DNA Methylation and Exome Chip Analysis in Complex Disease

Christian Magnus Page

Dissertation for the degree of Philosophiae Doctor

Department of Neurology
Institute of Clinical Medicine
Faculty of Medicine
Univeristy of Oslo
Norway

Series of dissertations submitted to the
Faculty of Medicine, University of Oslo

ISBN 978-82-8333-300-8

Cover: Hanne Baadsgaard Utigard.
Print production: Reprosentralen, University of Oslo.

Acknowledgements

First and foremost, I want to thank my supervisors, especially my main supervisor Bettina Andreassen for excellent guidance, motivation and support, and for always being there whenever I needed it. I would also thank my co-supervisor, Hanne Harbo for her attention to details, for always motivating me to do better, and for her inspiring enthusiasm for research.

I also want to thank my fellow colleagues with whom I shared an office, namely Mathias Toft and Steffan Bos-Haugen for many fruitful discussions, and especially Steffan for guiding me through the process of scientific writing. I would also thank all my colleagues in the MS group, including Hanne, Elisabeth, Tone, Steffan, Gro, Piotr, Pankaj, Pål, Marte, Ingvild, Sigrid, Stine Marit, Anna, Ina, Einar, Cecilia, and Heidi. Additionally, I would also like thank the other groups at the neurology department as well as the other groups at Domus Medica IV, for providing such a good working environment and ensuring that no two days were alike.

Furthermore, I would also like to thank my colleagues at the Oslo Center for Biostatistics and Epidemiology for many great seminars and inputs, especially Marissa LeBlanc and Arnoldo Frigessi for many good suggestions and discussions.

I am very grateful to University of Oslo for funding my doctoral research, and to the neurology departments at both Oslo University Hospital and the University of Oslo for hosting me during my time as doctoral candidate.

In the first year of my research I had the opportunity to stay in San Francisco for three months, at both Lisa Barcellos' lab at University of California, Berkeley, and at Sergio Baranzini's lab at University of California, San Francisco. I am most grateful for the hospitality I received at both labs, as well as the fruitful collaborations that continued afterwards.

I would also like to thank all of my co-authors for all the fruitful discussions we had and the contributions that they all made. Many others have also contributed during my time as a doctoral candidate and I would especially like to thank Benedicte, Jon, Ellen, Linda, Amine, and Bendik for their valuable contributions and for motivating me during my research.

Last but not least, I would like to give warm thanks to my family for all the support and motivation that I have received over the last years.

Oslo, June 2016

List of Papers

Paper I

Bos SD*, Page CM*, Andreassen BK, Elboudwarej E, Gustavsen MW, Briggs F, Quach H, Leikfoss IS, Bjølgerud A, Berge T, Harbo HF, Barcellos LF (2015) “Genome-Wide DNA Methylation Profiles Indicate CD8$^+$ T Cell Hypermethylation in Multiple Sclerosis.” *PLoS ONE* 10(3): e0117403. DOI:10.1371/journal.pone.0117403

Paper II

Page CM, Baranzini SE, Mevik BH, Bos SD, Harbo HF, Andreassen BK (2015) “Assessing the Power of Exome Chips.” *PLoS ONE* 10(10): e0139642. DOI:10.1371/journal.pone.0139642

Paper III

Page CM*, Vos L*, Rounge TB, Harbo HF, Andreassen BK, “Assessing genome-wide significance for the detection of differentially methylated regions.” Submitted to *Bioinformatics*

Table of Contents

Abbreviations

450k array	Illumina Infinium HumanMethylation450 BeadChip array
ACF	Autocorrelation function
ANOVA	Analysis of variance
AR	Autoregressive time series model
BMI	Body mass index
BMIQ	Beta-mixture quantile normalization
bp	Base pairs
CD	Cluster of differentiation
CDCV	Common disease – common variants
CDRV	Common disease – rare variants
ChIP-seq	Chromatin immunoprecipitation sequencing
CNS	Central nervous system
CpG	CG di-nucleotide
CPU	Central processing unit
DMP	Differentially methylated position
DMR	Differentially methylated region
DNA	Deoxyribonucleic acid
EDSS	Expanded disability status scale
EWAS	Epigenome-wide association study
FDR	False discovery rate
Fin-HIT	Finnish Health in Teens
FWER	Family-wise error rate
GWAS	Genome-wide association study
HLA	Human leukocyte antigen
IMSGC	International Multiple Sclerosis Genetics Consortium
LD	Linkage disequilibrium
MCMC	Markov chain Monte Carlo
mRNA	Messenger ribonucleic acid
MS	Multiple sclerosis
MSSS	Multiple sclerosis severity score
NK	Natural killer cells
PACF	Partial autocorrelation function
PC	Principal component
QQ-plot	Quantile-quantile plot
RNA	Ribonucleic acid
SKAT	Sequence kernel association test
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
WSS	Weighted sum statistics
WTCCC	Wellcome Trust Case Control Consortium

I PREFACE

Recent advances in biotechnology have led to an explosion in the amount of biological data available to researchers. Since the introduction of high-throughput technologies, a massive number of genetic markers can now be investigated for large numbers of study participants. This has led to the discovery of thousands of genetic markers that are associated with various human traits and diseases. Technological advances have made it possible to investigate not only diseases that are caused by alteration of a single gene, but also to explore the whole genome simultaneously. Most diseases are not only caused by a single genetic mutation, but by many genetic variants contributing to the disease risk, either on their own or in interaction with other variants or other environmental factors. In complex diseases, both genetic variants and environmental factors contribute to disease susceptibility, and identifying the underlying genetic risk variants for these diseases has been a major challenge in genomics.

Statistics is a tool for data analysis that has played an important role in the breakthroughs in genetic studies. Statistics have shaped experimental design by addressing issues such as false positive control, sample sizes requirements, and population heterogeneity. It has also played a prominent role in the development of quality-control protocols and different normalizations methods. New types of genetic data require development of both new methodologies as well as adaptations of existing methods, and this has led to the integration of statistical methodology into almost all aspects of genetic analyses.

The title of this thesis, DNA Methylation and Exome Chip Analysis in Complex Disease, is a broad title encompassing many aspects of both genetic and epigenetic epidemiology. Common themes in this thesis are methods for the identification of genetic biomarkers in complex diseases and the aggregation of genetic information across several genetic sites. Two papers involve DNA methylation data and one paper assesses the constraints in genetic studies involving low-frequency and rare variants using the Exome Chip.

The structure of the thesis is as follows. An introduction and background to the study of genetics and epigenetics of complex diseases is presented in Chapter II. Chapter III lists the aims of this thesis, and Chapter IV presents the materials and methods that were applied and developed. Chapter V outlines the results from the three papers. In Chapter VI, the methods and results are discussed, and the thesis finishes with suggestions for future extensions and a concluding remark.

II Introduction

II.1. Genetic variation in human health and disease

Variations in deoxyribonucleic acid (DNA) can explain a substantial part of the differences between human populations (Gibbs *et al.*, 2003; International HapMap Consortium, 2005; WTCCC, 2007). The simplest form of DNA variation is a substitution, insertion, or deletion of a singe base pair. This is called a single nucleotide variant (SNV). When such a variant is found in more than 1% of the population, it is commonly referred to as a single nucleotide polymorphism (SNP) (Brookes, 1999). Variants that have an frequency below 1% are usually referred to as rare variants (Bansal *et al.*, 2010), variants with a frequency range between 1% and 5% are often referred to as low frequent variant, and variants with frequency above 5% are referred to as common variants. Genetic variation can also result from structural variations that span sections of DNA, such as deletions, duplications, or insertions of segments or genes, or even duplications of whole chromosomes. All of these variations in DNA play an important role in human health and diseases (Stankiewicz & Lupski, 2010).

Mendel's second law of inheritance states that all alleles are passed independently from parents to their offspring. However, when variable loci are close to each other on a chromosome, the alleles tend to be inherited together (Ott, 1999). This deviation from Mendel's second law is called cosegregation. When two alleles cosegregate, they are said to be in linkage disequilibrium (LD) (Slatkin, 2008).

Tracing marker alleles that segregate with a disease is called linkage analysis and has been a successful method for identifying high penetrant variants involved in monogenetic diseases. However, most common diseases are not caused by a single genetic variant, but by a combination of multiple risk variants acting together with environmental components to modulate disease risk. When the number of causal variants in a disease increases or when the penetrance of the risk variants is low, linkage analysis is less feasible. In their landmark paper published in 1996, Risch and Merikangas show that in traits caused by multiple alleles with moderate frequency and small genotype relative risk, screening many genetic markers in an association study would be much more efficient than performing linkage analysis (Risch & Merikangas, 1996). However, it took another decade of development of the high-throughput genotyping technology before these analyses became feasible and the first genome-wide association studies (GWAS) were published (Hirschhorn & Daly, 2005).

II.2. Genome-wide association studies

The Human Genome Project and the HapMap project consortia were the first to publish complete drafts of the human genome for different populations (International Human Genome Sequencing Consortium, 2004; International HapMap Consortium, 2005). These drafts were used to design the first GWAS, targeting highly polymorphic marker alleles throughout the genome (Kennedy *et al.*, 2003; Matsuzaki *et al.*, 2004; WTCCC, 2007). The genetic variants included in the GWAS were selected primarily to capture a substantial degree of variation within the human genome, not because they were believed to be disease causing (Visscher *et al.*, 2012).

For many common traits and diseases, GWAS have been successful in identifying many variants influencing disease susceptibility (Hindorff *et al.*; Klein *et al.*, 2010). As mentioned above, the risk variants identified through GWAS may not represent the underlying causal variant, but instead associated variants in high LD with the true disease causing variant(s). Dense mapping or sequencing of these regions can identify the causal variant within the associated region (Faye *et al.*, 2013). Many of the significantly associated variants from GWAS turned out not to be located within the coding part of the genome, but to be situated in regulatory regions (Dunham *et al.*, 2012).

In complex diseases, heritability estimates from epidemiological studies have always exceeded the heritability estimates based on the findings from GWAS. The discrepancy between these two heritability estimates is referred to as "the missing heritability" (Maher, 2008). Several explanations have been proposed for this missing heritability, including that the remaining part of the heritability can be explained by epistasis, rare variants, and epigenetic and environmental factors, all of which contribute to disease risk (Maher, 2008; Manolio *et al.*, 2009; Zuk *et al.*, 2012). For this thesis, I have studied two of these suggestions more closely, namely the epigenetic and rare variant contribution to disease risk in complex diseases. Investigation and assessment of different factors in disease risk are common themes in all three papers included in this thesis.

GWAS targets mainly common variants, but it has long been known that most genetic variants are rare (Lander, 1996). In the final phase of the 1000 Genomes Project, it was noted that approximately 64 million of the 88 million discovered variants (~70%) had an allelic frequency below 0.5% (The 1000 Genomes Project Consortium *et al.*, 2015). However, it was first with the introduction of high throughput sequencing that rare variants could be genotyped in larger cohorts (Rabbani *et al.*, 2014). In order to investigate the contribution of rare variants in complex diseases on a population level, the Exome Chip was established. This chip was designed based on suggestions from the Exome Chip Consortium (Exome Chip Consortium, 2011). The variants on these chips are mainly rare variants affecting the amino acid sequence, with almost 85% of the variants having a minor allele frequency less than 0.5%. This low frequency range poses challenges that differ from those of common variants analysis. In **Paper II**, we investigate some of these challenges in rare variant association tests. In the study on which the paper is based, we paid attention to the sample size requirements for varying effect sizes and different causal variant scenarios based on two widely-used statistical approaches for identifying causal rare and low-frequency variants.

The investigation of the genetic contribution to disease risk in complex diseases has sparked a debate on the underlying genetic variation in complex diseases. The two main models suggested were the "common disease – common variants" (CDCV) and "common disease – rare variants" (CDRV) hypothesis (Gibson, 2011). GWAS have traditionally focused on the investigation of common variants (Andersson *et al.*, 2009). Hence, the design of GWAS was based on the hypothesis of "common disease – common variant" (CDCV). The CDCV hypothesis builds on the assumption that common diseases are caused by many common variants, each having a small effect on the phenotype (Fisher, 1930). One argument for this hypothesis is that highly penetrant deleterious rare variants are likely to be selected from a population through natural selection. This model is sometimes referred to as the infinitesimal model, since it models a scenario in which a large number of common variants each confer a small individual risk (Gibson, 2011).

An argument against the infinitesimal model is the lack of functional validation for the majority of the detected risk variants from GWAS (McClellan & King, 2010). This stands in contrast to many rare Mendelian traits where a rare causal functional SNV has been identified. Currently, the GWAS catalogue lists over

4,000 common risk variants for complex diseases, but few of those have been shown to exhibit functional properties that can be linked to disease risk (Huang, 2015). However, the variants selected to be interrogated in GWAS were not selected for their functional implications, but for their ability to tag variation in the genome (Weiss & Clark, 2002).

While large GWAS involving common variants have been successful in identifying risk loci, they have only explained a small fraction of the predicted heritability. The other proposed model for the underlying genetic risk of complex diseases is the "Common Disease – Rare Variants" hypothesis (CDRV). The assumption underlying this hypothesis is that under natural selection, deleterious or disease-causing variants will not segregate in the population, and these should therefore be rare (Pritchard & Cox, 2002; Gibson, 2011). This is in line with the known high impact disease-causing variants in Mendelian diseases, which tend to be rare (Altshuler *et al.*, 2008). However, investigations into rare variant contribution to disease risk in complex diseases have not helped to explain the missing heritability (Hunt *et al.*, 2013).

It has been suggested that some of the associations discovered in GWAS are a result of "synthetic association." This occurs when a rare causal variant is in weak LD with a common tag SNP (Dickson *et al.*, 2010). If this is the case, then what appears to be a common risk variant is in fact a high impact rare variant co-occurring with the genetic marker. However, the absence of any observed linkage indicates that rare causal variants with high impact are not likely for most complex diseases (Risch & Merikangas, 1996; Wray *et al.*, 2011). In most complex diseases, very few high impact rare causal variants have been identified through linkage analysis (Altshuler *et al.*, 2008). This indicates that a small number of highly penetrant variants are rarely the causative factor for the disease. For this reason, synthetic association alone cannot explain all the GWAS results (Gibson, 2011; Wray *et al.*, 2011).

The two different hypotheses of CDCV and CDRV are not necessarily mutually exclusive. It is likely that both rare and common variants contribute to disease risk for most complex diseases (Schork *et al.*, 2009; Gibson, 2011; Agarwala *et al.*, 2013; Lettre, 2014).

II.3. Epigenetics and DNA methylation

The term epigenetics was first used by Waddington in 1942, as a link between genetics and the phenotype (Waddington, 1942; Haig, 2004). Today, epigenetics is understood as chemical modifications of DNA that do not alter the base pair sequence, and are heritable over mitotic cell divisions (Russo *et al.*, 1996; Berger *et al.*, 2009).

Many studies have shown that the epigenome responds to environmental conditions, both to internal and external stimuli. This includes many factors such as smoking, diet, psychological stress, and socioeconomic status (Heijmans *et al.*, 2008; Hackman *et al.*, 2010; Feil & Fraga, 2012; Zeilinger *et al.*, 2013; Klengel *et al.*, 2014). Epigenetics could potentially serve as a link between environmental and genetic factors, and may also explain some of the missing heritability (Slatkin, 2009; Petronis, 2010; Feil & Fraga, 2012). This hypothesis together with recent advances in biotechnology has created a surge of studies investigating the epigenetic contribution to the risk of complex diseases (Dunham *et al.*, 2012; Paul & Beck, 2014).

The most frequently studied epigenetic modifier with respect to disease development is DNA methylation (Rakyan *et al.*, 2011). This is a modification of the DNA where a methyl group is added to the Cytosine base (C). This usually occurs on sites where the Cytosine base is adjacent to a Guanine base (G). This CG di-nucleotide is called a CpG site (Jones, 2012). The CpG sites are depleted throughout the genome, with more than a two-fold reduction in what would be expected if both C and G were distributed uniformly across the genome (Bell *et al.*, 2012). The CpGs tend to cluster together, appearing in regions highly enriched for CpGs, termed CpG islands (Jones, 2012). These islands co-localize with 60–70% of the known gene promoters, suggesting a highly important role in transcription regulation (Bell *et al.*, 2012). DNA methylation is implicated in DNA regulation, including regulation of messenger ribonucleic acid (mRNA) transcription, alternative splicing, ribonucleic acid (RNA) elongation, X chromosome inactivation, genomic imprinting, and cell linage proliferation (Jones, 2012). The DNA methylation can also be oxidized into hydroxymethylation (Kriaucionis & Heintz, 2009). However, most technologies cannot distinguish between methylation and hydroxymethylation, and report both in a single methylation measure (Huang *et al.*, 2010).

Since DNA methylation is a stationary epigenetic modification of the DNA, occurring only at specific loci, it can be unambiguously measured on a genome-wide scale. Studies analyzing DNA methylation genome-wide are referred to as epigenome-wide association studies (EWAS) and can range from including 27,000 loci (Illumina 27k array) to well over 5 million loci (whole genome bisulfite sequencing).

With the introduction of genome-wide methylation chips and bisulfite sequencing methodology, many EWAS of complex diseases have been published. These cover a wide range of complex diseases, such as cancer, obesity, diabetes, and autoimmune diseases (Brooks *et al.*, 2010; Meda *et al.*, 2011; Liu *et al.*, 2013; Lu, 2013; Farh *et al.*, 2015). The most robust findings have been in cancer research, where many studies have shown massive, genome-wide changes in DNA methylation (Jones, 2012). For most other complex diseases, epigenetic changes are expected to be much smaller, and thus more difficult to detect (Rakyan *et al.*, 2011). Most findings in studies of complex diseases have been of single CpG sites that are differentially methylated between cases and controls, referred to as differentially methylated positions (DMPs). While it has been suggested that a single CpG site can have an impact on the cell phenotype (Bell *et al.*, 2012), there might be a benefit of studying regions that are differentially methylated too.

The methylation status of CpGs in close proximity is usually found to be highly correlated (Eckhardt *et al.*, 2006). This correlation can lead to longer segments of DNA being differentially methylated between disease and normal tissue. If these differentially methylated regions (DMRs) overlap with regulatory regions or promoters, they can be of considerable biological interest. However, there is no consensus on what characterizes such a region in terms of methylation difference between cases and controls, length of the regions, distance between the CpGs, or other attributes. The two main approaches to identify DMRs involve the use of either static or dynamic aggregation. Static aggregation methods combine the data into predefined CpG islands or genes and include methods such as ProbeLasso (Butcher & Beck, 2015). The dynamic aggregation methods combine neighboring observations agnostic with gene annotation, and search the genome for DMRs relying only on the chromosomal position. This includes methods such as Bumphunter, DMRcate, BSmooth, and Comb-p (Hansen *et al.*, 2012; Jaffe *et al.*, 2012; Pedersen *et al.*, 2012; Peters *et al.*, 2015).

II.4. Genetics of complex diseases

Complex disease is a collective term used for diseases caused by a combination of multiple genetic, epigenetic, and environmental factors. Most of these diseases do not follow a clear Mendelian pattern of inheritance (Craig, 2008). In this thesis, I explore the genetics of complex diseases by using multiple sclerosis (MS) as an example.

MS and its epidemiology

MS is an inflammatory, demyelinating disease of the central nervous system (CNS), and is one of the leading causes of chronic neurological disability in young adults (Oksenberg *et al.*, 2008). Inflammation leads to a loss of the myelin covering the axons of the neurons within the CNS. The early stages of this inflammation are believed to be driven by autoreactive T cells (McFarland & Martin, 2007). The exact causes of MS are unknown, but there is strong evidence that both genetics and environmental factors contribute to disease susceptibility (Lassmann, 2013).

Multiple studies have shown that the prevalence of MS varies along a north–south axis, increasing with increasing distance from the equator (Milo & Kahana, 2010; Simpson *et al.*, 2011). However, there has been less observational evidence for variations in prevalence along a within-country axis, even for countries spanning many latitudes (Simpson *et al.*, 2011; Berg-Hansen *et al.*, 2014). MS affects twice as many women as men, but there is no clear biological explanation for this difference (Orton *et al.*, 2006; Bostrom *et al.*, 2014).

Studies of family recurrence rates in MS have estimated the narrow-sense heritability to be around 0.64 (Westerlind *et al.*, 2014). This means that the variation in the additive genetic risk factors account for 64% of the observed variation in disease risk (Tenesa & Haley, 2013). The human leukocyte antigen (HLA) risk variants account for approximately 8% of the sibling recurrence risk, and the identified non-HLA risk variants explain roughly 30% of sibling recurrence risk (Watson *et al.*, 2012; IMSGC, 2013). The heritability explained by common variants accounts for a large proportion of the observed disease heritability. This is in contrast to most other complex diseases, for which the known genetic risk variants explains much less of the observed heritability (Visscher *et al.*, 2012).

Genetics and epigenetics of MS

The strongest genetic risk variants of MS are located within the HLA complex (Sawcer *et al.*, 2014). In international GWAS, over 100 independent non-HLA risk variants have been identified (IMSGC & WTCCC 2, 2011; IMSGC, 2013). The majority of non-HLA genes associated with MS are situated in close proximity to immune regulating genes, which further adds to the evidence that MS is an immune-mediated disease (IMSGC & WTCCC 2, 2011; IMSGC, 2013). Many of the MS associated variants are also implicated in other autoimmune diseases (Sawcer *et al.*, 2014). Linkage studies of multiplex MS families have not found evidence of linkage outside the HLA complex, suggesting that rare variants carrying substantial risk of MS are uncommon (Modin *et al.*, 2003; Willer *et al.*, 2007; Sawcer *et al.*, 2014). Therefore, the main focus of the genetic research in MS has been on common variants assessed in large cohorts by international consortia. However, some associated genes have been reported to harbor rare variants. The best established examples are the genes *TYK2* and *CYP27B1*, with the associated *TYK2* risk variant has an allele frequency of 0.4% and the associated *CYP27B1* variant around 0.1% for the risk allele (Mero *et al.*, 2010; Sundqvist *et al.*, 2010). The *TYK2* gene harbors a rare non-synonymous variant that has been shown to be associated with MS in multiple studies (Ban *et al.*, 2009; Mero *et al.*, 2009; Mero *et al.*, 2010; Dyment *et al.*, 2012). The *CYP27B1* findings are more disputed, since rare MS risk variants within this gene have not been consistently replicated in independent cohorts (Sundqvist *et al.*, 2010; Ramagopalan *et al.*, 2011; Ban *et al.*, 2013; Barizzone *et al.*, 2013; Reinthaler *et al.*, 2014; Zhuang *et al.*, 2015).

Epigenetics studies focusing on MS are relatively recent, and compared with the MS studies of the genome, the epigenome has not been as thoroughly investigated for MS. Epigenetics may help to address several aspects of the disease, such as the missing heritability and disease heterogeneity in susceptibility and progression of disease. In the first published epigenetic report on MS, three discordant monozygotic twin pairs were compared using whole genome sequencing, targeting SNVs, DNA methylation, and mRNA (Baranzini *et al.*, 2010). An examination of 2 million CpG sites revealed that no loci were identified as associated with the disease or could help to explain disease discordance. Graves *et al.* investigated genome-wide methylation in CD4$^+$ T cells of MS patients and healthy controls (Graves *et al.*, 2013). However, many of the patients in Graves *et al.*'s study were treated with immunotherapeutic drugs, which may have modified the epigenetic profile of the immune cells. The authors identified multiple CpG sites associated with MS in their genome-wide methylation analysis. However, as reported in **Paper I**, we were not able to replicate these findings because the majority of their reported loci were not included in our analysis for technical reasons. Two other studies of epigenetics in MS have been published, one by Liggett *et al.* and one by Calabrese *et al* (Liggett *et al.*, 2010; Calabrese *et al.*, 2014). Liggett *et al.* identified 15 potential biomarkers for MS in a study of DNA methylation in cell-free plasma of MS cases and controls, and Calabrese *et al.* found a global reduction of hydroxymethylated cytosine in MS patients when investigating DNA methylation in blood.

As more genetic and epigenetic data become available for studies of complex diseases and new methodologies are introduced, much progress in this field can be expected. The '–omics' revolution that has integrated itself into medical research will in due time expand our knowledge of the genetic and epigenetic contributions to complex diseases, and may result in better understanding, prevention, and treatment of these diseases.

III Aims of the thesis

This thesis investigates different aspects of methodologies when analyzing genetic and epigenetic data with respect to identifying susceptibility to complex diseases.

Our first aim was to identify methylation patterns that can serve as biomarkers for MS in different purified immune cells and then replicate any findings in whole blood. This is explored in **Paper I**, where we compare the genome-wide methylation profile in different purified immune cells in MS patients with the methylation profile obtained from healthy controls. The comparison was done with regression analysis for each CpG, separately in purified immune cells and whole blood.

Our second aim was to investigate aggregation methods for rare variants and methylation sites when applied within rare variant association studies and epigenome-wide association studies in complex diseases. These methods are addressed in **Paper II** and **Paper III** respectively.

Rare variant association studies of complex diseases have become feasible with the introduction of rare variant genotyping chips, such as the Exome Chip. In **Paper II**, we investigated the performance of the Exome Chip by using an extensive computer simulation and static aggregation methods. The aim was to study the strengths and constraints of this chip, which contains mainly low-frequency and rare variants, by describing the relationship between sample size, effect size, and power.

In **Paper III**, we introduce a new method for identifying genome-wide significant DMRs in genome-wide DNA methylation studies. The method involves adapting scan statistics in the a similar way as was used before on another type of genetic data. The method is based on dynamic aggregation of neighboring methylation sites. Compared with the other two established methods, our method addressed the multiple testing issues in a better way.

IV Material and methods

IV.1. The datasets

In **Paper I**, we describe our investigation of the DNA methylation profile in relapsing remitting MS patients and how it compared with the methylation profile of healthy controls. For this purpose, we used blood samples from 16 treatment-naïve MS patients and 14 healthy age-matched controls. The clinical characteristics of the cohort are listed in Table 1 in **Paper I**.

From blood samples, we isolated two subpopulations of immune cells: CD4$^+$ T cells and CD8$^+$ T cells. We extracted DNA from the isolated cells, as well as from whole blood. All samples were assayed on the Illumina 450K methylation array and genotyped on the Illumina 660 Quad array. Genotypes were further imputed according to the central European reference panel (International HapMap Consortium, 2005). We removed probes that contained an observed or imputed SNP within the 50 base pairs (bp) long probe sequences on the 450k array. Based on the first two principal components (PCs), we did not observe any plate or batch effects in the methylation data. However, we could confirm cell-type specific methylation patterns on the basis of the first two PCs (see Figure 1A in **Paper I**). We then used the methylation data to study both single methylated positions and differentially methylated regions that could serve as a biomarker for MS.

The simulated data

The dataset used in **Paper III** is from the Finnish Health in Teens (Fin-HIT) study. This study recruited about 11,000 adolescents, in the age group 9–12 years, from schools throughout Finland. Different biometric information such as height, weight, and puberty scores, as well as epigenetic measurements from saliva was collected. In **Paper III**, we report how we used DNA methylation data from 100 girls aged 11 years, who were randomly selected amongst the 10th percentile with a both low and high end of the body mass index (BMI) distribution.

When benchmarking different methods for analyzing DNA methylation, we used the DNA methylation data from chromosome 22 as a backbone for a simulation study. This chromosome contained 58,910 observations, distributed over 1,071 CpG islands with a mean of 55 CpG per islands, ranging from 16 to 456 CpG sites. By adding an effect directly on the methylation data, we could investigate the performance of different methods for calling DMRs and then compare them with our new proposed method.

IV.2. Methods for the analysis of DNA methylation data

Pre-processing

Normalization and quality control of DNA methylation measurements are to a large degree technology-dependent. In **Paper I** and **Paper III** we describe how we used two different technologies to measure DNA methylation. For **Paper I**, we used a chip-based technology (Illumina 450k array) to measure the methylation levels, and for **Paper III** we used bisulfite sequencing to assess methylation.

For the Illumina 450k array used in **Paper I**, we considered three different methods for normalizing the observed methylation values: subset within array normalization, peak correction, and beta-mixture quantile normalization (Maksimovic *et al.*, 2012; Wang *et al.*, 2012; Teschendorff *et al.*, 2013). After testing all three normalization algorithms on our data and comparing the discrepancy between six technical replicates, all three normalization methods led to similar results. This indicated that it would make little difference which normalization method is chosen, and that other properties should be considered. Since the beta-mixture quantile (BMIQ) normalization allowed for the highest flexibility on which probes to keep and which to discard, this method became our preferred choice.

BMIQ is a normalization method for aligning measurement from two different chemistries (type I and type II probes) on the Illumina 450k array (Dedeurwaerder *et al.*, 2011; Teschendorff *et al.*, 2013). The two types of probes target different CpGs and have both different detection mechanisms and different properties, such as binding affinities. The type I probes have a wider dynamic range than the type II probes, and in BMIQ normalization the observations from the type II probes are mapped onto the distribution of the type I observations. This mapping is done using a mixture of three cumulative beta distributions, hence the name beta-mixture quantile normalization. The normalization was done independently for all samples, ensuring that no sample influenced the normalization of any other samples. We excluded probes that were missing in more than 10% of the individuals.

In **Paper III** we describe how we used reduced representation bisulfite sequencing to measure methylation levels in our samples. First, low-quality sequences were removed using Nesoni clip (Version 0.115, https://github.com/Victorian-Bioinformatics-Consortium/nesoni). The bisulfite converted sequence reads were then mapped to the human genome (hg19) using Bowtie2 Version 2.0.5 (Langmead & Salzberg, 2012) and Bismark Version 0.10 (Krueger & Andrews, 2011). Using Bismark methylation extractor, we could calculate the beta-methylation values for each CpG site. We retained only CpG sites with coverage above 10× and a call rate above 75% among all the samples.

Single site analysis

Paper I describes how whole blood was investigated separately from the CD4$^+$ T and CD8$^+$ T cells. The CD4$^+$ T and the CD8$^+$ T cells were analyzed together in one model, since the similarity in the methylation profiles between these cells types was assumed to be rather high. The two cell types originate from the same cell linage, and they appeared close in the PCA cluster plots (Figure 1A in **Paper I**). To analyze the two cell types in the same model, we used a linear mixed effects model in which each cell types had a random effect. By including an additional interaction between the two cell types, the heterogeneity between the cell types was accounted for in the same model.

For analysis of whole blood, a linear regression analysis was applied with case-control status as the independent variable and methylation as the dependent variable. We ran this model both with and without adjustment for possible confounders. We also estimated the white blood cell ratios in each individual from the methylation data, using a linear projection as described by Houseman *et al.* (Houseman *et al.*, 2012). This method required a training dataset for each cell type, which was combined with the whole blood methylation data, and informative markers for the different cell types were identified in the whole blood data. The training dataset we used is published in the article by Reinius *et al.*, and the cell types accounted for in our study were CD4$^+$ T cells, CD8$^+$ T cells, NK cells, B cells, monocytes, and granulocytes (Reinius *et al.*, 2012; Jaffe & Irizarry, 2014).

Static and dynamic aggregation of methylation sites

In this thesis, the terms static and dynamic aggregation are used to describe how genetic variables at neighboring sites, such as genetic variants or CpG sites, are aggregated into regions. In static aggregation, the variables are aggregated into predefined units, such as annotated genes or regulatory regions. Static aggregation was applied to identify genes which could be differentially methylated between MS cases and controls (see **Paper I**), and to investigate the power of gene-wise collapsing methods on the Exome Chip (see **Paper II**). For dynamic aggregation, the regions are not predefined, but rather defined based on the underlying dataset. This approach is applied in both **Paper I** and III to identify DMRs independent of genetic annotation beyond chromosomal position.

To identify potential DMRs in **Paper I** using a dynamic aggregation approach, we aggregated the top 5% of the test statistics into candidate regions stratified by cell type. A candidate region was defined to be any set of the top 5% CpGs with a maximal distance of 500bp between them. The candidate regions were permuted 10,000 times by shuffling the case-control status and recalculating the test statistics for each probe. The sum of the test statistics for each permutation was compared with the original sum in the candidate regions. We looked specifically for overlapping regions in the different cell types, such as whether CD4$^+$ and CD8$^+$ T cells shared any regions in the top 10 or top 100 regions. We applied the same permutation algorithm by clustering the CpGs to their annotated gene and permuting the gene regions. This can be considered a static aggregation of the CpG, where functional annotation dictates the aggregation, as opposed to letting the data drive the aggregation.

Bumphunter aggregates all CpG sites into regions based only on genomic annotation. There is no gap between the CpGs within a region that is larger than a limit specified by the user. Within the regions, the test statistic of each CpG is trimmed over a certain user-defined cut-off, often defined as a quantile of all the CpG-wise test statistics. The remaining sites that are in close proximity to each other are aggregated into subregions, which are the reported candidate DMRs. DMRcate uses a Gaussian smoothing kernel on the CpG-wise test statistics. The kernel bandwidth is equal to the maximum allowed gap between two CpG sites within the same region. Based on the smooth test statistics and using the Satterthwaite method (Satterthwaite, 1946), a new CpG-wise p-value can be calculated. The aggregation is done in such a way that no CpGs within the same region are more than a given number of base-pairs apart.

For the dynamic aggregation of variables reported in **Paper III**, we developed a novel application of a scan statistic for calling the DMRs. This method was benchmarked against established methods for DMR calling. In the benchmarking, we paid specific attention to the power as function of effect size, and we compared the convergence of power with increasing effect size for the different methods. There are different algorithms and methods for the dynamic aggregation of genetic observations into regions (Hansen *et al.*, 2012; Jaffe *et al.*, 2012; Pedersen *et al.*, 2012; Butcher & Beck, 2015; Peters *et al.*, 2015). We chose to compare our method with the technology-independent methods, which are appropriate for use on sequencing data. For the comparison we chose two methods: Bumphunter (Jaffe *et al.*, 2012) and DMRcate (Peters *et al.*, 2015).

Multiple testing penalty

A major issue with existing models for the identification of DMRs is that they do not properly adjust for multiple testing. Bumphunter uses permutations to construct a distribution of regions under the null hypothesis. The set of new regions arising from the permutations represents the expected distribution under the hypothesis of no association between the exposure and DNA methylation. The family-wise error rate (FWER) (Tukey, 1953) is estimated by the fraction of observed DMRs from the permutations having a larger area under the curve and spanning a larger number of probes than the candidate regions.

DMRcate calculates the region-wise p-value by combining the individual CpG-wise p-values within each candidate region, using Stouffer's method (Stouffer *et al.*, 1949). This is a method that has a close resemblance to Fisher's method for combining p-values, but operates on the test statistics instead of the p-values. The p-values from each CpG within the identified regions are combined into a region-wise p-value.

When combining p-values for a region, one has to take into account of whether the region has been pre-selected based on containing large test-statistics, as this may give an artificially low p-value if the preselection is not accounted for. To address this problem, we proposed a method based on scan statistics to identify DMRs. By building on a well-developed mathematical framework for scan statistics, we could identify DMRs while properly adjusting for multiple testing.

Scan statistics

Introduction

The scan statistic method presented in **Paper III** is a novel application of a method formulated by Zhang (Zhang, 2008). Zhang's method, which we extended, is a scan statistic method with empirically derived thresholds for the window sizes in the model. The method relies on Aldous's argument that for sufficiently large thresholds, the number of peaks over the thresholds follows a Poisson distribution (Aldous, 1989); this is sometimes referred to as Poisson heuristics. Zhang deduces a relationship between the significance level and the intensity rate of the peaks (Zhang, 2008). This relationship can be used to determine the threshold, given a predefined significance level α. Since Poisson heuristics assumes that all window observations are independent, overlapping windows may reduce the performance of this method. To account for the dependencies, no overlapping significant windows were allowed. The test with fixed window size and overlapping windows is referred to as the R-test in Zhang's notation. The extension of this test to the case with several window sizes, which Zhang referred to as the S-test, is also of interest. To construct independent observations in the S-test, no overlapping or nested significant windows were allowed in our study. If two overlapping windows were both above their respective thresholds, the smallest window was preferred as the significant one.

The mechanism that regulates genome-wide significance is the window thresholds. The window thresholds are the values with which the sums in each window are compared. If the sum within a window is above the threshold, it is regarded as significant. By testing different window thresholds on a null-simulation, we could identify the lowest window threshold that kept the overall false discovery rate (FDR) at a given level. For a given significance level, the thresholds for the different window sizes depend on the number of windows and the correlation structure in the null model of the single site test statistics. The window size for the scan statistic can be set by the user, but the correlation structure between the single site test statistics has to be determined in advance. For most correlation structures, no closed expression exists for the window thresholds, and thus the thresholds must be estimated numerically with a Monte Carlo method. However, a special case has been published by Siegmund *et al.* (Siegmund, 1985; Siegmund *et al.*, 2011), where the thresholds can be calculated analytically if the dependencies between the test statistics follow an Ornstein-Uhlenbeck process. This is a much faster way of determining the thresholds, since it does not rely on numerical approximation, but rather a closed-form expression. These thresholds can be tailored to account for multiple testing in the same way as the Monte Carlo test described above, by testing different thresholds and calculating the false discovery rate. The minimum thresholds that hold the alpha level for the false discovery rate after multiple testing are carried forward to the analysis.

Estimating window thresholds

There are different approaches to estimating the window thresholds for a given dependency structure and significance level for a scan statistic method. As described in **Paper III**, we chose to compare three different ways of constructing the window thresholds: two were based on Monte Carlo simulation, and one was based on Siegmund's analytical solution (Siegmund *et al.*, 2011). Of the two Monte Carlo simulated thresholds, one was a full-scale simulation of all the CpG-wise test statistics on chromosome 22, and one was based on the importance sampling procedure proposed by Zhang (Zhang, 2008). These three methods resulted in different window thresholds, which in turn led to some differences in the results. This can be seen in Figure 1 in **Paper III**, where the convergence in power for all the methods is plotted as a function of increasing effect size.

In the full Monte Carlo approach, we used an autoregressive (AR) process to simulate a null observation for all loci. Before doing the AR simulation, we investigated the autocorrelation function (ACF) and the partial autocorrelation function (PACF) for a large number of CpG islands on chromosome 22. Inspection of the ACF and PACF of the CpG islands indicated an AR(2) as the best common model. For each CpG island on this chromosome, we fitted the an AR(2) model and extracted the parameter estimates. The 75% quantile of these parameters was chosen as the overall AR parameter for the simulation. This was used to sample an AR(2) process of equal length to chromosome 22, and the sliding windows were applied to these simulated data.

When using importance sampling, we sampled a subset of observations. This subset was equivalent to two windows in length, simulated from a multinormal distribution. In the covariance matrix for this distribution, the off-diagonal elements represented dependencies in both spatial directions. In this simulation, we could sample the different window sums, weighted by the likelihood of observing the different sums. Based on this smaller segment and a given threshold, the number of peaks above the threshold was scaled to correspond to the full dataset.

A method for estimating the window threshold has been introduced by Siegmund *et al.* (Siegmund *et al.*, 2011), whereby the intensity of the peak over threshold is analytically calculated as a function of the threshold. The derivation of this equation is based on the assumption that the test statistic follows an Ornstein-Uhlenbeck process. A closed-form solution was first published by Siegmund (Siegmund, 2013), based on earlier work by Siegmund and Yakir (Siegmund & Yakir, 2007).

In all simulations, the exceedance rate for each window size was calculated on a grid of different threshold values. In these simulations, we could estimate the number of false positive findings for different thresholds and derive the relationship between the false positive rate, window size, and window thresholds. For a given false positive rate, we could distribute the false positive observations on the different window sizes in equal proportions by adjusting the individual window thresholds accordingly.

Benchmarking DMR calling methods

All efforts were taken to ensure that the parameters in Bumphunter and DMRcate were comparable with our method. However, we aimed to set as many parameters as possible to their default values without compromising the comparability. This was done to benchmark the methods in a realistic way, since in most analysis settings the methods are often used with their default parameter values.

In order to add an artificial causal effect to the data, we modified the M-values directly (Du *et al.*, 2010) by adding an offset on the causal CpGs in all cases. A random set of 100 causal DMRs of different lengths were picked from all CpG islands. All causal DMRs were shorter than the CpG islands they were located in. Thus, all CpG islands with a causal region also contained non-causal CpG sites. The CpG islands could only contain one causal DMR each, and therefore no CpG island had two or more causal DMRs.

Two different ways of adding the effect size were investigated. In one scenario, all CpGs within a causal DMR were shifted uniformly by the same value. In the second scenario the added effects were multiplied by a normal density kernel, making the added effect bell-shaped. In this scenario, the kernel contained a normalizing constant, such that the area of the added effect size (number of probes × effect) was the same as in the first scenario for each effect size. However, this made the maximum of the added effect in this scenario almost twice the value compared with the first scenario, and on the edges of the DMR a very small effect was added. The benchmarking was done on a set of effect sizes ranging from 0 to an added value of 5 on the logistic scale for all casual CpGs.

IV.3. Methods for the analysis of low-frequency and rare variants

Methods for aggregating rare variants

Several different methods for aggregated variants in rare variant association tests have been published (Madsen & Browning, 2009; Liu & Leal, 2010; Ionita-Laza *et al.*, 2011; Neale *et al.*, 2011; Wu *et al.*, 2011; Lettre, 2014; Lin, 2014). All methods for rare variants rely on static aggregation, where the variants are collapsed into annotated regions, which for rare variants are usually genes. Most methods fall into one of two categories: burden tests or variance tests (Auer & Lettre, 2015). Burden tests compare the burden of mutations in cases compared with controls. However, if a gene harbors both deleterious and protective variants, the variants might cancel out in a burden test, leaving the gene nonsignificant. In variance tests, the 2nd moment of the distribution of variants within a gene is compared between cases and controls. By considering higher order moments instead of sums, it is possible to account for variants with opposite effects.

In **Paper II**, we report a study of two methods: weighted sum statistic (WSS) and the sequence kernel association test (SKAT). These methods serve as surrogates for the two classes of models, where WSS is a commonly used burden test, and SKAT is a popular variance contrasting test. SKAT is a generalization of a well-known test for rare variants, the C(α) method (Neale *et al.*, 2011; Wu *et al.*, 2011). None of these methods are tailored to one specific genotyping technology, and are thus applicable to data from both genotyping chips as well as sequencing.

We chose WSS, both due to its simplicity and the ease with which it can be implement, and we chose SKAT because of is frequent use and its implementation in R (R Core Team, 2012; Seunggeun *et al.*, 2015). SKAT uses a kernel projection of the variance between cases and controls, and we used the beta distribution as the kernel. This kernel puts high weights on variants with few observations, and low weights on variants that are observed often.

Benchmarking rare variant methods on the exome chip

In **Paper II**, we describe how we simulated a large population pool according to the method used by Basu *et al.* (Basu & Pan, 2011), with the frequency thresholds obtained from the Exome Chip consortia (Exome Chip Consortium, 2011). From this pool of simulated individuals, we drew smaller cohorts to investigate the influence of sample size on power when using the Exome Chip.

The genotypes were constructed using multinormal random variables, where each vector of observations corresponded to a DNA strand within an individual. To model LD patterns between the variants, we used the covariance matrix within the multinormal distribution. The covariance matrix for the distribution was modeled with the Matern covariance function, where the dependency between variants was inversely proportional to their distance in base pairs (Matern, 1960). The model parameters were selected so that dependencies between variants in different genes would be negligible.

By dichotomizing the resulting vectors from the multinormal distribution with the allele frequencies reported by the Exome chip consortia, we could obtain a dataset with allele frequencies equal to that of the Exome Chip and still retain the LD structure between the simulated variants within the genes. The algorithm for simulating the variants can be found in the supporting information to **Paper II**.

While different ways of simulating effects on genetic data exist, there is no consensus on which method performs best under these circumstances. We chose the simulation approach presented by Madsen *et al.*, since it only has one free parameter that can be controlled (Madsen & Browning, 2009). The free parameter is the population attributable risk, which we considered as the effect size in our benchmarking. To restrict the possible scenarios to be explored, we only used one direction for the effect size and all variants had the same population attributable risk. Using the relationship between genotype relative risk and population attributable risk, one can construct the genotype relative risk of a variant, given the population attributable risk and the allele frequency. Based on the genotype relative risk of all the causal alleles and the carrier status of each individual (i.e., carrying 0, 1, or 2 of the variant in question), we could calculate the probabilities for each individual being a case. To determine the phenotype of each individual, a "loaded coin" was tossed, with the probability of being a case determining the load. The complete algorithm is described in detail in the supporting information to **Paper II**. When adding a simulated effect on the variants, we randomly selected 100 genes to be causal for the phenotype in question, and investigated two different scenarios. In the first scenario, all variants within the 100 causal genes were causally linked to the phenotype. However, having 100% of the variants casual for a disease is an unrealistic biological scenario. In the second simulation scenario, we therefore relaxed this assumption and randomly sampled 50% of the variants from the same 100 genes to be the causal variants.

We compared the power of the two methods for effect sizes between 0% and 8% population attributable risk, as shown in Figure 2 in **Paper II**. The retrieval rate for each gene was considered a surrogate for power. The retrieval rate for each of the 100 genes was assessed 50 times by drawing new case-control cohorts from the simulated pool of individuals. This allowed us to construct an empirical confidence interval for the power of the different methods. We compared the performance with increasing sample sizes for both scenarios. The different sample sizes were chosen to reflect realistic cohort sizes in rare variant association studies, and the case-control ratios were always equal to one for each simulation.

V Summary of results

V.1. Paper I

Genome-wide DNA methylation profiles indicate CD8$^+$ T cell hypermethylation in multiple sclerosis

This study aimed to identify MS-specific DNA methylation biomarkers in purified immune cells, which might also be used as biomarkers whole blood. We isolated DNA from CD4$^+$ T cells, CD8$^+$ T cells, and whole blood from 16 treatment-naïve female MS patients and 14 age-matched female controls. DNA methylation was assessed genome-wide using the Illumina 450k array.

The MS patients and the controls did not show any significant difference in phenotypic aspects important for DNA methylation studies, such as smoking status and age. The MS group was quite homogenous, as all had been diagnosed with relapsing-remitting MS, and all patients had low EDSS and MSSS scores.

To account for variation between the ratios of the different white blood cells in whole blood, we estimated the cell type proportions within each individual using the Houseman algorithm with a training dataset taken from Reinius *et al.* (Houseman *et al.*, 2012; Reinius *et al.*, 2012). No significant differences in the cell type proportions between MS cases and healthy controls were observed. When we inspected the effect size of the probes with nominally significant p-values in the differential DNA methylation analysis between MS patients and controls, we noted that they were largely shifted towards hypermethylation in the CD8$^+$ T cells for the MS patients. For the CD4$^+$ T cells and in whole blood, this hypermethylation was not observed; there was balanced hyper- and hypomethylation.

After analyzing the DNA methylation in the genes in close proximity to the MS-associated SNPs (IMSGC, 2013), no significant enrichment of differentially methylated genes was found. Following a single site analysis, no CpG was genome-wide significant in CD4$^+$ T cells, CD8$^+$ T cells, or whole blood, after correcting for multiple testing.

In our study, no individual CpG, gene, or candidate regions were genome-wide significantly different in MS patients and controls after correction for multiple testing. However, we did observe a striking number of hypermethylated probes in the CD8$^+$ T cells of MS patients compared with controls. We did not observe any systematic hypermethylation in either CD4$^+$ T cells or in whole blood.

V.2. Paper II

Assessing the power of Exome chips

Genotyping chips for rare and low frequency variants have become feasible with the introduction of Exome Chips. Our objective was to investigate the performance of these genotyping chips in different scenarios. We simulated 200,000 individuals with the same allele frequency spectrum as reported by the Exome Chip Consortium (Exome Chip Consortium, 2011). From this population pool, we drew cohorts of different sample size.

The methods were tested on a set of genes with no effect on the phenotype, to test whether the false positive level was acceptable. Both methods controlled the Type I error sufficiently, with SKAT being marginally more conservative than WSS.

For small effect size and small sample size, we observed that WSS converged marginally faster than SKAT, but when the population attributable risk reached 0.5%, SKAT outperformed WSS in power. We also noted that SKAT had a much slower convergence in power when all variants within each gene were given the same weight, than if the variants were given a weight that was inversely proportional to their allele frequencies.

The rate of convergence in a sample size of 100,000 samples (50k controls and 50k cases) was quite close for SKAT and WSS, but for small samples sizes there were substantial differences in the convergence rate for the power.

In line with earlier studies (Lin, 2014), we found that SKAT outperformed WSS in most cases. We also found that for small to moderate effect sizes, a sample size of 20,000 should be sufficient. This effect size should correspond to a population attributable risk around 0.5% on all variants within each gene. When only half of the observed variants within a gene are causal, a much larger sample size or effect size is needed to reach the same power.

V.3. Paper III

Assessing genome-wide significance for the detection of differentially methylated regions

We introduced a statistical approach to identify differentially methylated regions (DMR) with adjustment for multiple testing on the region-wise level. The latter has not been sufficiently addressed by methods to identify DMRs introduced earlier.

The window thresholds are the values that determine significance, and represent the value with which the window sums are compared in the scan statistic. For the method presented in **Paper III**, we evaluated three different ways of determining the window thresholds: first, using a full-scale MCMC simulating the null distribution of all test statistics; second, using the importance sampling algorithm outlined by Zhang (Zhang, 2008); and third, using Siegmund *et al.*'s (Siegmund *et al.*, 2011) analytical expression to determine the thresholds. All three methods gave different thresholds with the same basic input parameters, such as number of methylation sites and FDR limit. We compared three DMR methods with Bumphunter and DMRcate in two different scenarios of added effect size. Both of these scenarios included the same CpGs and 100 predetermined causal regions.

In the first scenario, we had a uniform effect size on all CpGs within the causal region. In the second scenario the effect size was weighted with a standard normal kernel. Due to the kernel smoothing, the effect size was lower on the border of the causal region and gradually increasing towards the midpoint of the causal region. In order to have a comparable scenario, we multiplied the kernel by a normalizing factor, making the area of added effects the same in the two scenarios.

Figure 1 in **Paper III** shows the convergence rates in power for the five different methods, where the first scenario is shown on the left-hand panel and the second scenario on the right-hand panel. Figure 1 in **Paper III** also shows the plotted false positive findings for each of the methods (in light colors). The top panels show the power to detect parts of the regions as a function of effect size, and the bottom panel shows the fraction of the recovered CpGs. Our method outperformed both DMRcate and Bumphunter in calling DMRs (top panel), and the difference in the convergence rate was substantial for small effect sizes. Investigating the proportion of true positive CpGs, DMRcate outperformed all other methods but at a cost of higher number of false positive rate. While the false positive probes in DMRcate were in close proximity to the causal regions, the proportion of false positive probes was considerable compared with the other methods. Bumphunter, DMRcate, and our method using the Siegmund thresholds did not called any false positive DMRs. Using the more liberal thresholds calculated by the Monte Carlo sampling and importance sampling resulted in less than 10 false positive independent regions out of 1,071 CpG islands assessed. Since the window thresholds in our method were independent of the effect size, the false discovery rate was constant for all the three methods.

The regions in which we added an effect varied in size from 5 CpGs to 100 consecutive CpGs. It is conceivable that the power depends both on the effect size and the length of the DMR, since longer regions tend to be easier to call. To investigate this, we plotted the effect size multiplied by the length of the causal region against the power, as show in Figure S1 in **Paper III**. Figure S1 shows the average of the power when accounting for both effect size and region length. We observed a similar pattern of convergence rates in power as shown in Figure 1 (**Paper III**). In the first scenario of uniform effect sizes for an effect size times length equal to 80, our method had reached a power of 100%, while Bumphunter and DMRcate never reached 100% power. In the second scenario, Bumphunter and DMRcate performed equally well, but our method performed much better compared with the first scenario. This could be explained by the peaks in the test statistics in this scenario and our method being more sensitive to small peaks than both Bumphunter and DMRcate.

In conclusion, our method outperformed both Bumphunter and DMRcate in detecting causal DMRs, especially for small effect sizes, while also keeping the false discovery rate under control.

VI Discussion

This thesis presents several broad aspects of genetic and epigenetic association testing in complex diseases, with a special focus on multiple sclerosis. It outlines some methods, challenges, and results for rare variant analysis as well as the analysis of methylation.

There have been numerous studies of both the genetic and environmental risk factors of MS, but there have been few publications on epigenetic factors in MS. **Paper I** is among the first publications on MS and DNA methylation in CD4$^+$ T cells, and the first to report simultaneous analysis of the genome-wide methylation profile of CD4$^+$ and CD8$^+$ T cells in treatment-naïve MS patients.

While association studies involving common variants have been successful, searches for rare variants involved in complex diseases have not been equally successful, with the exception of some psychiatric diseases (Walsh *et al.*, 2008; Neale *et al.*, 2012). Moreover, as more rare variants are discovered in the efforts to characterize the complete human genome, many more variants have the potential to be identified as risk alleles. However, for any given phenotype, this search will require a substantial number of individuals to be included, as shown in **Paper II**.

The technological developments for genome-wide detection of DNA methylation has resulted in many new epigenome-wide association studies in complex diseases. Many of these studies have identified regions that are differentially methylated between cases and controls. However, both the definitions of a DMR and the methodology for identification of DMRs have varied widely between studies, making comparisons of the results difficult. Few studies have compared the different methods for identifications of DMRs, and in this regard **Paper III** contributes to the growing field of statistical modeling in epigenetic studies.

Several aspects of the methodology are important to consider when analyzing genetic and epigenetic risk factors in complex diseases. These include study-specific problems such as power and confounders, as well as the influence of the measurement technology (e.g., the genotyping chips) on the results presented in each study.

VI.1. Power

Power is defined as the probability of correctly rejecting the null hypothesis when the alternative hypothesis is true. In both **Paper II** and **Paper III**, we report how we investigated power by simulation, thus taking an empirical approach when assessing the power. In both cases, power was estimated by the fraction of recovered causal loci. Using this approach, we had to make some assumptions about how the causal effects were distributed. All of the assumptions made in the simulations will have affected the estimated power in some way. For the study reported in **Paper II**, we investigated the performance of the Exome Chip in terms of power as a function of different sample sizes and effect sizes. We investigated 100 genes that were randomly chosen to represent the whole chip. The length of these genes, as well as the allele frequencies of the variants residing within them, potentially affects the estimated power.

In the study reported in **Paper I**, in order to have best chance of discovering differences between MS patients and healthy controls, we carefully selected our participants to be as homogeneous as possible. Because this was a small pilot study, we had the possibility to select such a homogeneous patient group, but for larger studies this may not always be the case. Increasing the sample size is usually preferable if all individuals can be selected from the same homogeneous group. However, population heterogeneity may reduce the power and should be addressed when increasing the sample size.

With population heterogeneity under control, significant contributors to power include effect size, sample size, and adequate statistical methods. The first component is usually outside the control of the experimenter, and the second component may be limited due to experimental constraints. This leaves the statistical methodology as often the most adjustable component. In both **Paper II** and III, we describe how we compared different methods applied to the same dataset and observed quite large differences in power for the different methods. This indicates that choice of methods can have a large effect on power, and that the optimal method for a design can have substantially higher power than all other methods.

When studying the contribution of rare variants to disease risk, population heterogeneity has to be given a special attention since rare variants tend to be restricted to a particular population or to a specific geographical location. Thus, a small shift in the geographical distribution between cases and controls can give sufficiently large differences in the allele frequencies to cause variants to be falsely associated with the disease. In **Paper II**, we explain how we circumvented this issue by simulating individuals from a homogeneous population. The target population was a mixture of the all genotypes from all individuals included in the design of the Exome Chip. However, this population is a mixture of different ethnicities. The majority of the cohorts contributing to the Exome Chips design consist of European-Americans, which made up roughly three-quarters of the individuals (Igartua *et al.*, 2015). The simulated population pool reflected an ideal scenario with no population stratification, since the genotypes were sampled independent of ancestry. This simplification may have led to an overestimation of the power, since population heterogeneity will generally reduce the power in an association study.

Population heterogeneity in both rare variant studies and GWAS can be seen as a structurally similar problem to tissue heterogeneity in EWAS. The solution to this problem is to estimate the main contribution to this unwanted variation and then adjust for it. In GWAS, this is usually done using the principal components as covariates, while in EWAS this is done using a training dataset to estimate the relative cell type proportions. However, removing too much of this variation carries the risks smoothing out truly informative variation between the cases and the controls, and therefore this should be borne in mind.

It is reasonable to assume that a single locus test for rare variants will perform worse than collapsing methods. This can be attributed to the fact that power decreases with decreasing allele frequency (Sham & Purcell, 2014). However, since informative markers are also combined into one unit, they will enhance the each other's signal. In the case where only one variant within a unit is causal, aggregating it with non-causal variants will not increase the power. Thus, the assumption of increased power from a collapsing test is only reasonable if one assumes that there is more than one casual variant within the gene and/or unit of interest.

The power of the methods assessed in **Paper II** was presented as a function of increasing population attributable risk, which was the only free parameter when simulating disease risk. However, each gene contributed a different number of variants, each with different genotype relative risks. Our plots presented a summary of power over all the genes, which may give a simplified representation of the correlation between population attributable risk and power. Depending on the allele frequency of the variants within each gene, the population attributable risk will affect the total risk from the genes differently. For a given population attributable risk, genes with many rare variants may have a much higher total load of genotype relative risks than genes with few variants. We could have attempted to summarize the genotype relative risks within each gene and presented the power as a function of both the genotype relative risk and the population attributable risk. However, this would have been needlessly complicated and could have obstructed the interpretation of the overall power for the Exome Chip. Since population attributable risk is easier to interpret than genotype relative risk on the population level, it is a more convenient effect measure to use when assessing the power of a method in larger cohorts.

In **Paper III**, we show that the choice of statistical methods has a large impact on power, as shown in Figure 1 in **Paper III**. However, power can be perceived from different angles, as illustrated in the upper and lower panels in Figure 1. In the upper panel, the number of causal region containing at least one observed DMP is presented as an estimation of the power. The lower panel displays the true positive CpGs as a fraction of the total number of causal loci. One explanation for the large discrepancies in our method (shown between the upper and lower panels) is that a sliding window is likely to have at least one significant window within the causal region, but does not identify the correct boundaries of the causal DMRs. When identifying the causal CpGs, the dominance of DMRcate shown in the lower panel is due to a quite aggressive smoothing of the test statistic. This smoothing also gave rise to the large number of false positive observations, which lay on the border of the causal regions.

VI.2. Confounders

A confounder is a factor that is a common cause of both the exposure and the outcome. The presence of an unmeasured confounder can profoundly impact the estimated association between the exposure and the outcome. This can lead to false associations between exposures and a disease, and should therefore be avoided by all possible means.

In **Paper I** we identify possible confounders, including smoking, disease modifying treatment, and tissue heterogeneity. Smoking is a factor that is known to affect disease risk and modulate DNA methylation, and is therefore a possible confounder (Zeilinger *et al.*, 2013; Kucukali *et al.*, 2015). However, when adjusting for smoking status in our regression analysis reported in **Paper I**, there were no observable changes in the main effects. Another important possible confounder mentioned in **Paper I** is the tissue heterogeneity of whole blood. Blood consists of a mixture of different cell types, each with a distinct methylation profile. If a disease is driven by one particular cell type in whole blood, this can result in observed differences in methylation between cases and controls when measured in blood. These differences in methylation may not be due to true methylation differences as such, but are rather a result of different ratios of cell types between the cases and the controls. The method we used to estimate cell type ratios relied on a training dataset to identify methylation loci that were informative for the different cell lineages. As a training dataset, we used the freely available dataset published by Reinius *et al.* (Reinius *et al.*, 2012). This dataset originates from seven Swedish males, and contains eight different cell types. The compatibility of this training dataset with our sample population is questionable, due to the sex and population difference between the two cohorts (Zhang *et al.*, 2011; Singmann *et al.*, 2015). However, adjusting for the estimated cell type ratios did not change the order of any of the main effects when ordered by p-value, nor did it change the p-value distribution. Due to lack of significant changes, and since this adjustment may add considerable noise to the model, this potential confounder was not included in the final model. The MS patients included in the study in **Paper I** came from a homogeneous group, and all had been recently diagnosed and were of Nordic genetic ancestry. A key strength of our study was that none of the MS patients were taking immunotherapeutic drugs, which also would have had the potential to influence the methylation patterns in immune cells in MS patients.

Given the large data size in genomics, it is not only important to identify possible confounders, but also to adjust for them without additional computational cost. In **Paper I** we describe how we tested each locus independently using a linear model and expanded the model to account for any possible number of confounders. In the collapsing methods used in the studies on which **Paper II** and **Paper III** are based, it may not always be the case that the models can be expanded to account for covariates without a large increase in the computational cost. All collapsing methods assessed for DMR calling as well as SKAT have the ability to be extended with additional covariates. However, for Bumphunter, this expansion comes at the cost of a substantial increase in computational time. The WSS method is the only collapsing test that cannot easily include covariates.

VI.3. Methods

Technological considerations

The Illumina 450k DNA methylation microarray (**Paper I**) used a 50bp probe sequence that hybridizes to the DNA in our samples. The probes sequences were designed based on a reference genome, which represented an "average" human genome. In an individual with a SNP in one of these probe sequences, the hybridization may not have been optimal and could have affected the methylation readout. A SNP residing in the probe sequence with a weak association to MS can cause quite large changes in the methylation measurement (see Figure S1B in The supporting information to **Paper I**). To account for this, we removed all probes where we observed a SNP residing in the probe sequence in our samples. This step removed almost one-fifth of the total data points. It is possible that this reduction may have caused us miss information that could have been relevant for the disease. However, the interpretation of such findings would have been difficult, due to the possible confounding of those SNPs. We could not rule out the possibility that we had not removed all probes harboring a SNP in our sample population. However, the lack of any strong signal in our remaining data suggests that the majority of SNPs affecting the readout had been removed.

In **Paper II**, we highlight an important consideration when using the Exome Chip, which is that the cohorts used when this chip was designed were enriched for a narrow set of disease groups. Many important disease groups were missed, such as autoimmune diseases, neuromuscular diseases, and others. The main groups of diseases included were lifestyle disorders such as type 2 diabetes, cardiovascular diseases, and BMI extremes. Thus, variants conferring risk for other diseases may have been missed. This may lead to reduced performance of the chip when applied to other diseases.

Simulation of genotypes

When simulating the genetic variants in the study reported in **Paper II**, a number of considerations had to be balanced. These included computational efficacy, correlation between variants due to LD, and variant frequency. Frequency tuning when simulating genetic variants is easy to do, but it comes at the expense of realistic LD modeling between the variants. When realistic modeling of the LD between the variants is considered more important than the allele frequencies, a coalescent simulation might be preferred. However, in each iteration of a coalescent simulation, there will be some fluctuations in the allele frequencies. For very rare variants, small fluctuations in the frequency will be large relative to its absolute value, and thus frequency tuning on a very fine scale may be difficult to obtain. For these reasons, we used the method outlined by Basu *et al.* to simulate the sample population in described in **Paper II** (Basu & Pan, 2011). This method generates a "snapshot" of the genetic variability in an outbreed population. Each individual is simulated once, and all individuals are simulated independently. Since each individual is simulated independently, the computational burden scales well with the number of individuals simulated for this method. When simulating the variants, the LD pattern was down-prioritized over exact frequency tuning, but was still present to some degree in the final simulated population pool. However, since the LD between rare variants tends to be low, this should not have distorted the simulated genotype too far away from a realistic setting.

Given a set of simulated genotypes, there is no general way of classifying the individuals into a binary trait based on their genotype. To stratify the individuals to a case-control status and to control the degree of added effects, we chose the algorithm presented by Madsen and Browning (Madsen & Browning, 2009), and later used by Lin (Lin, 2014). Given the genotype relative risk for the causal alleles in an individual, this algorithm constructs the probability for an individual of being a case. The genotype relative risk scales linearly with the population attributable risk, and are inversely proportional to the allele frequency. This means that the rarest variants will have highest impact on the disease probability, and that the contribution to disease risk will fall with increasing allele frequency. The interpretation of population attributable risk is the fraction of cases that can be explained by the exposure, where the exposure is the risk allele. In this setting, it is the fraction of cases that can be directly attributed to the risk variant. The population attributable risk indicates what fraction of the cases would not have developed if the risk variant had not existed in the population.

For computational reasons, we randomly selected a set of 100 genes to use for the power analysis. The number was picked to balance the computational burden and biological plausibility. The current results from most GWAS suggest that for most complex diseases, around 100–300 genetic variations may be involved with a measurable contribution. It is still not known whether the number of contributing rare variants to complex diseases is higher or lower than for common diseases. However, given our current knowledge, the number of causal genes used to simulate the phenotype in **Paper II** was not an unreasonable guess.

For large effect sizes, we did get enough spread of cases in our population pool to draw 100,000 samples with a 1:1 case-control ratio. However, for small effect sizes (below 0.2% population attributable risk), there were not enough cases in the population pool, and we needed to add additional causal alleles outside the investigated genes. These additional alleles contributed to the phenotype but were not assessed in the power analysis. Since this addition did not occlude the contribution from causal genes on the phenotype, the addition of (external) effects should not be problematic.

Single site analysis

In both GWAS and EWAS, the first assessment of the data is usually an independent investigation of each loci, adjusted for any identified confounders. In **Paper I**, when modeling each CpG site independently, we used a mixed model to account for heterogeneity between CD4^{+} and CD8^{+} T cells. This model allows for covariates to be included, and a better approach might have been to include the genotypes of any SNP that resided in the probe sequence in the model. Our approach may have left probes with a stronger signal in the analysis, but the interpretation could be influenced by the presence of such a SNP.

The significance cut-off for p-values after Bonferroni correction in genome-wide studies is usually fixed at 5 × 10-8, arising from the observations that after correcting for LD there are approximately 1 million independent loci (Pe'er *et al.*, 2008). However, rare variants are usually not assumed to be in equally strong LD as the older common variants, and thus this threshold may not be appropriate if many rare variants are tested (Auer & Lettre, 2015; Fadista *et al.*, 2016). However, if more stringent p-value threshold are used for rare variants, an unreasonably large sample size would be required to have the power to obtain any genome-wide significant findings.

Regional methods

DMR calling in methylation data

To identify possible DMRs in the study reported in **Paper I**, we trimmed the top 5% of data of the test statistic and aggregated them into candidate regions. This trimming of the data may have caused some inflation when calculating the permutation p-values for the regions, since only the probes that were most significant were used. When the candidate regions were then permuted, they were compared against an already inflated statistic. This may have underestimated the permutation p-values for the candidate regions presented in our paper. In Bumphunter, this issue is partly resolved by generating new independent regions from the permutations, and by adjusting for the length of the regions when comparing the permutation results to the candidate regions.

In **Paper III**, we describe how, when assessing our method for DMR calling, it became clear that the method would perform best when the signals in the data were distributed in distinct peaks of sufficient height. However, for most complex diseases, the DMR signals are not likely to take the shape of distinct peaks. For this reason, the window thresholds need to be close to the test static in order to pick up any signal. When the threshold approaches the test statistics, small changes in the threshold values can have large consequences for the Type I error rate. For this reason, a higher threshold might be preferred if more than one threshold value is available for the same significance level.

When constructing thresholds for the window sizes, different assumptions were made about the distribution and dependency for the test statistics (**Paper III**). For the closed form expression of the threshold by Siegmund, the test statistics were assumed to follow an Ornstein–Uhlenbeck process (Siegmund, 1985; Siegmund *et al.*, 2011). The Ornstein-Uhlenbeck process is equivalent to a continuous stationary AR(1) process, and may not have sufficiently long conditional dependencies between observations to model the methylation data adequately. For any other dependency structure, a Monte Carlo simulation is required to determine the thresholds. While an AR(p) model for the threshold might have given the most realistic dependency structure of the data, it was also by far the most computationally expensive to determine. For a large numbers of loci or a long sequence of different window sizes, the threshold estimation with this method may be unreasonably expensive in central processing unit (CPU) time. However, the AR(p) estimation could easily be done in parallel for each window size and for the different threshold grids, thereby reducing the perceived computational time. The middle ground between a full-scale Monte Carlo simulation and a closed form is the importance sampling algorithm. This is a subsampling of a small section of the data and estimates the window thresholds based on this subsection. Importance sampling had a very similar behavior to the Monte Carlo sampling of an AR(p) model, but with a substantial reduction in computational time.

When sampling an AR process, one assumes that all observations are equally spaced on the chromosome. This is clearly not the case, and may affect the estimation of the window thresholds. Additionally, the AR sampling considers only the total number of methylation sites, and is not tailored to the dependency structure within each CpG island. Investigation into the different CpG islands showed a large variation in the dependency structure. However, it is not feasible to account for this heterogeneity in dependency structures. To estimate the window thresholds, an overall dependency structure has to be applied to all CpG islands.

To prevent the window sums being canceled out by test statistics in opposite directions within the same window frame, we applied the absolute value operator. This operation can be accounted for in both the AR(p) and the importance sampling estimation of the thresholds. However, Siegmund *et al.* (Siegmund *et al.*, 2011) state that the test statistic is assumed to be symmetric, and taking the absolute value of the test statistics breaks this symmetry. However, Siegmund *et al.*'s method always gave more conservative window thresholds than the Monte Carlo simulations, and this transformation therefore seems to have no adverse effects.

Difference of aggregation methods in rare variant and epigenetic association studies

The studies described in **Paper II** and **Paper III** used aggregation of variables to increase the power of a genome-wide study. In rare variant association studies, the focus was on aggregating variants into genes (or other predetermined units), which were then tested sequentially. In EWAS, the most common methods have been more data-driven and annotation-free, allowing for investigation of subregions within genes or CpG islands.

The different approaches reflect differences in both biology and the data. Rare causal variants are not assumed to be in high LD, but are expected to distribute uniformly over the genes. Hence, no sub-unit in the genes is favored in the analysis and the location of the variants within the genes relative to each other is not considered important. What is important is the total variant load within each gene. However, in DNA methylation there is usually a high degree of correlation between the methylation levels at loci in close proximity. If methylation in the parts of the CpG islands does not contribute to disease risk, then including all observations within the CpG islands may introduce additional unwanted variation. This is why most of the methods for methylation data aim to identify subregions within the units of analysis.

Another essential difference is whether the observed variants themselves are collapsed or whether the corresponding summary statistics are collapsed. For rare variants, usually the genotypes are aggregated, whereas for most DMR methods the test statistics are aggregated. It is often more computationally feasible to aggregate the individual test statistics. However, for rare variant analysis, the power benefit may be marginal when already underpowered test statistics are aggregated, instead of aggregating the individual genotypes. Since power concerns are the main reason to aggregate variants, it is more sensible to aggregate the individual genotypes when considering rare variants.

When each unit or gene has been analyzed without considering any sub-units, the number of tests will be limited to the number of units. However, when subunits are considered as well, the number of possible test will increase with an increasing number of units. This issue can be approached in different ways. For our method described in **Paper III**, we determined the threshold for the sliding windows depending on the number of possible windows in the analysis. This should have the same effect as correcting for multiple testing.

VI.4. Results

Two published studies of DNA methylation in CD4^{+} T cells in MS have yielded conflicting results (Baranzini *et al.*, 2010; Graves *et al.*, 2013). The findings by Graves *et al.* could not be replicated in our study. This was mainly because four-fifths of their top hits were removed in our study due to technical reliability issues. The remaining observations were not genome-wide significant. Additionally, some of the patients included in Graves *et al.*'s study were on immunotherapeutic treatments, which may have confounded the analysis. Since the technology used in the study by Liggett *et al.* and the one used in **Paper I** differ, and there were no overlapping loci, the two studies could not be compared (Liggett *et al.*, 2010). Since the measurement of DNA methylation used in the **Paper I** study cannot distinguish between methylation and hydroxymethylation, we were not able to investigate any of the findings reported by Calabrese *et al.* (Calabrese *et al.*, 2014). If hydroxymethylation in MS cases differs from the controls but the total methylation levels (methylation plus hydroxymethylation) are the same, our study would not have observed any differences in methylation between the MS cases and controls.

Our findings reported in **Paper I** are in agreement with those from the study by Baranzini *et al.*, who found no large difference in DNA methylation in CD4$^+$ T cells between monozygotic twin pairs discordant for MS.

A substantial part (~84%) of the variants on the Exome Chip is below 0.5% in allele frequency, which is much lower than used in earlier power studies of rare variants. It is evident from **Paper II** that the dominance of the low frequencies caused a great reduction in power compared to earlier publications on rare variants with higher allele frequencies, such as that by Basu *et al.* or Lin (Basu & Pan, 2011; Lin, 2014). As we show in **Paper II**, the high number of very low frequency variants poses challenges that have not been covered in earlier publications. To our knowledge, this is the first published study that specifically investigated the performance of the Exome Chip.

In **Paper III**, our method built upon the method presented by Zhang (Zhang, 2008), with some extensions and a new application area. Earlier publications on DMR calling can roughly be divided into two groups—static and dynamic aggregation methods—with Bumphunter and DMRcate falling in the latter group.

Both DMRcate and Bumphunter provide some way of adjusting for multiple testing. In DMRcate, the region-wise p-values are calculated using Stouffer's method. This method is closely related to Fisher's method for combining p-values, and may also suffer from the same problems when combining dependent p-values by giving an inflated combined p-value. Calculating p-values using a permutation test, as is done in Bumphunter, may also inflate the estimated p-value if the pre-selection step when identifying the regions is not accounted for.

None of the methods for DMR calling starts with the aim of identifying DMRs adjusted for multiple testing, but rather by identifying DMRs and treating the issue of multiple testing in a somewhat ad hoc manner. Our method starts with the premise that post hoc adjustment for multiple testing may not be sufficient, and that correction for multiple testing should be considered in all steps of the DMR calling.

To our knowledge, there have not been any publications dedicated only to benchmark the different methods for DMR calling. In our paper, discrepancy in performance between Bumphunter and DMRcate was much smaller than in the article by Peters *et al.*, which also introduces the DMRcate method (Peters *et al.*, 2015).

In **Paper I**, we explain how, after stringent quality control, we did not identify any CpG sites with genome-wide significant association with MS. The main reasons for this negative finding are probably either the lack of power due to small sample size or the absence of a true biological signal at the observed loci. There are approximately 28 million CpG sites in the human genome, and only a small fraction of the CpG sites were targeted on the 450k array (Smith & Meissner, 2013), leaving the possibility that many important disease loci could have been overlooked in this study.

While smaller pilot studies can give a good indication of potential genetic or epigenetic variation that is important for a disease, they are also much more prone to variability and chance findings. The observed hypermethylation in CD8$^+$ T cells may be such a chance finding, and should be replicated in an independent cohort before it is considered as a significant result. This hypermethylation could not be observed in whole blood or in the CD4$^+$ T cells. In the case of whole blood, the likely reason for this is that whole blood is a heterogeneous tissue consisting of many different cell types, of which only a few may be relevant in MS pathology.

In **Paper II**, we describe how, using a large, homogenous simulated cohort, we estimated that for small effect sizes such as 0.5% population attributable risk on the causal variants, a cohort of approximately 20,000 individuals would be needed to obtain sufficient power. We estimated that in the presence of non-causal variants within the causal gene, a cohort larger than 30,000 individuals is needed to obtain similar power compared to the scenario that all variants within a gene were causal.

Figure 1D in **Paper II** shows the QQ-plot for a simulation with 0.3% population attributable risk on the causal genes. For the estimated p-values of the causal genes, SKAT are completely detached from the non-causal genes. This suggests that a population attributable risk of 0.3% on all variants within a causal gene is an unrealistically large effect size for most association studies with rare variants.

There is no common standard for how to add effect sizes to variants in genetic simulations, and two common ways are either risk ratio modeling or odds ratio modeling. For the study reported in **Paper II**, we used risk ratio modeling since it can be formulated with only one free parameter, which can be seen as the overall effect size. Modulating the risk of each variant based on the genotype relative risk was quite similar to the weighting we used in SKAT. This may explain some of the discrepancy between SKAT and WSS shown in Figure 2 in **Paper II**.

In studies of many complex diseases, the Exome Chips are still used and have resulted in some new associations (Huyghe *et al.*, 2013; Peloso *et al.*, 2014; Igartua *et al.*, 2015; Jackson *et al.*, 2016). However, sample size and power should be carefully considered when designing studies around this genotyping chip. It is also important to consider that several common disease groups were not included when the chip was designed, and applying the Exome Chip to these diseases may result in suboptimal performance.

When analyzing methylation in complex diseases, distinct peaks in the test statistics cannot be expected. Thus, when applying a sliding window to identify peaks, the significance threshold for the windows needs to be quite close to observed null statistics. For thresholds that lie close to the observed test statistics, a small change in the threshold level can have dramatic effects on Type I error rate.

The p-values reported by the three methods described in **Paper III** were calculated using very different approaches. DMRcate reports the minimum and a combined p-value for each significant region. These two quantities can sometimes be difficult to interpret since they may tend to be bias towards the alternative hypothesis (Brown, 1975). Bumphunter uses permutation to determine significance for each candidate region. This is done by shuffling the case/control status and finding new regions, which are compared to the candidate regions. Bumphunter was implemented for the Illumina 450k array, which has a relative small number of probes, and the permutation method does not scale well with larger datasets.

In **Paper III**, we show that with a sliding window approach and sensible thresholds, the detections of DMRs can be drastically improved without a large increase in the Type I error rate. The thresholds for the different window sizes can also be estimated such that all significant windows are adjusted for multiple testing. We show that proper adjustment for multiple testing is not an obstacle to the identification of DMRs. When benchmarking our method against DMRcate and Bumphunter, we found that our method had the fastest convergence in power.

VII Concluding remarks and future perspectives

Our studies of epigenetic changes CD4$^+$ T cells in MS patients (**Paper I**) indicated that there are no large-scale differences in the methylation patterns between MS patients and healthy controls in this cell type. Additionally, we did not identify any methylation biomarker for MS in whole blood. The findings of hypermethylation observed in CD8$^+$ T cells open up for new research concerning DNA methylation in MS within CD8$^+$ T cells.

It can be generally assumed that aggregation methods will increase the power compared to single site tests in rare variant studies. However, the simulation of the Exome Chip variants reported in **Paper II** revealed that a substantially large number of samples are still needed to show an association for the variants on this chip. Aggregation of methylation sites in genome-wide methylome studies may also increase the power compared to single site testing. However, when this is done dynamically, there are unresolved issues with p-value construction and multiple testing penalty that need further research. In **Paper III** we show that the choice of method will have a substantial impact on power, especially for smaller effect sizes. We also show that proper correction for multiple testing did in fact increase the power compared to other similar methods.

For many autoimmune diseases, the epigenome has not been as thoroughly investigated as DNA variants. Future investigation into the epigenetics of MS in larger sample sizes may shed light on unresolved questions, such as disease prognosis and response to treatment. Some genetic risk variants between different autoimmune diseases overlap (Cotsapas *et al.*, 2011), but no such investigations have been carried out for the epigenome. Shared epigenome analysis between different autoimmune diseases could give insight into the general biology of autoimmunity.

In the studies reported in the papers included in this thesis, the genetic and epigenetic factors were analyzed separately. The need for methodology integrating both data types and other data in one analysis is imminent. An extension of the scan statistics methodology, which incorporates both EWAS and GWAS results, can be envisioned.

VIII References

Agarwala, V., J. Flannick, S. Sunyaev, T. D. C. Go and D. Altshuler (2013). "Evaluating empirical bounds on complex disease genetic architecture." Nat Genet 45(12): 1418-1427.

Aldous, D. J. (1989). Probability approximations via the Poisson clumping heuristic. New York, Springer-Verlag.

Altshuler, D., M. J. Daly and E. S. Lander (2008). "Genetic mapping in human disease." Science 322(5903): 881-888.

Andersson, U., R. McKean-Cowdin, U. Hjalmars and B. Malmer (2009). "Genetic variants in association studies - review of strengths and weaknesses in study design and current knowledge of impact on cancer risk." Acta Oncologica 48(7): 948-954.

Auer, P. L. and G. Lettre (2015). "Rare variant association studies: considerations, challenges and opportunities." Genome Med 7(1): 16.

Ban, M., S. Caillier, I. L. Mero, K. M. Myhr, E. G. Celius, *et al.* (2013). "No evidence of association between mutant alleles of the CYP27B1 gene and multiple sclerosis." Ann Neurol 73(3): 430-432.

Ban, M., A. Goris, A. R. Lorentzen, A. Baker, T. Mihalova, *et al.* (2009). "Replication analysis identifies TYK2 as a multiple sclerosis susceptibility factor." European Journal of Human Genetics 17(10): 1309-1313.

Bansal, V., O. Libiger, A. Torkamani and N. J. Schork (2010). "Statistical analysis strategies for association studies involving rare variants." Nature Reviews Genetics 11(11): 773-785.

Baranzini, S. E., J. Mudge, J. C. van Velkinburgh, P. Khankhanian, I. Khrebtukova, *et al.* (2010). "Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis." Nature 464(7293): 1351-1356.

Barizzone, N., I. Pauwels, B. Luciano, D. Franckaert, F. R. Guerini, *et al.* (2013). "No evidence for a role of rare CYP27B1 functional variations in multiple sclerosis." Ann Neurol 73(3): 433-437.

Basu, S. and W. Pan (2011). "Comparison of statistical tests for disease association with rare variants." Genet Epidemiol 35(7): 606-619.

Bell, C. G., G. A. Wilson, L. M. Butcher, C. Roos, L. Walter, *et al.* (2012). "Human-specific CpG "beacons" identify loci associated with human-specific traits and disease." Epigenetics 7(10): 1188-1199.

Berg-Hansen, P., S. Moen, H. Harbo and E. Celius (2014). "High prevalence and no latitude gradient of multiple sclerosis in Norway." Multiple Sclerosis Journal 20(13): 1780-1782.

Berger, S. L., T. Kouzarides, R. Shiekhattar and A. Shilatifard (2009). "An operational definition of epigenetics." Genes Dev 23(7): 781-783.

Bostrom, I., L. Stawiarz and A. M. Landtblom (2014). "Age-specific sex ratio of multiple sclerosis in the National Swedish MS Register (SMSreg)." Mult Scler 20(4): 513-514.

Brookes, A. J. (1999). "The essence of SNPs." Gene 234(2): 177-186.

Brooks, W. H., C. Le Dantec, J. O. Pers, P. Youinou and Y. Renaudineau (2010). "Epigenetics and autoimmunity." J Autoimmun 34(3): J207-219.

Brown, M. B. (1975). "A method for combining non-independent, one-sided tests of significance." Biometrics: 987-992.

Butcher, L. M. and S. Beck (2015). "Probe Lasso: a novel method to rope in differentially methylated regions with 450K DNA methylation data." Methods 72: 21-28.

Calabrese, R., E. Valentini, F. Ciccarone, T. Guastafierro, M. G. Bacalini, *et al.* (2014). "TET2 gene expression and 5-hydroxymethylcytosine level in multiple sclerosis peripheral blood cells." Biochim Biophys Acta 1842(7): 1130-1136.

Cotsapas, C., B. F. Voight, E. Rossin, K. Lage, B. M. Neale, *et al.* (2011). "Pervasive sharing of genetic effects in autoimmune disease." PLoS Genet 7(8): e1002254.

Craig, J. (2008). "Complex diseases: Research and applications." Nature Education 1(1): 184.

Dedeurwaerder, S., M. Defrance, E. Calonne, H. Denis, C. Sotiriou, *et al.* (2011). "Evaluation of the Infinium Methylation 450K technology." Epigenomics 3(6): 771-784.

Dickson, S. P., K. Wang, I. Krantz, H. Hakonarson and D. B. Goldstein (2010). "Rare variants create synthetic genome-wide associations." PLoS Biol 8(1): e1000294.

Du, P., X. Zhang, C. C. Huang, N. Jafari, W. A. Kibbe, *et al.* (2010). "Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis." BMC Bioinformatics 11: 587.

Dunham, I., A. Kundaje, S. F. Aldred, P. J. Collins, C. Davis, *et al.* (2012). "An integrated encyclopedia of DNA elements in the human genome." Nature 489(7414): 57-74.

Dyment, D. A., M. Z. Cader, M. J. Chao, M. R. Lincoln, K. M. Morrison, *et al.* (2012). "Exome sequencing identifies a novel multiple sclerosis susceptibility variant in the TYK2 gene." Neurology 79(5): 406-411.

Eckhardt, F., J. Lewin, R. Cortese, V. K. Rakyan, J. Attwood, *et al.* (2006). "DNA methylation profiling of human chromosomes 6, 20 and 22." Nat Genet 38(12): 1378-1385.

Exome Chip Consortium. (2011). "Exome Chip Design." from http://genome.sph.umich.edu/wiki/Exome_Chip_Design.

Fadista, J., A. K. Manning, J. C. Florez and L. Groop (2016). "The (in) famous GWAS P-value threshold revisited and updated for low-frequency variants." European Journal of Human Genetics.

Farh, K. K., A. Marson, J. Zhu, M. Kleinewietfeld, W. J. Housley, *et al.* (2015). "Genetic and epigenetic fine mapping of causal autoimmune disease variants." Nature 518(7539): 337-343.

Faye, L. L., M. J. Machiela, P. Kraft, S. B. Bull and L. Sun (2013). "Re-ranking sequencing variants in the post-GWAS era for accurate causal variant identification." PLoS Genet 9(8): e1003609.

Feil, R. and M. F. Fraga (2012). "Epigenetics and the environment: emerging patterns and implications." Nature Reviews Genetics 13(2): 97-109.

Fisher, R. A. (1930). The genetical theory of natural selection: a complete variorum edition, Oxford University Press.

Gibbs, R. A., J. W. Belmont, P. Hardenbol, T. D. Willis, F. Yu, *et al.* (2003). "The international HapMap project." Nature 426(6968): 789-796.

Gibson, G. (2011). "Rare and common variants: twenty arguments." Nat Rev Genet 13(2): 135-145.

Graves, M., M. Benton, R. Lea, M. Boyle, L. Tajouri, *et al.* (2013). "Methylation differences at the HLA-DRB1 locus in CD4$^+$ T-Cells are associated with multiple sclerosis." Mult Scler 20(8): 1033-1041.

Hackman, D. A., M. J. Farah and M. J. Meaney (2010). "Socioeconomic status and the brain: mechanistic insights from human and animal research." Nat Rev Neurosci 11(9): 651-659.

Haig, D. (2004). "The (dual) origin of epigenetics." Cold Spring Harb Symp Quant Biol 69: 67-70.

Hansen, K. D., B. Langmead and R. A. Irizarry (2012). "BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions." Genome Biol 13(10): R83.

Heijmans, B. T., E. W. Tobi, A. D. Stein, H. Putter, G. J. Blauw, *et al.* (2008). "Persistent epigenetic differences associated with prenatal exposure to famine in humans." Proceedings of the National Academy of Sciences 105(44): 17046-17049.

Hindorff, L., J. MacArthur, J. Morales, H. Junkins, P. Hall, *et al.* A Catalog of Published Genome-Wide Association Studies. www.genome.gov/gwastudies.

Hirschhorn, J. N. and M. J. Daly (2005). "Genome-wide association studies for common diseases and complex traits." Nat Rev Genet 6(2): 95-108.

Houseman, E. A., W. P. Accomando, D. C. Koestler, B. C. Christensen, C. J. Marsit, *et al.* (2012). "DNA methylation arrays as surrogate measures of cell mixture distribution." BMC Bioinformatics 13: 86.

Huang, Q. (2015). "Genetic study of complex diseases in the post-GWAS era." Journal of Genetics and Genomics 42(3): 87-98.

Huang, Y., W. A. Pastor, Y. Shen, M. Tahiliani, D. R. Liu, *et al.* (2010). "The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing." PloS one 5(1): e8888.

Hunt, K. A., V. Mistry, N. A. Bockett, T. Ahmad, M. Ban, *et al.* (2013). "Negligible impact of rare autoimmune-locus coding-region variants on missing heritability." Nature 498(7453): 232-235.

Huyghe, J. R., A. U. Jackson, M. P. Fogarty, M. L. Buchkovich, A. Stančáková, *et al.* (2013). "Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion." Nature genetics 45(2): 197-201.

Igartua, C., R. A. Myers, R. A. Mathias, M. Pino-Yanes, C. Eng, *et al.* (2015). "Ethnic-specific associations of rare and low-frequency DNA sequence variants with asthma." Nat Commun 6: 5965.

IMSGC (2013). "Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis." Nature genetics.

IMSGC and WTCCC 2 (2011). "Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis." Nature 476(7359): 214-219.

International HapMap Consortium (2005). "A haplotype map of the human genome." Nature 437(7063): 1299-1320.

International Human Genome Sequencing Consortium (2004). "Finishing the euchromatic sequence of the human genome." Nature 431(7011): 931-945.

Ionita-Laza, I., J. D. Buxbaum, N. M. Laird and C. Lange (2011). "A new testing strategy to identify rare variants with either risk or protective effect on disease." PLoS Genet 7(2): e1001289.

Jackson, V. E., I. Ntalla, I. Sayers, R. Morris, P. Whincup, *et al.* (2016). "Exome-wide analysis of rare coding variation identifies novel associations with COPD and airflow limitation in MOCS3, IFIT3 and SERPINA12." Thorax 71(6): 501-509.

Jaffe, A. E. and R. A. Irizarry (2014). "Accounting for cellular heterogeneity is critical in epigenome-wide association studies." Genome Biol 15(2): R31.

Jaffe, A. E., P. Murakami, H. Lee, J. T. Leek, M. D. Fallin, *et al.* (2012). "Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies." Int J Epidemiol 41(1): 200-209.

Jones, P. A. (2012). "Functions of DNA methylation: islands, start sites, gene bodies and beyond." Nature Reviews Genetics 13(7): 484-492.

Kennedy, G. C., H. Matsuzaki, S. Dong, W. M. Liu, J. Huang, *et al.* (2003). "Large-scale genotyping of complex DNA." Nat Biotechnol 21(10): 1233-1237.

Klein, R. J., X. Xu, S. Mukherjee, J. Willis and J. Hayes (2010). "Successes of genome-wide association studies." Cell 142(3): 350-351; author reply 353-355.

Klengel, T., J. Pape, E. B. Binder and D. Mehta (2014). "The role of DNA methylation in stress-related psychiatric disorders." Neuropharmacology 80: 115-132.

Kriaucionis, S. and N. Heintz (2009). "The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain." Science 324(5929): 929-930.

Krueger, F. and S. R. Andrews (2011). "Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications." Bioinformatics 27(11): 1571-1572.

Kucukali, C. I., M. Kurtuncu, A. Coban, M. Cebi and E. Tuzun (2015). "Epigenetics of multiple sclerosis: an updated review." Neuromolecular Med 17(2): 83-96.

Lander, E. S. (1996). "The new genomics: global views of biology." Science 274(5287): 536-539.

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nat Methods 9(4): 357-359.

Lassmann, H. (2013). "Pathology and disease mechanisms in different stages of multiple sclerosis." Journal of the neurological sciences 333(1): 1-4.

Lettre, G. (2014). "Rare and low-frequency variants in human common diseases and other complex traits." J Med Genet 51(11): 705-714.

Liggett, T., A. Melnikov, S. Tilwalli, Q. Yi, H. Chen, *et al.* (2010). "Methylation patterns of cell-free plasma DNA in relapsing-remitting multiple sclerosis." J Neurol Sci 290(1-2): 16-21.

Lin, W. Y. (2014). "Association testing of clustered rare causal variants in case-control studies." PLoS One 9(4): e94337.

Liu, D. J. and S. M. Leal (2010). "A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions." PLoS Genet 6(10): e1001156.

Liu, Y., M. J. Aryee, L. Padyukov, M. D. Fallin, E. Hesselberg, *et al.* (2013). "Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis." Nature biotechnology 31(2): 142-147.

Lu, Q. (2013). "The critical importance of epigenetics in autoimmunity." J Autoimmun 41: 1-5.

Madsen, B. E. and S. R. Browning (2009). "A groupwise association test for rare mutations using a weighted sum statistic." PLoS Genet 5(2): e1000384.

Maher, B. (2008). "The case of the missing heritability." Nature 456(7218): 18-21.

Maksimovic, J., L. Gordon and A. Oshlack (2012). "SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips." Genome Biol 13(6): R44.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, *et al.* (2009). "Finding the missing heritability of complex diseases." Nature 461(7265): 747-753.

Matern, B. (1960). "Spatial variation." Meddelanden från Statens Skogsforskningsinstitut 49(5).

Matsuzaki, H., S. L. Dong, H. Loi, X. J. Di, G. Y. Liu, *et al.* (2004). "Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays." Nature Methods 1(2): 109-111.

McClellan, J. and M. C. King (2010). "Genetic heterogeneity in human disease." Cell 141(2): 210-217.

McFarland, H. F. and R. Martin (2007). "Multiple sclerosis: a complicated picture of autoimmunity." Nat Immunol 8(9): 913-919.

Meda, F., M. Folci, A. Baccarelli and C. Selmi (2011). "The epigenetics of autoimmunity." Cell Mol Immunol 8(3): 226-236.

Mero, I. L., A. R. Lorentzen, M. Ban, C. Smestad, E. G. Celius, *et al.* (2009). "The TYK2 gene is associated with susceptibility to multiple sclerosis." Multiple Sclerosis 15(9): S186-S186.

Mero, I. L., A. R. Lorentzen, M. Ban, C. Smestad, E. G. Celius, *et al.* (2010). "A rare variant of the TYK2 gene is confirmed to be associated with multiple sclerosis." European Journal of Human Genetics 18(4): 502-504.

Milo, R. and E. Kahana (2010). "Multiple sclerosis: geoepidemiology, genetics and the environment." Autoimmunity reviews 9(5): A387-A394.

Modin, H., T. Masterman, T. Thorlacius, M. Stefansson, A. Jonasdottir, *et al.* (2003). "Genome-wide linkage screen of a consanguineous multiple sclerosis kinship." Mult Scler 9(2): 128-134.

Neale, B. M., Y. Kou, L. Liu, A. Ma'Ayan, K. E. Samocha, *et al.* (2012). "Patterns and rates of exonic de novo mutations in autism spectrum disorders." Nature 485(7397): 242-245.

Neale, B. M., M. A. Rivas, B. F. Voight, D. Altshuler, B. Devlin, *et al.* (2011). "Testing for an unusual distribution of rare variants." PLoS Genet 7(3): e1001322.

Oksenberg, J. R., S. E. Baranzini, S. Sawcer and S. L. Hauser (2008). "The genetics of multiple sclerosis: SNPs to pathways to pathogenesis." Nat Rev Genet 9(7): 516-526.

Orton, S. M., B. M. Herrera, I. M. Yee, W. Valdar, S. V. Ramagopalan, *et al.* (2006). "Sex ratio of multiple sclerosis in Canada: a longitudinal study." Lancet Neurol 5(11): 932-936.

Ott, J. (1999). Analysis of human genetic linkage, JHU Press.

Paul, D. S. and S. Beck (2014). "Advances in epigenome-wide association studies for common diseases." Trends Mol Med 20(10): 541-543.

Pe'er, I., R. Yelensky, D. Altshuler and M. J. Daly (2008). "Estimation of the multiple testing burden for genomewide association studies of nearly all common variants." Genet Epidemiol 32(4): 381-385.

Pedersen, B. S., D. A. Schwartz, I. V. Yang and K. J. Kechris (2012). "Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values." Bioinformatics 28(22): 2986-2988.

Peloso, G. M., P. L. Auer, J. C. Bis, A. Voorman, A. C. Morrison, *et al.* (2014). "Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks." The American Journal of Human Genetics 94(2): 223-232.

Peters, T. J., M. J. Buckley, A. L. Statham, R. Pidsley, K. Samaras, *et al.* (2015). "De novo identification of differentially methylated regions in the human genome." Epigenetics Chromatin 8: 6.

Petronis, A. (2010). "Epigenetics as a unifying principle in the aetiology of complex traits and diseases." Nature 465(7299): 721-727.

Pritchard, J. K. and N. J. Cox (2002). "The allelic architecture of human disease genes: common disease-common variant...or not?" Hum Mol Genet 11(20): 2417-2423.

R Core Team (2012). "R: A language and environment for statistical computing."

Rabbani, B., M. Tekin and N. Mahdieh (2014). "The promise of whole-exome sequencing in medical genetics." J Hum Genet 59(1): 5-15.

Rakyan, V. K., T. A. Down, D. J. Balding and S. Beck (2011). "Epigenome-wide association studies for common human diseases." Nat Rev Genet 12(8): 529-541.

Ramagopalan, S. V., D. A. Dyment, M. Z. Cader, K. M. Morrison, G. Disanto, *et al.* (2011). "Rare variants in the CYP27B1 gene are associated with multiple sclerosis." Ann Neurol 70(6): 881-886.

Reinius, L. E., N. Acevedo, M. Joerink, G. Pershagen, S. E. Dahlen, *et al.* (2012). "Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility." PLoS One 7(7): e41361.

Reinthaler, E., G. Machetanz, C. Hotzy, M. Reindl, F. Fazekas, *et al.* (2014). "No evidence for a role of rare CYP27B1 variants in Austrian multiple sclerosis patients." Mult Scler 20(3): 391-392.

Risch, N. and K. Merikangas (1996). "The future of genetic studies of complex human diseases." Science 273(5281): 1516-1517.

Russo, V. E. A., R. A. Martienssen and A. D. Riggs (1996). Epigenetic mechanisms of gene regulation. Plainview, N.Y., Cold Spring Harbor Laboratory Press.

Satterthwaite, F. E. (1946). "An Approximate Distribution of Estimates of Variance Components." Biometrics Bulletin 2(6): 110-114.

Sawcer, S., R. J. Franklin and M. Ban (2014). "Multiple sclerosis genetics." The Lancet Neurology 13(7): 700-709.

Schork, N. J., S. S. Murray, K. A. Frazer and E. J. Topol (2009). "Common vs. rare allele hypotheses for complex diseases." Curr Opin Genet Dev 19(3): 212-219.

Seunggeun, L., L. Miropolsky and M. Wu (2015). SKAT: SNP-Set (Sequence) Kernel Association Test.

Sham, P. C. and S. M. Purcell (2014). "Statistical power and significance testing in large-scale genetic studies." Nat Rev Genet 15(5): 335-346.

Siegmund, D. (1985). Sequential analysis : tests and confidence intervals. New York, Springer-Verlag.

Siegmund, D. (2013). Sequential analysis: tests and confidence intervals, Springer Science & Business Media.

Siegmund, D. and B. Yakir (2007). The statistics of gene mapping, Springer Science & Business Media.

Siegmund, D. O., N. R. Zhang and B. Yakir (2011). "False discovery rate for scanning statistics." Biometrika 98(4): 979-985.

Simpson, S., L. Blizzard, P. Otahal, I. Van der Mei and B. Taylor (2011). "Latitude is significantly associated with the prevalence of multiple sclerosis: a meta-analysis." Journal of Neurology, Neurosurgery & Psychiatry 82(10): 1132-1141.

Singmann, P., D. Shem-Tov, S. Wahl, H. Grallert, G. Fiorito, *et al.* (2015). "Characterization of whole-genome autosomal differences of DNA methylation between men and women." Epigenetics Chromatin 8: 43.

Slatkin, M. (2008). "Linkage disequilibrium--understanding the evolutionary past and mapping the medical future." Nat Rev Genet 9(6): 477-485.

Slatkin, M. (2009). "Epigenetic inheritance and the missing heritability problem." Genetics 182(3): 845-850.

Smith, Z. D. and A. Meissner (2013). "DNA methylation: roles in mammalian development." Nat Rev Genet 14(3): 204-220.

Stankiewicz, P. and J. R. Lupski (2010). "Structural variation in the human genome and its role in disease." Annual review of medicine 61: 437-455.

Stouffer, S. A., A. A. Lumsdaine, M. H. Lumsdaine, R. M. Williams Jr, M. B. Smith, *et al.* (1949). "The American soldier: combat and its aftermath.(Studies in social psychology in World War II, Vol. 2.)."

Sundqvist, E., M. Baarnhielm, L. Alfredsson, J. Hillert, T. Olsson, *et al.* (2010). "Confirmation of association between multiple sclerosis and CYP27B1." Eur J Hum Genet 18(12): 1349-1352.

Tenesa, A. and C. S. Haley (2013). "The heritability of human disease: estimation, uses and abuses." Nature Reviews Genetics 14(2): 139-149.

Teschendorff, A. E., F. Marabita, M. Lechner, T. Bartlett, J. Tegner, *et al.* (2013). "A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data." Bioinformatics 29(2): 189-196.

The 1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, *et al.* (2015). "A global reference for human genetic variation." Nature 526(7571): 68-74.

Tukey, J. W. (1953). The problem of multiple comparisons.

Visscher, P. M., M. A. Brown, M. I. McCarthy and J. Yang (2012). "Five years of GWAS discovery." Am J Hum Genet 90(1): 7-24.

Waddington, C. H. (1942). "Endeavour." 18-20.

Walsh, T., J. M. McClellan, S. E. McCarthy, A. M. Addington, S. B. Pierce, *et al.* (2008). "Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia." science 320(5875): 539-543.

Wang, D., L. Yan, Q. Hu, L. E. Sucheston, M. J. Higgins, *et al.* (2012). "IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data." Bioinformatics 28(5): 729-730.

Watson, C. T., G. Disanto, F. Breden, G. Giovannoni and S. V. Ramagopalan (2012). "Estimating the proportion of variation in susceptibility to multiple sclerosis captured by common SNPs." Sci Rep 2: 770.

Weiss, K. M. and A. G. Clark (2002). "Linkage disequilibrium and the mapping of complex human traits." Trends Genet 18(1): 19-24.

Westerlind, H., R. Ramanujam, D. Uvehag, R. Kuja-Halkola, M. Boman, *et al.* (2014). "Modest familial risks for multiple sclerosis: a registry-based study of the population of Sweden." Brain: awt356.

Willer, C. J., D. A. Dyment, S. Cherny, S. V. Ramagopalan, B. M. Herrera, *et al.* (2007). "A genome-wide scan in forty large pedigrees with multiple sclerosis." J Hum Genet 52(12): 955-962.

Wray, N. R., S. M. Purcell and P. M. Visscher (2011). "Synthetic associations created by rare variants do not explain most GWAS results." PLoS Biol 9(1): e1000579.

WTCCC (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." Nature 447(7145): 661-678.

Wu, M. C., S. Lee, T. Cai, Y. Li, M. Boehnke, *et al.* (2011). "Rare-variant association testing for sequencing data with the sequence kernel association test." Am J Hum Genet 89(1): 82-93.

Zeilinger, S., B. Kuhnel, N. Klopp, H. Baurecht, A. Kleinschmidt, *et al.* (2013). "Tobacco smoking leads to extensive genome-wide changes in DNA methylation." PLoS One 8(5): e63812.

Zhang, F. F., R. Cardarelli, J. Carroll, K. G. Fulda, M. Kaur, *et al.* (2011). "Significant differences in global genomic DNA methylation by gender and race/ethnicity in peripheral blood." Epigenetics 6(5): 623-629.

Zhang, Y. (2008). "Poisson approximation for significance in genome-wide ChIP-chip tiling arrays." Bioinformatics 24(24): 2825-2831.

Zhuang, J. C., Z. Y. Huang, G. X. Zhao, H. Yu, Z. X. Li, *et al.* (2015). "Variants of CYP27B1 are associated with both multiple sclerosis and neuromyelitis optica patients in Han Chinese population." Gene 557(2): 236-239.

Zuk, O., E. Hechter, S. R. Sunyaev and E. S. Lander (2012). "The mystery of missing heritability: Genetic interactions create phantom heritability." Proc Natl Acad Sci U S A 109(4): 1193-1198.

PAPER I

RESEARCH ARTICLE

Genome-Wide DNA Methylation Profiles Indicate CD8+ T Cell Hypermethylation in Multiple Sclerosis

Steffan D. Bos[1,2*‡], **Christian M. Page**[1,2‡], **Bettina K. Andreassen**[3,4], **Emon Elboudwarej**[5], **Marte W. Gustavsen**[1,2], **Farren Briggs**[5], **Hong Quach**[5], **Ingvild S. Leikfoss**[1,2], **Anja Bjølgerud**[1,2], **Tone Berge**[1], **Hanne F. Harbo**[1,2], **Lisa F. Barcellos**[5]

1 Department of Neurology, Oslo University Hospital, Oslo, Norway, **2** Institute of Clinical Medicine, University of Oslo, Oslo, Norway, **3** Oslo Centre for Biostatistics and Epidemiology, Department of Biostatistics, University of Oslo, Oslo, Norway, **4** Epi-Gen, Institute of Clinical Medicine, Akershus University Hospital, University of Oslo, Oslo, Norway, **5** Genetic Epidemiology and Genomics Laboratory, Division of Epidemiology, School of Public Health, University of California, Berkeley, United States of America

‡ These authors contributed equally to the manuscript.
* s.d.bos@medisin.uio.no

OPEN ACCESS

Citation: Bos SD, Page CM, Andreassen BK, Elboudwarej E, Gustavsen MW, Briggs F, et al. (2015) Genome-Wide DNA Methylation Profiles Indicate CD8+ T Cell Hypermethylation in Multiple Sclerosis. PLoS ONE 10(3): e0117403. doi:10.1371/journal.pone.0117403

Academic Editor: Angela H Ting, Cleveland Clinic Foundation, UNITED STATES

Received: July 21, 2014

Accepted: December 21, 2014

Published: March 3, 2015

Data Availability Statement: Data are publicly available in the Harvard Dataverse: doi:10.7910/DVN/27694.

Funding: Funding for this study was provided through "The Norwegian Research Council" grant ID 196776/V50 (SDB, AB, MWG, HFH), the National Institutes of Health/National Institute of Environmental Health Sciences grant ID R01 ES017080 (LB) and the Molecular Life Sciences fund at the University of Oslo (SDB, CMP). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Objective

Determine whether MS-specific DNA methylation profiles can be identified in whole blood or purified immune cells from untreated MS patients.

Methods

Whole blood, CD4+ and CD8+ T cell DNA from 16 female, treatment naïve MS patients and 14 matched controls was profiled using the HumanMethylation450K BeadChip. Genotype data were used to assess genetic homogeneity of our sample and to exclude potential SNP-induced DNA methylation measurement errors.

Results

As expected, significant differences between CD4+ T cells, CD8+ T cells and whole blood DNA methylation profiles were observed, regardless of disease status. Strong evidence for hypermethylation of CD8+ T cell, but not CD4+ T cell or whole blood DNA in MS patients compared to controls was observed. Genome-wide significant individual CpG-site DNA methylation differences were not identified. Furthermore, significant differences in gene DNA methylation of 148 established MS-associated risk genes were not observed.

Conclusion

While genome-wide significant DNA methylation differences were not detected for individual CpG-sites, strong evidence for DNA hypermethylation of CD8+ T cells for MS patients was observed, indicating a role for DNA methylation in MS. Further, our results suggest that large DNA methylation differences for CpG-sites tested here do not contribute to MS

Competing Interests: The authors have declared that no competing interests exist.

susceptibility. In particular, large DNA methylation differences for CpG-sites within 148 established MS candidate genes tested in our study cannot explain missing heritability. Larger studies of homogenous MS patients and matched controls are warranted to further elucidate the impact of CD8+ T cell and more subtle DNA methylation changes in MS development and pathogenesis.

Introduction

Multiple sclerosis (MS) is a chronic, inflammatory disease of the central nervous system (CNS) and the leading cause of disability in the young Western population[1]. The knowledge of the underlying mechanisms is sparse, but points to a complex interplay between common genetic and environmental factors. Genome-wide association studies (GWAS) and earlier genetic studies have identified 110 MS-associated loci and alleles of the *HLA-DRB1* (most frequently ∗15:01) and *HLA-A* (∗02) loci[2, 3]. Immunologically relevant genes, particularly those involved in T-helper cell differentiation, are significantly overrepresented among MS-associated variants[4]. Clinical and para-clinical evidence indicate MS results at least in part from inflammatory reactions in the CNS[5]. CD4+ T cells predominate in acute CNS lesions[6], whereas CD8+ T cells predominate in chronic lesions[7, 8], indicating an active role for these lymphocyte subclasses in MS.

Recently, epigenetic modifications have been shown to influence predisposition to complex diseases[9]. DNA methylation, the addition a methyl group to the cytosine in C-G dinucleotides (CpG-sites) modulates expression of nearby genes. DNA methylation associations have been reported for several autoimmune diseases, including Sjogren's syndrome, systemic lupus erythematous and rheumatoid arthritis[10–12]. Investigation of genome-wide DNA methylation can be performed by the Infinium HumanMethylation450 BeadChip (450K)[13]. DNA methylation of different tissues is highly diverse and influenced by environmental factors, therapy or on-going disease processes[14]. Therefore, sample homogeneity is a requirement for successful investigations of the relationship between DNA methylation and phenotypes. However, in a clinical setting heterogeneous whole blood (WB) is easily accessible for MS patients, and whether disease relevant changes can be reliably detected in WB has not been determined.

DNA methylation studies of WB, or purified blood cells from MS patients have been performed for a small number of discordant twin pairs and siblings at genome-wide scale[15], or for candidate genes and a limited numbers of CpG-sites[16, 17]. Huynh *et al.* have shown that pathogen-free brain regions of MS patients have a different global and specific DNA methylation profile as compared to healthy donor brain samples[18]. More detailed DNA methylation profile studies in carefully characterized, homogenous MS samples are highly warranted. Here we present genome-wide DNA methylation results from purified CD4+ and CD8+ T cells and WB of female MS patients and healthy controls.

Materials and Methods

Samples and genotyping

A homogenous collection of 16 untreated, female Norwegian MS patients with relapsing remitting MS (RRMS) and 14 age-matched female controls were included (Table 1). All patients and controls were of self-declared Nordic ancestry. Patients were between ages 18 and 63 and recruited from the MS clinic at the Oslo University Hospital, Oslo, Norway. Controls were

Table 1. Characteristics of individual MS patients and summaries of patients and controls.

Patient	Age category[1]	Years MS[2]	EDSS[2]	MSSS[2]	OCB[3]	MRI lesions	Contrast lesions MRI[4]
1	3	11	3.50	4.13	Yes	>20	No
2	3	11	2.00	2.11	Yes	>20	No
3	6	33	2.50	1.14	No	>20	No
4	1	2	0.00	0.53	Yes	10–20	No
5	5	1	0.00	0.64	Yes	>20	No
6	2	8	1.50	1.90	Yes	>20	No
7	2	11	1.50	1.38	No	>20	No
8	4	6	5.00	7.61	Yes	10–20	Yes
9	5	11	0.00	0.17	Yes	>20	Yes
10	3	9	1.00	0.86	Yes	>20	No
11	2	6	2.00	3.51	Yes	>20	Yes
12	4	16	1.00	0.38	Yes	10–20	No
13	2	6	1.50	2.30	Yes	>20	Yes
14	5	3	2.50	5.98	Yes	10–20	No
15	2	6	2.00	3.51	Yes	10–20	No
16	4	1	1.00	2.34	Yes	10–20	No
Summarized							
Patients Mean (S.D.; range)	38.9(25–63)	8.8(7.7; 1–33)	1.7(1.3; 0–5)	2.4(2.1; 0.2–7.6)	14/16(87.5%)	N/A	4/16(25%)
Controls Mean (S.D.; range)	39.2(28–58)	N/A	N/A	N/A	N/A	N/A	N/A

[1] Age category: 1 = 25–29, 2 = 30–34, 3 = 35–39, 4 = 40–44, 5 = 45–49, 6 = 60–64.

[2] At inclusion in this study.

[3] Oligoclonal bands present in cerebrospinal fluid taken at time of diagnosis.

[4] Contrast enhancing lesions on MRI.

Abbreviations: EDSS = Expanded Disability Status Scale, MSSS = Multiple Sclerosis Severity Score, OCB = oligoclonal bands, MRI = Magnetic Resonance Imaging, S.D. standard deviation

doi:10.1371/journal.pone.0117403.t001

recruited either through the patients or among hospital employees. None of the patients had ever received immune-modulatory drugs. Patients had not experienced a relapse or received steroids in the three months prior to enrollment and fulfilled the updated McDonald MS criteria[19]. MRI of the CNS was performed within four weeks of blood sampling and the number of lesions and contrast-enhancing lesions was counted. The Extended Disability Status Scale (EDSS) was performed on the day of blood sampling.

Genome-wide single nucleotide polymorphism (SNP) genotypes for patients and controls were assessed using the Human Omni Express BeadChip (Illumina, San Diego, CA, USA). A large Norwegian GWAS dataset published earlier[20] was used to confirm Nordic ancestry of our MS patients and controls by principal component analysis (PCA) as implemented in the R (version3.0.3) software package[21] (S1A Fig.). Genotypes were imputed against the European 1000-genomes data using IMPUTE2[22]. Details on procedures are provided in S1 Materials and Methods.

Ethics statement

The Regional Committee for Medical and Health Research Ethics South East, Norway, approved this study. Written informed consent was obtained from all study participants.

DNA methylation profiling and data normalization

CD4+ and CD8+ T cells from WB were isolated for MS patients and controls in a semi-automated manner using the autoMACS Pro Separator (Miltenyi Biotec, Germany). DNA from WB and purified CD4+ and CD8+ T cell samples was extracted and treated with bisulphite. DNA methylation levels were assessed using the 450K (Illumina, USA). Raw data were exported from Illumina's BeadStudio and normalized using the 'BMIQ' algorithm described previously[23]. Analyses were performed using beta values of methylation[24]. The CD4+ sample from donor 8 and both the CD8+ and WB sample from donor 3 had technical issues and were excluded before further analysis.

In order to prevent false positive signals due to genetic variation other than DNA methylation at probes, all probes that had an observed SNP in their target sequence (N = 60,106; see S1 Materials and Methods) in our data were removed before analysis[25] (S1B Fig.). To assess consistency of cell type specific methylation profiles, PCA of overall DNA methylation was applied (Fig. 1).

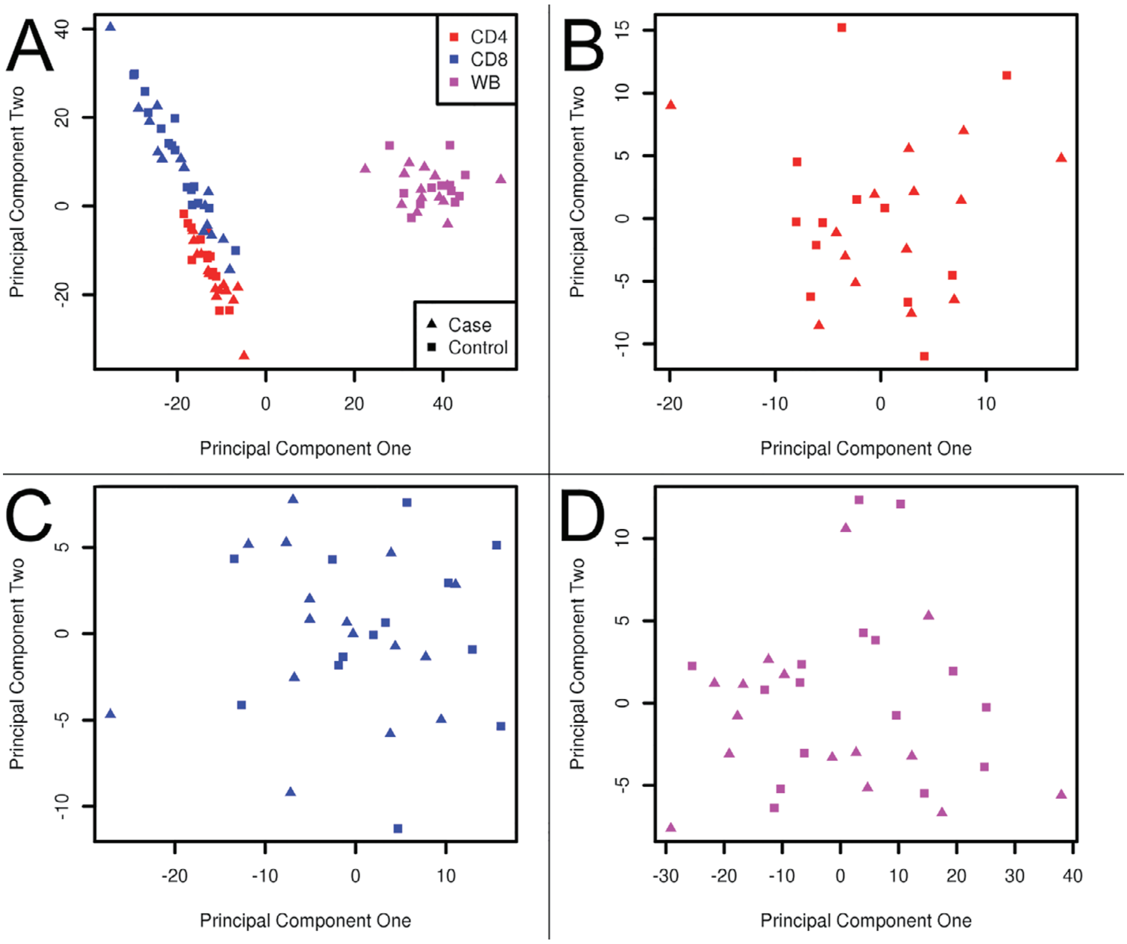


Fig 1. Principal component analyses. For samples in analyses a PCA was performed on overall methylation levels of CpG-sites that passed both quality controls and SNP filtering in (**A**) whole blood (Red), CD4+ T cells (Blue) and CD8+ T cells (Magenta) for all cases (squares) and controls (triangles). (**B**) PCA of DNA methylation data from whole blood only. (**C**) PCA of DNA methylation data from CD4+ T cells only. (**D**) PCA of DNA methylation data from CD8+ T cells only.

doi:10.1371/journal.pone.0117403.g001

To account for cellular heterogeneity of WB, we adjusted for cell type distribution in our regression models. Sample-specific estimates of the cell type proportions were obtained by adapting the algorithm from Houseman *et al.*[26] using reference information on cell-specific methylation signatures[27]. Details on the procedures above are provided in S1 Materials and Methods.

CpG-site differential methylation analysis

Two regression models were used in the analysis CpG-sites. In the first model we analyzed CD4+ T cell, CD8+ T cell or WB data separately, with 'case-control' status as a factor. Secondly, a two-way interaction model that utilized data from both CD4+ and CD8+ T cells was applied. In this model three factors were included; the 'cell type', the 'group' effect (case-control status), and an 'interaction' factor, which tested for statistical interaction between the cell type and case-control status. In case of statistical interaction between these two main factors, the DNA methylation directions are different between cell types across groups. To account for multiple testing we employed the Benjamini and Hochberg false discovery rate (FDR)[28]. CpG-sites with the lowest nominal p-values and at least 5% absolute difference in methylation [29] between MS patients and controls were examined. We examined the differences prioritized by lowest p-values to ensure the most consistently changing CpG-sites between MS cases and controls were considered. Fisher's exact test was used to test for differences in distribution of all CpG-sites that reached nominal significance.

For the 5% of probes with the lowest p-values in the CD4+ and CD8+ T cell specific analyses, we determined whether support for any observed signal was present at neighboring CpG-sites. Our approach was based on the method described recently by Jaffe *et al.*[30]. Briefly, we defined a neighbor probe to be of interest if its p-value was also in the 5% of probes with lowest p-values for the respective cell type analyses, and the maximum distance between CpG-sites was not greater than 500 base pairs. If a neighbor hit was identified the algorithm then extended over the next 500 base pairs until no additional hits were present. We then grouped these individual CpG-sites into differentially methylated regions (DMRs). By permutation testing based on the area under the curve with respect to the test statistic we calculated p-values for these DMRs.

Per-gene differential methylation analysis

The recently published list of MS-associated SNPs was used to define candidate genes (N = 148) for methylation differences given their putative role in the genetic predisposition to MS [4]. To account for multiple testing we also applied the FDR procedure[28]. CpG-sites were assigned to specific genes (N = 21,115) based on the provided Illumina manifest for the 450K. CpG-sites that mapped to multiple genes were included in analyses of all these genes. We used a permutation test based on the sum of the test statistics for each CpG-site within a gene.

Results

MS patient and control characteristics

Study characteristics are provided in Table 1. There were no significant differences between mean age or smoking status of MS patients compared to controls. All patients were diagnosed having RRMS, and the mean duration of disease was 8.8 years. The majority of patients had oligoclonal bands in their cerebrospinal fluid. All patients had modest EDSS and MSSS scores, and more than 10 typical MS lesions on cerebral MRI.

Cell type specific DNA methylation profiles

PCA analysis of the DNA methylation profiles of CD4+ and CD8+ T cells as well as WB samples identified differences in the overall DNA methylation patterns between these cell types (Fig. 1A). Within each cell type, we did not observe clustering of the MS patients and controls, indicating that on a global level there are no large, consistent DNA methylation differences that distinguish individuals according to disease status. (Fig. 1B-D)

Single CpG-site methylation analyses

In total 424,990 CpG-sites were considered after removal of CpG-sites with a low detection signal or SNPs in the probe sequence. Complete results from the per-CpG-site analysis using linear regression models are provided in S1 Table. We examined whether methylation differences observed in the T cell subsets were correlated with WB. Correlation of absolute mean differences from the WB data and either CD4+ and CD8+ T cell data was only moderate (respectively $R^2 = 0.51$ and $R^2 = 0.56$), whereas a higher correlation coefficient ($R^2 = 0.70$) was observed for CD4+ and CD8+ T cells (S1C Fig.).

The 40 CpG-sites with the lowest nominal p-values and >5% absolute difference in methylation between MS patients and controls are listed in Table 2–4. For CD4+ and CD8+ T cells we also listed whether associated CpG-sites were in a DMR as defined above. All DMRs are provided in S2 Table. Two CpG-sites occurred in the top-40 for all three analyses, both were hypermethylated in MS patients compared to controls. The first of these two probes; cg05821046, is annotated at *TMEM48*, 622 base pairs upstream from the gene transcription start site. This CpG-site is located in a DMR of three CpG-sites, which was identified in both CD4+ and CD8+ T cell analyses (S2 Table, Chr1:54304846–54305115). *TMEM48* encodes a protein involved in the nuclear pore complex formation. The second probe; cg22560193, is located in the first exon of *APC2*, a gene predicted to be involved in microtubule and beta-catenin binding. Furthermore, several CpG-sites within *DNHD1* were also among the top 40 most differentially methylated in all three datasets. This gene encodes the dynein heavy chain domain like 1, which is a protein complex that is involved in microtubule movement. We note that after adjustment for multiple testing, none of these findings reached a genome-wide significance level (lowest adjusted p-value = 0.88, S1 Table).

Interestingly, for CD8+ T cells, 38 of the 40 most differentially methylated CpG-sites (95%) showed evidence for hypermethylation in MS patients when compared to controls. The *DNHD1* gene contained one of the only two hypomethylated CpG-sites in CD8+ T cells (Table 3). In contrast, a more balanced pattern was observed for both CD4+ T cells and WB; a much lower number of CpG-sites, 55% and 52.5%, respectively showed evidence for hypermethylation in MS patients, compared to controls (Table 2 and Table 4 respectively). When considering all CpG-sites with nominal p-values below 0.05 from the patient-control comparison, the proportion of hypermethylated CD8+ T cell CpG-sites in MS patients is significantly greater than hypomethylated CpG-sites (Fisher's exact test p-value <0.01, Fig. 2A). DNA methylation of CpG-sites at different genomic features with respect to genes may provide additional insights in specific roles of the observed DNA hypermethylation in CD8+ T cells. When we considered genomic features for CpG-sites with p-values below 0.05, an overrepresentation of hypermethylated CpG-sites was slightly more frequent in 1,500 base pair regions upstream of the transcription start site (TSS-1500) and 1st exon of genes (>76% hypermethylated sites) whereas the gene body and 3'-UTR show less evidence for hypermethylation; the lowest proportion (63%) of hypermethylated CpG-sites was observed in the 3'-UTR (data not shown). Furthermore, when we compared the more recently diagnosed patients (<7 years from diagnosis) with patients diagnosed earlier (>8 years from diagnosis) the more recently diagnosed patients showed a slightly higher proportion of

DNA hypermethylation of their CD8+ T cells (proportion of hypermethylated sites 73% in recently diagnosed patients vs. 68% in the earlier diagnosed patients). We also examined CpG-sites for which patient-control comparisons did not yield p-values below 0.05, and the observation that CD8+ T cells are more likely to be hypermethylated remained, although less significant

Table 2. Top 40 results sorted by p-values from linear regression analysis models of DNA methylation in CD4+ T cells.

CD4+ T cells					
probeID[1]	**Gene**[2]	**p-value**[3]	**Effectsize**[4]	**stdev**[5]	**p-value DMR (# probes in DMR)**[6]
cg20585410	*DCX*	3.86E-05	-0.074	0.015	-
cg13988338	No gene	7.30E-05	-0.093	0.020	-
cg15552461	*RDH13*	9.58E-05	-0.069	0.015	-
cg01833234	***DNHD1***	1.49E-04	**0.145**	0.033	-
cg07937631	No gene	1.51E-04	**0.144**	0.033	-
cg24637308	***DNHD1***	1.63E-04	**0.108**	0.025	-
cg27419327	No gene	2.29E-04	-0.073	0.017	-
cg26477117	*TEKT5*	2.57E-04	-0.242	0.058	-
cg02336026	No gene	2.78E-04	-0.065	0.016	-
cg24431033	*TXNL1*	2.78E-04	**0.072**	0.017	5.5E-02 (3)
cg12543766	*MAGI2*	2.84E-04	-0.194	0.046	-
cg03700679	*TTC30B*	2.94E-04	**0.053**	0.013	-
cg06346838	***APC2***	3.88E-04	-0.062	0.015	-
cg05821046	*TMEM48*	4.03E-04	-0.065	0.016	7.0E-04 (3)
cg11213150	*ANGPTL2/RALGPS1*	4.06E-04	-0.054	0.013	-
cg08633479	*USP29*	4.11E-04	**0.066**	0.016	-
cg12243267	*USP29*	5.40E-04	**0.064**	0.016	-
cg06154311	*C20orf151*	5.68E-04	-0.075	0.019	-
cg27246129	*DLL1*	6.50E-04	-0.095	0.025	-
cg15627136	No gene	6.65E-04	-0.060	0.016	-
cg16288318	No gene	6.81E-04	-0.096	0.025	-
cg16259355	*DACH2*	7.49E-04	**0.064**	0.017	-
cg17332091	No gene	8.03E-04	-0.051	0.013	1.0E-03 (3)
cg23023970	*INPP5A*	8.82E-04	-0.061	0.016	-
cg08682625	*LOC727677*	9.72E-04	**0.116**	0.031	-
cg04587084	No gene	1.03E-03	-0.070	0.019	-
cg10208301	***DNHD1***	1.08E-03	**0.129**	0.035	-
cg07733481	*SEMA5B*	1.15E-03	**0.148**	0.041	-
cg14667685	No gene	1.34E-03	-0.078	0.022	2.0E-03 (5)
cg22560193	*APC2*	1.39E-03	-0.089	0.025	-
cg14759977	*SUGT1L1*	1.44E-03	**0.051**	0.014	-
cg01413790	No gene	1.45E-03	-0.057	0.016	1.0E-03 (3)
cg06942183	*HOXB2*	1.51E-03	**0.068**	0.019	-
cg20954971	No gene	1.53E-03	-0.067	0.019	-
cg15015426	*OR10J5*	1.64E-03	-0.074	0.021	-
cg19285525	*RBMS1*	1.65E-03	-0.395	0.113	-
cg07019386	No gene	1.66E-03	-0.080	0.023	5.0E-02 (3)
cg17976205	*C20orf151*	1.74E-03	-0.052	0.015	-
cg22687569	No gene	1.79E-03	-0.120	0.035	-

(*Continued*)

Table 2. (*Continued*)

CD4+ T cells					
probeID[1]	Gene[2]	p-value[3]	Effectsize[4]	stdev[5]	p-value DMR (# probes in DMR)[6]
cg00506935	*AEN*	1.87E-03	**0.062**	0.018	-

[1]Probe ID on 450K chip.

[2]Gene annotated to probe.

[3]p-value for specified probe in CD4+ T cells.

[4]Effect size of beta difference for specified probe. Positive values indicate hypomethylation of MS samples (i.e. controls DNA methylation higher than MS patients)

[5]Standard deviation for specified probe.

[6]Permutation-derived p-values for DMR in case the indicated probes is located in a DMR, in brackets we provided the number supportive CpG-sites in the respective DMRs.

Formatting legend

"**Bold probeID**" Specific probe occurs in all three data top-40 (see Tables 3, 4)

"***Bold Italic Gene***" Gene occurs in all three data top-40 (see Tables 3, 4)

"**Bold Effectsize**" Hypermethylation of probe in MS patients

Results shown are restricted to methylation differences of at least 5% (absolute beta difference). Full lists are provided in S1 Table.

doi:10.1371/journal.pone.0117403.t002

(Fig. 2B). For blood and CD4+ T cells, the distributions of hyper vs. hypomethylated CpG-sites were nearly identical (~50%) and not significantly different (Fig. 2A).

Methylation differences between cell types

As expected, we observed large differences in DNA methylation profiles between CD4+ and CD8+ T cells. This was illustrated by the high total number of CpG-sites showing significant differences and the large differences of beta levels for these sites. Table 5 shows the 20 most significantly different CpG-sites among cell types, adjusted for disease status and possible interaction between disease status and cell type. Among these 20 CpG-sites none showed a case-control or interaction effect in the combined model. The CpG-sites showing the greatest differences among cell types had beta differences of up to 0.85, translating to an almost full switch of methylation status. Furthermore, the genes near or containing these CpG-sites have known roles in CD4+ T cell and CD8+ T cell regulation.

MS candidate genes and exploratory per-gene analyses

Analysis of MS patients versus controls was performed at gene-level using a per-gene DNA methylation summary statistic for either CD4+ or CD8+ T cells. When considering CpG-sites annotated to genes of all established MS-associated SNPs[2], we observed no significant differences between MS patients and controls following correction for multiple testing (S3 Table). Similarly, no significant genes were observed when all genes covered by the 450K were taken into consideration (S3 Table).

Discussion

Using a robust genome-wide DNA methylation profiling approach, we show no consistent large-effect DNA methylation differences for CD4+ T cells, CD8+ T cells or WB in a homogenous collection of MS patients and controls. However, while nominally significant methylation differences were small, CD8+ T cell DNA from MS patients showed strong evidence for hypermethylation at a large number of these CpG-sites. Furthermore, we confirmed large-effect, genome-wide significant DNA methylation differences between CD4+ T cells and CD8+ T cells, underscoring the importance of separating different immune cell subpopulations in DNA methylation studies. Although none of the MS patient-control DNA methylation analyses reached genome-wide significance, we observed two CpG-sites with low p-values for all the three different sample types. We cannot exclude the possibility that genetic variation other

Table 3. Top 40 results sorted by p-values from linear regression analysis models of DNA methylation in CD8+ T cells.

CD8+ T cells					
probeID[1]	**Gene[2]**	**p-value[3]**	**Effectsize[4]**	**stdev[5]**	**p-value DMR (# probes in DMR)[6]**
cg06346838	***APC2***	2.91E-06	-0.087	0.015	-
cg22560193	***APC2***	2.16E-05	-0.101	0.020	-
cg17332091	No gene	2.22E-05	-0.066	0.013	2.0E-05 (3)
cg13988338	No gene	4.61E-05	-0.093	0.019	-
cg10673318	No gene	5.39E-05	-0.062	0.013	-
cg19432993	*HOXA2*	6.94E-05	-0.066	0.014	1.3E-02 (5)
cg21995652	*HRNBP3*	1.43E-04	-0.055	0.012	-
cg24998110	*HEXDC*	1.47E-04	**0.060**	0.014	-
cg18772882	*NTRK3*	1.74E-04	-0.051	0.012	-
cg20971998	No gene	1.79E-04	-0.078	0.018	-
cg12580893	No gene	2.00E-04	-0.066	0.015	-
cg20585410	*DCX*	2.18E-04	-0.088	0.021	-
cg13560901	*TRIL*	2.86E-04	-0.072	0.017	-
cg20864214	*ARHGEF17*	2.95E-04	-0.090	0.022	-
cg07311615	*ESPNP*	2.95E-04	-0.068	0.016	2.0E-03 (2)
cg02225599	*HOXA2*	2.99E-04	-0.064	0.016	1.3E-02 (5)
cg09309261	*LHX5*	3.68E-04	-0.063	0.016	-
cg11902995	No gene	3.80E-04	-0.063	0.016	-
cg26477117	*TEKT5*	4.59E-04	-0.241	0.061	-
cg19225422	No gene	4.80E-04	-0.052	0.013	-
cg09213964	*LRRC43*	4.82E-04	-0.051	0.013	-
cg10173124	*CYP27C1*	5.21E-04	-0.052	0.013	-
cg05821046	*TMEM48*	5.36E-04	-0.097	0.025	2.2E-01 (3)
cg18782774	No gene	5.59E-04	-0.052	0.013	-
cg24938727	*HHATL*	6.39E-04	-0.061	0.016	-
cg00402910	*AMMECR1*	6.54E-04	-0.062	0.016	-
cg08065835	No gene	6.67E-04	-0.051	0.013	-
cg04764898	*C19orf45*	6.77E-04	-0.056	0.015	-
cg21686577	*SRRM3*	6.81E-04	-0.058	0.015	-
cg08387780	No gene	6.90E-04	-0.058	0.015	2.0E-05 (3)
cg01573321	*PSD3*	7.23E-04	-0.064	0.017	-
cg14531668	No gene	7.23E-04	-0.050	0.013	-
cg22970003	*PTPRN2*	7.63E-04	-0.073	0.019	-
cg14828182	*LOC654342*	7.63E-04	-0.062	0.016	-
cg20692922	No gene	7.65E-04	-0.078	0.021	-
cg16017089	*ARHGEF17*	7.79E-04	-0.059	0.016	-
cg24637308	***DNHD1***	7.84E-04	**0.086**	0.023	-
cg09307264	*KIF1C/INCA1*	8.08E-04	-0.052	0.014	-
cg05280762	*VSIG1*	8.08E-04	-0.054	0.014	-

(*Continued*)

Table 3. (*Continued*)

CD8+ T cells

probeID[1]	Gene[2]	p-value[3]	Effectsize[4]	stdev[5]	p-value DMR (# probes in DMR)[6]
cg25512439	*CNTN4*	9.38E-04	-0.060	0.016	-

[1] Probe ID on 450K chip.

[2] Gene annotated to probe.

[3] p-value for specified probe in CD8+ T cells.

[4] Effect size of beta difference for specified probe. Positive values indicate hypomethylation of MS samples (i.e. controls DNA methylation higher than MS patients)

[5] Standard deviation for specified probe.

[6] Permutation-derived p-values for DMR in case the indicated probes is located in a DMR, in brackets we provided the number supportive CpG-sites in the respective DMRs.

Formatting legend

"**Bold probeID**" Specific probe occurs in all three data top-40 (see Tables 2, 4)

"***Bold Italic Gene***" Gene occurs in all three data top-40 (see Tables 2, 4)

"**Bold Effectsize**" Hypermethylation of probe in MS patients

Results shown are restricted to methylation differences of at least 5% (absolute beta difference). Full lists are provided in S1 Table.

doi:10.1371/journal.pone.0117403.t003

than DNA methylation could underlie such consistent results; however, given the dense genotype information we obtained, and lack of a known SNP in the probe sequences[31], our evidence strongly suggests a consistent DNA methylation difference between MS patients and controls is present. The first CpG-site, measured by probe cg05821046 resides in a DMR including two additional probes for both CD4+ and CD8+ T cells (Tables 2 and 3). The lead CpG-site is localized upstream of *TMEM48*, a gene encoding the nuclear pore complex protein NDC1. Little is known about this protein and its potential role in MS. The second consistent CpG-site difference was measured by probe cg22560193 and is annotated to the last exon of gene *APC2*. This CpG-site is not located in a DMR when considering the CpG-sites covered by the 450K. *APC2* encodes the protein adenomatosis polyposis coli 2, which is mainly expressed in neuronal tissue. The relevance of increased DNA methylation of CpG-sites within this gene in immune cells from MS patients is unclear.

Remarkably, the CD8+ T cells of MS patients showed a predominantly higher level of DNA methylation compared to controls for those CpG-sites with the lowest p-values. Since the canonical role of DNA methylation at gene promoters is gene silencing and we observed a slightly higher percentage of hypermethylated sites in these promoter regions, it is possible that gene silencing in circulating CD8+ T cells of MS patients may be present. Whether this observation persists in a larger study warrants further investigation.

After correcting for multiple testing, we did not find significant evidence for association between per-gene DNA methylation within specifically candidate genes[2], or when all genes on the 450K were considered. It is important to note that the 450K covers only a portion of the CpG-sites present in the human genome. Although the array is gene centric and largely encompasses potential regulatory regions, it is possible that MS-associated DNA methylation differences exist outside the CpG-sites covered by this array. Given the complex disease aetiology in MS, at individual patient level, changes in DNA methylation may still contribute to disease-risk.

While the sample size in this study is modest, we had at least 80% power to detect beta-value differences of 0.05 and larger, assuming per-CpG-site median standard deviations (S1D Fig.). Thus, for half of the CpG-sites, the power to detect a beta difference over 0.05 was over

80%. Therefore, our study had power to detect large-effect, consistent methylation differences between MS patients and controls. The observed hypermethylation in CD8+ T cells has small effect sizes and none of the CpG-sites reached genome-wide significance individually. A PCA of genome-wide SNP data[20] allowed us to verify Nordic ancestry and excluded systematic genetic differences between patients and controls in the study. Methylation levels for specific loci

Table 4. Top 40 results sorted by p-values from linear regression analysis models of DNA methylation in whole blood samples.

Whole Blood				
probeID[1]	**Gene[2]**	**p-value[3]**	**Effectsize[4]**	**stdev[5]**
cg16259355	*DACH2*	6.95E-05	**0.109**	0.023
cg24493834	*LAMA2*	8.65E-05	**0.059**	0.012
cg23023844	*TTLL8*	1.16E-04	**0.138**	0.030
cg04903509	*GALNT9*	1.27E-04	**0.058**	0.013
cg20373036	*POU3F4*	2.25E-04	-0.059	0.014
cg00827196	No gene	3.63E-04	-0.051	0.012
cg16288318	*No gene*	3.98E-04	-0.147	**0.035**
cg00420742	NLRP12	5.07E-04	0.051	0.013
cg02336026	No gene	5.78E-04	-0.076	0.019
cg05052271	*PLS3*	5.87E-04	-0.070	0.018
cg01262952	*ANKRD1*	5.88E-04	**0.078**	0.020
cg02313554	No gene	7.35E-04	-0.138	0.036
cg13834112	No gene	7.86E-04	-0.051	0.013
cg25031670	No gene	8.17E-04	-0.084	0.022
cg25671428	*CLSTN2*	8.26E-04	-0.051	0.013
cg05141400	*MAGEB4*	8.60E-04	-0.086	0.023
cg01281231	No gene	8.85E-04	-0.054	0.014
cg25488749	No gene	8.92E-04	-0.052	0.014
cg22560193	***APC2***	9.08E-04	-0.091	0.024
cg27571374	No gene	9.31E-04	**0.137**	0.036
cg06076512	No gene	9.76E-04	**0.054**	0.014
cg11837293	No gene	1.02E-03	**0.058**	0.015
cg02851397	*PCDHA7*	1.06E-03	-0.081	0.022
cg17140469	No gene	1.08E-03	-0.066	0.018
cg20410114	No gene	1.08E-03	**0.053**	0.014
cg11336696	*TMEM27*	1.15E-03	-0.064	0.017
cg11185456	*DNHD1*	1.19E-03	**0.152**	0.041
cg06833709	*LGI1*	1.19E-03	-0.061	0.017
cg08243619	*PTCHD2*	1.19E-03	**0.081**	0.022
cg18618432	No gene	1.22E-03	-0.382	0.104
cg25523580	*MMD2*	1.24E-03	-0.089	0.024
cg24938727	*HHATL*	1.33E-03	-0.063	0.017
cg05821046	***TMEM48***	1.37E-03	-0.087	0.024
cg00399951	*NXPH1*	1.39E-03	-0.085	0.023
cg14336566	*TDRD9*	1.44E-03	**0.072**	0.020
cg23266594	*CDX1*	1.48E-03	-0.078	0.022
cg07465864	*YTHDC2*	1.51E-03	**0.066**	0.018
cg22351833	No gene	1.52E-03	-0.069	0.019
cg02778467	*RGPD1/PLGLB2*	1.58E-03	-0.091	0.025

(*Continued*)

Table 4. (*Continued*)

Whole Blood				
probeID[1]	**Gene**[2]	**p-value**[3]	**Effectsize**[4]	**stdev**[5]
cg25584862	No gene	1.62E-03	-0.052	0.015

[1] Probe ID on 450K chip.

[2] Gene annotated to probe.

[3] p-value for specified probe in whole blood.

[4] Effect size of beta difference for specified probe. Positive values indicate hypomethylation of MS samples (i.e. controls DNA methylation higher than MS patients)

[5] Standard deviation for specified probe.

Formatting legend

"Bold probeID" Specific probe occurs in all three data top-40 (see Tables 2, 3)

"*Bold Italic Gene*" Gene occurs in all three data top-40 (see Tables 2, 3)

"Bold Effectsize" Hypermethylation of probe in MS patients

Results shown are restricted to methylation differences of at least 5% (absolute beta difference). Full lists are provided in S1 Table.

doi:10.1371/journal.pone.0117403.t004

might change with age and differ between gender[32]; therefore, only female MS patients and female, age matched controls were included in this study. The clinical data show these MS patients are representative of an average MS population with a relative benign disease course. Importantly, since medication may influence DNA methylation[33], the MS patients selected for this study had never used immune-modulatory drugs at time of sampling or received steroids for at least three months prior to inclusion. Furthermore, since tobacco smoke is a known driver of methylation differences in peripheral blood cells[34], we also performed an analysis including smoking status as a covariate; however, this did not substantially change the results (data not shown).

A recent study by Graves *et al.* reported significant DNA methylation changes within CD4+ T cells of the MHC region in MS patients using the 450K[35]. In our study, we noted 18 of 19 (95%) of these CpG- sites within the MHC were compromised by the presence of at least one

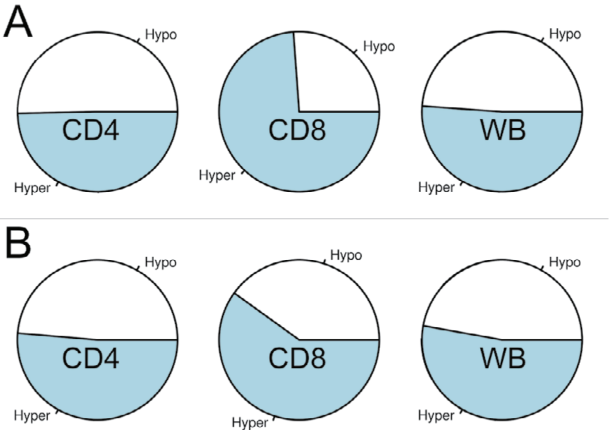


Fig 2. Pie charts of overall methylation levels for the three sample types. A. Pie-charts of DNA hyper- and hypomethylation for all CpG sites with p-values less then or equal to 0.05. **B**. Pie-charts of DNA hyper- and hypomethylation for all CpG-sites with p-values above 0.05. Abbreviations: Hypo – hypomethylation, Hyper – hypermethylation, CD4 – CD4+ T cell data, CD8 – CD8+ T cell data, WB – whole blood data.

doi:10.1371/journal.pone.0117403.g002

Table 5. Distinct differences between CD4+ and CD8+ T-cells observed in the 'cell type' term when applying a linear regression two-way interaction model including both the CD4+ and CD8+ T cell methylation data, including the terms 'cell type', 'group' (case-control status), and 'interaction' (case-control status x cell type).

				p-values[5]			
probeID[1]	Gene[2]	Effect size[3]	SD[4]	Cell Type	Cell Type BH corrected[6]	Group	Interaction
cg22505006	*ZBTB7B*	0.849	0.008	1.61E-40	6.85E-35	0.992	0.502
cg24955196	*ZBTB7B*	0.724	0.007	3.27E-40	6.94E-35	0.408	0.799
cg16871561	*SLC25A3*	0.709	0.010	4.03E-35	5.71E-30	0.918	0.290
cg25939861	*CD8A*	-0.754	0.012	1.29E-34	1.37E-29	0.314	0.824
cg06935361	*BRCA2*	-0.669	0.011	1.97E-34	1.67E-29	0.779	0.602
cg00219921	*CD8A*	-0.764	0.013	3.41E-33	2.41E-28	0.870	0.904
cg01782486	*ZBTB7B*	0.656	0.012	6.61E-33	4.01E-28	0.718	0.632
cg06449334	No gene	-0.533	0.010	2.74E-32	1.46E-27	0.299	0.557
cg25350872	*LOC154822*	-0.530	0.010	4.35E-32	1.87E-27	0.314	0.062
cg17343167	*N4BP3*	-0.448	0.009	4.82E-32	1.87E-27	0.503	0.467
cg24345747	*CD8A*	-0.638	0.012	4.85E-32	1.87E-27	0.370	0.396
cg19453665	*SERPINH1*	-0.309	0.006	9.20E-32	3.16E-27	0.392	0.891
cg03318654	*CD8A*	-0.559	0.011	9.66E-32	3.16E-27	0.408	0.947
cg03505866	*KIAA0247*	0.437	0.009	1.14E-31	3.46E-27	0.092	0.769
cg08934126	*CTNNBIP1*	-0.309	0.006	1.42E-31	4.04E-27	0.829	0.264
cg10837404	*DCP2*	0.574	0.012	2.33E-31	6.19E-27	0.460	0.357
cg26986871	No gene	-0.565	0.011	3.33E-31	7.96E-27	0.400	0.664
cg14477767	No gene	0.716	0.015	3.37E-31	7.96E-27	0.144	0.386
cg24462702	*CD40LG*	0.378	0.008	4.38E-31	9.80E-27	0.749	0.191
cg13798679	No gene	-0.446	0.010	1.22E-30	2.59E-26	0.326	0.835

[1] Probe ID on 450K chip.

[2] Gene annotated to probe.

[3] Effect size of beta difference for specified probe.

[4] standard deviation for specified probe.

[5] p-value for specified probe in respective models.

[6] Benjamini-Hochberg corrected p-values for factor "cell type".

The top 20 highest-ranking probes sorted by p-values for differences of the 'cell type' term are listed, full lists are provided in S1 Table.

doi:10.1371/journal.pone.0117403.t005

SNP in the probe sequence[25]. For the remaining CpG-site in the MHC, we did not observe a nominally significant difference. Furthermore, a SNP was present in the probes for 8 of 55 associated CpG-sites outside the MHC region. None of the remaining 47 non-MHC CpG-sites reached significance in our study. Therefore, we could not confirm the findings reported by Graves *et al.*[35]. Notably, our sample was smaller, though more clinically homogeneous with respect gender and disease course. The high number of excluded CpG-sites due to the presence of a SNP in the probe sequence underscores the need for genotype-based filtering of chip-based DNA methylation data. Alternatively, probes that might contain SNPs[25] can be identified by utilizing publicly available data[36].

Our results are in agreement with Baranzini *et al.*, who applied reduced bisulphite sequencing covering over 2 million CpG-sites, and showed no consistent large-scale methylation differences in MS discordant twins and siblings[15]. The reported switch of methylation from 20% to 80% for CpG-sites close to the *TMEM1* or *PEX14* genes between discordant twins could not be examined, since these CpG-sites are not included on the 450K.

Temporality must be considered in DNA methylation studies. It remains possible that MS patient DNA methylation profiles deviated from healthy controls at disease onset and are no longer detectable. When we consider the more recently diagnosed patients these showed a high proportion of DNA hypermethylation of their CD8+ T cells. The patients that were diagnosed earlier also show a profound DNA hypermethylation, though the proportion is slightly lower as compared to the recently diagnosed patients. We cannot exclude the possibility that the disease process in itself affects DNA methylation. This possibility must be investigated in a longitudinal cohort of MS patients.

For use as possible biomarkers of MS in the clinic, characteristic DNA methylation profiles should preferably be identified in easily obtainable WB. After correction of the WB methylation profiles in our dataset according to Houseman *et al.*[26], the correlation coefficients of WB compared to T cells remained moderate (S1C Fig.). Therefore, we cannot conclude that WB will reliably reflect disease relevant changes in T cells, however additional studies on the biomarker value of DNA methylation profiles derived from WB are warranted.

In conclusion, this is the first study of genome-wide DNA methylation profiles derived from WB, CD4+ and CD8+ T cells, in homogenous, untreated female MS patients and matched controls. We identified strong evidence for DNA hypermethylation in CD8+ T cells of MS patients. The significant methylation differences observed between CD4+ T cells, CD8+ T cells and WB underscore the importance of considering cell-based profiles. Further, more sophisticated algorithms for correction of individual variability in cell proportions are needed, if DNA methylation profiles from WB are to be used reliably. Based on available power, we excluded large-scale individual and per-gene DNA methylation differences between patients and controls, for CpG-sites tested here. In particular, large DNA methylation differences for CpG-sites within 148 established MS candidate genes tested in the current study do not explain missing heritability. Larger studies of homogenous MS patients and controls are warranted to further elucidate the impact of smaller DNA methylation changes that may be important in MS pathogenesis.

Supporting Information

S1 Fig. Supplementary figures S1A-D. A. Principal component analysis (PCA) of MS patients and controls used in the methylation analyses (respectively triangles and squares in color). The principal components for samples in current study were plotted against those derived from an earlier large GWAS study of Norwegian MS patients and controls. Results showthe samples in the DNA methylation study cluster within the Nordic population. **B**. SNPs in methylation probes influence reported beta values; example of a SNP located in the sensing probe sequence of CpG-site cg21139150 resulting correlation between reported beta-values and sample genotype. **C**. Scatterplot of –log(p-values) of the per-probe patient-control analysis for CD8+ T cell test statistics against CD4+ T cell test statistics, resulting in a correlation coefficient $R^2 = 0.70$. **D**. Post-hoc power calculations for increasing quintiles of observed probe variance.
(TIF)

S1 Materials and Methods. Detailed materials and methods for procedures briefly described in manuscript.
(DOCX)

S1 Table. Per-probe analyses details.
(ZIP)

S2 Table. All DMR analyses details.
(XLSX)

S3 Table. Per-gene analyses details.
(XLSX)

Acknowledgments

The authors acknowledge Elisabeth Gulowsen Celius, Xiaorong Shao, Gro Nygård, Aslaug Muggerud, Piotr Sowa, Dag Undlien, Anne Spurkland, Hanne Saether, Kristina Gervin and the Norwegian Metacenter for Computational Science for their contributions in patient recruitment, data collection, use of facilities and/or technical assistance.

Author Contributions

Conceived and designed the experiments: SDB BKA TB HFH LB. Performed the experiments: SDB MWG ISL AB HQ EE. Analyzed the data: SDB CMP BKA EE. Contributed reagents/materials/analysis tools: SDB CMP BKA TB HFH LB. Wrote the paper: SDB CMP BKA EE MWG FB HQ ISL AB TB HFH LB. Technical assistance and quality assurance: HQ AB.

References

1. Compston A, Coles A (2008) Multiple sclerosis. Lancet 372: 1502–1517. doi: 10.1016/S0140-6736(08)61620-7 PMID: 18970977
2. International Multiple Sclerosis Genetics Consortium (2013) Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nat Genet 45: 1353–1360. doi: 10.1038/ng.2770 PMID: 24076602
3. Gourraud PA, Harbo HF, Hauser SL, Baranzini SE (2012) The genetics of multiple sclerosis: an up-to-date review. Immunological reviews 248: 87–103. doi: 10.1111/j.1600-065X.2012.01134.x PMID: 22725956
4. International Multiple Sclerosis Genetics Consortium (2013) Network-based multiple sclerosis pathway analysis with GWAS data from 15,000 cases and 30,000 controls. Am J Hum Genet 92: 854–865. doi: 10.1016/j.ajhg.2013.04.019 PMID: 23731539
5. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, et al. (2012) Systematic localization of common disease-associated variation in regulatory DNA. Science 337: 1190–1195. doi: 10.1126/science.1222794 PMID: 22955828
6. Chitnis T (2007) The role of CD4 T cells in the pathogenesis of multiple sclerosis. International review of neurobiology 79: 43–72. PMID: 17531837
7. Huseby ES, Huseby PG, Shah S, Smith R, Stadinski BD (2012) Pathogenic CD8 T cells in multiple sclerosis and its experimental models. Frontiers in immunology 3: 64. doi: 10.3389/fimmu.2012.00064 PMID: 22566945
8. Broux B, Stinissen P, Hellings N (2013) Which immune cells matter? The immunopathogenesis of multiple sclerosis. Critical reviews in immunology 33: 283–306. PMID: 23971528
9. Nielsen HM, Tost J (2012) Epigenetic changes in inflammatory and autoimmune diseases. Sub-cellular biochemistry 61: 455–478.
10. Altorok N, Coit P, Hughes T, Koelsch KA, Stone DU, et al. (2014) Genome-Wide DNA Methylation Patterns in Naive CD4+ T Cells From Patients With Primary Sjogren's Syndrome. Arthritis & rheumatology 66: 731–739. doi: 10.1016/j.bone.2015.01.007 PMID: 25603545
11. Absher DM, Li X, Waite LL, Gibson A, Roberts K, et al. (2013) Genome-wide DNA methylation analysis of systemic lupus erythematosus reveals persistent hypomethylation of interferon genes and compositional changes to CD4+ T-cell populations. PLoS genetics 9: e1003678. doi: 10.1371/journal.pgen.1003678 PMID: 23950730
12. Whitaker JW, Shoemaker R, Boyle DL, Hillman J, Anderson D, et al. (2013) An imprinted rheumatoid arthritis methylome signature reflects pathogenic phenotype. Genome medicine 5: 40. doi: 10.1186/gm444 PMID: 23631487
13. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, et al. (2011) High density DNA methylation array with single CpG site resolution. Genomics 98: 288–295. doi: 10.1016/j.ygeno.2011.07.007 PMID: 21839163
14. Feinberg AP (2007) Phenotypic plasticity and the epigenetics of human disease. Nature 447: 433–440. PMID: 17522677

15. Baranzini SE, Mudge J, van Velkinburgh JC, Khankhanian P, Khrebtukova I, et al. (2010) Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. Nature 464: 1351–1356. doi: 10.1038/nature08990 PMID: 20428171

16. Kumagai C, Kalman B, Middleton FA, Vyshkina T, Massa PT (2012) Increased promoter methylation of the immune regulatory gene SHP-1 in leukocytes of multiple sclerosis subjects. Journal of neuroimmunology 246: 51–57. doi: 10.1016/j.jneuroim.2012.03.003 PMID: 22458980

17. Calabrese R, Zampieri M, Mechelli R, Annibali V, Guastafierro T, et al. (2012) Methylation-dependent PAD2 upregulation in multiple sclerosis peripheral blood. Multiple sclerosis 18: 299–304. doi: 10.1177/1352458511421055 PMID: 21878453

18. Huynh JL, Garg P, Thin TH, Yoo S, Dutta R, et al. (2014) Epigenome-wide differences in pathology-free regions of multiple sclerosis-affected brains. Nature neuroscience 17: 121–130. doi: 10.1038/nn.3588 PMID: 24270187

19. Polman CH, Reingold SC, Banwell B, Clanet M, Cohen JA, et al. (2011) Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. Annals of neurology 69: 292–302. doi: 10.1002/ana.22366 PMID: 21387374

20. International Multiple Sclerosis Genetics Consortium (2007) Risk alleles for multiple sclerosis identified by a genomewide study. The New England journal of medicine 357: 851–862. PMID: 17660530

21. R Core Team (2013) A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, Available: http://www.R-project.org/. doi: 10.1007/s13197-013-0993-z PMID: 25598746

22. Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS genetics 5: e1000529. doi: 10.1371/journal.pgen.1000529 PMID: 19543373

23. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J et al. (2013) A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics 29: 189–196. doi: 10.1093/bioinformatics/bts680 PMID: 23175756

24. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, et al. (2011) Evaluation of the Infinium Methylation 450K technology. Epigenomics 3: 771–784. doi: 10.2217/epi.11.105 PMID: 22126295

25. Price ME, Cotton AM, Lam LL, Farre P, Emberly E, et al. (2013) Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. Epigenetics & chromatin 6: 4. doi: 10.1038/tp.2014.145 PMID: 25603415

26. Houseman EA, Molitor J, Marsit CJ (2014) Reference-free cell mixture adjustments in analysis of DNA methylation data. Bioinformatics 30: 1431–1439. doi: 10.1093/bioinformatics/btu029 PMID: 24451622

27. Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlen SE, et al. (2012) Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PloS one 7: e41361. doi: 10.1371/journal.pone.0041361 PMID: 22848472

28. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate—a Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc B Met 57: 289–300.

29. Dayeh T, Volkov P, Salo S, Hall E, Nilsson E, et al. (2014) Genome-wide DNA methylation analysis of human pancreatic islets from type 2 diabetic and non-diabetic donors identifies candidate genes that influence insulin secretion. PLoS genetics 10: e1004160. doi: 10.1371/journal.pgen.1004160 PMID: 24603685

30. Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, et al. (2012) Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. International journal of epidemiology 41: 200–209. doi: 10.1093/ije/dyr238 PMID: 22422453

31. Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65. doi: 10.1038/nature11632 PMID: 23128226

32. Boks MP, Derks EM, Weisenberger DJ, Strengman E, Janson E, et al. (2009) The relationship of DNA methylation with age, gender and genotype in twins and healthy controls. PloS one 4: e6767. doi: 10.1371/journal.pone.0006767 PMID: 19774229

33. Szyf M (2009) Epigenetics, DNA methylation, and chromatin modifying drugs. Annual review of pharmacology and toxicology 49: 243–263. doi: 10.1146/annurev-pharmtox-061008-103102 PMID: 18851683

34. Zeilinger S, Kuhnel B, Klopp N, Baurecht H, Kleinschmidt A, et al. (2013) Tobacco smoking leads to extensive genome-wide changes in DNA methylation. PloS one 8: e63812. doi: 10.1371/journal.pone.0063812 PMID: 23691101

35. Graves M, Benton M, Lea R, Boyle M, Tajouri L, et al. (2013) Methylation differences at the HLA-DRB1 locus in CD4+ T-Cells are associated with multiple sclerosis. Multiple sclerosis 20: 1033–1041. PMID: 24336351

36. International HapMap Consortium (2003) The International HapMap Project. Nature 426: 789–796. PMID: 14685227

Paper II

PLOS | ONE

RESEARCH ARTICLE

Assessing the Power of Exome Chips

Christian Magnus Page[1,2], Sergio E. Baranzini[3], Bjørn-Helge Mevik[4], Steffan Daniel Bos[1,2], Hanne F. Harbo[1,2], Bettina Kulle Andreassen[1,5]*

1 Institute of Clinical Medicine, University of Oslo, 0316, Oslo, Norway, **2** Department of Neurology, Oslo University Hospital, 0424, Oslo, Norway, **3** Department of Neurology, University of California San Francisco, San Francisco, California, 94158, United States of America, **4** University Center for Information Technology, University of Oslo, 0316, Oslo, Norway, **5** Department of Research, Cancer Registry of Norway, 0304, Oslo, Norway

* b.k.andreassen@kreftregisteret.no

Abstract

Genotyping chips for rare and low-frequent variants have recently gained popularity with the introduction of exome chips, but the utility of these chips remains unclear. These chips were designed using exome sequencing data from mainly American-European individuals, enriched for a narrow set of common diseases. In addition, it is well-known that the statistical power of detecting associations with rare and low-frequent variants is much lower compared to studies exclusively involving common variants. We developed a simulation program adaptable to any exome chip design to empirically evaluate the power of the exome chips. We implemented the main properties of the Illumina HumanExome BeadChip array. The simulated data sets were used to assess the power of exome chip based studies for varying effect sizes and causal variant scenarios. We applied two widely-used statistical approaches for rare and low-frequency variants, which collapse the variants into genetic regions or genes. Under optimal conditions, we found that a sample size between 20,000 to 30,000 individuals were needed in order to detect modest effect sizes (0.5% < PAR > 1%) with 80% power. For small effect sizes (PAR <0.5%), 60,000–100,000 individuals were needed in the presence of non-causal variants. In conclusion, we found that at least tens of thousands of individuals are necessary to detect modest effects under optimal conditions. In addition, when using rare variant chips on cohorts or diseases they were not originally designed for, the identification of associated variants or genes will be even more challenging.

OPEN ACCESS

Citation: Page CM, Baranzini SE, Mevik B-H, Bos SD, Harbo HF, Andreassen BK (2015) Assessing the Power of Exome Chips. PLoS ONE 10(10): e0139642. doi:10.1371/journal.pone.0139642

Editor: Joseph Devaney, Children's National Medical Center, Washington, UNITED STATES

Received: June 8, 2015

Accepted: September 14, 2015

Published: October 5, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The study was funded by the University of Oslo (www.uio.no), grant number 531217/1231 (CMP); and Norwegian Research Council (NOTUR), grant number NN9311K/1734 (CMP, BKA, BHM). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Since the introduction of Genome Wide Association Studies (GWAS), a large number of common single nucleotide variants (SNVs) have successfully been associated to many complex diseases [1]. However, both the proportion of the phenotypic variability explained by these variants and the effect sizes are rather small for most studied traits. This issue has been widely discussed and is referred to as "missing heritability" [2–5]. This term suggests that genetic causes that are difficult to detect with a classic SNV array design are involved in the phenotype of interest. Such causes may be gene-gene and gene-environment interactions, chromosomal aberrations, epigenetic differences, or less frequent causal variants with minor allelic frequencies of 0.5% to 5% (low-frequency variants) or less than 0.5% (rare variants). Several of such rare and low frequent SNVs have been shown to associate with complex diseases with odds ratios (ORs) around 3 (e.g. [6–8]). Some structural variants associated to psychiatric disorders have been reported with even higher ORs (e.g. [9–12]). For example, a structural variant has been shown to give as much as a 20 fold increased risk for autism spectrum disorder [10].

The importance of considering allelic variants in coding regions, as well as budgetary and practical restrictions for whole exome sequencing in large studies, motivated the construction of the "exome chips" [13, 14]. A number of studies that used exome chips have already been published [15–29], with several of the studies reporting negative findings. However, phenotype-associations of some variants and genes have been discovered using this chip. Igartua *et. al.* [20] found one low-frequent variant associated to asthma when using a single variant test in a multi-ethnic cohort of 11,225 individuals. By using a collapsing approach (Sequence Kernel Association Test [30]), two additional genes were identified. Within a cohort of 8,229 Finnish individuals, Huyghe *et.al.* [16] identified new associations of low-frequent loci to fasting glucose levels. In a follow up case-control study by Wessel et.al., including more than 158,000 individuals, and by using statistical approaches designed for low-frequency or rare variants, one novel genetic association was discovered, driven by four rare non-synonymous SNVs within this gene [21]. With a multi-ethnic cohort of 56,000 individuals typed on the exome chip, four low-frequent variants were identified to be associated with coronary heart disease, using a single variant test. Furthermore, Tachmazidou *et.al.* identified a significant cardio-protective variant which was common in an isolated population, however this variant is assumed to be rare in outbred European populations [28].

The design of the exome chip was based on pooled exome sequencing data of 16 contributing studies [31], which comprised 12,031 individuals. These studies were highly enriched for European Americans, which accounted for approximately three-quarters of the sequenced individuals [20]. This has caused a concern concerning the generalizability of using low-frequency and rare variants in studies across populations. These variants are more likely to be evolutionary young [32], and thus, population specific. Approximately 65% of the contributing individuals were enriched for lifestyle disorders (Cardiovascular diseases, Type 2 Diabetes, Overweight, Lipid extremes, Body Mass Index extremes). Additionally, 20% of the samples were collected from psychiatric disorder cohorts (autism spectrum, schizophrenia and depression). The remaining 15% were samples from the thousand genomes project, a Sardinian cohort (SardiNIA sequencing project), and two cancer studies (S1 Table). In the design of this chip many common disease groups were absent, including autoimmune and neurodegenerative diseases. The exome chip consortia focused on capturing low-frequency and rare, non-synonymous variants, which were observed more than three times in at least two different cohorts. Most of the variants assayed on the exome chips were rare (84%), 9.2% were low-frequent, and 5.8% were common. Both the companies Illumina and Affymetrix produced a genotyping chip for low-frequent and rare, exonic variants based on the proposed list of SNVs from the Exome Chip Consortia, leading to the Illumina HumanExome BeadChip Array and the Axiom Exome Genotyping Array, respectively.

Since the power to detect an association between a single SNV and a phenotypic trait decreases with decreasing minor allelic frequency, there has been a need for new statistical tools for analysing low-frequent and rare variants. These variants often occur at different locations throughout the considered genes. Therefore, methods for this type of variants have been developed, which aim to collapse variants along a meaningful biological unit (i.e. gene, promoter, enhancer, etc.) into one test statistic. This includes methods which contrast the mean number of observed variants between cases and controls, such as Weighted Sum Statistic (WSS) [33] and

Replication Based Test (RBT) [34], or adaptive burden tests, like the Kernel Based Adaptive Clustering Method [35]. Another general class of methods comprises variance contrasting methods, which compare the variation of alleles between cases and controls, such as the C(α)–method [36] and Sequence Kernel Association Test (SKAT) [30]. While several different methods have been compared extensively (e.g.[37–42]), no single gold standard has been established. On the contrary, it is also recommended to use different kinds of methods [37, 41].

With respect to the increasing use of exome genotyping chips, we aimed to investigate the sample size requirements for association studies using these chips. The power for different statistical approaches for analysing low-frequent and rare variants has been investigated and compared to each other by others [30, 33, 37, 40, 41]. The corresponding simulations were performed for varying properties of a single unit (i.e. gene), thereby focusing on the comparison of statistical methods with respect to the detection of rare and low-frequency variants. These simulations did not take the whole variety of possible allelic frequencies into account, neither the dependencies between the variants corresponding to a real chip design. Thus, these power simulations are only representative for certain allelic frequencies, ignoring the underlying realistic allele frequency distribution and dependency patterns.

We developed a simulation pipeline, which relies on simulations based on all variants of the underlying chip design, thereby capturing the entire allele frequency spectrum and underlying dependencies between the variants. In this paper, we mimicked the structure of the Illumina HumanExome BeadChip array, but the available pipeline can also be applied to any other (future) chip designs.

Material and Methods

Simulation of genotypes

As a starting point for the simulations in this paper, we simulated a data pool of genotypes for 200,000 unrelated individuals using the approach described in Basu *et.al.*[41], with some modifications. To mimic the chip as accurately as possible, we used the publicly available allele frequencies reported by the Exome Chip Consortia. From their documentation [31], we reproduced the allele frequency of 212,353 non-synonymous SNVs, thus including 96% of the coding variants on the chip. In order to mimic the dependency structure between the variants, we applied a correlation function based on the position of the variants on the exome chip [43].

Simulation of phenotypes

To construct case-control phenotypes, we used the same approach as Madsen *et.al.* [33, 37] fixating the population attributable risk (PAR) for all variants, and calculating a genotype relative risk (GRR) based on the given PAR and the minor allele frequency (MAF) (see S1 Algorithm Eq 1). We assumed that all causal rare variants were deleterious, and that no variants had any protective effect. The probability of an individual being diseased based on their genotype, was calculated as the product of their GRRs, multiplied by a fixed incidence (see S1 Algorithm Eq 2). This was done for each individual separately. The relation between GRR and PAR is such that for a given MAF, and PAR, a linear increase in PAR corresponds to a linear increase in GRR. If the PAR was fixed, then an increase in the MAF corresponded to an inverse proportional decrease in GRR.

We considered two different scenarios for the structure of the simulated causal genes. In the first scenario, 100% of the SNVs in each analysed gene were causally linked to the phenotype. In the second scenario, the same genes were analysed, but only 50% of the SNVs within each gene where causally linked to the phenotype, thus decreasing the signal to noise ratio.

Statistical Methods

To assess the sample size required to obtain sufficient power, we applied two widely used statistical methods for rare variants: SKAT [30] and WSS [33]. In SKAT which is a generalization of the variance contrast test (C(α) method [36]), we used an adaptive weighting for each variant (the *Beta*(MAF, 0.5, 0.5) kernel). The WSS test is an adaptive sum test, for each unit, it calculates a weighted sum for all individuals, and then permutes the ranking of those sums, if the cases are consistently ranked on top, this will correspond to a low p-value. The weight for each variant is determined by the MAF and the case-control ratio. The two statistical methods used here were chosen as representatives for two common classes of methods for rare variant analysis; variance contrasting tests and sum tests. In both methods, all the genes are tested independent of each other. The weights applied to all variants have similar structure for both SKAT and WSS. In both methods, the weighing is such that common alleles will receive a low weight, while empirically rare variants will have a high weight.

Power Simulations. We investigated the power performance by drawing sample sizes of 10,000 (10k), 20k, 30k, 60k, and 100k individuals from the genotype pool described above. The simulated case-control ratio was 1:1. To assess the power under the different scenarios, we randomly selected a set of 100 genes. The distribution of allelic frequencies of this subset was similar to the corresponding allelic frequency distribution of all SNVs on the chip (S1A Fig). The mean number of SNVs per drawn gene was 18. 50 simulated datasets including 100 genes were generated for each combination of effect size, scenario and number of individuals. For each simulated dataset, the genes were tested on the Bonferroni adjusted genome-wide threshold based on the number of reproduced genes on the chip (19,975), thus neglecting findings in the genes without any simulated effect. The power was defined as the percentage of true discovered genes within one replicate. The overall power was presented as the mean power over all replications along with the empirical 95%—confidence interval.

Null simulations. We provide two types of simulations without adding an effect on the simulated genotypes (0% PAR on all causal variants). First, we aim to characterize the implemented statistical methods with respect to their ability to detect false positive findings. To achieve that, we used the 50 simulated datasets for 60k samples including 100 genes described above and assigned case-control status randomly. For each simulated dataset, we evaluate the percentage of false positives and present the mean percentage across all simulated datasets. In this simulation, we choose a 5% threshold for the p-values of each gene. A genome-wide threshold could have been simulated here as well, but would require a much larger number of null genes and thus dramatically increase the computational burden. Second, we wanted to show the genome-wide performance of the tests with no underlying effect present for the underlying chip structure considered in this paper. Thus, we simulated 10 datasets including all genes (19,975) for two different numbers of individuals (10k, 60k), assigning the case-control status arbitrarily.

The simulations and power assessment where done using the computer program R 3.2.1 [44], with the additional packages: Matrix [45], MultiPhen [46] and snpStats [47].

The simulation program can be received from the authors by request.

Results

Both of the statistical methods keep the Type-I error level, with SKAT being slightly more conservative than WSS. The corresponding estimated mean percentages of false positives were 0.0465 (SKAT) and 0.0503 (WSS) when applying the 5% Type-I error threshold (see Material and Methods). In order to understand the performance of the exome chip when no effects are present, the distribution of the p-values across all 19,975 genes was visualized in a QQ-plot

(Fig 1A and 1B). It can be seen that SKAT is more conservative, with no false positive observations, while WSS had an average of 4.7 false positive for a 10k sample, and an average of 5.4 false positive in a 60k sample, in a genome wide scan.

To assess the power of the underlying chip under the non-null distribution, we simulated an increasing effect size (PAR) for different sample sizes based on two statistical approaches

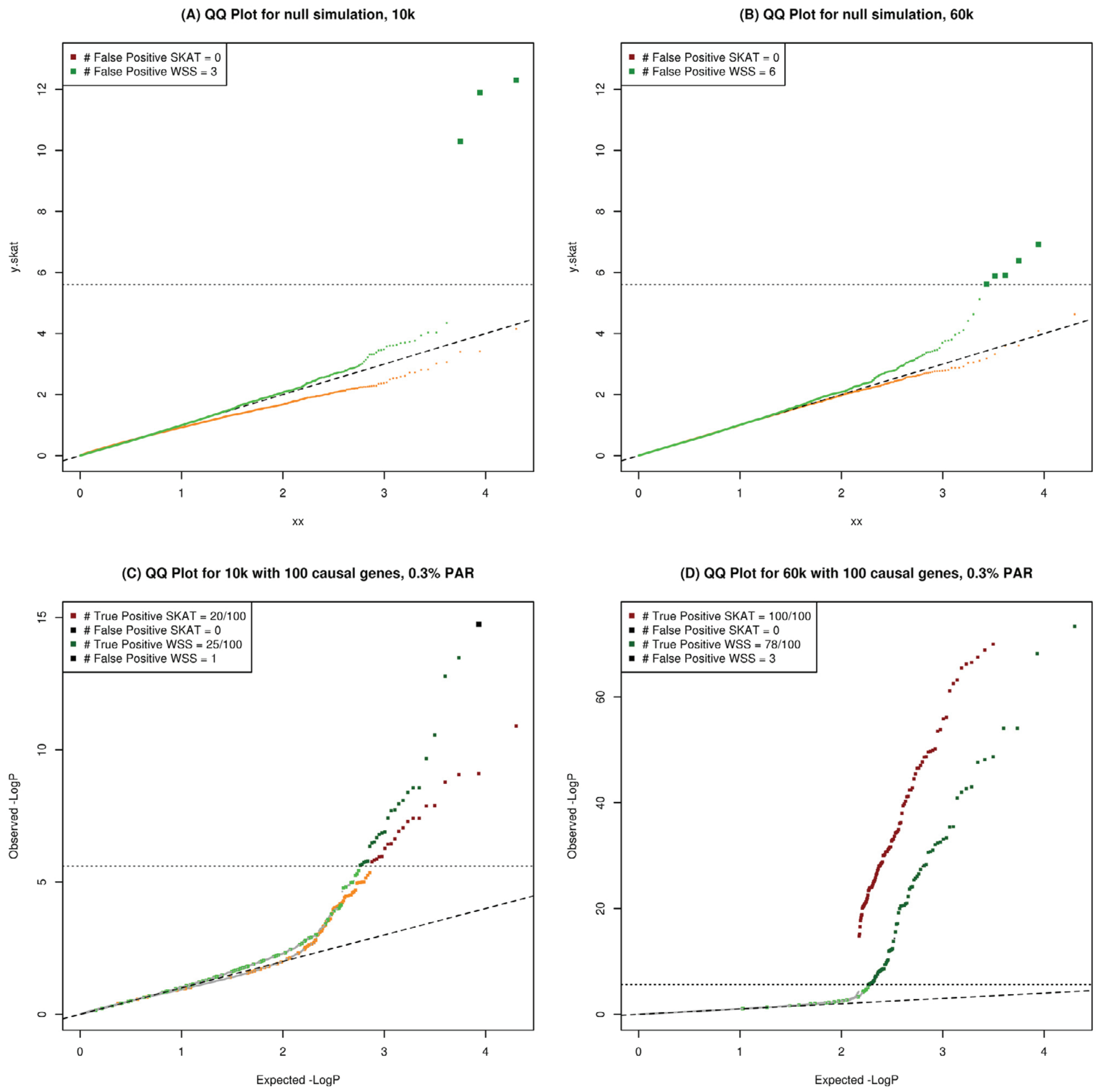


Fig 1. QQ-plot for power analysis and null simulation, the diagonal line represent the expected value and the horizontal line the Bonferroni cut-off. (A) QQ-plot for one realization of the null simulation for 10k, SKAT is plotted in red/orange and WSS in dark green/light green. (B) QQ-plot for one realization of the null simulation for 60k, SKAT is plotted in red/orange, and WSS in dark green/light green. (C) QQ-plot of—log p-values for SKAT and WSS, given 100% causal SNVs within the causal genes, and a sample size of 10k. False negative is in lighter colors (SKAT; light green, WSS; orange) and true negative is colored in gray. (D) QQ-plot of—log p-values for SKAT and WSS, given 100% causal SNVs within the causal genes, and a sample size of 60k. False negative is in lighter colors (WSS; orange) and true negative is colored in gray.

doi:10.1371/journal.pone.0139642.g001

(SKAT and WSS). We first investigated a scenario where all variants within a gene were causal. In this scenario, both SKAT and WSS reached a power of 80% with a PAR less than 1.4% (SKAT) and 2.4% (WSS) per SNV, and SKAT converged to maximum power for 1.5% for sample sizes above 10k (Fig 2A and 2B). SKAT and WSS had approximately the same speed of convergence when all variants were assigned the same weight in SKAT (data not shown). For sample sizes larger than 20k individuals, the rate of convergence of power evolved more than twice as fast in SKAT as compared to WSS. The increase in power for sample sizes above 60k individuals was marginal in SKAT. However, in WSS, the rate of convergence between the different sample sizes was more pronounced, with notable differences in convergence for sample sizes of 60k and 100k individuals. For small effect sizes ($PAR < 0.5\%$) and a sample size of 10k, WSS converged marginally faster than SKAT. To evaluate the global performance of the chip for a given PAR in this scenario, we applied both WSS and SKAT to all genes, with a sample size of 10k and 60k. The result for $PAR = 0.3\%$ on all causal variants is presented in Fig 1C and 1D. Fig 1 shows that SKAT is more conservative in its p-value estimation than WSS, both for the null simulation and with an effect size of 0.3% PAR for a sample size of 10k.

When only 50% of the SNVs within each unit were causal, a much slower convergence was observed for both methods (Fig 2C and 2D). For 10k individuals, a PAR of 6.5% was needed to obtain 80% power with SKAT, whereas WSS reached 70% power within 8.0% PAR on each causal variant. To reach 80% power with WSS within 6% PAR, a sample size of at least 30k was needed. For SKAT, a sample size of at least 30k did converge to 100% power for PAR up to 7%. This is a substantial loss of power, compared to the assumption that all SNVs within each gene were causal. In that case, half of the effect sizes were sufficient to reach the same power. When considering sample sizes larger than 10k individuals, 80% power is reached within a PAR of 1.4% for SKAT. WSS reached 80% power within 5.6% PAR for sample size of 30k. For sample sizes above 10k, SKAT converged to maximum power at 2% PAR. In WSS, the convergence was substantially slower, with none of the sample sizes converging to 100% within their tested range of PAR.

In order to assess how many individuals would be needed for a given power, we plotted power as a function of sample size (Fig 3). Under the assumption of 100% causal variants per unit, the best performance was reached with a sample size of 60k individuals or more, where both SKAT and WSS were above the 80% threshold in for the two biggest effect sizes presented ($PAR = 0.5\%$ and 1%). For WSS, a larger sample size was consistently needed to obtain the same power as SKAT in the same scenarios (Fig 3A and 3B). When 100% of the SNVs were causal, the power of WSS was comparable to the power of SKAT when only 50% of the variants were causal (Fig 3B and 3C). For effect sizes of 0.2% PAR in the 50% scenario, a sample size of 60k was sufficient to reach 80% with SKAT, but for WSS, 100k was needed (Fig 3C and 3D).

In order to investigate the relationship and distribution of the effect sizes GRR and PAR of the causal variants, we plotted a histogram of the GRR for different PAR (S1B Fig). For a fixed PAR of 0.5% on all causal variants, the GRR range in our simulated data was between 1 and 70, with a median of approximately 12. Since the GRR scales linearly for small PAR, a doubling of PAR to 1%, resulted in a doubling of the GRR. The corresponding GRR then had a maximum of 140 and a median of 24, as seen in S1B Fig.

Discussion

In this work we addressed the utility of genotyping chips for rare variants under optimal conditions, illustrated by simulating the content of the Illumina HumanExome BeadChip array

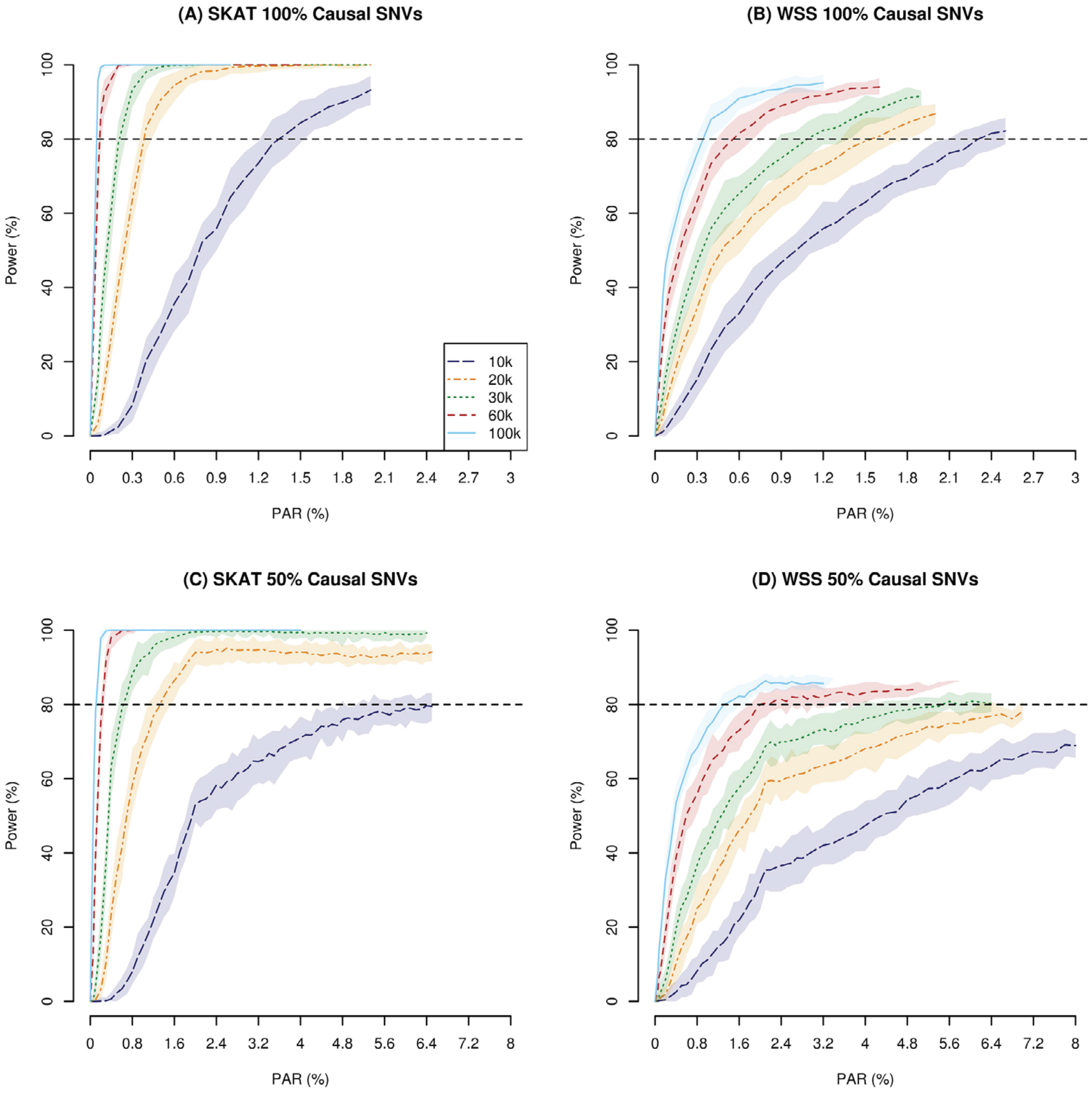


Fig 2. Power plots for increasing PAR for SKAT and WSS for multiple testing adjusted analyses, for different sample sizes. The dashed line represent the median power, with the covered area is the inter quantile range of 25% to 75% power. (A) 100% causal SNVs within all genes, estimated with SKAT. (B) 100% causal SNVs within each gene, estimated with WSS. (C) 50% causal SNVs within each gene, estimated with SKAT. (D) 50% causal SNVs within each gene, estimated with WSS.

doi:10.1371/journal.pone.0139642.g002

under different scenarios. Given a homogenous population (as was used for the design of the chip), we found that to detect a true association with 80% power, for a PAR around 1.5% on each causal variant in the presence of noise, a sample size of at least 20k individuals were needed under optimal conditions. Thus, the chip performance was acceptable for large (PAR > 1%) effects even in relatively small cohorts (10-20k). For small effect sizes

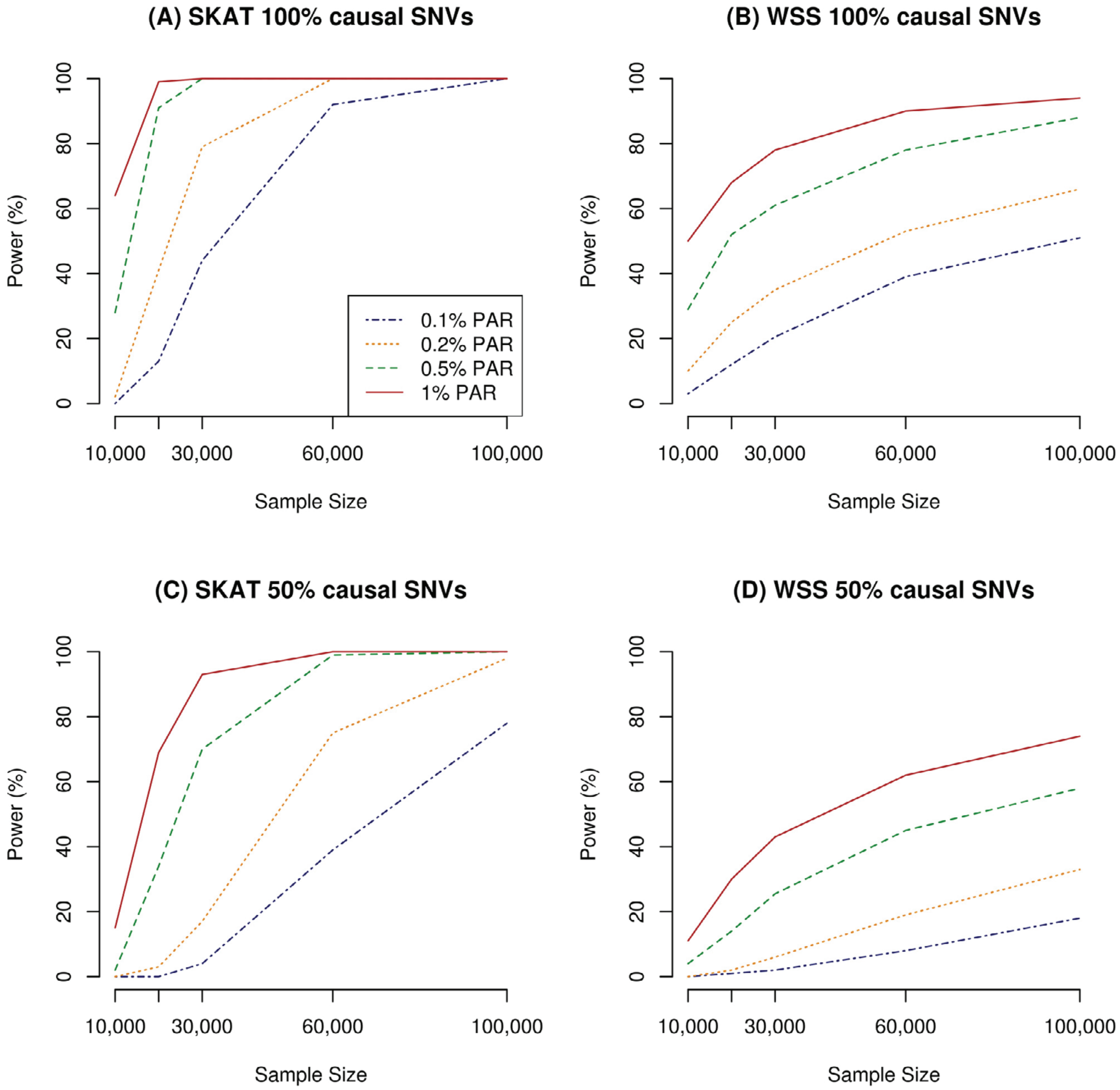


Fig 3. Power for increasing sample sizes and different PAR values after multiple testing adjusted analyses. (A) 100% causal SNVs within all genes, estimated with SKAT. (B) 100% causal SNVs within each gene, estimated with WSS. (C) 50% causal SNVs within each gene, estimated with SKAT. (D) 50% causal SNVs within each gene, estimated with WSS.

doi:10.1371/journal.pone.0139642.g003

(PAR < 0.5%) in the presence of noise, a balanced case-control study with a total sample size of 30k to 50k individuals would be required.

Our assumption of sample homogeneity of conferred risk for the SNVs in this analysis is not likely to be met in most association studies. This is mainly due to population specific rare variants. We also assumed that all rare coding causal variants were deleterious. Although some variants may be protective, the majority of rare coding alterations are believed to be either harmful or have low phenotypic effect [14], thus making our assumption a reasonable choice. We have focused our simulations on two different scenarios, one where 100% of the assayed alleles within the gene were deleterious, and the other where 50% of the alleles were deleterious. A scenario where all detected variants within a gene are causal to disease is very unlikely, but represents an upper bound on the power estimate for collapsing methods. In earlier studies which identified associated genes, the fraction of causal rare and low-frequent SNVs within the gene was estimated to be as low as 5% [29]. For the gene discovered to be associated with psychophysiological endophenotypes by Vrieze *et.al.* the association seemed to be driven by two alleles, which represent 10% of the variants in this specific gene on the exome chip [27]. One of these variants was low-frequent (MAF = 1.25%) and the other one rare (MAF = 0.3%) [27]. 40% of the rare coding exome chip variants within a gene associated to higher fasting glucose levels showed a strong individual association to this phenotype [21].

An important issue is the enrichment of SNVs associated with certain diseases in the design of the exome chip. Many common complex disease groups were not represented in the cohorts used to design the chip, leaving the possibility that rare variants which may be strong risk factors for these diseases were not included on the chip. The rest of the SNVs included may be neutral or have very small phenotypic effects. This will most likely result in sub-optimal performance in studies of diseases that were not considered when designing the exome chips.

Although the difference in mean power between SKAT and WSS analyses was not substantial, SKAT consistently outperformed WSS, which is in line with previous studies [37]. Furthermore, SKAT also outperformed WSS on elapsed computational time, where the R implementation of SKAT could benefit from parallelization on a cluster computer infrastructure.

The simulation pipeline developed here could be adapted to different chip designs. This program is only dependent on allele frequencies and positions, since the linkage disequilibrium (LD) between variants was modelled with a distance function. Furthermore, the algorithm is flexible in its implementation, so it can be applied to assess the performance of any other chip design, under different scenarios. In some simulation studies for assessment of rare variant methods, a genome wide p-value cut-off was not used [37, 41] and the allele frequency range was much wider. Simulations were performed on a single unit (i.e. genes) which included several variants, leaving out valuable information about realistic underlying allele frequencies and dependency patterns. In our study, we mimicked the properties of the exome chip, increasing both usability and reliability of our results.

There is no standard algorithm for simulating effects on genetic variants, this has led to a situation where the reported results can vary depending on the implemented methods and assumptions. Two popular approaches for emulating effect sizes are Odds Ratio (OR) modelling and Risk Ratio (RR) modelling. Although these approaches are quite different, when the number of observed genotypes is small, both the OR and the RR will be approximately the same.

When the GRR was empirically estimated from the simulated data set, they were consistently lower than expected from the equation used to generate them (S1 Algorithm Eq 1). This indicates that the GRR presented in S1B Fig was overestimated, since it is theoretically calculated, and not empirically assessed.

The collapsing methods tests each gene (unit) and the power presented in Figs 2 and 3 on the y-axis are gene-wise. However, the effect applied on the genotypes (x-axis in Fig 2), was per SNV and not per gene. When considering genes, it is important to note that the disruption of any coding element may be contributing to disease risk, and different variants within a gene can all disrupt the gene product, with observed mild effect sizes for each variant. For this reason, many different variants within the same gene may be underlying the same trait or disease. By using the collapsing statistic on a gene instead of testing individual variants, the sample sizes requirement may therefore be smaller. By selecting genes at random in the data simulation process, we study the variety of genes on the chip, without using the entire data set, thereby decreasing the computational load. Since the underlying allelic frequencies were properly presented, this gives a good indication of the overall performance of the chip, however, the actual performance for each particular gene may vary from gene to gene.

While recent studies using exome chips have identified associations between low-frequent or rare variants and disease, the identified variants have not yet contributed substantially to explaining "The missing heritability". Few of the studies have reported variants with minor allele frequency below 0.5% to be associated with disease [21, 24, 26–28]. Our study indicates that some negative reports may suffer from insufficient sample sizes and the special design of the exome chips as explained above.

In our study we have only considered "perfectly called" variants, i.e. we have not introduced any errors in the genotype calling algorithms. This may be an important issue for rare and low-frequency genotyping chips, where calling the variants has proved to be challenging [48].

In conclusion, we found that a very large sample size, in the order of tens of thousands is needed to detect modest effects under optimal conditions. For effect sizes less than 0.2% PAR, around 100,000 individuals should be studied to have enough power to reach genome wide significant results.

Supporting Information

S1 Table. Reproducing the table from the exome chip consortia [31], showing the different contributions to the design of the exome Chips.
(DOCX)

S1 Fig. (A) Histogram of the distribution of the allele frequencies on the exome chip, plotted on log 10 scale. The histogram is split into three bins, depending on their allele frequency. The lines indicated the allele frequencies of the causal alleles in each scenario. The blue line represents the allele distribution of the SNVs selected in the scenario where 100 of the SNVs within each causal gene were causal. The orange line represents the allele distribution of the causal SNVs, where 50% of the SNVs within the casual genes are causal. (B) Histogram of Genotype Relative Risk (GRR) for all causal variants in the 100% scenario, for two different Population Attributable Risks (PAR). This is the GRR used to construct the phenotypes for those two PARs.
(TIF)

S1 Algorithm. Algorithm for variant simulation and phenotype construction.
(DOCX)

S1 Source Code. R code for reproducing the variant simulation and phenotype construction.
(TAR.GZ)

Acknowledgments

We acknowledge the HPC centre at University of Oslo, for support and usage of their computer infrastructure. The authors would also like to acknowledge; Pouya Khankhanian, Lothith Madireddy both at UCSF, and Marissa LeBlanc at UiO for stimulating discussions and suggestions.

Author Contributions

Conceived and designed the experiments: CMP BKA HFH SDB SEB. Performed the experiments: CMP BHM BKA. Analyzed the data: CMP BKA. Contributed reagents/materials/analysis tools: BHM CMP. Wrote the paper: CMP SEB SDB HFH BKA.

References

1. Hindorff LA, Junkins HA, Mehta J, Manolio T. A catalog of published genome-wide association studies. National Human Genome Research Institute. 2010.
2. Maher B. Personal genomes: The case of the missing heritability. Nature. 2008; 456(7218):18–21. Epub 2008/11/07. doi: 10.1038/456018a PMID: 18987709.
3. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009; 461(7265):747–53.
4. McClellan J, King M-C. Genetic heterogeneity in human disease. Cell. 2010; 141(2):210–7. doi: 10.1016/j.cell.2010.03.032 PMID: 20403315
5. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. The American Journal of Human Genetics. 2012; 90(1):7–24. doi: 10.1016/j.ajhg.2011.11.029 PMID: 22243964
6. Gudmundsson J, Sulem P, Gudbjartsson DF, Masson G, Agnarsson BA, Benediktsdottir KR, et al. A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. Nat Genet. 2012; 44(12):1326–9. doi: 10.1038/ng.2437 PMID: 23104005; PubMed Central PMCID: PMC3562711.
7. Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet. 2011; 43(11):1066–73. doi: 10.1038/ng.952 PMID: 21983784; PubMed Central PMCID: PMC3378381.
8. Jonsson T, Atwal JK, Steinberg S, Snaedal J, Jonsson PV, Bjornsson S, et al. A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. Nature. 2012; 488(7409):96–9. doi: 10.1038/nature11283 PMID: 22801501.
9. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. science. 2008; 320 (5875):539–43. doi: 10.1126/science.1155174 PMID: 18369103
10. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature. 2012; 485(7397):242–5. doi: 10.1038/nature11011 PMID: 22495311; PubMed Central PMCID: PMC3613847.
11. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, et al. Strong association of de novo copy number mutations with autism. Science. 2007; 316(5823):445–9. doi: 10.1126/science.1138659 PMID: 17363630; PubMed Central PMCID: PMC2993504.
12. Malhotra D, McCarthy S, Michaelson JJ, Vacic V, Burdick KE, Yoon S, et al. High frequencies of de novo CNVs in bipolar disorder and schizophrenia. Neuron. 2011; 72(6):951–63. doi: 10.1016/j.neuron.2011.11.007 PMID: 22196331; PubMed Central PMCID: PMC3921625.
13. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. The American Journal of Human Genetics. 2014; 95(1):5–23. doi: 10.1016/j.ajhg.2014.06.009 PMID: 24995866
14. Lettre G. Rare and low-frequency variants in human common diseases and other complex traits. Journal of medical genetics. 2014:jmedgenet-2014-102437.
15. Chung SJ, Kim M-J, Kim J, Kim YJ, You S, Koh J, et al. Exome array study did not identify novel variants in Alzheimer's disease. Neurobiology of aging. 2014; 35(8):1958. e13–e14. doi: 10.1016/j.neurobiolaging.2014.03.007 PMID: 24685331
16. Huyghe JR, Jackson AU, Fogarty MP, Buchkovich ML, Stančáková A, Stringham HM, et al. Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. Nature genetics. 2013; 45(2):197–201. doi: 10.1038/ng.2507 PMID: 23263489
17. Holmen OL, Zhang H, Zhou W, Schmidt E, Hovelson DH, Langhammer A, et al. No large-effect low-frequency coding variation found for myocardial infarction. Human molecular genetics. 2014; 23 (17):4721–8. doi: 10.1093/hmg/ddu175 PMID: 24728188
18. Chen JA, Wang Q, Davis-Turak J, Li Y, Karydas AM, Hsu SC, et al. A Multiancestral Genome-Wide Exome Array Study of Alzheimer Disease, Frontotemporal Dementia, and Progressive Supranuclear Palsy. JAMA neurology. 2015.
19. Vrieze SI, Feng S, Miller MB, Hicks BM, Pankratz N, Abecasis GR, et al. Rare nonsynonymous exonic variants in addiction and behavioral disinhibition. Biological psychiatry. 2014; 75(10):783–9. doi: 10.1016/j.biopsych.2013.08.027 PMID: 24094508
20. Igartua C, Myers RA, Mathias RA, Pino-Yanes M, Eng C, Graves PE, et al. Ethnic-specific associations of rare and low-frequency DNA sequence variants with asthma. Nature communications. 2015; 6.

21. Wessel J, Chu AY, Willems SM, Wang S, Yaghootkar H, Brody JA, et al. Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. Nat Commun. 2015; 6:5897. doi: 10.1038/ncomms6897 PMID: 25631608; PubMed Central PMCID: PMC4311266.

22. Peloso GM, Auer PL, Bis JC, Voorman A, Morrison AC, Stitziel NO, et al. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. American journal of human genetics. 2014; 94(2):223–32. doi: 10.1016/j.ajhg.2014.01.009 PMID: 24507774; PubMed Central PMCID: PMC3928662.

23. Mauer AC, Hwang S-J, Yao J, Smith AV, Thanassoulis G, Budoff M, et al. RARE VARIANT ASSOCIATION STUDY FINDS NO LARGE-EFFECT, LOW-FREQUENCY VARIANTS FOR AORTIC AND MITRAL VALVE CALCIFICATION. Journal of the American College of Cardiology. 2015; 65(10_S).

24. Lunetta KL, Day FR, Sulem P, Ruth KS, Tung JY, Hinds DA, et al. Rare coding variants and X-linked loci associated with age at menarche. Nat Commun. 2015; 6:7756. doi: 10.1038/ncomms8756 PMID: 26239645.

25. Zuo X, Sun L, Yin X, Gao J, Sheng Y, Xu J, et al. Whole-exome SNP array identifies 15 new susceptibility loci for psoriasis. Nat Commun. 2015; 6:6793. doi: 10.1038/ncomms7793 PMID: 25854761; PubMed Central PMCID: PMC4403312.

26. Lim ET, Liu YP, Chan Y, Tiinamaija T, Karajamaki A, Madsen E, et al. A novel test for recessive contributions to complex diseases implicates Bardet-Biedl syndrome gene BBS10 in idiopathic type 2 diabetes and obesity. American journal of human genetics. 2014; 95(5):509–20. doi: 10.1016/j.ajhg.2014.09.015 PMID: 25439097; PubMed Central PMCID: PMC4225638.

27. Vrieze SI, Malone SM, Pankratz N, Vaidyanathan U, Miller MB, Kang HM, et al. Genetic associations of nonsynonymous exonic variants with psychophysiological endophenotypes. Psychophysiology. 2014; 51(12):1300–8. doi: 10.1111/psyp.12349 PMID: 25387709; PubMed Central PMCID: PMC4231532.

28. Tachmazidou I, Dedoussis G, Southam L, Farmaki AE, Ritchie GR, Xifara DK, et al. A rare functional cardioprotective APOC3 variant has risen in frequency in distinct population isolates. Nat Commun. 2013; 4:2872. doi: 10.1038/ncomms3872 PMID: 24343240; PubMed Central PMCID: PMC3905724.

29. Hallengren E, Almgren P, Engstrom G, Persson M, Melander O. Analysis of Low Frequency Protein Truncating Stop-Codon Variants and Fasting Concentration of Growth Hormone. PLoS One. 2015; 10(6):e0128348. doi: 10.1371/journal.pone.0128348 PMID: 26086970; PubMed Central PMCID: PMC4472854.

30. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. The American Journal of Human Genetics. 2011; 89(1):82–93. doi: 10.1016/j.ajhg.2011.05.029 PMID: 21737059

31. Exome Chip Consortia. Exome Chip Design http://genome.sph.umich.edu2013. Available from: http://genome.sph.umich.edu/wiki/Exome_Chip_Design.

32. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013; 493(7431):216–20. doi: 10.1038/nature11690 PMID: 23201682

33. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS genetics. 2009; 5(2):e1000384. doi: 10.1371/journal.pgen.1000384 PMID: 19214210

34. Ionita-Laza I, Buxbaum JD, Laird NM, Lange C. A new testing strategy to identify rare variants with either risk or protective effect on disease. PLoS Genet. 2011; 7(2):e1001289. doi: 10.1371/journal.pgen.1001289 PMID: 21304886; PubMed Central PMCID: PMC3033379.

35. Liu DJ, Leal SM. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. PLoS Genet. 2010; 6(10):e1001156. doi: 10.1371/journal.pgen.1001156 PMID: 20976247; PubMed Central PMCID: PMC2954824.

36. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, et al. Testing for an unusual distribution of rare variants. PLoS genetics. 2011; 7(3):e1001322. doi: 10.1371/journal.pgen.1001322 PMID: 21408211

37. Lin W-Y. Association Testing of Clustered Rare Causal Variants in Case-Control Studies. PloS one. 2014; 9(4):e94337. doi: 10.1371/journal.pone.0094337 PMID: 24736372

38. Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, et al. Pooled association tests for rare variants in exon-resequencing studies. American journal of human genetics. 2010; 86(6):832–8. doi: 10.1016/j.ajhg.2010.04.005 PMID: 20471002; PubMed Central PMCID: PMC3032073.

39. Moutsianas L, Morris AP. Methodology for the analysis of rare genetic variation in genome-wide association and re-sequencing studies of complex human traits. Briefings in functional genomics. 2014; 13(5):362–70. doi: 10.1093/bfgp/elu012 PMID: 24916163; PubMed Central PMCID: PMC4168660.

40. Larson NB, Schaid DJ. Regularized rare variant enrichment analysis for case-control exome sequencing data. Genet Epidemiol. 2014; 38(2):104–13. doi: 10.1002/gepi.21783 PMID: 24382715; PubMed Central PMCID: PMC3985431.

41. Basu S, Pan W. Comparison of statistical tests for disease association with rare variants. Genetic epidemiology. 2011; 35(7):606–19. doi: 10.1002/gepi.20609 PMID: 21769936

42. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. Nature reviews Genetics. 2014; 15(5):335–46. doi: 10.1038/nrg3706 PMID: 24739678.

43. Matérn B. Spatial variation. 1960.

44. R Core Team. R: A lanuguage and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2015.

45. Bates D, Maechler M. Matrix: Sparse and Dense Matrix Classes and Methods. R package version 11-4. 2014:http://CRAN.R-project.org/package=Matrix.

46. Coin L, O'Reilly P, Pompyen Y, Hoggart C, Calboli F. MultipPhen: MultipPhen, a packaage for the genetic association testing of multiple phenotypes. R package version 200. 2014:http://CRAN.R-project.org/package=MultiPhen.

47. Clayton D. snpStats: SnpMatrix and XSnpMatrix classes and methods. R package version 1140. 2013: http://CRAN.R-project.org/package = snpStats.

48. Grove ML, Yu B, Cochran BJ, Haritunians T, Bis JC, Taylor KD, et al. Best practices and joint calling of the HumanExome BeadChip: the CHARGE Consortium. PLoS One. 2013; 8(7):e68095. doi: 10.1371/journal.pone.0068095 PMID: 23874508

Supplement to Paper I

Blood collection

EDTA coated vacuum tubes (Greiner Bio-One, Frickenhausen, Germany) were used to collect 64 ml of whole blood, which was transferred to a 100 ml culture flask containing 1 ml of 100 mM EDTA solution (Life Technologies, Paisley, UK). The total volume of the blood was adjusted to 140 ml using RPMI culture medium (Life Technologies, Paisley, UK). Four equal volumes of diluted blood were carefully pipetted onto 15 ml of lymphoprep (Sigma-Aldrich, Oslo, Norway) in 50 ml tubes (Greiner Bio-One, Frickenhausen, Germany) and the peripheral blood mononuclear cells (PBMCs) were separated from the other blood constituents by centrifugation at 800g for 30 minutes. The PBMCs were washed three times in ice-cold PBS before suspension in 2 ml ice cold PBS.

Cell separation

PBMCs were counted using nucleocount NC-100 (ChemoMetec A/S, Denmark) and pelleted, followed by suspension in a volume of 80 µl MACS buffer per 10 million cells. Per 10 million cells, 20 µl of magnetic anti-CD8$^+$ beads (Miltenyi Biotec, Lund, Sweden) was mixed with the cells and incubated for 15 minutes at 4°C, washed with MACS buffer and resuspended in 500 µl of MACS buffer. The autoMACS cell separator was used to separate positive and negative cell fractions (positive selection) and the positive fraction was counted and kept on ice. The CD8$^+$ negative fraction was centrifuged for 10 minutes at 300g and the supernatant removed. The pellet was suspended in 30 µl per 10 million cells MACS buffer and 10 µl per 10 million cells of biotin labeled CD4$^+$ negative antibody cocktail was added and incubated for 10 minutes at 4°C. To this suspension 10 µl per million cells and 10 µl per million cells anti-biotin magnetic beads were added and incubated at 4°C for 15 minutes. The cells were washed in MACS buffer before resuspension in 500 µl of MACS buffer. The autoMACS cell separator (Miltenyi Biotec, Lund, Sweden) was used to separate positive and negative cell fractions (negative selection) and the negative fraction was counted and kept on ice, whereas the positive fraction was discarded. For each separated cell type, aliquots of 1-3 million cells were stored at -20°C.

Flow Cytometry assessment of cell fraction purities

During collecting for the majority of samples flow cytometry was performed as detailed below. All CD4$^+$ and CD8$^+$ T cell fractions demonstrated 95% or greater purity. Cells were labelled with FITC-conjugated mouse anti-human CD4 (Clone RTF-4g), mouse anti-human CD8 (clone RTF-8) or mouse IgG1 isotype control (clone 15H6) (all from Southern Biotech), and cell purity was assessed by flow cytometry on FACS Calibur (BD Biosciences) and data analysed by Cell Quest Pro (BD Biosciences). During collecting for the majority of samples flow cytometry was performed and all CD4$^+$ and CD8$^+$ T cell fractions demonstrated 95% or greater purity.

DNA isolation and QC

DNA was isolated from the cell pellets using Qiamp DNA mini kit (Qiagen, Sweden) by adding 200 µl of lysis buffer to the thawed cell pellets before following the instructions as provided by the manufacturer. DNA was quantified using the nanodrop spectrophotometer (Thermo Scientific, Wilmington, DE 19810 USA) and samples with 260/280 values below 1.7 were subjected to additional purification by precipitation and dissolving in extraction buffer.

Imputation of genotypes against the 1,000 genomes reference panel

Genotypes of the Illumina 660 Quad array were pre-phased using phase-it. Phased data was then imputed using impute2[22], applying the 1,000 genomes central European reference panel. We applied a 90% information threshold for calling genotypes of imputed SNPs.

Illumina 450K Methylation Array measurements

750 ng of the DNA was used as input for the bisulfite conversion using the Zymo EZ-96DNA Methylation Kit (Catalog #D5004) Deep-Well Format. Then 4 µl of the bisulfite converted DNA was used as input for the Illumina Infinium HD Methylation Assay according to the manufacturer's protocol. Samples are transferred to Illumina's Infinium HumanMethylation450K DNA Analysis BeadChip before scanning on the Illumina HiScan.

Once scanning was completed, the data was uploaded into GenomeStudio for preliminary analysis and QC. Target success rate was determined. The detection p-value is the 1-p computed from the background model characterizing the chance that the target sequence was distinguishable from the negative controls. CpG-sites with more than 20% missing values were excluded from the analysis (CD4^{+} T cells N=72; CD8^{+} T cells N=72; whole blood (WB) N=67; overlap for all categories N=49). Sample replicates and Jurkat cell DNA control replicates are checked to ensure an r2 value of greater than 0.99. No irregularities were observed in the on-array internal controls provided by Illumina.

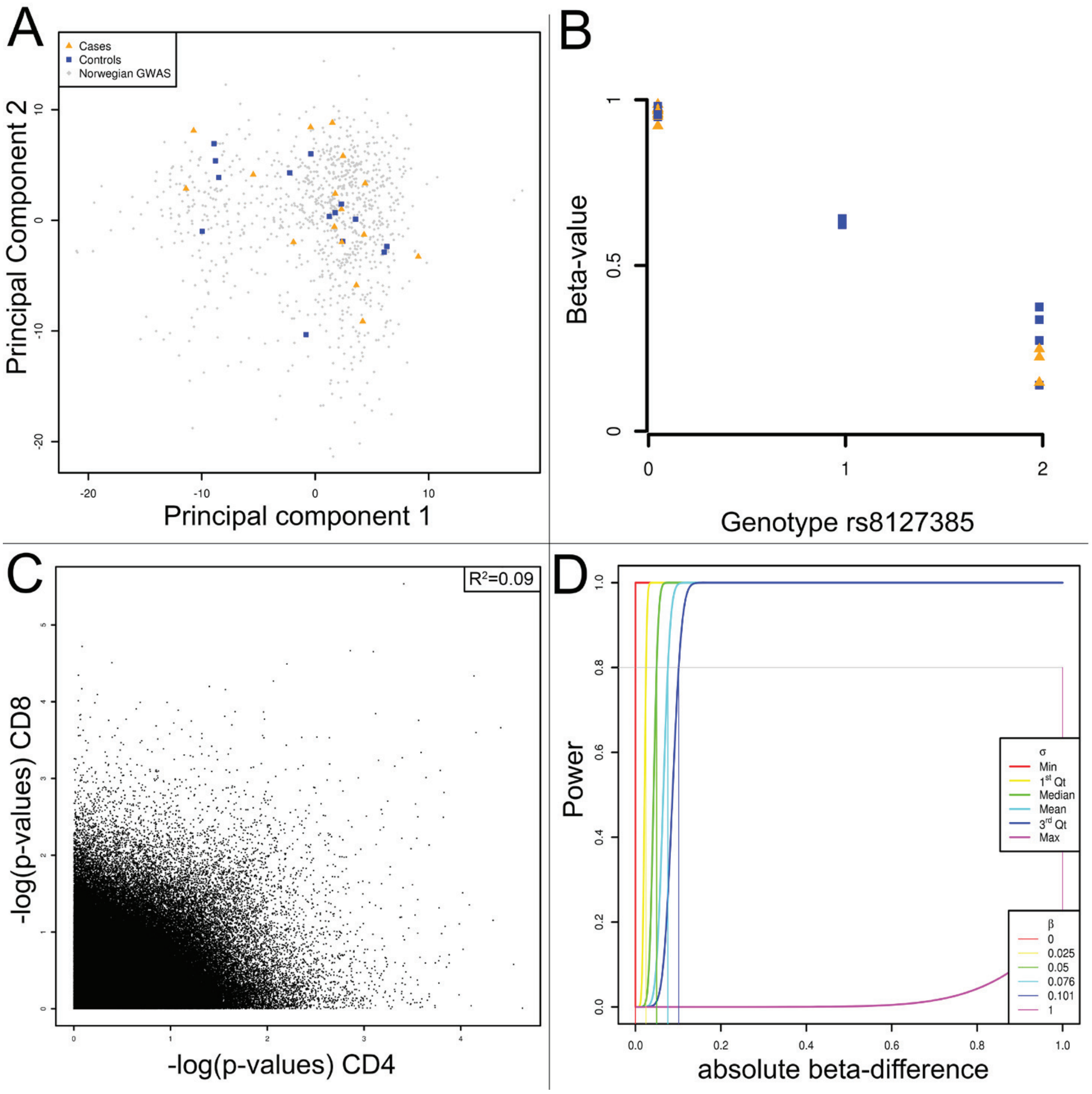
SNP assessment for probes

By overlaying the probe genomic locations as provided by Illumina with the imputed and genotyped per-sample SNP map, we identified all probes that contain at least one observed or imputed polymorph site in their sequences. These probes (N=60,106) were removed from analyses in all samples to prevent false methylation readouts (Figure S1B).

Correlation of WB, CD4^{+} T and CD8^{+} T cell DNA profiles

For each CpG-site in the WB, CD4^{+} and CD8^{+} T cells data, absolute differences in beta values were calculated. Correlations of these absolute differences were assessed for the WB to either CD4^{+} or CD8^{+} T cell data, and for CD4^{+} T cell data to CD8^{+} T cell data. The latter comparison is illustrated by Figure S1C.

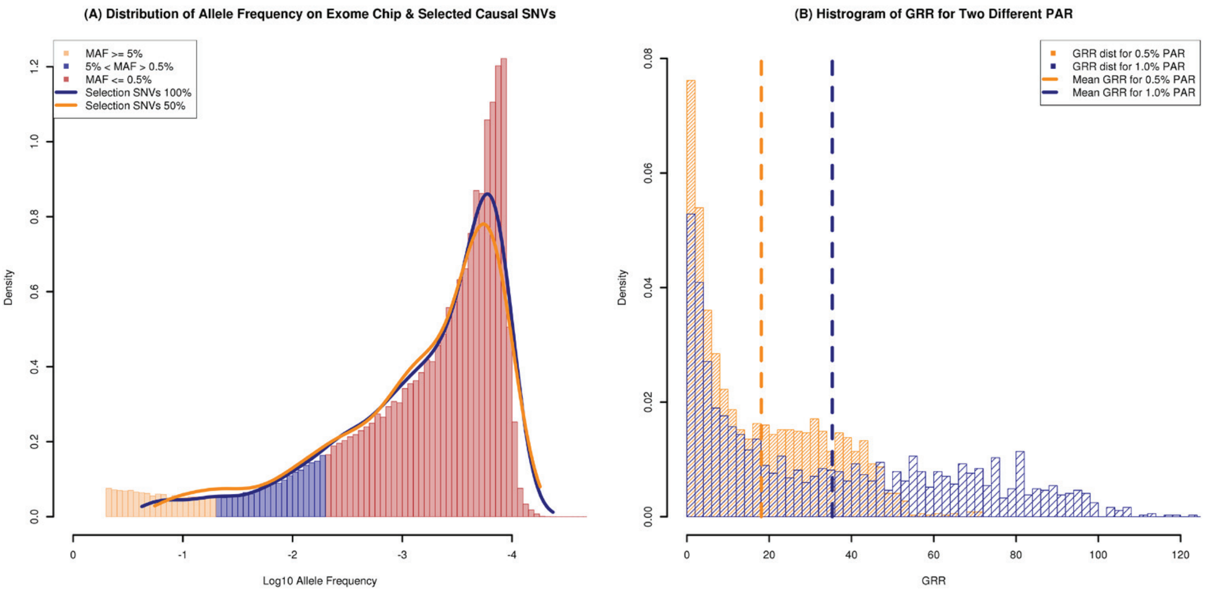
Supplemetary Figure S1



A. Principal component analysis (PCA) of MS patients and controls used in the methylation analyses (respectively triangles and squares in color). The principal components for samples in current study were plotted against those derived from an earlier large GWAS study of Norwegian MS patients and controls. Results showthe samples in the DNA methylation study cluster within the Nordic population. B. SNPs in methylation probes influence reported beta values; example of a SNP located in the sensing probe sequence of CpG-site cg21139150 resulting correlation between reported beta-values and sample genotype. C. Scatterplot of –log(p-values) of the per-probe patient-control analysis for CD8^{+} T cell test statistics against CD4^{+} T cell test statistics, resulting in a correlation coefficient R2 = 0.70. D. Post-hoc power calculations for increasing quintiles of observed probe variance.

Supplement to Paper II

Supplementary figure 1



(A) Histogram of the distribution of the allele frequencies on the exome chip, plotted on log 10 scale. The histogram is split into three bins, depending on their allele frequency. The lines indicated the allele frequencies of the causal alleles in each scenario. The blue line represents the allele distribution of the SNVs selected in the scenario where 100 of the SNVs within each causal gene were causal. The orange line represents the allele distribution of the causal SNVs, where 50% of the SNVs within the casual genes are causal. (B) Histogram of Genotype Relative Risk (GRR) for all causal variants in the 100% scenario, for two different Population Attributable Risks (PAR). This is the GRR used to construct the phenotypes for those two PARs.

Algorithm for simulation and phenotype construction:

1. We constructed a set of 2*N = 400,000 independent multivariate normally distributed vectors, of length p = 212,353 (representing each SNV). The correlation within the vectors, where modeled using the Matern covariance function, with parameters (σ, ϑ, ρ) equal to (1.9, 10, 15) respectively.
2. All the 2*N vectors were dichotomised with a threshold, using the allele frequency for each SNV, reported by the exome chip consortia. The vector now represents the independent maternal and paternal haplotypes.
3. The vectors were added together in groups of two, resulting in data set of N vectors, representing each individual, with the number of alleles (Basu and Pan 2011).
4. The alleles were now clustered together into their respective genes.

5. We selected a set of 100 genes at random to be causally linked to the phenotype.
6. Two scenarios were constructed. In the first scenario, all SNVs within the select causal genes, where themselves causal. In the second scenario, we picked 50% of the SNVs within the causal genes at random to be causally linked to the phenotype.
7. For a given Population Attributed Risks (PAR), we calculated the Genotype Relative Risk (GRR) for each causal variant, using the equation below. Observe that this equation is only dependent on the PAR and the allele frequency, which is considered the exposure of the allele. (**Equation 1**)

$$GRR_j = \frac{PAR}{(1 - PAR)MAF_j} + 1$$

8. Given the set of k GRRs for one PAR, we calculated the probability for each individuals of being a case, using the equation below. This resulted in a vector of N entries with either zero or one, corresponding to the phenotype for each individual. (**Equation 2**)

$$P(Y = 1|G = \{a\}) = b_0 \prod_{j \in \{a\}} GRR_j^{a_j}$$

9. Here b_0 is the base line population risk (incidence), k is the number of causal alleles, and a_j is the allelic count of the allele number j, {0,1,2}.
10. For a set of increasing PAR (p_1, p_2, … ,p_m), we calculated the corresponding phenotype vector, giving a matrix of dimension N*m.
11. The genotypes for the causal genes along with the set of different phenotypes, where given to SKAT and WSS
12. For each phenotype vector (corresponding to all the PAR analysed), drew a random subset of a given sample size and asses the detection percent over all genes, and repeated this for 50 replicates. We then calculated the mean and 95% empirical confidence interval of the power over all the replications.

Relationship between PAR and GRR

The relationship between Population Relative Risk (PAR) and Genotype Relative Risk (GRR) used in the simulation is clarified in the equations below. If we define PAR as a ratio between an Exposure (E) and the Relative Risk (RR), as in the formula below:

$$PAR = \frac{E(RR - 1)}{1 + E(RR - 1))}$$

We can now substitute the Relative Risk with the Genotype Relative Risk (GRR), and the Exposure is substituted with the minor allele frequency (MAF), after some rearrangement we get:

$$PAR = \frac{E(RR-1)}{1+E(RR-1))}$$

$$PAR = \frac{MAF(GRR-1)}{1+MAF(GRR-1))}$$

$$PAR(1+MAF(GRR-1)) = MAF(GRR-1)$$

$$GRR(MAF(PAR-1)) = MAF(PAR-1) - PAR$$

$$GRR = 1 - \frac{PAR}{MAF(PAR-1)}$$

This leads to the final relation which is used in the simulation alforithm;

$$GRR = \frac{PAR}{MAF(1-PAR))} + 1$$

The GRR can also be inverted to emulate a protective effect, by substituting the effect sizes with 1/effect size

$$GRR = \left(\frac{PAR}{MAF(1-PAR))} + 1\right)^{-1}$$

Phenotype construction

To construct phenotypes, we calculated the probability of an individual being affected as the product of all the GRRs, given in the equation below(Equation 2)

Where b_0 is the baseline risk (incidence), and a is the set of all causal alleles, such that g_j is the allelic count (0, 1, 2), for allele number j.

To justify Equation 2 that the probability of being affected is the multiplication of GRR, we start with the (Genotype) Relative Risk for one variant, given an exposure (E), which

$$P(Y=1|G=\{a\}) = \min\left\{1, b_0 \prod_{j\in\{a\}} GRR^{a_j}\right\}$$

can also be an allele:

$$RR = \frac{R(Y=1|E)}{P(Y=1|E^c)}$$

$$P(Y=1|E) = RR \times P(Y=1|E^c)$$

Now the last part of the equation ($P(Y|E^c)$) is the probability of being affected given no exposure, which is the incidence or background risk (b_0). For many different variants, we can assume that the probability of being affected is the intersection of the probabilities. If all the variants act independently, and the incidence is assumed to be constant for the trait, this reduces to multiplication of all relative risks:

$$\bigcap P(Y=1|E_i) = \bigcap_i \times P(Y|E_i^c) = b_0 \prod_{\forall i} RR_i$$

Supplemantary tabel 1

Reproducing the table from the exome chip consortia, showing the different contributions to the design of the exome Chips.

Study name	Phenotype	Decent	Sample Size
NHLBI ESP (5 tranches)	CD, Lung traits, Obesity	(American) Europeans, AA	4260
ARRA	Autism	(American) Europeans	1778
GO T2D (2 tranches)	T2D	(American) Europeans	1618
KG (2 tranches)	Healthy individuals	Diverse	1128
Sweden Schizophrenia Study	Schizophrenia	European (Swedes)	525
SardiNIA	LDL-c, HDL-c, TG	Sardinia population	508
CoLaus	Overweigh, Diabetes, Fasting Glucose	European (UK)	456
Cancer Genome Atlas	Cancer	European	422
T2D GENES	T2D	Hispanic (Mexico)	362
Cancer Cohort Study (SMWHS*)	Cancer	Chinese	327
Pfizer/MGH/Broad	T2D Extreme risk	European	182
Lipid Extremes	Lipid Extremes	European	131
Int'l HIV Controllers	HIV Controllers	(American) Europeans	121
SAEC DILI	Augmentin DILI	European	117
I2B2	Major Depression	European	50
BMI Extremes	BMI Extremes	European	46

*SMWHS= Shanghai Men and Women Health Study

Supplement to Paper III

Appendix

Threshold selection

The threshold window size k is chosen such that the expected number of significant tests for each window size k is equal and the significance level of the total test is α. Using extreme value theory Zhang deduces the following simple relationship between the significance level and the total intensity λ, $\alpha = 1 - e^{-\lambda}$. Hence $\lambda = -\log(1-\alpha)$. The number of non-overlapping windows of size k ($\#window_k$) is equal to number of observation (nobs) in the CpG-island divided by window size k. Therefor the threshold t_k of window size k is chosen such that the intensity of window size k is $\lambda_k = k \cdot \lambda_1$. The total intensity $\lambda = \sum\limits_{k \in K} \lambda_k = \sum\limits_{k \in K} k \cdot \lambda_1$, here K is the set of window sizes to be examined. The expected number of significant windows of size k is given by

$$E[sign.window] = \lambda_k \cdot \#window_k = k \cdot \lambda_1 \cdot \frac{nobs}{k} = \frac{\lambda}{\sum\limits_{k \in K} k} \cdot nobs = \frac{-\log(1-\alpha)}{\sum\limits_{k \in K} k} \cdot nobs \tag{1}$$

hence only depending on the sum of all window sizes tested and not on a particular window size. So we can derive number of significant windows allowed keeping the significance level α. Using simulation we can find the optimal threshold such that

$$E[sign.window] = \frac{-\log(1-\alpha)}{\sum\limits_{k \in K} k} \cdot nobs \tag{2}$$

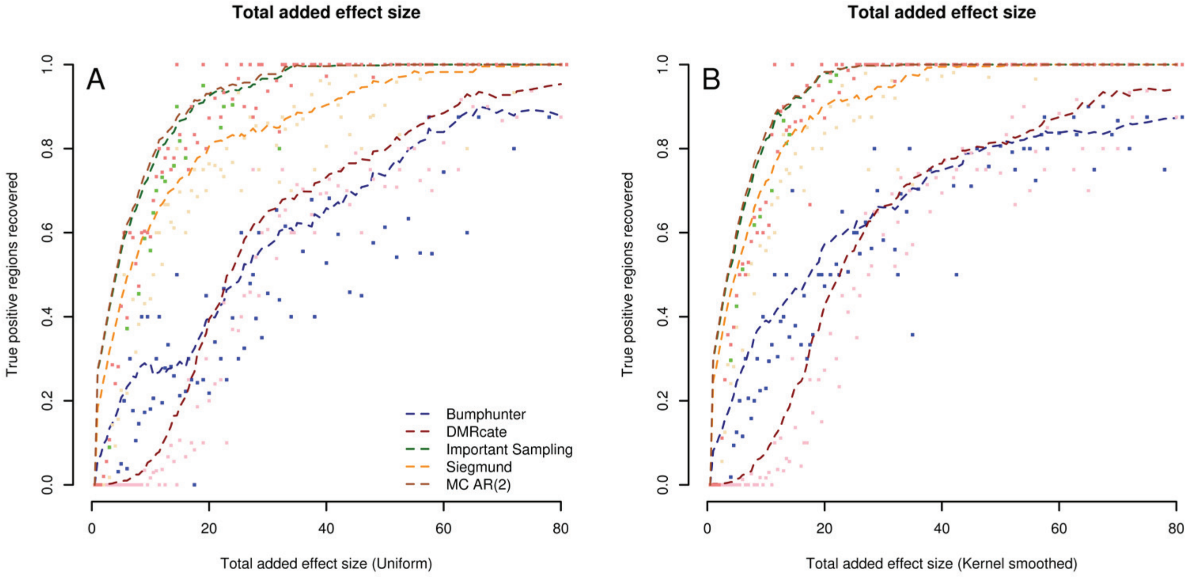
for the various window sizes.

p-value DMR Scan(MCMC)

T^k_{DMR} is the average of T_{CPG} in a window sized k, where T_{CPG} is following an AR(2) process, i.e.

$$T^k_{\mathrm{DMR}_i} = \frac{1}{k} \sum_{i}^{i+k} T_{\mathrm{CPG}_i}. \tag{3}$$

Using simulation one is able to find an emperical estimate for the variance $\hat{\sigma}^k_{\mathrm{DMR}}$ of T^k_{DMR} with non-overlapping T_{CPG}. Using the assumption of DMR Scan that no two overlapping windows can be significant we know that selected DMRs are i.i.d. Gaussian with mean zero and variance $\hat{\sigma}^k_{\mathrm{DMR}}$. p-values can be calculated using the Gaussian distribution function.



Supplementary Figure 1 *Estimated power when total effect size is considered. Long DMRs have more added effect than shorter regions. Here long regions with small effects are collapsed with short regions with high effect.*

Supplementary Material & Methods

Materials

The data used in this study consists of targeted bisulfite-sequenced (84 MB target - SureSelect Methyl-Seq, Agilent Technologies, Santa Clara, CA, USA) saliva DNA from samples of 100 11 years-old Finnish females. Per individual, we measured methylation in 2,947,202 loci, mainly residing in CpG islands. The size of these islands ranged from 2 to 791 CpGs.

A custom script produced β-methylation values from raw sequencing data by combining open-source softwares. First, low quality sequences and adaptors were removed using the Nesoni clip (version 0.115) (http://www.vicbioinformatics.com/software.nesoni.shtml). The bisulfite-converted sequence reads were then mapped to the human genome (hg19) using Bowtie2 (version 2.0.5) (Langmead & Salzberg, 2012) and Bismark (version 0.10) (Krueger & Andrews, 2011). The Bismark methylation extractor and custom formatting scripts were used to calculate β-methylation values for CpG sites. The first seven bases of each sequence were ignored, as a strong bias toward non-methylation was caused by the insertion of unmethylated cytosines during end-repair in the sequencing library preparation. CpG sites at which more than 25% of the samples had less than 10x coverage were discarded.

To make the benchmarking computationally feasible, we extracted all annotated regions from chromosome 22. This chromosome had 58,910 measured CpGs, distributed over 1,071 regions, with a mean of 55 observations per region, and a range of 16 to 456 CpGs per region. Chromosome 22 represented a reduced dataset without any a priori significant regions for the phenotypes sampled in earlier study.

Methods

Simulation parameters for Bumphunter and DMRcate

Bumphunter

The main parameters in Bumphunter are, the maximum allowed gap between probes within a cluster, the trimming coefficient, and the number of permutations. The trimming coefficient gives the quantile of the test statistics that are aggregated into new clusters in each permutation. This was set to the default value of 0.99, thus setting the threshold to include 1% of the top hits. We only considered regions with a "fwer" value below 0.05, after 2000 permutations.

DMRcate

The genome wide significance level for the probe-wise p-values is set by the user as well as the multiple testing adjustment method for the probe-wise p-value. The scaling factor (C) is inverse proportional to the standard deviation of the kernel smoothing. Empirical testing by the authors of DMRcate showed that when smoothing parameter is 1kb, the optimal scaling factor was close to 2 (Peters *et al.*, 2015).

References:

Krueger, F. and S. R. Andrews (2011). "Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications." Bioinformatics 27(11): 1571-1572.

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nature methods 9(4): 357-359.

Peters, T. J., M. J. Buckley, A. L. Statham, R. Pidsley, K. Samaras, *et al.* (2015). "De novo identification of differentially methylated regions in the human genome." Epigenetics Chromatin 8: 6.