GENETIC VARIATION AND DNA METHYLATION IN THE CONTEXT OF NEUROLOGICAL DISEASE

Dena Michelle Godwin Hernandez

I, Dena M.G. Hernandez confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy to the University College London

Acknowledgements

I would like to thank my graduate supervisors, Dr. Andrew Singleton and Dr. John Hardy. I am truly grateful for this opportunity to grow as a scientist and I cannot thank you enough for your encouragement, guidance and support. Special gratitude and appreciation is given to Dr. Singleton for your time, wisdom and calm direction. Many thanks to members of the Laboratory of Neurogenetics who contributed and extended their valuable assistance in the preparation and completion of these studies, Dr. Mark Cookson, Dr. Bryan Traynor, Dr. Raphael Gibbs, Dr. Michael Nalls, Sampath Arepalli and Chris Letson. Finally, a special thanks to my family for continued enthusiasm, support, patience, humor and more patience. Thank you, I love you all very much.

Published Work

Parts of this thesis have been published in the following peer reviewed manuscripts:

1. Gibbs JR*, van der Brug MP*, **Hernandez DG***, Traynor BJ, Nalls MA, Lai SL, Arepalli S, Dillman A, Rafferty IP, Troncoso J, Johnson R, Zielke HR, Ferrucci L, Longo DL, Cookson MR, Singleton AB. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS Genet. 2010 May 13;6(5):e1000952. PubMed PMID: 20485568.
2. **Hernandez DG**, Nalls MA, Gibbs JR, Arepalli S, van der Brug M, Chong S, Moore M, Longo DL, Cookson MR, Traynor BJ, Singleton AB. Distinct DNA methylation changes highly correlated with chronological age in the human brain. Hum Mol Genet. 2011 Mar 15;20(6):1164-72. PubMed PMID: 21216877
3. International Parkinson Disease Genomics Consortium, Nalls MA, Plagnol V, **Hernandez DG**, Sharma M, Sheerin UM, Saad M, Simón-Sánchez J, Schulte C, Lesage S, Sveinbjörnsdóttir S, Stefánsson K, Martinez M, Hardy J, Heutink P, Brice A, Gasser T, Singleton AB, Wood NW. Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet. 2011 Feb 19; 377 (9766) :641-9. PubMed PMID: 21292315.
4. International Parkinson's Disease Genomics Consortium (IPDGC), Wellcome Trust Case Control Consortium 2 (WTCCC2). A two-stage

meta-analysis identifies several new loci for Parkinson's disease. PLoS Genet. 2011 Jun; 7 (6): e1002142. PubMed PMID: 21738488;

5. **Hernandez DG**, Singleton AB. Using DNA methylation to understand biological consequences of genetic variability. Neurodegener Dis. 2012:9 (2): 53-9 PubMed PMID: 22123027

Abstract

Understanding genetic control of biological processes is an important goal in the post-genome era. GWA studies have been successful in identifying loci linked to disease and recently in revealing disease risk alleles. However, pinpointing key genes and their alterations within associated loci remains a central challenge. Described is the completion of a large, international meta-analysis and replication study of existing PD GWA data, confirming 6 previously identified loci and identifying and confirming an additional 5 novel loci; thus, expanding our understanding of the genetic basis of PD. Following this is an effort to annotate the consequences of genetic variation within the context of normal human brain tissue by generating and integrating data to investigate the effects of common genetic variability on DNA methylation in four brain regions of 150 neurologically normal individuals, 600 samples total. Genome-wide SNP data is generated and 27,578 CpG sites assessed in each brain region. Results show methylation patterns differ between brain regions, genotype is correlated with methylation levels and DNA methylation QTL occur more often in sites outside of CpG islands. Next, an expanded map of DNA methylation in human brain assessing 486,428 CpG sites is generated and proximal CpG sites are integrated with known PD loci implicated by GWA studies; thus, gaining potential mechanistic insight into pathogenesis of disease. Significant DNA methylation QTL for 19 of 28 PD risk loci are identified, demonstrating the correlation of risk alleles for neurological disease with a biologically relevant trait in human brain tissue is a manageable goal.

Lastly, analyses show CpG sites within normal human brain exhibit significant age-associated increases in methylation with an enrichment of changes at CpG islands of functionally related transcripts; thus, providing a footing for future integration of age-related epigenetic changes into disease models exhibiting age as a primary risk factor.

Table of Contents

Table of Figures

Table of Tables

Abbreviations

AD	Alzheimer’s disease
ALS	Amyotrophic lateral sclerosis
ARJP	Autosomal recessive juvenile parkinsonism
ARPD	Autosomal recessive Parkinson’s disease
CDCV	Common disease common variant
CDRV	Common disease rare variant
CH_3	Methyl group
CRBLM	Cerebellum
dmQTL	DNA methylation quantitative trait locus
DNA	Deoxyribonucleic acid
EOPD	Early onset Parkinson’s disease
eQTL	Expression quantitative trait locus
FCTX	Frontal cortex
FDR	False discovery rate
GWA	Genome wide association
HapMap	Human haplotype map
HCL	Hierarchical clustering
HSP	Hereditary spastic paraplegia
IBD	Identity by descent
IPDGC	International Parkinson’s disease genomics consortium
LB	Lewy body
LD	Linkage disequilibrium
LOPD	Late onset Parkinson’s disease
MDS	Multi dimensional scaling
mRNA	Messenger ribonucleic acid
PAR	Pseudoautosomal dominant region

PCA	Principal Component Analysis
PCR	Polymerase chain reaction
PD	Parkinson’s disease
QTL	Quantitative trait locus
Q-Q plot	Quantile-quantile plot
RNA	Ribonucleic acid
RRBS	Reduced representation bisulfite sequencing
SD	Standard deviation
SEM	Standard error of the mean
SNP	Single nucleotide polymorphism
SN	Substantia Nigra
TCTX	Temporal cortex
WTCCC	Wellcome Trust Case-Control Consortium

1 Introduction

1.1 Specific Aims of this Thesis

This thesis focuses on understanding the basis of a complex, late-onset neurological disease, Parkinson's disease (PD), through the identification of genetic risk variants, and an assessment of the effects of these variants on DNA methylation. The first part of this effort is the generation and analysis of genome wide association data in a large series of PD cases and controls. This work aims to specifically test the common disease common variant hypothesis, and to identify loci that contain risk variants for disease. The second part of this thesis aims to generate a dataset that tests the hypothesis that common DNA variability may be correlated with DNA methylation in the human brain, and to provide a reference dataset that can be used to examine these correlations. The next aim centers on integrating the PD genome wide association data and the brain DNA methylation quantitative trait data to detect whether the identified risk alleles are also correlated with differential DNA methylation, potentially identifying a biological basis for association. The final aim of this thesis was to examine the relationship between chronological age and DNA methylation in the human brain in order to understand if predictable DNA methylation changes occurred with aging.

1.2 Understanding Neurological Disease

For the majority of neurological diseases current treatment options are limited, mostly restricted to symptomatic or palliative therapies. The ultimate goal of disease-based research is to change this paradigm by creating therapies that are directed at the underlying disease process, rather than the manifesting symptoms. A central part of this effort to understand etiology has been the identification of disease related genes, which in turn provides a tool with which to investigate the molecular events that are the disease process.

Traditionally gene identification efforts have been centered on the isolation of disease causing mutations, usually in rare familial forms of disease. These mutations, which most often affect the protein coding sequence of the gene, are subsequently used to model disease within cell and animal based systems. While this has been a fruitful area of research, there is a need to more fully understand the genetic architecture of typical diseases, including the identification of variants that increase risk for, rather than cause disease. It is believed that a more general understanding of the genetic architecture of complex disease and the identification of genes involved in lifetime risk for disease, will not only inform on the etiologic level, but will also be a key step in the identification of patients at risk, biomarkers for disease, and disease subtypes.

As discussed below, there has been a great deal of progress in methodological and analytical approaches to the identification of risk loci. This success not only creates exciting opportunities to understand the disease

process, but also significant challenges. A critical challenge lies not only in the identification of the disease-linked gene at a particular risk locus, but also in understanding the biological effect mediated through the gene and locus. In this introduction I will discuss the recent evolution of gene discovery efforts in complex traits, the state of the field in gene discovery within a complex, age-related disorder, Parkinson's disease, and the role that DNA methylation may play in understanding the disease process.

1.3 Genetic Variability and Complex Disease

Complex diseases likely arise due to a combination of many genetic, environmental, and lifestyle factors. There are two principle theories regarding the genetic component of complex disease: the common disease rare variant (CDRV) hypothesis and the common disease common variant (CDCV) hypothesis. At the inception of these theories, there was much discussion regarding the likely validity of these hypotheses, and in particular much argument in favor of one over the other; however it is important to note that these two hypotheses are not mutually exclusive, and it is likely that any complex disease will contain both genetic components.

The CDRV hypothesis speculates that a contributing risk component for complex disease will be rare genetic variants; these are strictly defined as any genetic allele with a frequency of 1% or less. The CDRV hypothesis suggests that because low frequency variants are abundant in human populations, and because there has been little opportunity for the removal of these functional but rare variants through natural selection (as a consequence of recent

population growth) deleterious rare alleles are likely to exist. This phenomenon may be particularly pronounced in late onset diseases, where selective pressures, mainly driven through traits that occur before reproductive age is reached, do not apply. A primary limitation in the investigation of the CDRV hypothesis has been a technical one; it has traditionally been extremely difficult to identify rare variants in suitably powered sample series. However, there is now a great deal of interest in the investigation of the CDRV hypothesis due to the affordability of second generation sequencing methods to identify these rare alleles in large groups of individuals. This is one of the fastest growing fields of disease genetics and has had some initial success [1, 2]. However, the rarity of these alleles requires very large sample numbers to detect significant effects and therefore, it is expensive to execute well-powered studies. As a consequence, the CDRV hypothesis remains largely untested in most common disorders, a situation that will inevitably change over the next few years.

Kathleen Merikangas and Neil Risch proposed the idea of ‘common disease—common variant’ in 1996, suggesting that common, modest-risk SNPs could mediate genetic susceptibility to common diseases, such as Parkinson’s disease and other complex disorders [3]. The CDCV hypothesis posits that a significant proportion of risk for common diseases is mediated through common genetic variants (i.e. variants present at greater than 1% allele frequency within a population) [4, 5]. By definition these variants are common and therefore will have been within the population for a significant amount of time. Thus, conversely to rare variants, highly functional, deleterious alleles

are more likely to have been selected out of the population. Therefore, the CDCV hypothesis accepts that the effect of individual common alleles on any deleterious trait is likely to be quite small, but numerous such common alleles may contribute to that trait, so collectively the contribution of common alleles to a trait may be substantial. Testing of the CDCV has been well established, primarily because the technology available to genotype a large number of common variants has been available, and relatively affordable since ~2005. This has led to an extremely large number of identified common risk loci as evidenced by the large catalog of positive genome wide association studies (http://www.genome.gov/gwastudies/).

1.4 Knowledge and Technology: Enabling Complex Genetics

The Human Genome Project (http://www.genome.gov/10001772), the mapping of genetic variability across populations via projects such as the International Haplotype Map Project (http://hapmap.ncbi.nlm.nih.gov), and improvements in technology combined to facilitate the development of tools that revolutionized the study of human genetics. During this time, the core underlying data and methods used to discover genetic causes and risk for complex human disease were fundamentally transformed, resulting in a great deal of genetic discovery.

Before the scope of human genetic variability was well-documented and prior to the advent of GWA studies and microarray technology, candidate gene association studies were widely applied in the search for genes underlying complex disease. The candidate gene approach allowed researchers to

investigate variants in selected genes based upon a hypothesis of the etiological role of the gene(s) in the disease, with little known about the human genome or the variability at that locus. As the gene(s) being tested were hypothesized to have a biological role in the disease mechanism, one advantage was that finding a genetic association also informed on the functionality in the pathway. However, there are many limitations to candidate gene studies. For example, bias can be inserted into a study as a researcher must know at least some of the biology underlying the disease and therefore, may choose genes based on their own specialty, limiting the ability to search for variants outside a specified region. Therefore, causative variants could be missed. In this regard, GWA studies have a distinct advantage being hypothesis-free, simultaneously assessing SNPs throughout the entire genome versus focusing on loci where there may be a suggestive causal relationship to the disease [6].

Additionally, without microarray technology, candidate gene studies focused on genotyping small numbers of common variants in small numbers of cases and controls, resulting in limited statistical power. Thus, the effects of some tested variants were likely not detectable, even if the variants truly contributed to disease susceptibility. The best indication of the general failure of candidate gene association studies may be the large number of such studies from the mid 1990's and the lack of genuine associations identified by those studies.

Enabling genetic studies, the Human Genome project was an international effort launched in 1990 aiming to sequence and map the genome of *Homo*

sapiens. The first draft was completed and released in April 2003. Following the successful effort by the Human Genome Project, the International HapMap Consortium was founded in October of 2002. The goal of the HapMap project was to develop a haplotype map of the human genome defining common patterns of genetic variation in worldwide populations. The first phase of the HapMap project included four populations from Africa, China, Japan, and US residents with European ancestry. Additional populations were later assembled to include cohorts from Tuscany, Italy, Kinyawa, Kenya and distinct US populations with Chinese, African, Indian and Mexican ancestry. The focus of the HapMap project was to map common alleles occurring in at least 1% of the different populations [7, 8].

These two ambitious, international collaborations and their public release of data provide the basis of knowledge for the majority of human genetic experiments that are performed today. While the creation of these genetic maps was necessary to facilitate a new era of discovery in human genetics, so too was the development of technologies capable of capitalizing on this knowledge.

In addition to having a reference human genome sequence and a map of sequence variation in several populations, a timely technological advancement allowed faster, accurate and much cheaper genotyping of single nucleotide polymorphisms (SNP). These methods allowed the realization of high throughput, high content genotyping, and were critical for the execution of GWA studies.

1.4.1 DNA microarrays

DNA microarray technology permits the accurate and efficient typing of hundreds of thousands to millions of variants in a large series of subjects. Array technology was developed in the early 1990s [9], but its potential in the context of SNP genotyping was not fully realized until 2005. The basic principle underlying DNA microarray technology is hybridization between two DNA strands. Specifically, complementary nucleic acid pairs form hydrogen bonds with each other on a substrate containing allele-specific oligonucleotide probes in a specific pattern (in the instance of Illumina technology the substrate is a bead). The DNA from each subject is fragmented, amplified and hybridized to the array. Non-specific binding sequences are washed off the slide and fluorescently labeled target sequences that bind to an allele-specific oligonucleotide probe sequence generate a fluorescent signal. The signal is then detected by scanning software and SNP genotypes are determined for individuals. Today, this technology allows for up to 5 million known genetic variants (Illumina Omni5M, Illumina Inc., CA, USA) to be simultaneously typed in a single human DNA sample.

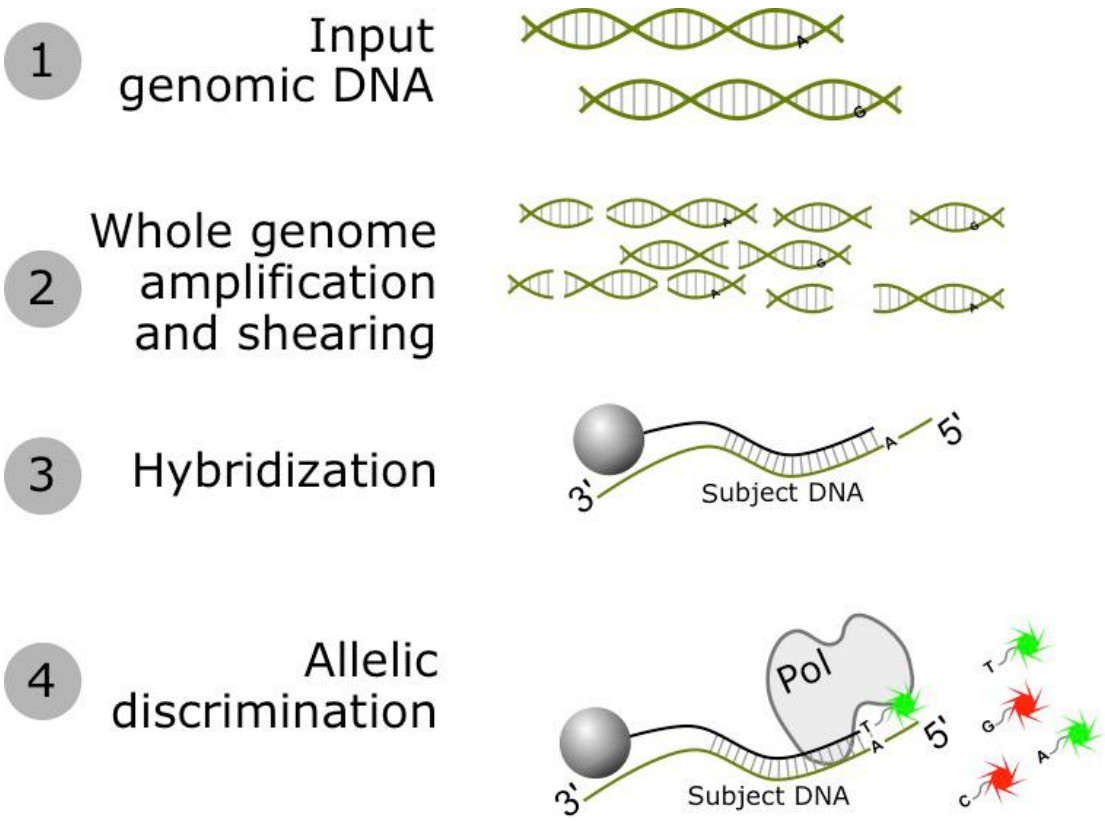


Figure 1. Graphic representation of an Infinium HD SNP array genotyping experiment.

In step one, genomic DNA from the sample is extracted, purified, and input into the experiment (~250ng). In step two, the genomic DNA is amplified using a non-PCR based method, and the resulting product is sheared into small fragments. In step 3 the DNA is hybridized to beads with a target sequence abutting the variant of interest. In step 4, a single base pair extension occurs and depending upon the variant at this position in the source DNA the complimentary base pair tagged with a fluorescent label is incorporated by polymerase.

There exist now a large number of arrays with different content and slightly different purposes. The majority of arrays, and those that have facilitated the majority of advances through GWA, are aimed at maximizing the information content of interrogated SNPs across the human genome. Although genome-wide SNP arrays allow typing of a very large number of variants, it is possible to use the understanding of the haplotype block structure of the human

genome to predict or impute, the genotypic state of SNPs that are not typed directly on an array, but that are physically close to directly typed variants [8, 10]. Thus many SNPs are selected for inclusion on arrays based on the information that the SNP provides on neighboring variants, with variants that efficiently provide information being called tagging SNPs. The incorporation of tagging SNPs can be used to capture a major fraction of the genetic variation present within a population while reducing genotyping demands, data handling and computation time [11, 12].

While genome wide SNP chips have greatly facilitated the application of genome wide association (this will be discussed further below) they have also played a significant role in other genetic discovery efforts [13]. Although individually SNPs are less informative than the microsatellite markers traditionally used in linkage, the density of SNP genotyping chips provides more information and finer scale mapping than microsatellites. Because SNP arrays are very informative, and due to their ease of use, speed, and low cost, they are now the predominant technology for generating genotypes in genome wide linkage studies. A related application is the use of these arrays in rapidly identifying regions of homozygosity, a methodology used in autozygosity mapping for recessive traits, identifying loss of heterozygosity in tumor samples, and in defining regions of uniparental isodisomy [13]. In addition, because the arrays provide information not only on genotypic state, but also generate an intensity value, the data can be abstracted from these arrays to identify copy number variants, although it should be noted that there is a high error rate in this regard [13]. Additionally, and critical for genome wide

association studies, the genotypes from genome wide SNP arrays can be used in identifying population structure and outliers within a group. This is performed using principal component analysis and is used routinely in GWA analysis to remove outliers and to correct for population stratification within an analysis [14]. This is discussed further below.

1.4.1.1 Genome Wide Association

GWA studies examine variants across the genomes of large cohorts of disease cases and control subjects. On the most basic level, a test of association between variability at a particular SNP and a trait such as disease is performed. Commonly an additive model is assumed where the risk for disease for a heterozygote is intermediate between the risks at opposite homozygotes, although other tests such as the genotypic, allelic, dominant, and recessive tests may be executed. Often implemented is the Cochran-Armitage trend test, which is used in the whole genome association toolset plink (http://pngu.mgh.harvard.edu/~purcell/plink), and this toolset has been a critical in the execution of GWA studies, because it is free to use, easy to access, and relatively straightforward. More recently, several studies use home grown solutions to calculating significance implemented through R, and this is particularly apparent for studies that use imputed data, as the association with disease should most appropriately be regressed against allele dosage rather than against absolute genotype (this is discussed more below).

Although modern genotyping arrays contain a large number of variants, even greater genetic information may be gleaned from these arrays using imputation. Imputation is based on linkage disequilibrium (LD) and uses genotypes at neighboring SNPs to predict (or impute) the likely values of neighboring variants [8, 10].

LD is the non-random association between two or more alleles where certain combinations of alleles are more likely to occur together on a chromosome. This generally occurs because the SNPs are close together and recombination events are not likely to occur within the few base pairs between them. The HapMap Project was able to catalog this phenomenon and provided an understanding of LD in the human genome; thus allowing successful imputation of millions of genotypes to be used in association studies [7, 11]. Further, in 2012 the 1,000 Genomes Project released 1092 genomes sequenced from a number of different ethnic groups [15, 16]. This novel dataset greatly added to our understanding of LD and has become a key resource for modern imputation methods (Figure 2).

Notably in most cases the results for imputed SNPs are assigned as allele dosages rather than genotypes, and because they are probabilistic are often not absolute, i.e. a genotype of A/A, A/B, and B/B would not be imputed, but rather an allele dosage would be produced for the variant allele ranging from 0 to 2, and most often not a whole number.

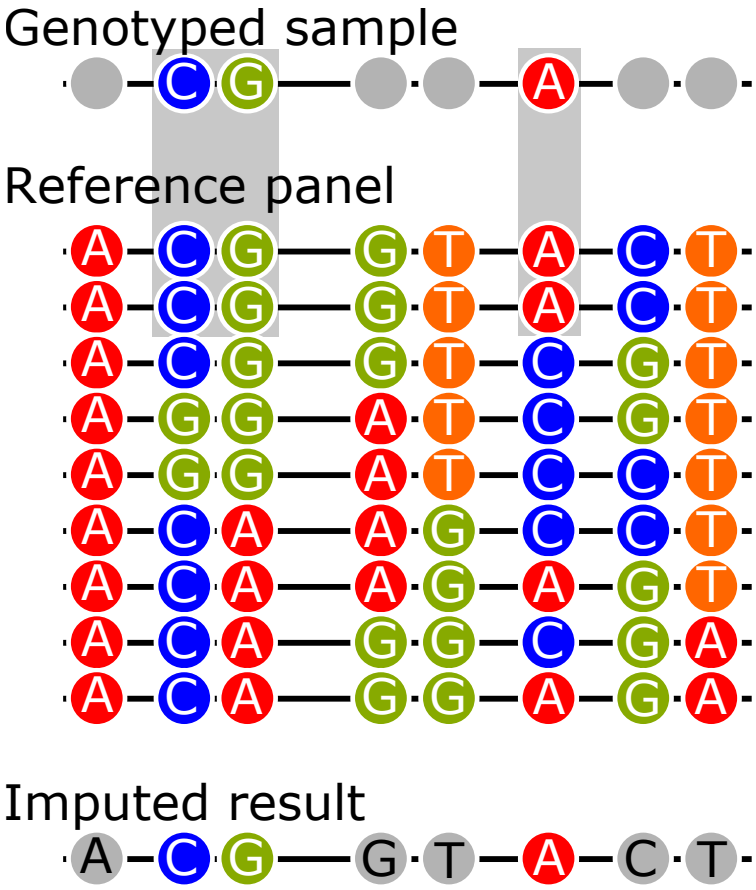


Figure 2. Representation of haplotype based imputation.

Known variants within a genotyped sample are compared to those within a reference set, and the reference set of haplotypes is then used to impute or predict proximal untyped variants. While this representation shows alleles, in most instances the probability of an allele is predicted, rather than a haplotype call.

These technological and analytical advances were necessary precursors for genome-wide association (GWA) studies to come of age as a novel approach to the study of genetic variability linked to complex disease. Coupled with imputation, it is routine now to interrogate more than 10 million SNPs in a GWA.

1.4.1.1.1 Considerations in Implementing Genome Wide Association Studies

There are a number of best practice considerations in the implementation of GWA studies. Well-characterized cohorts of cases and controls are the foundation of successful GWA studies. To ensure maximum benefit from genome-wide genotyping efforts and/or sequencing efforts prior to genetic association analysis, it is important that sample collection for GWA studies involve detailed phenotypic classification that may involve additional genetic, physiological and clinical tests. Phenotypic misclassification between cases can reduce the power to detect association in GWA studies. Manchia and colleagues quantified the impact of phenotypic heterogeneity in the analysis and interpretation of GWAS results noting that the presence of “non-cases” reduces the statistical power to identify genetic association and significantly decreases the estimates of risk attributed to genetic variation. Their results also suggest that accurate phenotype delineation may be more important for detecting true genetic associations than increase in sample size [17].

Many of the quality control steps taken in GWAS focus on the removal of systematic errors in order to limit the number of false positive and false negative association signals. Initial quality control includes basic measures including the removal of poor performing SNPs and samples. Because genotyping using arrays is quite robust, the threshold for SNP and sample failure remains high, with any SNP failing in more than 2% of samples, being removed, and any sample with a call rate of less than 95-98% also being removed. Following initial quality control, an assessment of heterozygosity

across all non-sex chromosomes is performed, with outliers being removed (any sample outside of +/- 6 standard deviations). Excess or low heterozygosity is considered indicative of poor genotyping or sample contamination.

This is followed by an assessment of genotype gender compared to sample gender, which is easily performed using genotypes on the non pseudoautosomal dominant region (PAR) of chromosome X; simply genotypes from males will appear hemizygous, whereas females will have a proportion of heterozygous calls on this chromosome. Because arrays also provide intensity data, which is related to the original copy number of template DNA, the signal intensity for chromosome X SNPs will be higher in females than in males (Figure 3).

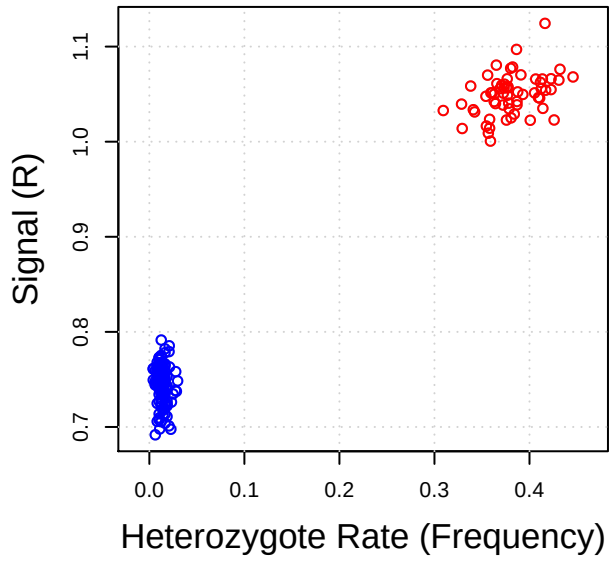


Figure 3. Example quality control plot of genotype gender and reported gender.

This was generated from 200 samples (red, female; blue, male) and average heterozygote frequency for SNPs on the non-PAR region of chromosome X was plotted against average signal intensity for the same SNPs. In this instance the underlying genotype data is from the NeuroX array. There are no gender discrepancies.

Another critical step in quality control of GWA focuses on estimating population stratification within the cohort. This is performed using a principal component analysis (PCA); one common form is the EIGENSTRAT method [14]. The varied PCA methods operate on a comparable construct, first a PCA is performed on genotype data across all cases and controls, and often, on reference populations. This creates a defined number of axes of variation, where each axis describes as much inter sample variation as possible; with this being based solely on genotype information these axes of variation represent inter sample differences in genotype. Because of the prior removal of genotype artifacts and poorly performing SNPs, one of the remaining key

differences between samples should be genetic diversity due to population differences. At this point, the PCAs explaining the majority of inter sample variation can be used to view population stratification within the test cohort and if a geographically diverse reference sample is included, the samples can be viewed in the context of known population structure. Samples that are evidently outside of the expected population (typically 6 standard deviations) can be removed. For the remaining samples, the values from the calculated PCAs can be used to adjust genotypes, thereby minimizing the effects of population stratification (Figure 4). Because most GWA are looking for genetic differences between cases and controls, it is important to be aware that over correction for too many PCAs may remove genuine trait related association signals.

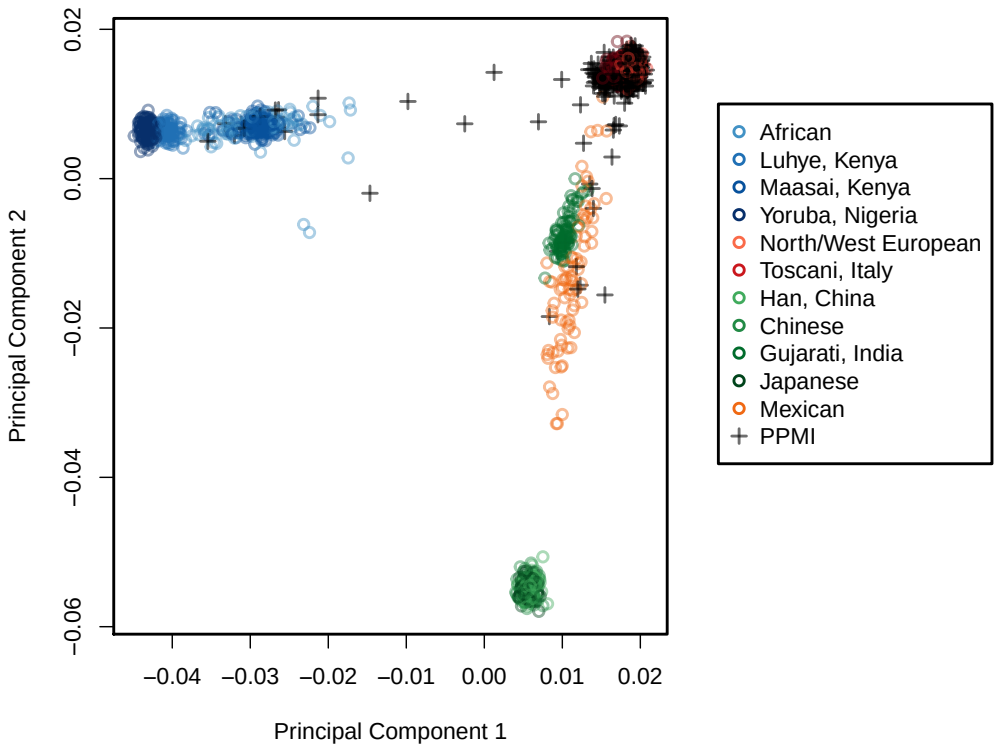


Figure 4. Plotting of the first two principal components for a study on population and reference populations.

In this example the first two genetic principal components are plotted for the study population (PPMI, internal data from LNG) along with the same PCAs for 11 reference populations for which data was downloaded from the 1000 genomes project (http://www.1000genomes.org). As can be seen these PCAs separate distinct populations; these data illustrate the study population is largely European, but also reveal some African, and Asian/South American ancestry.

An additional common quality control step is the application of a quantile-quantile (Q-Q) analysis. This step also searches for systematic bias within the data set. Essentially this involves plotting the observed versus expected p values across the whole genome and determining whether there is a skew in the distribution of observed p values. An excess of significant p values would suggest systematic bias in the data and that many of the significant p values are a result of false positive association due to genotyping errors or

population stratification. Conversely a lower than expected number of associations would indicate that the data might be overcorrected (for example by applying too many PCAs) and that genuine association signals may have been removed.

Lastly, a critical step in GWA is the application of correction for multiple testing. Because many hundreds of thousands, or millions of statistical tests are being performed, there will be a wide range of p values generated simply by chance, and not as a result of a genuine association. The threshold for genome wide significance is considered to be $5x10^{-8}$; however, ultimately the most important consideration in GWA is the need for a large and independent replication of any putative associations. Many of the considerations for GWA apply to this replication, including adequate power, correction for multiple tests, and standard sample and genotype quality control. A replication needs to fulfill these criteria and replicated signals should show a consistent direction of effect compared to the discovery (genome wide) data. Notably, the effect size is often smaller in the replication cohort, a consequence of the 'winners curse' effect, where the magnitude of a number of effects discovered as significant in stage I will be overestimated (this is a principle adapted from economics) [18]. This is most evident for variants with an effect size close to the limit of power for stage I. Because of this, it is preferable that a replication phase include a larger sample series than the discovery phase, although in practice this is often difficult to achieve.

1.4.1.1.2 Limitations of Genome Wide Association Studies

While GWA has been an extremely useful tool in the identification of novel risk loci for disease, there are several limitations to this approach. First, in its current form the method is designed to test the CDCV hypothesis, and is not designed to identify rare risk variants for disease, an approach better suited to sequence based approaches.

Second, each individual study is limited in power by the sample size used, and the expected effect sizes. Often, this has resulted in the combining of datasets generated in different laboratories and different countries. Several methods have been developed to facilitate data combining, ranging from simple aggregation of data and combined (or joint) analysis, through to a formal meta-analysis. In the latter design, association statistics are calculated individually for each study, or each population, and then compared across studies, typically these statistics are effect sizes, but p values can also be used. Meta analysis increases the power and facilitates the identification of consistent effects across studies, even if the effect is not statistically significant within individual studies.

The paramount limitation of GWA studies is perhaps that, unlike linkage and positional cloning approaches used to find disease-causing mutations, GWA identifies loci, and not genes or individual risk variants. The next steps from the identification of a robust and replicated association signal center on identifying the critical variant of interest, and on understanding the biological effect of the risk variant. For the most part, GWA identified risk loci are not

explained by protein coding variants such as splice site variants, frame shift variants, stop/start mutations, or non-synonymous amino acid changes. In fact, less than 30% of identified GWA signals are in linkage disequilibrium with such changes, and it is likely that only a proportion of these represent the genuine trait-influencing variant [19]. Therefore, the majority of identified GWA loci must exert an effect on the disease process through expression, including effecting basal expression, induced expression, transcript specific expression, and temporospatial expression. Possibly such effects could be mediated through a number of mechanisms, such as altering the epigenetic regulation around a gene and altering transcription regulatory elements. Thus, while a great deal of investment continues in performing GWA with some success, there is increasing focus in the development of methods and data that will shed light on the biological consequences of identified risk variants.

1.5 Parkinson's Disease

1.5.1 Clinical Background

Parkinson's disease is a common, progressive neurodegenerative disorder, affecting 3% of those older than 75 years of age [20] . James Parkinson first described the disease in 1817, where he illustrated six cases of "shaking palsy" [21]. PD is the second most common neurodegenerative disorder, (following Alzheimer's disease) and is the most common neurodegenerative

movement disorder with an estimated 10 million individuals worldwide living with the disease [22] More males than females are affected with the disease in a ratio of 1.5:1.0 [20, 23]. Clinically PD is associated with resting tremor, postural instability, rigidity, bradykinesia and a good response to levodopa therapy [24]. In the past, PD was considered a ‘sporadic’ disorder; but, approximately 15% of patients report a positive family history of parkinsonism. The rare, familial forms of PD tend to be early onset, occurring before the age of 50, and are often monogenic in origin [25, 26].

Typically, onset of PD is subtle, asymetrical and steadily progressive as neuronal dysfunction and cell death lead to a significant reduction in the neurotransmitter dopamine in the striatum, a vital section of the basal ganglia accountable for the initiation and control of movement [27, 28]. For this reason, patients show a dramatic response to dopamine replacement therapy using the metabolic precursor of dopamine, L-DOPA, which is the chief treatment for PD. The neuropathologic hallmarks of PD are loss of dopaminergic neurons of the nigrostriatal system, the cause of the primary movement abnormalities, and deposition of intracytoplasmic aggregates termed Lewy bodies (LB). In PD, Lewy bodies are particularly prevalent in the substantia nigra; but may also be seen in neocortical areas [27, 29]. While clinical diagnosis of PD is relatively sensitive and specific, a true diagnosis can only be made neuropathologically [27].

1.5.2 The Role of Genetics in Parkinson's Disease

Parkinson's disease was long thought to be a sporadic disorder without genetic causation. However, in 1997 mutations responsible for the disease were identified in the alpha-synuclein gene (*SNCA)* [30, 31]. This landmark discovery revealed the first indisputable, heritable component to PD and launched 15+ years of successful research into the genetics of PD. Quickly following detection of the first mutations in *SNCA,* additional genetic links were identified at two novel chromosomal regions and linkage of *SNCA* was excluded in >200 PD families [32-36]. Therefore, by 1998, it was evident that PD was a genetically heterogeneous disease. Several genes have since been linked to inherited forms of parkinsonism and several monogenic forms of the disease and numerous genetic risk factors have been identified. In this section, I will first provide a brief overview of the monogenic forms of disease and move forward to discuss how our view of PD etiology has matured since 1997 to now include risk alleles.

1.5.2.1 Monogenic Forms of Parkinson's Disease

While the work in this thesis centers on risk loci for complex genetic forms of PD, it is useful to understand the monogenic forms of this disease. This provides insight more broadly into the genetic architecture of this disease, and as described later, there appears to be overlap in the genes that contain disease causing mutations, and those that contain risk variants. Mutations in three genes, *SNCA* (*PARK1*; encoding α-synuclein), *LRRK2* (*PARK8*;

encoding dardarin) and *VSP35* (encoding vacuolar protein sorting 35) have been shown to cause autosomal dominant forms of PD. Mutations in six genes, *PINK 1* (*PARK6*; pten induced kinase 1), *DJ1* (*PARK7*), *Parkin* (PARK2) and *ATP13A2* (PARK9) *FBXO7* and *PLA2GB* have been shown to cause autosomal recessive PD and/or parkinsonism. The mutations in these genes, with the exception of LRRK2, cause PD in a small subset of patients. All known monogenic forms of PD combined explain only about 30% of familial and 3-5% of sporadic cases [26].

1.5.2.1.1 Alpha-Synuclein

Traditional linkage mapping was used to identify the first PD gene, *SNCA*, which encods alpha-synuclein [31]. The mutation underlying disease (Ala53Thr in exon 4) was discovered in a large Italian family and subsequently in three Greek families with familial PD. The primary Greek families found to harbor the p.A53T mutation, orginated from a very small geographical area in southern Greece. Eight additional families, located in central and southwestern Greece, were also confirmed to have mutations in a-synuclein, suggesting the presence of a founder mutation [37, 38]. A decade later, two Korean, and one Swedish family were shown to have the mutation [39-41].

Shortly following the discovery of *SNCA* mutations causing a rare familial form of PD, Spillantini and colleagues determined that α-synuclein was a dominant constituent of Lewy bodies, the pathological hallmark of PD [42]. This profound finding in the brains of typical sporadic PD patients distinctly tied

together the etiology and pathogenesis of rare familial forms of PD with typical sporadic PD. This crucial link between the two disease phenotypes, ultimately established that examination of rare forms of familial PD, even those that differed clinically and neuropathologically from typical PD, was pertinent to the study of the common form.

Mutations in *SNCA* are rare. As yet, only five (G51D, G50D) autosomal dominant, missense mutations have been discovered in a-synuclein along with duplications and triplications of the complete gene [43]. The first identified missense mutation, A53T is the most frequent and has been found in seven families throughout the world. The remaining two missense mutations were found in only one family each: p.A30P was reported in a German family with autosomal dominant PD and p.E46K in a Spanish family from the Basque country [44, 45]. Triplication of the entire genomic region containing SNCA was first discovered in 2003 and has since been reported as a cause of disease in three separate families [46-48]. Duplications of the entire coding region of SNCA have been reported as a cause of disease in 13 PD families and four sporadic cases [47, 49-55].

The clinical phenotype associated with *SNCA* mutations consists of progressive L-DOPA responsive parkinsonism with cognitive decline, autonomic dysfunction and dementia. The average age of onset for those patients with the p.A53T mutation is 46 years of age, which is younger than typical sporadic PD and the disease is fully penetrant [56]. In contrast, families with the A30P mutation have an age of onset that is slightly later (age 52) and the disease is not fully penetrant [44], while the E46K mutation causes

dominant PD and Lewy body dementia with symptoms beginning between the ages of 50 and 65 years with dementia presenting within 2 years of diagnosis [45].

Genomic duplications and triplications at the *SNCA* locus cause early-onset PD with the age of onset and severity of the disease phenotype correlating with the SNCA copy number, suggesting a gene-dose effect. Triplication of alpha-synuclein causes a fully penetrant, early onset and rapidly progressive dopa-responsive parkinsonism, accompanied or followed by dementia. Clinical presentation ranges widely from severe idiopathic PD to PD with dementia or diffuse Lewy body disease. PD patients with duplication of *SNCA*, therefore generating three copies of the gene, develop the disease about a decade later than those with four copies of synuclein and the disease course, while still aggressive, is generally more benign [46, 48, 51, 57, 58].

1.5.2.1.2 Leucine Rich Repeat Kinase 2

In addition to *SNCA*, autosomal-dominant PD-causing mutations have been found in the gene encoding *Leucine-rich repeat kinase 2 (LRRK2).* Linkage of PD to a region on chromosome 12 was originally mapped in a large, Japanese family with autosomal-dominant, late-onset PD showing incomplete penetrance [59]. Within two years, the locus was verified and further delineated in several European families [60]. In 2004, positional cloning was performed by two groups that identified mutations in *LRRK2* as the root cause

of chromosome 12-linked PD [60, 61]. Mutations in *LRRK2* now represent the most common known cause of familial PD.

More than 100 distinct missense and nonsense mutations have been reported in *LRRK2* to date [62]; however, only for a small minority is there overwhelming proof of pathogenicity (p.R1441C/G/H, p.Y1699C, P.G2019S, p.I2020T, p.S1761R, and p.I2012T) [63-67]. Of these pathogenic modifications, clustered in exons encoding the carboxy-terminal region of the protein, the G2019S mutation is the most frequent and most well studied. G2019S is common across many populations and has been identified in up to 42% of familial cases, depending on the ethnic background [65, 68]. Importantly, G2019S has also been detected in sporadic PD cases. The mutation is seen in approximately 2% of sporadic cases in Northern European and U.S. populations and up to 10% of sporadic cases worldwide [64]. The penetrance of the G2019S mutation is age dependent and can vary from 28% at 59 years, 51% at 60 years to 74% at 79 years of age. The G2019S mutation is frequent in North African, Middle Eastern and Ashkenazi Jewish PD patients and it is believed that most LRRK2 G2019S mutation carriers are from a common founder originating in North Africa and spreading with the Jewish diaspora [65, 69-74].

Overall, mutations in *LRRK2* are the most common known cause of late-onset, autosomal dominant and sporadic PD. Mutations are found in ~10% of patients with autosomal dominant familial PD [75-78], 3.6% of patients with sporadic PD and 1.8% of healthy controls [78]. Phenotypically, *LRRK2* mutation carriers are essentially indistinguishable from sporadic PD [79]

demonstrating mid to late onset of disease around 60 years of age, with a slow progression and a good response to levodopa therapy. Dementia is not common. Neuropathologic features are consistent with typical PD and Lewy bodies (LB) are present in the brainstem and there is loss of dopaminergic neurons in the substantia nigra [79].

LRRK2 encodes a large, multi-domain protein of 2527 amino acids and is a member of the ROCO protein family [80, 81]. LRRK2 is widely expressed in brain tissue and localized to LBs in the brainstem where it is associated with the endoplasmic reticulum of dopaminergic neurons [82, 83]. Several functional domains, including a GTPase domain and a kinase domain, characterize the LRRK2 protein. Pathogenic mutations occur in both domains and the G2019S mutation is consistently shown to increase kinase activity [84]. The identification of LRRK2 mutations has proven to be a landmark discovery that has profoundly impacted our understanding of Parkinson's disease.

1.5.2.1.3 Vacuolar Protein Sorting 35

Zimprich and colleagues were the first to use next-generation sequencing methods to detect a PD causing gene. They identified a mutation in *vacuolar protein sorting 35 homolog gene (VPS3*5); encoding Vacuolar protein sorting 35, as a cause of late-onset autosomal dominantly inherited Parkinsonism. A family from Australia with 16 affected individuals was investigated. Exomes from pairs of affected cousins were compared and a list of rare, shared

heterozygous coding mutations was generated. Subsequently, only one mutation Asp620Asn was validated by Sanger sequencing and found to segregate with disease in a Mendelian dominant manner. Several thousand PD and control subjects have since been screened for the c.1858G>A (p.Asp620Asn) mutation . The mean age of onset in affected individuals is 53 years, causing approximately 1% of familial parkinsonism and 0.2% of sporadic PD [61, 85-87].

Autosomal Recessive Parkinsonism

Mutations in six genes: PINK 1 (*PARK6*; pten induced kinase 1), DJ1 (PARK7), *PARK2* (encoding parkin), ATP13A2 (*PARK9*; ATPase type 13A2), PLA2G6 (PARK14; phospholipase A2, group VI) and FBX07 (*PARK15*; F-box only protein 7) have been shown to cause autosomal recessive (AR) PD/parkinsonism. The mutations in these genes cause PD in a small subset of patients. All known monogenic forms of PD combined explain only about 20% of early-onset PD and less than 3% of late-onset PD, although as will be discussed below, this proportion varies drastically across ethnic groups.

1.5.2.1.4 Parkin

Parkin was the second gene identified to cause parkinsonism and the first gene decisively shown to cause an autosomal recessive (AR) form of the disease. A homozygous deletion of exons 3-7 in the *parkin* gene was first reported by Kitada and colleagues in Japanese families with autosomal-recessive juvenile on-onset parkinsonism (ARJP); disease onset often

occurring before twenty years of age [33]. Mutations in Parkin are the primary cause of ARJP and early onset, recessive parkinsonism. Numerous unique mutations in all 12 exons of Parkin have been identified throughout various ethnic populations. These mutations consist of point mutations and exon rearrangements, including both deletions and duplications [67, 88-92]. To date, approximately 147 different exonic mutations have been described of which a third are single-nucleotide changes, 13% are minor deletions and 54% are larger deletions or duplications comprised of one or more exons [93]. The number of exon rearrangements in *parkin* is likely still to increase as many exon rearrangements were often omitted due to the early labor intensive and expensive methods for identification of exon rearrangements.

Mutations are present in approximately 50% of patients with recessive, EOPD in the age range of 7-58 years of age and present in up to 77% of sporadic cases with disease onset younger than 20 years [94]. *Parkin* is the second largest gene in the human genome and codes for a 465-amino acid protein. It is an E3 ubiquitin ligase [95].

Key clinical features of *parkin* disease have been reported to include age at onset <40 years, foot dystonia, psychiatric symptoms and a dramatic response to treatment [96]. However, these symptoms can mirror those of typical young onset PD cases without *parkin* mutations [97].

The pathology of *parkin* disease consists of severe neuronal loss in the substantia nigra, occasional tau pathology and a distinct lack of postmortem LBs, found in only a minority of genetically confirmed parkin patients [98-100].

A possible explanation for the lack of LBs, a pathological hallmark of PD, may be the young age of *parkin* disease onset [101]. It is notable that no cases of juvenile-onset PD have been reported with postmortem LBs [102, 103] and patients with LBs had a significantly older age of disease onset (mean age of onset: without postmortem LBs = 27 years; with postmortem LBs = 46 years) [101, 104].

1.5.2.1.5 PTEN-Induced Putative Kinase 1

The PARK6 locus was first mapped to a 12.5-centimorgan (cM) region on chromosome 1p35-p36 in a large consanguineous family from Sicily [105]. In 2004, two homozygous mutations were identified in the *PTEN-induced putative kinase 1* (*PINK1*) gene by Valente and colleagues. A p.G309D missense mutation and a p.W437X truncating mutation, were found in a Spanish family and two Italian families respectively. All three families shared a common haplotype, demonstrating a shared ancestry [106].

Both homozygous and compound heterozygous loss of function mutations in the phosphatase and tensin homolog (*PTEN)*-induced putative kinase 1 (*PINK1)* gene are the second most common cause of autosomal-recessive, early on-set Parkinson's disease (EOPD) [106-111]. The clinical phenotype of *PINK1*-related PD strongly resembles levodopa responsive, classic idiopathic PD with no reports of dementia [112].

1.5.2.1.6 DJ-1

The PARK7 locus on chromosome 1p was established through the discovery of a consanguineous pedigree with autosomal recessive PD. In 2003, DJ-1 became the third gene associated with ARPD. Bonifati and colleagues identified recessively inherited missense and exonic deletions in DJ-1 in two European families [113]. Homozygosity mapping and positional cloning performed on a consanguineous pedigree from a genetically isolated population in the Northern Netherlands revealed a homozygous deletion of several exons in DJ-1 causing disease. Subsequently, a missense mutation in a highly conserved residue (Leu166Pro) of DJ-1 was found to cause disease in an Italian ARPD family. Mutations in DJ-1 are extremely rare, found in <1% of early-onset PD cases. The mutations are found in both homozygous and compound heterozygous states, resulting in loss of protein function. Phenotypically, DJ1 mutations cause Levodopa responsive disease onset in the mid-twenties, resembling Parkin and PINK-1 linked forms [113-119].

The DJ1 gene spans 24 kb in length and includes 8 exons encoding a domain protein of 189 amino acids [120]. DJ-1 belongs to the pepidase C56 family of proteins and has been reported to protect cells against oxidative stress and to play a role in maintaining normal dopaminergic function in the nigrostriatal pathway [115, 121-126]. Besides the importance of DJ-1 in dopamine neurotransmission and signaling, it has been reported to have multiple functions associated with PD pathogenesis, such as chaperone activity and the ability to inhibit α-synuclein aggregation, which is thought to be a key event in Lewy body formation [127]. It has also been suggested that DJ-1

may be involved in transcriptional regulation of neuroprotective or anti-apoptotic genes [128].

1.5.2.1.7 PLA2G6, ATP13A2, and FBXO7

More rare, recessively inherited forms of PD are caused by mutations in three genes: ATP13A2 (ATPase type 13A2), PLA2G6 (phospholipase A2, group VI) and FBX07 (F-box only protein 7). Mutations in *ATP13A2* cause the rare, juvenile-onset disorder Kufor-Rakeb syndrome, which is characterized by a lower response to levodopa and additional atypical features of disease such as dystonia and supranuclear palsy [129, 130]. Mutations in the gene *PLA2G6* cause autosomal recessive, levodopa-responsive parkinsonism with dystonia [131]. Brain iron accumulation is found in most but not all affected individuals [131-133]. Shojaee and colleagues first identified mutations in *FBXO7* through linkage mapping followed by gene sequencing in an Iranian family with AR recessive juvenile-onset parkinsonian-pyramidal syndrome [134].

1.5.2.2 Genetic Risk Factors in Parkinson's disease

During recent years several susceptibility genes and numerous risk loci associated with PD have been identified. In this section, I will discuss how variability associated with genetic risk for PD was first identified through candidate gene-based assessments in three genes: *SNCA*, *LRRK2* and *GBA*. In the Results section, I will outline how GWA studies extended upon these

three susceptibility genes and our known understanding of risk to help identify genetic risk loci for PD. The work associated with the identification of several novel risk loci will be discussed in the results section.

1.5.2.2.1 Alpha-synuclein

As previously noted, families with rare mutations in alpha-synuclein enabled the novel discovery of a genetic link to PD [31]. Following this hallmark finding, Kruger and colleagues examined common variability within alpha-synuclein to establish whether the gene was also associated with risk for the sporadic form of PD [135]. Kruger's study initially reported that APOE genotype, a major risk factor for late-onset Alzheimer's disease [136], interacted with a variable dinucleotide repeat within SNCA. The combination of the APOE4 allele and NACP allele 1 of the promoter polymorphism were shown to be significantly different between sporadic PD patients and controls. PD patients presenting this genotype had a 12.8-fold increased relative risk for developing PD over the course of their lives. Unfortunately, this interaction between synuclein and APOE genotypes was not replicated; however, risk for PD within alpha-synuclein was later demonstrated by Maraganore and colleagues using meta-analysis of existing REP1 genotype data [137]. Maraganore showed an unequivocal association between genetic variability within the *SNCA* locus and PD [137]. Since then, association of Parkinson's disease with Alpha-synuclein has been overwhelmingly established in GWA studies, revealing more about the architecture of genetic risk at this locus. The

Results section of this thesis will examine these GWAS studies in greater detail.

1.5.2.2.2 Leucine Rich Repeat Kinase 2

As with SNCA, subsequent to the identification of LRRK2 mutations as a cause of PD [60, 61] common variability across LRRK2 was examined in several populations. Within Asian populations, the variant p.G2385R was first identified as a cause of PD [138]. However, this variant was present in 5% of the population and later shown to be a risk allele that doubled the risk of PD in individuals [139]. This finding was replicated extensively in Asian populations including those from Singapore, Taiwan, China, Korea and Japan [39, 140-150]. An additional variant described in 2008, p.R1628P was also shown to be associated with a ~2 fold increase risk for developing PD and has been replicated in several Asian populations including Thai, Chinese and Taiwanese populations [141, 151-154]. Several additional variants within LRRK2 have been assessed and have varying levels of support for association with risk for PD [152, 155].

1.5.2.2.3 Glucocerebrosidase

Remarkably, thorough clinical observation versus a previously known genetic association lead to the discovery of PD risk variants within the gene encoding glucocerebrocidase (GBA); a gene long tied to the autosomal recessive lysosomal storage disorder, Gaucher's disease [156]. Tayebi and colleagues observed that a portion of Gaucher's disease patients manifested with

parkinsonism, compelling an early hypothesis that GBA deficiency may lead to a predisposition to parkinsonism [157, 158]. This idea was sussed out a year later with convincing evidence. Aharon-Peretz and colleagues were able to show that a single mutation in GBA increased the risk for PD [159]. Further, meta-analysis of existing data was later used to show in Ashkenazi Jewish populations the frequency of two common mutations in GBA (p.L444P and p.N370S) was 15% in PD and 3% in controls; whereas non-Ashkenazi Jewish populations demonstrated a 3% frequency of these mutations in cases and <1% in controls [160]. Overall, these data indicate that a single mutation in GBA escalates the risk for PD ~5 fold, while remaining inadequate to cause Gaucher's disease. These two variants have also been linked to risk for dementia with Lewy bodies and PD with dementia [161].

Table 1. Loci involved in monogenic forms of PD and risk loci identified prior to the advent of genome wide association studies.

***Not yet verified.**

Locus	Gene	Protein	Model
Park1	*SNCA*	α-synuclein	Autosomal Dominant
Park2	*PARK2*	Parkin	Autosomal Recessive
*Park3**	*unknown*	unknown	Autosomal Dominant
Park4	*SNCA*	α-synuclein	Autosomal Dominant
*Park5**	*UCHL1*	Ubiquitin c terminal hydrolase	Autosomal Dominant
Park6	*PINK1*	Pten-induced putative kinase 1	Autosomal Recessive
Park7	*PARK7*	DJ-1	Autosomal Recessive
Park8	*LRRK2*	Leucine rich repeat kinase 2	Autosomal Dominant
Park9	*ATP13A2*	lysosomal type 5 ATPase	Autosomal Recessive
Park10	*unknown*	unknown	Risk locus
*Park11**	*GIGYF2*	GRB interacting GYF protein 2	Autosomal Dominant
Park12	*unknown*	unknown	X-linked
*Park13**	*HTRA2*	HTRA serine peptidase 2	Autosomal Dominant
Park14	*PLA2G6*	Phospholipase A2	Autosomal Recessive
Park15	*FBXO7*	F-box only protein 7	Autosomal Recessive
Park17	*VPS35*	Vacuolar protein sorting 35	Autosomal Dominant
*Park18**	*EIF4G1*	Eukaryotic translation initiation factor 4 gamma 1	Autosomal Dominant
*Park19**	*DNAJC16*	DNAJ/HSP40 homolog subfamily C member 6	Autosomal Recessive
-	*SNCA*	α-synuclein	Risk locus
-	*LRRK2*	Leucine rich repeat kinase 2	Risk locus
-	*GBA*	Glucocerbrocidase	Risk locus

1.6 Understanding Pathobiology in Complex Disease

Using traditional gene cloning to find genetic mutations provides a penetrant, often coding mutation with which to model disease. The path of biological investigation in these cases, while difficult, is clear. Creation of model systems that parallel some aspect of the aberrant gene have traditionally been used as tools to study the disease mechanism. The same is not true for GWA studies. In comparison to highly penetrant alleles associated with monogenic disease, the risk alleles implicated by GWA are associated with small affect sizes. Although a substantial gap still exists between SNP associations from GWA studies and understanding how loci contribute to disease, clues are emerging through the study of gene expression and epigenetic mechanisms, such as DNA methylation. Given that many GWA loci do not map to coding changes or protein open reading frames, it is likely that a great deal of biologically and clinically important genetic variation exerts pathobiological effect through differential gene expression and/or splicing, rather than point mutations in protein sequence. In this manner, genetic variability can have a direct impact on gene expression either quantitatively or qualitatively. Gene expression QTL mapping has been used in an attempt to catalog, map and understand these effects; however, an intermediate and plausible effect of genetic variability could be the influence on transcriptional potential and transcriptional assignment through varying levels of critical epigenetic mark, local DNA methylation.

1.6.1 Epigenetics

In addition to a direct influence on gene expression, genetic variability could affect gene expression and/or splicing through epigenetic mechanisms. Epigenetics is the study of heritable changes in gene function caused by mechanisms other than changes in the underlying DNA sequence. Epigenetic modifications, are heritable but potentially reversible, may alter throughout life and can be affected by the environment, such as lifestyle, diet and toxin exposure [162]. The study of epigenetics is an expanding field of research where technical breakthroughs have recently allowed the success of large-scale epigenomic studies. For example the discovery of CpG island shores was made [163], the human methylome was characterized at single nucleotide resolution [164], the putative identification of non-CpG methylation was made [165], and the roles of novel histone variants and modification have been defined [166-168]. Two major categories of epigenetic modifications that initiate and sustain epigenetic change are chemical modifications to the cytosine residues of DNA (DNA methylation) and chemical modifications to histone proteins associated with DNA (histone modifications). Additional epigenetic modifications include the effect of small and non-coding RNA mediated regulation. Functionally, signatures of these epigenetic modifications can serve as epigenetic indicators representing gene activity and expression as well as chromatin state [169, 170].

Histone proteins and DNA form a complex of chromatin that comprises chromosomes. When histone proteins are modified via post-translational modification, they can influence how chromatin is arranged and can therefore

determine whether the associated chromosomal DNA will be transcribed. Histone proteins act as a spool around which DNA tightly winds and is bundled into the nucleus. This repeating DNA-histone complex consists of 146 base pairs of double-stranded DNA wrapped around eight histone proteins and is called a nucleosome. Generally, tightly folded or condensed heterochromatin tends to be inactive, or not expressed, while more open euchromatin tends to be active, or expressed [169]. Histones can be modified by the addition of an acetyl or methyl group to the amino acid lysine that is located in the histone. Acetylation is generally associated with euchromatin, while deacetylation is more associated with heterochromatin. Conversely, histone methylation can be a marker for both active and inactive regions of the chromatin. For example, X-chromosome inactivation in females is achieved by methylation of a distinct lysine (K9) on a specific histone (H3) that marks silent DNA and is spread throughout heterochromatin. On the other hand, methylation of a different lysine (K4) on the same histone (H3) is an indicator of active genes [169].

Enzymes and different forms of RNA can also transform chromatin. RNA in the form of antisense transcripts, noncoding RNAs or RNA interference can turn off gene expression. These may influence gene expression by causing heterochromatin to form or by activating histone modifications and DNA methylation [169, 171].

It is important to emphasize that the observed outcome of epigenetic modifications is the sum of their interactions and feedback mechanisms.

However, the study of DNA methylation alone has the ability to convey important epigenetic information by distinguishing regions of transcriptional silence or transcriptional potential. Because only a subset of potential target CpG sites are methylated within the genome, the signature of methylated sites is easily distinguished, making the study of CpG methylation attractive, especially on a genome-wide level. CpG methylation serves as an area of focus in this thesis and is discussed in more detail below.

1.6.1.1 DNA methylation

DNA methylation is an important epigenetic regulator of chromatin structure and function making it a key regulator of gene expression, splicing, growth, and differentiation in virtually all tissues, including brain [172]. DNA methylation is perhaps the most widely studied epigenetic modification and is the oldest epigenetic mechanism known to correlate with gene expression [173]. In its most fundamental form, DNA methylation consists of the covalent addition of a methyl group (CH_3) at the 5-carbon of the cytosine ring within the context of a CpG dinucleotide, resulting in 5-methylcytosine (5-mC). The addition of methyl groups is governed at several distinct levels in cells and is carried out by a family of enzymes called DNA methyltransferases (DNMTs). The combined action of three DNMTs (DNMT1, DNMT3A, and DNMT3B) mediate the establishment and maintenance of DNA methylation patterns. DNMT1 is responsible for the preservation of established patterns of DNA methylation, while DNMT3A and DNMT3B facilitate the formation of de novo

methylation patterns [174]. The added CH_3 group projects into the major groove of DNA inhibiting transcription; therefore, 5-mC is associated with the repression and silencing of gene expression. DNA methylation is central to genomic imprinting and X chromosome inactivation. In human, 5-mC is present in roughly 1.5% of genomic DNA [175] .

In general, the CpG sites within the landscape of genomic DNA in mammals tend to be methylated [176]. The distribution of DNA methylation throughout the genome shows enrichment at non-coding regions and interspersed repetitive elements, but not in CpG islands of active genes [177]. CpG islands are clusters of CpG dinucleotides that have a strong association with gene promoters and housekeeping genes [178]. CpG islands are largely un-methylated throughout the genome in normal cells, allowing access to the transcriptional machinery, facilitating transcription. Thus, while an unmethylated CpG island in a gene promoter does not necessarily mean active expression of the associated gene, it does suggest there is transcription potential. There are approximately 30,000 CpG islands in the human genome, and recent studies have identified a growing number of methylated islands in non-pathological somatic tissues [179].

Traditionally, a CpG island is defined as having a G+C content greater than 50%, an observed versus expected ratio for the occurrence of CpGs of more than 0.6 and a minimum size of 200bp. However, the definition of CpG island has and continues to evolve. A recent study revised the traditional rules of CpG island prediction in order to exclude other GC-rich genomic sequences such as Alu repeats. In comparison to previous definitions, it was shown that

DNA regions immediately 5' to genes with a G+C content >55% and an observed versus expected ratio of CpG dinucleotides of >0.65, both in a track of 500bp or longer, are more likely to be true CpG islands [180]. Seventy-five percent of transcription start sites (TSS) and 88% of active promoters are associated with CpG–rich sequences [181]. Although, research has typically focused on CpG islands spanning the 5′ end of the regulatory region of genes, it is now evident that variation in methylation occurs more often in the 'shores' of CpG islands versus within the islands themselves. It appears that around 76% of methylated sites occur a short distance away from CpG islands, with only six percent found within the islands themselves. Interestingly, most tissue-specific DNA methylation occurs in these CpG island shores, up to 2,000 base pairs away from CpG islands [163]. Additionally, a recent study revealed an important role for intergenic DNA methylation in the regulation of alternative promoters within gene bodies [182]. Intergenic methylation appears to modulate gene expression and splice variants and CpG islands in introns can serve as promoters for non-coding RNA (ncRNA) regulatory functions [183]. As research focuses on the most widely studied epigenetic modification, the complexity and significance of DNA methylation will continue to be highlighted.

1.6.1.2 DNA Methylation analysis tools

The methylation signature in a genomic DNA sample is complex; it represents the CpG methylation levels from a compilation of cells that were used to provide the DNA sample. Within each cell the DNA methylation signature at an individual site is in one of three states: both parental strands are

methylated, both parental strands are unmethylated, or one is methylated and one is unmethylated. Assessing CpG methylation can be done for the pattern of methylated CpG sites along a sequence for individual DNA molecules or as an average methylation signal at a single genomic locus across many DNA molecules. Techniques to comprehensively characterize DNA methylation patterns are the most highly developed of the epigenetic methods.

Standard molecular biology techniques such as PCR and cloning erase DNA methylation marks; therefore, DNA must be pretreated to reveal the presence or absence of the methyl group at cytosine residues. There are three different initial treatments that can be used: endonuclease digestion, affinity enrichment, and bisulfite conversion. Techniques designed to pretreat DNA for methylation analysis were initially confined to localized regions of the genome; however, many methods now enable DNA methylation analysis on a genome-wide scale, including the bisulfite treatment of DNA. Bisulfite conversion is the most conventional approach for pretreatment and is considered the gold standard for determining DNA methylation status because it offers single CpG resolution [184]. Bisulfite treatment converts unmethylated cytosines to uracil while leaving methylated cytosines unconverted [185-187]. DNA can then be amplified or hybridized to arrays [188, 189].

One microgram of bisulfite-converted DNA can now be used to ascertain quantitative measurements of DNA methylation for up to 450,000 CpG dinucleotides on genome-wide methylation microarrays, such as the Illumina Infinium Human Methylation 450 array. The Infinium HumanMethylation450

Assay uses two different bead types to detect CpG methylation. One bead type matches the unmethylated CpG site, and the other type matches the methylated site. The level of methylation for the interrogated locus is determined by calculating the ratio of the fluorescent signals from the methylated versus unmethylated sites. The field of epigenomics has flourished with the use of microarray hybridization techniques adopted from gene expression and genome-based assays to profile whole-genome DNA methylation patterns [190-193].

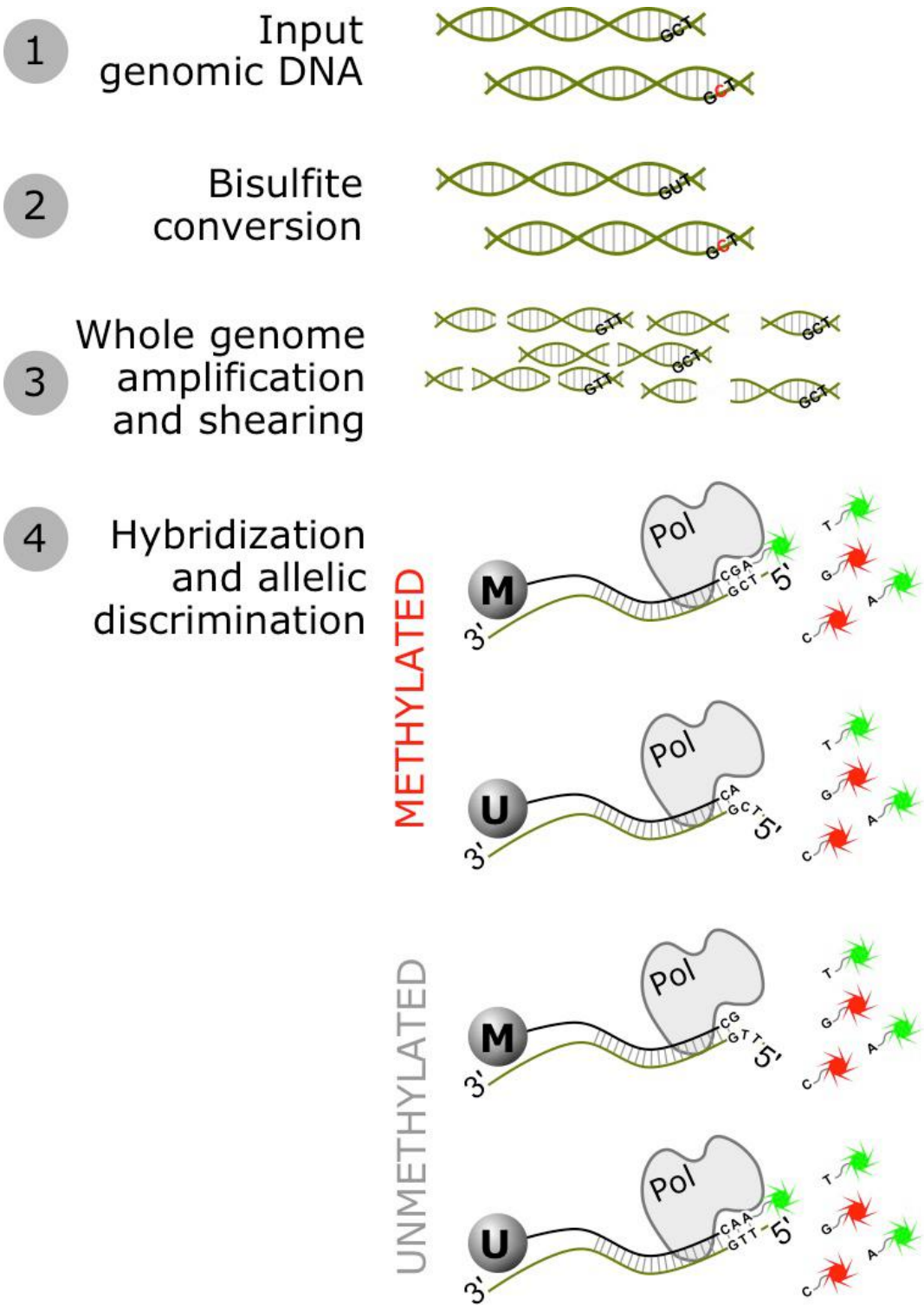


Figure 5. Schematic of the Infinium DNA methylation assay.

Input genomic DNA is bisulfite converted, which leads to deamination of unmethylated cytosine to produce uracil (the methylated cytosine is shown in red). Following conversion the DNA is amplified and hybridized to Illumina beadarrays. There are two classes of bead types, those against the methylated site (M), and those against the unmethylated site (U). For the M type the single base extension will occur where a cytosine is intact, but not where a thymine (created from a uracil during amplification) exists. The converse occurs for the U type bead.

Bisulfite treatment of DNA in conjunction with next-generation sequencing can be used to decode the methylation status of the entire genome [165]. However, due to the high costs currently associated with large-scale sequencing, other methods concentrating on more limited sequencing of genomic regions have been developed, such as reduced representation (RRBS) bisulfite sequencing [194]. RRBS is an approach for large-scale high-resolution DNA methylation analysis, where only a subset of the genome is analyzed. DNA is digested with methylation insensitive restriction enzyme (MspI) to remove much of the unmethylated regions of the genome, subsequently only DNA fragments of a specified length are bisulfite sequenced, consisting mostly of methylated DNA.

Although bisulfite treatment of DNA has long been considered a superior technique for measuring DNA methylation status, it does have disadvantages. Bisulfite conversion typically calls for larger quantities of sample DNA, which can degrade following chemical treatment, it can be limited by incomplete conversion of all unmethylated cytosines to uracils, and bisulfite conversion can not discriminate between methylcytosine and hydroxyl methylcytosine. Alternative assessments of methylation status are based on enrichment of methylated DNA with immunoprecipitation (MeDIP-seq) or affinity purification (MethylCap-seq) and subsequent analysis of enriched sequences using microarrays or sequencing [195].

The HELP assay (HpaII fragment enrichment by ligation- mediated PCR)

pretreats DNA with methylation sensitive and insensitive restriction endonuclease digestion. Subsequent comparative analysis of the resulting fragments using microarray or sequencing is used for the determination of the methylation state of restriction sites [196].

Several methods have been developed to map DNA methylation on a genome-wide scale. These methods, while diverse in technique, have been shown to produce concordant results [197].

1.6.2 Epigenetics and complex disease

Until recently, the majority of epigenetic research focused on the study of cancer and many advances have been made in defining the disease pathogenesis. Not only has global DNA hypomethylation consistently been observed in many cancers but, research shows that all three types of normal epigenetic modifications of DNA, including chromatin modifications, DNA methylation, and genomic imprinting are altered in cancer cells [198, 199]. As the field of epigenetics research has expanded over the last few years, epigenetic alterations have been found to be linked to disorders such as metabolic disorders [200] cardiovascular diseases [201-203] and myopathies [204].

There is also evidence suggesting a relationship between epigenetic alterations and neurological disorders. For example, hyper-methylation of the *FMR1* promoter has been described in Fragile X syndrome [205] and hyper-methylation of gene promoters *FXN* in Friedreich's ataxia, *SMN2* in spinal

muscular atrophy, and neprilysin in Alzheimer's disease (AD) [206]. Conversely, the overexpression of tumor necrosis factor alpha in the substantia nigra of Parkinson's disease (PD) patients is associated with promoter hypo-methylation, inducing apoptosis in neuronal cells [207]. Neurodegenerative disorders such as AD and PD are believed to have a multi-factorial origin arising from a combination of risk factors and susceptibility genes, where age, diet, lifestyle and level of education are all correlated with the onset and severity of the sporadic forms [208, 209]. The mode(s) in which environmental factors and susceptibility genes interact to cause disease are not fully understood; however, epigenetic mechanisms may provide a link between genes and environment.

1.6.3 Epigenetics and Neurological Disease

Our understanding of the roles played by different epigenetic elements in distinct neurodegenerative diseases is still progressing. It remains unclear whether epigenetic dysregulation is involved in pathogeneisis generally or whether epigenetic dysregulation may characterize a common pathway for neurodegeneration.

The analysis of discordant twins in 2005 by Fraga and colleagues provided important results for the field of epigenetics. Fraga et al was able to show that twins sharing common genotype displayed divergent penetrance of numerous diseases, even neurological disorders [210]. Neurodegenerative diseases are one category of neurological disorder. DNA methylation has now been linked

to at least four major neurodegenerative diseases, Alzheimer's disease (AD) [211-213], Huntington's disease [214], amyotrophic lateral sclerosis (ALS) [215] and Parkinson's disease [216, 217].

1.6.3.1 Parkinson's Disease and Epigenetics

In terms of epigenetics, PD has been less studied than other neurodegenerative disorders such as AD. However, epigenetic links to PD are developing. Genes associated with Parkinson's disease have been shown to be regulated by epigenetic mechanisms and to also regulate/modulate the function of certain epigenetic elements. For example, in brain regions such as substantia nigra, putamen and cortex of PD patients, alpha synuclein (SNCA) showed decreased levels of DNA methylation compared to controls [218]. This finding suggests a relationship between hypomethylation of CpG sites in the promoter region of SNCA and increased expression of SNCA. If this study holds true, the relationship between DNA methylation and SNCA is a very important one considering alpha-synuclein is the primary structural component of Lewy bodies; thus contributing greatly to the pathogenesis of PD. It is interesting to note, that Matsumoto and colleagues found methylation levels to be significantly reduced in substantia nigra tissue from PD patients versus healthy controls [219]. They did not find a significant difference in methylation levels between PD and controls in two additional brain regions: the putamen and anterior cingulate [219].

1.6.4 DNA Methylation as a Quantitative Trait

The expression profile of a cell and its response to environmental signals effectively defines its overall phenotype. Using gene expression as a quantitative trait has successfully identified genetic modifiers of gene expression. However, assaying DNA methylation as an intermediary between genetic variation and gene expression patterns provides a new, and unlike gene expression, a stable measure of the cellular phenotype. Quantitative measures of DNA methylation provide a chromatin signature of cellular transcriptional potential that is preserved and can be regenerated during cell division. DNA methylation has been shown to influence gene expression in an age-dependent and tissue-dependent manner [220, 221], characteristics that are potentially important for the study of neurodegenerative disease, where distinct regions of brain tissue and/or cell types are compromised in an age-dependent manner. Therefore, the study of quantitative trait loci that are influencing epigenetic regulators of gene expression such as the covalent modifications of DNA are highly attractive as a means to further explore the molecular pathology of neurodegenerative disease beyond RNA quantification.

As my focus is on the study of age-related neurodegenerative disease, one critical goal is to determine the immediate biological consequences of disease-associated common genetic variation in the human brain. As discussed above, two functional, quantitative variables that can efficiently be investigated from a genome-wide perspective are mRNA expression and DNA methylation. Combining these data with quantitative trait locus (QTL) analysis

allows a systematic, genome-wide, relatively hypothesis free investigation into the effect of common genetic variability on important functional variables.

1.6.5 DNA Methylation and Aging

Aging is the primary risk factor for the majority of neurodegenerative diseases. AD affects 11% of people age 65 or older and 32% of people over age 85 [222]; while PD affects 1–2% of people 60 years of age and over, and increases to 3–5% in people 85 years and older [223]. The fact that neurodegenerative diseases are frequently late-onset implies there is a biological characteristic that changes as a person ages. One unvarying factor for each disease is that neurons steadily lose function as the disease progresses with age. The biochemistry of aging is complex with significant alterations occurring in proteins, lipids and nucleic acids. The process is thought to encompass many dynamic, interacting factors including oxidative DNA damage, nuclear and mitochondrial genome mutations, depletion of stem cells, and shortening of telomeres [224]. In addition to these diverse interactions, there is a strong link between DNA methylation and human aging. Fraga and colleagues describe one aspect of this link in a simple but well-designed study analyzing global and locus-specific differences in the DNA methylation patterns of monozygotic twins. The authors found that younger twins had indistinguishable methylomes, whereas older twins demonstrated significantly divergent methylomes, revealing an epigenetic drift with aging [210]. Differential DNA methylation has been shown to be age-related [225, 226] and methylation of DNA sequences within or near

regulatory elements has been shown to suppress gene expression through effects on DNA binding proteins and chromatin structure [227]. There is global reduction of DNA methylation across the genome with aging in union with hypermethylation of distinct loci, often promoter-associated CpG islands [228, 229].

Indeed, both increases and decreases in DNA methylation are reported to occur with aging contingent upon the tissue and the gene [227]. Bocklandt and colleagues assessed genome-wide methylation levels in both monozygotic twins and non-twin samples and reported 88 CpG sites within or near 80 different genes whose methylation levels substantially changed with age. Methylation of CpG sites within three genes: *TOM1L1, NPTX2* and *EDARADD* were particularly associated with age to the point where the authors were able to build a model from CpG sites within these loci to predict the age of an individual within an error of 5.2 years [230]. These three genes, highly correlated with age, are also implicated in human disease [231].

TOM1L1 expression is reduced in esophageal squamous cell carcinoma [232] and *EDARADD* mutations slow or decrease wound healing [233]. Interestingly, the third gene, *NPTX2* is reported to be upregulated in both pancreatic cancer [234] and sporadic PD [235]. Moran and colleagues report that *NPTX2* is a novel component of Lewy bodies (a pathological hallmark of PD) and is found in close proximity to alpha-synuclein aggregates in the cerebral cortex and substantia nigra (SN) and is profoundly (> 800%) upregulated in the parkinsonian SN [235]. As NPTX2 has an established role in synaptic plasticity as well as dopaminergic nerve cell death [236], Moran

and colleagues hypothesize that NPTX2 is involved in the disease pathogenesis underlying PD.

DNA methylation levels of NPTX2 are used to predict the age of an individual, while it's over expression is tied to neuronal death and PD pathology. Clearly evidence is mounting that the cellular mechanisms associated with aging and those that are related to neuronal degeneration are intertwined. It is possible that aberrant patterns of DNA methylation accumulated during aging promotes or exacerbates pathobiological consequences, perhaps explaining why aged organisms generally have a higher risk for disease. Teschendorff and colleagues also show a degree of overlap between genes differentially methylated during healthy aging and during disease development [229].

Studies illustrate that aberrant methylation can both contribute to disease and also act as a by-product of disease. For example, oxidative stress can modulate the regulation of gene expression by triggering DNA lesions; thus, stimulating genomic hypomethylation by preventing DNMTs from binding to cytosine [237]. This example of aberrant methylation (hypomethylation) as a by-product of disease leads to the prospect that age-related hypomethylation could be the result of free-radical induced DNA damage. Whether or not aberrant methylation and its association with aging are causal for age-related neurodegenerative disease, a by-product of disease or both is not clear. Therefore, it is an important undertaking to understand patterns of DNA methylation in the normal aging brain as a foundation for gaining biological insight into age-related neurodegenerative diseases such as Parkinson's and Alzheimer's diseases. Because this goal is central to my thesis, the last

analysis presented here aims to identify and map the landscape of age-related DNA methylation changes in the context of normal human brain tissue. Ultimately, this type of analysis should be incorporated into models aimed at understanding the genetic, and molecular basis of disease.

2 Genetic risk for Parkinson's disease: meta-analysis of genome wide association studies and replication of results

STATEMENT OF CONTRIBUTIONS TO THIS RESEARCH:

In this section, I describe a series of experiments that were performed to identify and replicate novel risk loci for Parkinson's disease. These experiments encompass several scientific disciplines and are the collective effort of several investigators. I was involved in the inception, planning and design of the experiments and analyses. I performed experimental work and quality control for genome-wide SNP datasets for both the meta-analysis and replication stage. I am a co-author on the manuscript and member of the International Parkinson's Disease Genomics Consortium (IPDGC).

2.1 Introduction

Efforts to understand the genetic basis of PD have resulted in the identification of several genes that contain disease causing mutations, and a number of risk loci [31, 33, 61, 106, 238]. As with other complex late-onset neurodegenerative diseases, this genetic knowledge has served as the foundation of investigation into the molecular pathogenesis in PD. The creation of cell and animal based models has predominantly used genetic manipulation, focusing on *SNCA* and *LRRK2.* In this context, the emphasis has been placed on identifying a unifying pathway for disease gene related dysfunction with the aim of pinpointing a nexus for therapeutic intervention for all forms of PD.

Recent work in hereditary spastic paraplegia (HSP) has illustrated that as the number of known genetic causes of disease increases, this information can be synthesized to better understand the pathobiology of the disease, particularly using functional network analysis. In turn, functional networks can then be used to gain information regarding the genetic basis of disease by nominating candidate genes [239, 240]. For such methods to work, a large number of disease-linked genes are required, and thus, a key goal in PD research is to understand more about the genetic basis of this disease.

In general, GWA studies have been applied as a means of identifying risk loci since the first successful published GWA study in 2005, identifying *CFH* polymorphisms as a significant risk factor for age related macular degeneration [241]. At the time of performing the experiments outlined in this

chapter there were a number of GWA studies completed in PD [31, 242-249]. These early efforts in PD failed to convincingly identify risk loci as these studies were, like many other studies at the time, of low power, only examining ~300 cases [244, 250]. However, in 2009 two collaborative studies examining Caucasian and Asian subjects were the first to reveal genome wide significant risk alleles for PD [247, 248]. The Caucasian study identified risk loci at *SNCA* and *MAPT*, and provided supporting evidence for association at *LRRK2* and *PARK16* [248]. The study in Asian subjects revealed association at *SNCA*, *LRRK2*, *PARK16*, and *BST1* [247]. Both studies genotyped more than 1000 cases in the original stage I (genome-wide) analysis, which became widely characteristic of the sample size required to begin to see genome-wide significant effects using GWA. Over the next several years these loci were replicated, and two additional risk loci were nominated at *HLA-DRB5* and *GAK*. As with GWA detected loci in other complex diseases, the effect sizes for each of these loci are individually modest [31, 242-249]. Until this point, studies were relying on the use of directly typed variants because standard methods for imputation were not yet available.

GWA SNP data is powerful in that data from varied sample series and data generated at different sites can be combined. Thus, the identification of additional risk loci can be facilitated by the analysis of existing genome wide data sets. Here, a meta-analysis of five existing PD GWAS datasets from the U.S. and Europe was pursued by the formation of a collaborative group, the International Parkinson's Disease Genomics Consortium (IPDGC). Discovery

phase results were replicated in a large, independent samples series using a custom array that included the most significant variants from stage I.

2.2 Materials and Methods

The study consisted of two stages: (1) a meta-analysis of five independent GWA studies (2) a replication stage.

2.2.1 Discovery (stage I)

2.2.1.1 Samples

The IPDGC was formed with the express goal of discovering new loci linked to PD through meta-analysis of each consortium member's independent GWA study. The discovery phase of this project included five datasets. Four of these were from the IPDGC including data from U.S. National Institute on Aging, UK, Germany, and France. A fifth dataset was downloaded from the database of genotypes and phenotypes (dbGAP; http://www.ncbi.nlm.nih.gov/gap). This dataset represented the CIDR: Genome Wide Association Study in Familial Parkinson Disease (dbGaP Study Accession: phs000126.v1.p1; CIDR study), a US family based GWA study.

The NIA data set was genotyped using Illumina 550Kv1 and 550Kv3 Arrays as described previously [248]. The CIDR study, accessed via dbGAP was initially genotyped at the Center for Inherited Disease Research at Johns Hopkins University (CIDR) using Illumina 370K arrays [246]. The German study comprised PD samples collected from the University of Munich and the University of Tuebingen in addition to controls from the KORA and Popgen studies [251, 252]. These samples were genotyped using Illumina 550Kv1 arrays as previously described [248]. The United Kingdom samples were

derived from movement disorder centers within the UK. Controls were genotyped by the Wellcome Trust Case Control Consortium (1958 Birth Cohort, and blood donors). Cases were genotyped with the Illumina Human660-Quad BeadChip and controls were genotyped with the Illumina 1.2M Duo BeadChip as previously described [242]. Samples from the French group were ascertained from across 15 university hospitals in France, the controls were derived from the French Three City Cohort (http://www.three-city-study.com). Summaries of these datasets are shown in Table 2.

2.2.1.2 Quality Control Procedures

The stage I datasets each went through a series of quality control steps, and these were performed separately for each cohort at the respective center. Initial quality control centered on the exclusion of poor performing SNPs and samples; following clustering within BeadStudio (Illumina, CA), data were exported in a standard Final Report format and imported into the PLINK toolset (http://pngu.mgh.harvard.edu/~purcell/plink/; version 1.07). First, SNPs that failed in more than 5% of samples were removed; second any sample with a call rate of less than 95% was removed. Any variant that was out of Hardy-Weinberg equilibrium based on a p value threshold of $<1x10^{-5}$ in controls and $<1x10^{-7}$ in cases was removed. Next, the missingness rate for each SNP was compared between cases and controls, any variant with skewed missingness based on $p<1x10^{-4}$ was removed (based on χ^2 test). Self reported gender was then compared to genotype gender, which was predicted based on heterogeneity of SNPs from the non-PAR of the X chromosome. Samples discordant for gender were removed. The next quality control steps

centered on removing samples based on excessive population stratification and cryptic relatedness, followed by the calculation of population substructure to be used as a covariate in the analysis. First, for each cohort individually the samples were merged with public genotype data from the HapMap phase II for samples of Caucasian, Japanese, Chinese, and African ancestry (CEU, JPT, CHB, and YRI respectively; http://hapmap.ncbi.nlm.nih.gov). For the overlapping SNPs between the cohort and the HapMap samples, LD pruning was performed to remove any SNP with an r^2 of >0.2 with any other SNP. Samples were then clustered based on multidimensional scaling (MDS; performed using the R package; http://www.r-project.org). We removed any sample that was > 3 standard deviations from the mean component vector estimates for the first two components from this MDS for the CEU group. The remaining samples were then tested for cryptic relatedness, for any pair of samples that shared more than 15% of alleles, a single member of the pair was removed. Again, this was performed individually for each cohort, with the exception of the NIA and CIDR cohorts, which were both of North American origin, and for which samples had been drawn in part from the NINDS Neurogenetics Repository at Coriell (https://catalog.coriell.org/1/NINDS/). These two groups were combined for this analysis, and this resulted in the removal of 42 duplicates between the cohorts. At this point components 1 and 2 were calculated for each case-control cohort, with the exception of the UK dataset, and these measures were recorded for use as covariates in the study level association analyses.

Following these quality control steps the total number of meta-analysis (Discovery stage) samples was 5,333 PD cases and 12,019 controls Table 2.

Table 2: Characteristics of the stage I cohorts in the meta-analysis of PD GWA studies

	Cases			Controls			Imputed SNPs
	Sample Size	**Mean AAO (SD)**	**Percent Female**	**Sample Size**	**Mean AAE (SD)**	**Percent Female**	
USA-NIA	971	55.9 (15.1)	40.5	3034	62 (15.6)	52.8	7590773
UK	1705	65.8 (10.8)	43.3	5200	NA	49.5	7678643
Germany	742	56 (11.6)	39.8	944	NA	48	7589890
France	1039	48.9 (12.8)	41.2	1984	73.7 (5.4)	33	7340040
USA-dbGAP	876	61.5 (9.2)	40.4	857	NA	60.2	7482040
	5333			**12019**			

2.2.1.3 Imputation

Following sample collection and quality control of collaborator datasets, genotypes imputed for all subjects in a two-stage process in order to flag and exclude imputed SNPs of poor quality. Genotypes for all subjects of European ancestry were imputed using haplotypes derived from low coverage sequencing on 112 European ancestry samples from the 1,000 Genomes Project (University of Michigan Center for Statistical Genetics. 1000G 2009-08 download) [253]. The Markov Chain based Haplotyper (MACH; version 1.0.16) was used

(http://www.sph.umich.edu/csh/abescasis/MACH/download/1000G-Sanger-0908.html). The first stage of this analysis centered on generating error and crossover maps to be used as parameter estimates for imputation. This was performed using a subset of 200 samples from each of the 5 studies, where for each study the 200 samples were randomly selected and the analysis performed 100 times. The estimates derived from these 100 iterations were then used to generate maximum likelihood estimates of allele numbers for each variant for the second round of imputation. In order to remove low quality imputed variants, any variant with an R^2 estimate of less than 0.3 as indicated by MACH was excluded from further analysis. Notably, this method generates non-integer allele counts for each variant for each sample; thus in an individual, each imputed variant is not scored as a minor homozygote, heterozygote, or major homozygote, but rather is represented by a score for minor allele load. While this approach fails to provide categorical genotypes it does incorporate the uncertainty inherent within imputation [10]. This provided a total of between 7,340,040 to 7,678,643 post-QC imputed variants for each the five studies.

2.2.1.4 Meta Analysis

A formal meta-analysis was performed, versus a grouped analysis because, the samples not only were sent from diverse sites, but also because different samples series were genotyped with different array types. A fixed-effects inverse variance weighted meta-analyses was executed using METAL (http://www.sph.umich.edu/csg/abecasis/metal). The standard errors of the beta coefficients were scaled by the square root of study-specific genomic

inflation factor estimates prior to combining the summary statistics across studies in an attempt to control for genomic inflation. A secondary random-effects meta-analyses was performed for each SNP using R, not to generate primary outcome statistics but rather to estimate the possible influence of study heterogeneity on results by generating I^2. Tests of effect heterogeneity (Cochran's Q) were also performed using METAL.

2.2.2 Replication (Stage II)

2.2.2.1 Existing Data

In silico replication data was provided by two consortium members representing case-control cohorts with GWAS data provided after the initial discovery phase. These data were available in Dutch and Icelandic cohorts [249].

2.2.2.1.1 Dutch *in silico* Replication

Genome wide genotyping data comprising 587,388 SNPs typed in 2,082 population control participants from the Rotterdam study III was used. The Rotterdam study is a population based cohort study started in 1990 with focus on the prevalence and determinants of neurologic, ophthalmologic, locomotor and cardiovascular disease prevalence and determinants in the elderly. Subject are 45 years of age or older (http://www.epib.nl/research/ergo.htm). This dataset had been through similar quality control procedures to those

described above, with some minor differences (2.2.1.2). As above, discrepancies between reported and assayed gender were detected, and samples were removed. SNP failure cutoff was set at 5% and individual sample failure rate for genotyping was set at 2.5%. Heterozygosity outliers were excluded from the dataset. Population outliers were removed by clustering with 210 HapMap samples and using a threshold of >4 standard deviations outside of the mean for principle components 1 through 4. As before, sample duplicates, and subjects who were cryptically related were removed by calculating identity by descent distances.

Data for 824 Dutch PD cases genotyped at 559,589 variants was available. These data had also been filtered through a similar quality control procedure.

2.2.2.2 Samples

Three replication cohorts were available for replication genotyping. The replication stage consisted of independent, Dutch, French, German, UK and US cohorts genotyped on the Illumina Immunochip.

The US cohort consisted of 2,807 PD cases and 2,215 controls after quality control, with samples contributed from collaborators at the Parkinson's, Genes and Environment (PAGE) study, the PostCept Study, from investigators at Washington University of Saint Louis (WUSTL), and samples from the NINDS Neurogenetics Repository at Coriell.

The UK cohort consisted of 1,271 cases contributed by the University College London, Cardiff University and 1,864 Wellcome Trust population control samples. Dutch case-control samples that were assayed on the ImmunoChip,

included 304 cases and 402 controls. An additional 1,153 PD cases and 712 controls were included from collaborators at the Universität Tübingen. 267 French cases and 363 French controls were provided to complete the replication series. Following quality-control procedures, the independent replication sample set included 7,053 PD cases and 9,007 controls.

Assay Design

The ImmunoChip was designed as a cost effective solution for replication of GWA loci across a variety of traits [254]. The ImmunoChip is based on the Illumina Infinium genotyping chip, and contains 196,524 polymorphisms (718 small insertion deletions, 195,806 SNPs). Wellcome Trust Case-Control Consortium (WTCCC) created this array as a cost and labor effective solution to replicate and fine map GWA identified loci across a large number of traits; primarily related to major autoimmune and inflammatory diseases. The UK portion of the GWA in PD was part of the WTCCC; therefore, a segment of the ImmunoChip was used to type PD related variants. A total of 1920 PD related SNPs were placed on the ImmunoChip by the IPDGC for analysis, with the remainder of variants available for use in quality control procedures. Of these 1920 SNPs, 1200 were genotyped for the purpose of replication of signals identified within the meta-analysis.

2.2.2.3 Replication Genotyping

Genotyping was performed at the NIH per manufacturers protocol [255]. Briefly, A total of 200-400ng of genomic DNA for each sample was suspended in TE solution (10mM Tris, 1mM EDTA) and normalized to 50ng/ul. The

samples were then denatured with NaOH and neutralized before amplification. The denatured DNA was placed at 37 degrees Celsius for 20-24 hours to isothermally amplify the amount of DNA by several thousand-fold. Following amplification, the DNA was enzymatically fragmented to an endpoint at 37°C for 1 hour. An isopropanol precipitation was then performed and the DNA collected by centrifugation at 4°C, followed by resuspension in hybridization buffer (RA1, Illumina). The re-suspended DNA was then hybridized to the Immuno BeadChips by incubation overnight at 48°C where the amplified and fragmented DNA samples annealed to locus-specific 50-mers. The following day the unhybridized and non-specifically hybridized DNA was washed off the BeadChips using RA1 solution. The BeadChips were then 'X-stained' using a Tecan Freedom EVO robot (Tecan, Switzerland). Here, single-base extention of the oligos on the BeadChip occurs using the hybridized DNA as a template, incorporating detectable labels on the BeadChip and enabling the genotype call for each sample to be read on the iScan (Illumina, San Diego). After X-staining and before chips are placed on the iScan, the BeadChips were washed in PB1 solution (Illumina) and coated with XC3 polymer (Illumina) then vacuum-dried for 1 hour. The BeadChips were read on the iScan (Illumina, Inc San Diego), which uses a laser to excite the fluorophore of the single-base extension on the beads, creating a high-resolution image of the emitted light.

2.2.2.3.1 Quality Control of ImmunoChip Genotyping

The Genotyping Analysis Module within Genome Studio version 1.9.4 was used to analyze data read by the iScan (illumina). The threshold call rate for

sample inclusion was 97%. All subjects with call rates greater than 97% were included in the analysis. Quality control of sample handling was determined by comparing the subjects' reported gender with the genotypic gender determined using Genome Studio's 'estimate gender' algorithm. 'Estimate gender' determines genotypic gender based on heterozygosity across the X chromosome. Samples with gender discrepancies were excluded from the analyses. Prior to our secondary round of quality control and the generation of covariates to adjust for population sub-structure, the regions of interest outside of the 11 genome-wide significant loci (+/-2MB) and all A/T and G/C SNPs were removed.

2.2.2.4 Statistical Analyses

For the replication stage, SNPs that passed genome-wide significance (fixed effects $p<5x10^{-8}$) and quality control on the Immunochip array (Illumina, San Diego, CA USA) were included. SNPs with inconsistent results across the datasets ($I^2>75\%$) were removed [256, 257].

Due to technical constraints of the Immnunochip assay and the likelihood that not all SNPs would genotype successfully, several proxy SNPs were included in the design of the chip to capture the association signal at each locus. After quality control of SNPs generated from Immunochip experiments, either the originally nominated SNP from each locus, or the best proxy SNP was selected, contingent upon genotyping at the proxy SNP surpassing quality control threshold. All proxy SNPs were in strong LD with main effects.

Specificity of genotyping was confirmed by visual inspection of clusters within Genome Studio Software (Figure 9, Figure 10).

2.3 Results

2.3.1 Discovery (stage I) Meta Analysis

Based on an assessment of the fixed effects p values we identified 12 loci that exceeded the threshold for genome-wide significance in the discovery phase ($p<5x10^{-8}$) (Figure 6). A single locus on chromosome 17 showed an extremely high level of heterogeneity across studies, possibly indicative of an experimental artifact due to different genotyping platforms, distinctly sourced cases and controls, or batch effects in the execution of the discovery genotyping. This variant was removed from further analysis. The remaining 11 genome-wide significant loci included six previously identified PD GWA loci: *SNCA*, *MAPT*, *BST1*, *LRRK2*, *GAK* and *HLA-DRB5* (Table 3) [245, 247, 248]. In addition, five putative novel loci were identified at; *GBA/SYT11*, *ACMSD/TMEM163*, *STK39*, *MCCC1*, and *CCDC62*. As indicated by the heterogeneity of effect scores all loci showed consistent effects across studies, with the exception of previously identified *BST1* locus ($I^2 = 74.7$; $p = 0.0041$); however, the p value threshold did not meet our criteria for exclusion. *BST1* had been previously identified using genotypes not included

in the current meta-analysis; thus, we chose to leave this association in our study for replication.

Locus plots for each of the associated loci are shown in the appendix (Supplementary Figure 1 to Supplementary Figure 11, pages 214 though 224 inclusive).

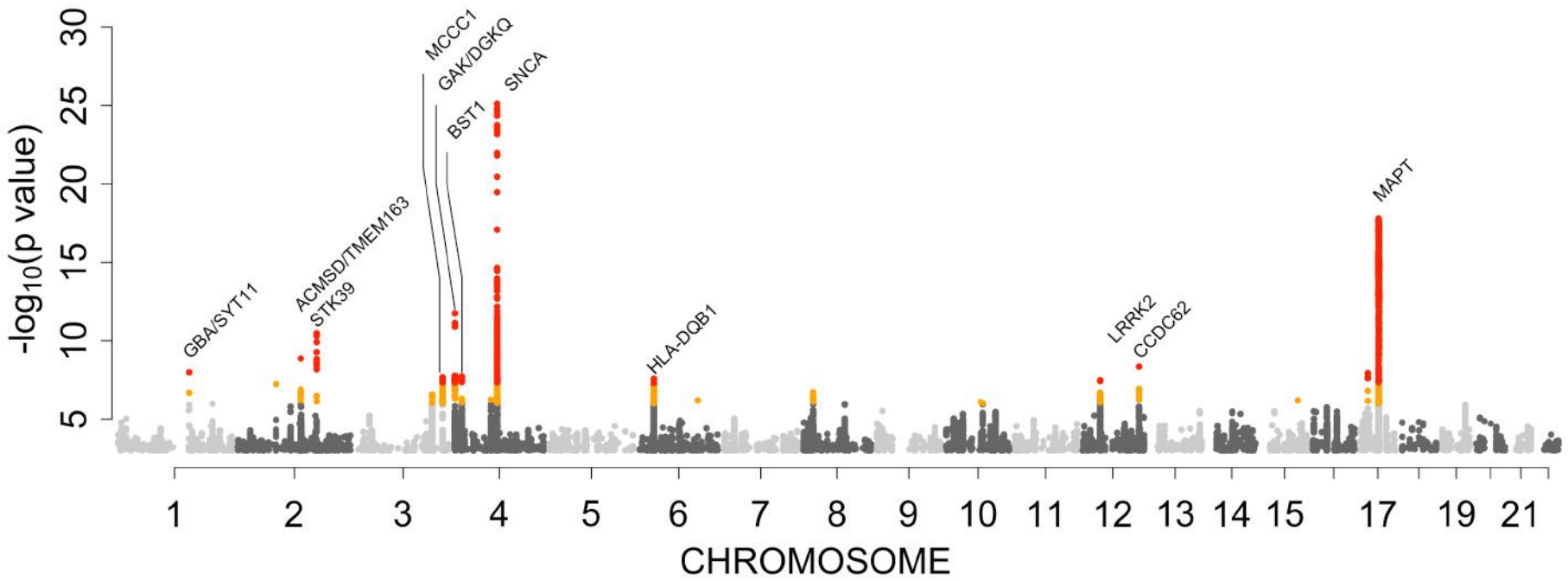


Figure 6: Manhattan plot generated from stage I (discovery) of a meta-analysis of PD GWA studies.

Shown in orange are SNPs with a p value of <1x10^{-6} and >5x10^{-8} (suggestive loci), and in red are SNPs with a p value of <5x10^{-8} (genome wide significant loci). Each of the genome wide significant loci is labeled with the nearest or most

Table 3: Genome wide significant results for discovery phase genotyping, and the results of replication of these loci.

Locus information				Discovery Phase Results						Replication Phase Results		
Most Significant SNP in Region	**C**	**Position (bp)**	**Candidate Gene**	**OR**	**SE**	**Fixed Effects P-value**	**Random Effects P-value**	**I^2**	**Het P-value**	**OR**	**SE**	**Fixed Effects P-value**
chr1:154105678	1	154105678	*SYT11*	1.67	0.09	1.02E-08	5.70E-09	0	0.77	1.44	0.08	1.18E-06
rs6710823	2	135308851	*AMCSD*	1.38	0.05	1.35E-09	1.61E-05	48.26	0.11	1.07	0.02	0.003161
rs2102808	2	168825271	*STK39*	1.28	0.04	3.31E-11	1.54E-11	0	0.72	1.12	0.04	0.001639
rs11711441	3	184303969	*MCCC1/LAMP3*	0.82	0.04	2.10E-08	1.17E-08	0	0.97	0.87	0.03	6.92E-05
chr4:911311	4	911311	*GAK*	1.21	0.03	1.80E-12	2.96E-07	51.58	0.09	1.14	0.02	7.46E-08
rs11724635	4	15346199	*BST1*	0.87	0.03	1.85E-08	0.001407	74.77	4.1E-03	0.87	0.02	2.43E-09
rs356219	4	90856624	*SNCA*	1.30	0.03	7.90E-26	1.11E-26	0	0.58	1.27	0.02	4.23E-23
chr6:32588205	6	32588205	*HLA-DR*	0.70	0.06	2.58E-08	1.44E-08	0	0.88	0.80	0.04	9.30E-08
rs1491942	12	38907075	*LRRK2*	1.19	0.03	3.23E-08	5.24E-06	35.52	0.20	1.30	0.05	1.06E-08
rs12817488	12	121862247	*CCDC62/HIP1R*	1.16	0.03	4.43E-09	2.99E-06	34.97	0.20	1.13	0.03	9.06E-07
rs2942168	17	41070633	*MAPT*	0.76	0.03	1.62E-18	3.91E-19	0	0.74	0.80	0.03	1.37E-13

C – chromosome; OR – Odds ratio per dose of the minor allele; SE – standard error of the odds ratio; Het P-value – heterogeneity p value

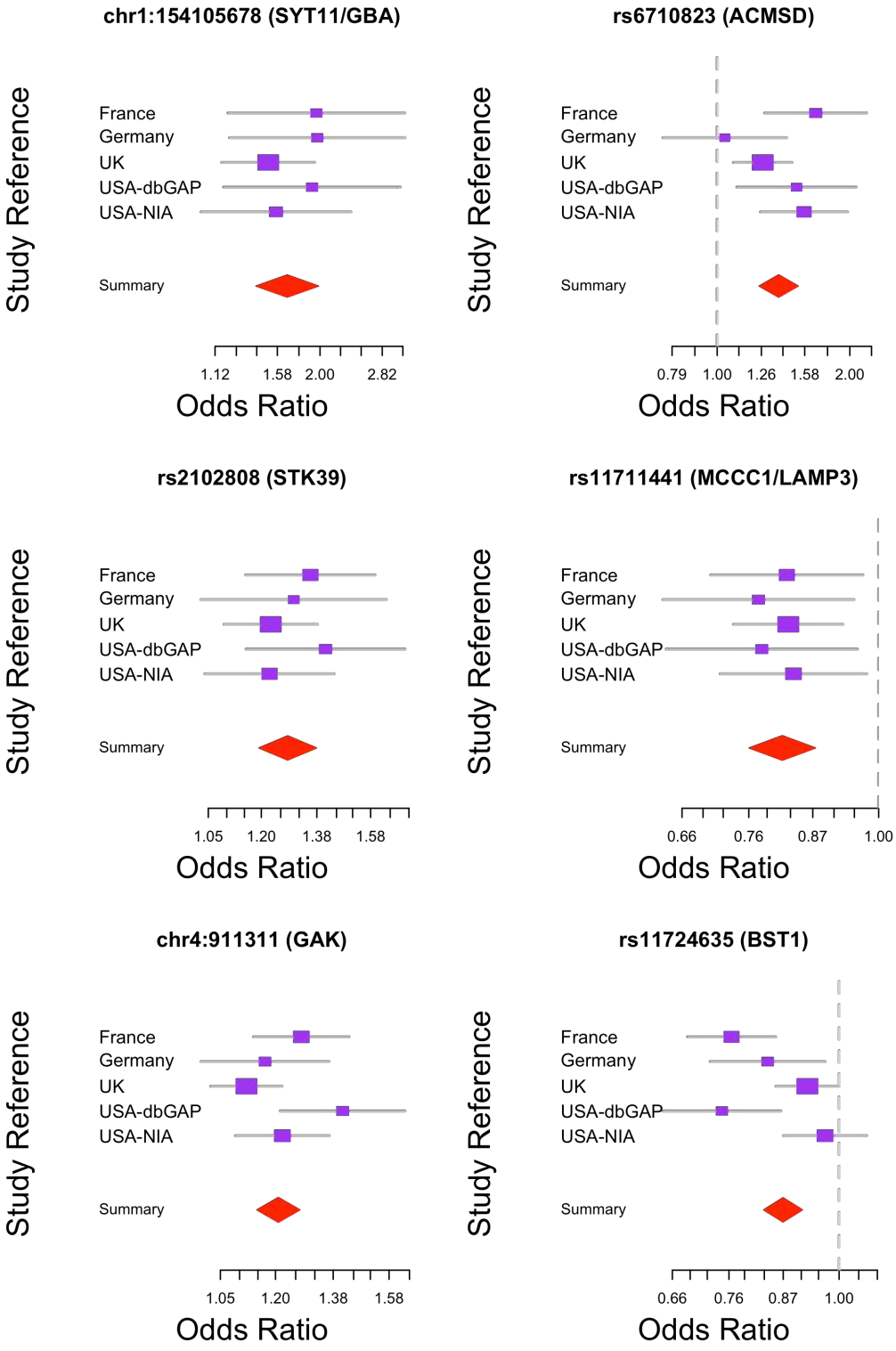


Figure 7: Forest plots showing discovery phase results from a meta-analysis of GWA studies.

These plots illustrate the lack of heterogeneity of effect across studies, even for *BST1*, suggesting the loci are genuine, and are generalizable across populations.

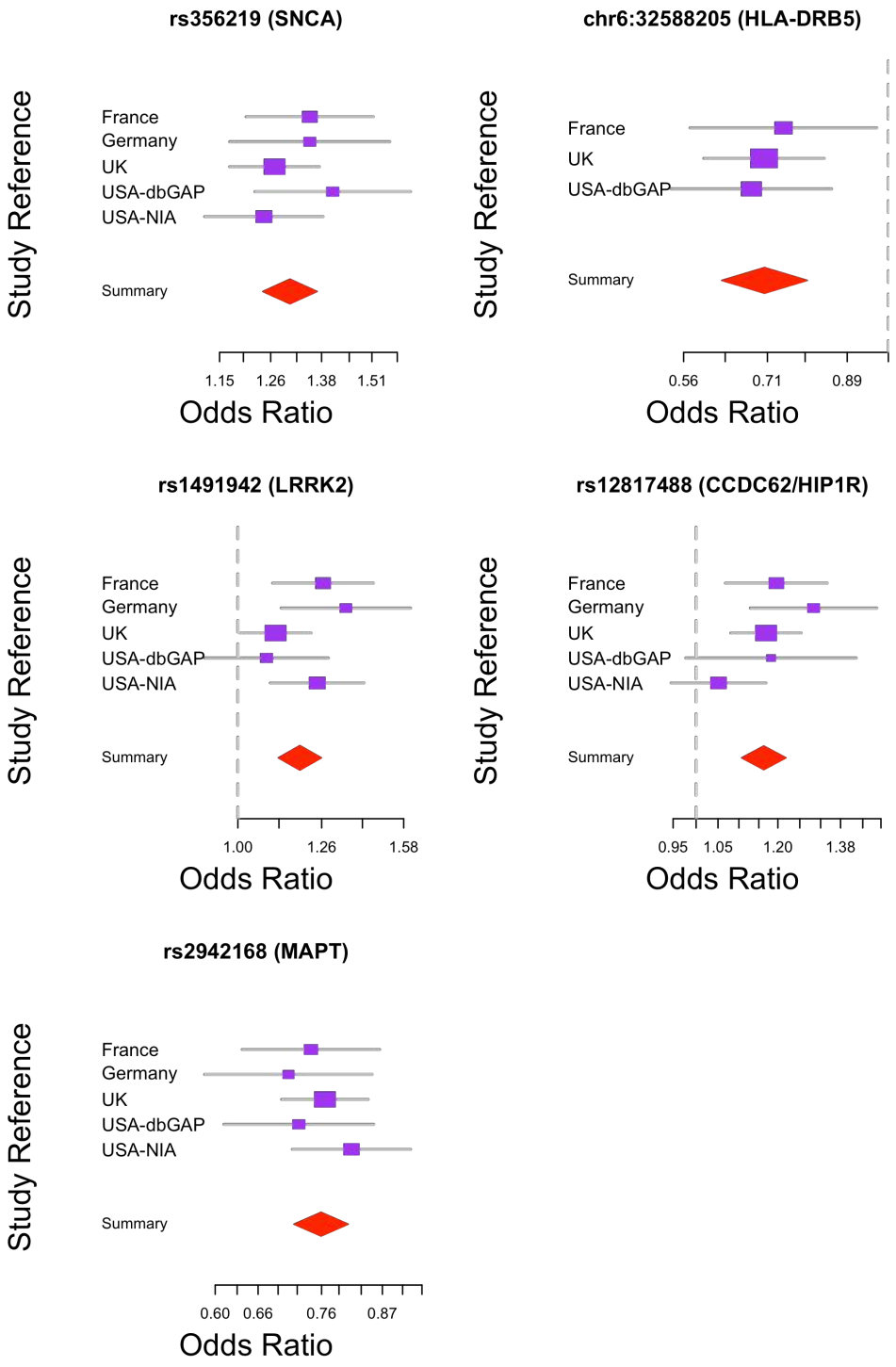


Figure 8: Forest plots showing discovery phase results from a meta-analysis of GWA studies

These plots illustrate the lack of heterogeneity of effect across studies, suggesting the loci are genuine, and are generalizable across populations

2.3.2 Replication of Loci

The most associated variants at the eleven identified genome wide significant loci were examined using the ImmunoChip data in an independent set of 7,053 cases and 9,007 cases. Cluster plots were visually inspected within BeadStudio (Illumina, Inc). This revealed that 2 of these SNPs had failed to perform well. For these variants the proxy markers that had been included on ImmunoChip at these loci were selected (Figure 9 and Figure 10).

Analysis of the resulting genotypes at these loci confirmed the discovery phase signals based on a Bonferroni corrected replication p value cutoff of < 0.0045 (0.05 divided by 11 independent tests) (Table 3). For each of these loci the direction of effect was consistent with that seen in the discovery phase and displayed low heterogeneity across studies (Figure 11 and Figure 12). Consistent with previous GWA studies in complex late-onset neurodegenerative diseases, the effect sizes detected were modest, ranging from 0.8 to 0.87 and from 1.07 to 1.44.

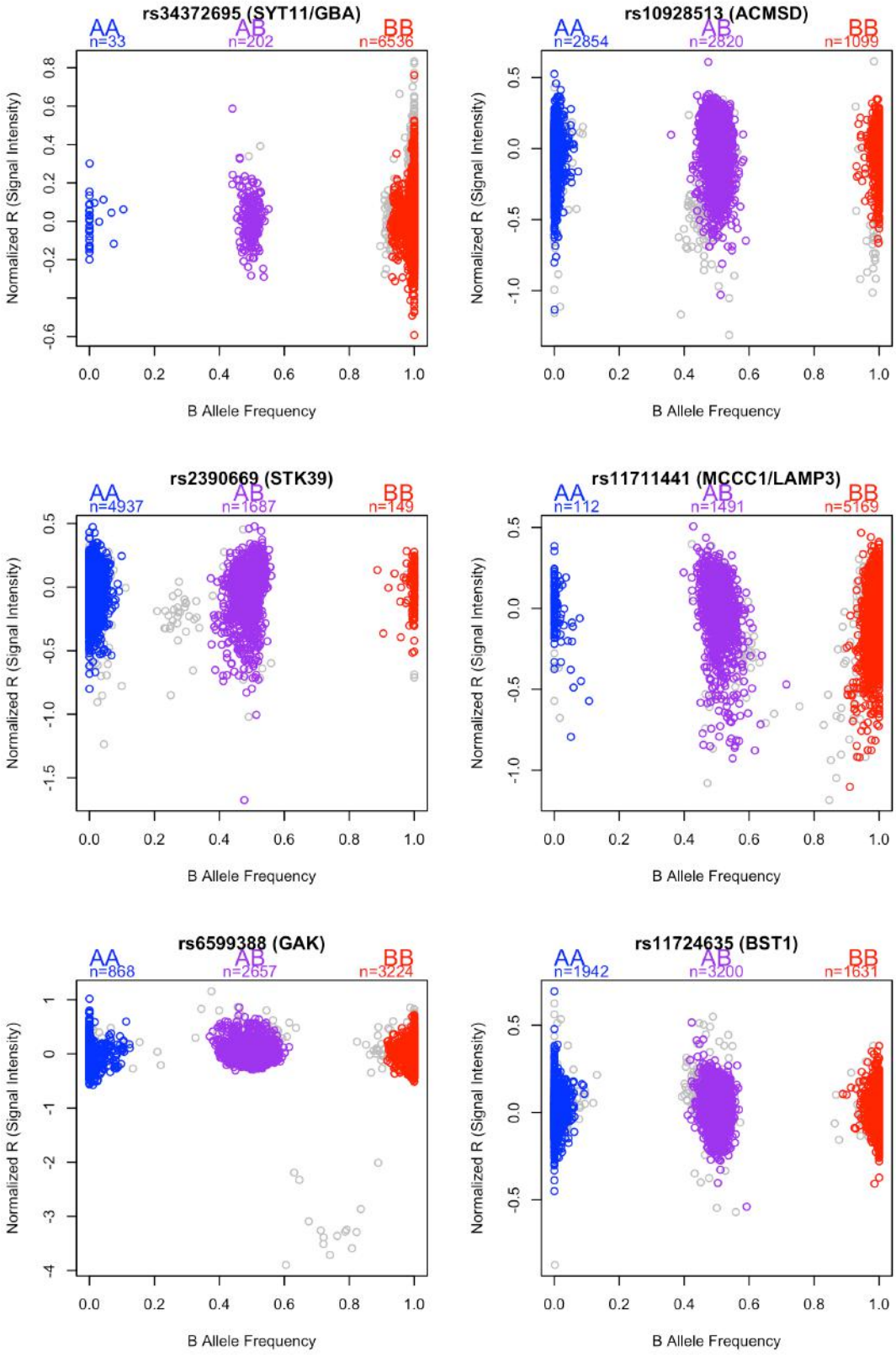


Figure 9: Cluster plots from SNPs significantly associated with PD, typed on ImmunoChip during replication phase

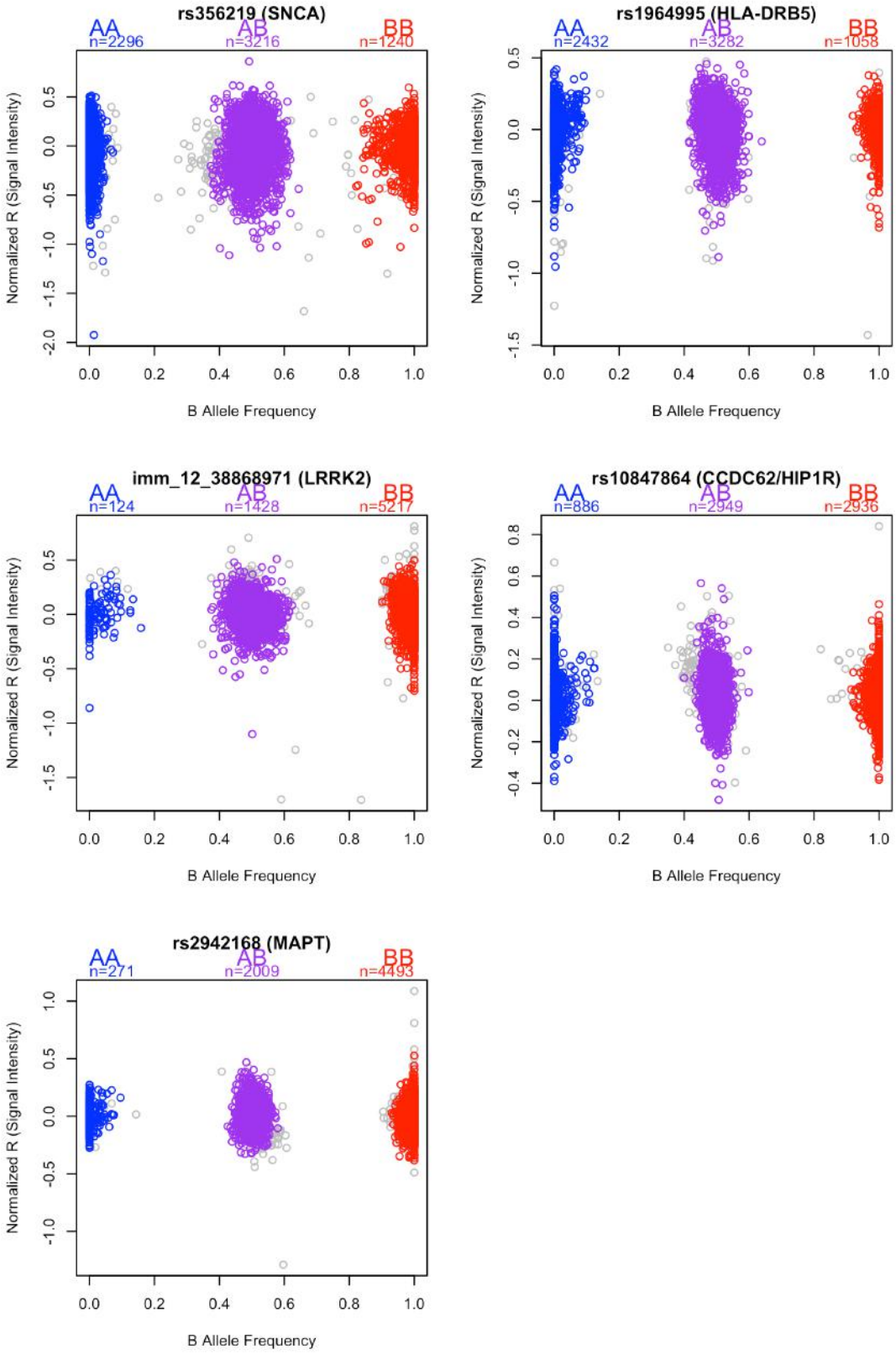


Figure 10: Cluster plots from SNPs significantly associated with PD, typed on ImmunoChip during replication phase

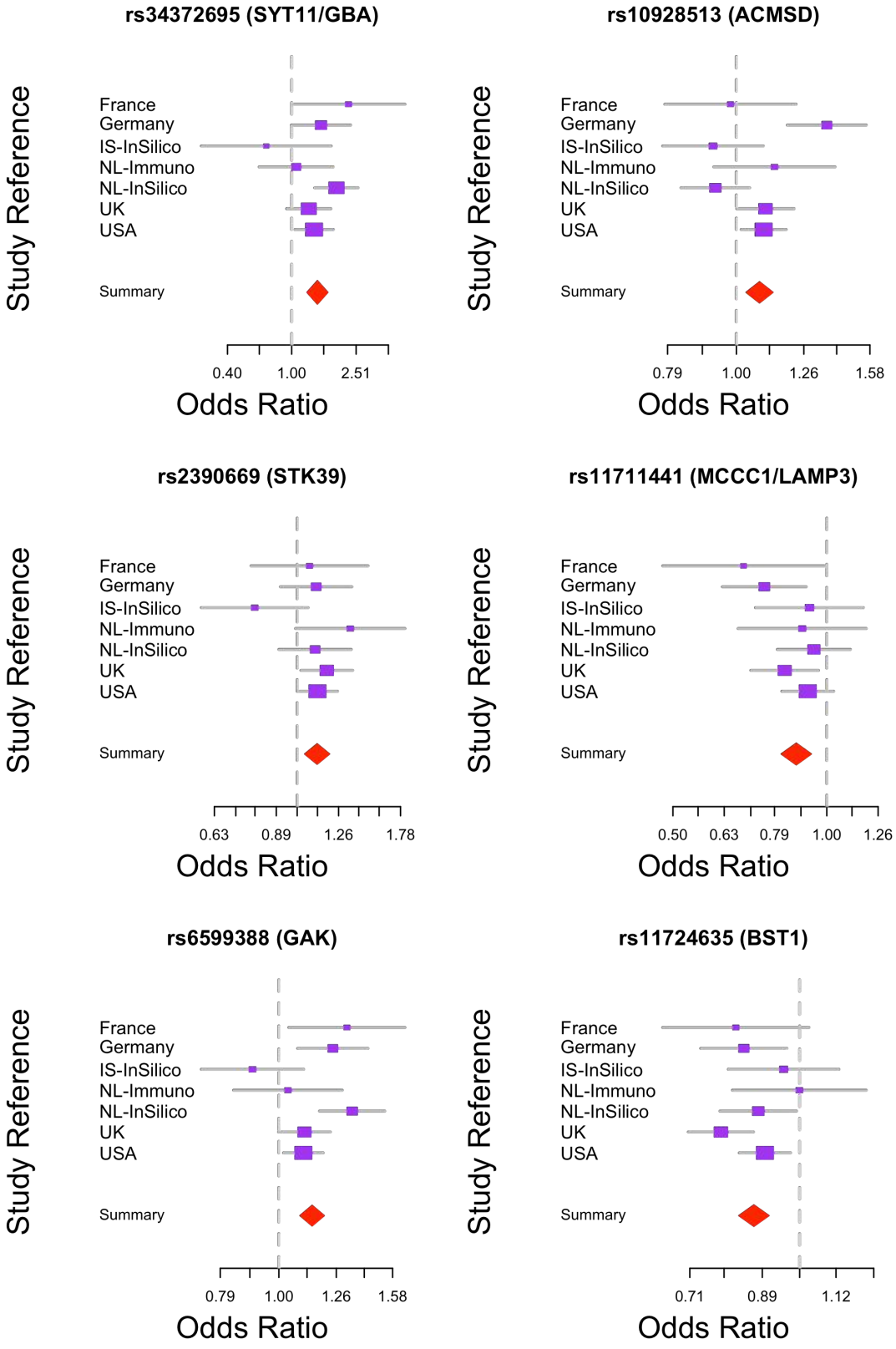


Figure 11: Forest plots of replication phase SNPs.

These show minimal heterogeneity between studies and consistent effect direction with the discovery phase signals.

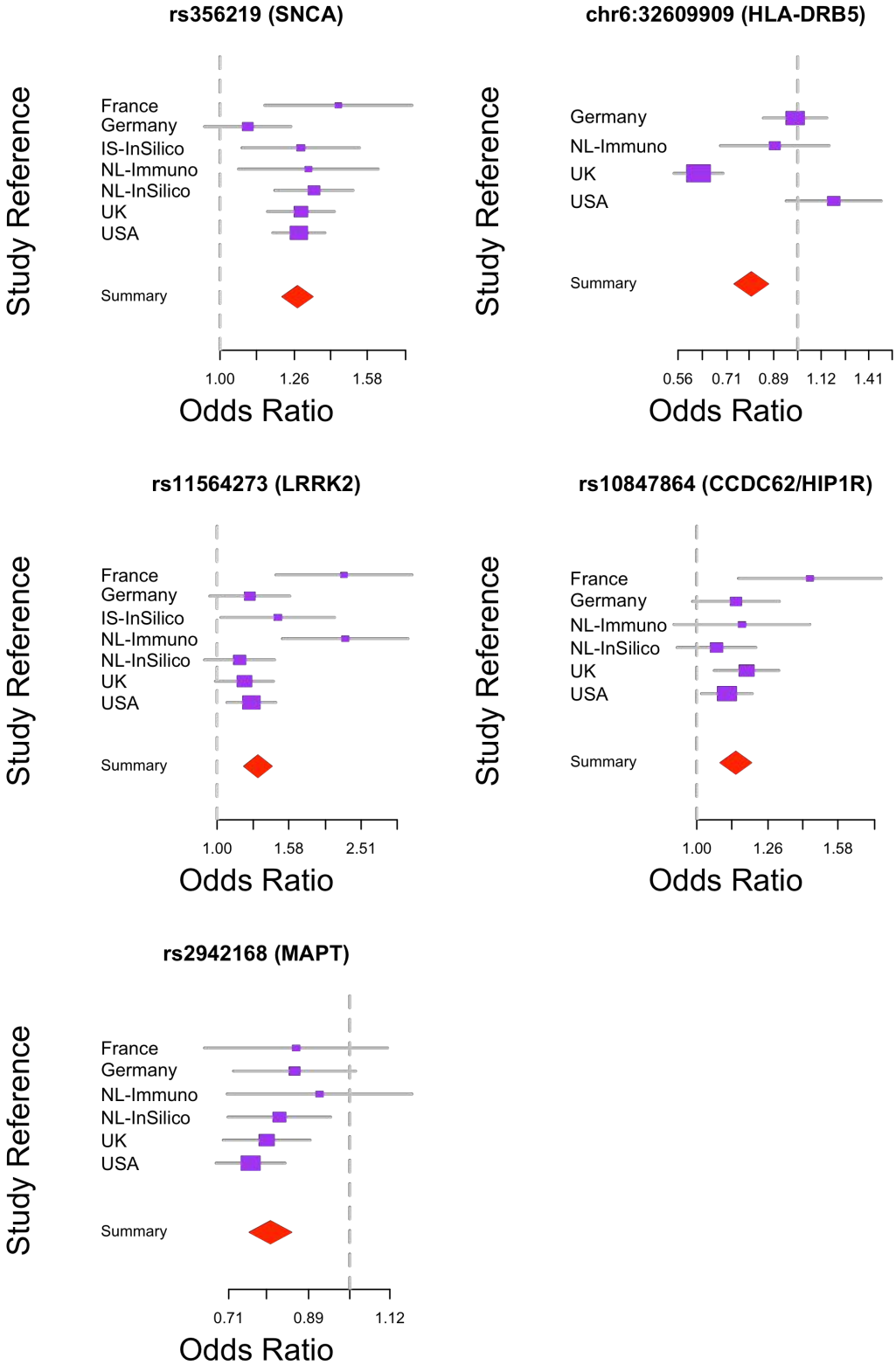


Figure 12: Forest plots of replication phase SNPs.

These plots illustrate consistent effects across each series and demonstrate a direction of effect consistent with that observed within the discovery series.

2.4 Discussion

This chapter describes the completion of a large International meta-analysis and replication of existing GWA data for PD. At the time of publication this was the largest GWA study in PD and confirmed 6 previously identified loci, and identified and subsequently confirmed an additional 5 novel loci. Of the six previously identified loci only *SNCA*, *MAPT*, and *HLA-DRB5* had met strict criteria for genome wide association. These loci were confirmed in this study, both in terms of significance and direction of effect. Notably, while *BST1*, *GAK*, and *LRRK2* were each previously implicated by association, these loci had formerly failed to meet the criteria for genome wide significance. Thus, this study is the first to provide unequivocal evidence for association at these loci [242, 246-249, 258]. This study also linked 5 novel loci conferring risk for PD: *ACMSD*, *STK39*, *HIP1R*, *MCCC1*, and *SYT11*.

Two of the identified loci, at *LRRK2* and *SYT11* are close to known disease causing mutations or moderate risk factors. The most significant risk SNP in *SYT11* lies approximately 500kb from *GBA*, which (as discussed in the introduction) is known to contain mutations that substantially increase risk for PD [259]. The phenomenon of synthetic association has been suggested to exist at some GWA signals [260]. Synthetic association describes a hypothetical scenario where rare disease related variants occur by chance more often on the background of a common allele, which may be quite distant from the rare variant. This creates a potential for misattribution of association to common variants, and mislocalization of the association signal. To investigate this possibility the GWA signals at *SYT11* and *LRRK2* were tested

in the context of known mutation status at *GBA* and *LRRK2* respectively (data not shown). These analyses suggested that the GWA signals were independent of these mutations.

It is important to understand that GWA studies identify loci rather than genes, and that while genes were nominated for each locus based on proximity or function, these genes may not be the functionally relevant effector of association. With this caveat in mind it is useful to briefly discuss the nominated genes at these loci in the context of their potential role in disease.

SYT11 encodes synaptotagmin XI a protein important in the maintenance of synaptic function. *SYT11* has been investigated previously in a negative mutation screening study in 393 familial and sporadic PD patients because of a work showing an interaction between the protein products of *SYT11* and *PARK2* [261, 262]; the latter data suggest a functional role for *SYT11* in the pathogenesis of PD.

ACMSD has been shown to be associated with picolinic and quinolinic acid homeostasis and is described as a possible therapeutic target for a number of conditions affecting the central nervous system [263]. It is worth noting some concern regarding this locus, not only because of the modest replication p value (p=0.0032) but also because the estimate of heterogeneity of effect across cohorts was moderately high (I^2=48.3).

The locus identified near the gene *STK39* has been cited as being associated with autism, and defined as an expression QTL, although no investigations of this locus contributing to neurodegenerative phenotypes have been reported

[264, 265].

LAMP3 within the *MCCC1/LAMP3* locus has been implicated in a study suggesting *LAMP3* is partly responsible for modulation of the neuronal/neurosecretory function in PC12 cell lines [266].

The *HIPR1R* locus is a biologically plausible association, as the protein product of *HIPR1R* is functionally implicated in the intrinsic cell death pathways and believed to interact with huntingtin to modulate polyglutamine-induced neuronal dysfunction in transgenic worm and mouse models [267].

The *HLA-DRB5* is a very interesting locus, as this suggests a role for immunity and inflammation within the pathogenesis of PD [268]. The *HLA* locus has been shown to be associated with numerous neurological diseases, including multiple sclerosis, and Alzheimer's disease [269, 270].

Association of the *SNCA*, *LRRK2*, and *MAPT* loci indicates that genetic variation around these loci is important not only for rare familial forms of neurodegenerative disease, but also for typical sporadic PD. Given the lack of disease associated protein coding polymorphisms at these loci, these data also suggest that expression of these genes is the pathologically relevant effector.

In summary, these data substantially expand our understanding of the genetic basis of PD. The data provide support to the common disease common variant hypothesis, reinforcing the prediction that for common diseases, one form of genetic risk will be mediated by myriad common polymorphisms that

individually confer a modest effect.

3 Abundant Quantitative Trait Loci Exist for DNA Methylation in the Human Brain

STATEMENT OF CONTRIBUTIONS TO THIS RESEARCH:

In this section, I describe a large series of experiments performed to comprehensively assess the relationship between common genetic variability across the human genome with DNA methylation and gene expression in human brain. These experiments encompass several scientific disciplines and are the collective effort of several investigators. I was involved in the inception, planning and design of the experiments and analyses. Additionally, I performed all experimental work and quality control measures for the genome-wide genotyping and genome-wide methylation datasets. I drafted parts of the original manuscript and made substantial contributions to the critical revision of the entire published manuscript, of which I am, one of three equally contributing first authors.

3.1 Introduction

In an assessment of published GWA studies Hindorff and colleagues showed that at most, 30% of GWA loci could be linked to a protein-coding variant [19]. Therefore, the majority, and likely more than 70%, of GWA loci must be explained by changes outside of simple amino-acid sequence changes. If single amino-acid changes are not the primary driver of pathobiologic effects associated with GWA loci, then the primary alternative is that such effects must be driven by expression. Such an effect on expression may be driven at the level of the gene, the transcript, in the context of basal expression, induced expression, or localization of mRNA. While there has been some success in examining the role of risk variants as expression quantitative trait loci (eQTL), there are several limitations that suggest this approach in isolation is not likely to succeed. One limitation is that the expression differences between genotypes are state dependent, and require a response to a stimulus of some kind, and may not be reflected in an expression profile examined in a single tissue at a single time point. It may therefore be useful to also look at the potential for expression, which can feasibly be assessed by examining DNA methylation. Further, many current eQTL studies use array based expression assays in which transcript splicing and exon usage are not well captured. In this context, DNA methylation may also provide an indicator of whether risk SNPs may be mediating a pathobiologically relevant effect through splicing.

In an attempt to efficiently provide high dimensional data that allows the rapid identification of DNA methylation QTLs (dmQTL), a series of experiments were performed to assay the genetic variability and DNA methylation levels in frozen human brain samples. The effects of common genetic variability on DNA methylation in the frontal cortex, temporal cortex, pons and cerebellum were investigated using brain tissue from subjects who were neurologically normal at time of death. All subjects were clinically identified as normal controls at the time of death with no diagnostically confirmed neurological disease. Neuropathological data were not available and therefore not taken into consideration for control subjects.

Frozen tissue for each of these regions from 150 subjects for a total of 600 tissue samples was used. Genome wide genotyping (>500,000 SNPS) and whole-genome methylation profiling (>27,000 CpG methylation sites) was carried out in each of the four regions of the brain. Results of these assays characterizing dmQTL in the human brain are presented in this chapter. Notably eQTL data was also generated in the same tissue samples. This work was not performed as part of this thesis, and the detailed methods and results are not presented, but are discussed briefly to place the current results in context where appropriate.

3.2 Materials and Methods

3.2.1 Samples and Assays

For each of the six hundred samples (150 brains x four regions), approximately 5 grams of frozen tissue was sub-dissected at either the University of Maryland Brain Bank or at the Department of Neuropathology, Johns Hopkins University, and sent on dry ice to the Laboratory of Neurogenetics (LNG), NIA. At LNG, 100-200mg aliquots of frozen tissue were sub-dissected from each sample. Samples were kept on dry ice to avoid thawing. Separate pieces were cut for DNA extraction to be used in SNP genotyping assays and for methylation assays. Each tissue aliquot was stored at -80°C until use.

Genomic DNA extraction for genotyping was performed using the DNeasy Blood and Tissue Kit as per the manufacturer's instructions (Qiagen Inc., Valencia, CA). Genomic DNA for the Infinium methylation assays was extracted using phenol-chloroform and ethanol precipitation. DNA concentration was determined using a Nanodrop ND-1000 spectrophotometer (Thermo Scientific, Wilmington, DE), and DNA extraction was repeated using a new tissue aliquot for samples with DNA concentration less than 50ng/ul, or with a 260/280 ratio less than 1.7. 100-200mg aliquots of frozen tissue were sub-dissected from the frontal cortex, temporal cortex, cerebellum and caudal pons from the brains of each of the 150 neurologically normal Caucasian donors, resulting in 600 samples used for CpG methylation.

3.2.1.1 SNP Genotyping

SNP genotyping was performed using DNA extracted from cerebellar tissue for each subject using Infinium HumanHap550 version 3 BeadChips (Illumina Inc., San Diego, CA) according to the manufacturer's instructions and as described previously (2.2.2.3). Genotype data was analyzed using the Genotyping Analysis Module 3.2.32 within the BeadStudio software version 3.1.4 (Illumina Inc.).

3.2.1.2 DNA Methylation

Genome-wide methylation profiles were generated using Infinium HumanMethylation27 BeadChips (Illumina Inc.), which measure DNA methylation at 27,578 CpG dinucleotides spanning 14,495 genes. A GenePaint automatic slide processor robotic system was used to simultaneously process twenty-four BeadChips. Briefly, 1ug of genomic DNA underwent bisulfite conversion using the WZ-96 DNA Methylation Kit according to the manufacturer's protocol (Zymo Research Corp, Orange, CA). Incubation conditions used during conversion were as follows: (95°C for 30 seconds, 50°C for 1 hour) for 16 cycles, hold overnight at 4°C. Unmethylated cytosines were chemically deaminated to uracil in the presence of bisulfite, while methylated cytosines were refractory to the effects of bisulfite and remained as cytosines. Methylation was then detected as a C/T nucleotide polymorphism at each CpG site. After bisulfite conversion, each sample was whole-genome amplified, fragmented, and hybridized to the BeadChip. DNA molecules anneal to locus-specific DNA oligomers. Two bead types correspond to each CpG locus, one to the methylated (C) state, and the other

to the unmethylated (T) state. After extension, the BeadChip was fluorescently stained, and scanned using a BeadArray laser confocal scanner to measure the intensities of unmethylated and methylated bead types for each locus. Data were analyzed using the Methylation Analysis Module 3.2.0 within BeadStudio. Intensities were not normalized. DNA methylation for each locus was recorded as a beta value, which is a continuous variable between 0 and 1 representing the ratio of the intensity of the methylated bead type to the combined locus intensity (for a schematic of this procedure see Figure 5).

3.2.2 Quality Control

3.2.2.1 Genotype Data

The threshold call rate for inclusion of the sample in analysis was 95%. Two samples initially had a call rate below this threshold, but were successfully re-genotyped using fresh DNA aliquots. Thus all 150 brain samples had a call rate greater than 95%, and were included in the subsequent analyses (average call rate = 99.86%; range 97.72% - 99.95%, based on the missing procedure within the PLINK v1.04 software toolset [271].

The gender of the samples reported to LNG by the brain banks was compared against their genotypic gender using PLINK 's check-sex algorithm, which determines a sample’s genotypic gender based on heterozygosity across the X chromosome. Two samples with gender discrepancies were detected. One of these arose from a clerical error at the brain bank and was included in the

analysis after correction of the clinical information, whereas the other sample (UMARY1496) was removed from subsequent analysis.

To confirm the ethnicity of the samples, Identity-By-State (IBS) clustering and multidimensional scaling analyses were performed within PLINK using the genotypes from the brain samples that had been merged with data from the four HapMap (http://hapmap.ncbi.nlm.nih.gov) populations (n = 32 Caucasian (CEU), 12 Han Chinese, 16 Japanese and 24 Yoruban non-trio samples previously genotyped by Illumina and assayed on the Infinium HumanHap500 version genotyping chips). Two samples were outliers based on population and excluded from further analysis (UMARY4545, UMARY927). Results are shown in Supplemental Figure 1.

Genotype data of the samples were compared for cryptic relatedness using the Identity-By-Descent (IBD) procedure within PLINK. No samples were found to be from related individuals.

Mach software version 1.0.16 [272] and HapMap CEU phase data (release 22) were used to impute genotypes for ~2.5 million SNPs. Imputed SNPs were excluded if the linkage disequilibrium r^2 values between imputed and known genotypes was less than 0.3, and if their posterior probability averages were less than 0.8 for the most likely imputed genotype. For each of the four tissue regions, SNPs were also excluded if: (a) call rate was less than 95%, (b) Hardy-Weinberg equilibrium (HWE) p-value was less than 0.001, and (c) the SNP had less than 3 minor homozygotes present. Exact numbers of SNPs used are shown in Supplemental Table 2 A.

Two samples were outliers based on population ethnicity and excluded from further analysis (Figure 13).

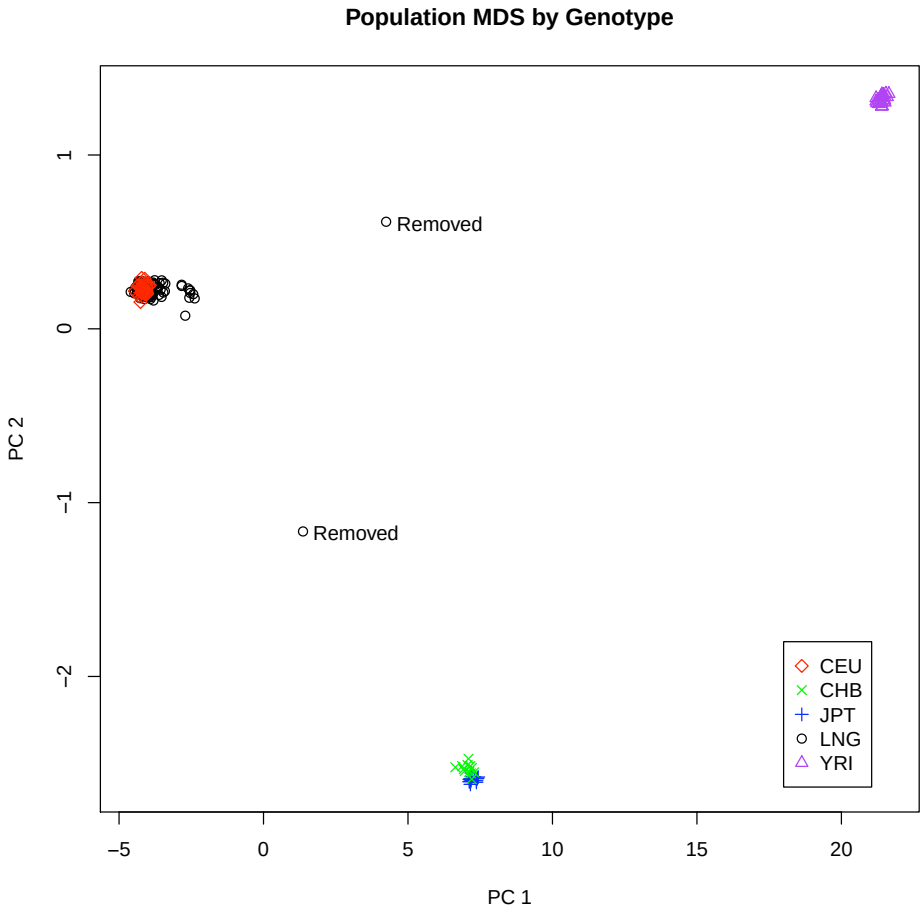


Figure 13: Population MDS plot for brain samples used in dmQTL analyses.

These are based on genotype from genome wide Identity-By-State pairwise distances between the 150 samples used in this study and HapMap samples (CEU, CHB, JPT and YRI). The plot shows that of the 150 samples originating from reported Caucasian individuals from the United States, two samples (indicated by Removed labels) are ethnic outliers relative to the cohort used in this study.

Cryptic relatedness of samples was determined by comparing genotypes using the Identity-By-Descent (IBD) procedure within PLINK. No samples were found to be from related individuals. Mach software version 1.0.16 [273] and HapMap CEU phase data (release 22) were used to impute genotypes for

~2.5 million SNPs. Imputed SNPs were excluded if the linkage disequilibrium r^2 values between imputed and known genotypes was less than 0.3, and if their posterior probability averages were less than 0.8 for the most likely imputed genotype. For each of the four tissue regions, SNPs were also excluded if: (a) call rate was less than 95%, (b) Hardy-Weinberg equilibrium (HWE) p-value was less than 0.001, and (c) the SNP had less than 3 minor homozygotes present. The final numbers of SNPs used per brain tissue for CpG methylation assay are shown below (Table 4) and a Venn diagram is shown below illustrating which detected probes overlapped per region.

Table 4: Summary count of the number of samples, DNA methylation sites, and SNPs tested per tissue

	Region			
	CRBLM	**FCTX**	**PONS**	**TCTX**
Samples	108	133	125	127
Probes	27,310	27,532	27,476	27,538
SNPs	1,540,472	1,624,830	1,602,245	1,607,740

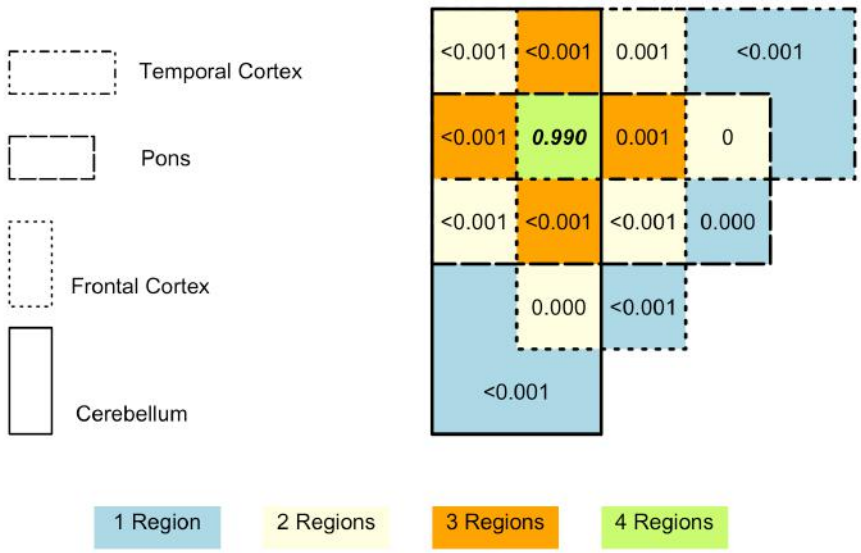


Figure 14: Venn diagram of probes detected across regions.

This represents the number of probes detected in 95% of samples between the four brain tissues. The rectangles with different orientations and border, shown on the left legend represent the different tissue and the different squares represent overlapping frequencies between different tissues. The green square iillustrates that 99.0% of the 27,551 CpG probes detected in at least one tissue region were also detected in all four tissues regions.

3.2.2.2 CpG Methylation Data

The threshold call rate for inclusion of samples in the analysis was 95%. Based on this metric, 9 cerebellar samples (UMARY933, UMARY1465, UMARY4593, UMARY4640, UMARY4726, UMARY4727, UMARY4842, JHU1361, and BLSA1640), 6 pons samples (UMARY384, UMARY1541, UMARY1583, UMARY1613, UMARY1907 and UMARY4903), 2 frontal cortex samples (UMARY1712, BLSA1840) and 8 temporal cortex samples (UMARY1865, UMARY1866, UMARY1867, UMARY4545, UMARY4598, UMARY4726, UMARY4727 and UMARY5087) were excluded from analysis.

The remaining brain samples had an average detection rate of 99.84% (range 95.0% to 99.98%).

The gender of the samples reported to LNG by the brain banks were compared against their methylation gender based on beta values of methylation probes across the X chromosome [185]. Four samples with gender discrepancies were detected and were removed from subsequent analysis (BLSA2102-crblm, UMARY1862-crblm, UMARY880-pons, UMARY4540-tctx). The resulting HCL sample tree is based on Chromosome X sites, after removal of these four individuals (Figure 13).

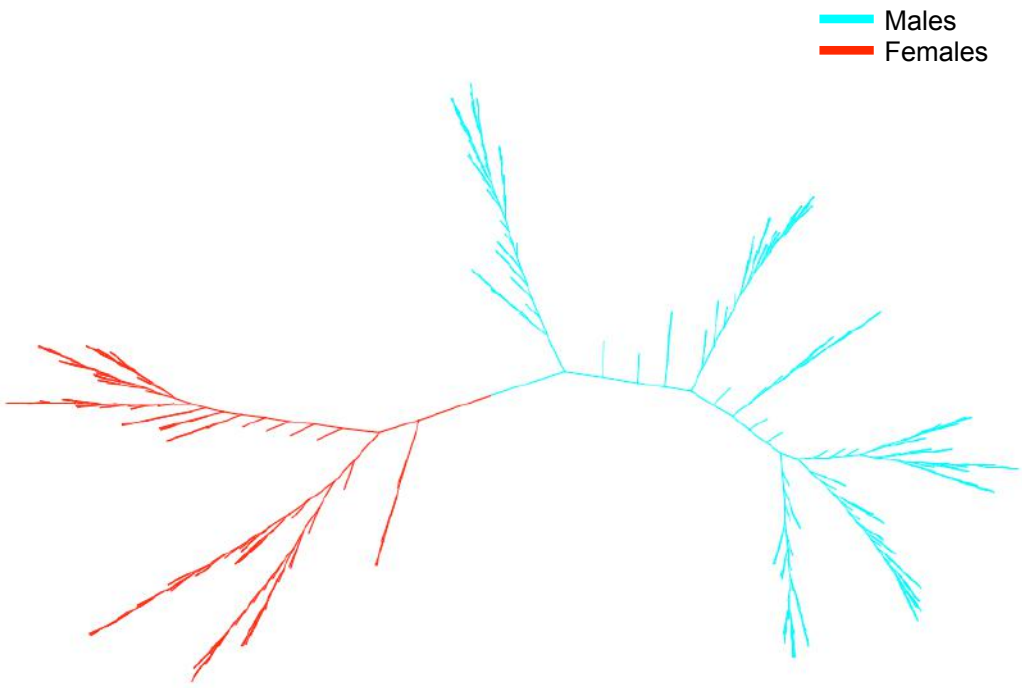


Figure 15: Unsupervised hierarchical cluster plot using DNA methylation data from 4 brain regions.

The plot was generated in HyperTree using the samples tree generated from an HCL of Chromosome X methylation data using 'Average Linkage clustering'. Male and female status was not used as a variable within the cluster analysis, and the colors represent post-hoc labeling based on brain back reported gender. The plot is after removal of the individuals that appeared to be gender mismatches.

3.2.3 Clustering of Samples by Brain Region

HCL [274] of methylation profiles was performed using the TM4 MeV version 4.1.01 tool [275] with Euclidian distances and Average Linkage clustering. Detected data for autosomal probes only was used in clustering the CpG data, otherwise sub-clusters based on gender appeared. The HCL sample tree was saved as a Newick tree file and plotted again using the HyperTree tool (http://hypertree.sourceforge.net/).

3.2.4 Correction for Known Biological and Methodological Covariates

In an effort to remove the influence of potentially confounding variables, each trait was adjusted for the following covariates: age, gender, post-mortem interval (PMI), brain bank and sample hybridization batch. This adjustment was performed prior to QTL analysis. In R [276] each trait was regressed with Y representing the trait profile (log_2 normalized mRNA expression intensities or raw values of CpG DNA methylation) and $X_1 \ldots X_n$ representing the biological and methodological covariates. Within this model gender, tissue bank and batch were treated as categorical covariates. After fitting each trait to the model, the residuals were kept and represent the trait in the present study. Thus variance attributable to gender, age, post-mortem interval, tissue source and hybridization batch were controlled for prior to QTL analysis.

3.2.5 Quantitative Trait Locus Analysis

For each of the four brain regions, a regression analysis was performed on the residuals described in the preceding section for the CpG methylation levels. The trait residuals were then used as the quantitative phenotype for each probe in a genome-wide association analysis investigating quantitative trait loci. These analyses were performed using the *assoc* function within Plink [271], correlating allele dosage with change in the trait. Each of the four brain regions was analyzed separately.

3.2.6 Correction for Multiple Tests

In an attempt to correct for the number of variants tested per trait, genome-wide empirical p-values were computed for the asymptotic p-value for each SNP; this was performed using maxT permutation [271], by means of 1,000 permutations of swapping sample labels of the traits. This method is not dependent on these quantitative traits having a normal distribution and also allows the linkage disequilibrium of the genomic regions being tested against the traits to be maintained [277].

In an attempt to correct for the number of traits (i.e. CpG sites) being tested in each tissue region, a false discovery rate (FDR) threshold was determined based on empirical p-values using the *fwer2fdr* function of the multtest package in R [276]. Empirical p-values were allowed to exceed this threshold if their linkage disequilibrium r^2 was greater than or equal to 0.7 with a SNP with empirical values within the FDR threshold.

Following regression analyses, significant QTLs were designated as *cis* if the SNP was within 1MB of the CpG methylation site being tested. All other SNP –dependent variables were designated as *trans.*

3.2.7 Polymorphism(s) in Assay Probes

Sequence variants within probes used to assay DNA methylation levels may cause differential hybridization and inaccurate expression and methylation measurements. To exclude this possible confounding variable, the sequences of probes with significant correlation to a trait were examined for the presence of polymorphisms using CEU HapMap data, and, if present, that QTL was removed from the result set.

3.3 Results

3.3.1 CpG Methylation Levels Prove to be Different Between Brain Regions.

In order to determine whether CpG methylation levels differed among the four brain regions, a global comparison of CpG methylation was performed across the four different tissues. An unsupervised cluster analysis [274] revealed that the four regions have unique DNA methylation profiles. These data revealed a distinction in DNA methylation patterns between the pons and cerebellum with the two cortices overlapping considerably. Previous work showed a distinct pattern of DNA methylation in the human cerebellum compared with cortical tissues [278, 279].

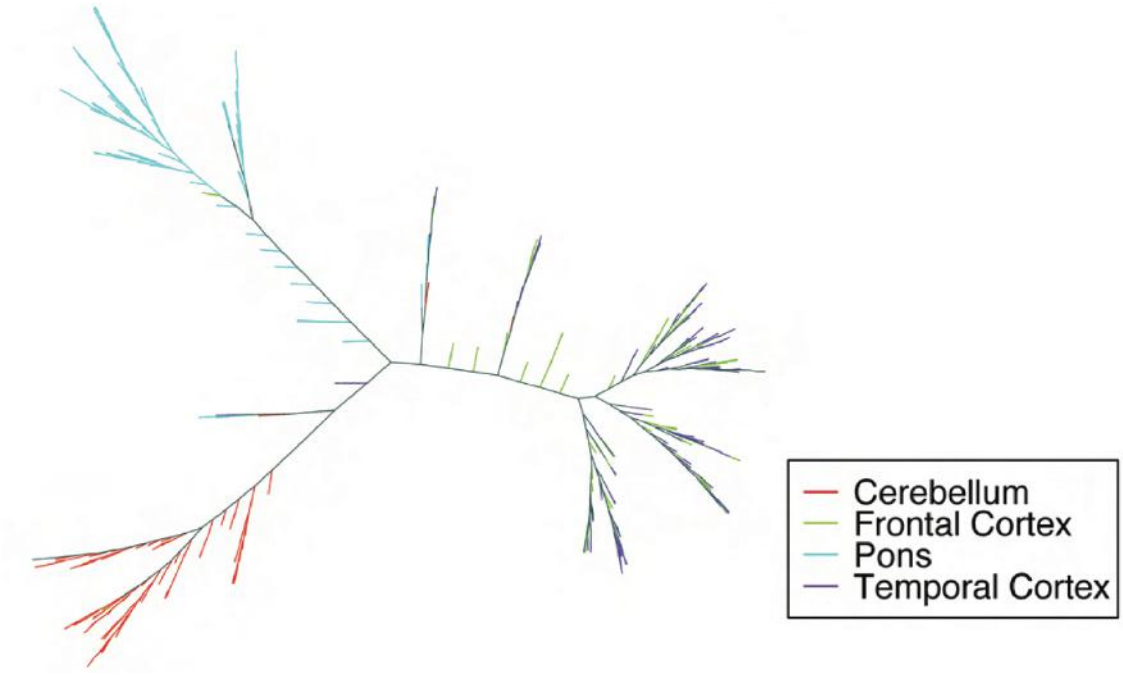


Figure 16: Unsupervised cluster analysis of CpG methylation levels at autosomal loci.

These data show separation of the cerebellum, pons and cortices, but a general inability of the clustering to separate data derived from frontal cortex and temporal cortex.

Next, an analysis was performed in a dataset including only probes that showed an Illumina detection p-value of less than or equal to 0.01 in 95% of samples. This subset of probes was analyzed in each of the four brain tissues and included 27,551 unique CpG methylation sites (Figure 17).

The distribution of the observed CpG methylation levels was plotted as a histogram for each brain region (Figure 17). This analysis revealed a large number of CpG sites infrequently methylated and a smaller number of highly methylated sites, across all tissues, exhibiting a similar pattern of DNA methylation to that reported earlier by Zhang and colleagues (2009) and Meissner and colleagues (2008) [280, 281]. CpG sites within islands were

primarily unmethylated [178, 282]. Next, a direct comparison of CpG methylation at individual loci was made between each possible pair of brain regions. On the whole, DNA methylation levels were similar between tissues. The levels within the frontal and temporal cortices were consistently the most alike with the cerebellar region showing the most divergent profile.

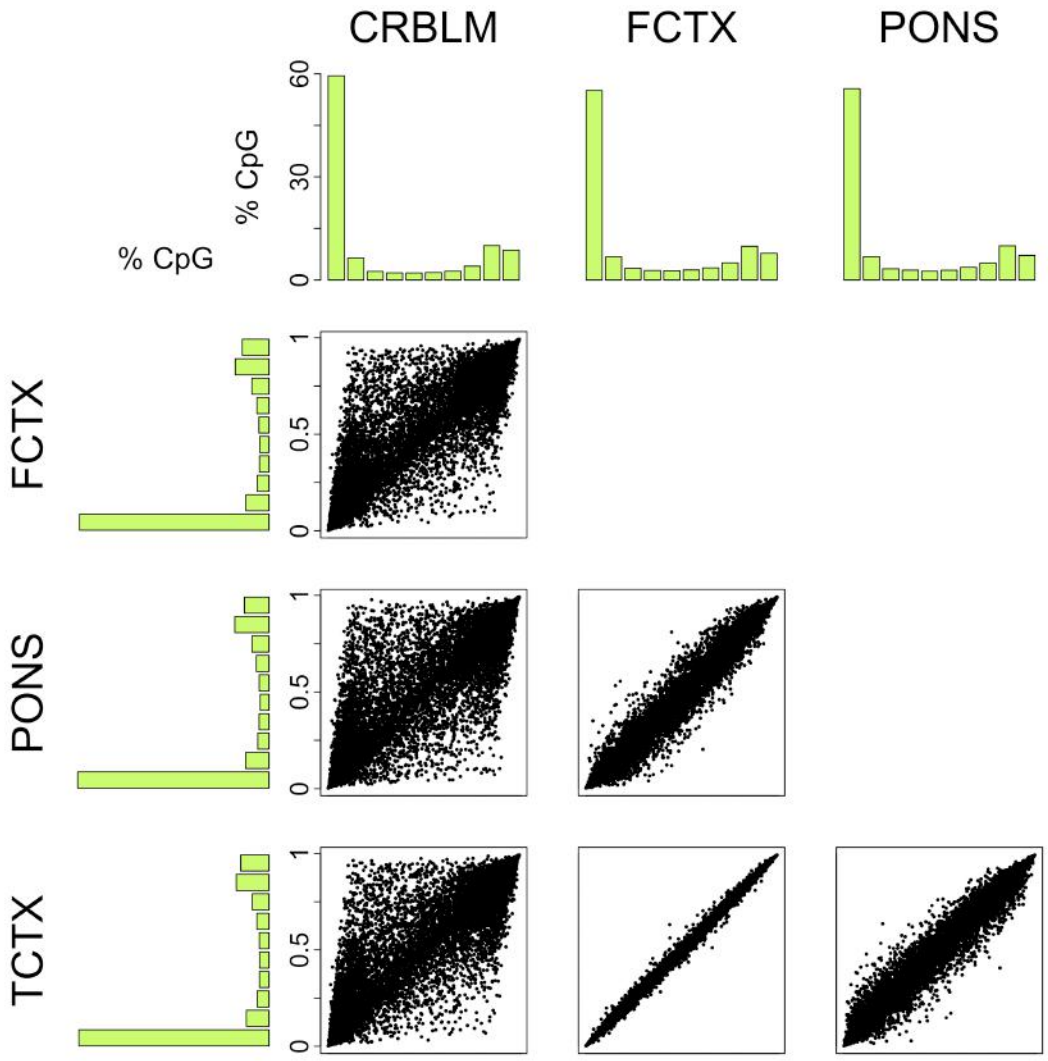


Figure 17: Comparison of DNA methylation levels between tissues.

These data show only data from those probes with an Illumina detection p-value of less than or equal to 0.01 in 95% of the samples assayed. This illustrates that the DNA methylation levels are most similar across probes for the temporal and frontal cortices, with the signal being most dissimilar from other tissues in the cerebellum.

3.3.2 Genetic Control of DNA Methylation

A principal objective of these experiments was to examine the influence of genetic variability on DNA methylation and expression within brain tissues. Therefore, a series of QTL analyses was carried out. For each sample genotyped, 537,411 SNPs passed quality control. Those genotypes were then imputed to 2,545,178 SNPs using MACH and HapMap CEU phase data. After additional quality control measures were applied, an average of 1,629,853 SNPs were used in the QTL analysis. Each brain region was considered separately for the QTL analysis. This allowed for the inclusion of CpG probes that were detected in 95% of the samples for individual brain regions versus 95% of samples for all four regions. The number of CpG sites tested for each brain region is listed above (Table 4).

It has been shown that SNPs in close proximity to genes can have a greater influence on gene expression [283, 284]. In order to find out if this could also be applied to QTLs linked to DNA methylation and to examine if this could be generalized across tissues, all significant QTLs were plotted by genomic position of the SNP versus that of the CpG site (Figure 18 and Figure 19).

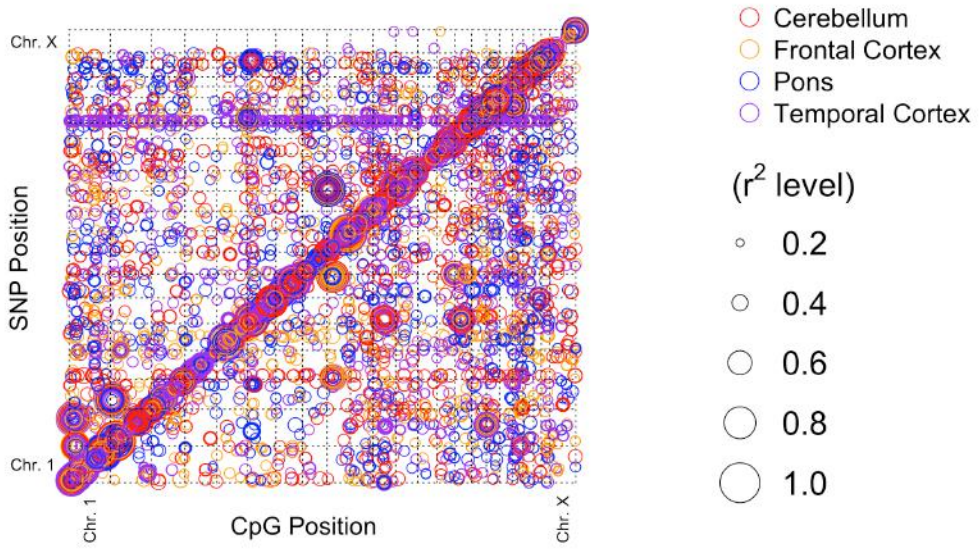


Figure 18: Positions of dmQTL based on SNP and associated CpG site.

This figure illustrates that for all regions the vast majority of SNP-CpG dmQTL pairs are physically close.

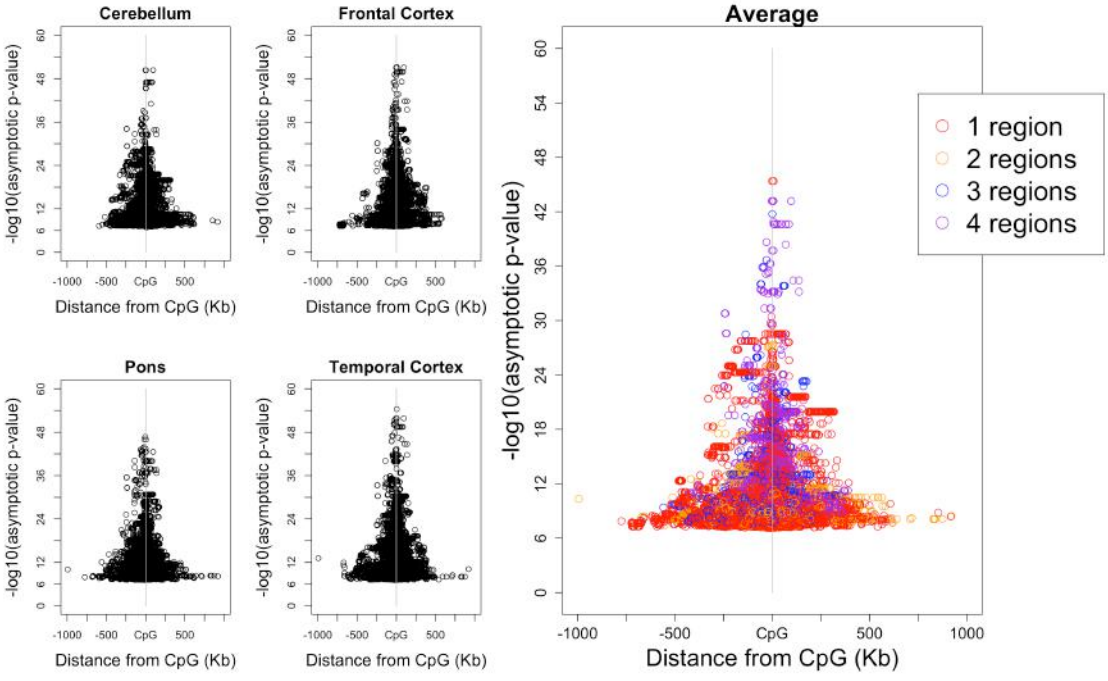


Figure 19. dmQTL within and across regions.

***Cis* mdQTL demonstrate a symmetric association between methylation level and variants on both sides of the CpG site in all four brain regions. Notably, the figure displaying a summary of dmQTLs seen in 1, 2, 3, or 4 regions shows that the most significant QTLs tend to be present in all regions.**

Previous work by Bock and colleagues [282] showed that SNPs were only faintly correlated with the methylation status of CpG islands when compared with the much stronger influence of DNA sequence or structure. However, a large number of significant correlations between genetic variation and the methylation status of individual CpG loci were found in the present study (Table 5). There is a robust positional effect seen genome-wide with an excess of the number and magnitude of *cis* associations. Interestingly, detection of *cis* QTLs for DNA methylation occurred more when the CpG site was located outside of an island [178].

Table 5: . Summary counts of significant methQTL results found in each brain tissue.

	Cerebellum	**Frontal Cortex**	**Pons**	**Temporal Cortex**
***cis* dmQTL**	9,117	9,242	7,966	12,081
***trans* dmQTL**	2,985	2,893	3,408	4,653
Total dmQTL	12,102	12,135	11,374	16,374

Exploring the distribution of the numerous *cis* dmQTL revealed that the number of significant dmQTL and strength of association between the SNP and DNA methylation level were inversely correlated with the physical distance between the genetic and epigenetic variants in question. The number of significant QTL and the strength of association for those loci increased the closer the SNP was to the CpG site. The average distance between correlated *cis* SNP and the correlated CpG site was 81kB.

In order to determine whether there was an enrichment of *cis* QTLs relative to those in *trans,* the number of observed and possible *cis* and *trans* QTL for each of CpG methylation and mRNA expression levels was calculated. The outcome showed a tremendous enrichment of *cis* dmQTLs (4,400-fold) compared to *trans* dmQTLs. The peak level of *cis* enrichment was observed when the threshold distance was dramatically decreased from 1MB to ~45bp (Figure 6).

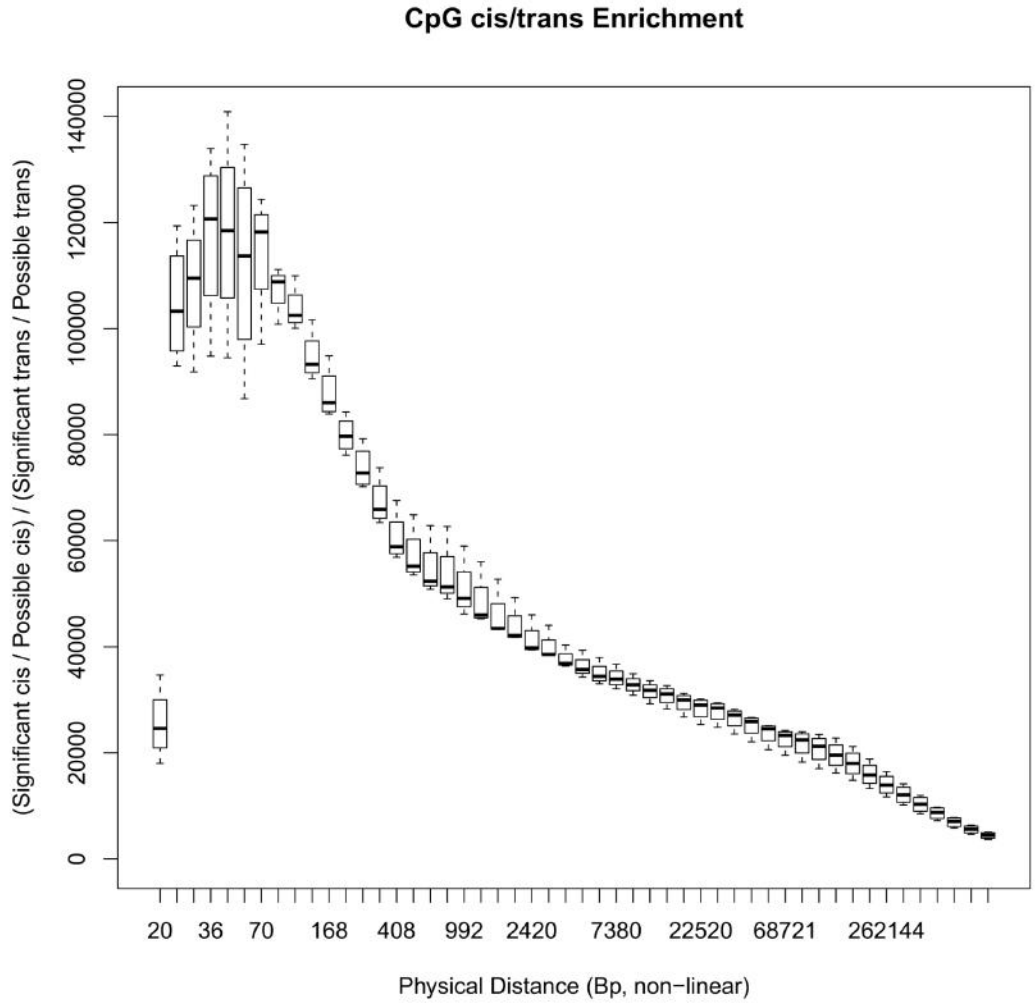


Figure 20: Enrichment of observed *cis* dmQTL relative to observed *trans* dmQTL.

This plot displays the proportions of significant *cis* and *trans* SNP/Probe pairings from the possible pairings at different distances. At 1 Mb the enrichment of cis to trans for CpG QTLs is 4,427-fold. This plot shows how this enrichment changes when physical distance changes as a threshold for cis, where the x-axis represents the physical distance, resulting in a non-linear relationship at 50 distances.

3.3.3 Highly Significant dmQTL Are Consistent Across Brain Regions

Figure 19 above shows that the most significant dmQTLs tend to occur across all regions. Taking statistically significant results into account, 49% of CpG sites with a significant *cis* QTL are detected in only one tissue; thus, appearing to be tissue specific. However, these results can be misleading due to the reliance on using a threshold for significance. It is possible that dmQTL that did not reach this threshold were excluded.

In order to compare detected dmQTLs between tissues in a more complete way, every SNP-CpG methylation pair that passed the defined threshold for significance in at least one tissue was selected. Using ternary plots, the R^2 values were then compared for each SNP-CpG pair in all four tissues, including those pairs that were considered non-significant (Figure 21). The majority of large effect and many moderate effect dmQTLs were shared across the four brain regions when significant effects from a tissue were compared with corresponding (potentially non-significant) effects from another tissue. A subset of *trans* dmQTLs was also shared between regions and had high R^2 values. This shows, even as there is enrichment for *cis* dmQTLs in our dataset, there are also strong effects where SNP and CpG methylation show significant association but are physically distant from one another (Figure 21 E-F).

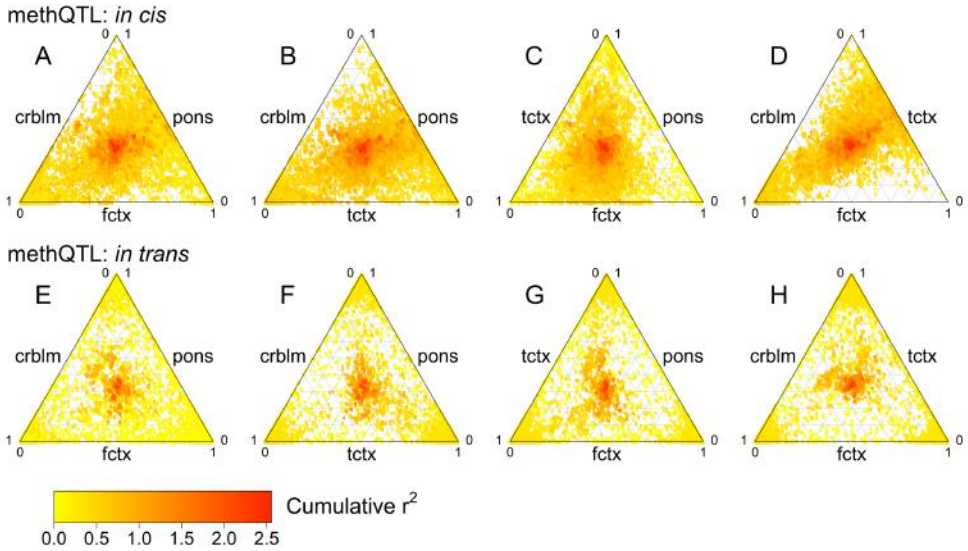


Figure 21: Ternary plots of dmQTLs across regions.

dmQTL that passed the threshold for significance in at least one tissue were included in the ternary plots. The color of the points in the ternary plots reflects the cumulative R^2 value for all tissues tested within each plot. Points toward the center indicate an equal R^2 value across the three regions under investigation. Points toward the corner of a plot indicate a high R^2 in one of the three tissues; points toward the edges of the plot indicate a high R^2 in two of the three tissues

3.4 Discussion

In an attempt to understand and map DNA methylation and the relationship of DNA methylation to genetic variability, epigenetic variation was investigated from a whole-genome perspective in human brain tissue. This work integrated common genetic variability typed and imputed from whole genome genotyping with genome-wide methylation levels. Four brain regions; the cerebellum, frontal cortex, temporal cortex, and caudal pons were studied in tissues from 150 neurologically normal individuals.

Findings show that DNA methylation patterns differ between brain regions, sufficiently so that it is possible to separate cortical tissue from that derived from the pons or cerebellum based on DNA methylation profile alone. These data show that there are region specific differences in DNA methylation and suggest that these differences may require investigators to use a tissue directly relevant to their disease or trait of interest. However, these data also show that highly significant and high effect dmQTLs occur across multiple brain tissues. Thus, for many of the large effect dmQTLs it may be sufficient to examine only one brain tissue.

A large number of dmQTLs were identified in the current study and there is a clear enrichment for *cis* dmQTLs when compared to *trans* dmQTLs. Notably, the SNP and CpG methylation site tend to be close physically, as indicated by the enrichment of highly significant SNPs close to the correlated CpG site (shown in Figure 19) and the peak enrichment of observed to expected dmQTLs at only ~50bp from the CpG site in question (Figure 20).

Although not shown here, expression QTL analysis was performed in the same tissues as the current study. A comparison was performed to test whether transcripts with proximal dmQTLs were more likely to have eQTLs, and whether the same variant was mediating both effects. Strikingly, this was not the case. There is relatively little overlap between the identified eQTLs and dmQTLs possibly reflecting some of the limitations of current eQTL analyses, in that most studies examine transcription at a single time, rather than transcriptional potential. It is also conceivable that dmQTL may exert an effect through altering transcript ratios, UTR usage, and splicing, effects that may not be seen using expression arrays. Regardless, these data suggest that dmQTL analysis may serve as a valuable tool in understanding the consequences of genetic variation.

There are several limitations to the current study; notably the sample size, while large for dmQTL analysis, is not of sufficient size to reliably detect modest effects. Also, while the DNA methylation assays used here are reliable, it is important to assay additional sites. Perhaps an even more significant limitation is the use of brain tissue, which represents a highly heterogeneous mix of cells. This would suggest that the dmQTLs observed here are more likely to be of large effect and/or generalizable across cell types, and that cell type specific dmQTLs may have been missed.

A major reason for performing these experiments was to allow the creation of a dataset that could be used to ‘lookup’ the effects of genetic variability on DNA methylation, particularly for studies interested in understanding the effects of risk alleles. In order to do this, data have been made public through

dbGAP (www.ncbi.nlm.nih.gov/gap/) and GEO (http://www.ncbi.nlm.nih.gov/geo/). These data, and the companion gene expression data, have been downloaded and used by numerous investigators, primarily to help understand functional consequences of disease linked risk variants for disorders such as Tourette's syndrome, coffee consumption, obsessive compulsive disorder, body mass index, and PD [285-287].

4 Assessment of Parkinson's Disease Genetic Risk Loci as DNA Methylation QTLs

STATEMENT OF CONTRIBUTIONS TO THIS RESEARCH:

In this section, I describe a series of experiments that were performed to identify Parkinson's Disease risk variants as dmQTLs. These experiments encompass several scientific disciplines and are the collective effort of several investigators. I was involved in the inception, planning and design of the experiments and analyses. I performed experimental work and quality control for genome-wide SNP datasets for the PD meta-analysis and NeuroX replication stage genotyping. I performed the DNA methylation analysis, and genotyping in human brain tissue as described in Section 3 above and drafted a manuscript (in submission).

4.1 Introduction:

Achieving a complete understanding of the pathobiological mechanisms underlying complex neurological disease has been the emphasis of genetic studies for well over a decade. As described previously, GWA studies have had considerable success identifying Mendelian genes and now numerous genetic risk factors for Parkinson's disease [286]. However, the next step following the discovery of genetic risk-associated markers is the challenging interpretation of such risk within the context of disease pathobiology. A primary aim of meta-analysis of PD GWA studies (described in Chapter 2) was not only to identify novel risk variants for PD, but to also to assess the biological consequences of those variants.

One approach is to integrate genotyping and a quantitative trait, such as DNA methylation levels, in a manner in which biological meaning can be derived from the association between the two datasets. A direct relationship between the underlying genetic sequence and CpG methylation levels was described and defined as DNA methylation quantitative trait loci (dmQTL) in Chapter 3 of this thesis. Additional studies have also reported the close interplay between allele load and DNA methylation levels [288-290] including descriptions of natural human variation [290-292], neurological disorders [288] and rheumatoid arthritis [293] emphasizing the significance of DNA methylation patterns for diverse phenotypes, including those related to diseases.

DNA methylation is cell and tissue specific [294]; therefore, it is critical to analyze methylation patterns in the tissue of interest. For degenerative

diseases such as PD, in which cell death is a key element of the disease pathology, it is also important to assess DNA methylation levels (or other quantitative trait, such as gene expression) in tissue derived from healthy individuals to avoid detection of alterations caused by cell loss. Here, following a large-scale meta-analysis of GWA data [286], we evaluate the association of previously confirmed and novel PD risk loci with CpG methylation levels in the frontal cortex region of 309 normal human brain samples. We examine CpG methylation levels within 1Mb of 28 reported PD risk variants [286] as potential methylation QTLs. Results of these assays characterizing dmQTLs in the human brain within the context of PD risk are presented in this chapter.

4.2 Materials and Methods:

4.2.1 Samples

Aliquots of frozen tissue were sub-dissected from the frontal cortex, from the brains of 432 neurologically normal Caucasian donors from the US and UK. Genomic DNA was extracted from 100-200mg of tissue using phenol-chloroform. DNA concentration was measured using the Qubit 2.0 Fluorometer.

4.2.2 SNP Genotyping

Genotyping for the 28 SNPs with known genome-wide significance in PD GWA studies was performed using the Illumina NeuroX genotyping array (Illumina Inc, San Diego) [295]. The NeuroX array is an Illumina Infinium iSelect HD Custom Genotyping array containing 267,607 Illumina standard content exonic variants and an additional 24,706 custom variants designed for neurological disease studies [295]. Of the custom variants, approximately 9,000 are designed to study Parkinson's disease and are applicable to both large population studies of risk factors and to investigations of familial disease and known mutations. The custom SNPs include, tagging SNPs, proxies and technical replicates for the PD associated loci identified in the discovery phase of the PD GWA meta-analysis recently published by our laboratory [286].

Genotyping on Illumina NeuroX array was performed per manufacturers protocol (Illumina, Inc. San Diego) and as described in Chapter 2. The Genotyping Analysis Module within Genome Studio version 1.9.4 was used to analyze data. The threshold call rate for sample inclusion was 97%. All 28 SNPs analyzed were manually clustered and visually inspected. Genotypes for standard exome content variants were called using a cluster file from the CHARGE consortium based on more than 60,000 samples [296].

4.2.3 CpG methylation

Genomic DNA was bisulfite converted using Zymo EZ-96 DNA Methylation Kit (Zymo Research Corp., Irvine, CA) per manufacturer's protocol. CpG methylation status of 485,577 CpG sites was determined using the Illumina Infinium HumanMethylation450 BeadChip (Illumina Inc., San Diego, CA) per the manufacturer's protocol, outlined in Section 3.2.1.2.

The HumanMethylation450K BeadChip uses both Infinium I and Infinium II chemistries. Infinium I chemistry uses two probes per locus: one unmethylated and one methylated as described in Chapter 3. Infinium II chemistry is designed to use only one bead type with the methylated state defined at the single base pair extension following hybridization. The 3' end of the 50bp probe complements the base directly upstream of the CpG site; subsequently, the single base pair extension results in an addition of a labeled (A) or (G) nucleotide, complementary to either the methylated (C) or the unmethylated (T) state (Figure 5 page 60).

The HumanMethylation450 BeadChip queries 99% of NCBI Reference Sequence Database genes (http://www.ncbi.nlm.nih.gov/refseq/) with sites in the promoter region, 5'UTR, first exon, gene body and 3'UTR. 96% of CpG islands within the human genome are queried with multiple probes within islands, shores and shelves.

Initial data analysis was performed using the Methylation Analysis Module within Genome Studio 2011.1 (Model M Version 1.9.0 Illumina, Inc., San Diego, CA). CpG probes within 1Mb either side of the 28 SNPs reported as

most significant in GWA studies for PD were analyzed for QTL analysis. This included 20,616 CpG probes.

Threshold call rate for inclusion of samples was 95%. Quality control of sample handling included comparison of brain bank reported sex versus sex of the same samples determined by analysis of methylation levels of CpG sites on the X chromosome. Beta values were extracted for sites on the X chromosome. Subject mean methylation versus subject mean intensity levels were plotted in R V2.11.1. Based on methylation levels for chromosome X loci, these data split into two primary groups. Calls generated by this method were then compared with sample information reported by brain bank(s). Samples not matching between clinical reported sex and methylation data were excluded from analyses. Eight samples with gender errors and 5 samples considered outliers were removed from the analysis (Figure 22).

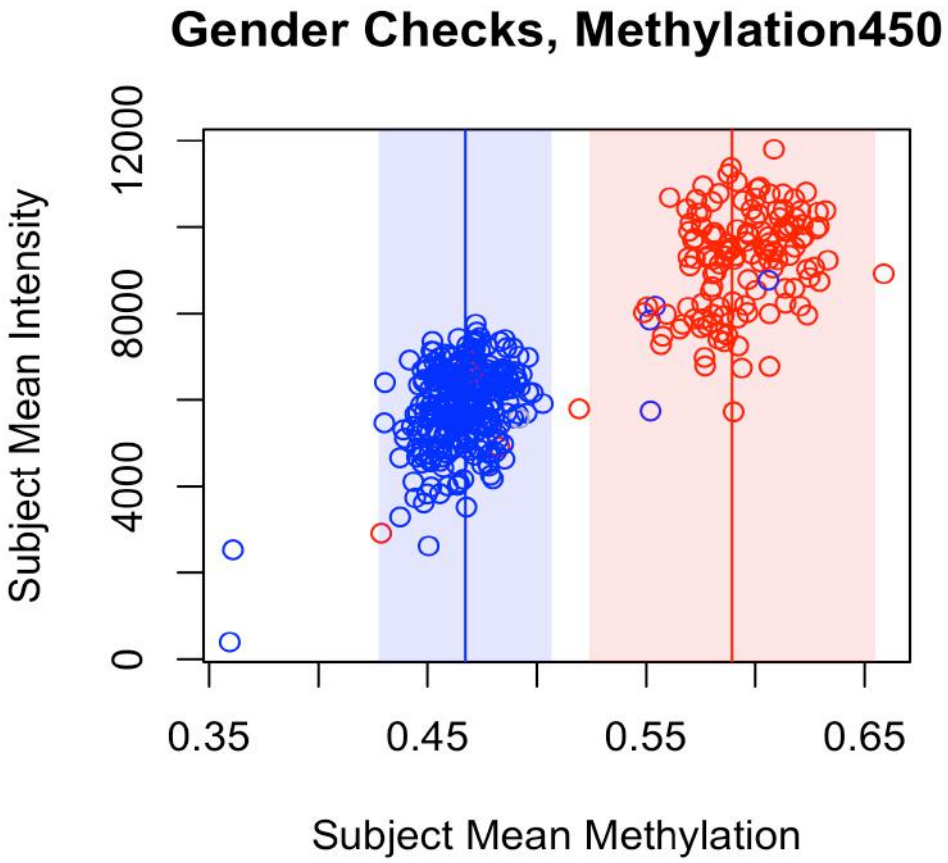


Figure 22:Sample handling quality control for CpG methylation samples.

Beta values were extracted for sites on the X chromosome. Subject mean methylation versus subject mean intensity levels are plotted for 11,232 CpG probes on Chr X. Males are depicted in Blue. Females are depicted in Red. Plot shows 8 gender discrepancies and 5 outliers.

4.2.4 *Cis* Quantitative Trait Locus Analysis

Genome-wide methylation results were filtered to include CpG probes within 1Mb of the most significant SNP in each PD associated region. The filtered results included methylation levels for 20,615 CpG probes. Any probes with >5% missingness were excluded from the analysis leaving a total 17,620 probes tested. The inverse variance normalization method was applied to stabilize the data into a normal distribution using R software 3.0.2. Linear

regression of allele dosage was performed using CpG methylation level as the dependent variable and genotype as the independent variable. Linear regression analysis was carried out using R software 3.0.2 to associate allele dosage with quantitative trait. For these analyses, significance was determined based on standard FDR adjustments for multiple testing.

Preceding regression analyses, biological and methodological covariates including gender, age, post-mortem interval, brain bank, experimental batch and the first two principle component vectors from multi-dimensional scaling were taken into consideration. Each trait was adjusted in an effort to reduce the influence of systematic confounding effects using mach2qtl v1.11.

A secondary round of analysis was performed because it was noted that two independent PD risk variants within *SNCA* were strongly associated with DNA methylation at the same CpG (cg08767460). This was performed in order to test whether the association between these risk alleles and DNA methylation was independent. To test this, the linear modeling was performed for the second SNP (rs3910105) including the allele dosage at the first SNP (rs356181) as a covariate in the adjustment of methylation values.

4.3 Results

Of the 486,428 assayed CpG sites on the Illumina HumanMethylation450 array, 20,615 were within 1Mb of the top SNP within the loci of interest. Of these 17,620 passed initial quality control. Some of the risk loci overlapped physically; therefore, a total of 25,292 SNP-CpG pairs were tested. In order to be conservative and due to polymorphisms within several probes, which can create significant bias in QTL analysis, all probes against a sequence that contained a polymorphism at >2% allele frequency in Caucasians was removed. This resulted in 3,580 testable CpG sites, and 5,473 testable CpG-SNP pairs. Of the remaining CpG-SNP pairs, 3,302 CpG sites were annotated as being within a CpG island or a CpG island shore, with the remaining 2,171 sites being inter CpG island (1,432 within a CpG Island, 299 in a North Shelf, 659 in a North Shore, 338 in a South Shelf, and 574 within a South Shore).

Analysis of the dmQTL data revealed that significant dmQTLs existed for top PD associated SNPs at 19 of the 28 identified PD risk loci (Table 6, Figure 23 to Figure 41 inclusive). Within these significant loci, the most significant CpG site was within an annotated feature 68% of the time, slightly (but not significantly) more than expected (60% based on chance).

Top PD associated SNPs at the remaining nine loci failed to show an association with proximal DNA methylation levels based on an FDR corrected p value (raw p value shown, Supplementary Figure 12 through Supplementary Figure 20 inclusive).

Before removal of probes targeting a polymorphic region, the number of significantly associated loci was 24, highlighting the potential extent of the 'polymorphism in probe' problem, which has been observed extensively in expression QTL analysis [297].

The most associated probes with each of the 19 SNPs ranged in significance from p=0.0009 on chromosome 16 to $p=3.4x10^{-79}$ at the *MAPT* locus.

SNP	Chr	CpG	Beta	SD	p	pAdj	Annotation	UCSC RefGene Name	UCSC RefGene Group
rs823118	1	cg14893161	-0.83	0.06	3.5E-33	5.0E-30	South Shore	*PM20D1*	TSS200
rs10797576	1	cg16244711	-0.41	0.11	0.0002	0.0117	-	*SIPA1L2*	TSS200
rs6430538	2	cg02741327	-0.38	0.08	4.2E-06	0.0004	-	*TMEM163*	Body
rs34884217	4	cg06157924	-0.92	0.11	5.4E-15	1.6E-12	South Shore	*TMEM175*	Body
rs34311866	4	cg14530993	0.71	0.08	6.7E-19	3.1E-16	South Shelf	*GAK*	Body
rs6812193	4	cg20432211	-0.26	0.06	5.2E-05	0.004	Island	*STBD1*	TSS1500
rs356181	4	cg08767460	0.36	0.08	9.7E-06	0.0008	South Shore	*SNCA*	TSS1500 5'UTR
rs3910105	4	cg15133208	-0.48	0.08	4.7E-10	7.8E-08	North Shore	*SNCA*	5'UTR
exm535099	6	cg08188698	-1.80	0.18	2.0E-20	1.1E-17	-	*ATF6B*	Body
rs115462410	6	cg08265274	1.34	0.10	1.6E-30	1.6E-27	South Shore	*HLA-DRB5*	Body
rs199347	7	cg14444376	0.71	0.07	5.7E-20	3.0E-17	North Shore	*GPNMB*	5'UTR 1stExon
rs329648	11	cg11936536	-0.82	0.08	2.8E-23	1.9E-20	North Shelf	-	-
rs76904798	12	cg25382486	-0.28	0.08	0.0006	0.031	South Shelf	-	-
rs11060180	12	cg06742321	-0.40	0.08	2.0E-06	0.0002	-	*PITPNM2*	TSS200
rs11158026	14	cg22955899	-0.50	0.08	3.7E-09	5.3E-07	-	*DLGAP5*	Body
rs2414739	15	cg07084345	-0.56	0.09	5.7E-10	9.3E-08	-	-	-
rs14235	16	cg16747885	0.22	0.07	0.0009	0.04	South Shelf	-	-
rs11868035	17	cg15030378	-0.87	0.08	1.8E-24	1.3E-21	Island	*SREBF1*	Body
rs17649553	17	cg24801230	1.34	0.05	3.4E-79	4.3E-75	South Shelf	*MAPT*	5'UTR

Table 6: Significant dmQTLs identified at PD associated loci.

Shown are the SNP used, and the CpG site for which methylation was modeled as the trait. Also shown are the raw p value, and the pAdj, which is the locus specific FDR corrected p value. CpG sites were annotated as within a CpG island or on the North/South shores and shelves of CpG Islands.

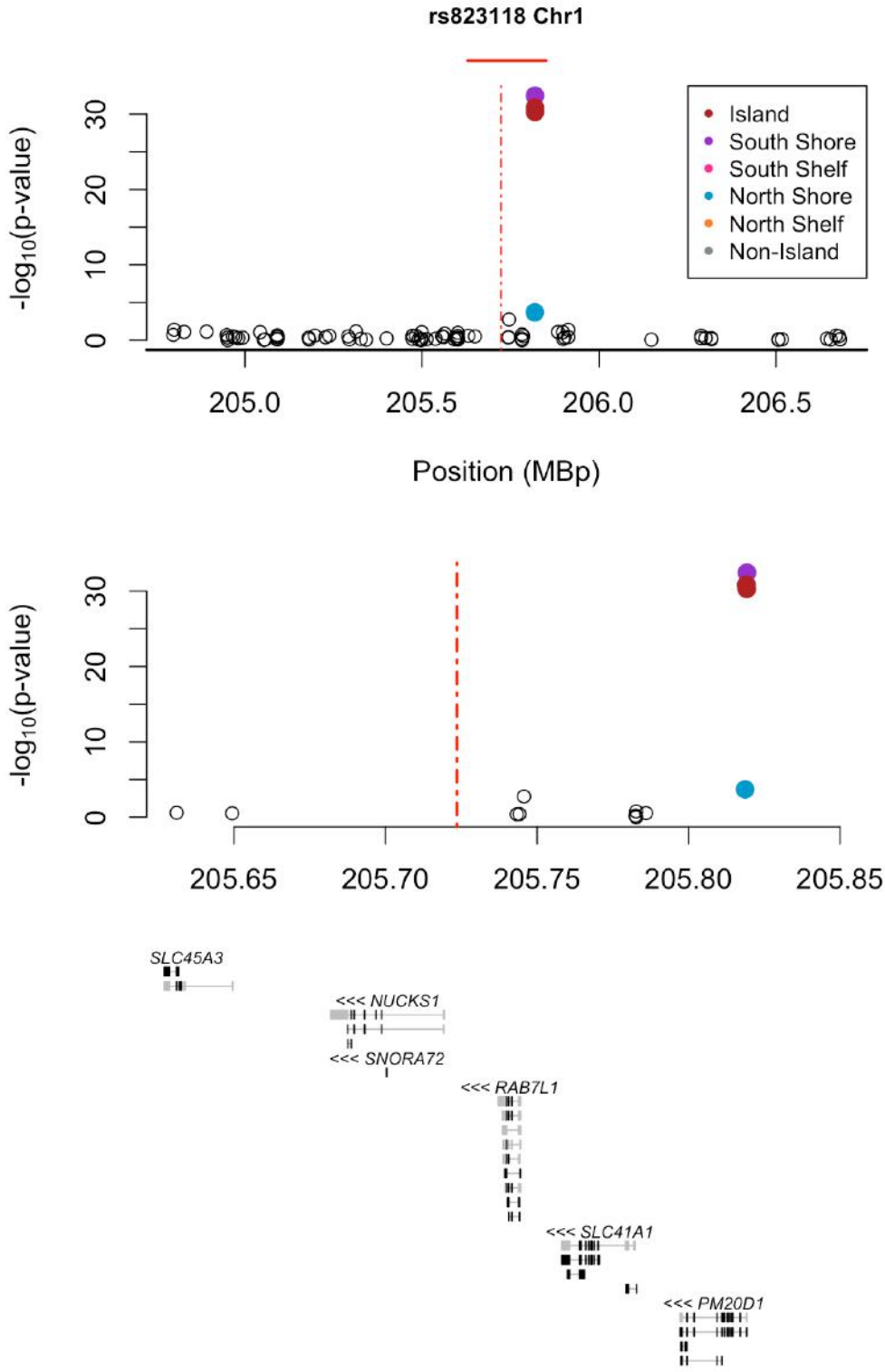


Figure 23: PD risk allele rs823118 on chromosome 1 is a dmQTL.

Allele burden at rs823118 was assessed for association with DNA methylation levels at proximal CpG sites. Non-associated CpG sites are shown as open symbols, significantly associated CpG sites are shown as filled symbols, colored according to annotated functional position.

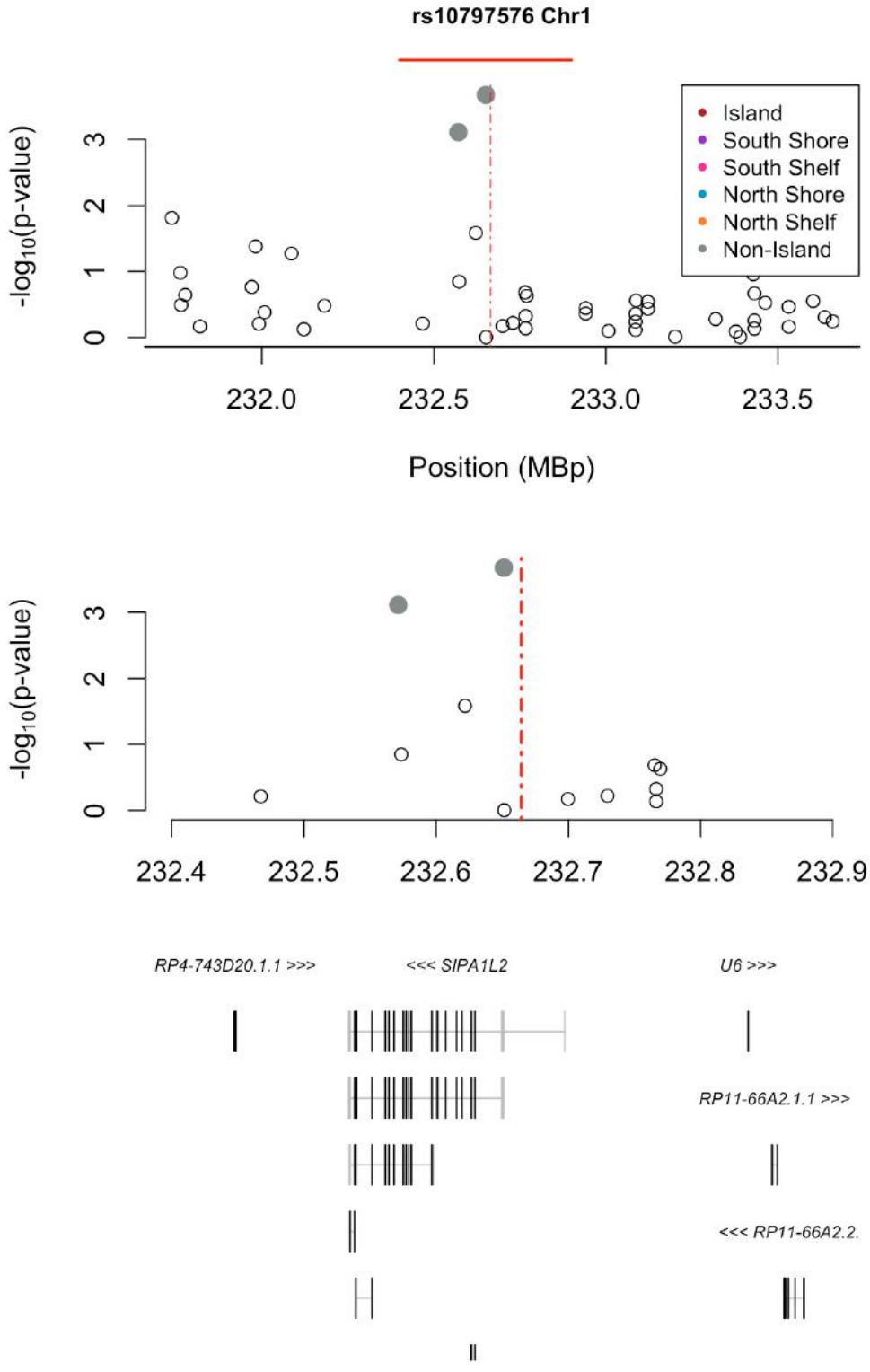


Figure 24: PD risk allele rs10797576 on chromosome 1 is a dmQTL.

Allele burden at rs10797576 was assessed for association with DNA methylation levels at proximal CpG sites. Non-associated CpG sites are shown as open symbols, significantly associated CpG sites are shown as filled symbols, colored according to annotated functional position.

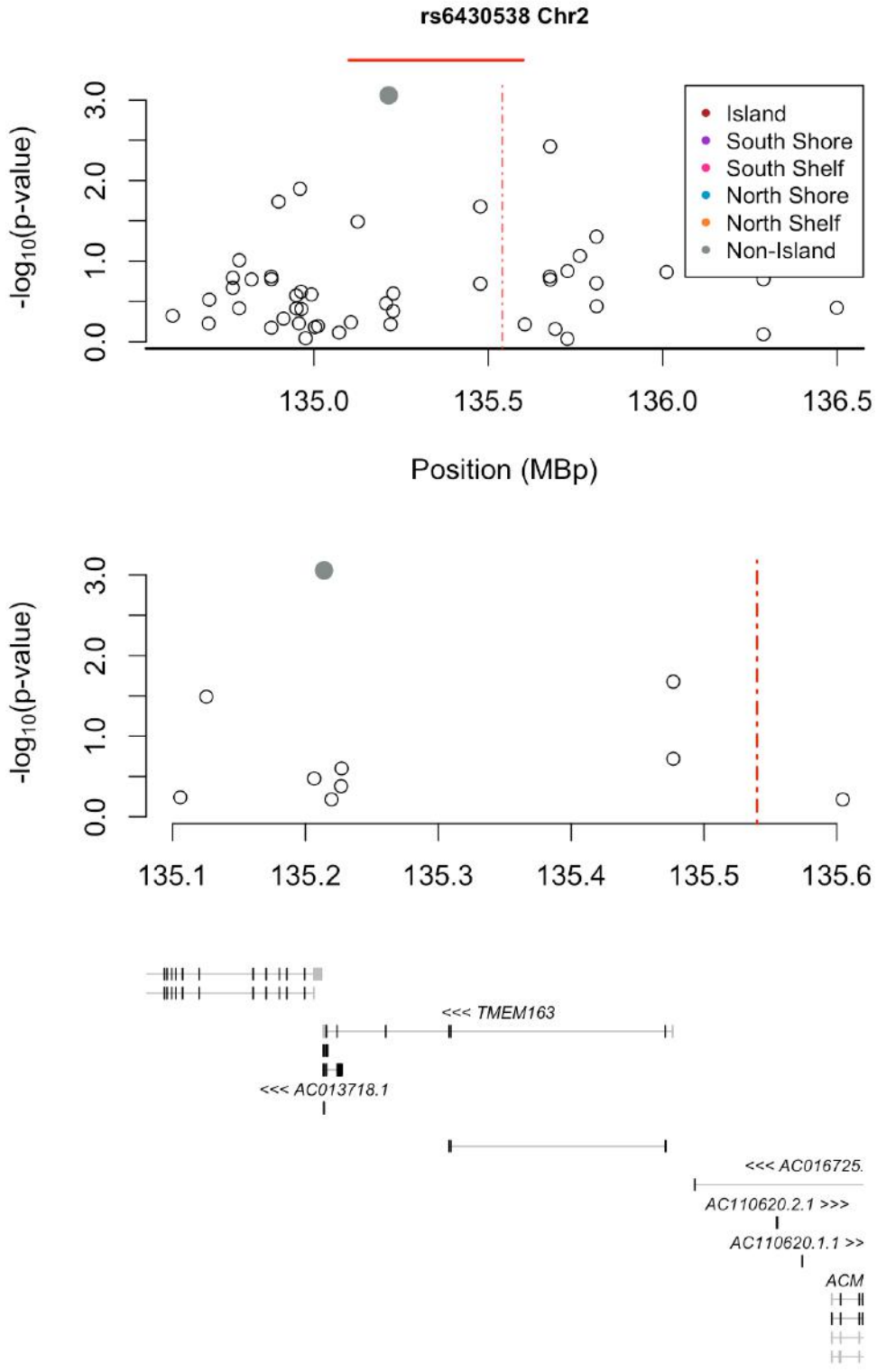


Figure 25: PD risk allele rs6430538 on chromosome 2 is a dmQTL.

Allele burden at rs6430538 was assessed for association with DNA methylation levels at proximal CpG sites. Non-associated CpG sites are shown as open symbols, significantly associated CpG sites are shown as filled symbols, colored according to annotated functional position.

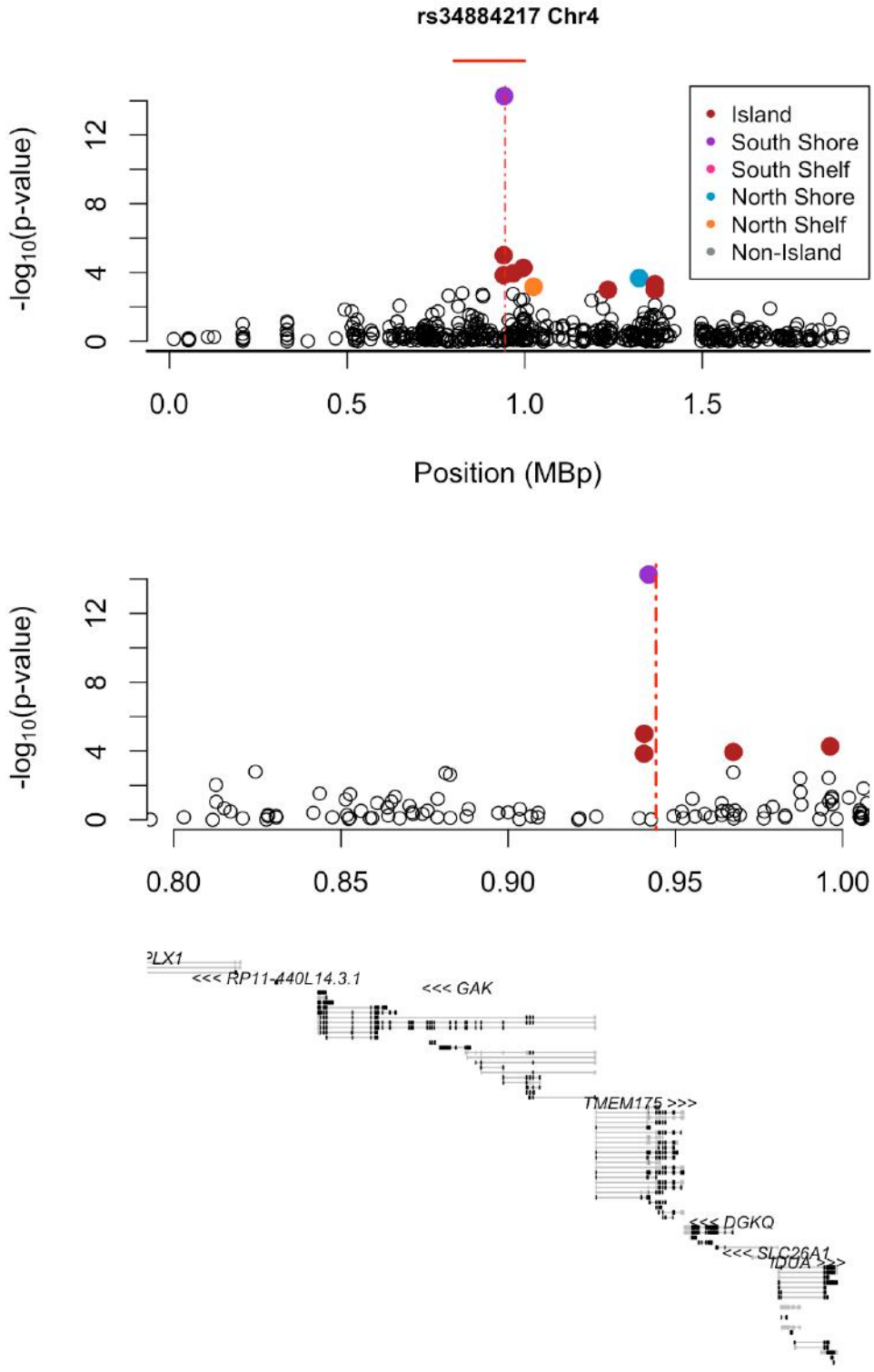


Figure 26: PD risk allele rs34884217 on chromosome 4 is a dmQTL.

Allele burden at rs34884217 was assessed for association with DNA methylation levels at proximal CpG sites. Non-associated CpG sites are shown as open symbols, significantly associated CpG sites are shown as filled symbols, colored according to annotated functional position.

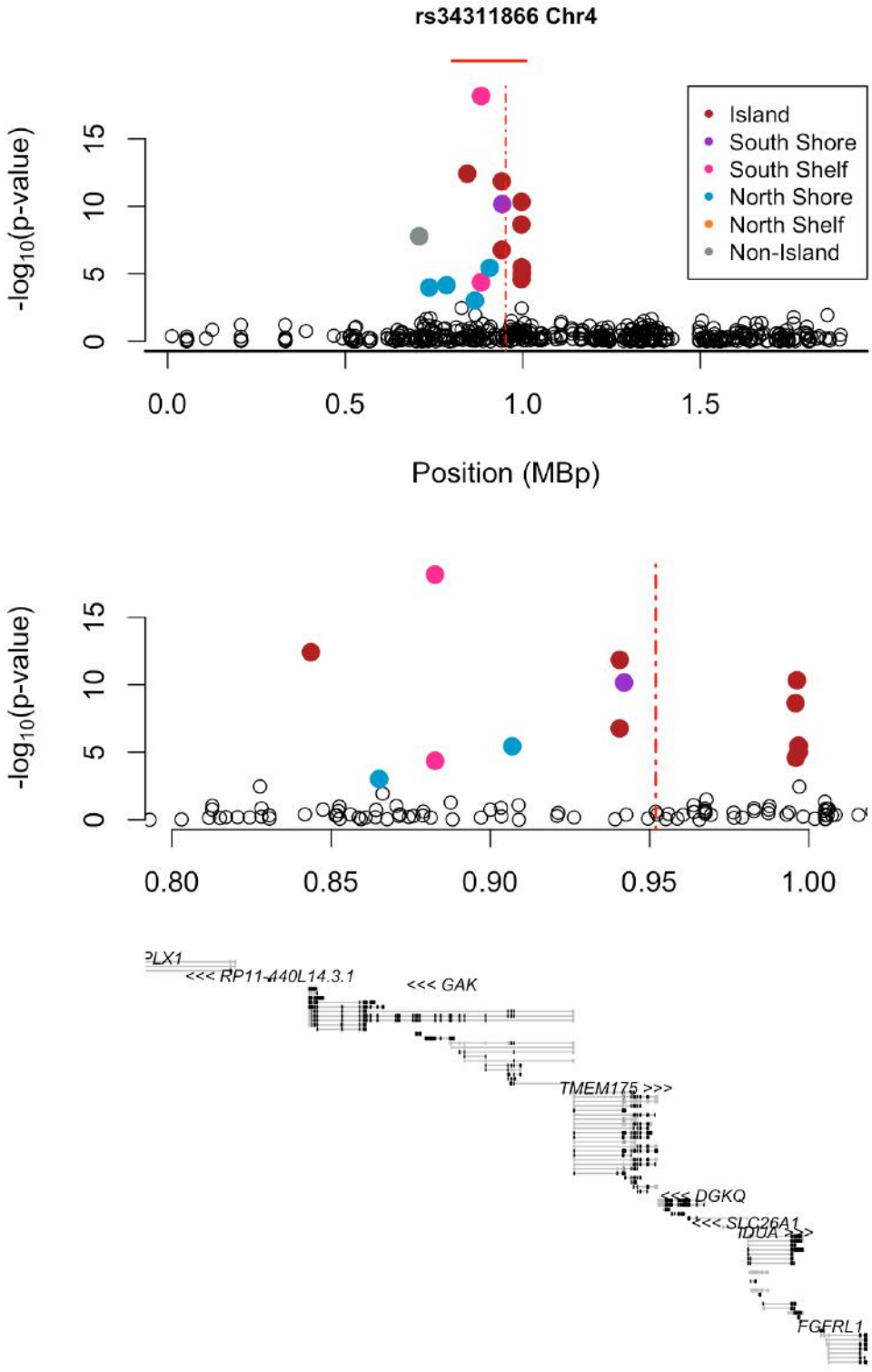


Figure 27: PD risk allele rs34311866 on chromosome 4 is a dmQTL.

Allele burden at rs34311866 was assessed for association with DNA methylation levels at proximal CpG sites. Non-associated CpG sites are shown as open symbols, significantly associated CpG sites are shown as filled symbols, colored according to annotated functional position

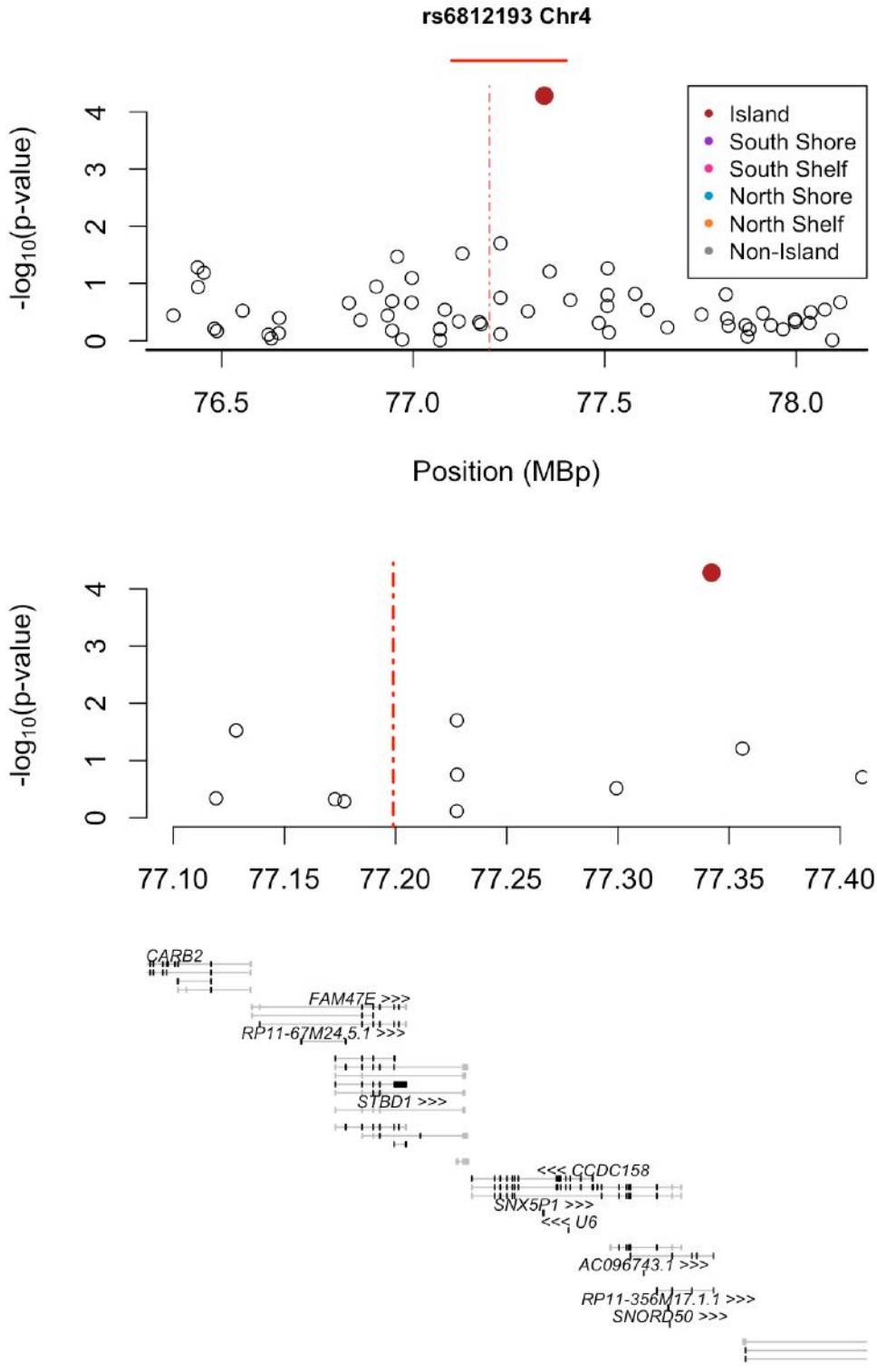


Figure 28: PD risk allele rs6812193 on chromosome 4 is a dmQTL.

Allele burden at rs6812193 was assessed for association with DNA methylation levels at proximal CpG sites. Non-associated CpG sites are shown as open symbols, significantly associated CpG sites are shown as filled symbols, colored according to annotated functional position

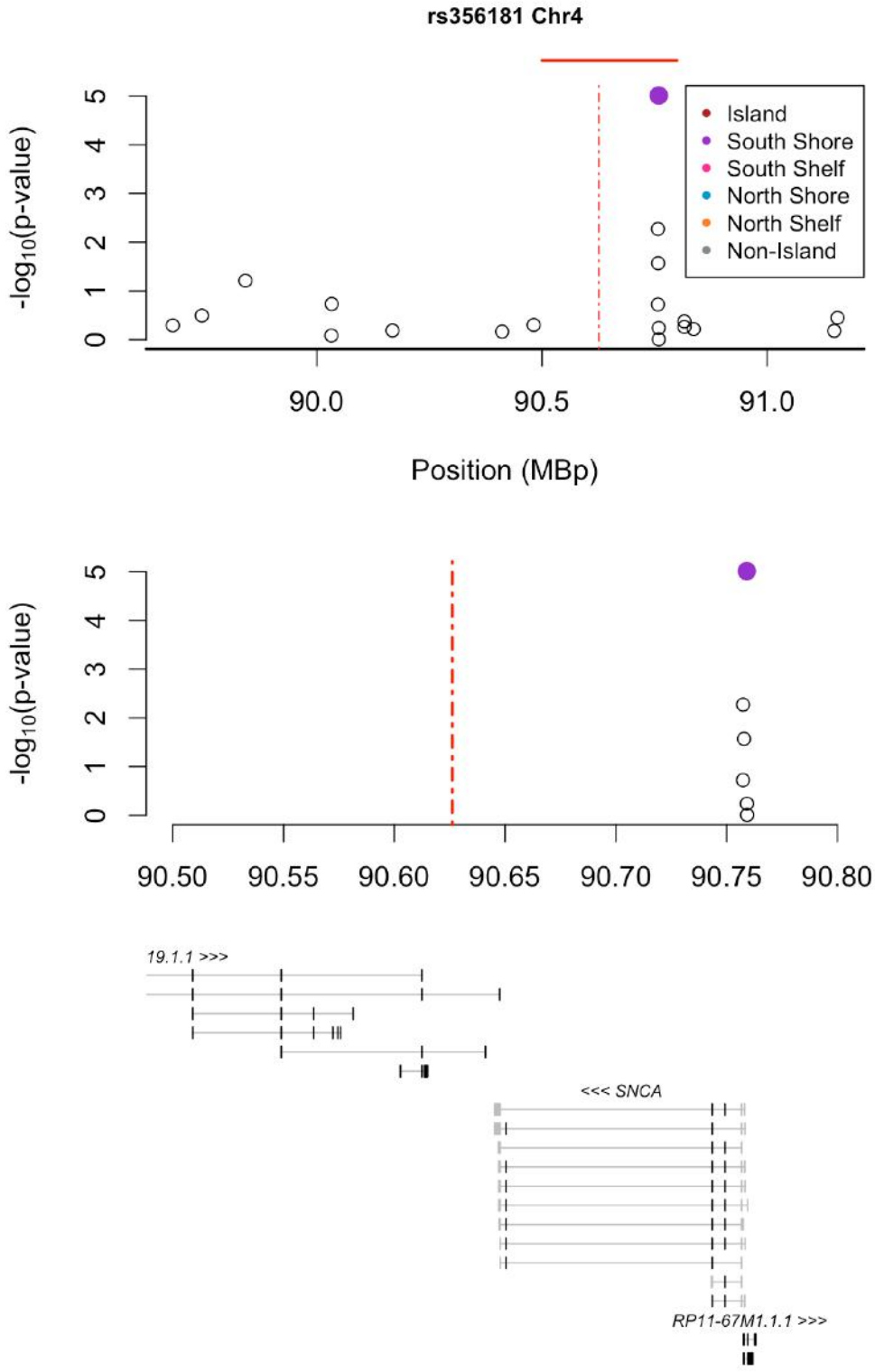


Figure 29: PD risk allele rs356181 on chromosome 4 is a dmQTL.

Allele burden at rs356181 was assessed for association with DNA methylation levels at proximal CpG sites. Non-associated CpG sites are shown as open symbols, significantly associated CpG sites are shown as filled symbols, colored according to annotated functional position

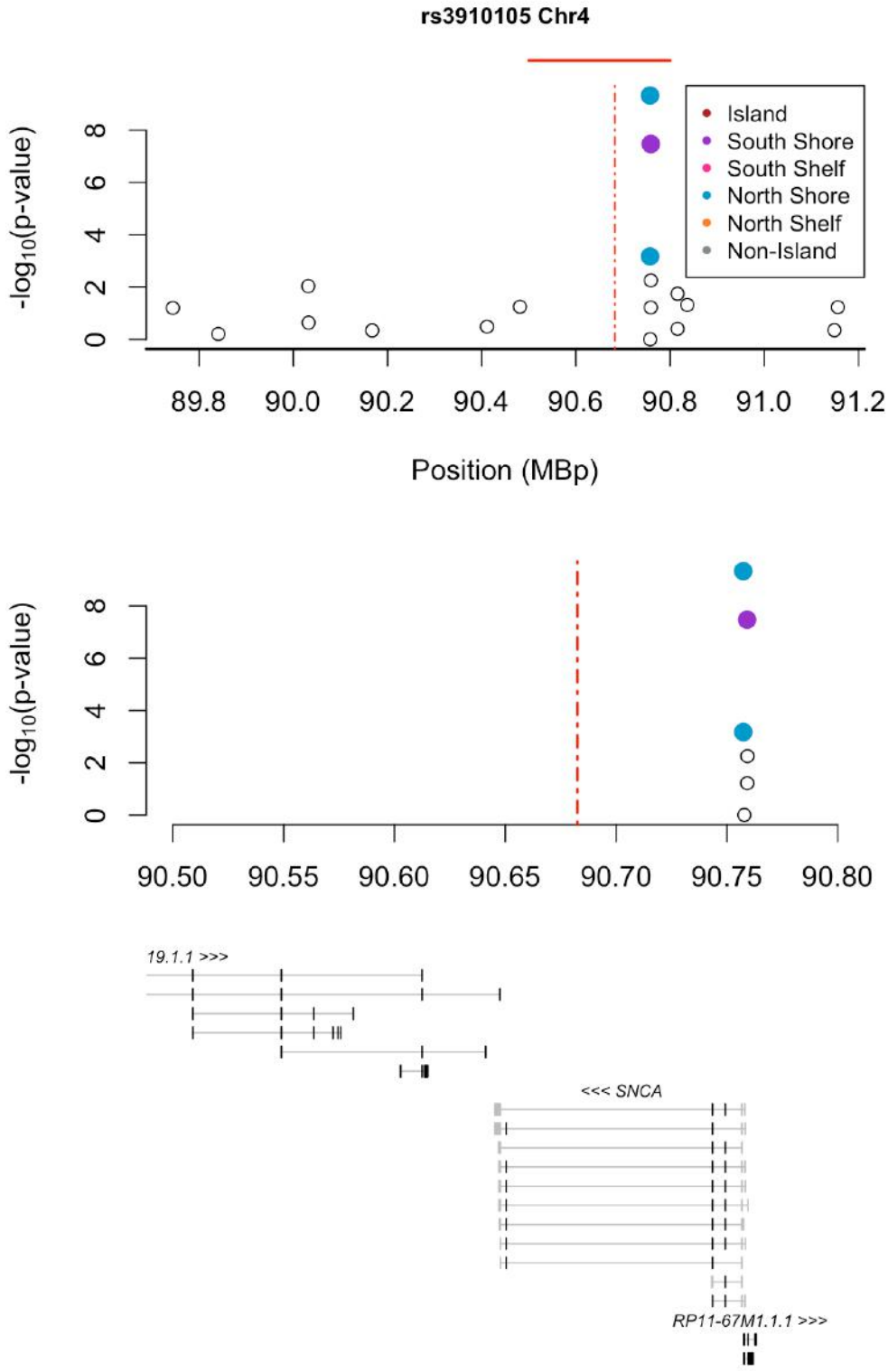


Figure 30: PD risk allele rs3910105 on chromosome 4 is a dmQTL.

Allele burden at rs3910105 was assessed for association with DNA methylation levels at proximal CpG sites. Non-associated CpG sites are shown as open symbols, significantly associated CpG sites are shown as filled symbols, colored according to annotated functional position

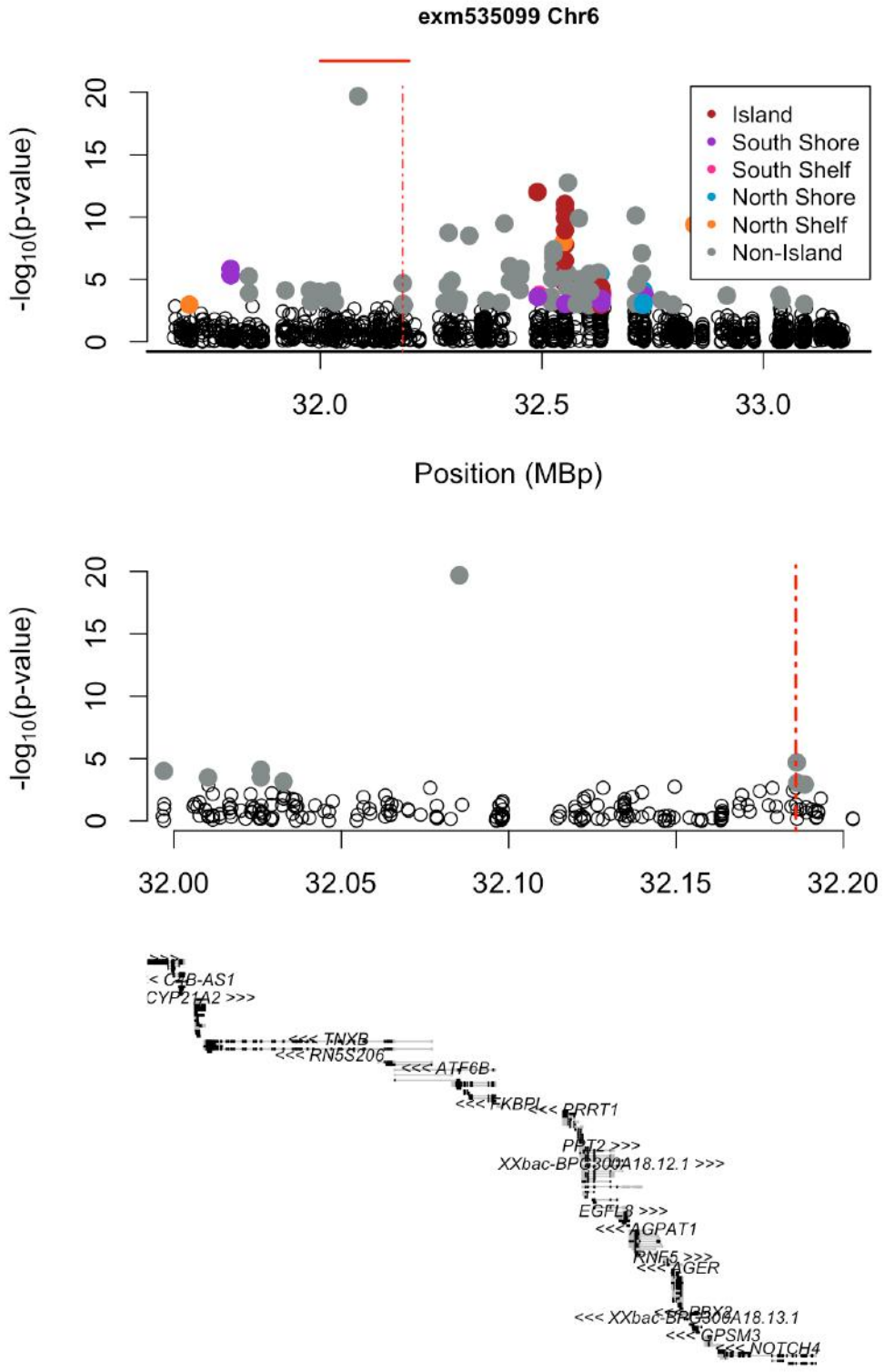


Figure 31: PD risk allele exm535099 on chromosome 6 is a dmQTL.

Allele burden at exm535099 was assessed for association with DNA methylation levels at proximal CpG sites. Non-associated CpG sites are shown as open symbols, significantly associated CpG sites are shown as filled symbols, colored according to annotated functional position

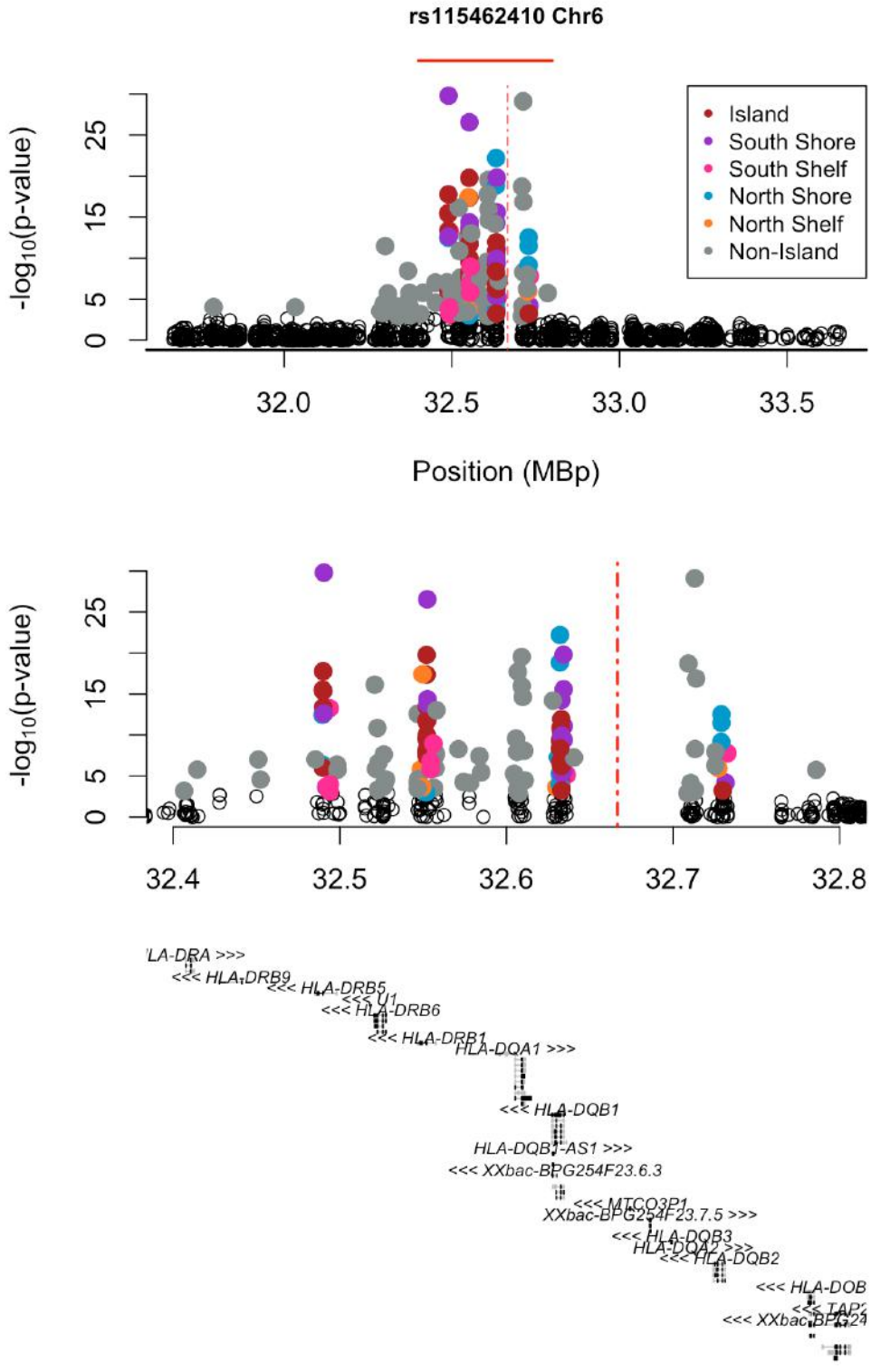


Figure 32: PD risk allele rs115462410 on chromosome 6 is a dmQTL.

Allele burden at rs115462410 was assessed for association with DNA methylation levels at proximal CpG sites. Non-associated CpG sites are shown as open symbols, significantly associated CpG sites are shown as filled symbols, colored according to annotated functional position

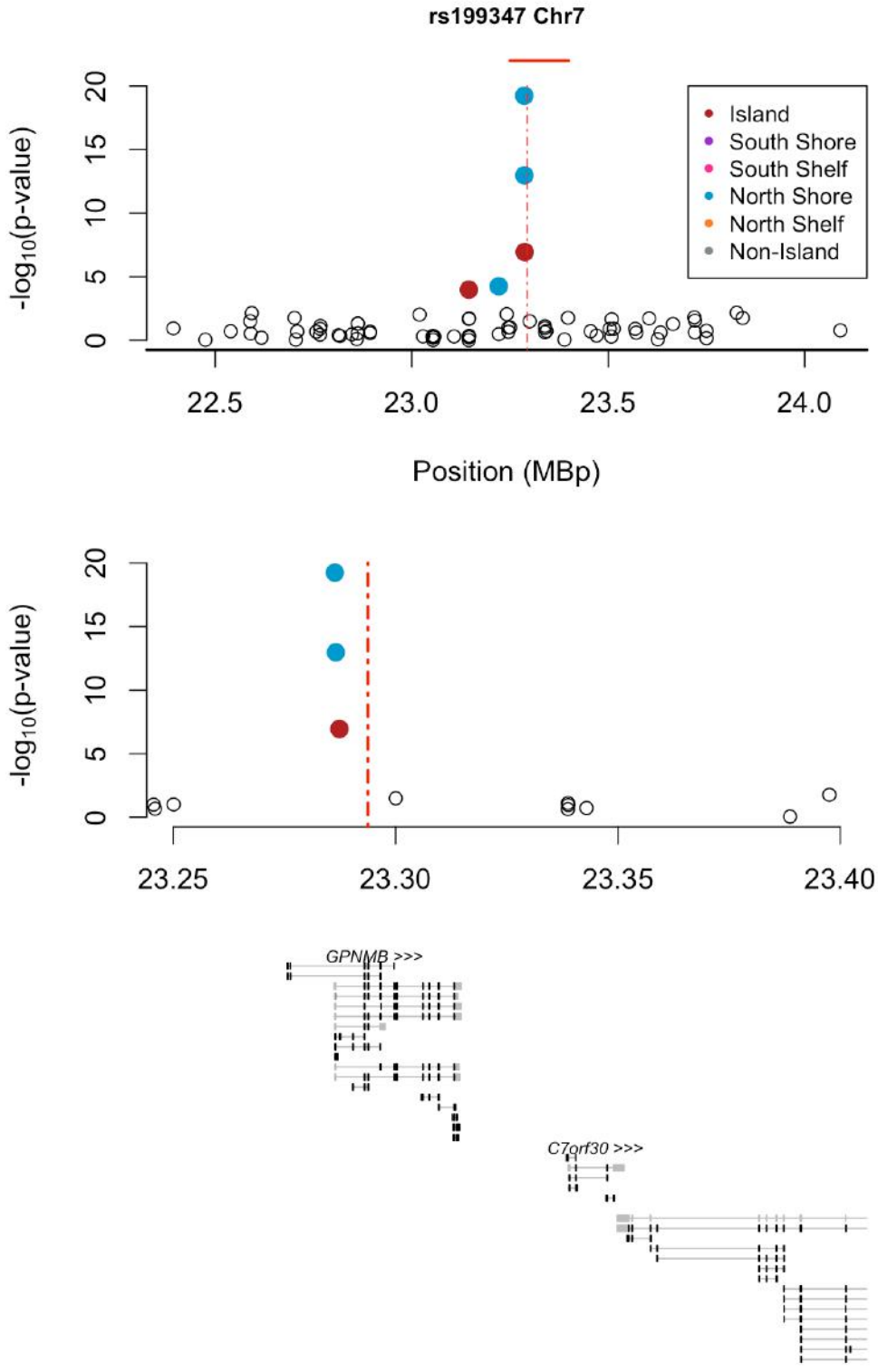


Figure 33: PD risk allele rs199347 on chromosome 7 is a dmQTL.

Allele burden at rs199347 was assessed for association with DNA methylation levels at proximal CpG sites. Non-associated CpG sites are shown as open symbols, significantly associated CpG sites are shown as filled symbols, colored according to annotated functional position

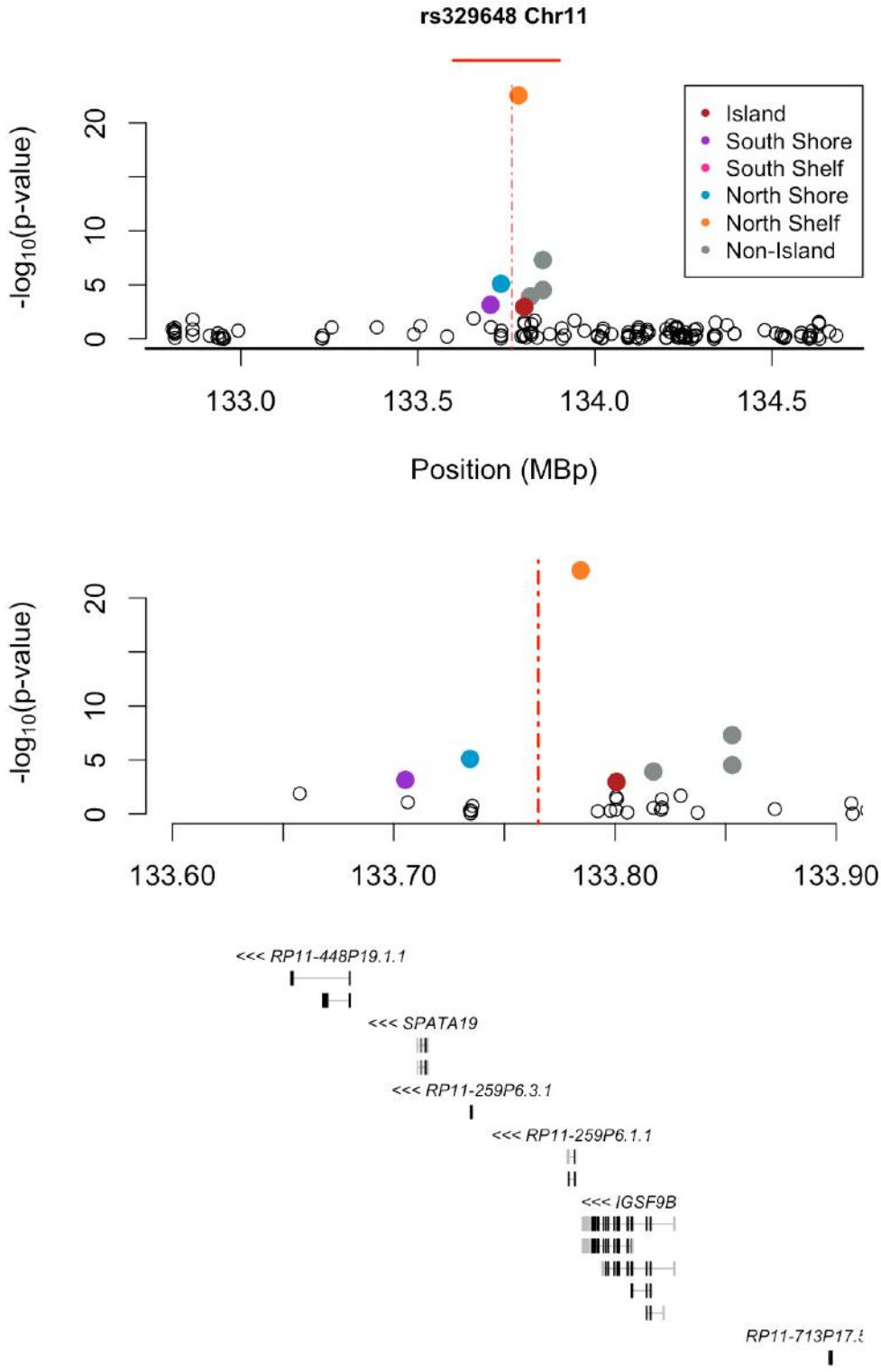


Figure 34: PD risk allele rs329648 on chromosome 11 is a dmQTL.

Allele burden at rs329648 was assessed for association with DNA methylation levels at proximal CpG sites. Non-associated CpG sites are shown as open symbols, significantly associated CpG sites are shown as filled symbols, colored according to annotated functional position

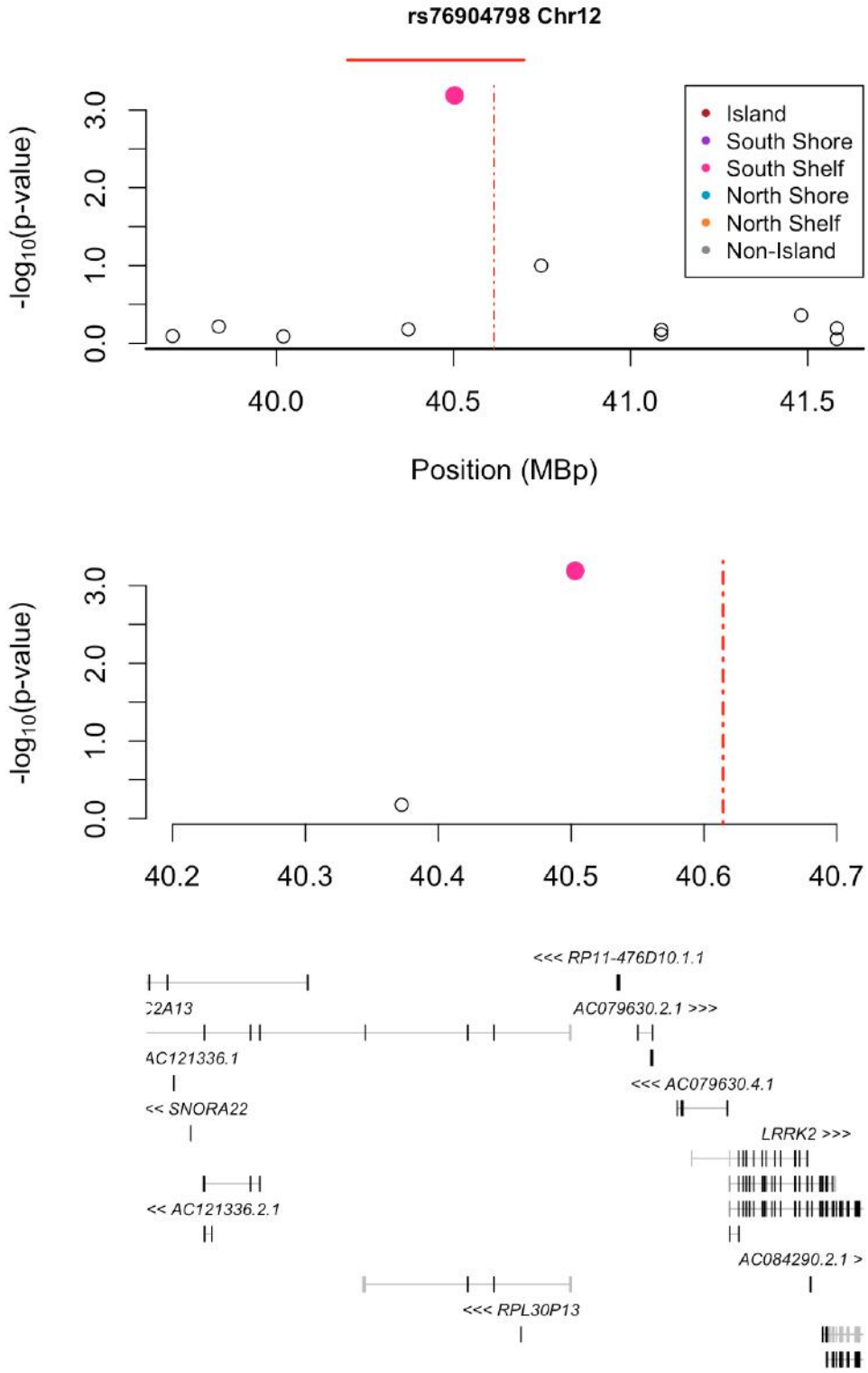


Figure 35: PD risk allele rs76904798 on chromosome 12 is a dmQTL.

Allele burden at rs76904798 was assessed for association with DNA methylation levels at proximal CpG sites. Non-associated CpG sites are shown as open symbols, significantly associated CpG sites are shown as filled symbols, colored according to annotated functional position

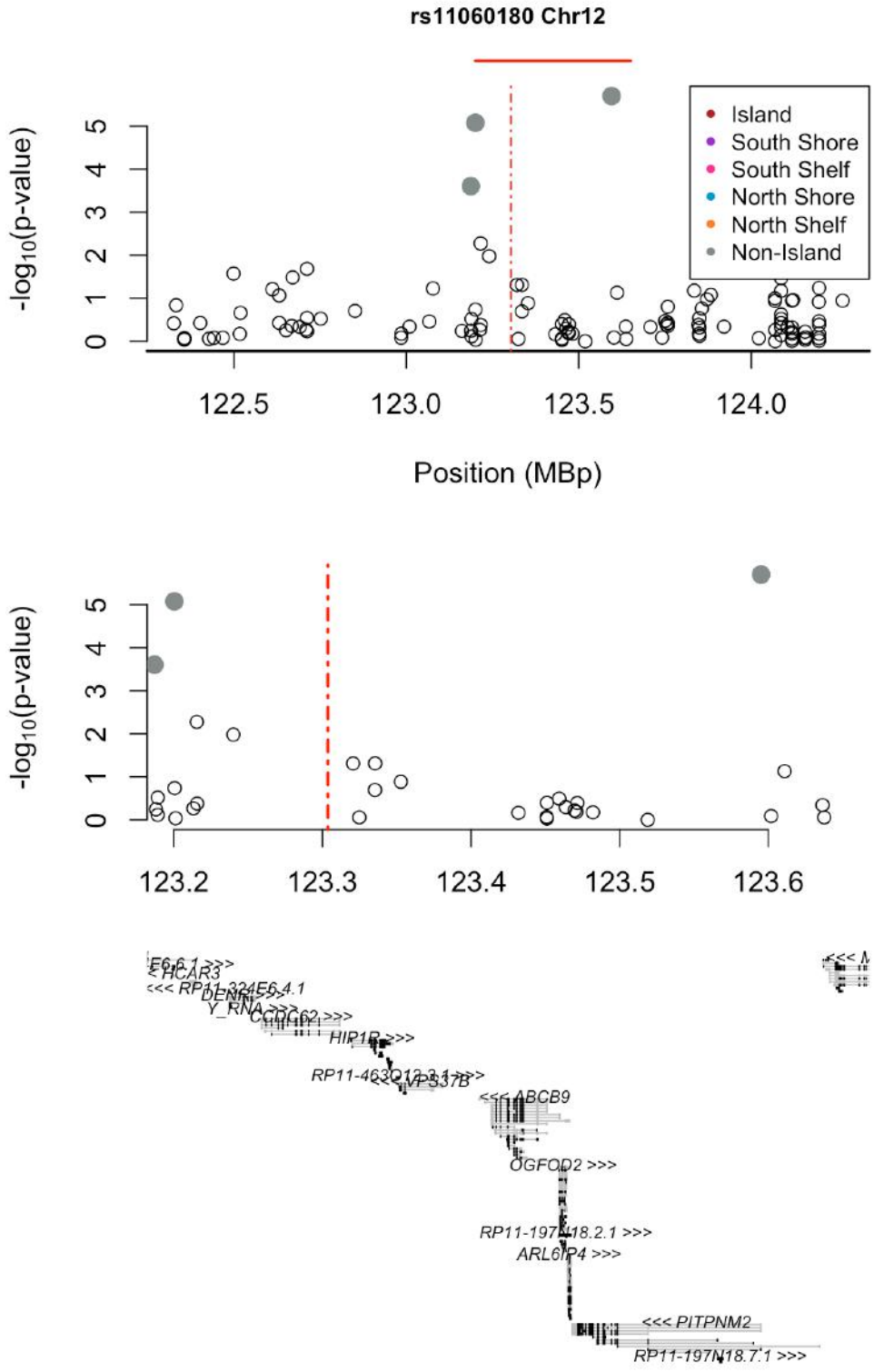


Figure 36: PD risk allele rs11060180 on chromosome 12 is a dmQTL.

Allele burden at rs11060180 was assessed for association with DNA methylation levels at proximal CpG sites. Non-associated CpG sites are shown as open symbols, significantly associated CpG sites are shown as filled symbols, colored according to annotated functional position

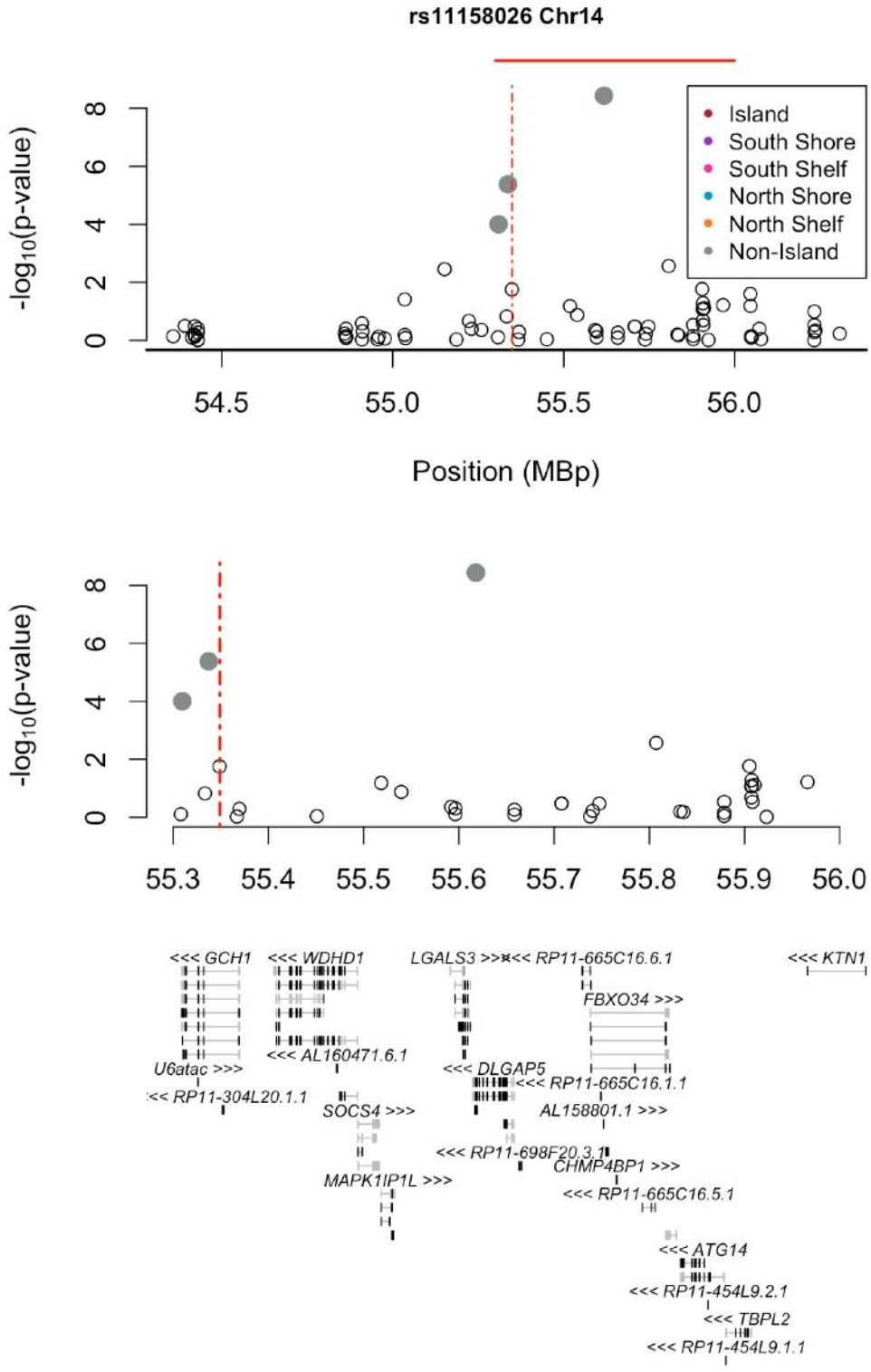


Figure 37: PD risk allele rs11158026 on chromosome 14 is a dmQTL.

Allele burden at rs11158026 was assessed for association with DNA methylation levels at proximal CpG sites. Non-associated CpG sites are shown as open symbols, significantly associated CpG sites are shown as filled symbols, colored according to annotated functional position

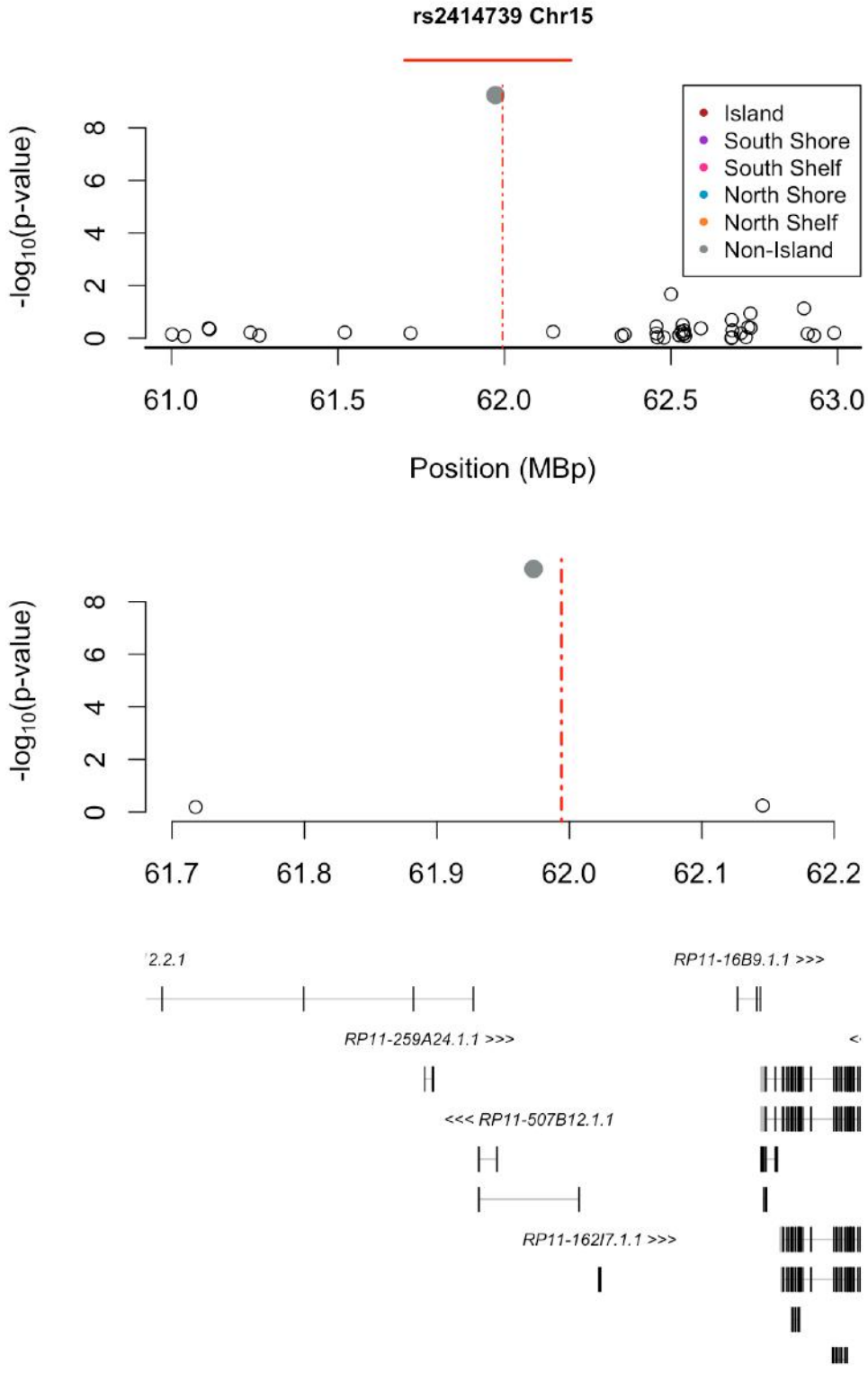


Figure 38: PD risk allele rs2414739 on chromosome 15 is a dmQTL.

Allele burden at rs2414739 was assessed for association with DNA methylation levels at proximal CpG sites. Non-associated CpG sites are shown as open symbols, significantly associated CpG sites are shown as filled symbols, colored according to annotated functional position

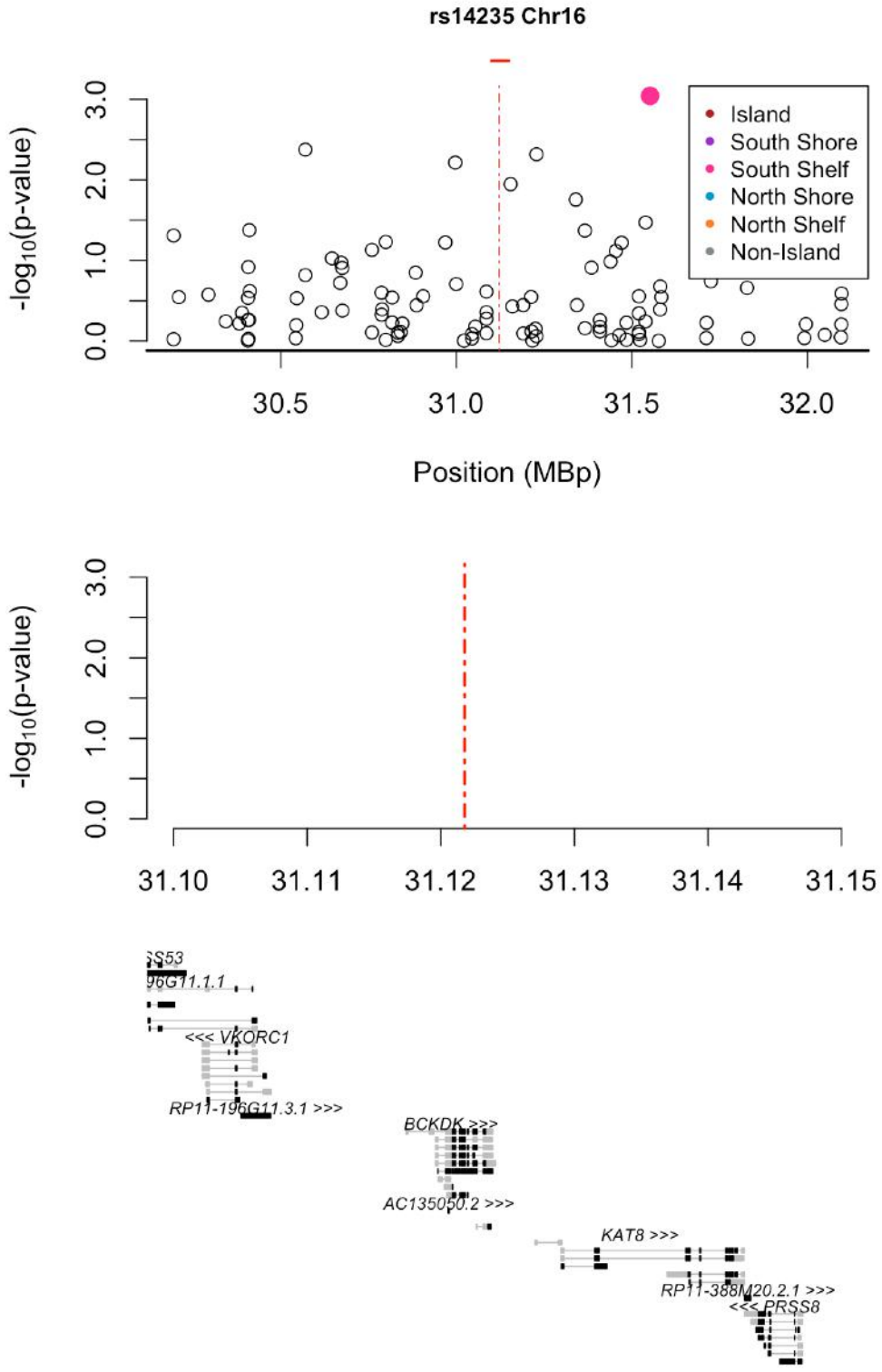


Figure 39: PD risk allele rs14235 on chromosome 16 is a dmQTL.

Allele burden at rs14235 was assessed for association with DNA methylation levels at proximal CpG sites. Non-associated CpG sites are shown as open symbols, significantly associated CpG sites are shown as filled symbols, colored according to annotated functional position

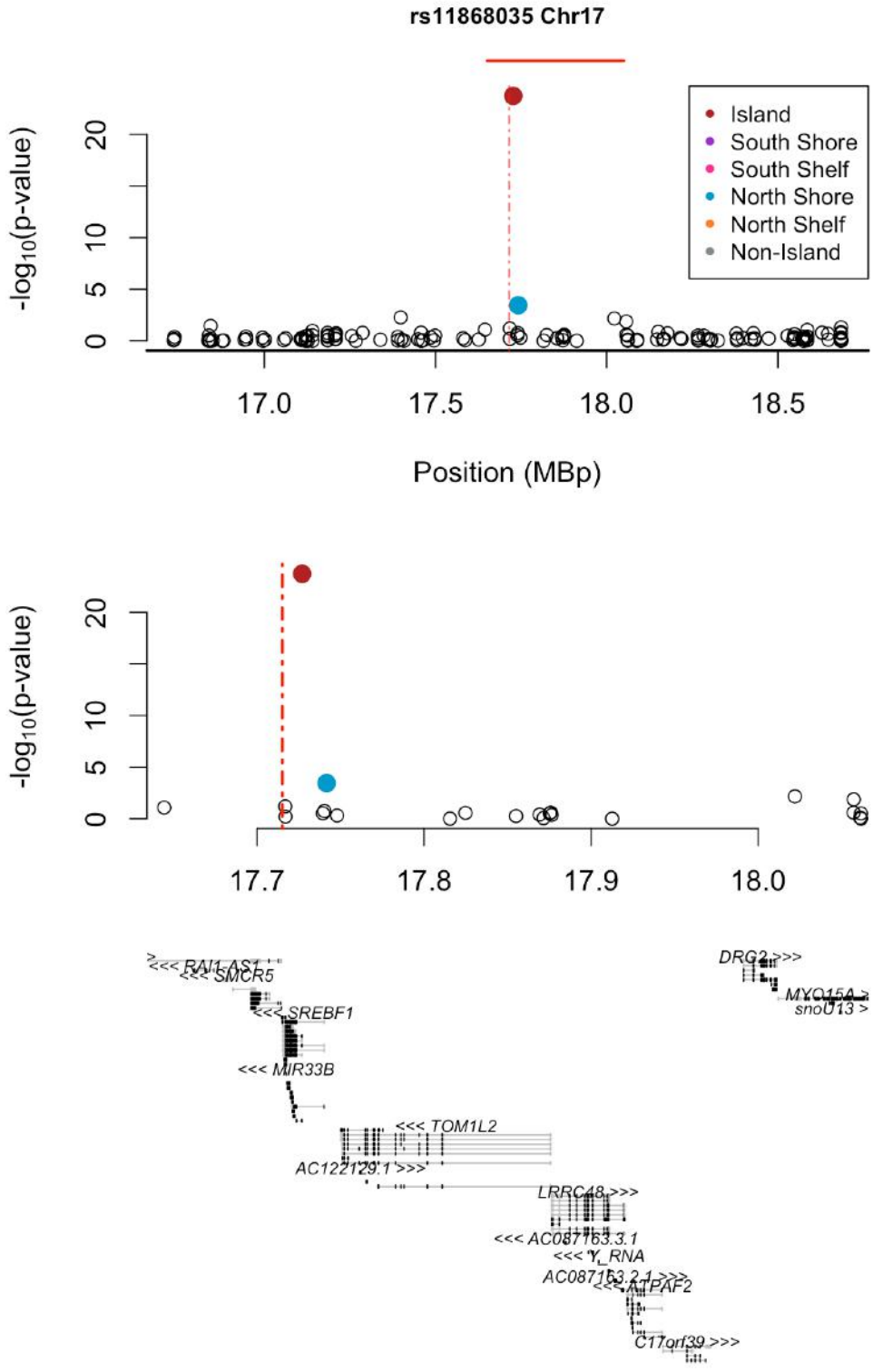


Figure 40: PD risk allele rs11868035 on chromosome 17 is a dmQTL.

Allele burden at rs11868035 was assessed for association with DNA methylation levels at proximal CpG sites. Non-associated CpG sites are shown as open symbols, significantly associated CpG sites are shown as filled symbols, colored according to annotated functional position

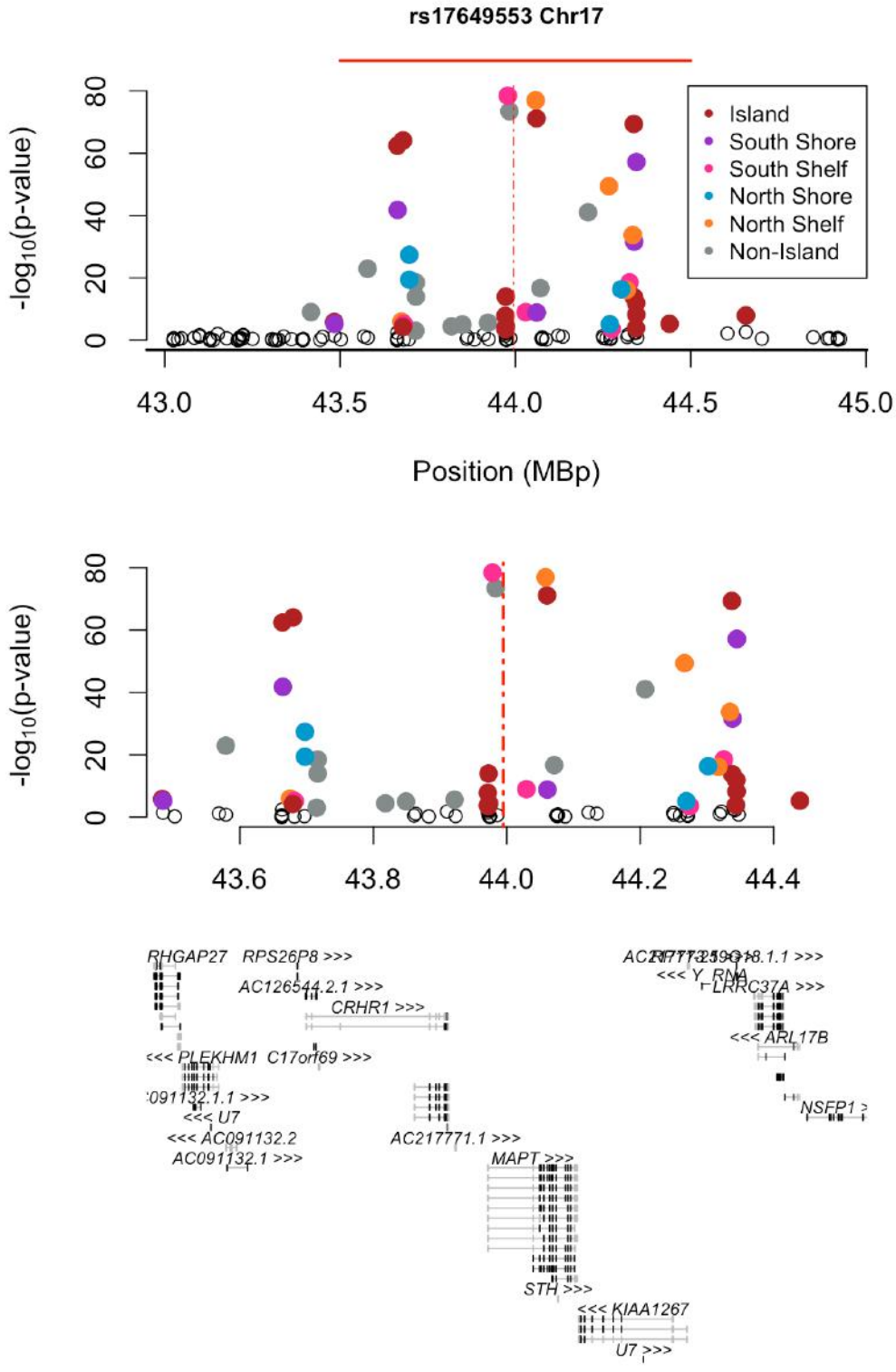


Figure 41: PD risk allele rs17649553 on chromosome 17 is a dmQTL.

Allele burden at rs17649553 was assessed for association with DNA methylation levels at proximal CpG sites. Non-associated CpG sites are shown as open symbols, significantly associated CpG sites are shown as filled symbols, colored according to annotated functional position.

Notably there are two confirmed independent PD risk variants within *SNCA*, one 3' to the gene (rs356181), and another within intron 4 (rs3910105). The current analysis revealed that both of these variants are also dmQTLs with the top associated CpG being cg15133208 for the intronic variant, and cg08767460 for the 3' variant.

It is also notable that while the top CpG association with rs3910105 is at CpG site cg15133208, there is also a strong association between allele dosage at this SNP and methylation at site cg08767460 ($p=3.8x10^{-8}$) (Figure 30). Thus both risk SNPs in *SNCA,* despite being ~55kb apart, exert an effect on methylation at the same CpG site. Annotating the alleles at these SNPs in terms of risk for PD reveals that the association with DNA methylation is in the same direction for both SNPs, with high-risk alleles being associated with higher DNA methylation (Figure 42).

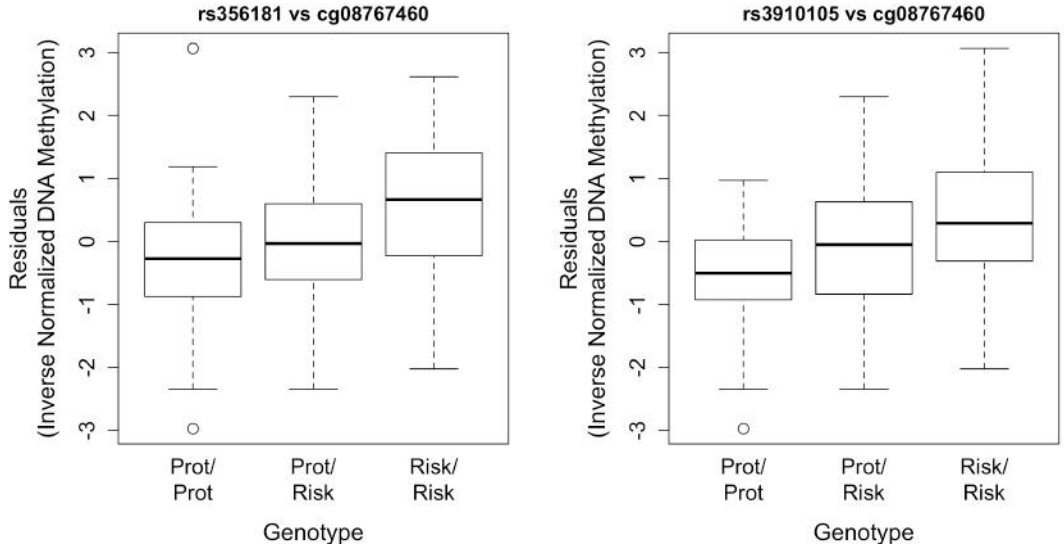


Figure 42: Association of *SNCA* risk alleles with DNA methylation at cg08767460.

At both loci within *SNCA* the PD risk alleles are associated with increased methylation at CpG site cg08767460.

Covariate analysis was performed in an attempt to correct for any potential confounding of the association at rs3910105 because of linkage disequilibrium with rs356181, which is also strongly associated with cg08767460 methylation. This adjusted analysis confirmed that a statistically significant association remains between rs3910105 and methylation at cg08767460 suggesting that this is a truly independent effect, and not one driven by residual LD between rs3910105 and rs356181 (p=0.001, Beta = 0.55, Standard Error = 0.16; Figure 43). The r^2 between these two SNPs based on 1000 genomes Caucasian data is nominal (r^2=0.271).

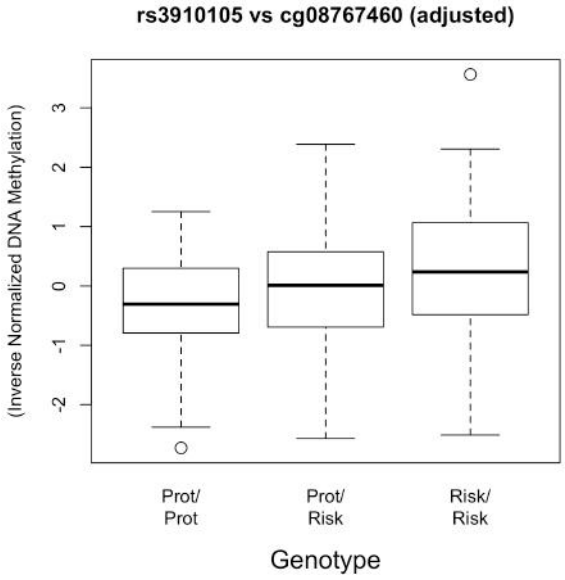


Figure 43: Association of allele dosage at rs3910105 and methylation at cg08767460.

In this analysis methylation levels have been adjusted for allele dosage at the proximal CpG methylation associated risk variant rs3910105. Analysis of this association using a linear model adjusting for genotype at rs356181 and other biological covariates revealed a significant association (p=0.001).

4.4 Discussion

This chapter describes the integration of known PD loci implicated by large GWA studies, with a map of the genetic control of DNA methylation in human brain. The aim of this effort was to link definitively identified risk alleles with a biological effect in order to gain potential mechanistic insight into the pathogenesis of disease.

The current work identified significant dmQTLs for 19 of 28 identified PD risk loci, demonstrating that the association of risk alleles for neurological disease with a biologically relevant trait in human brain tissue is a tractable goal.

Examining the loci here, there are several that are of immediate interest. Perhaps most notable is the *SNCA* locus, which contains two independent PD risk alleles [242, 286]. In the current work, both alleles show significant association with methylation of a CpG Island and Island Shore close to the promoter region of *SNCA*. It is also notable that these variants, which confer independent risk effects for PD, also confer independent effects on CpG methylation. Lastly, in both instances the risk allele at each SNP is associated with increased DNA methylation, showing consistency of effect. Previous data has suggested an association between *SNCA* risk alleles and increased *SNCA* expression [298], although these data were not definitive. The current data however suggests a mechanism for how expression may be modulated at *SNCA*.

The status of DNA methylation at *SNCA* has been assessed by several groups previously; however, these studies in general examined DNA methylation in PD cases versus controls, rather than methylation levels linked to disease risk alleles [218, 299]. Previous studies show inconsistent results, although this may be a feature of the generally small sample size used and because only limited and varied CpG sites were studied.

As with *SNCA*, expression at *MAPT* has been associated with genetic risk alleles, although it was not clear until recently, that the observed effects were not simple changes in expression but rather represented changes in exon usage [300]. Examining the *MAPT* locus in the current data reveals a complex picture, with multiple significant dmQTLs across the entirety of the locus. While it is notable that the most significant dmQTL is with a CpG site within the 5' end of *MAPT,* the extensive linkage disequilibrium across this locus means that a large number of CpG sites are also associated with the risk alleles at this locus.

Another particularly interesting locus, given recent functional work, is the PARK16 locus on chromosome 1 (Figure 23). Most recently, *RAB7L1* within the PARK16 locus was functionally implicated in PD, as the protein product interacts with the protein product of another PD gene, *LRRK2* [301, 302]. However, in the current work an extremely strong association with DNA methylation at a CpG site distal to *RAB7L1* and 5' to *PM20D1* is observed. There are several potential reasons for this observation. First, DNA methylation at this locus may be a genuine effect, but simply unrelated to PD pathogenesis; second, although unlikely, the lrrk2-rab7l1 interaction could be

spurious, or unrelated to disease; or third, the methylation change may exert an effect on *RAB7L1.* With regard to this last point, it is clear that regulatory elements may be quite distal to the regulated gene, and in one particular example at the *FTO* locus, a DNA regulatory element in one gene (*FTO*) exerts an effect through long-range functional connections with a distant gene (*IRX3*) [303]. Thus, while the identified regulated CpG site at the *PARK16* locus is quite distant from *RAB7L1,* this could indeed be the pathologically relevant effected gene.

This last illustrative example raises the question of where to go next with this work in order to further understand pathogenesis. Following up on individual loci, prioritizing those with the most biologically interesting signal/target is one route that could provide pertinent information. Although, a more comprehensive approach may be warranted, one that integrates further high content data to create high dimensional datasets. The addition of RNA sequencing data, including data generated in neuronal cells with varied stimuli and the mapping of regulatory elements and other epigenetic marks in disease relevant tissues/cells would add an extremely important element to this work with the potential to reveal the sequence of events leading from genotype to protein phenotype.

As a last point, revealing the epigenetic mediator of genotype may also have the potential to inform regarding other risk factors for disease. It is plausible that environmental and lifestyle risk factors for a slowly progressive, late onset disease may also exert pathogenic effects through an epigenetic effector and

therefore, it will be worth an investigation of the effects of candidate exposures on the epigenetic marks identified here.

5 Distinct DNA methylation changes highly correlated with chronological aging in the human brain

STATEMENT OF CONTRIBUTIONS TO THIS RESEARCH:

In this section, I describe a series of experiments that were performed to assess the relationship between DNA methylation and chronological age in the human brain. I was involved in the inception, planning and design of the experiments and analyses. I performed all of the experimental work and quality control for the genome-wide methylation dataset. I drafted the original published manuscript. I am one of two equally contributing first authors.

5.1 Introduction

A major risk factor for common neurodegenerative disease is aging [222, 304] In addition, DNA methylation patterns have been shown in many studies to alter with age [210, 228, 229, 305, 306] and vary globally or locally across different brain regions [307]. In a recent study, Steve Horvath reports on an epigenetic clock, using methylation levels at 353 CpG sites to accurately predict an individual's age. The method can be used to predict age in several tissues, with much improved accuracy using cortical brain tissue, where the clocks median error is 1.5 years [308]. Neurodegenerative diseases exhibit many non-Mendelian variances such as late age of onset, suggesting an epigenetic component may contribute to disease etiology. Therefore, it is important to better understand the landscape of DNA methylation patterns in the normal aging brain as a foundation for insight into the functional etiology of neurodegenerative diseases, where the primary risk factor is aging.

Analysis of the methylation data produced in neurologically normal controls (Chapter 2) is expanded here to test the effect of age on DNA methylation status. Methylation levels at approximately 27,000 CpG sites were assayed throughout the human genome in four brain regions: the frontal cortex, temporal cortex, pons and cerebellum tissues from 150 human donors. This was extended to include CpG methylation data generated in an additional 237 cerebellum and 237 frontal cortex samples. Results of these assays characterizing DNA methylation in the human brain of donors 0.4-102 years are presented in this chapter.

5.2 Materials and Methods

5.2.1 Tissue samples

For stage I analysis, fresh, frozen tissue samples of the frontal and temporal cortices, caudal pons and cerebellum regions were obtained from 150 neurologically normal Caucasian subjects, resulting in 600 tissue samples. For stage two analysis, fresh, frozen tissue samples of the frontal cortex and cerebellum regions were obtained from an additional 237 neurologically normal Caucasian subjects. Genomic DNA was phenol-chloroform extracted from brain tissues [309] and quantified on the Nanodrop1000 spectrophotometer prior to bisulfite conversion.

5.2.2 CpG Methylation

Bisulfite conversion of 1 microgram of genomic DNA was performed using Zymo EZ-96 DNA Methylation Kit per the manufacturers protocol. CpG methylation status of >27,000 sites was determined using Illumina Infinium HumanMethylation27 BeadChip, per the manufacturers protocol. Data were analyzed in BeadStudio software (Illumina Beadstudio v.3.0). The threshold call rate for inclusion of samples in the analysis was 95%. Quality control of sample handling included comparison of genders reported by the brain banks with the gender of the same samples determined by analyzing methylation levels of CpG sites on the X chromosome. Beta values were extracted from BeadStudio (Illumina, Inc) for sites on chromosome X and loaded into TM4

MeV tool. This data was then clustered by sample. Based on methylation levels for Chromosome X loci, these data split into two primary groups correlated with gender. Calls generated by this method were then compared with sample information reported by the brain bank. Samples where genders did not match between brain bank and methylation data were excluded from our analyses. Forty-seven tissue samples from subjects were excluded due to low methylation call rate or gender discrepancies.

5.2.3 CpG Methylation Analysis

For all available samples, stratified by brain region, multivariate linear regression was performed to test the effect of age on CpG methylation at each CpG site in the publicly available data. Regression models were adjusted for the following covariates: hybridization and amplification batch, study center responsible for sample collection, post-mortem interval and gender. Bonferroni correction of 1.8e-6 was used to account for the effects of multiple testing phenomenon after testing the associations of > 27,000 CpG sites per brain region in the stratified analyses (27,476 in pons, 27,310 in cerebellum, 27,532 in frontal cortex, and 27,538 in temporal cortex).

Any CpG site passing the Bonferroni thresholds for significance (1.8e-6) in all four brain regions was carried forward from the discovery phase of the project. Ten CpG sites that met these criteria and were analyzed using the same statistical models as implemented in the discovery phase, in an independent set of frontal cortex and cerebellum samples.

Post hoc, we categorized CpG sites as within or outside of CpG islands. This categorization was based upon annotation as a CpG island if the CpG site was described as an island in at least two resources out of three used for annotation: EPI score [310], UCSC genome browser sequence based annotation of CpG sites [311] or Illumina documentation. Non-island CpG sites were defined as sites not annotated as within an island in any of the three resources used for annotation.

5.2.4 DAVID analysis

Functional relationships were investigated using DAVID (http://david.abcc.ncifcrf.gov/). Enrichment of selected gene ontology (GO) terms among age-associated CpG sites was examined using the functional annotation clustering module. Six hundred and eighty three unique Entrez Gene identifiers in the David database were cross-referenced from the Illumina gene annotation for significantly associated CpG sites from our discovery analyses; where a CpG site passed Bonferroni correction in any brain region specific analysis. These 683 genes were considered our experimental pool in the clustering analysis.

To account for possible bias in the Illumina array design (i.e., bias introduced by the array being enriched for CpG sites nearby a certain functional class of gene), 14495 unique Entez Gene identifers were cross referenced between the entire Illumina CpG array annotation and the DAVID database, with this second gene set serving as the background level of enrichment for genes on

the array. Default settings were used for the derivation of clusters and false-discovery rates were used to correct for multiple testing. A total of 228 clusters were generated, with six clusters with enrichment scores showing a greater than four-fold enrichment of clustered terms.

5.2.5 Additional analyses

Replication was deemed successful if the association between age and methylation passed the Bonferroni threshold for significance of 1.8e-6 in analyses of both the frontal cortex and cerebellum datasets. Since the replication dataset included a significant number of individuals in the lower age ranges compared to the data used in the discovery phase, two additional iterations of the replication model were utilized to further scrutinize results, first by excluding all samples with age at sampling under 16 years, then excluding all samples under 18 years. Neither of these secondary models caused any marked attenuation of the p-values in the replication results. In addition, a fourth set of models using additional covariates of component vectors 1 and 2 from multidimensional scaling of genotype data for these samples did not significantly alter the results of the regression models.

5.3 Results

A series of experiments was performed to map the association of DNA methylation levels in human brain with chronological age. The first set of experiments included tissue from four brain regions: the frontal cortex, temporal cortex, pons and cerebellum from 150 individuals varying in age

from 16-101 years. The second (or replication) stage included tissue from the two most disparate regions of the brain, the frontal cortex and cerebellum regions and included 237 human brains collected from donors ranging from 0.4 to 102 years of age

5.3.1 Association between CpG methylation levels and chronological age across brain regions

Analysis of the association between chronological age and DNA methylation levels at individual CpG sites in the Stage 1 sample set revealed a considerable number of strongly associated loci (Figure 44).

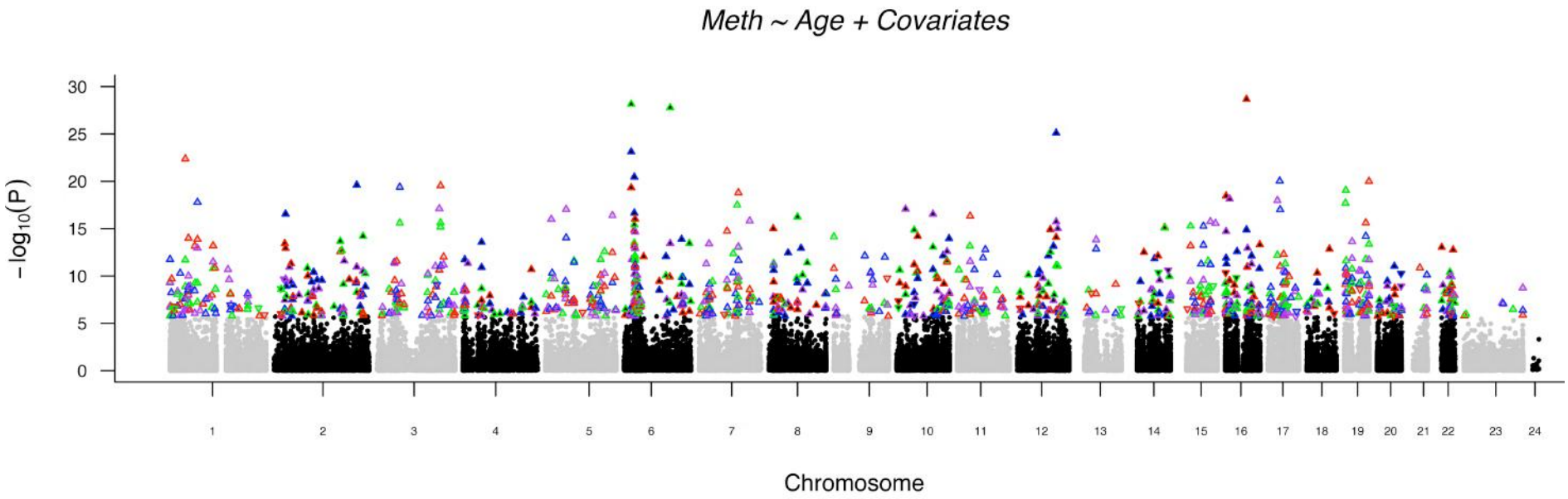


Figure 44: Manhattan plot representing >27,000 CpG sites/probes and their respective p-values associated with age.

Shown are data for cerebellum (purple), frontal cortex (green), pons (blue) and temporal cortex (red). For each point, a positive association between DNA methylation and chronological age is indicated by upward pointing triangles; a negative association is indicated by a downward pointing triangles

After applying correction for multiple testing, 1,141 significant associations between DNA methylation and age at CpG sites in the Stage 1 sample set were revealed. Of these 1141 associations, 589 loci were significant in only one region, 167 loci were significant in two regions, 86 loci were significant in three brain regions, and DNA methylation levels at ten CpG loci were significantly correlated with age in all four brain regions. Of the total number of significant CpG sites detected (1,141), 932 were considered to be within a strict definition of CpG islands, 129 were not within CpG islands, and 80 were in regions that could not unequivocally be defined as islands or non-islands.

The ten CpG sites that showed significant genome-wide association with chronological age across all four brain regions were further examined. An emphasis was placed on these ten loci as they were significant in all brain regions tested, suggesting these associations were genuine and that the potential confound of cellularity most likely did not play a roll. All ten of these significantly associated loci were located within CpG islands and the DNA methylation levels at these sites showed an increase with age across each of the four tissues Figure 45. The analysis of the independently ascertained stage 2 sample series confirmed the strong associations at all of these loci Table 7. It is important to note that the direction and magnitude of effect was consistent in both sample series.

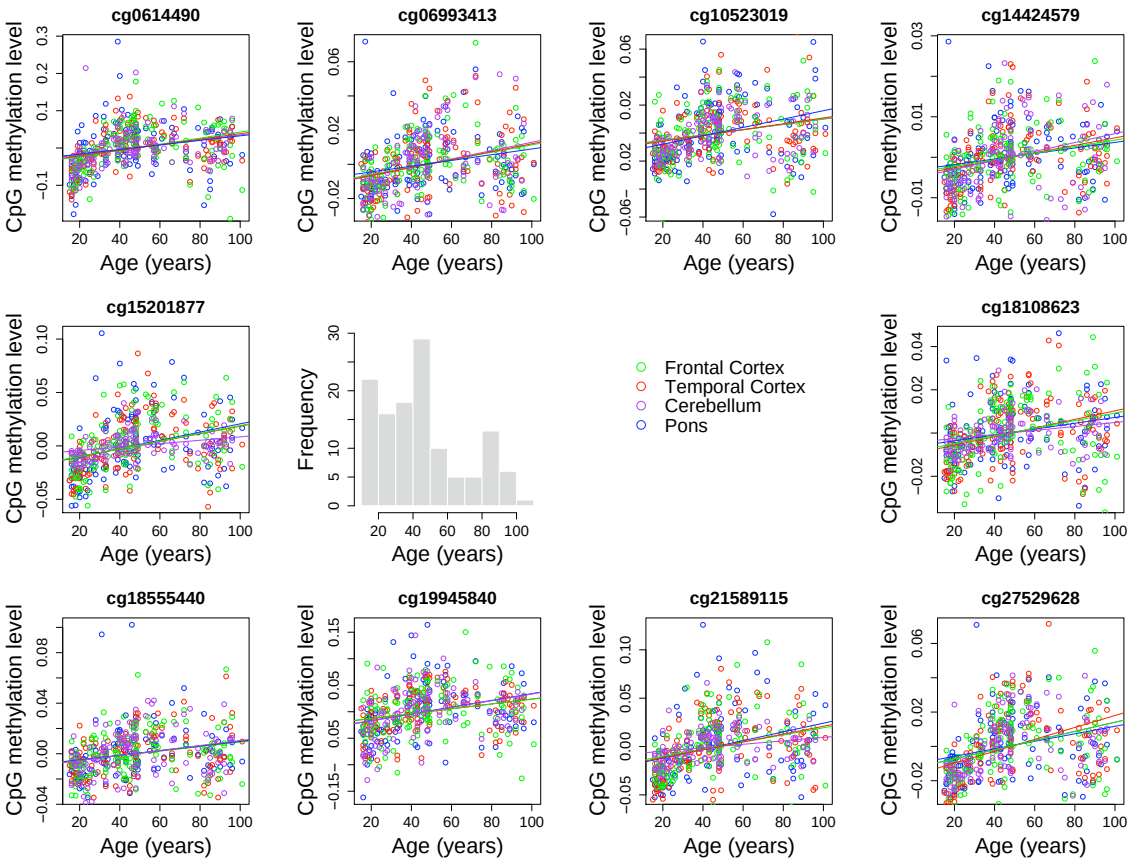


Figure 45: Covariate adjusted methylation levels brain.

Shown are data in cerebellum (purple), frontal cortex (green), pons (blue) and temporal cortex (red) for ten CpG sites where methylation levels increase significantly with age in all four brain regions (based on a Bonferroni threshold for significance of $p = 1.8 \times 10^{-6}$). Notably for all 10 loci that met our conservative threshold for significance, methylation levels were positively associated with age.

Table 7: Ten DNA methylation sites identified as significantly associated with chronological age in all tissues from stage I.

Notably DNA methylation level at each of these CpG sites is significantly and consistently associated with chronological age in stage II. CBLM – cerebellum; FCTX – frontal cortex; PONS – pons; TCTX – temporal cortex. Genomic position is based on hg18 of the human genome.

Name	Chr	Genomic position (bp)	Symbol	Dist. to TSS	Stage I p value				Stage II p value		Coefficient Range
					CBLM	FCTX	PONS	TCTX	CBLM	FCTX	
cg06144905	17	24393906	*PIPOX*	138	$7.3x10^{-7}$	$6.2x10^{-13}$	$3.4x10^{-7}$	$1.1x10^{-18}$	$5.8x10^{-8}$	$3.2x10^{-18}$	$[1.5x10^{-3},3.4x10^{-3}]$
cg06993413	15	63597257	*DPP8*	162	$2.7x10^{-10}$	$5.4x10^{-16}$	$9.0x10^{-8}$	$5.4x10^{-14}$	$5.5x10^{-21}$	$3.8x10^{-10}$	$[4.3x10^{-4},9.8x10^{-04}]$
cg10523019	2	227408702	*RHBDD1*	315	$7.7x10^{-7}$	$9.2x10^{-9}$	$6.1x10^{-15}$	$3.2x10^{-11}$	$2.0x10^{-16}$	$5.5x10^{-15}$	$[5.9x10^{-4},1.2x10^{-3}]$
cg14424579	2	27127813	*FLJ21839*	240	$5.9x10^{-7}$	$3.5x10^{-8}$	$3.2x10^{-7}$	$3.7x10^{-14}$	$2.5x10^{-15}$	$9.4x10^{-12}$	$[2.8x10^{-4},4.0x10^{-4}]$
cg15201877	1	71285561	*PTGER3*	518	$5.4x10^{-10}$	$1.3x10^{-14}$	$1.1x10^{-13}$	$1.6x10^{-18}$	$5.1x10^{-8}$	$8.5x10^{-26}$	$[3.4x10^{-4},1.6x10^{-3}]$
cg18108623	17	30725434	*FLJ34922*	701	$1.7x10^{-6}$	$9.3x10^{-12}$	$1.8x10^{-7}$	$9.6x10^{-18}$	$3.1x10^{-8}$	$1.2x10^{-22}$	$[3.0x10^{-4},8.4x10^{-4}]$
cg18555440	11	17698263	*MYOD1*	528	$1.8x10^{-6}$	$2.2x10^{-8}$	$3.6x10^{-7}$	$2.6x10^{-10}$	$1.0x10^{-26}$	$6.0x10^{-11}$	$[5.7x10^{-4},8.0x10^{-4}]$
cg19945840	1	1157899	*B3GALT6*	391	$4.2x10^{-10}$	$1.7x10^{-6}$	$1.8x10^{-12}$	$5.0x10^{-10}$	$6.9x10^{-20}$	$2.3x10^{-17}$	$[1.9x10^{-3},2.6x10^{-3}]$
cg21589115	19	54558926	*DKKL1*	72	$1.2x10^{-6}$	$1.7x10^{-12}$	$6.2x10^{-15}$	$2.4x10^{-16}$	$5.1x10^{-13}$	$3.2x10^{-26}$	$[5.8x10^{-4},1.8x10^{-3}]$
cg27529628	12	99491350	*GAS2L3*	270	$8.1x10^{-15}$	$1.8x10^{-16}$	$8.2x10^{-12}$	$7.8x10^{-26}$	2.3x10-^{27}	$3.8x10^{-26}$	$[7.1x10^{-4},1.5x10^{-3}]$

5.3.2 Substantial enrichment of CpG Methylation sites positively correlated with chronological age

The preliminary assessment of the 10 loci where DNA methylation was associated with chronological age in all four brain regions revealed that each of the associations was in the positive direction, increasing DNA methylation. Further, the majority of all the significantly associated loci in each tissue showed that this positive association was the trend. This is illustrated in Figure 44, where positive correlations between chronological age and DNA methylation appear to be in tremendous excess and are indicated by upward pointing triangles. 95.4% of significant results passing Bonferroni correction in stage I showed a positive correlation with age, whereas only 56.0% of non-significant results had positive regression coefficients, illustrating that this consistent direction of effect far exceeds chance (Z-statistic = 26.7, p-value < 0.0001). This enrichment of positive associations was also seen in the replication dataset with 78.6% of significant associations having a positive direction of effect, and 55.1% of non-significant results having a positive direction of effect (Z-statistic = 20.4, p-value < 0.0001).

It was recently shown by Christensen and colleagues that (among solid tissues studied, including the brain) the direction of correlation between age and methylation differs upon whether the CpG site is located within an island [312], observing that loci within CpG islands showed significant increases in

methylation with advancing age while CpG sites located outside of islands showed significant losses of methylation with aging. In the present analysis the regression coefficients from the stage I data showed an excess of CpG sites where DNA methylation positively correlated with age within islands compared to those sites outside of CpG islands. Of the age associated sites within CpG islands, the correlation between DNA methylation and chronological age was positive in more than 98% of sites. Conversely, a substantially lower proportion of associated sites outside of CpG islands showed a positive correlation between DNA methylation levels and age (Figure 46).

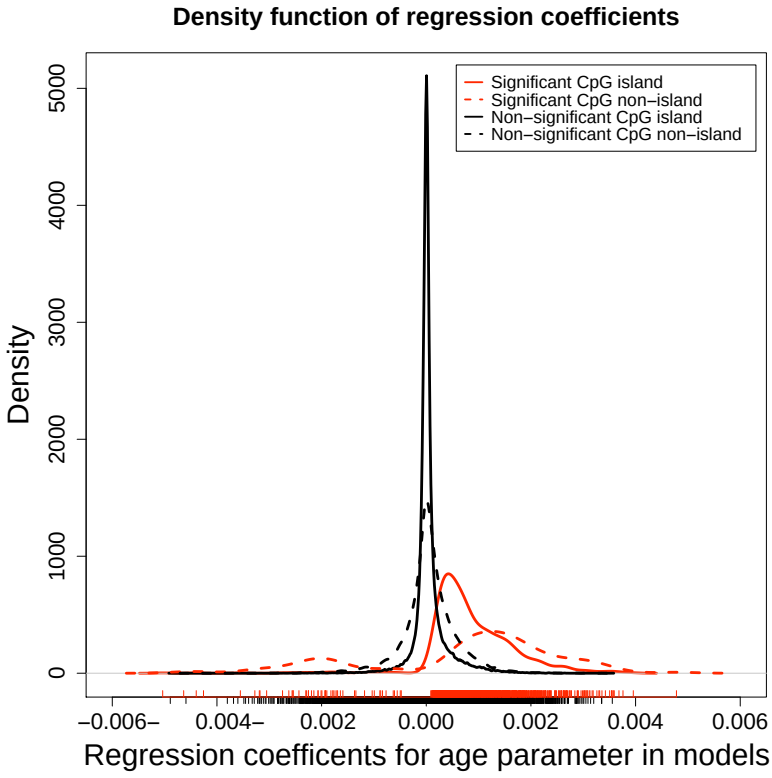


Figure 46: Significant results show an excess of CpG sites positively associated with age.

This relationship is particularly pronounced at CpG sites within canonical islands (red solid line). Both positive and negative correlation between CpG sites and chronological age exist at non-island CpG sites.

Due to the potential confounding effect of unreliable designation into island/non-island status, we repeated this analysis using a more restrictive definition of CpG sites inside and outside of islands, requiring that a CpG site meet the criteria for being located within an island in all three resources used for annotation: EP score [282], UCSC genome browser sequence-based annotation of CpG sites [178] and Illumina documentation. Restricting the definition did not change the excess of positive correlations within islands versus non-islands (Figure 47).

These data are supported by previous work performed in human blood [313]. Our analysis illustrates that those sites where DNA methylation was negatively correlated with age are sixteen times more likely to be located outside of a CpG island versus within a CpG island.

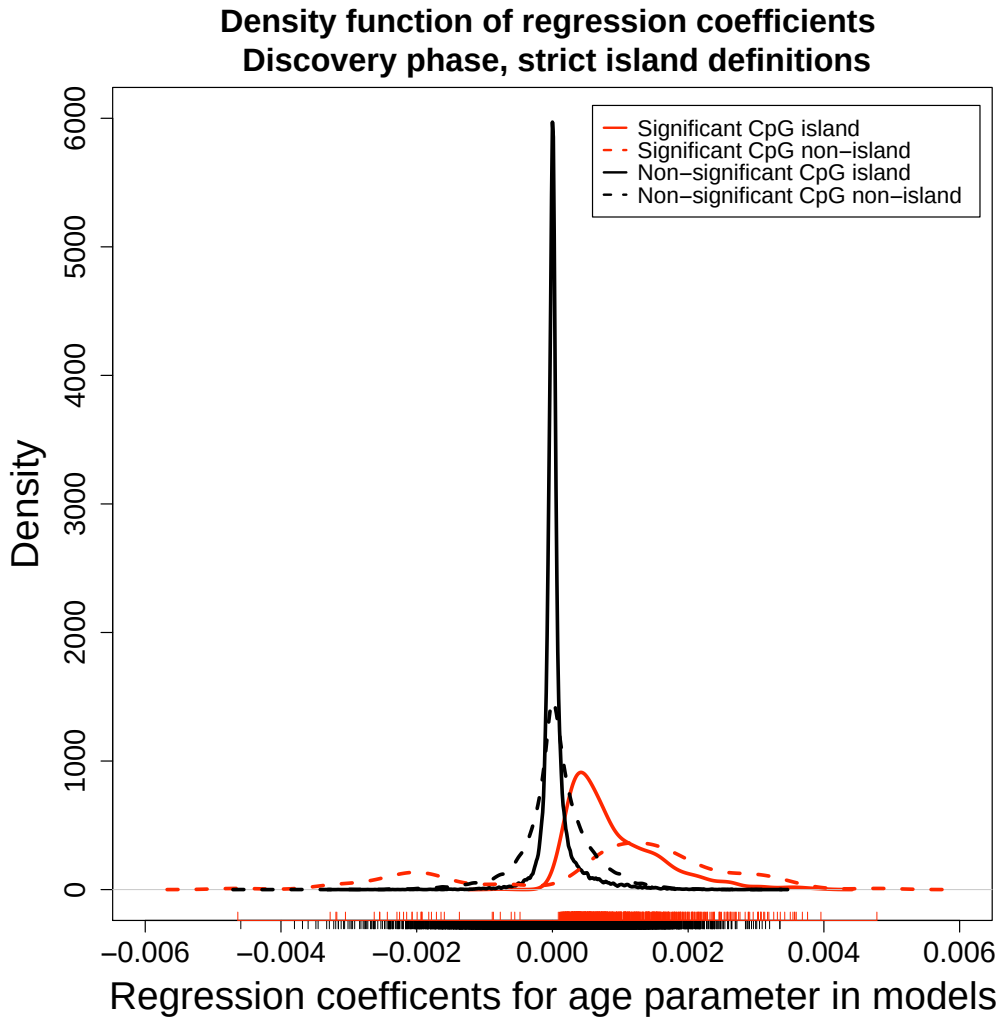


Figure 47: Analysis of the regression coefficients from stage 1 data continue to show the excess of positive correlations within islands versus non-islands using a more restrictive definition of sites inside and outside of islands.

5.3.3 Comparison of age-related CpG methylation changes across brain regions

In order to determine whether the associations found between DNA methylation at individual CpG sites and chronological age were consistent across brain regions, a comparison of association *p*-values was performed across cerebellum, frontal cortex, pons, and temporal cortex datasets. Individual CpG sites where DNA methylation level showed a significant association with chronological age in at least one of the four tissues were included in the analysis, revealing that age associated CpG sites are most similar in frontal cortex and temporal cortex and that these two tissues are in turn quite similar to pons (Figure 48). In contrast, the pattern of age associated CpG sites observed in the cerebellum was by far the most distinct of the four regions tested.

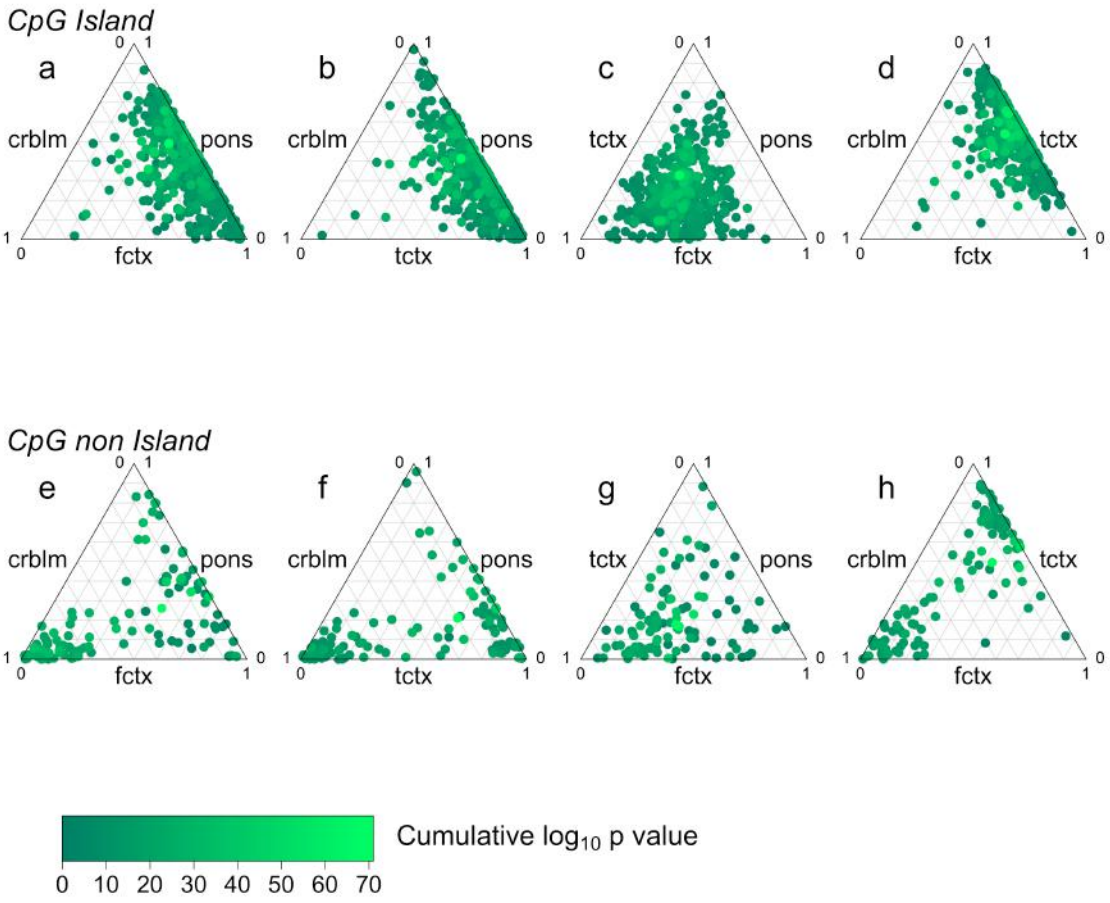


Figure 48: Ternary plots showing concordance of combined p values from phase I analyses across all four brain regions stratified by CpG island or non-island status.

The number and identity of samples in stage I were marginally different between the four tissue regions tested due to occasional sample or assay failure. To ensure the observed differences were not a result of power, a comparison was performed among age-associated CpG sites across tissues on a subset of donors from stage I for whom data on each of the four tissues were available (n = 84). These analyses revealed that uniqueness of associations in cerebellar tissue

was not due to sampling bias. The relative similarity between the frontal cortex and temporal cortex tissues remained (Figure 49).

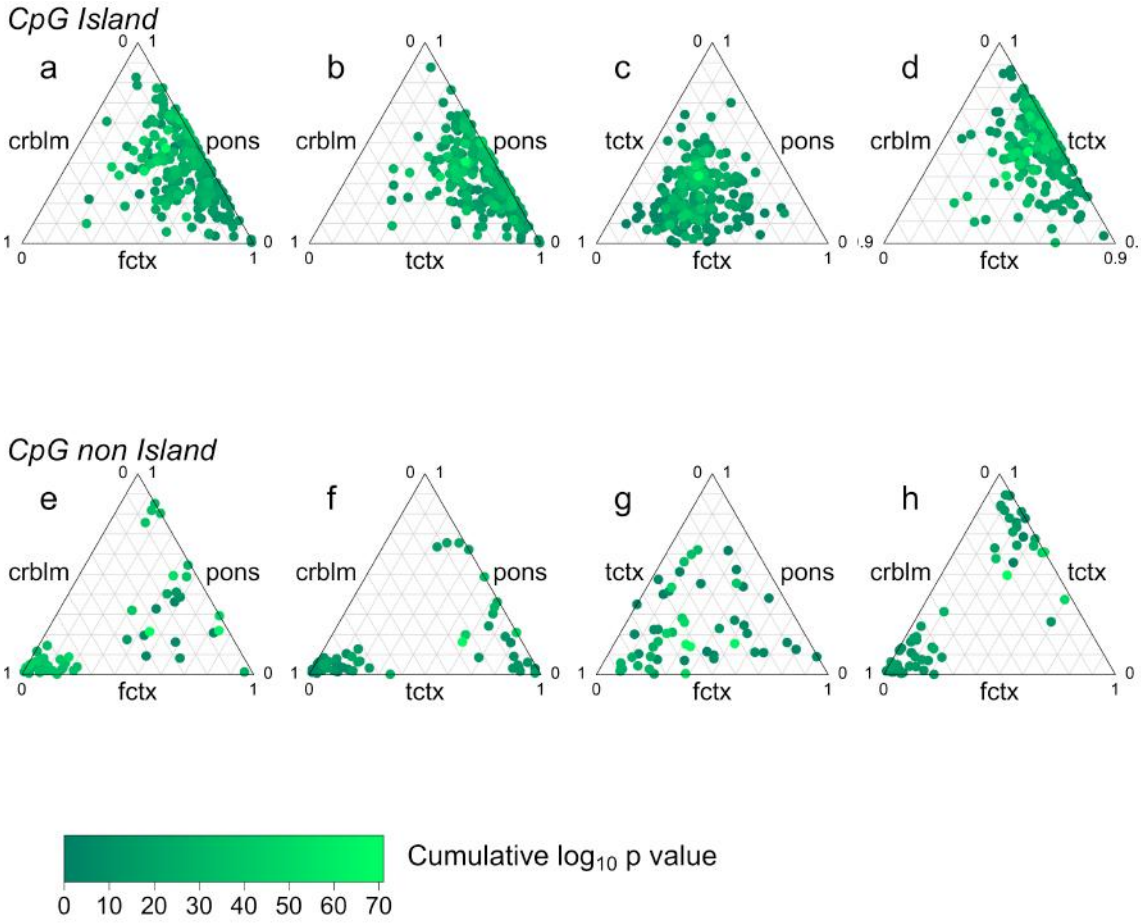


Figure 49: Analyses on a subset of donors from stage 1 for whom data on each of the four tissues were available revealed the uniqueness of associations in cerebellar tissue was not due to sampling bias

Next, the analysis was expanded to include data derived from the additional frontal cortex and cerebellar samples typed in stage II (Figure 14). Again significant associations occurred in both the frontal and cerebellar datasets (as well as in all 4 regions). As before, the most associated methylation sites (with

the highest p-values) displayed comparatively similar methylation levels in both tissues. These findings are consistent with previous reports from our group and others showing that patterns of both DNA methylation and expression are quite different in cerebellum compared to other brain tissues [279].

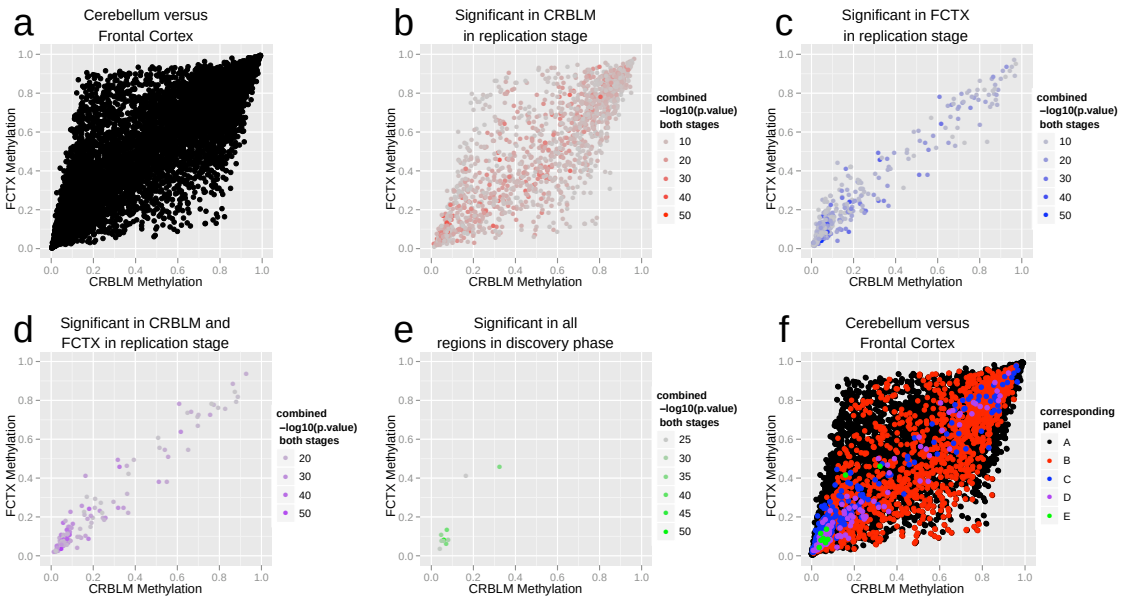


Figure 50: Data from stage II comparing results from cerebellum and frontal cortex.

These data show that significant results across two or more tissues often occur at similar methylation levels across various tissues, seemingly robust to the magnitude of methylation.

5.3.4 Gene ontology/functional annotation analysis

In order to provide insight into the biological function of age-associated CpG sites and to determine whether genes within close proximity to these sites were functionally similar, the Database for Annotation, Visualization and Integrated Discovery (DAVID) (http://david.abcc.ncifcrf.gov/) was use to investigate for enrichment of gene ontology (GO) terms. Six hundred and eighty three unique EntrezGene identifiers were cross-referenced in the DAVID database, using Illumina gene annotation for significantly associated CpG sites from our initial stage 1 analysis. These 683 were considered our experimental pool in the clustering analysis. A total of 228 clusters were generated. Six clusters with the highest degrees of enrichment are shown in Figure 51. These clusters illustrate a strong enrichment for genes related to DNA binding, morphogenesis and regulation of transcription.

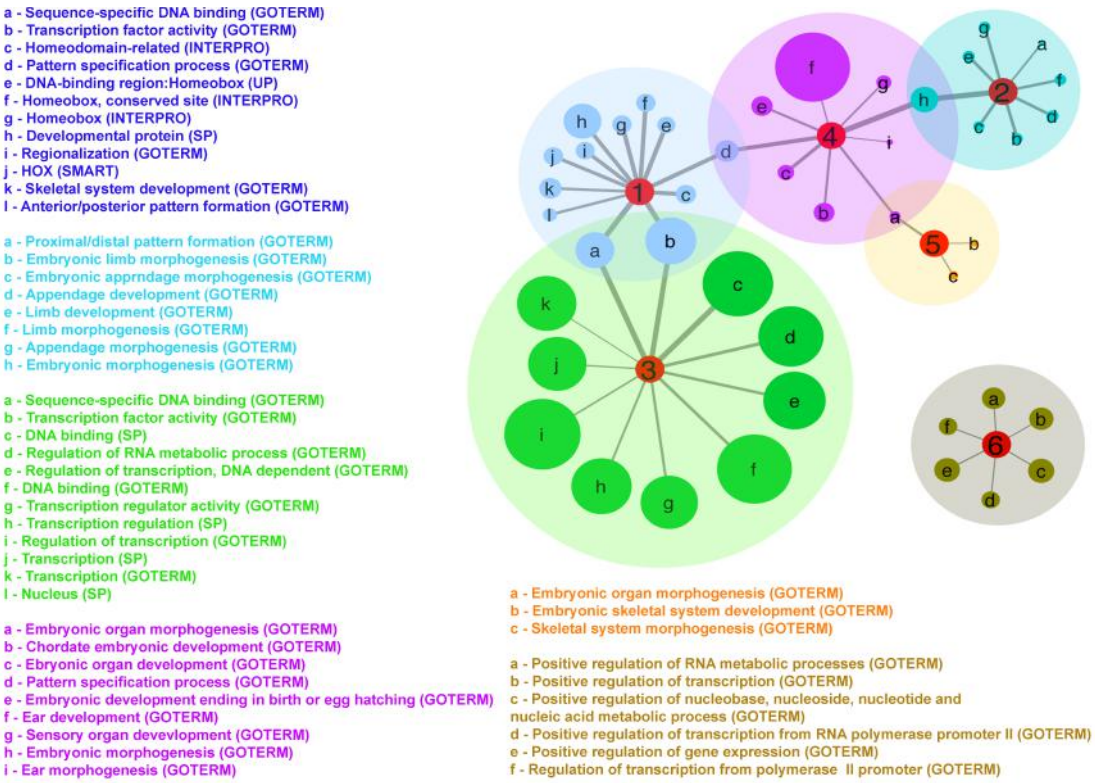


Figure 51: A map of enriched functional clusters for genes proximal to age associated CpG sites based on DAVID functional annotation clustering.

This figure shows the inferred functional relatedness of clustered genes at a greater than four-fold level of enrichment among significant results in phase I of this study. Line thickness connecting nodes indicates relative p-value of the term within the cluster, with the thickest line representing the most significant term (p-value = 1.6x10-19) and the thinnest lines representing the least significant term (p-value = 2.5x10-5).

5.4 Discussion

In an attempt to understand and map DNA methylation levels in normal brain with chronological age, genome-wide analysis of DNA methylation across four distinct brain regions was performed in two stages, a discovery stage in four brain regions and a replication in two brain regions.

CpG sites were shown to exhibit strong age-associated changes in DNA methylation. Many of the significantly associated CpG sites were found in multiple tissues and occurred at higher frequencies than expected by chance. Assessing DNA methylation levels in frozen brain tissue samples and attempting to associate these levels with chronological age presents a unique confound in that brain tissue has a dynamic cellular composition. Of note, is that the proportion of neurons to glia may change with chronological aging. However, we have identified consistent results across multiple brain regions and therefore, it is not likely to be a confounding point among the 10 CpG sites showing significant genome-wide association with chronological age across all four brain regions.

Further, four of the ten loci identified in all four brain regions of the analyses were shown previously to be associated with age-related methylation changes in pleura and blood [312, 313]. Age-related increases in methylation levels for discrete CpG loci was also shown in DNA isolated by fluorescence activated cell sorting (FACS) from cerebral cortex neuronal nuclei of 125 subjects [314].

Although this particular study was not a genome-wide examination of CpG loci, it did show similar results of increasing DNA methylation in the brain correlated with age. This study examined 50 CpG sites in samples consisting solely of neuronal DNA, thus removing the possible confound of non-neuronal cells contributing to the signal. Together, these data provide evidence that the age-related CpG loci detected here are not artifacts of age-associated alterations in brain tissue cellularity, but indicate a biological change in DNA methylation state.

In the present study, an enrichment of age-associated methylation changes at CpG islands of functionally related transcripts was shown. DNA binding factors and transcription factors were identified as classes of genes linked with age-associated CpG sites. The clustering of age-associated CpG methylation sites proximal to genes associated with DNA binding and predominantly transcription with homeobox proteins suggests that age-related alterations in methylation might be important for the maintenance of transcriptional programs in aging tissues. It is the tendency for gene expression to show higher variance as organisms age, versus a linear association with aging [315]. Therefore, it is an interesting possibility that the increase of DNA methylation may be important in maintenance of consistent gene expression patterns with age and conceivably relevant to the study of late onset neurodegenerative disease mechanism.

6 General Summary and Conclusions

A principal reason for performing genetic analysis of disease is to use the insights provided by genetics to understand the biological processes underlying disease etiology. Disease genetics has followed a route for a number of years, discovering genetic mutations that cause disease and subsequently using those mutations to model and understand the disease process in cells and animals. While progress on this path has been challenging, there have been successes and science has benefited from this research paradigm. However, as new approaches to understanding genetic influences in disease emerge, such as genetic risk factors, the requirements for a novel means to study the biological basis of these new genetic influences must be established.

To this end, the work in this thesis illustrates the beginning of one new path using genetics to understand disease. This research focused on using modern genomic and epigenomic approaches to generate a dataset mapping the influence of common genetic variability on CpG methylation levels in human brain and investigating trends in CpG methylation levels associated with chronological age in normal human brain.

Genome wide association studies have provided information on a new form of genetic influence in disease and now require a novel set of methods to effectively leverage this information. Here, GWA has been performed in a large series of PD

cases and controls, resulting in the identification of 11 novel risk loci for this disease. This work has provided fundamental knowledge regarding the genetic architecture of PD. It has shown that the common disease common variant hypothesis is testable and true in PD, revealing that genes containing dominant PD causing mutations such as *LRRK2* and *SNCA,* also contain non-coding variability that confers risk for the disease in the general PD population. This observation suggests that similar mechanisms leading to disease in mutation carriers also occur in genetically complex PD. This effort has also delivered a large number of genes putatively involved in risk for PD. As with other GWA, it is extremely challenging to distinguish the key gene within a locus that may be altered and operating in disease pathogenesis.

In an attempt to begin to address this obstacle, this thesis describes work that centered on creating a reference set of data, linking common genetic variation with DNA methylation levels in the human brain. These data showed that methylation levels at a large number of CpG sites are associated with genotype, and further that when the two were correlated, they also tended to be physically close and generalizable across brain regions.

Integrating the PD GWA with dmQTL data revealed that the majority of PD risk alleles were also dmQTLs for proximal CpG sites. These results suggest the investigation of dmQTL could be a fruitful area of initial investigation on the path toward understanding disease etiology. Most striking of the dmQTLs presented, were the dmQTLs observed for the two independent risk alleles at *SNCA*.

Despite being ~55kb apart, both risk alleles influence DNA methylation at the same CpG site within the promoter region of *SNCA*. Notably, both risk alleles influence CpG methylation at this site in the same direction (increasing risk being associated with increasing methylation).

The last portion of my thesis describes work that aims to identify whether there are age-associated changes in DNA methylation in the human brain. This effort represents the most comprehensive analysis of CpG methylation levels and chronological age in neurological tissue performed to date and provides insight into coordinated changes in DNA methylation during aging providing a starting point for understanding the underlying mechanisms of aging. This work showed consistent and highly significant changes in DNA methylation at 10 loci across all brain regions. A portion of these age-associated methylation loci had been previously reported in other tissues. Therefore, it is an interesting possibility that predictable changes in DNA methylation may be important in maintenance of consistent gene expression patterns with age.

There are many directions in which this research can be taken further. In the context of the discovery of genetic risk factors, much is already being done. During the time of performing this thesis work the number of loci identified for PD went from 5 to 28 and it is likely that ongoing analyses using existing GWA data will identify additional loci. Second generation sequencing will be a factor in the identification of new PD linked genes, and this will likely impact in the field through resequencing of candidate loci (including GWA loci), exome sequencing,

and ultimately whole genome sequencing. There is also room to apply each of these techniques beyond simple disease association, analyzing other key phenotypes such as disease onset, disease progression, disease course, response to medication, and the appearance of other symptoms (such as depression, and dementia).

Many challenges remain in our endeavor to translate genetics to an understanding of disease pathogenesis. Possibly the most difficult of these challenges today is understanding the biological effects of risk alleles both in the context of which gene is mediating risk, and in what way that gene is altered. The current work examining DNA methylation provides some insight; however, there are limitations of these data. While this work is the most comprehensive in sample size and number of CpG sites examined in brain tissue, there are still a large number of CpG sites that remain to be assessed. Examining the complete methylation signature across each of the associated PD risk loci could provide valuable information. Further, this work examines tissue of mixed cellular composition; therefore, cell specific changes are difficult to observe. The application of more comprehensive sequence based methods, particularly in defined cell populations, would address some of these issues. However, the application (for example) of DNA methylation sequencing in a series of iPS cells differentiated into dopaminergic neurons is currently time and cost prohibitive. Enhancing our work by expanding the brain sample series and/or assessing a more comprehensive list of CpG sites will be useful; however, significant

information could likely be gained by integrating other forms of data. Total RNA sequencing data in the same series of brains described here is currently being generated. Given that DNA methylation can be associated with gene expression levels, transcript expression levels, splicing, and UTR usage, RNA sequencing will add a valuable set of data to the current work. Investigation into many of the identified loci using other genetic and functional efforts will be performed over the next 2-3 years. These efforts include resequencing of the PD risk loci (which will help in gene identification through the identification of rare coding mutations) and mapping of transcriptional regulatory elements within cells of interest at these specific loci.

Lastly, the work examining the association between chronological age and DNA methylation is currently being followed up within a longitudinal series of blood samples collected at 3 time points in an epidemiological cohort (InCHIANTI study: www.inchiantistudy.net/). Undoubtedly, as the relationship between aging and methylation is better defined, there will be an opportunity to integrate these epigenetic changes (and those within PD loci in particular) into our models of PD risk and pathogenesis.

7 Bibliography

1. Jonsson, T., et al., *Variant of TREM2 associated with the risk of Alzheimer's disease.* N Engl J Med, 2013. **368**(2): p. 107-16.

2. Guerreiro, R., et al., *TREM2 variants in Alzheimer's disease.* N Engl J Med, 2013. **368**(2): p. 117-27.

3. Risch, N. and K. Merikangas, *The future of genetic studies of complex human diseases.* Science, 1996. **273**(5281): p. 1516-7.

4. Reich, D.E. and E.S. Lander, *On the allelic spectrum of human disease.* Trends Genet, 2001. **17**(9): p. 502-10.

5. Pritchard, J.K. and N.J. Cox, *The allelic architecture of human disease genes: common disease-common variant...or not?* Hum Mol Genet, 2002. **11**(20): p. 2417-23.

6. Brookfield, J.F., *Q&A: promise and pitfalls of genome-wide association studies.* BMC Biol, 2010. **8**: p. 41.

7. International HapMap, C., *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.

8. International HapMap, C., et al., *Integrating common and rare genetic variation in diverse human populations.* Nature, 2010. **467**(7311): p. 52-8.

9. Norrgard, K. *Genetic variation and disease: GWAS*. Nature Educatio, 2008. **1**, 87.

10. Li, Y., et al., *MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes.* Genet Epidemiol, 2010. **34**(21058334): p. 816-834.

11. Hinds, D.A., et al., *Whole-genome patterns of common DNA variation in three human populations.* Science, 2005. **307**(5712): p. 1072-9.

12. Peiffer, D.A. and K.L. Gunderson, *Design of tag SNP whole genome genotyping arrays.* Methods Mol Biol, 2009. **529**: p. 51-61.

13. Gibbs, J.R. and A. Singleton, *Application of genome-wide single nucleotide polymorphism typing: simple association and beyond.* PLoS Genet, 2006. **2**(10): p. e150.

14. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nat Genet, 2006. **38**(8): p. 904-9.

15. Genomes Project, C., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.

16. Genomes Project, C., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.

17. Manchia, M., et al., *The impact of phenotypic and genetic heterogeneity on results of genome wide association studies of complex diseases.* PLoS One, 2013. **8**(10): p. e76295.

18. E. C. Capen, R.V.C., W. M. Campbell, *Competitive Bidding in High-Risk Situations.* Journal of Petroleum Technology, 1971. **23**(6): p. 641-653.

19. Hindorff, L.A., et al., *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.* Proc Natl Acad Sci U S A, 2009. **106**(23): p. 9362-7.

20. Bower, J.H., et al., *Incidence and distribution of parkinsonism in Olmsted County, Minnesota, 1976-1990.* Neurology, 1999. **52**(6): p. 1214-20.

21. Kempster, P.A., B. Hurwitz, and A.J. Lees, *A new look at James Parkinson's Essay on the Shaking Palsy.* Neurology, 2007. **69**(5): p. 482-5.

22. Hindle, J.V., *Ageing, neurodegeneration and Parkinson's disease.* Age Ageing, 2010. **39**(2): p. 156-61.

23. de Rijk, M.C., et al., *Prevalence of Parkinson's disease in the elderly: the Rotterdam Study.* Neurology, 1995. **45**(12): p. 2143-6.

24. Jankovic, J., *Parkinson's disease: clinical features and diagnosis.* J Neurol Neurosurg Psychiatry, 2008. **79**(4): p. 368-76.

25. Corti, O., S. Lesage, and A. Brice, *What genetics tells us about the causes and mechanisms of Parkinson's disease.* Physiol Rev, 2011. **91**(4): p. 1161-218.

26. Kumar, K.R., A. Djarmati-Westenberger, and A. Grunewald, *Genetics of Parkinson's disease.* Semin Neurol, 2011. **31**(5): p. 433-40.

27. Forno, L.S., *Neuropathology of Parkinson's disease.* J Neuropathol Exp Neurol, 1996. **55**(3): p. 259-72.

28. Orr, C.F., D.B. Rowe, and G.M. Halliday, *An inflammatory review of Parkinson's disease.* Prog Neurobiol, 2002. **68**(5): p. 325-40.

29. Spillantini, M.G., et al., *Filamentous alpha-synuclein inclusions link multiple system atrophy with Parkinson's disease and dementia with Lewy bodies.* Neurosci Lett, 1998. **251**(3): p. 205-8.

30. Polymeropoulos, M.H., et al., *Linkage of the locus for cerebral cavernous hemangiomas to human chromosome 7q in four families of Mexican-American descent.* Neurology, 1997. **48**(3): p. 752-7.

31. Polymeropoulos, M.H., et al., *Mutation in the alpha-synuclein gene identified in families with Parkinson's disease.* Science, 1997. **276**(5321): p. 2045-7.

32. Farrer, M., et al., *Low frequency of alpha-synuclein mutations in familial Parkinson's disease.* Annals of neurology, 1998. **43**(3): p. 394-7.

33. Kitada, T., et al., *Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism.* Nature, 1998. **392**(6676): p. 605-8.

34. Munoz, E., et al., *Identification of Spanish familial Parkinson's disease and screening for the Ala53Thr mutation of the alpha-synuclein gene in early onset patients.* Neurosci Lett, 1997. **235**(1-2): p. 57-60.

35. Scott, W.K., et al., *Genetic complexity and Parkinson's disease. Deane Laboratory Parkinson Disease Research Group.* Science, 1997. **277**(5324): p. 387-8; author reply 389.

36. Scott, W.K., et al., *The alpha-synuclein gene is not a major risk factor in familial Parkinson disease.* Neurogenetics, 1999. **2**(3): p. 191-2.

37. Athanassiadou, A., et al., *Genetic analysis of families with Parkinson disease that carry the Ala53Thr mutation in the gene encoding alpha-synuclein.* Am J Hum Genet, 1999. **65**(2): p. 555-8.

38. Spira, P.J., et al., *Clinical and pathological features of a Parkinsonian syndrome in a family with an Ala53Thr alpha-synuclein mutation.* Annals of neurology, 2001. **49**(3): p. 313-9.

39. Choi, J.M., et al., *Analysis of PARK genes in a Korean cohort of early-onset Parkinson disease.* Neurogenetics, 2008. **9**(4): p. 263-9.

40. Ki, C.S., et al., *The Ala53Thr mutation in the alpha-synuclein gene in a Korean family with Parkinson disease.* Clin Genet, 2007. **71**(5): p. 471-3.

41. Puschmann, A., et al., *A Swedish family with de novo alpha-synuclein A53T mutation: evidence for early cortical dysfunction.* Parkinsonism Relat Disord, 2009. **15**(9): p. 627-32.

42. Spillantini, M.G., et al., *Alpha-synuclein in Lewy bodies.* Nature, 1997. **388**(6645): p. 839-40.

43. Klein, C. and M.G. Schlossmacher, *The genetics of Parkinson disease: Implications for neurological care.* Nat Clin Pract Neurol, 2006. **2**(3): p. 136-46.

44. Kruger, R., et al., *Ala30Pro mutation in the gene encoding alpha-synuclein in Parkinson's disease.* Nat Genet, 1998. **18**(2): p. 106-8.

45. Zarranz, J.J., et al., *The new mutation, E46K, of alpha-synuclein causes Parkinson and Lewy body dementia.* Annals of neurology, 2004. **55**(2): p. 164-73.

46. Farrer, M., et al., *Comparison of kindreds with parkinsonism and alpha-synuclein genomic multiplications.* Annals of neurology, 2004. **55**(2): p. 174-9.

47. Ibanez, P., et al., *Alpha-synuclein gene rearrangements in dominantly inherited parkinsonism: frequency, phenotype, and mechanisms.* Arch Neurol, 2009. **66**(1): p. 102-8.

48. Singleton, A.B., et al., *alpha-Synuclein locus triplication causes Parkinson's disease.* Science, 2003. **302**(5646): p. 841.

49. Chartier-Harlin, M.C., et al., *Alpha-synuclein locus duplication as a cause of familial Parkinson's disease.* Lancet, 2004. **364**(9440): p. 1167-9.

50. Ibanez, P., et al., *Causal relation between alpha-synuclein gene duplication and familial Parkinson's disease.* Lancet, 2004. **364**(9440): p. 1169-71.

51. Nishioka, K., et al., *Clinical heterogeneity of alpha-synuclein gene duplication in Parkinson's disease.* Annals of neurology, 2006. **59**(2): p. 298-309.

52. Fuchs, J., et al., *Phenotypic variation in a large Swedish pedigree due to SNCA duplication and triplication.* Neurology, 2007. **68**(12): p. 916-22.

53. Uchiyama, T., et al., *Prominent psychiatric symptoms and glucose hypometabolism in a family with a SNCA duplication.* Neurology, 2008. **71**(16): p. 1289-91.

54. Troiano, A.R., et al., *Re: Alpha-synuclein gene duplication is present in sporadic Parkinson disease.* Neurology, 2008. **71**(16): p. 1295; author reply 1295.

55. Ikeuchi, T., et al., *Patients homozygous and heterozygous for SNCA duplication in a family with parkinsonism and dementia.* Arch Neurol, 2008. **65**(4): p. 514-9.

56. Papapetropoulos, S., et al., *Clinical phenotype in patients with alpha-synuclein Parkinson's disease living in Greece in comparison with patients with sporadic Parkinson's disease.* J Neurol Neurosurg Psychiatry, 2001. **70**(5): p. 662-5.

57. Ross, O.A., et al., *Genomic investigation of alpha-synuclein multiplication and parkinsonism.* Annals of neurology, 2008. **63**(6): p. 743-50.

58. Singleton, A. and K. Gwinn-Hardy, *Parkinson's disease and dementia with Lewy bodies: a difference in dose?* Lancet, 2004. **364**(9440): p. 1105-7.

59. Funayama, M., et al., *A new locus for Parkinson's disease (PARK8) maps to chromosome 12p11.2-q13.1.* Ann Neurol, 2002. **51**(3): p. 296-301.

60. Zimprich, A., et al., *Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology.* Neuron, 2004. **44**(4): p. 601-7.

61. Paisan-Ruiz, C., et al., *Cloning of the gene containing mutations that cause PARK8-linked Parkinson's disease.* Neuron, 2004. **44**(4): p. 595-600.

62. Rubio, J.P., et al., *Deep sequencing of the LRRK2 gene in 14,002 individuals reveals evidence of purifying selection and independent origin of the p.Arg1628Pro mutation in Europe.* Hum Mutat, 2012. **33**(7): p. 1087-98.

63. Aasly, J.O., et al., *Novel pathogenic LRRK2 p.Asn1437His substitution in familial Parkinson's disease.* Mov Disord, 2010. **25**(13): p. 2156-63.

64. Bardien, S., et al., *Genetic characteristics of leucine-rich repeat kinase 2 (LRRK2) associated Parkinson's disease.* Parkinsonism Relat Disord, 2011. **17**(7): p. 501-8.

65. Healy, D.G., et al., *Phenotype, genotype, and worldwide genetic penetrance of LRRK2-associated Parkinson's disease: a case-control study.* Lancet Neurol, 2008. **7**(7): p. 583-90.

66. Lorenzo-Betancor, O., et al., *LRRK2 haplotype-sharing analysis in Parkinson's disease reveals a novel p.S1761R mutation.* Mov Disord, 2012. **27**(1): p. 146-51.

67. Nuytemans, K., et al., *Genetic etiology of Parkinson disease associated with mutations in the SNCA, PARK2, PINK1, PARK7, and LRRK2 genes: a mutation update.* Hum Mutat, 2010. **31**(7): p. 763-80.

68. Correia Guedes, L., et al., *Worldwide frequency of G2019S LRRK2 mutation in Parkinson's disease: a systematic review.* Parkinsonism Relat Disord, 2010. **16**(4): p. 237-42.

69. Brice, A., *Genetics of Parkinson's disease: LRRK2 on the rise.* Brain, 2005. **128**(Pt 12): p. 2760-2.

70. Farrer, M., et al., *LRRK2 mutations in Parkinson disease.* Neurology, 2005. **65**(5): p. 738-40.

71. Kachergus, J., et al., *Identification of a novel LRRK2 mutation linked to autosomal dominant parkinsonism: evidence of a common founder across European populations.* Am J Hum Genet, 2005. **76**(4): p. 672-80.

72. Lesage, S., A. Durr, and A. Brice, *[LRRK2 is a major gene in North African parkinsonism].* Med Sci (Paris), 2006. **22**(5): p. 470-1.

73. Lesage, S., et al., *LRRK2 G2019S as a cause of Parkinson's disease in North African Arabs.* N Engl J Med, 2006. **354**(4): p. 422-3.

74. Ozelius, L.J., et al., *LRRK2 G2019S as a cause of Parkinson's disease in Ashkenazi Jews.* N Engl J Med, 2006. **354**(4): p. 424-5.

75. Johnson, J., et al., *Comprehensive screening of a North American Parkinson's disease cohort for LRRK2 mutation.* Neurodegener Dis, 2007. **4**(5): p. 386-91.

76. Khan, N.L., et al., *Mutations in the gene LRRK2 encoding dardarin (PARK8) cause familial Parkinson's disease: clinical, pathological, olfactory and functional imaging and genetic data.* Brain, 2005. **128**(Pt 12): p. 2786-96.

77. Lesage, S., et al., *Molecular analyses of the LRRK2 gene in European and North African autosomal dominant Parkinson's disease.* J Med Genet, 2009. **46**(7): p. 458-64.

78. Paisan-Ruiz, C., et al., *Comprehensive analysis of LRRK2 in publicly available Parkinson's disease cases and neurologically normal controls.* Hum Mutat, 2008. **29**(4): p. 485-90.

79. Haugarvoll, K. and Z.K. Wszolek, *Clinical features of LRRK2 parkinsonism.* Parkinsonism Relat Disord, 2009. **15 Suppl 3**: p. S205-8.

80. Li, Y., et al., *The R1441C mutation alters the folding properties of the ROC domain of LRRK2.* Biochim Biophys Acta, 2009. **1792**(12): p. 1194-7.

81. Greggio, E., et al., *The Parkinson's disease kinase LRRK2 autophosphorylates its GTPase domain at multiple sites.* Biochem Biophys Res Commun, 2009. **389**(3): p. 449-54.

82. Vitte, J., et al., *Leucine-rich repeat kinase 2 is associated with the endoplasmic reticulum in dopaminergic neurons and accumulates in the core of Lewy bodies in Parkinson disease.* J Neuropathol Exp Neurol, 2010. **69**(9): p. 959-72.

83. Alegre-Abarrategui, J., et al., *LRRK2 is a component of granular alpha-synuclein pathology in the brainstem of Parkinson's disease.* Neuropathol Appl Neurobiol, 2008. **34**(3): p. 272-83.

84. Greggio, E., *Role of LRRK2 kinase activity in the pathogenesis of Parkinson's disease.* Biochem Soc Trans, 2012. **40**(5): p. 1058-62.

85. Lesage, S., et al., *Identification of VPS35 mutations replicated in French families with Parkinson disease.* Neurology, 2012. **78**(18): p. 1449-50.

86. Nuytemans, K., et al., *Whole exome sequencing of rare variants in EIF4G1 and VPS35 in Parkinson disease.* Neurology, 2013. **80**(11): p. 982-9.

87. Zimprich, A., et al., *A mutation in VPS35, encoding a subunit of the retromer complex, causes late-onset Parkinson disease.* Am J Hum Genet, 2011. **89**(1): p. 168-75.

88. Foroud, T., et al., *Heterozygosity for a mutation in the parkin gene leads to later onset Parkinson disease.* Neurology, 2003. **60**(5): p. 796-801.

89. Hedrich, K., et al., *Distribution, type, and origin of Parkin mutations: review and case studies.* Mov Disord, 2004. **19**(10): p. 1146-57.

90. Hedrich, K., et al., *The importance of gene dosage studies: mutational analysis of the parkin gene in early-onset parkinsonism.* Hum Mol Genet, 2001. **10**(16): p. 1649-56.

91. Lesage, S., et al., *Deletion of the parkin and PACRG gene promoter in early-onset parkinsonism.* Hum Mutat, 2007. **28**(1): p. 27-32.

92. West, A.B., et al., *Functional association of the parkin gene promoter with idiopathic Parkinson's disease.* Hum Mol Genet, 2002. **11**(22): p. 2787-92.

93. Grunewald, A., et al., *Mutant Parkin impairs mitochondrial function and morphology in human fibroblasts.* PLoS One, 2010. **5**(9): p. e12962.

94. Lucking, C.B., et al., *Association between early-onset Parkinson's disease and mutations in the parkin gene.* N Engl J Med, 2000. **342**(21): p. 1560-7.

95. Scuderi, S., et al., *Alternative splicing generates different parkin protein isoforms: evidences in human, rat, and mouse brain.* Biomed Res Int, 2014. **2014**: p. 690796.

96. Khan, N.L., et al., *Parkin disease: a phenotypic study of a large case series.* Brain, 2003. **126**(Pt 6): p. 1279-92.

97. Periquet, M., et al., *Parkin mutations are frequent in patients with isolated early-onset parkinsonism.* Brain, 2003. **126**(Pt 6): p. 1271-8.

98. van de Warrenburg, B.P., et al., *Clinical and pathologic abnormalities in a family with parkinsonism and parkin gene mutations.* Neurology, 2001. **56**(4): p. 555-7.

99. Farrer, M., et al., *Lewy bodies and parkinsonism in families with parkin mutations.* Ann Neurol, 2001. **50**(3): p. 293-300.

100. Mori, H., N. Hattori, and Y. Mizuno, *Genotype-phenotype correlation: familial Parkinson disease.* Neuropathology, 2003. **23**(1): p. 90-4.

101. Doherty, K.M. and J. Hardy, *Parkin disease and the Lewy body conundrum.* Mov Disord, 2013. **28**(6): p. 702-4.

102. Takahashi, H., et al., *Familial juvenile parkinsonism: clinical and pathologic study in a family.* Neurology, 1994. **44**(3 Pt 1): p. 437-41.

103. Mizutani, Y., M. Yokochi, and S. Oyanagi, *Juvenile parkinsonism: a case with first clinical manifestation at the age of six years and with neuropathological findings suggesting a new pathogenesis.* Clin Neuropathol, 1991. **10**(2): p. 91-7.

104. Doherty, K.M., et al., *Parkin disease: a clinicopathologic entity?* JAMA Neurol, 2013. **70**(5): p. 571-9.

105. Valente, E.M., et al., *Localization of a novel locus for autosomal recessive early-onset parkinsonism, PARK6, on human chromosome 1p35-p36.* Am J Hum Genet, 2001. **68**(4): p. 895-900.

106. Valente, E.M., et al., *Hereditary early-onset Parkinson's disease caused by mutations in PINK1.* Science, 2004. **304**(5674): p. 1158-60.

107. Bonifati, V., et al., *Early-onset parkinsonism associated with PINK1 mutations: frequency, genotypes, and phenotypes.* Neurology, 2005. **65**(1): p. 87-95.

108. Healy, D.G., et al., *The gene responsible for PARK6 Parkinson's disease, PINK1, does not influence common forms of parkinsonism.* Ann Neurol, 2004. **56**(3): p. 329-35.

109. Klein, C., et al., *PINK1, Parkin, and DJ-1 mutations in Italian patients with early-onset parkinsonism.* Eur J Hum Genet, 2005. **13**(9): p. 1086-93.

110. Li, Y., et al., *Clinicogenetic study of PINK1 mutations in autosomal recessive early-onset parkinsonism.* Neurology, 2005. **64**(11): p. 1955-7.

111. Marongiu, R., et al., *Whole gene deletion and splicing mutations expand the PINK1 genotypic spectrum.* Hum Mutat, 2007. **28**(1): p. 98.

112. Ibanez, P., et al., *Mutational analysis of the PINK1 gene in early-onset parkinsonism in Europe and North Africa.* Brain, 2006. **129**(Pt 3): p. 686-94.

113. Bonifati, V., et al., *DJ-1(PARK7), a novel gene for autosomal recessive, early onset parkinsonism.* Neurol Sci, 2003. **24**(3): p. 159-60.

114. Abou-Sleiman, P.M., et al., *The role of pathogenic DJ-1 mutations in Parkinson's disease.* Annals of neurology, 2003. **54**(3): p. 283-6.

115. Canet-Aviles, R.M., et al., *The Parkinson's disease protein DJ-1 is neuroprotective due to cysteine-sulfinic acid-driven mitochondrial localization.* Proc Natl Acad Sci U S A, 2004. **101**(24): p. 9103-8.

116. Mitsumoto, A. and Y. Nakagawa, *DJ-1 is an indicator for endogenous reactive oxygen species elicited by endotoxin.* Free Radic Res, 2001. **35**(6): p. 885-93.

117. Mitsumoto, A., et al., *Oxidized forms of peroxiredoxins and DJ-1 on two-dimensional gels increased in response to sublethal levels of paraquat.* Free Radic Res, 2001. **35**(3): p. 301-10.

118. Pankratz, N., et al., *Mutations in DJ-1 are rare in familial Parkinson disease.* Neurosci Lett, 2006. **408**(3): p. 209-13.

119. Zhou, W., et al., *The oxidation state of DJ-1 regulates its chaperone activity toward alpha-synuclein.* J Mol Biol, 2006. **356**(4): p. 1036-48.

120. Nagakubo, D., et al., *DJ-1, a novel oncogene which transforms mouse NIH3T3 cells in cooperation with ras.* Biochem Biophys Res Commun, 1997. **231**(2): p. 509-13.

121. Taira, T., et al., *DJ-1 has a role in antioxidative stress to prevent cell death.* EMBO Rep, 2004. **5**(2): p. 213-8.

122. Kinumi, T., et al., *Cysteine-106 of DJ-1 is the most sensitive cysteine residue to hydrogen peroxide-mediated oxidation in vivo in human umbilical vein endothelial cells.* Biochem Biophys Res Commun, 2004. **317**(3): p. 722-8.

123. Menzies, F.M., S.C. Yenisetti, and K.T. Min, *Roles of Drosophila DJ-1 in survival of dopaminergic neurons and oxidative stress.* Curr Biol, 2005. **15**(17): p. 1578-82.

124. Meulener, M., et al., *Drosophila DJ-1 mutants are selectively sensitive to environmental toxins associated with Parkinson's disease.* Curr Biol, 2005. **15**(17): p. 1572-7.

125. Martinat, C., et al., *Sensitivity to oxidative stress in DJ-1-deficient dopamine neurons: an ES- derived cell model of primary Parkinsonism.* PLoS Biol, 2004. **2**(11): p. e327.

126. Takahashi-Niki, K., et al., *Reduced anti-oxidative stress activities of DJ-1 mutants found in Parkinson's disease patients.* Biochem Biophys Res Commun, 2004. **320**(2): p. 389-97.

127. Shendelman, S., et al., *DJ-1 is a redox-dependent molecular chaperone that inhibits alpha-synuclein aggregate formation.* PLoS Biol, 2004. **2**(11): p. e362.

128. Xu, J., et al., *The Parkinson's disease-associated DJ-1 protein is a transcriptional co-activator that protects against neuronal apoptosis.* Hum Mol Genet, 2005. **14**(9): p. 1231-41.

129. Bruggemann, N., et al., *Recessively inherited parkinsonism: effect of ATP13A2 mutations on the clinical and neuroimaging phenotype.* Arch Neurol, 2010. **67**(11): p. 1357-63.

130. Ramirez, A., et al., *Hereditary parkinsonism with dementia is caused by mutations in ATP13A2, encoding a lysosomal type 5 P-type ATPase.* Nat Genet, 2006. **38**(10): p. 1184-91.

131. Morgan, N.V., et al., *PLA2G6, encoding a phospholipase A2, is mutated in neurodegenerative disorders with high brain iron.* Nat Genet, 2006. **38**(7): p. 752-4.

132. Paisan-Ruiz, C., et al., *Characterization of PLA2G6 as a locus for dystonia-parkinsonism.* Ann Neurol, 2009. **65**(1): p. 19-23.

133. Sina, F., et al., *R632W mutation in PLA2G6 segregates with dystonia-parkinsonism in a consanguineous Iranian family.* Eur J Neurol, 2009. **16**(1): p. 101-4.

134. Shojaee, S., et al., *Genome-wide linkage analysis of a Parkinsonian-pyramidal syndrome pedigree by 500 K SNP arrays.* Am J Hum Genet, 2008. **82**(6): p. 1375-84.

135. Kruger, R., et al., *Increased susceptibility to sporadic Parkinson's disease by a certain combined alpha-synuclein/apolipoprotein E genotype.* Ann Neurol, 1999. **45**(5): p. 611-7.

136. Chouraki, V. and S. Seshadri, *Genetics of Alzheimer's disease.* Adv Genet, 2014. **87**: p. 245-94.

137. Maraganore, D.M., et al., *Collaborative analysis of alpha-synuclein gene promoter variability and Parkinson disease.* JAMA, 2006. **296**(6): p. 661-70.

138. Mata, I.F., et al., *Lrrk2 pathogenic substitutions in Parkinson's disease.* Neurogenetics, 2005. **6**(4): p. 171-7.

139. Di Fonzo, A., et al., *A common missense variant in the LRRK2 gene, Gly2385Arg, associated with Parkinson's disease risk in Taiwan.* Neurogenetics, 2006. **7**(3): p. 133-8.

140. Chan, D.K., et al., *LRRK2 Gly2385Arg mutation and clinical features in a Chinese population with early-onset Parkinson's disease compared to late-onset patients.* J Neural Transm, 2008. **115**(9): p. 1275-7.

141. Fu, X., et al., *LRRK2 G2385R and LRRK2 R1628P increase risk of Parkinson's disease in a Han Chinese population from Southern Mainland China.* Parkinsonism Relat Disord, 2013. **19**(3): p. 397-8.

142. Funayama, M., et al., *Leucine-rich repeat kinase 2 G2385R variant is a risk factor for Parkinson disease in Asian population.* Neuroreport, 2007. **18**(3): p. 273-5.

143. Kim, J.M., et al., *The LRRK2 G2385R variant is a risk factor for sporadic Parkinson's disease in the Korean population.* Parkinsonism Relat Disord, 2010. **16**(2): p. 85-8.

144. Li, C., et al., *The prevalence of LRRK2 Gly2385Arg variant in Chinese Han population with Parkinson's disease.* Mov Disord, 2007. **22**(16): p. 2439-43.

145. Tan, E.K., et al., *The LRRK2 Gly2385Arg variant is associated with Parkinson's disease: genetic and functional evidence.* Hum Genet, 2007. **120**(6): p. 857-63.

146. Tan, E.K., et al., *Analysis of LRRK2 Gly2385Arg genetic variant in non-Chinese Asians.* Mov Disord, 2007. **22**(12): p. 1816-8.

147. Zabetian, C.P., et al., *LRRK2 mutations and risk variants in Japanese patients with Parkinson's disease.* Mov Disord, 2009. **24**(7): p. 1034-41.

148. Tan, E.K., et al., *LRRK2 G2385R modulates age at onset in Parkinson's disease: A multi-center pooled analysis.* Am J Med Genet B Neuropsychiatr Genet, 2009. **150B**(7): p. 1022-3.

149. Wang, C., et al., *Penetrance of LRRK2 G2385R and R1628P is modified by common PD-associated genetic variants.* Parkinsonism Relat Disord, 2012. **18**(8): p. 958-63.

150. Farrer, M.J., et al., *Lrrk2 G2385R is an ancestral risk factor for Parkinson's disease in Asia.* Parkinsonism Relat Disord, 2007. **13**(2): p. 89-92.

151. Pulkes, T., et al., *Frequencies of LRRK2 variants in Thai patients with Parkinson's disease: evidence for an R1628P founder.* J Neurol Neurosurg Psychiatry, 2011. **82**(10): p. 1179-80.

152. Wu, X., et al., *Quantitative assessment of the effect of LRRK2 exonic variants on the risk of Parkinson's disease: a meta-analysis.* Parkinsonism Relat Disord, 2012. **18**(6): p. 722-30.

153. Wu-Chou, Y.H., et al., *Genetic variants of SNCA and LRRK2 genes are associated with sporadic PD susceptibility: a replication study in a Taiwanese cohort.* Parkinsonism Relat Disord, 2013. **19**(2): p. 251-5.

154. Ross, O.A., et al., *Analysis of Lrrk2 R1628P as a risk factor for Parkinson's disease.* Ann Neurol, 2008. **64**(1): p. 88-92.

155. Ross, O.A., et al., *Association of LRRK2 exonic variants with susceptibility to Parkinson's disease: a case-control study.* Lancet Neurol, 2011. **10**(10): p. 898-908.

156. Tsuji, S., et al., *A mutation in the human glucocerebrosidase gene in neuronopathic Gaucher's disease.* N Engl J Med, 1987. **316**(10): p. 570-5.

157. Tayebi, N., et al., *Gaucher disease and parkinsonism: a phenotypic and genotypic characterization.* Mol Genet Metab, 2001. **73**(4): p. 313-21.

158. Tayebi, N., et al., *Gaucher disease with parkinsonian manifestations: does glucocerebrosidase deficiency contribute to a vulnerability to parkinsonism?* Mol Genet Metab, 2003. **79**(2): p. 104-9.

159. Aharon-Peretz, J., H. Rosenbaum, and R. Gershoni-Baruch, *Mutations in the glucocerebrosidase gene and Parkinson's disease in Ashkenazi Jews.* N Engl J Med, 2004. **351**(19): p. 1972-7.

160. Sidransky, E., T. Samaddar, and N. Tayebi, *Mutations in GBA are associated with familial Parkinson disease susceptibility and age at onset.* Neurology, 2009. **73**(17): p. 1424-5, author reply 1425-6.

161. Nalls, M.A., et al., *A multicenter study of glucocerebrosidase mutations in dementia with Lewy bodies.* JAMA Neurol, 2013. **70**(6): p. 727-35.

162. Mill, J., *Toward an integrated genetic and epigenetic approach to Alzheimer's disease.* Neurobiol Aging.

163. Doi, A., et al., *Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts.* Nat Genet, 2009. **41**(12): p. 1350-3.

164. Irizarry, R.A., et al., *The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores.* Nat Genet, 2009. **41**(2): p. 178-86.

165. Lister, R., et al., *Human DNA methylomes at base resolution show widespread epigenomic differences.* Nature, 2009. **462**(7271): p. 315-22.

166. Dawson, M.A., et al., *JAK2 phosphorylates histone H3Y41 and excludes HP1alpha from chromatin.* Nature, 2009. **461**(7265): p. 819-22.

167. Hurd, P.J., et al., *Phosphorylation of histone H3 Thr-45 is linked to apoptosis.* J Biol Chem, 2009. **284**(24): p. 16575-83.

168. Zhao, Q., et al., *PRMT5-mediated methylation of histone H4R3 recruits DNMT3A, coupling histone and DNA methylation in gene silencing.* Nat Struct Mol Biol, 2009. **16**(3): p. 304-11.

169. Egger, G., et al., *Epigenetics in human disease and prospects for epigenetic therapy.* Nature, 2004. **429**(6990): p. 457-63.

170. Portela, A. and M. Esteller, *Epigenetic modifications and human disease.* Nat Biotechnol, 2010. **28**(20944598): p. 1057-1068.

171. Weinhold, B., *Epigenetics: the science of change.* Environ Health Perspect, 2006. **114**(3): p. A160-7.

172. Portela, A. and M. Esteller, *Epigenetic modifications and human disease.* Nat Biotechnol. **28**(10): p. 1057-68.

173. Razin, A. and A.D. Riggs, *DNA methylation and gene function.* Science, 1980. **210**(4470): p. 604-10.

174. Goll, M.G. and T.H. Bestor, *Eukaryotic cytosine methyltransferases.* Annu Rev Biochem, 2005. **74**: p. 481-514.

175. Lister, R., et al., *Human DNA methylomes at base resolution show widespread epigenomic differences.* Nature, 2009. **462**(19829295): p. 315-322.

176. Herman, J.G. and S.B. Baylin, *Gene silencing in cancer in association with promoter hypermethylation.* N Engl J Med, 2003. **349**(21): p. 2042-54.

177. Bird, A.P., *CpG-rich islands and the function of DNA methylation.* Nature, 1986. **321**(6067): p. 209-13.

178. Gardiner-Garden, M. and M. Frommer, *CpG islands in vertebrate genomes.* J Mol Biol, 1987. **196**(2): p. 261-82.

179. Shen, L., et al., *Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters.* PLoS Genet, 2007. **3**(10): p. 2023-36.

180. Takai, D. and P.A. Jones, *Comprehensive analysis of CpG islands in human chromosomes 21 and 22.* Proc Natl Acad Sci U S A, 2002. **99**(6): p. 3740-5.

181. Zhou, H., H. Hu, and M. Lai, *Non-coding RNAs and their epigenetic regulatory mechanisms.* Biol Cell. **102**(12): p. 645-55.

182. Maunakea, A.K., et al., *Conserved role of intragenic DNA methylation in regulating alternative promoters.* Nature. **466**(7303): p. 253-7.

183. Tost, J., *Epigenetics*. 2008, Norfolk, UK: Caister Academic Press. xi, 404, 3 p. of plates.

184. Ehrich, M., et al., *Cytosine methylation profiling of cancer cell lines.* Proc Natl Acad Sci U S A, 2008. **105**(12): p. 4844-9.

185. Bibikova, M., et al., *High-throughput DNA methylation profiling using universal bead arrays.* Genome Res, 2006. **16**(3): p. 383-93.

186. Clark, S.J., et al., *High sensitivity mapping of methylated cytosines.* Nucleic Acids Res, 1994. **22**(15): p. 2990-7.

187. Clark, S.J., et al., *DNA methylation: bisulphite modification and analysis.* Nat Protoc, 2006. **1**(5): p. 2353-64.

188. Fraga, M.F. and M. Esteller, *DNA methylation: a profile of methods and applications.* Biotechniques, 2002. **33**(3): p. 632, 634, 636-49.

189. Pomraning, K.R., K.M. Smith, and M. Freitag, *Genome-wide high throughput analysis of DNA methylation in eukaryotes.* Methods, 2009. **47**(3): p. 142-50.

190. Hatada, I., et al., *A microarray-based method for detecting methylated loci.* J Hum Genet, 2002. **47**(8): p. 448-51.

191. van Steensel, B., J. Delrow, and S. Henikoff, *Chromatin profiling using targeted DNA adenine methyltransferase.* Nat Genet, 2001. **27**(3): p. 304-8.

192. Huang, T.H., M.R. Perry, and D.E. Laux, *Methylation profiling of CpG islands in human breast cancer cells.* Hum Mol Genet, 1999. **8**(3): p. 459-70.

193. Beck, S., A. Olek, and J. Walter, *From genomics to epigenomics: a loftier view of life.* Nat Biotechnol, 1999. **17**(12): p. 1144.

194. Gu, H., et al., *Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling.* Nat Protoc, 2011. **6**(4): p. 468-81.

195. Rauch, T.A. and G.P. Pfeifer, *The MIRA method for DNA methylation analysis.* Methods Mol Biol, 2009. **507**: p. 65-75.

196. Suzuki, M. and J.M. Greally, *DNA methylation profiling using HpaII tiny fragment enrichment by ligation-mediated PCR (HELP).* Methods, 2010. **52**(3): p. 218-22.

197. Bock, C., et al., *Quantitative comparison of genome-wide DNA methylation mapping technologies.* Nat Biotechnol, 2010. **28**(10): p. 1106-14.

198. Fazzari, M.J. and J.M. Greally, *Introduction to epigenomics and epigenome-wide analysis.* Methods Mol Biol. **620**: p. 243-65.

199. Fazzari, M.J. and J.M. Greally, *Epigenomics: beyond CpG islands.* Nat Rev Genet, 2004. **5**(6): p. 446-55.

200. Symonds, M.E., et al., *Nutritional programming of the metabolic syndrome.* Nat Rev Endocrinol, 2009. **5**(11): p. 604-10.

201. Turunen, M.P., E. Aavik, and S. Yla-Herttuala, *Epigenetics and atherosclerosis.* Biochim Biophys Acta, 2009. **1790**(9): p. 886-91.

202. Hang, C.T., et al., *Chromatin regulation by Brg1 underlies heart muscle development and disease.* Nature. **466**(7302): p. 62-7.

203. Movassagh, M., et al., *Differential DNA methylation correlates with differential expression of angiogenic factors in human heart failure.* PLoS One. **5**(1): p. e8564.

204. Zeng, W., et al., *Specific loss of histone H3 lysine 9 trimethylation and HP1gamma/cohesin binding at D4Z4 repeats is associated with facioscapulohumeral dystrophy (FSHD).* PLoS Genet, 2009. **5**(7): p. e1000559.

205. Gheldof, N., T.M. Tabuchi, and J. Dekker, *The active FMR1 promoter is associated with a large domain of altered chromatin conformation with embedded local histone modifications.* Proc Natl Acad Sci U S A, 2006. **103**(33): p. 12463-8.

206. Urdinguio, R.G., J.V. Sanchez-Mut, and M. Esteller, *Epigenetic mechanisms in neurological diseases: genes, syndromes, and therapies.* Lancet Neurol, 2009. **8**(11): p. 1056-72.

207. Pieper, H.C., et al., *Different methylation of the TNF-alpha promoter in cortex and substantia nigra: Implications for selective neuronal vulnerability.* Neurobiol Dis, 2008. **32**(3): p. 521-7.

208. Gao, H.M. and J.S. Hong, *Gene-environment interactions: Key to unraveling the mystery of Parkinson's disease.* Prog Neurobiol. **94**(1): p. 1-19.

209. Chouliaras, L., et al., *Epigenetic regulation in the pathophysiology of Alzheimer's disease.* Prog Neurobiol. **90**(4): p. 498-510.

210. Fraga, M.F., et al., *Epigenetic differences arise during the lifetime of monozygotic twins.* Proc Natl Acad Sci U S A, 2005. **102**(30): p. 10604-9.

211. Yu, L., et al., *Association of Brain DNA Methylation in SORL1, ABCA7, HLA-DRB5, SLC24A4, and BIN1 With Pathological Diagnosis of Alzheimer Disease.* JAMA Neurol, 2014.

212. Di Francesco, A., et al., *Global changes in DNA methylation in Alzheimer's disease peripheral blood mononuclear cells.* Brain Behav Immun, 2014.

213. Humphries, C.E., et al., *Integrated Whole Transcriptome and DNA Methylation Analysis Identifies Gene Networks Specific to Late-Onset Alzheimer's Disease.* J Alzheimers Dis, 2014.

214. Lee, J., et al., *Epigenetic mechanisms of neurodegeneration in Huntington's disease.* Neurotherapeutics, 2013. **10**(4): p. 664-76.

215. Martin, L.J. and M. Wong, *Aberrant regulation of DNA methylation in amyotrophic lateral sclerosis: a new target of disease mechanisms.* Neurotherapeutics, 2013. **10**(4): p. 722-33.

216. Tan, Y.Y., et al., *Methylation of alpha-synuclein and leucine-rich repeat kinase 2 in leukocyte DNA of Parkinson's disease patients.* Parkinsonism Relat Disord, 2014. **20**(3): p. 308-13.

217. Coupland, K.G., et al., *DNA methylation of the MAPT gene in Parkinson's disease cohorts and modulation by vitamin E In Vitro.* Mov Disord, 2014. **29**(13): p. 1606-14.

218. Jowaed, A., et al., *Methylation regulates alpha-synuclein expression and is decreased in Parkinson's disease patients' brains.* J Neurosci, 2010. **30**(18): p. 6355-9.

219. Matsumoto, L., et al., *CpG demethylation enhances alpha-synuclein expression and affects the pathogenesis of Parkinson's disease.* PLoS One, 2010. **5**(11): p. e15522.

220. DePinho, R.A., *The age of cancer.* Nature, 2000. **408**(6809): p. 248-54.

221. Issa, J.P., *CpG-island methylation in aging and cancer.* Curr Top Microbiol Immunol, 2000. **249**: p. 101-18.

222. Alzheimer's, A., *2013 Alzheimer's disease facts and figures.* Alzheimers Dement, 2013. **9**(2): p. 208-45.

223. de Lau, L.M. and M.M. Breteler, *Epidemiology of Parkinson's disease.* Lancet Neurol, 2006. **5**(6): p. 525-35.

224. Kirkwood, T.B., *Understanding the odd science of aging.* Cell, 2005. **120**(4): p. 437-47.

225. Gopisetty, G., K. Ramachandran, and R. Singal, *DNA methylation and apoptosis.* Mol Immunol, 2006. **43**(11): p. 1729-40.

226. Ottaviano, Y.L., et al., *Methylation of the estrogen receptor gene CpG island marks loss of estrogen receptor expression in human breast cancer cells.* Cancer Res, 1994. **54**(10): p. 2552-5.

227. Richardson, B., *Impact of aging on DNA methylation.* Ageing Res Rev, 2003. **2**(3): p. 245-61.

228. Bell, J.T., et al., *Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population.* PLoS Genet, 2012. **8**(4): p. e1002629.

229. Teschendorff, A.E., et al., *Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer.* Genome Res, 2010. **20**(4): p. 440-6.

230. Bocklandt, S., et al., *Epigenetic predictor of age.* PLoS One, 2011. **6**(6): p. e14821.

231. Johnson, A.A., et al., *The role of DNA methylation in aging, rejuvenation, and age-related disease.* Rejuvenation Res, 2012. **15**(5): p. 483-94.

232. Qi, Y., et al., *Decreased Srcasm expression in esophageal squamous cell carcinoma in a Chinese population.* Anticancer Res, 2010. **30**(9): p. 3535-9.

233. Langton, A.K., S.E. Herrick, and D.J. Headon, *An extended epidermal response heals cutaneous wounds in the absence of a hair follicle stem cell contribution.* J Invest Dermatol, 2008. **128**(5): p. 1311-8.

234. Park, J.K., et al., *Quantitative analysis of NPTX2 hypermethylation is a promising molecular diagnostic marker for pancreatic cancer.* Pancreas, 2007. **35**(3): p. e9-15.

235. Moran, L.B., et al., *Neuronal pentraxin II is highly upregulated in Parkinson's disease and a novel component of Lewy bodies.* Acta Neuropathol, 2008. **115**(4): p. 471-8.

236. Dorsey, D.A., et al., *Ultrastructural characterization of alpha-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid-induced cell death in embryonic dopaminergic neurons.* Apoptosis, 2006. **11**(4): p. 535-44.

237. Franco, R., et al., *Oxidative stress, DNA methylation and carcinogenesis.* Cancer Lett, 2008. **266**(1): p. 6-11.

238. van Duijn, C.M., et al., *Park7, a novel locus for autosomal recessive early-onset parkinsonism, on chromosome 1p36.* Am J Hum Genet, 2001. **69**(3): p. 629-34.

239. Singleton, A.B., *Genetics. A unified process for neurological disease.* Science, 2014. **343**(6170): p. 497-8.

240. Novarino, G., et al., *Exome sequencing links corticospinal motor neuron disease to common neurodegenerative disorders.* Science, 2014. **343**(6170): p. 506-11.

241. Klein, R.J., et al., *Complement factor H polymorphism in age-related macular degeneration.* Science, 2005. **308**(5720): p. 385-9.

242. Consortium, U.K.P.s.D., et al., *Dissection of the genetics of Parkinson's disease identifies an additional association 5' of SNCA and multiple associated haplotypes at 17q21.* Hum Mol Genet, 2011. **20**(2): p. 345-53.

243. Edwards, T.L., et al., *Genome-wide association study confirms SNPs in SNCA and the MAPT region as common risk factors for Parkinson disease.* Ann Hum Genet, 2010. **74**(2): p. 97-109.

244. Fung, H.C., et al., *Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data.* Lancet Neurol, 2006. **5**(11): p. 911-6.

245. Hamza, T.H., et al., *Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease.* Nat Genet, 2010. **42**(9): p. 781-5.

246. Pankratz, N., et al., *Genomewide association study for susceptibility genes contributing to familial Parkinson disease.* Hum Genet, 2009. **124**(6): p. 593-605.

247. Satake, W., et al., *Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease.* Nat Genet, 2009. **41**(12): p. 1303-7.

248. Simon-Sanchez, J., et al., *Genome-wide association study reveals genetic risk underlying Parkinson's disease.* Nat Genet, 2009. **41**(12): p. 1308-12.

249. Simon-Sanchez, J., et al., *Genome-wide association study confirms extant PD risk loci among the Dutch.* Eur J Hum Genet, 2011. **19**(6): p. 655-61.

250. Maraganore, D.M., et al., *High-resolution whole-genome association study of Parkinson disease.* Am J Hum Genet, 2005. **77**(5): p. 685-93.

251. Wichmann, H.E., et al., *KORA-gen--resource for population genetics, controls and a broad spectrum of disease phenotypes.* Gesundheitswesen, 2005. **67 Suppl 1**: p. S26-30.

252. Yamauchi, K. and S. Arimori, *Basophilic crisis in chronic myelogenous leukemia: case report and literature review in Japan.* Jpn J Med, 1990. **29**(3): p. 334-40.

253. Li, Y., et al., *Genotype imputation.* Annu Rev Genomics Hum Genet, 2009. **10**: p. 387-406.

254. Cortes, A. and M.A. Brown, *Promise and pitfalls of the Immunochip.* Arthritis Res Ther, 2011. **13**(1): p. 101.

255. Steemers, F.J. and K.L. Gunderson, *Whole genome genotyping technologies on the BeadArray platform.* Biotechnol J, 2007. **2**(1): p. 41-9.

256. Higgins, J.P., et al., *Measuring inconsistency in meta-analyses.* BMJ, 2003. **327**(7414): p. 557-60.

257. Ioannidis, J.P., N.A. Patsopoulos, and E. Evangelou, *Heterogeneity in meta-analyses of genome-wide association investigations.* PLoS One, 2007. **2**(9): p. e841.

258. Saad, M., et al., *Genome-wide association study confirms BST1 and suggests a locus on 12q24 as the risk loci for Parkinson's disease in the European population.* Hum Mol Genet, 2011. **20**(3): p. 615-27.

259. Sidransky, E., et al., *Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease.* N Engl J Med, 2009. **361**(17): p. 1651-61.

260. Dickson, S.P., et al., *Rare variants create synthetic genome-wide associations.* PLoS Biol, 2010. **8**(1): p. e1000294.

261. Glass, A.S., et al., *Screening for mutations in synaptotagmin XI in Parkinson's disease.* J Neural Transm Suppl, 2004(68): p. 21-8.

262. Huynh, D.P., et al., *The autosomal recessive juvenile Parkinson disease gene product, parkin, interacts with and ubiquitinates synaptotagmin XI.* Hum Mol Genet, 2003. **12**(20): p. 2587-97.

263. Garavaglia, S., et al., *The crystal structure of human alpha-amino-beta-carboxymuconate-epsilon-semialdehyde decarboxylase in complex with 1,3-dihydroxyacetonephosphate suggests a regulatory link between NAD synthesis and glycolysis.* FEBS J, 2009. **276**(22): p. 6615-23.

264. Ramoz, N., et al., *An analysis of candidate autism loci on chromosome 2q24-q33: evidence for association to the STK39 gene.* Am J Med Genet B Neuropsychiatr Genet, 2008. **147B**(7): p. 1152-8.

265. Cunnington, M.S., et al., *STK39 polymorphisms and blood pressure: an association study in British Caucasians and assessment of cis-acting influences on gene expression.* BMC Med Genet, 2009. **10**: p. 135.

266. Malosio, M.L., et al., *Dense-core granules: a specific hallmark of the neuronal/neurosecretory cell phenotype.* J Cell Sci, 2004. **117**(Pt 5): p. 743-9.

267. Parker, J.A., et al., *Huntingtin-interacting protein 1 influences worm and mouse presynaptic function and protects Caenorhabditis elegans neurons against mutant polyglutamine toxicity.* J Neurosci, 2007. **27**(41): p. 11056-64.

268. Wersinger, C. and A. Sidhu, *An inflammatory pathomechanism for Parkinson's disease?* Curr Med Chem, 2006. **13**(5): p. 591-602.

269. Karch, C.M. and A.M. Goate, *Alzheimer's Disease Risk Genes and Mechanisms of Disease Pathogenesis.* Biol Psychiatry, 2015. **77**(1): p. 43-51.

270. Cree, B.A., *Multiple sclerosis genetics.* Handb Clin Neurol, 2014. **122**: p. 193-209.

271. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.

272. Li, M., M. Boehnke, and G.R. Abecasis, *Efficient study designs for test of genetic association using sibship data and unrelated cases and controls.* Am J Hum Genet, 2006. **78**(5): p. 778-92.

273. Li, Y., et al., *MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes.* Genet Epidemiol. **34**(8): p. 816-34.

274. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14863-8.

275. Saeed, A.I., et al., *TM4: a free, open-source system for microarray data management and analysis.* Biotechniques, 2003. **34**(2): p. 374-8.

276. R Development Core Team, *R: A language and environment for statistical computing*. Vol. 2. 2005, Vienna: R Foundation for Statistical Computing.

277. Churchill, G.A. and R.W. Doerge, *Empirical threshold values for quantitative trait mapping.* Genetics, 1994. **138**(3): p. 963-71.

278. Khaitovich, P., et al., *Regional patterns of gene expression in human and chimpanzee brains.* Genome Res, 2004. **14**(15289471): p. 1462-1473.

279. Ladd-Acosta, C., et al., *DNA methylation signatures within the human brain.* Am J Hum Genet, 2007. **81**(6): p. 1304-15.

280. Zhang, Y., et al., *DNA methylation analysis of chromosome 21 gene promoters at single base pair and single allele resolution.* PLoS Genet, 2009. **5**(3): p. e1000438.

281. Meissner, A., et al., *Genome-scale DNA methylation maps of pluripotent and differentiated cells.* Nature, 2008. **454**(7205): p. 766-70.

282. Bock, C., et al., *CpG island mapping by epigenome prediction.* PLoS Comput Biol, 2007. **3**(6): p. e110.

283. Myers, A.J., et al., *A survey of genetic human cortical gene expression.* Nat Genet, 2007. **39**(12): p. 1494-9.

284. Veyrieras, J.B., et al., *High-resolution mapping of expression-QTLs yields insight into human gene regulation.* PLoS Genet, 2008. **4**(10): p. e1000214.

285. Monda, K.L., et al., *A meta-analysis identifies new loci associated with body mass index in individuals of African ancestry.* Nat Genet, 2013. **45**(6): p. 690-6.

286. Nalls, M.A., et al., *Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease.* Nat Genet, 2014. **46**(9): p. 989-93.

287. The, C., et al., *Genome-wide meta-analysis identifies six novel loci associated with habitual coffee consumption.* Mol Psychiatry, 2014.

288. Zhang, D., et al., *Genetic control of individual differences in gene-specific methylation in human brain.* Am J Hum Genet, 2010. **86**(3): p. 411-9.

289. Shoemaker, R., et al., *Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome.* Genome Res, 2010. **20**(7): p. 883-9.

290. Heyn, H., et al., *DNA methylation contributes to natural human variation.* Genome Res, 2013. **23**(9): p. 1363-72.

291. Bell, J.T., et al., *DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines.* Genome Biol, 2011. **12**(1): p. R10.

292. Fraser, H.B., et al., *Population-specificity of human DNA methylation.* Genome Biol, 2012. **13**(2): p. R8.

293. Liu, Y., et al., *Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis.* Nat Biotechnol, 2013. **31**(2): p. 142-7.

294. Lokk, K., et al., *DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns.* Genome Biol, 2014. **15**(4): p. r54.

295. Nalls, M.A., et al., *NeuroX, a fast and efficient genotyping platform for investigation of neurodegenerative diseases.* Neurobiol Aging, 2014.

296. Grove, M.L., et al., *Best practices and joint calling of the HumanExome BeadChip: the CHARGE Consortium.* PLoS One, 2013. **8**(7): p. e68095.

297. Ramasamy, A., et al., *Resolving the polymorphism-in-probe problem is critical for correct interpretation of expression QTL studies.* Nucleic Acids Res, 2013. **41**(7): p. e88.

298. International Parkinson Disease Genomics, C., et al., *Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies.* Lancet, 2011. **377**(9766): p. 641-9.

299. Ai, S.X., et al., *Hypomethylation of SNCA in blood of patients with sporadic Parkinson's disease.* J Neurol Sci, 2014. **337**(1-2): p. 123-8.

300. Trabzuni, D., et al., *MAPT expression and splicing is differentially regulated by brain region: relation to genotype and implication for tauopathies.* Hum Mol Genet, 2012. **21**(18): p. 4094-103.

301. Beilina, A., et al., *Unbiased screen for interactors of leucine-rich repeat kinase 2 supports a common pathway for sporadic and familial Parkinson disease.* Proc Natl Acad Sci U S A, 2014. **111**(7): p. 2626-31.

302. MacLeod, D.A., et al., *RAB7L1 interacts with LRRK2 to modify intraneuronal protein sorting and Parkinson's disease risk.* Neuron, 2013. **77**(3): p. 425-39.

303. Smemo, S., et al., *Obesity-associated variants within FTO form long-range functional connections with IRX3.* Nature, 2014. **507**(7492): p. 371-5.

304. de Lau, L.M. and M.M. Breteler, *Epidemiology of Parkinson's disease.* Lancet neurology, 2006. **5**(6): p. 525-35.

305. Gopisetty, G., K. Ramachandran, and R. Singal, *DNA methylation and apoptosis.* Mol Immunol, 2006. **43**(16500705): p. 1729-1740.

306. Ottaviano, Y.L., et al., *Methylation of the estrogen receptor gene CpG island marks loss of estrogen receptor expression in human breast cancer cells.* Cancer Res, 1994. **54**(8168078): p. 2552-2555.

307. Ladd-Acosta, C., et al., *DNA methylation signatures within the human brain.* Am J Hum Genet, 2007. **81**(17999367): p. 1304-1315.

308. Horvath, S., *DNA methylation age of human tissues and cell types.* Genome Biol, 2013. **14**(10): p. R115.

309. Strauss, W.M., *Preparation of genomic DNA from mammalian tissue.* Curr Protoc Mol Biol, 2001. **Chapter 2**: p. Unit2 2.

310. Bock, C., et al., *CpG island mapping by epigenome prediction.* PLoS Comput Biol, 2007. **3**(17559301).

311. Gardiner-Garden, M. and M. Frommer, *CpG islands in vertebrate genomes.* J Mol Biol, 1987. **196**(3656447): p. 261-282.

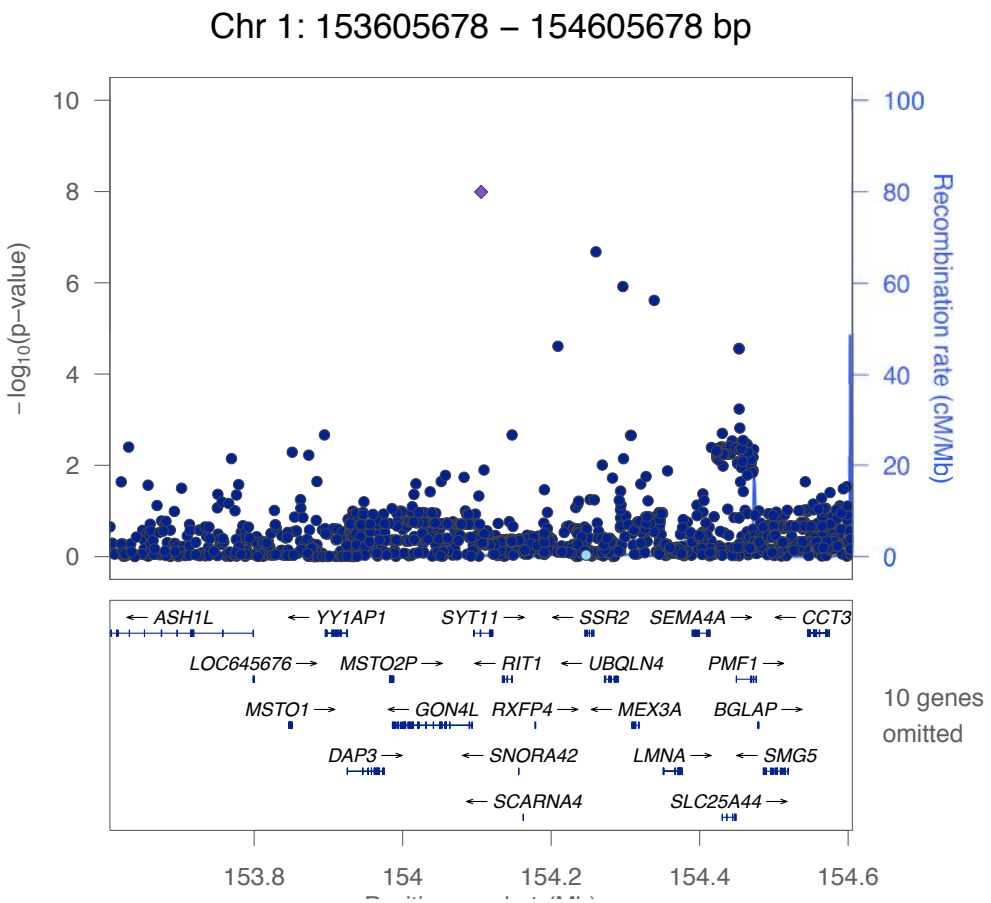
312. Christensen, B.C., et al., *Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context.* PLoS Genet, 2009. **5**(8): p. e1000602.

313. Rakyan, V.K., et al., *Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains.* Genome Res. **20**(4): p. 434-9.

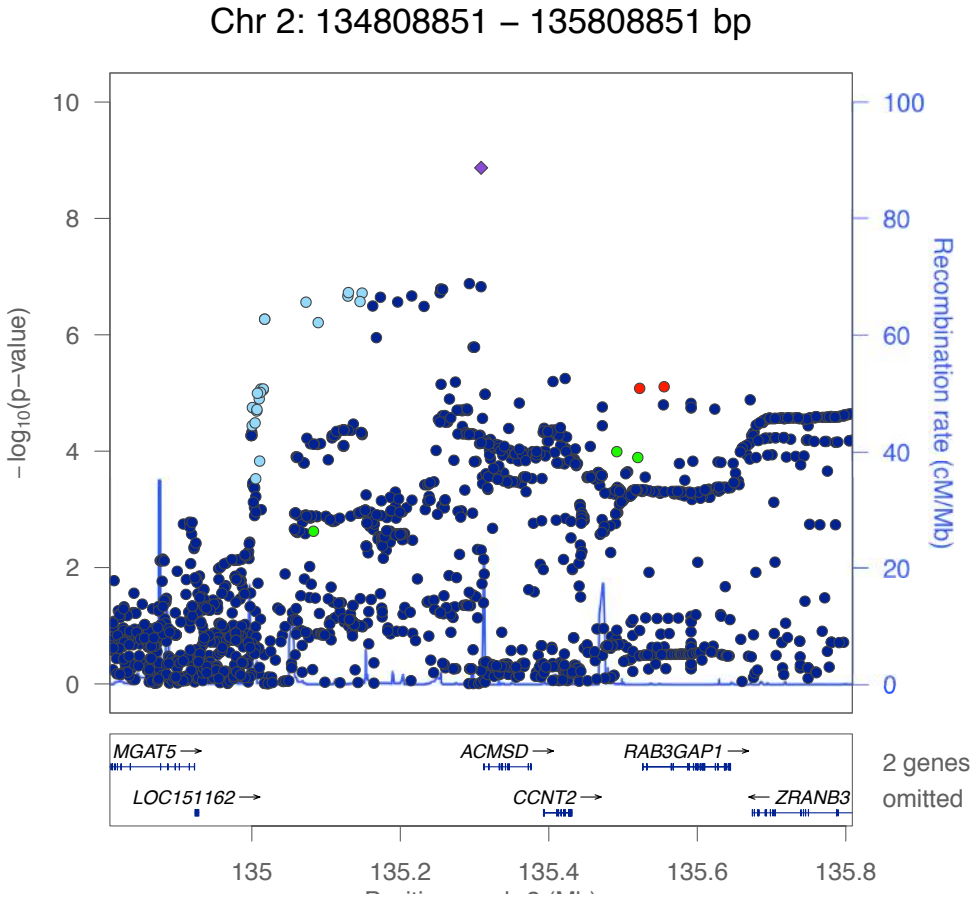
314. Siegmund, K.D., et al., *DNA methylation in the human cerebral cortex is dynamically regulated throughout the life span and involves differentiated neurons.* PLoS One, 2007. **2**(9): p. e895.

315. Southworth, L.K., A.B. Owen, and S.K. Kim, *Aging mice show a decreasing correlation of gene expression within genetic modules.* PLoS Genet, 2009. **5**(12): p. e1000776.

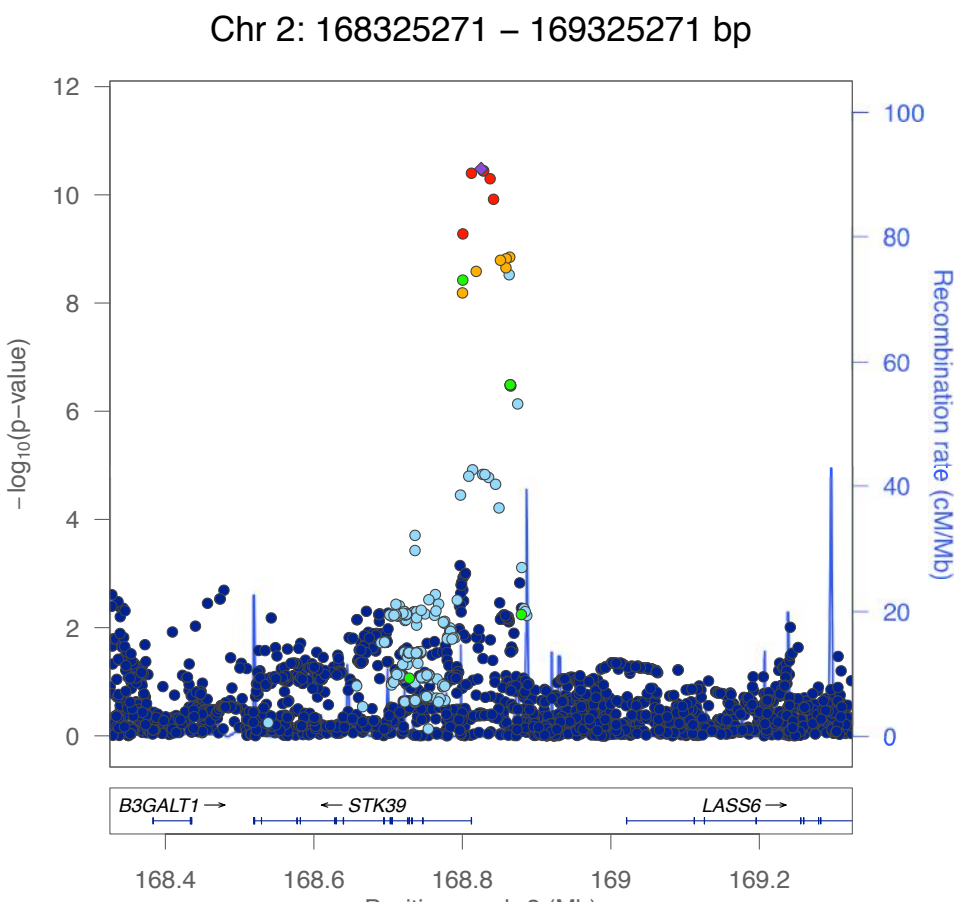
8 Supplementary Information



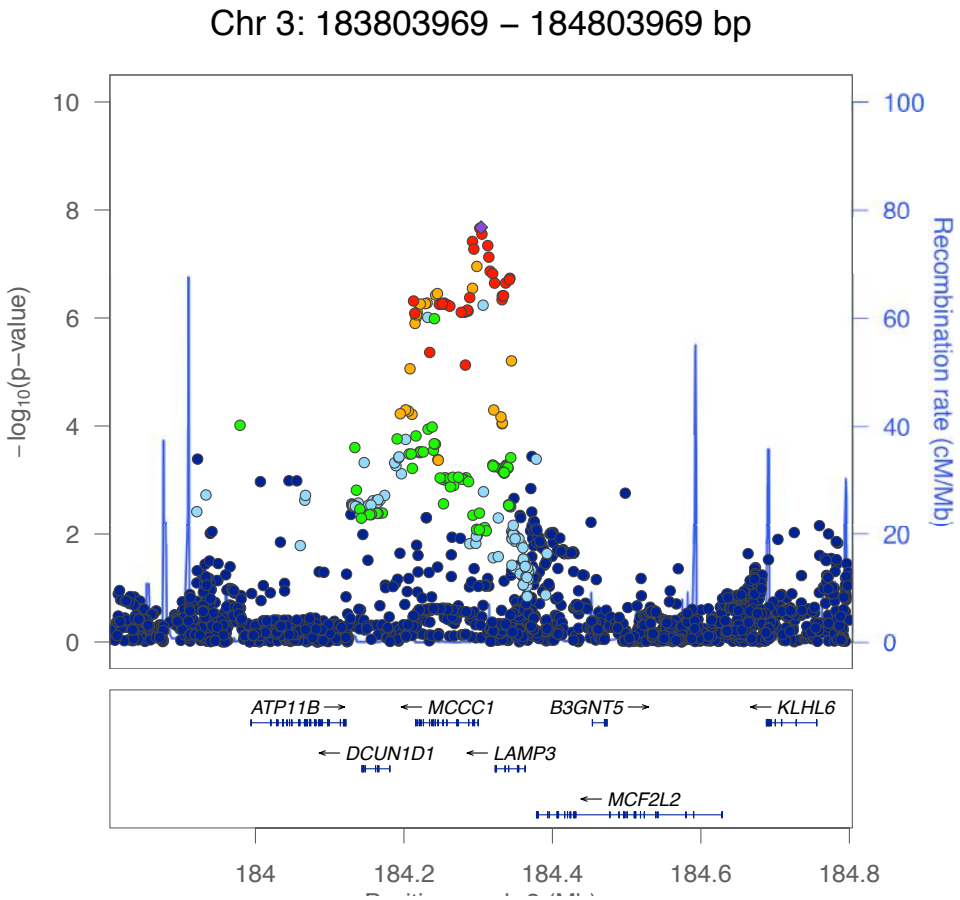
Supplementary Figure 1: Locus plot for the *SYT11* locus on chromosome 1.



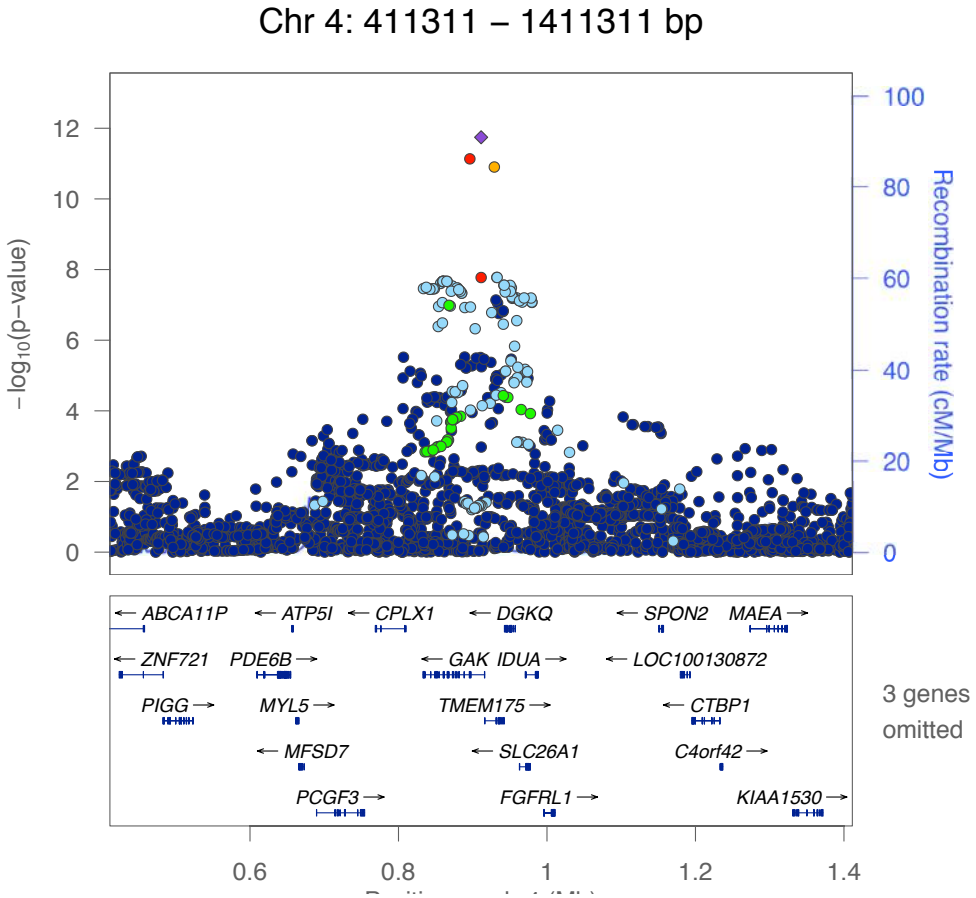
Supplementary Figure 2: Locus plot for the *ACMSD* locus on chromosome 2



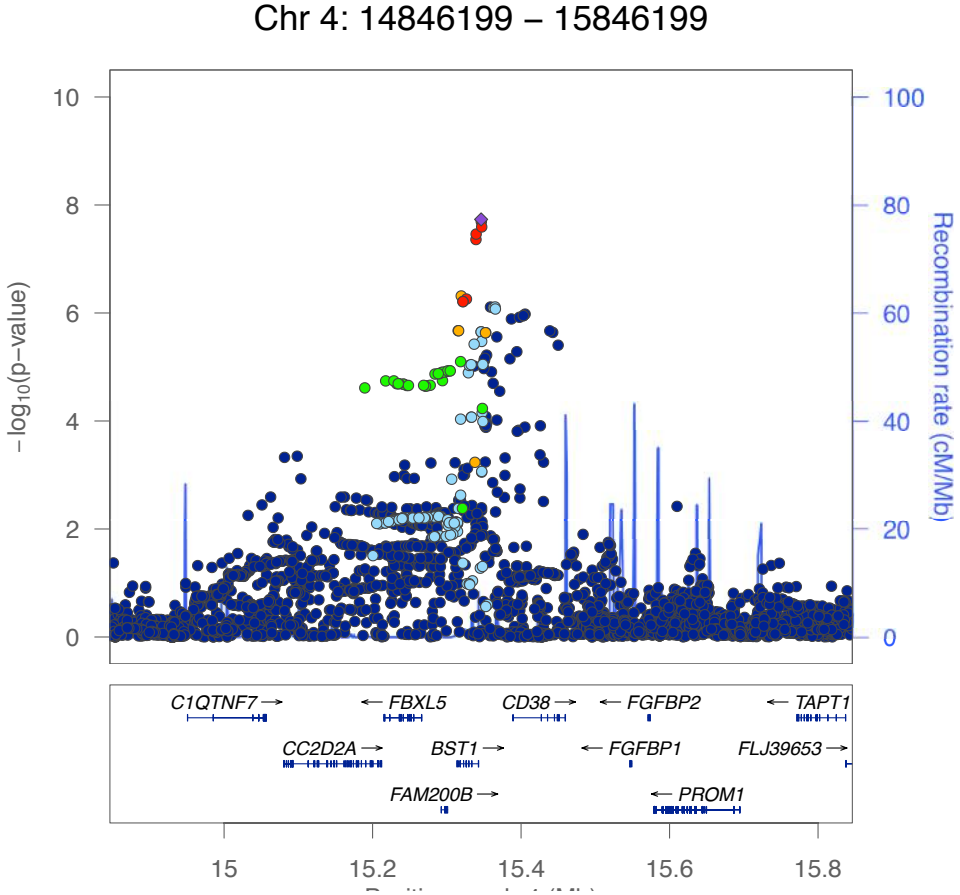
Supplementary Figure 3: Locus plot for the *STK39* locus on chromosome 2



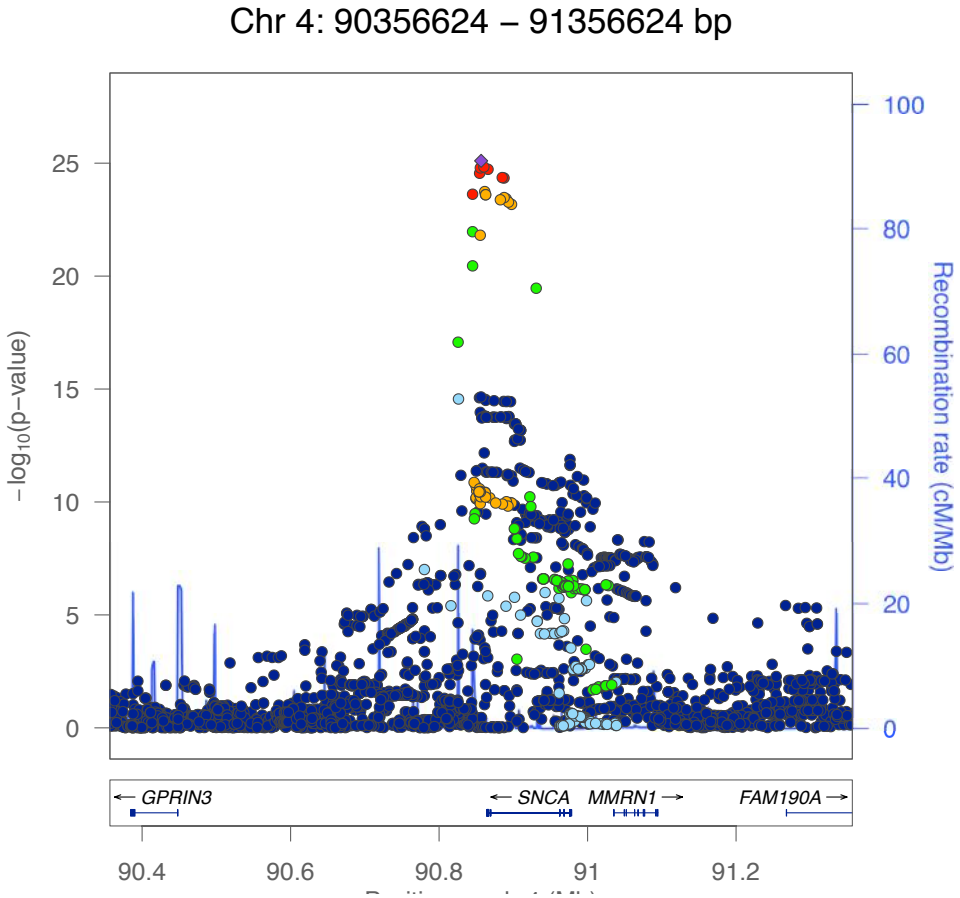
Supplementary Figure 4: Locus plot for the *MCCC1* locus on chromosome 3



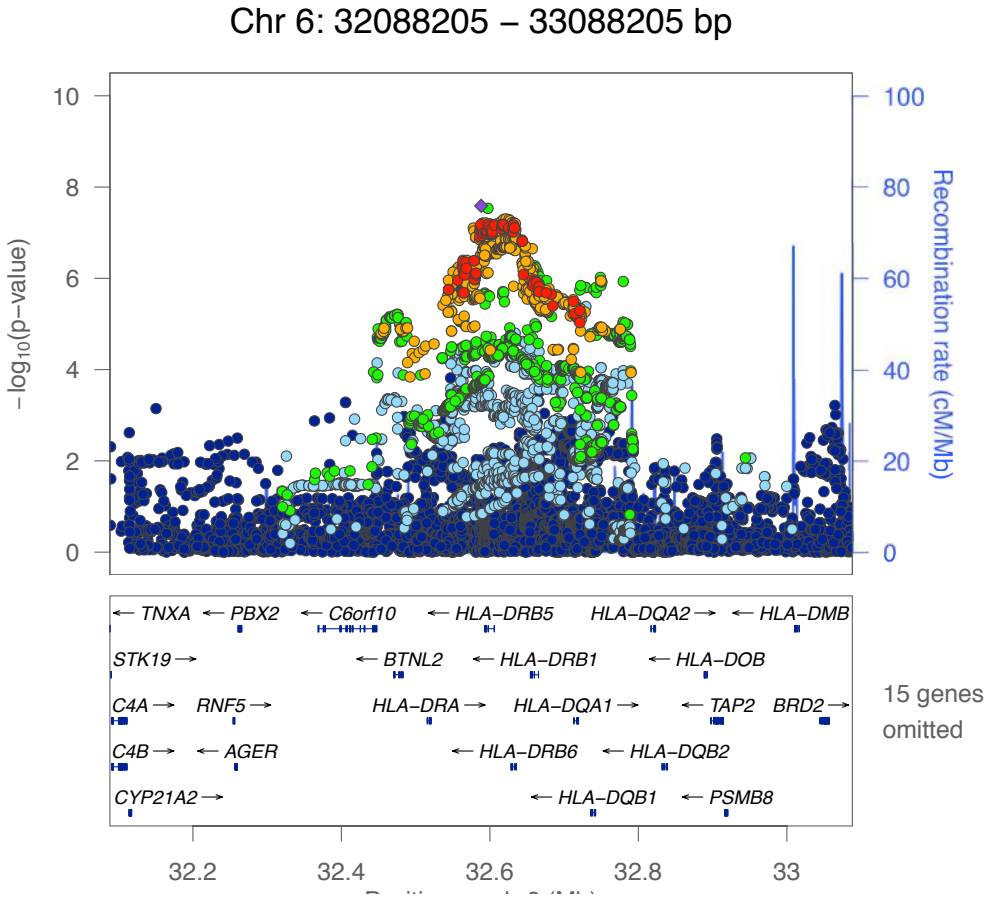
Supplementary Figure 5: Locus plot for the *GAK/DGKQ* locus on chromosome 4



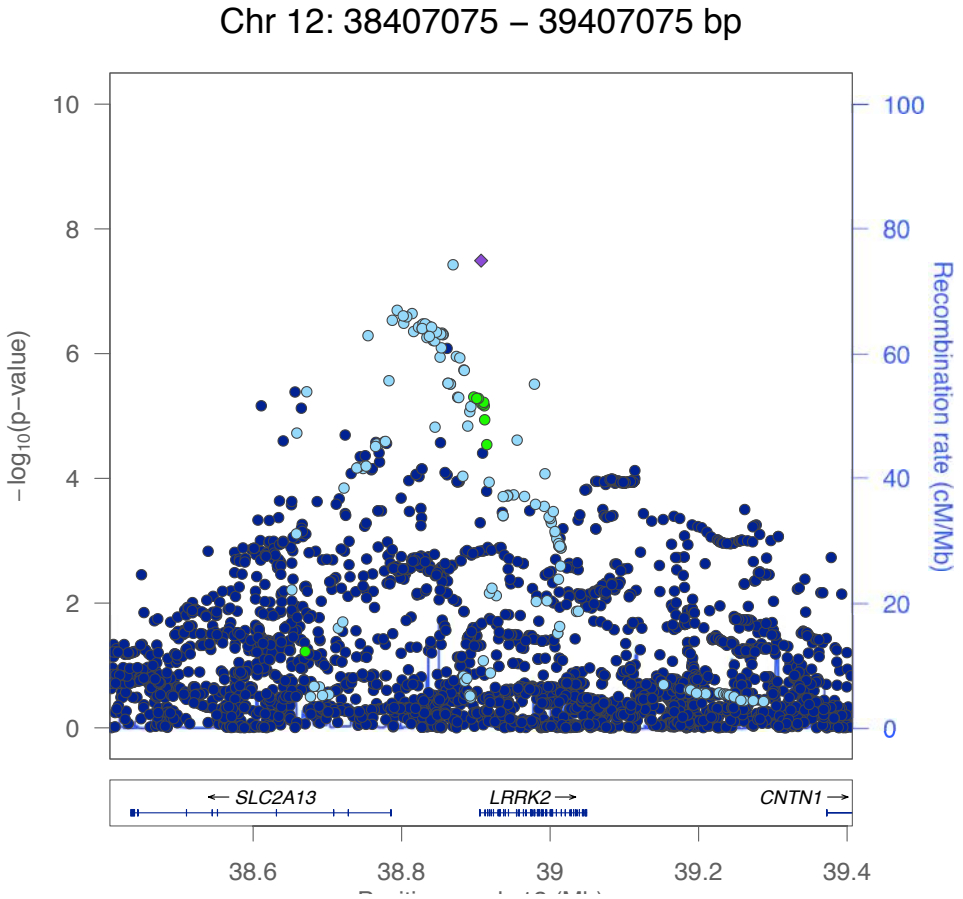
Supplementary Figure 6: Locus plot for the *BST1* locus on chromosome 4



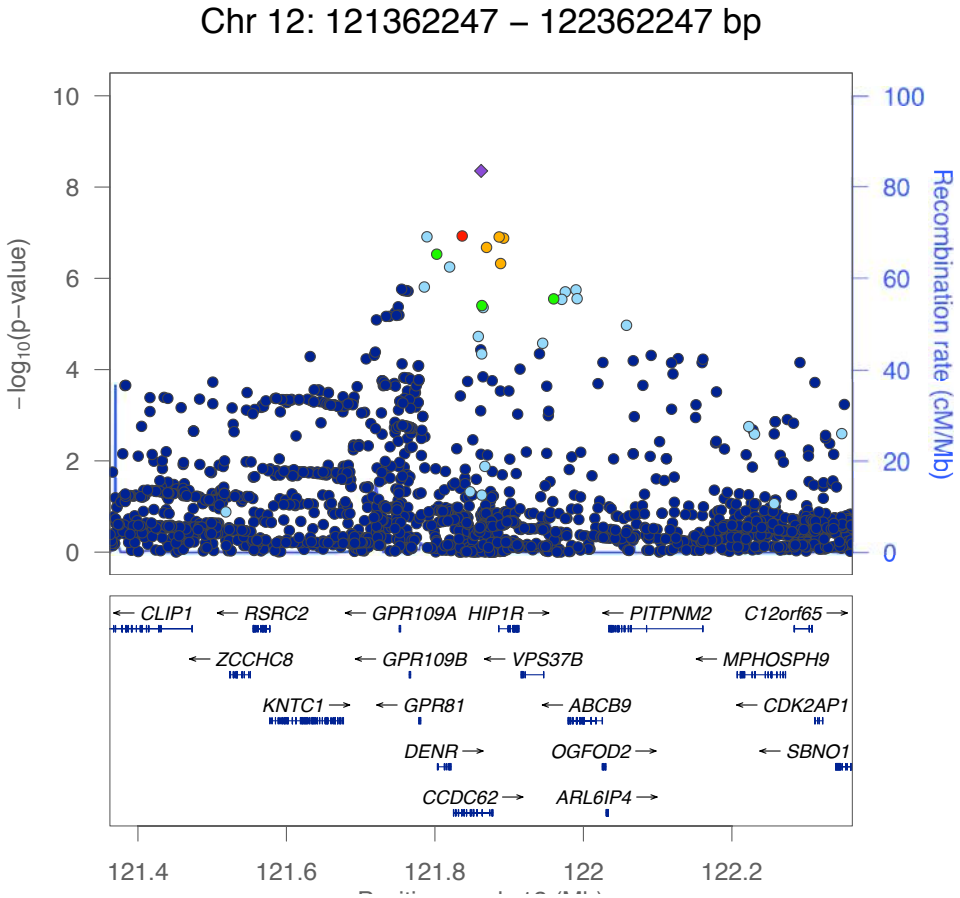
Supplementary Figure 7: Locus plot for the *SNCA* locus on chromosome 4



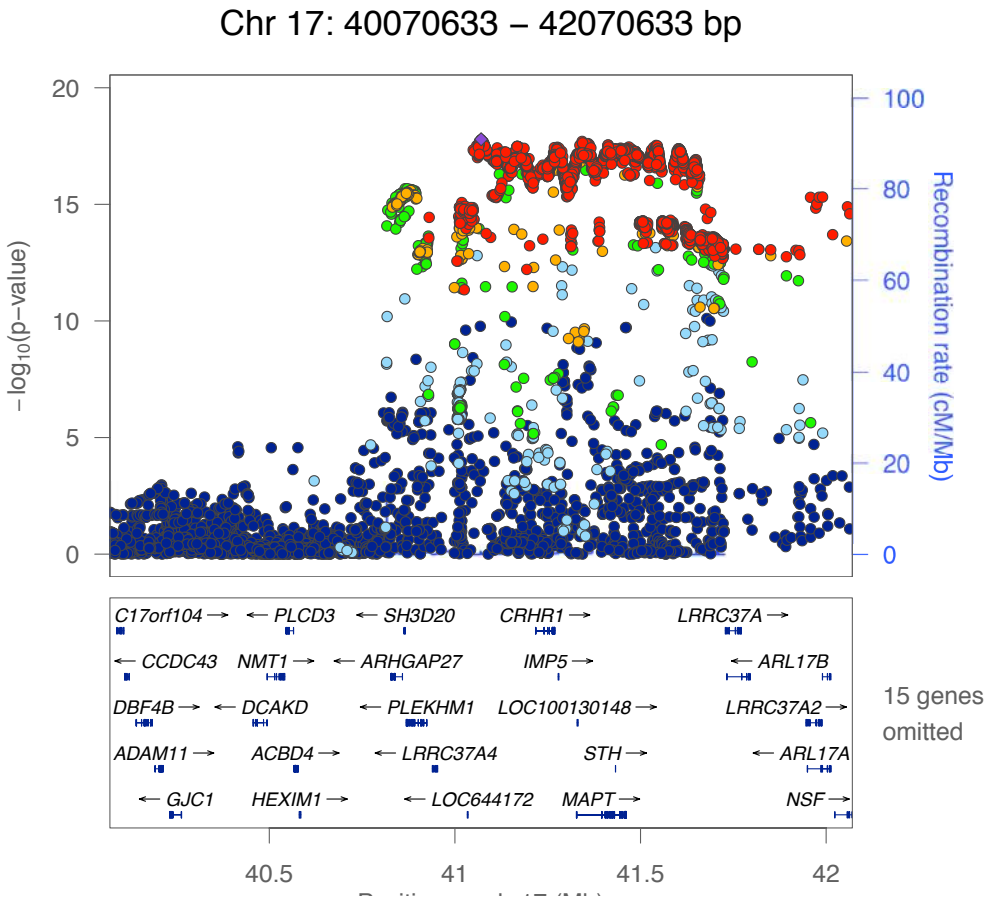
Supplementary Figure 8: Locus plot for the *HLA-DRB* locus on chromosome 6



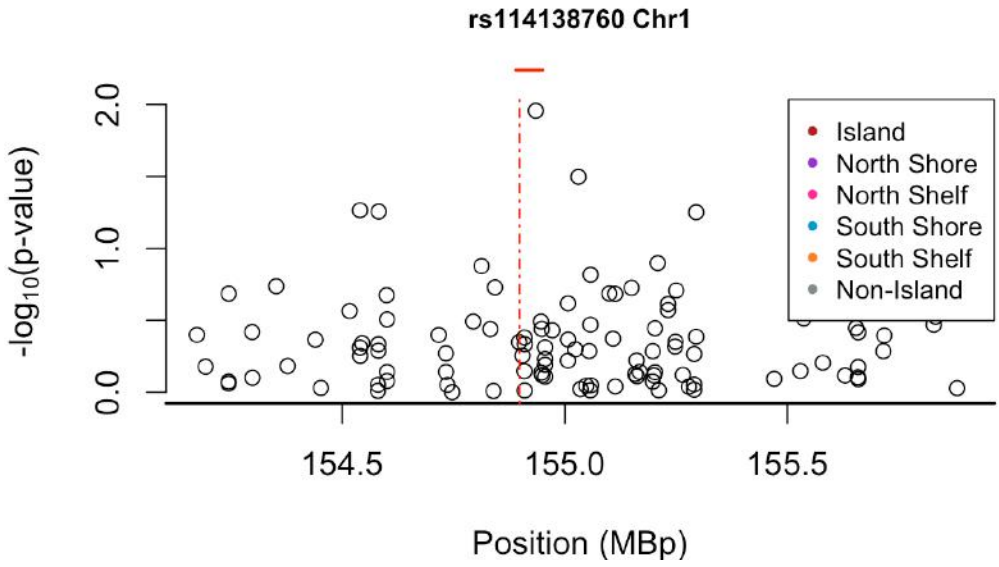
Supplementary Figure 9: Locus plot for the *LRRK2* locus on chromosome 12



Supplementary Figure 10: Locus plot for the *HIP1R* locus on chromosome 12

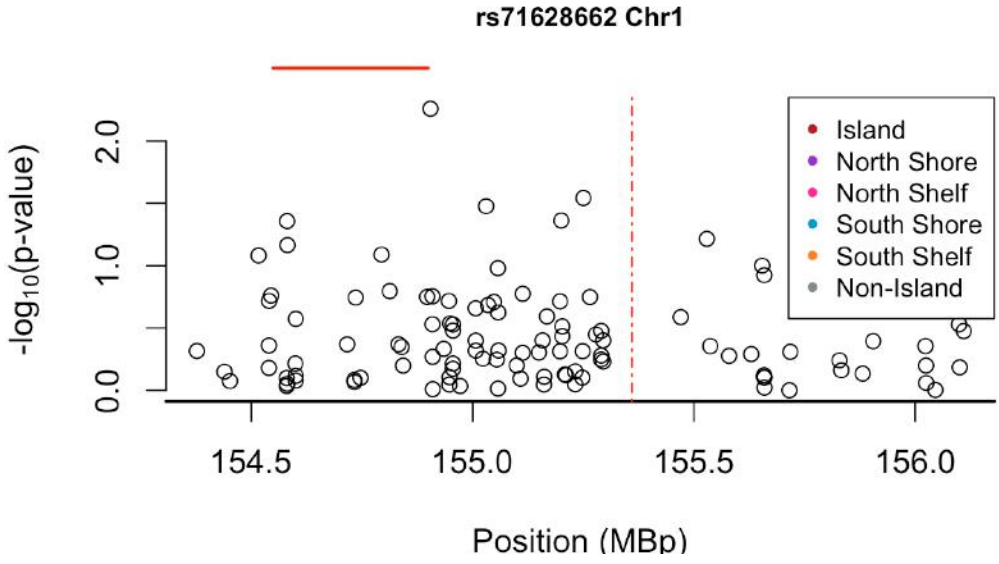


Supplementary Figure 11: Locus plot for the *MAPT* locus on chromosome 17



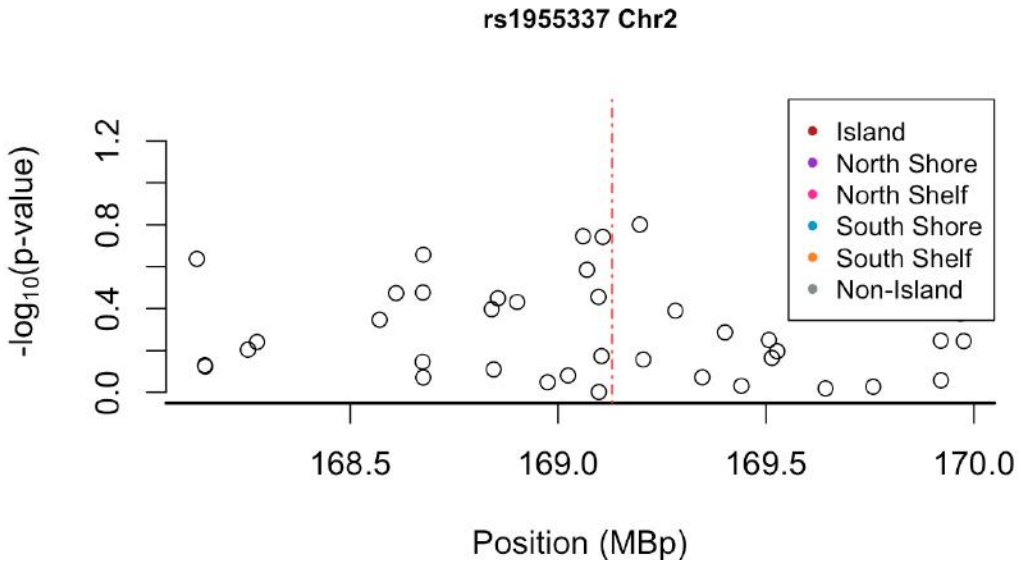
Supplementary Figure 12: PD risk allele rs114138760 on chromosome 1 does not significantly associate with DNA methylation.

Allele burden at rs114138760 was assessed for association with DNA methylation levels at proximal CpG sites. Non-associated CpG sites are shown as open symbols



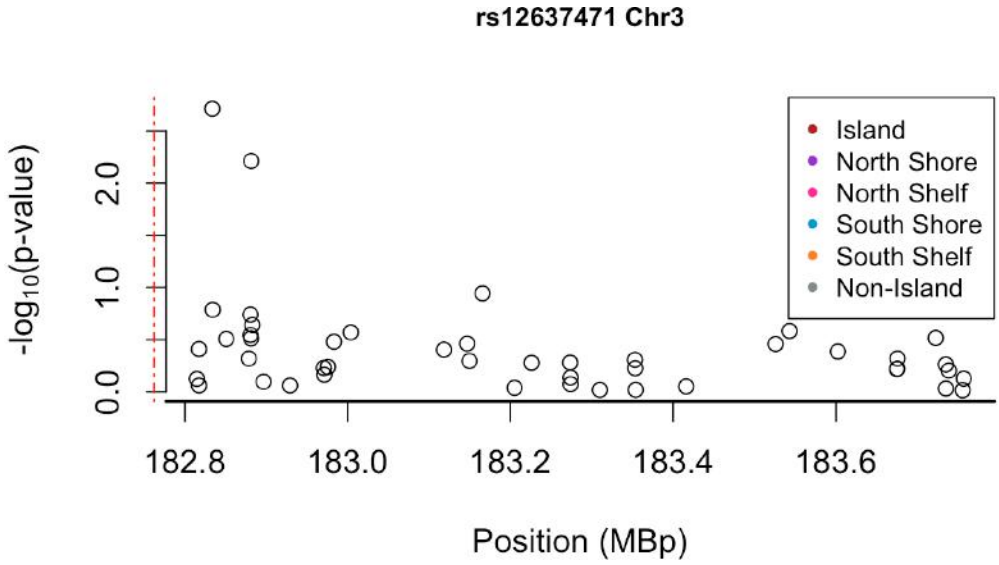
Supplementary Figure 13: PD risk allele rs71628662 on chromosome 1 does not significantly associate with DNA methylation.

Allele burden at rs71628662 was assessed for association with DNA methylation levels at proximal CpG sites. Non-associated CpG sites are shown as open symbols



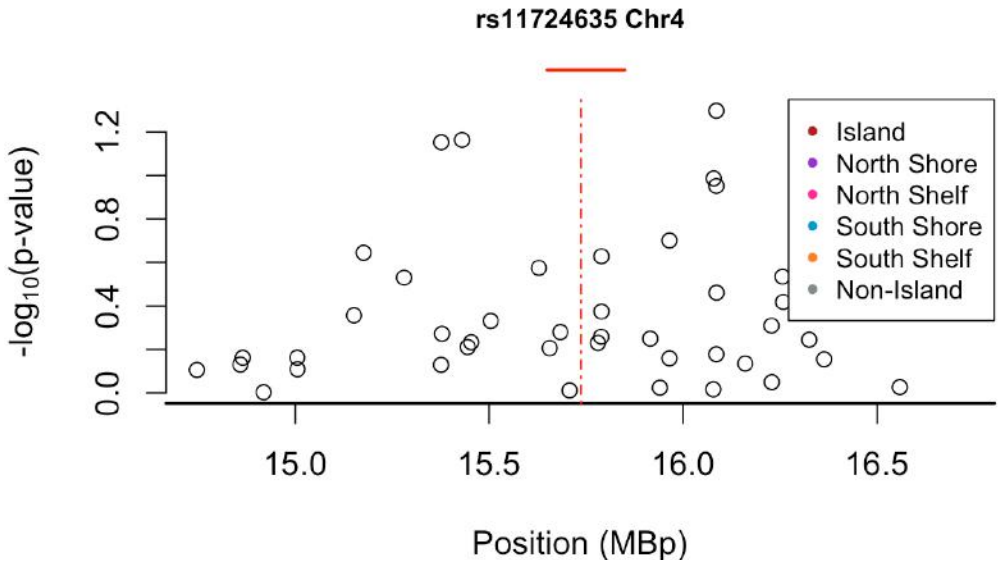
Supplementary Figure 14: PD risk allele rs1955337 on chromosome 2 does not significantly associate with DNA methylation.

Allele burden at rs1955337 was assessed for association with DNA methylation levels at proximal CpG sites. Non-associated CpG sites are shown as open symbols



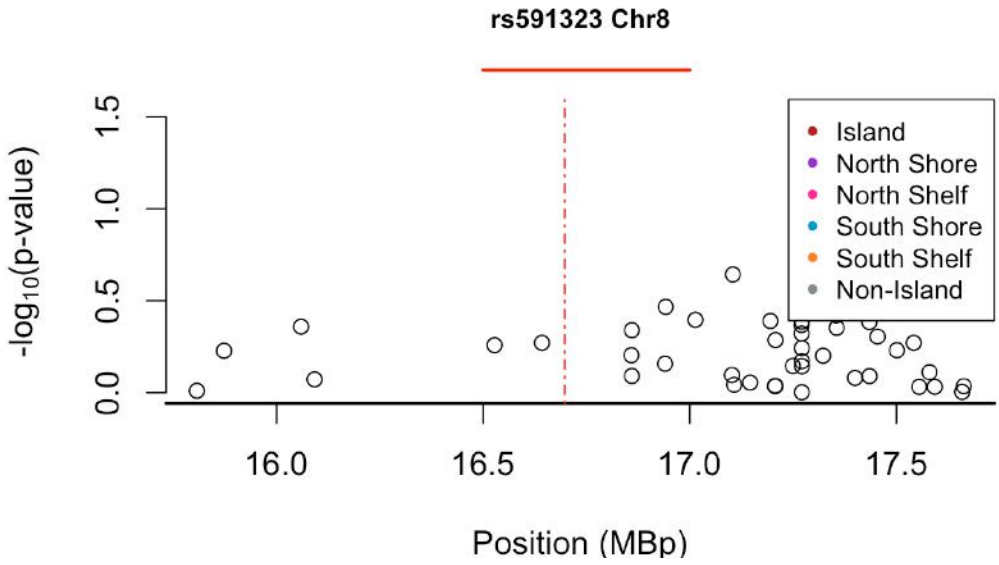
Supplementary Figure 15: PD risk allele rs12637471 on chromosome 3 does not significantly associate with DNA methylation.

Allele burden at rs12637471 was assessed for association with DNA methylation levels at proximal CpG sites. Non-associated CpG sites are shown as open symbols



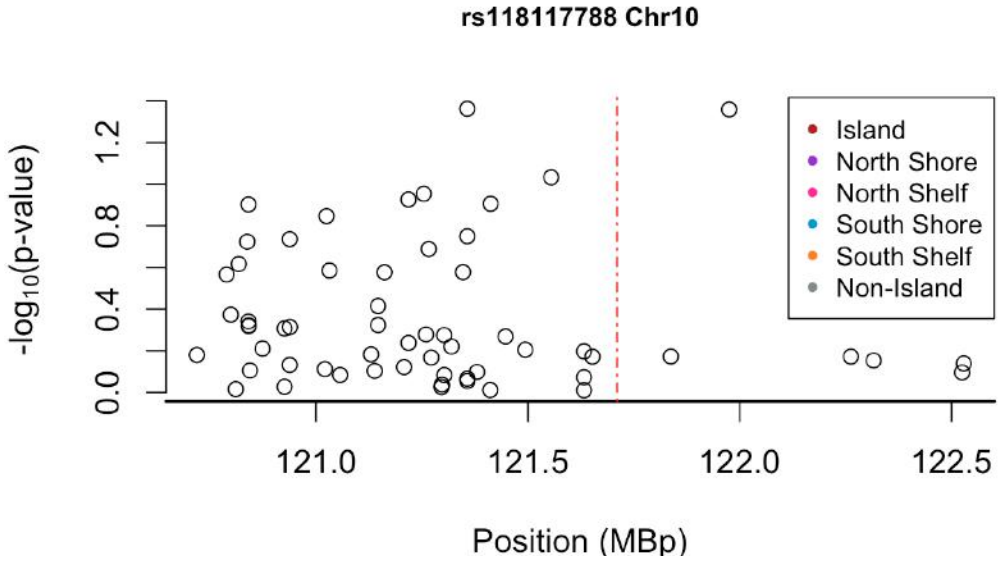
Supplementary Figure 16: PD risk allele rs11724635 on chromosome 4 does not significantly associate with DNA methylation.

Allele burden at rs11724635 was assessed for association with DNA methylation levels at proximal CpG sites. Non-associated CpG sites are shown as open symbols



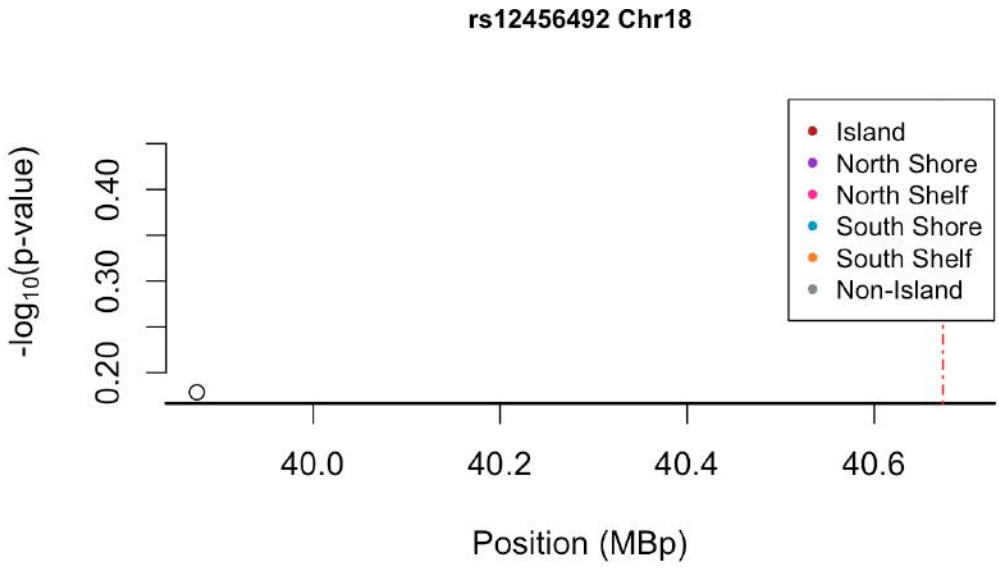
Supplementary Figure 17: PD risk allele rs591323 on chromosome 8 does not significantly associate with DNA methylation.

Allele burden at rs591323 was assessed for association with DNA methylation levels at proximal CpG sites. Non-associated CpG sites are shown as open symbols



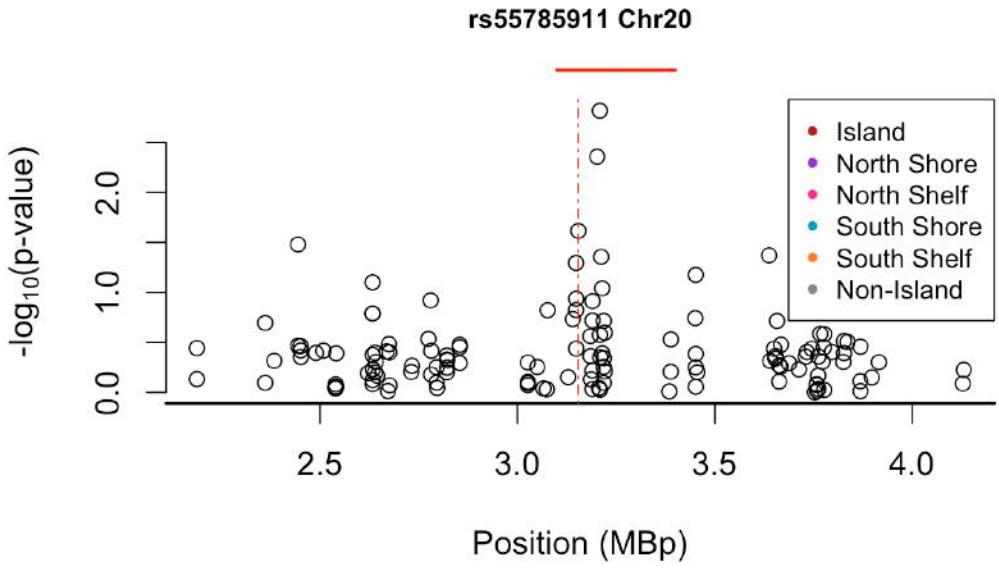
Supplementary Figure 18: PD risk allele rs118117788 on chromosome 10 does not significantly associate with DNA methylation.

Allele burden at rs118117788 was assessed for association with DNA methylation levels at proximal CpG sites. Non-associated CpG sites are shown as open symbols



Supplementary Figure 19: PD risk allele rs12456492 on chromosome 18 does not significantly associate with DNA methylation.

Allele burden at rs12456492 was assessed for association with DNA methylation levels at proximal CpG sites. Non-associated CpG sites are shown as open symbols



Supplementary Figure 20: PD risk allele rs55785911 on chromosome 20 does not significantly associate with DNA methylation.

Allele burden at rs55785911 was assessed for association with DNA methylation levels at proximal CpG sites. Non-associated CpG sites are shown as open symbols