peterson and B. Domingue, "Analysis of polygenic risk score usage and performance in diverse human populations," *Nature Communications,* vol. 10, 2019.

[43] S. Gunn, "How to: perform polygenic risk score analysis," 08 August 2020. [Online]. Available: https://frontlinegenomics.com/how-to-perform-polygenic-risk-score-analysis/. [Accessed 11 January 2021].

[44] Schizophrenia Working Group of the Psychiatric Genomics Consortium, "Biological insights from 108 schizophrenia-associated genetic loci," *Nature,* vol. 511, pp. 421-427, 2014.

[45] N. B. Larson, J. Chen and D. J. Schaid, "A Review of Kernel Methods for Genetic Association Studies," *Genetic Epidemiology,* vol. 43, no. 2, pp. 122-136, 2019.

[46] T. Hofmann, B. Schölkopf and A. J. Smola, "Kernel methods in machine learning," *The Annals of Statistics,* vol. 36, no. 3, pp. 1171-1220, 2005.

[47] J. H. Manton, "A Primer on Reproducing Kernel Hilbert Spaces," *Foundations and Trends® in Signal Processing,* vol. 8, no. 1-2, pp. 1-126, 2015.

[48] D. J. Schaid, "Genomic similarity and kernel methods II: methods for genomic information," *Human Heridity,* vol. 70, no. 2, pp. 132-140, 2010.

[49] K. Gohrani, "Different Types of Distance Metrics used in Machine Learning," 10 November 2019. [Online]. Available: https://medium.com/@kunal_gohrani/different-types-of-distance-metrics-used-in-machine-learning-e9928c5e26c7. [Accessed 14 March 2021].

[50] D. J. Schaid, "Genomic similarity and kernel methods I: advancements by building on mathematical and statistical foundations," *Human Heridity,* vol. 70, no. 1, pp. 109-131, 2010.

[51] S. Yasmeen, S. Papiol, P. Falkai, T. G. Schulze and H. Bickeböller, "Polygenic Risk for Schizophrenia and Global Assessment of Functioning—A Comparison with In-Silico Data," *Journal of Psychiatry and Brain Science,* vol. 4, 2019.

[52] M. Budde, H. A. Schmidt, K. Gade, D. . R. Erkelenz, K. Adorjan, J. . L. Kalman, F. Senner, . S. Papiol, T. . F. M. Andlauer , A. L. Comes, E. C. Schulte , P. Falkai, T. G. Schulze and U. Heilbronner , "A longitudinal approach to biological psychiatric research: The PsyCourse study," *American Journal of Medical Genetics Part B Neuropsychiatric Genetics,* vol. 180, no. 2, pp. 89-102, 2019.

[53] B. Bulik-Sullivan, H. K. Finucane, V. Anttila, A. Gusev, F. R. Day, P.-R. Loh, ReproGen Consortium and Psychiatric Genomics Consortium, "An atlas of genetic correlations across human diseases and traits," *Nature Genetics,* vol. 47, p. 1236–1241, 2015.

[54] A. Reichenberg, P. D. Harvey, C. R. Bowie, R. Mojtabai, J. Rabinowitz, . R. K. Heaton and . E. Bromet, "Neuropsychological Function and Dysfunction in Schizophrenia and Psychotic Affective Disorders," *Schizophrenia Bulletin,* vol. 35, no. 5, p. 1022–1029, 2009.

[55] S. Yasmeen, P. Burger, S. Friedrichs, S. Papiol and H. Bickeböller, "Relating drug response to epigenetic and genetic markers using a region-based kernel score test," *BMC Proceedings,* vol. 12, 2018.

[56] M. R. Irvin, D. Zhi, R. Joehanes, M. Mendelson, S. Aslibekyan, S. A. Claas, K. S. Thibeault, N. Patel, K. Day, L. W. Jones, L. Liang, B. H. Chen, C. Yao, H. K. Tiwari, J. M. Ordovas, D. Levy, D. Absher and D.

K. Arnett, "Epigenome-Wide Association Study of Fasting Blood Lipids in the Genetics of Lipid Lowering Drugs and Diet Network Study," *Circulation,* vol. 130, no. 7, pp. 565-572, 2014.

[57] S. Papiol, D. Keeser, A. Hasan, T. Schneider-Axmann, F. Raabe, F. Degenhardt, M. J. Rossner, H. Bickeböller, L. Cantuti-Castelvetri, M. Simons, T. Wobrock, A. Schmitt, B. Malchow and P. Falkai, "Polygenic burden associated to oligodendrocyte precursor cells and radial glia influences the hippocampal volume changes induced by aerobic exercise in schizophrenia patients," *Translational Psychiatry,* vol. 9, no. 284, 2019.

[58] A. C. Ruocco , S. Amirthavasagam and K. K. Zakzanis, "Amygdala and hippocampal volume reductions as candidate endophenotypes for borderline personality disorder: a meta-analysis of magnetic resonance imaging studies," *Psychiatry Research,* vol. 201, no. 3, pp. 245-252, 2012.

[59] W. G. Iacono, "Endophenotypes in psychiatric disease: prospects and challenges," *Genome Medicine,* vol. 10, no. 11, 2018.

[60] S. Ranlund, S. Calafato, J. H. Thygesen, K. Lin, W. Cahn, B. Crespo-Facorro, S. M. de Zwarte, Á. Díez, M. D. Forti, GROUP*, C. Iyegbe, A. Jablensky, R. Jones, M.-H. Hall, R. Kahn, L. Kalaydjieva, E. Kravariti, C. McDonald, A. M. McIntosh, A. McQuillin, PEIC, M. Picchion, D. P. Prata, D. Rujescu, K. Schulze, M. Shaikh, T. Toulopoulou, N. v. Haren, J. v. Os, E. Vassos, M. Walshe, WTCCC2, C. Lewis, R. M. Murray, J. Powell and E. Bramon, "A polygenic risk score analysis of psychosis endophenotypes across brain functional, structural, and cognitive domains," *American Journal of Medical Genetics,* vol. 177, no. 1, p. 21–34, 2017.

[61] S. Stuart G Baker, "Metrics for Evaluating Polygenic Risk Scores," *JNCI Cancer Spectrum,* vol. 5, no. 1, 2021.

[62] N. Zhao, X. Zhan, Y.-T. Huang, L. M. Almli, A. Smith, M. P. Epstein, K. Conneely and M. C. Wu, "Kernel machine methods for integrative analysis of genome-wide methylation and genotyping studie," *Genetic Epidemiology,* vol. 42, no. 2, pp. 156-167, 2018.

[63] T. Cai, "Kernel Machine Approach to Testing the Significance of Multiple Genetic Markers for Risk Prediction," *Biometrics,* vol. 67, no. 3, p. 975–986, 2011.

# 6 APPENDIX

## I. References, Web-Links and Digital Object Identifiers of Original Articles

**S Yasmeen, P Burger, S Friedrichs, S Papiol, and H Bickeböller.**

Relating drug response to epigenetic and genetic markers using a region-based kernel score test.
BMC proceedings 2018; 12 (9), 47.
URL: https://pubmed.ncbi.nlm.nih.gov/30275895/
DOI: 10.1186/s12919-018-0154-5

**S Yasmeen, S Papiol, P Falkai, TG Schulze, and H Bickeböller.**

Polygenic Risk for Schizophrenia and Global Assessment of Functioning - A Comparison with In-Silico Data.
Journal of Psychiatry and Brain Science, 2019; 4: e190003.
URL: https://doi.org/10.20900/jpbs.20190003
DOI: 10.20900/jpbs.20190003

***PDF copies of both articles are compiled with the thesis.***