Inheritance of DNA Methylation Level in Healthy Human Tissues

Amy E. Rowlatt, BSc., MSc.

PhD – The University of Edinburgh - 2015

Declaration

(a) This thesis has been completed by me

(b) The work is my own except where indicated at the end of each chapter

(c) The work has not been submitted for any other degree or professional qualification

……………………………

Amy Rowlatt 20th August 2015

Acknowledgements

I thank my supervisor Albert Tenesa for proposing the project, obtaining funding and his overall supervision. I thank my second supervisor, Chris Haley, for his guidance throughout. I thank the members of Albert Tenesa's group for their support and feedback pertaining to research presented during group meetings. In addition, I thank Isobel Stewart and Konrad Rawlik for answering helpful questions. I thank the members of my thesis committee, Bernadette Dutia and Kim Summers for their encouragement, support and insight during yearly review meetings. I thank Peter Joshi and Ian White for discussions regarding statistics. I thank Andy Law for help with Perl. I thank James Prendergast for insightful comments on a copy of a manuscript containing the research presented in chapter 4.

I would also like to thank our collaborators', Gustavo Hernandez-Sanchez, Maria Carolina Sanabria and Martha Serrano-Lopez, Jose Luis Soto, Adela Castillejo, Cristina Alenda and Eva Hernandez-Illan who collected samples and provided me with the colon dataset. I would like to thank Suzanne Rowe and our colleagues [1] for their collaboration and for providing me with the fluid intelligence dataset.

In addition, I would like to give recognition to those researchers' that made publically available the data they had collected, Gibbs et al. [2]. We thank them and the Division of Aging Biology, the Division of Geriatrics and Clinical Gerontology (NIA) and the NCBI Data Repositories dbGaP (Accession Number: phs000249.v1.p1) and GEO (GSE15745) for the availability of the brain dataset.

Finally, I thank The Roslin Institute and the BBSRC for funding my project.

Contents

Abstract

DNA methylation (DNAm) is the covalent modification of DNA by addition of a methyl group primarily at the cytosine directly upstream of a guanine. DNAm level plays a central role in transcriptional regulation and is linked to disease. Therefore, understanding genetic and environmental influences on DNAm level in healthy tissue is an important step in the elucidation of trait and disease etiology. However, at present only a minority of easy to access human tissues and ethnicities have been investigated.

Therefore, we studied DNAm level measured in five human tissues: cerebellum, frontal cortex, pons, temporal cortex and colon in either North American or South American samples. We applied a novel statistical approach to estimate the heritability attributable to genomic regions (regional heritability, $\hat{h}_{r,g}^2$) for DNAm level at thousands of individual DNAm sites genome-wide.

In all five tissues, DNAm level was significantly associated with the local genomic region for more DNAm sites than expected by chance. Moreover, DNAm level could be predicted from the local sequence variants with an accuracy that scaled with the estimated $\hat{h}_{r,g}^2$.

Our results inform on molecular mechanisms regulating DNAm level and trait etiology in several ways. Firstly, DNAm level at DNAm sites located in genomic risk regions and measured in a tissue relevant to the disease can be influenced by the local genetic variants. Specifically, we found that genetic variation within a region associated with Fluid Intelligence was also associated with local DNAm level at the proline-rich coiled-coil 1 (PRRC1) gene in healthy temporal cortex tissue. Additionally, we replicated the finding of a Colorectal Cancer risk variant (rs4925386) associated with two DNAm sites in healthy colon tissue. More generally, we showed that DNAm sites located within a susceptibility region and measured in a relevant tissue exhibit a similar overall pattern of estimated $\hat{h}_{r,g}^2$ to DNAm sites outwith a susceptibility region. Secondly, the propensity for DNAm level to be associated with the local sequence variation differs with respect to CpG dinucleotide density and genic location. Most notably, DNAm sites located in CpG dense regions of the genome are less likely to be heritable than DNAm sites located in CpG sparse regions of the genome. Additionally, within both CpG dense and CpG sparse regions of the genome intergenic DNAm sites are more likely to be heritable than intragenic DNAm sites. Overall, our study suggests that variation in DNAm level at some DNAm sites is at least partially controlled by nuclear genetic variation. Moreover, DNAm level in healthy tissue has the potential to act as an intermediary in trait variation and etiology.

Lay Summary

The DNA sequence encodes the instructions that give rise to life. Extensive research has shown that some differences in the DNA sequence between people can influence observable characteristics such as susceptibility to disease. Recently studies have revealed that there are DNA modifications, known as epigenetic marks, which do not alter the sequence but can affect the way in which it is decoded. One such epigenetic mark is DNA methylation (DNAm). Differences in DNAm level between people within a tissue has also been linked to disease susceptibility. However, what is less known is the extent to which DNAm in healthy tissue is influenced by nature (inherited genetic variation) and by nurture (the environment).

Therefore, we studied DNAm level measured in five human tissues: cerebellum, frontal cortex, pons, temporal cortex and colon in either North American or South American people. We used a novel statistical approach to determine the degree that DNAm level at thousands of individual locations in the DNA sequence is affected by genetic variation.

We found that in all tissues studied, DNAm level at some locations is at least partially controlled by the surrounding genetic variation. Therefore, in some cases DNAm level can be inherited across generations. In addition we obtained specific examples where DNAm level in a relevant tissue is affected by the same genetic variation that influences disease. Together these results indicate that DNAm level in healthy tissue has the potential to act as an intermediary between genetic variation and disease. Our results have implications for studying how disease develops.

List of Abbreviations

BMI	Body Mass Index
BR	Brain Region
CP	Cellular Phenotype
CRBL	Cerebellum
CRC	Colorectal Cancer
CV	Coefficient of Variation
DNMT	DNA Methyltransferase
DNAm	DNA Methylation
DZ	Dizygotic
eQTL	Expression Quantitative Trait Loci
EWAS	Epigenome-wide Association Study
FCTX	Frontal Cortex
FI	Fluid Intelligence
GRM	Genetic Relationship Matrix
GWAS	Genome-wide Association Study
HM27K	Illumina Infinium HumanMethylation27 BeadChip
HM450K	Illumina Infinium HumanMethylation450 BeadChip
ICR	Imprinting Control Region
LD	Linkage Disequilibrium
LCM	Laser Capture Microdissection
MAF	Minor Allele Frequency
MZ	Monozygotic
methQTL	Methylation Quantitative Trait Loci
PMI	Post-mortem Interval
PONS	Pons
QTL	Quantitative Trait Loci
REML	Restricted Maximum Likelihood
RH	Regional Heritability
SNP	Single Nucleotide Polymorphism
TCTX	Temporal Cortex
TF	Transcription Factor
TSS	Transcription Start Site
UP	Ultimate Phenotype
WCB	Whole Colorectal Biopsy

Chapter 1 Introduction

1.1 The Discovery of 5-methylcytosine

5-methylcytosine is formed by the biochemical process of the addition of a methyl group (CH_3) at the 5th position of the atom ring that constitutes the pyrimidine base, cytosine. At the turn of the twentieth century experiments in organic chemistry led to the synthesis of 5-methylcytosine in the laboratory [3]. Scientists seeking to understand the nature of nucleotide bases knew that hydrolysis of cytosine led to the removal of an amine group (deamination) and the formation of uracil, a constituent of nucleic acid. At this time, thymine had also been realized as a constituent of nucleic acid and chemists hypothesized that a pyrimidine base with that of the chemical formula of 5-methylcytosine could deaminate and lead to the formation of thymine. Despite success in synthesizing and exploring the chemical properties of 5-methylcytosine a further 21 years passed before 5-methylcytosine was discovered in the nucleic acid of a living organism; *Mycobacterium tuberculosis* [4]. By adding picric acid to a filtrate containing cytosine obtained from the nucleic acid of *Mycobacterium tuberculosis*, 5-methycytosine could be separated from cytosine as salt crystals with observable differences [4].

1.2 Current Techniques for Interrogation of 5-methycytosine

With the dawn of the 21st century came the development of assays that could interrogate cytosine for methylation at basepair resolution, genome-wide. These assays utilize traditional molecular techniques for assaying 5-methylcytosine such as immunoprecipitation, methyl-sensitive restriction enzymes or sodium bisulfite conversion, coupled with either the latest sequencing or microarray technology (reviewed in [5]). These methods, of which there are many variants, all treat DNA prior to amplification or hybridization. This is a necessary requirement because 5-

methylcytosine is not maintained throughout polymerase chain reaction and cannot be identified from hybridization alone due to the location of the methyl group with respect to the DNA structure (reviewed in [6]). Following treatment of the DNA, microarray based methods are efficient in assaying large numbers of samples while whole genome sequencing remains a more costly gold standard.

The treatment of DNA with sodium bisulfite is a popular approach taken by researchers' to interrogate cytosine for methylation, genome-wide. In 1992, a seminal paper revealed the utility of sodium bisulfite for measuring 5-methylcytosine [7]. In the presence of sodium bisulfite, unmethylated cytosine deaminated to uracil whereas methylated cytosine remained essentially unreactive. This changed the epigenetic modification into a genetic modification amenable to established assays for genetic polymorphism. However, both 5-methylcytosine and the oxidation product, 5-hydroxymethylcytosine, do not react to treatment with sodium bisulfite; therefore, this assay cannot distinguish between these two forms of methylated cytosine (reviewed in [8]). Methylation of cytosine is a binary outcome within a cell at a given genomic location on one strand of DNA. However, two double strands of DNA within a cell, input of multiple cells within a sample, and polymerase chain reaction following bisulphite treatment, means that the occurrence of 5-methylcytosine is reported as a variable indicating a level of 5-methylcytosine. This variable is often the proportion of methylated to total cytosine bases at a genomic location (Beta-value) or a transformation of that proportion (M-value) [9].

1.3 Biological Roles and Importance of 5-methylcytosine

5-methylcytosine is a reversible mark that can affect how the DNA sequence is decoded. DNA methyltransferases (Table 1) and agents that lead to demethylation of 5-methylcytosine (Table 2) respectively add and remove 5-methylcytosine at sites across the genome. A DNA methyltransferase (DNMT) acts by covalently bonding with cytosine to create a negative charge that attracts a CH_3 group from the methyl donor, 5-adenosyl-L-methionine. After the CH_3 bonds to the cytosine the DNMT is released [10]. The process of demethylation is less well characterized than the process of methylation; however, several mechanisms have been identified (Table 2).

Table 1 DNA Methyltransferases

	Function	Reference
DNMT1	Maintenance across mitosis	[11-13]
DNMT2	Non CpG 5-methylcytosine	Reviewed in [14]
DNMT3A	De novo (imprinting) De novo (development)	[15-18]
DNMT3B	De novo (development)	[15]
DNMT3L	De novo (Imprinting)	[16-18]

Table 2 Agents of Demethylation

Mechanism	Description	Examples	Reference
Oxidation	To 5-hydroxymethylcytosine (and other products sequentially) by Tet dioxygenases	Paternal genome, preimplantation embryo	Reviewed in [19]
Base Excision Repair and DNA Glycosylases	Bases mismatches resulting from deamination of 5-methylcytosine or oxidized product are targeted for removal	Imprinted loci in Primordial Germ Cells	Reviewed in [19,20]
Methyl Binding Proteins	Direct	MBD2	Reviewed in [14]
Replication Dependent	No maintenance throughout mitosis	Maternal genome, preimplantation embryo	Reviewed in [19]

5-methylcytosine has a functional role in several crucial biological processes including regulation of gene expression level. As early as 1975 Holiday and Pugh hypothesized that 5-methylcytosine could be responsible for the changes in the expression of developmental genes in a time dependent manner that led to cell differentiation [21]. Since this hypothesis, experiments have confirmed that 5-methylcytosine and chromatin modifications are associated with the transition of cells from a pluripotent to unipotent state [8,22]. For instance, a large increase in global 5-methylcytosine and methylation of H3K9me3 by G9a has been observed at the first differentiation event in embryogenesis when genes that maintain pluripotency are down regulated [8,22]. Additionally, 5-methylcytosine at imprinting control regions (ICR) has been shown to regulate gene expression level in a parent of origin manner. This means that 5-methylcytosine, which occurs on either the paternal or maternal haplotype at an ICR, leads to mono-allelic expression at the locus (reviewed in [23]). At fertilization the paternal and maternal DNA undergo extensive chromatin remodelling and epigenetic changes. These epigenetic changes include a global loss of 5-methylcytosine by embryonic day 4 (reviewed in [19]). However, at a minority of sites namely in imprinted loci, 5-methylcytosine escapes demethylation. In human germ cells, parental imprints are erased in a second wave of demethylation occurring between embryonic day 7 and embryonic day 13.5. Research showed that the reestablishment of methylation at imprinted loci in female mice germ cells occurred after birth and was dependent on the developmental stage of the oocyte [24]. In contrast, in male mice, reestablishment of methylation at imprinted loci began at embryonic day 14.5 and was almost complete by embryonic day 17.5 [25].

Despite the important changes in 5-methylcytosine that occur during development, after development 5-methylcytosine is stably inherited throughout mitotic cell divisions. Studies have assayed the human methylome in differentiated cells. These studies provided several results 1) 5-methylcytosine occurred almost exclusively at CpG dinucleotides in differentiated cells and on both the forward and reverse strand at complementary CpG dinucleotides [26]. 2) level of 5-methylcytosine was highly correlated for CpG sites located up to 1kb apart [27]. 3) Average level of 5-methylcytosine within amplicons was bimodal with the majority of amplicons exhibiting either low or high levels of 5-methylcytosine [27,28]. 4) Level of 5-methylcytosine varied across different genomic locations. Specifically, CpG sites within transcription start sites (TSS) showed lower levels of 5-methylcytosine than intragenic CpG sites [27,28] and centromeric CpG sites showed lower levels of 5-methylcytosine than CpG sites located close to the telomeres [26]. 5) Different tissues exhibited a different pattern of level of 5-methylcytosine across the genome [26-29]. 6) In some cases, level of 5-methylcytosine associated with gene expression level [30,31]. Together these studies highlighted that the level of 5-methylcytosine is extensive and variable throughout the genome and across cell types. Moreover, level of 5-methylcytosine and level of gene expression are related at some loci in differentiated somatic cells.

Aberrant 5-methylcytosine can be characteristic of disease. Several rare imprinting disorders can result from an unusual pattern of 5-methylcytosine. For instance, research showed that 2-5% of cases of Angelman syndrome and Prader-Willi syndrome resulted from aberrant 5-methylcytosine on the respective maternal and paternal allele at 15q11-13 (reviewed in [32]). In addition, defective 5-methylcytosine at 11p15.5 led respectively to 40% and 40-60% of cases of Silver Russel Syndrome and Beckwith-Wiedemann syndrome (reviewed in [32]). Changes in 5-methylcytosine are linked to common disease such as cancer. Global hypomethylation in tumour cells has been associated with decreased genomic stability and changes in level of gene expression (reviewed in [33]). In contrast global hypermethylation has been linked to the silencing of genes that are important for normal cell maintenance and tumour suppression (reviewed in [33]). Moreover, level of 5-methylcytosine at a specific locus can be used as an indicator (biomarker) of cancer (reviewed in [34]). Epigenome-wide association studies (EWAS) are now a popular method for assessing how changes in 5-methylcytosine relate to variation in a common phenotype. An EWAS tests the association of a phenotype with level of 5-methylcytosine at many individual sites of 5-methylcytosine across the genome.

EWASs have been conducted for a number of common diseases including systemic lupus, major psychosis and type 1 diabetes nephropathy. These studies identified small changes in 5-methylcytosine that were associated with changes in disease risk (reviewed in [35]). The role of 5-methylcytosine in disease makes it of interest to clinical and human geneticists. Moreover, now that it is feasible to assay 5-methylcytosine at hundreds of thousands of individual sites genome-wide, human quantitative geneticists can study phenotypic variation in 5-methylcytosine in healthy tissue.

1.4 Quantitative Genetics

Quantitative genetics is the measurement and classification of the inherited variation that comprises variation in complex phenotypes (or traits), which are phenotypes that are influenced by multiple genes [36]. Variance in a phenotype can be classified into genetic and environmental variance (Equation 1) [37]. The genetic variance is a measure of the deviation of genotypic effect from the population mean. The genotypic effect for an individual in the population is the total effect of their genotype on the phenotype. The genetic variance can be further decomposed into additive variance and non-additive variance (Equation 2) [37]. Additive genetic variance arises from variation in breeding value within a population. The breeding value of an individual within a population is a measure of their genetic effect that can be inherited by their offspring. The environmental variance can be further classified into general and special environmental variance (Equation 2) [37]. General environmental variance arises from variation in the effect of an environmental factor that is common to individuals within the population. Special environmental variance results from variation in environmental effects that arise from environmental factors unique to the individual. The emergence of quantitative genetics can be traced back to the turn of the twentieth centenary. In 1918, building on the work of Darwin, Mendel, Galton and Pearson, Fisher hypothesized that many genes each of small effect and of Mendelian inheritance lead to the heritable variation observed for normally distributed traits [38].

Equation 1

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2$$

Equation 2

$$\sigma_P^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2 + \sigma_{EG}^2 + \sigma_{ES}^2$$

Human quantitative genetics, as the name suggests, is the study of quantitative genetics in human populations. The development of sequencing and array technology, at the turn of the twenty first century, has enabled the discovery of a vast number of genetic variants (or Single Nucleotide Polymorphisms, SNPs) in human populations. For instance, in 1990 the landmark U.S Human Genome Project began [39]. The Human Genome Organization included over 20 groups worldwide that over a time period of 13 years, worked to develop efficient strategies to sequence the human genome. The draft sequence from the U.S Human Genome Project was published in 2001 [39], and the completed sequence (99% of the genome) uncovering more than 1.4 million SNPs followed in 2004 [40]. Following the U.S Human Genome Project came studies seeking to assess the extent of variation within multiple different human populations, such studies include the International HapMap project [41] and The 1000 Genomes Project [42]. These studies and smaller scale studies with similar goals (of which there are too many to reference) have generated large amounts of genetic data. This has led to the establishment of public on-line data repositories for storing genomic data. The ability to download genomic data and assay genetic variation in a high throughput manner has revolutionized human quantitative genetics. With genotypic data readily attainable human quantitative geneticists can now use novel methods to explore the genetic architecture of complex traits.

1.5 Understanding the Genetic Contribution to Variation in Complex Quantitative Traits

One fundamental focus for human quantitative genetics is estimation of the heritability of a phenotype. The broad sense heritability is the proportion of the phenotypic variance that can be explained by variance in the total genetic effect (Equation 3) [37]. In contrast, the narrow sense heritability is the proportion of the phenotypic variance explained by variance in additive genetic effects (Equation 4) [37]. Several methods have been used to estimate the heritability and the method used can influence the estimation. Therefore, results from different study designs must be interpreted differently. Common and traditional methods involve using measurements of the phenotype and pedigree relationship among relatives to obtain an estimate of heritability due to the pedigree relationship ($\hat{h}^2_{ped}$). These pedigree based studies capture the full extent of the additive genetic heritability but can be biased by un-modelled sources of variation [37]. For instance, the offspring phenotypic value can be regressed on the phenotypic value of one of the parents to provide an estimate equal to half of the heritability [37]. However, general environmental effects can make parents and offspring appear more similar to one another than expected due to the sharing of alleles. This scenario would bias the heritability estimate upwards. As a second example, human studies often use monozygotic (MZ) and dizygotic (DZ) twins to estimate heritability. Twice the difference of correlation of MZ twin pairs minus the correlation of DZ twin pairs provides an estimate of heritability. However, this estimate of the heritability includes genetic variance arising from non-additive effects [37]. Moreover, it assumes that the general environmental variance is equal for MZ and DZ sib pairs. Un-modelled sources of environmental variance that result in MZ twins being more similar to one another than DZ twins will bias the heritability estimate upwards. Overall, setting aside sampling variance of individual experiments, pedigree based methods to estimate the heritability can be thought of as providing an upper limit for the additive genetic heritability.

Equation 3

$$H^2 = \frac{\sigma_G^2}{\sigma_P^2}$$

Equation 4

$$h^2 = \frac{\sigma_A^2}{\sigma_P^2}$$

Within Human Quantitative Genetics a second question is to determine the causal effects acting on heritable phenotypes. One commonly used method to tackle this question, that has been enabled by the discovery and assay of abundant SNPs within the human genome, is a Genome-wide Association Study (GWAS, Single SNP Association Approach). A GWAS is a population-based method where SNPs across the genome are individually tested for correlation with a phenotype [43]. A SNP is used as a marker for causal variants with which it is in linkage disequilibrium (LD). This method estimates the effects of causal variation in LD with a marker SNP. GWAS have revealed important findings regarding the genetic architecture of common complex phenotypes. In some cases, such as with age related macular degeneration, several genetic variants have been discovered which have relatively large effects (reviewed in [44]). However, in most cases, common complex phenotypes appear to be influenced by many genetic variants of intermediate frequency and small effect. A classic example is human height where 423 causal loci have been identified [45] and together explain 0.16 of the phenotypic variance. Moreover, for the majority of traits, estimates of SNP effects from GWASs have fallen short of the heritability estimated from family based methods and this is called the “missing heritability” problem [46].

Several explanations have been proposed for the elusive heritability. Firstly, it is possible that family based estimates are inflated estimates of the narrow sense heritability. As discussed above, this can occur when factors such as shared environmental effects and non-additive genetic effects, including dominance, epistasis and genetic by environment interactions, are not accounted for and lead to phenotypic covariance among relatives [46]. While this may be the case for some phenotypes, it has been shown that is unlikely to be the case for human height [47]. Secondly, small effect sizes may not be detected at stringent significance thresholds even if the causal variant is in LD with the marker SNP [46]. This hypothesis has spurred researchers' to seek an increased sample size, as this will increase the power to detect variants of small effect. Thirdly, the effect of causal genetic variants not tagged by marker SNPs on genotyping arrays cannot be captured by GWAS [46]. This type of genetic variation includes both rare and common variants. The majority of marker SNPs on genotyping arrays are of intermediate frequency in the human population [48]. An allele of intermediate frequency cannot have a high correlation with an allele of low frequency. Thus a GWAS that uses markers of intermediate frequency does not effectively capture the phenotypic variance explained by rare genetic variants [46]. The effect of common variants can be missed due to sampling variation where by the causal variant is not tagged by a SNP on the genotyping array [49].

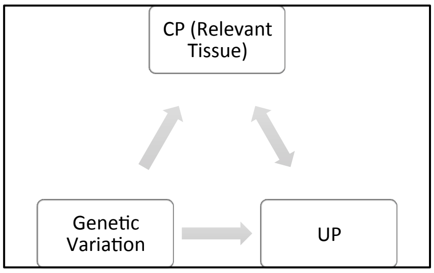
In 2010 a pivotal study [49] used human height as an example and provided evidence for both the second and third of the above three hypotheses. At this time, GWAS had uncovered ~50 genetic variants associated with human height and together these variants explained 0.05 of the phenotypic variation. In the 2010 study [49], random effects of 294,831 SNPs on height were fit simultaneously within a restricted maximum likelihood (REML) framework. The most likely estimate of the variance in the phenotype due to additive SNP effects ($\hat{h}_g^2$) was 0.45. Fitting SNPs simultaneously led to a 9-fold increase in the proportion of the phenotypic variance explained compared to that explained by the individual effects of the known associated variants. The authors' [49] concluded that this result came from variants with small effect that individually would not surpass the significance threshold. Therefore, the variants had not been detected in previous GWASs. When the SNP effects were added together they explained a substantial proportion of the phenotypic variance. The total proportion of the phenotypic variance explained by the SNPs could be detected by fitting the effects of SNPs simultaneously. The authors' [49] went on to show that even when fitting all SNPs simultaneously some causative variation could be missed. Firstly, positioning of the tagging SNPs with respect to the causative variants meant that some causative effects were missed and the magnitude of the discrepancy depended on the number of tagging SNPs. Secondly, the effects of causative variation with a minor allele frequency (MAF) < 0.1 were not captured as well as causative variation with a MAF < 0.5, when the tagging SNPs each have a MAF < 0.5. This indicated that if causative variants had a different allele frequency spectrum to that of tagging SNPs then their effects would not be captured. Indeed, when the authors' adjusted for insufficient LD between causative variants with MAF < 0.1 and tagging variants with MAF < 0.5 they found $\hat{h}_g^2$ of height to be 0.80 [49]. Overall, this study revealed that extensive common variation of small effect and rare variation could explain the missing heritability. Additionally, fitting the effects of SNPs simultaneously could capture common causative variation not identified by GWASs. However, the genomic heritability due to (common) tagging SNPs on an array may be lower than the additive genetic heritability. This is because the genomic heritability is unlikely to capture the effects of rare variants.

1.6 Challenges and Prospects for Human Quantitative Genetics

Current challenges for human quantitative genetics are finding the source of the missing heritability, identifying the genetic variants that underlie phenotypic variation and elucidating the relevant biological pathways. There are several prospects that could address the questions raised by these challenges. Firstly, insight could be gained from the development and application of novel statistical approaches to analyse genomic data. Secondly, the relationship between a phenotype (ultimate phenotype, UP), a cellular phenotype (CP) measured in a tissue relevant to the UP and genetic variation could be explored (Figure 1).

Figure 1 Possible Relationships Among Genetic Variation, A Cellular Phenotype and An Ultimate Phenotype.

Arrows indicate potential causal pathways.



A relatively well-characterized CP is gene expression level and several approaches have been taken to implicate genes that may be involved in the aetiology of an UP. Firstly, level of gene expression can be tested for association with an UP. This can be done in a high throughput manner. For instance, the expression levels of 23,720 genes measured in adipose tissue were tested for association with body mass index (BMI). A significant association was found for 72% of the genes tested and variation in gene expression level at 2,784 genes explained 0.10 of the variation in BMI. This suggested a relationship between a large number of genes in adipose tissue and BMI. Secondly, the comparison of SNPs associated with an UP (Quantitative Trait Loci, QTL) and SNPs associated with gene expression (expression Quantitative Trait Loci, eQTL) in a tissue relevant to the UP has been used to identify genes that may be involved in the aetiology of the UP. For instance, genetic variants associated with childhood asthma were also associated in *cis* with expression levels of ORMDL3 in lymphoblastoid cells [50]. This result elucidated a potential mechanism by which the genetic variants could influence childhood asthma. Similarly, SNPs associated with the expression of genes in the macrophage-enriched metabolic pathway measured in adipose tissue associated with BMI, which suggested a link between the group of genes and obesity [51]. Overall, these studies have suggested a link between an UP, the expression level at genes measured in a relevant tissue and in some cases, genetic variation. The results indicted that it was possible that the CP, gene expression level, could mediate genetic effects on an UP. However, it should be noted that these studies have not proved a causal relationship between gene expression level and the UP.

Information regarding the level of association between SNPs and a CP can be used to increase the power to detect the association of SNPs with an UP. In the case of gene expression levels, eQTLs have been used to prioritize SNPs to test for an association with an UP. For instance, eQTLs from a public database were used to select a subset, out of 500 SNPs most strongly associated with Crohn's disease, for replication studies. Out of the 10 SNPs selected three were found significant in an independent study [52]. Using eQTLs to prioritize SNPs for association with an UP is predicated on the hypothesis that in comparison to the general pool of SNPs, trait associated SNPs are more likely to be eQTLs. Under this assumption, using eQTLs will provide an enrichment of significant associations between SNPs and a UP. Moreover, there may be increased power to detect a specified SNP effect on an CP than on the trait the CP influences [53]. This could occur if the CP resulted from a lower number of molecular interactions than the UP. The CP would be subject to less biological noise than the UP and the estimated SNP effect on the CP would be more precise compared the estimated SNP effect on the UP [53]. Therefore, overall genetic determination of CPs in a relevant tissue could be helpful in identifying SNPs that influence an UP.

1.7 Genetic Variation Associated with Level of 5-methylcytosine

Level of 5-methylcytosine is a CP less well studied than level of gene expression. However, studies have begun to assess inter-individual variability in the level of 5-methylcytosine and the relationship with genetic variability [2,54-58]. These studies have been of small sample size and have been conducted in a minority of tissues, typically tissues that are easy to access. In the tissues assayed, $\hat{h}^2_{ped}$ has been calculated or individual SNPs genome-wide have been tested for association with level of 5-methylcytosine at basepair resolution across the genome to identify SNPs associated with 5-methylcytosine (methQTLs). In a minority of cases [57,58], $\hat{h}^2_{ped}$ have been compared for CpG sites located in different functional genomic contexts. Studies utilizing the pedigree information for MZ and DZ twins have suggested a wide range of $\hat{h}^2_{ped}$ for 5-methylcytosine [56-58]. Additionally, these studies have suggested that the average $\hat{h}^2_{ped}$ may be significantly different for different tissues [57] and for CpG sites located in different functional genomic contexts [56,58]. GWASs for 5-methylcytosine have revealed that a SNP can explain up to 88% of the phenotypic variation in level of 5-methylcytosine [2]. Moreover, SNPs local to the 5-methylcytosine were enriched for higher effect sizes than SNPs distal to the 5-methylcytosine [2] and the majority of associations detected for 5-methylcytosine resulted from local SNPs [54,55].

1.8 The Aim of this Study

The overall aim of this work was to explore the heritability of 5-methylcytosine, genome-wide in differentiated somatic cells. We used a newly established statistical methodology that we anticipated would have increased power over a GWAS approach to detect variants with a small effect. Moreover, unlike pedigree-based studies this methodology allowed us to determine the contribution of specific regions of the genome to variation in level of 5-methylcytosine. We investigated both the effect of local and distal genomic regions on level of 5-methylcytosine. We examined variation in the estimate of local genetic effects between CpG sites located within different genomic contexts. Finally, we assessed the extent to which genetic effects local to a CpG site differ among samples taken from different tissues. Our analyses informed on the basic mechanisms that lead to variation in the level of 5-methylcytosine in differentiated somatic cells.

1.9 Contributions

Both my supervisors provided comments on drafts of this chapter.

Chapter 2 The Effect of Local Genetic Variation on DNA Methylation Level Measured in Four Regions of the Human Brain

2.1 Introduction

The occurrence of 5-methylcytosine, from here on defined as DNA methylation (DNAm), can be dependent on the genome sequence. A classic example is the random methylation of one of the two X chromosomes in females to silence genes and to compensate for the higher levels of gene expression that would otherwise occur in females relative to males. A second well noted example is DNAm at ICRs in a parent of origin manner. Furthermore, studies [2,54-61] have established that variation within the genome sequence can affect the binary outcome of DNAm.

Family based studies [56-60] (Table 3) and GWASs [2,54,55,61] (Table 4) have been critical in establishing that genetic variation can influence DNAm level at individual CpG sites throughout the genome The results from different GWASs (Table 4) are not directly comparable because typically each study uses a different sample size, applies a different multiple testing correction, assays different DNAm sites and different tissues and defines local variation by a specific number of basepairs surrounding a DNAm site. Despite the difficulties in comparing findings from multiple GWASs, these studies (Table 4) and others [62,63] have suggested that a substantial proportion of the phenotypic variation for DNAm level can be explained by SNPs local to the DNAm site. However, currently, published results from GWASs (Table 4) are insufficient in providing an estimate of the local heritability for the DNAm sites studied. This is because while these studies have provided estimates of the proportion of the phenotypic variance explained by SNPs they have not taken into account the effects of rare variants, combinations of variants of small effect, or the dependence of effects estimated from SNPs in LD with one another (Table 4).

Table 3 Twin Studies Assessing Heritability of DNAm Level

The table shows twin studies where the pedigree relationship was used to calculate $\hat{h}^2_{ped}$ *for individual DNAm sites. The cell or tissue type in which DNAm level was measured is listed (column one) along with location of the DNAm sites assayed (column two). The* $\hat{h}^2_{ped}$ *is the average for all the DNAm sites assayed (column three). The number of twin pairs provided (column 4) includes pairs of monozygotic (MZ) and dizygotic (DZ) twins.*

Cell or Tissue Type	DNAm Sites Considered	Average $\hat{h}^2_{ped}$	Number of Pairs	Reference
Buccal Epithelial	Genome-wide	0.30	20 MZ, 20 DZ	[57]
White Blood	Genome-wide	0.01	20 MZ, 19 DZ	[57]
CD4+	MHC Complex CpG Islands	0.07	49 MZ, 40 DZ	[56]
CD4+	MHC Complex 5 Prime	0.16	49 MZ, 40 DZ	[56]
CD4+	MHC Complex CNC	0.12	49 MZ, 40 DZ	[56]
CD4+	MHC Complex Random	0.02	49 MZ, 40 DZ	[56]
Peripheral Blood Lymphocytes	Genome-wide	0.20	67 MZ	[58]
Whole Blood	Genome-wide	0.22	23 MZ, 23 DZ	[59]
Cord Blood Mononuclear	Genome-wide (n=19204)	0.12	18 MZ, 9 DZ	[60]
Human Umbilical Vascular Endothelial	Genome-wide (n=19350)	0.07	14 MZ, 20 DZ	[60]
Placenta	Genome-wide (n=26353)	0.05	8 MZ, 7 DZ	[60]

Table 4 GWAS Assessing Association of SNPs with DNAm level

GWAS testing the association of SNPs with DNAm level at individual DNAm sites. The cell or tissue type in which DNAm level was measured is listed (column one) along with location of the DNAm sites assayed (column two) and the SNPs tested for association (column three). The proportion of the phenotypic variance explained by a SNP is reported (column four) and the number of DNAm sites with at least one significantly associated SNP (column five).

Cell/Tissue Type	DNAm Sites Considered	SNP Considered	R^2	Number Sign. DNAm Sites	Number Samples	Ref.
Cerebellum	Genome-wide (n=8590)	Within +/- 1MB	0.17 - 0.73	2046	153	[54]
Cerebellum	Genome-wide (n=8590)	Outwith +/- 1MB	Not Reported	372	153	[54]
Lymphoblastoid	Genome-wide (n=22290)	Within +/- 0.5MB	0.22 - 0.83	180	77	[55]
Lymphoblastoid	Genome-wide (n=22290)	Outwith +/- 1MB	Not Reported	10	77	[55]
Whole Blood	Associated with mRNA Levels (n=517)	Within +/- 0.5MB	Not Reported	69	148	[61]
Whole Blood	Associated with mRNA Levels (n=705)	Outwith +/- 0.5MB	Not Reported	1	148	[61]
Cerebellum	Genome-wide (n=27310)	Within +/- 1MB	0.23 - 0.88	444	108	[2]
Cerebellum	Genome-wide (n=27310)	Outwith +/- 1MB	0.22 – 0.76	657	108	[2]
Frontal Cortex	Genome-wide (n=27532)	Within +/- 1MB	0.20 – 0.83	420	133	[2]
Frontal Cortex	Genome-wide (n=27532)	Outwith +/- 1MB	0.19 – 0.77	740	133	[2]
Pons	Genome-wide (n=27476)	Within +/- 1MB	0.21 – 0.83	359	125	[2]
Pons	Genome-wide (n=27476)	Outwith +/- 1MB	0.19 – 0.76	774	125	[2]
Temporal Cortex	Genome-wide (n=27538)	Within +/- 1MB	0.21 – 0.86	547	127	[2]
Temporal Coretx	Genome-wide (n=27538)	Outwith +/- 1MB	0.20 – 0.78	886	127	[2]

Recently, the genomic heritability method (see 1.5) was adapted so that the effects of causal genetic variation could be partition into genomic regions (regional heritability, RH, $\hat{h}^2_{g,r}$). The effects of SNPs within a genomic region of interest were simultaneously modelled as random effects. Simulation experiments revealed that the RH approach could capture an increased proportion of the genetic variance relative to a GWAS or gene based approaches that simultaneously estimated the effect of SNPs within a genomic region [64,65]. The RH approach outperformed a GWAS and gene based approaches except when the number of causal variants was extremely low [64,65], for instance, less than three within a 10 to 100 SNP window spanning between 103 and 1031 KB respectively [65]. In conjunction, the RH approach could capture the phenotypic variance explained by low MAF much more effectively than GWAS or gene based approaches [64,65]. Moreover, $\hat{h}^2_{g,r}$ obtained when the number of causal loci was not extremely small (less than three), were not dramatically affected by the number of causal variants, the variance contributed by individual causal variants, the size of the genomic window or the overall heritability of the trait [64,65]. Therefore, the RH approach can be used to obtain robust $\hat{h}^2_{g,r}$ that are likely to be more accurate estimates of the true additive heritability than those obtained from a GWAS.

We used the RH approach to estimate the proportion of the phenotypic variance in DNAm level that can be explained by genetic variation +/- 1MB of a DNAm site. We chose a region of +/- 1MB surrounding a DNAm site in keeping with previous GWASs [2,54] that have investigated the effect of *cis* acting genetic variation on variation in DNAm level. We conducted this analysis using publically available data, on which others have conducted a GWAS for DNAm level [2]. This dataset, which we refer to as the brain dataset, was composed of DNAm level measured at 27,578 DNAm sites in the cerebellum (CRBL), frontal cortex (FCTX), pons (PONS) and temporal cortex (TCTX) for a total of 148 unrelated individuals. We selected a subset of the total assayed DNAm sites for RH analysis, choosing the DNAm sites located in risk regions for a disease or disorder related to brain function. We prioritized these DNAm sites for analysis because we were interested to establish if genetic variation that associated with disease also influenced a change in DNAm level measured in a tissue relevant to the disease. We conducted simulation experiments that validated the accuracy of the RH approach for obtaining $\hat{h}^2_{g,r}$ for DNAm level in our sample of unrelated individuals. Subsequently, we reported the RH results for the selected DNAm sites. Then we used the RH framework and all DNAm sites measured within the TCTX to show the extent to which local SNP effects could predict DNAm level for the TCTX in our population. Finally, from our work we have provided an example where genetic variation may mediate the outcome of a trait, fluid intelligence (FI), through changes in DNAm level.

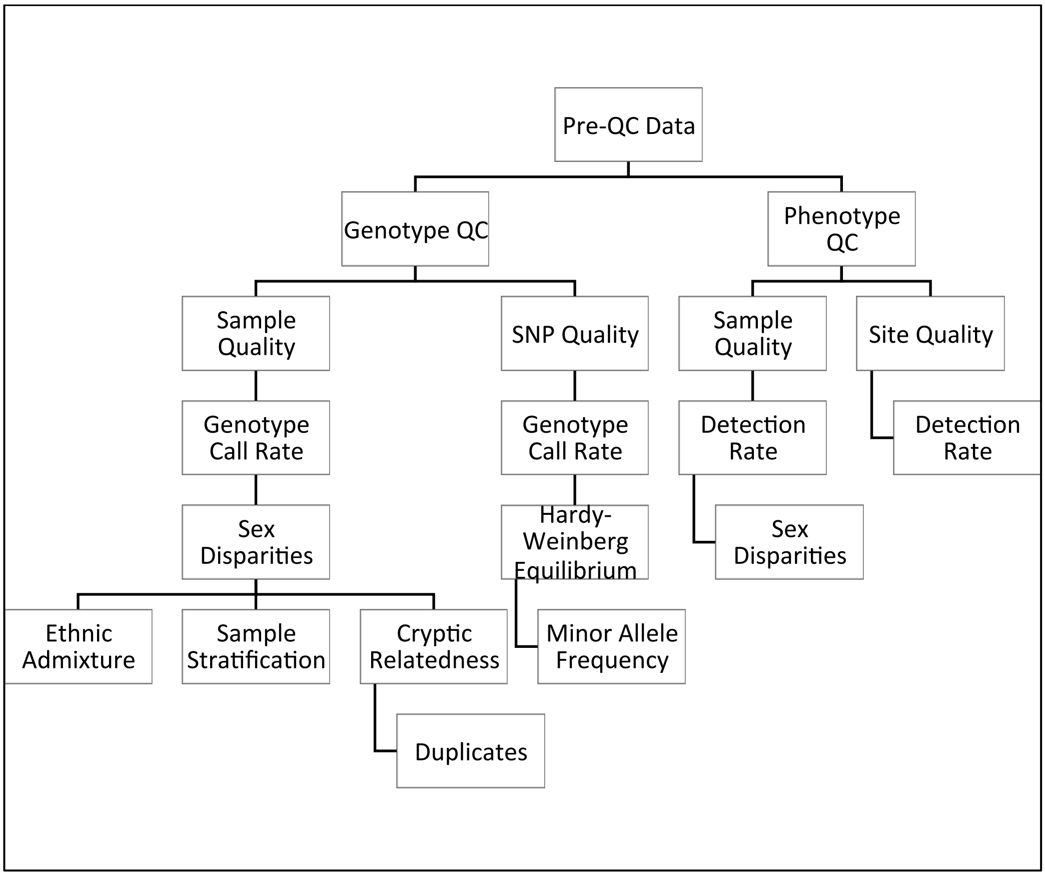
2.2 Materials and Methods

2.2.1 Quality Control

Obtaining accurate analytical results that reflect the true biological situation of study depend on the quality of the genotypic and phenotypic data. Technical factors and human error could increase variation in the data set, which if not accounted for could mask true biological signal. One ramification of this is false positive association, an association that falsely leads to the rejection of the null hypothesis of no association. A second and less identifiable consequence is false negative association, falsely accepting the null hypothesis of no association when in fact there was an association. Application of appropriate quality control procedures (Figure 2) can lead to reliable measurements of genetic markers and traits. This can reduce the probability of obtaining spurious results such as false negatives and false positives in downstream analysis.

Figure 2 Schematic Diagram of the Quality Control Procedure

The data obtained from the online repositories consisted of both phenotype and genotype data that must be subject to quality control (QC) measures.



2.2.1.1 Genotype Quality Control

Genotypes for 148 individuals at 561466 marker SNPs obtained from authorized access to dbGaP were made available to myself on September 12th 2011. These genotypes were subjected to the quality control procedures outlined below using PLINK v1.07 [66].

2.2.1.1.1 Genotyping Algorithms

Genotype calling algorithms for SNP arrays determine a genotype at a given SNP by using the intensity measure from the probe binding the major allele (X) and the intensity measure from the probe binding the alternative allele (Y) and comparing the intensity measures (X, Y) across all assayed individuals in 2D space.

Three clusters representing the three genotypes, A1A1, A1A2, A2A2 should be present for a SNP measured with high accuracy. Proximity to a cluster determines an individual's genotype [67,68]. Problems genotyping specific SNPs, poor quality DNA or low concentration of DNA can mean that genotype clusters are not well defined. In this case genotyping algorithms cannot specify individual genotypes with certainty and genotypes may be miscalled or set to missing. If the proportion and type of genotypes set to missing or miscalled is correlated with the trait of interest then false positive and false negative associations can occur [67]. There are several precautions that can be taken to minimize the effect of missing or miscalled genotypes on obtaining false results. Firstly, individuals for which a substantial proportion of SNPs are set to missing and SNPs that do not type in substantial proportion of the individuals assayed can be removed. Secondly, SNPs that have genotype frequencies that deviate from the expectation in a randomly mating population and are not in Hardy-Weinberg Equilibrium can be removed. Thirdly, SNPs with a low MAF can be excluded [68]. Following classical protocol, samples and SNPs with a call rate below 95% were excluded from downstream analysis [69]. The expected genotype frequencies were calculated from the observed allele frequencies and compared to the observed genotype frequencies using a chi-squared test. SNPs were removed if significantly out of Hardy-Weinberg Equilibrium with a P-value < 0.0001 [69]. SNPs with a MAF < 0.01 were also removed. This protocol reduced the number of SNPs available for downstream analysis from 561466 to 530632. No samples were removed based on the missingness threshold set, leaving 101 male and 47 female samples for analysis. This confirmed that the previous authors' [2] working with this dataset had removed poor quality samples prior to submission to dbGaP.

2.2.1.1.2 Genotype Sample Mix-up

If left undetected accidental mislabelling or swapping of samples during the genotyping process will result in misalignment of the genotypes with the phenotypes. Sample mix-ups can be identified by comparing the recorded sex to the estimate of homozygosity by descent on the X chromosome [68,69]. Homozygosity by descent is the excess in the number of SNPs that are monomorphic in a sample relative to the number of SNPs that are expected by chance to be monomorphic in the sample [66]. Homozygosity by descent for an individual based on the X chromosome can be calculated from the observed number of homozygous SNPs and the expected number of homozygous SNPs due to chance [66]. The expected number of homozygous SNPs due to chance is calculated as the sum of the probability of each SNP being homozygous following the expectation under Hardy-Weinberg Equilibrium [66]. Human males have one X chromosome and with the exception of variation due to genotyping error, genotype calling algorithms should observe that these individuals are completely homozygous at SNPs on the X chromosome. In contrast, females have two X chromosomes and the observed number of homozygous SNPs on the X chromosome is likely to be lower than that observed for males. Based on SNPs on the X chromosome females typically have a low homozygosity by descent (~0.2) and males typically have a high homozygosity by descent (~0.8) [66]. The recorded and predicted homozygosity by descent sex for each of the 148 individuals was concordant. This indicated that the genotype data was free from human error relating to a male sample being swapped with a female sample and vice-versa.

2.2.1.1.3 Cryptic Relatedness, Population Structure and Ethnicity

Cryptic relatedness, population structure and ethnicity can be identified by assessing the relationship of each individual with each other individual in the population sample. The relationship between two individuals in a population can be calculated as the proportion of SNPs at which the two individuals have the same alleles (identity by state) or as the proportion of the genome that two individuals share from a common ancestor (identity by descent). Identity by state is used to infer identity by descent [66]. Calculations of allelic sharing for investigating duplication of samples, population structure and ethnicity can be conducted using a set of SNPs pruned for LD to increase computational efficiency.

The pairwise estimates of identity by descent can reveal cryptic relatedness. For example, in a population sample containing unrelated individuals, identity by descent = 1, an estimate which would be obtained if the two individuals being compared where MZ twins, is not expected and is likely to be the result of a duplication of one sample's DNA [68,69]. Using $r^2>0.1$ (r=0.31) we pruned sets of SNPs within a 50Kb window for LD. We found a total of 35116 SNPs in linkage equilibrium from an initial number of 530632 and we found a maximum identity by descent between two samples of 0.07319. Since all individuals shared substantially less of their genome than third degree relatives (0.125), we did not remove any samples due to cryptic relatedness.

The average proportion of the genome shared identity by state can be used to determine if samples look more dissimilar to one another than what is expected if the population sample is homogeneous. The matrix of pairwise identity by state estimates can be visually interpreted by multidimensional scaling and by plotting the first two multidimensional scaling components. This technique can be conducted within the population of interest to identify structure. Additionally, this technique can be used to examine the ethnicity of samples when samples of known ethnicity, such as samples from the HapMap populations are included. In both cases, the distance between samples on the plot represents the genetic distance between the samples. We used plink [66] to calculate pairwise identity by state and to conduct multidimensional scaling with our samples. By plotting the first two multidimensional scaling components we identified visually that there was no apparent population stratification. Additionally, pooling our samples and those from four HapMap populations [41] (YRI, Yoruba in Ibadan, Nigeria; JTP and CHB, (Japanese in Tokyo and Han Chinese in Beijing; Utah residents with ancestry from northern and western Europe, CEU) validated that our samples all clustered with HapMap individuals of European ancestry.

2.2.1.2 DNAm Level Quality Control

DNAm level measured on the Beta-value scale (see 1.2) at 27,578 DNAm sites and sampled from the four brain regions (BRs: CRBL, FCTX, PONS, TCTX) were downloaded from the GEO database (http://www.ncbi.nlm.nih.gov/geo/) (Table 5). Two samples from the CRBL region were excluded prior to the phenotype quality control procedure because these two individuals were not represented in the genotype data. All other individuals present in the phenotypic data were present in genotype data (Table 5).

2.2.1.2.1 Detection of DNA Methylation Level Above Background Noise

Samples had been assayed for DNAm level using the Illumina Infinium HumanMethylation27 BeadChip (HM27K). The HM27K assay reports a P-value for the significance of the detection of the level of DNAm at each DNAm site above the background level of noise measured from negative control probes on the array [70]. Following others [2], samples were removed if DNAm level at greater than 5% of DNAm sites were not detected above background levels of variation (P = 0.01) (Table 5). Similarly, DNAm sites were removed if DNAm level at that site was not detected above the background level of variation (P=0.01) in greater than 5% of samples (Table 6).

Table 5 Samples Available for Analysis after Application of Sequential Quality Control Steps

Each cell in the table gives the number of samples assayed for DNAm level and remaining after the sequential quality control steps. The number of samples for each brain region is listed separately.

	CRBL	FCTX	PONS	TCTX
Downloaded from GEO	121	133	125	127
No Genotype Data in dbGaP	119	133	125	127
Detection Above Background Variation	109	132	125	127
Sample Sex Discrepancy	108	132	125	127
Cluster Analysis	103	130	121	126
No Covariate Information	102	129	120	125

Table 6 DNAm Sites Available for Analysis after the Sequential Application of Quality Control Procedures

	CRBL	FCTX	PONS	TCTX
Downloaded	27579	27579	27579	27579
Detection Above Background Variation	27393	27545	27516	27554
SNPs Within Assay Probe	20722	20843	20820	20848
+/- 1MB GWAS SNP (Autosomal)	3057 (3049)	3072 (3064)	3070 (3062)	3073 (3065)

2.2.1.2.2 Removal of DNAm Sites Assayed by Probes Containing SNP

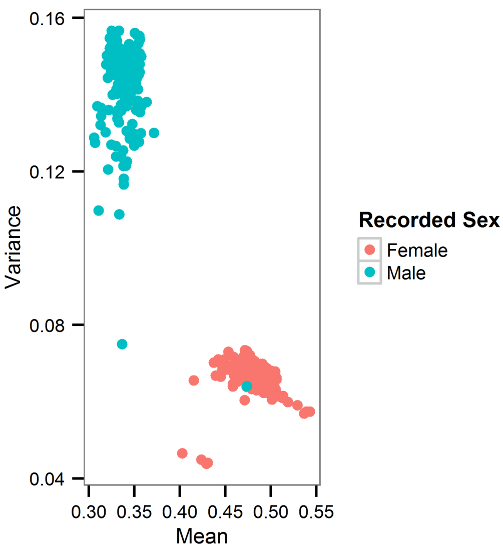
Only DNAm sites that did not contain a SNP within the assay probe sequence, as documented in the manifest file for the Illumina Infinium HumanMethylation450 BeadChip (HM450K), were retained for analysis (Table 6)

2.2.1.2.3 Sample Sex Inconsistencies

Similar to genotype quality control, the process of phenotyping a sample taken from an individual is subject to human error that can lead to samples being incorrectly labelled. One way of identifying mislabelled samples is to check the recorded sex for an individual against the predicted sex for an individual based on the level of DNAm of DNAm sites along the X chromosome [71]. Female samples are expected to exhibit hemimethylation of the X chromosome in accordance with the mechanism of dosage compensation [71]. Male samples are expected to show low DNAm level on the X chromosome in comparison to the hemimethylation exhibited by female samples. We assessed the average DNAm level on the X chromosome for each sample. The results indicated that the samples clustered into two groups. Indeed, samples recorded as being male typically had a lower estimate of average DNAm level and higher variance of the mean estimate of DNAm level than samples recorded as female (mean for all 4 BR: males = 0.34, females = 0.48, variance for all 4 BR: males 0.14, females = 0.07). The results indicated that one sample from the CRBL region was likely to have been mislabelled and this individual was removed from further analysis (Figure 3, Table 5).

Figure 3 Recorded and Phenotypic Estimated Sex for CRBL Samples

Recorded sex and sex estimated from DNAm level on the X chromosome revealed that a sample was potentially mislabelled. One sample reported as male appears to have the DNAm level profile of a female. This sample was removed from further analyses, as the assay may not have been conducted on the correct sample.

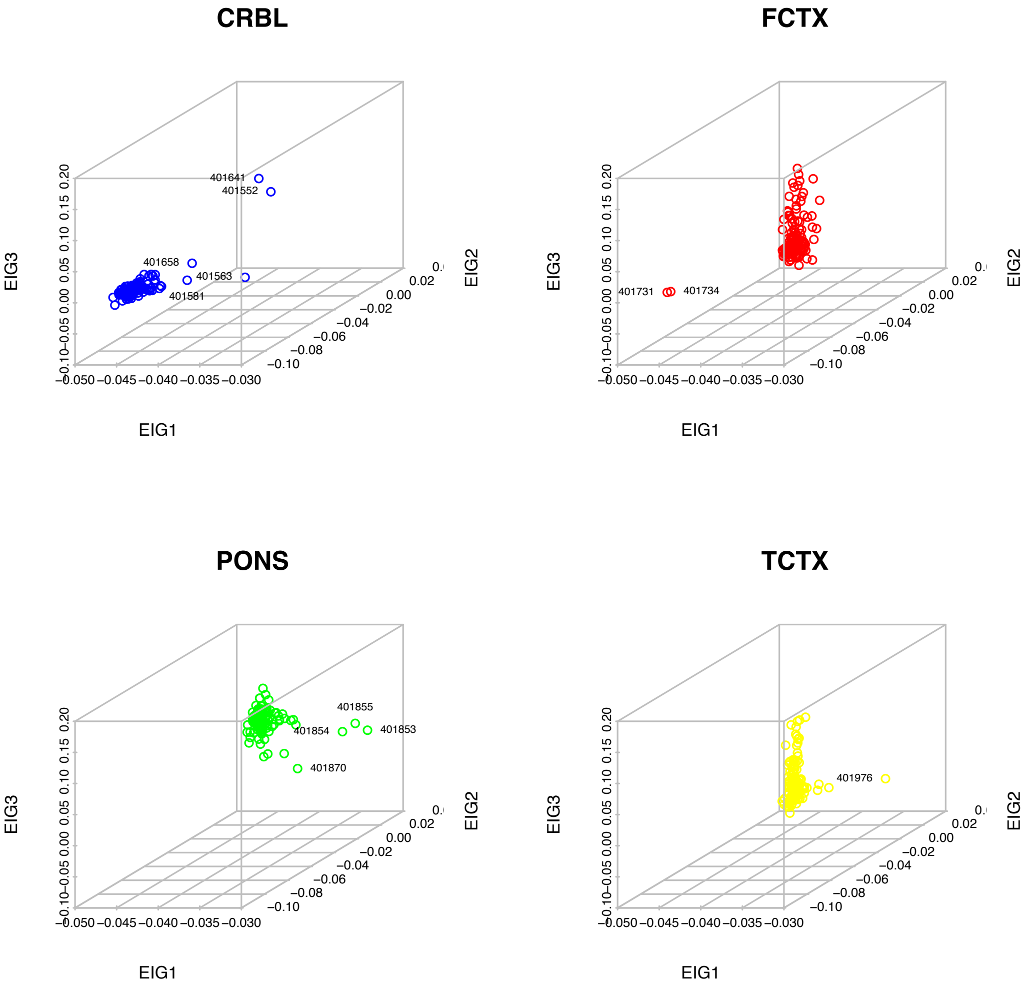


2.2.1.2.4 Whole Genome Methylation Profiling to Detect Outliers

We used principal component analysis with autosomal DNAm sites, common to all BR, to investigate the phenotypic data for outlying data points. Between 1 and 5 samples were removed from each BR based on an abnormal DNAm level profile as determined by the first three major axis of variation (Figure 4, Table 6)

Figure 4 Principal Component Analysis of DNAm Level

Principal component analyses examining the covariance of DNAm level for all autosomal DNAm sites and samples. Within the plots the numbers indicate the identity of the samples that were removed due to an aberrant DNAm profile.



2.2.1.3 Gene Expression Level Quality Control

The level of gene expression measured for genes across the genome and in the CRBL, FCTX, PONS and TCTX was available for the samples in the brain dataset and downloaded from the GEO database (http://www.ncbi.nlm.nih.gov/geo/). The limma [72], Biobase [73] and arrayQualityMetrics [74] packages for the statistical program 'R' [75] were used to conduct quality control measures for the mRNA expression data. Probes measuring mRNA expression were kept for subsequent analysis if they were detected above background noise levels at $p < 0.05$ in $\geq 95\%$ of the samples. A Kolmogorov-Smirnov test comparing the intensity distribution for each sample to that of all samples pooled for each tissue was used to detect and subsequently remove sample outliers. A total of 145, 146, 144 and 147 samples from the CRBL, FCTX, PONS, and TCTX remained for analysis. The mRNA levels for samples were quantile normalized within brain tissue type using limma [72]. Additionally, the expression levels of each mRNA probe were adjusted for sex, age, time lapse between death and taking the sample (post-mortem interval, PMI), brain bank each sample came from (study) and genotyping plate (plate) prior to analyses.

2.2.2 Testing the Effect of Explanatory Variables on DNAm Level

In order to accurately assess the strength of an association between the phenotype and genotype data it is necessary to include the effects of variables that can explain a significant proportion of the phenotypic variation. As discussed in 2.2.1.3 included in the brain dataset were measurements for the 5 variables: sex, age, PMI, plate, and study for each individual. Sex and age have been found to correlate with DNAm level [59]. Additionally, variation in PMI (ranging from 0 to 28 hours after death), plate (7 levels), and study (3 levels) may correlate with inter-individual variation in DNAm level.

The proportion of the variance explained by a full model, including all 5 explanatory variables, was compared to a reduced model including only four of the five explanatory variables. To test if the fit of the full model was significantly better than the fit of the reduced model we used an F-test, $P < 0.05$ and the drop1 function in R [75]. The results for the inclusion of the effect of sex, age and plate were highly enriched for significant associations with DNAm level measured in all four BR (results not shown). The results for the inclusion of the variable, study, were highly enriched for significant associations in the CRBL (results not shown). The results for the inclusion of PMI were moderately enriched for significant associations in the FCTX and TCTX (results not show). Based on these results, and the work of others who have analysed this dataset [2], we fit all five variables: sex, age, PMI, plate and study in subsequent analyses. The consequence was a loss of 11 degrees of freedom.

Furthermore, we determined that cause of death was not a significant factor influencing variation in DNAm level. A total of 50 different causes of death were reduced to the four categorical variables that showed highest representation in the dataset (control, multiple injuries, cardiovascular disease and other). The proportion of variance in DNAm level explained by cause of death was determined by comparison of a full model containing sex, age, PMI, plate, study and cause of death and a reduced model that did not contain cause of death. Overall, we did not find an enrichment of results where the fit of the full model was significantly better than the fit of the null model (results not shown). Therefore, in accordance with previous authors' [2] who have analysed this dataset we did not fit cause of death as a covariate in further analyses.

2.2.3 Transformation of DNAm Level

The results from fitting a linear model are reasonable descriptive statistics if several conditions regarding the dataset were met prior to analysis [76]. Firstly, the dependent variable must have a linear relationship with the independent variable, secondly, for any value of the independent variable the values of the dependent variable should be independent and normally distributed, and thirdly the probability distribution of the dependent variable for each value of the independent variable should have the same standard deviation. In practice, visualization of a plot of the residual deviation for each set of data points can be used to check that the third assumption is met. However, this is not feasible when one has a large number of linear regressions to conduct. Therefore, we used an alternative method to determine if our data met the third criteria stated above. We regressed DNAm level on sex, age, PMI, study and plate at each DNAm site and we used the Shapiro-wilk test in the statistical package R [75] to test if the residual distribution was significantly different from a normal distribution. The results suggested that for each brain region approximately 80% of the traits did not have normally distributed residuals. Therefore, we used a rank transformation to normalize the phenotypic data for each DNAm site. After normalization I found that approximately 90% of the DNAm sites for each BR had normally distributed residuals as determined by the Shapiro-wilk test. The rank transformed Beta-values were used in all analyses in this Chapter unless otherwise stated.

2.2.4 Selection of a Subset of DNA Methylation Sites Local to Risk Variants for Disorders and Disease of the Brain

The National Institute of Health full Catalogue of Published Genome-Wide Association Studies (http://www.genome.gov/gwastudies/) was downloaded on September 30th 2011. A total of 202 SNPs, representing multiple loci across the genome, were listed as being associated with at least one of the following seven brain related traits: parkinson’s disease, alzheimer’s disease, bipolar disorder, schizophrenia, major depressive disorder, glioma, and neuroblastoma. We identified DNAm sites within a 1MB window upstream or downstream of the risk associated SNPs.

2.2.5 Simulating DNAm Level from the Genotype Data

The phenotype, DNAm level, was simulated for individuals within the brain dataset using causal SNPs selected from the genotype data and for a specified heritability. We used the GCTA software [77] , which is an earlier version of the REACTA software [78] mentioned in 2.2.6. With the --simu-causal-loci command GCTA [77] selects effect sizes for each of the causal SNPs from the standard normal distribution. Using allele frequencies for these SNPs, calculated from the genotype data, and the effect sizes, the total additive genetic variance can be calculated. Specifying a heritability for the simulated phenotype facilitates calculation of the environmental variance. An environmental deviation for each sample can be obtained from a distribution with mean of zero and variance equal to the calculated environmental variance.

2.2.6 Estimating Regional Heritability and Prediction of DNAm Level

Estimation of RH and prediction of DNAm level at an individual DNAm site was conducted within a mixed linear model and REML framework using the publically available software: REACTA [78]. Below, “total genetic effects” refers to the total genetic effects within a region of interest, which except where specified, is +/- 1MB of a DNAm site.

2.2.6.1 Estimating Regional Heritability

Let the following variables be represented:

$y =$ vector of DNAm levels for an individual DNAm site and all samples

$X =$ incidence matrix

$I =$ identity matrix

$\beta =$ vector of independent effects

$g =$ vector of total genetic effects, one for each sample

$\varepsilon =$ vector of error terms

$V =$ variance-covariance of y

$A =$ matrix of pairwise genetic relationships for the samples

$\sigma_g^2 =$ variance of g

$\sigma_\varepsilon^2 =$ variance of ε

The variance-covariance in the phenotype, DNAm level, is a function of the variance-covariance of the total genetic effects of the samples and the variance-covariance of an error term (Equation 5)

Equation 5

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{g} + \boldsymbol{\varepsilon}; \mathbf{V} = \mathbf{A}\boldsymbol{\sigma}_{\mathbf{g}}^{\mathbf{2}} + \mathbf{I}\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{\mathbf{2}}$$

The covariance-variance of the total genetic effects of the samples can be related to the pairwise genetic relationships between the samples (Equation 6). In our analyses the pairwise genetic relationships were calculated from SNP within a region of interest (often +/- 1MB of a DNAm site) following a previously defined equation [79].

Equation 6

$$g \sim MVN(0, A\sigma_g^2)$$

Restricted maximum likelihood provides the most likely estimate of the variance in the genetic effects for specified phenotypic and genotypic data. When the genotypic data provided refers to pairwise genetic relationships calculated from SNP within a region of interest then the genetic variance relates to the genetic variance within the region of interest. Once the variance components have been estimated $\hat{h}^2_{g,r}$ can be calculated (Equation 7).

Equation 7

$$\hat{h}^2_{g,r} = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\varepsilon^2}$$

The null hypothesis that $\hat{h}^2_{g,r}$ was not significantly different from zero was tested with a Likelihood Ratio Test distributed as 50:50 mixture of Chi-squared distributions with 0 and 1 degrees of freedom [80]. If P < 0.05 we rejected the null hypothesis and concluded that the DNAm site was heritable.

2.2.6.2 Prediction of DNAm Level

Assume the following variables in addition to those specified above:

$W =$ matrix of the number of reference alleles for each SNP and sample
$N =$ total number of SNPs
$u =$ vector of SNP effects
$\sigma_u^2 =$ variance of SNP effects

Then the mixed model and REML framework (Equation 5) can be utilized to obtain individual SNP effects within the genomic region of interest for a given fold of samples (Equation 8). In this case the total genetic effect of a sample within the fold is a function of the individual SNP effects and the alleles the sample carries at the SNPs.

Equation 8

$$g = Wu;\ \sigma_g^2 = N\sigma_u^2; u \sim N(0, I\sigma_u^2)$$

Once the effects of SNP within a fold of samples have been estimated they can be used to predict the phenotype of samples outwith the fold (Equation 9). The predicted phenotype is a product of the SNP effects estimated within the fold and the alleles at the SNPs carried by the sample outwith the fold.

Equation 9

$$\hat{y} = Wu$$

2.3 Results

2.3.1 Validation of the Accuracy of Regional Heritability Estimates

We designed simulation experiments to determine if we could accurately detect heritability within our dataset that is a homogeneous population sample of up to 129 unrelated individuals (where sample size depends on BR) of European Ancestry.

All our models were based on using non-causal array SNPs to calculate the GRM for individuals within our dataset. Overall, we examined a total of four different general models, the difference between each being the localization of the causal variation and the genomic division used for calculation of the GRM. The four general models were 1) genome-wide causal variations and a genome-wide GRM, 2) causal variations +/- 1MB region surrounding a DNAm site ($\hat{h}_g^2$) and a genome-wide GRM ($\hat{h}_g^2$), 3) causal variations +/- 1MB region surrounding a DNAm site and a GRM constructed from all SNPs outwith the 2MB region containing the causal variations ($\hat{h}_{g,trans}^2$ *trans*) and 4) causal variations +/- 1MB region surrounding a DNAm site and GRM constructed from all non-causal SNPs within the 2MB causal region ($\hat{h}_{g,cis}^2$ *cis*). We contrasted how accurately each approach could capture simulated heritability.

Following model 1) we simulated 100 phenotypes from 100 causal SNPs randomly distributed across the genome and specifying a heritability of 0.2 and 0.9. In the majority of cases, the heritability estimated from genome-wide non-causal SNPs did not match the simulated heritability of 0.2 or 0.9 (Figure 5). Most often, the estimated heritability was an extreme value of less than 0.05 or greater than 0.95 (Figure 5) and none of the heritability estimates were significant at $P < 0.05$.

Following models 2), 3) and 4) we used genotype information from 200 genomic regions of 2MB size to simulate the effect of *cis* acting SNPs on DNAm level. This ensured that we captured different LD structures between SNPs within a 2MB region and this was important because lower LD between causal and tagging SNP will result in an under estimation of the heritability. Each of the 200 simulations can be thought of as a replicate experiment and I refer to them subsequently as simulation experiments. Additionally, we sought to simulate multiple different biologically plausible models that represented how SNPs may affect DNAm level. Others [2] have conducted a GWAS with this dataset and they have shown that on average 20 SNPs within +/- 1MB of a DNAm site associated with level of DNAm at a DNAm site. From the aforementioned analysis it is not clear to what extent the SNPs within the +/-1MB region that were associated with DNAm level were capturing signal from the same causal variant and it is possible that the total number of causal variants is less than 20. Therefore, for each of the simulation experiments we used five different numbers of causal variants (20, 10, 5, 3, 1). For each simulation experiment the 20 causal variants were selected randomly from the genotyped SNPs within the 2MB region and the lower number of causal variants was always a random subset of the next largest number of causal variants. Each of the five sets of causal variants was used to generate a phenotype with a known heritability following the protocol outlined in 2.2.5. Due to the reported wide range of SNP effects acting in *cis* to affect DNAm level [2] we specified the simulation of phenotypes based on the following heritabilities: 0, 0.2, 0.4 and 0.9. The results indicated that on average, for all numbers of causal variants, causal genetic effects simulated from SNPs within the 2MB region could be captured by non-causal SNPs within the 2MB region (Figure

6). In contrast, genome-wide or *trans* SNPs do not on average capture the heritability simulated within a 2MB region (Figure 6). Moreover, the standard errors of $\hat{h}^2_{g,cis}$ are substantially smaller than of that from $\hat{h}^2_{g,trans}$, $\hat{h}^2_g$ (Figure 6).

As discussed above, with our sample size only one of the four simulation models that we tested accurately captured the effect of the simulated causative variation. Therefore, we took this model forward for downstream analysis.

Figure 5 Genome-wide SNP Effects were not Captured by Genome-wide Tagging SNPs

The x-axis is binned in increments of 0.05.

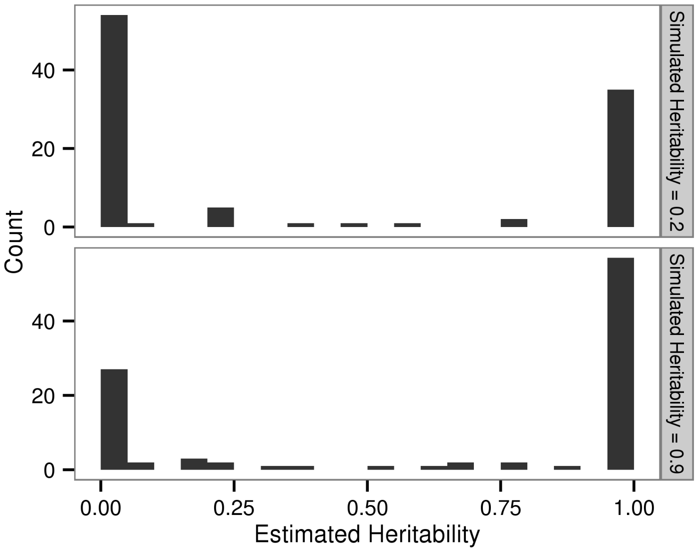
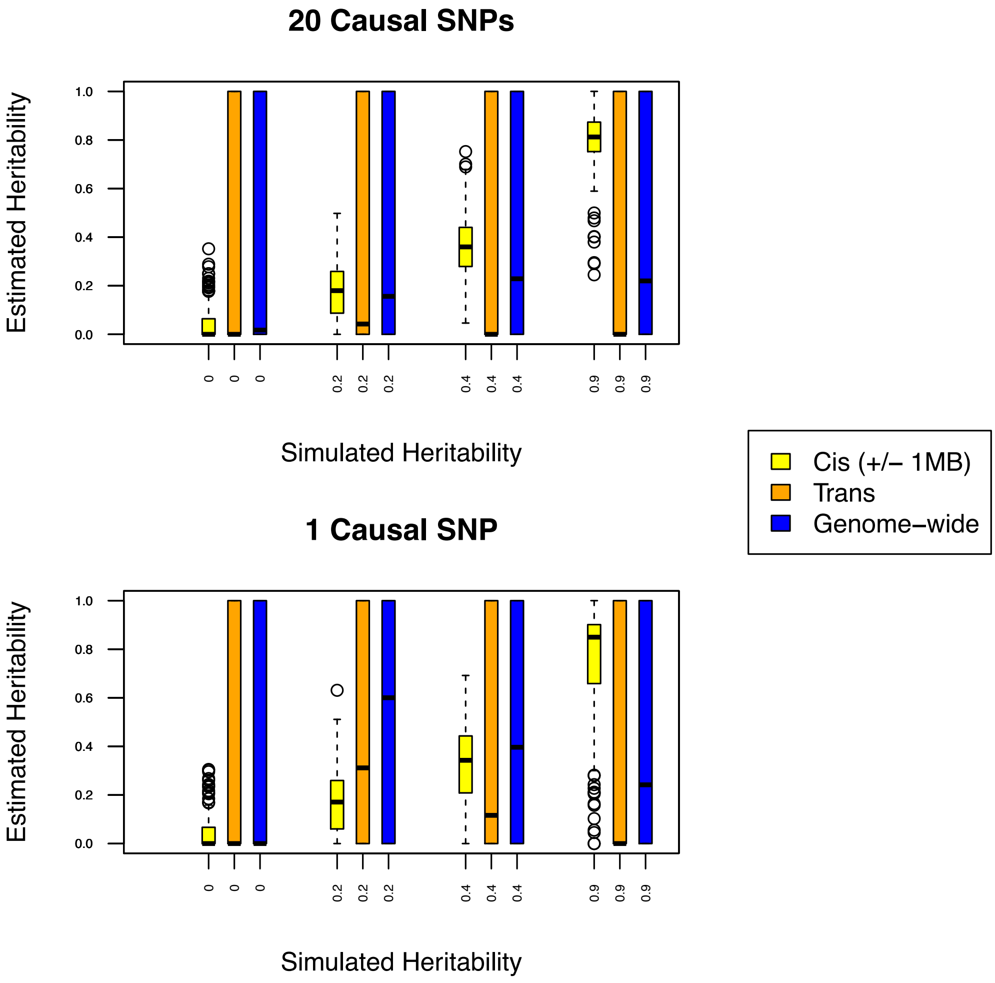


Figure 6 Heritability Estimated from Simulation Experiments

Phenotypes were simulated from different numbers of causal SNPs (20, 10, 5, 3 and 1) located within a 2MB region for a range of heritabilities (0.0, 0.2, 0.4, and 0.9). Non-causal SNPs in cis, trans, or genome-wide were used following the method described in 2.2.6 to obtain $\hat{h}^2_{g,r}$. The results from the simulations using the two most extreme numbers of causal SNPs are shown and they indicated that in comparison to SNPs in cis, SNPs genome-wide or in trans to the causal variation did not accurately or precisely capture the genetic effects. This result is reflected by a discrepancy in the mean estimated (black dots within boxplot) and simulated heritability and extensive interquartile range when genome-wide and trans SNP are used.



2.3.2 Heritability of DNA Methylation Level Due to Local Genetic Variation

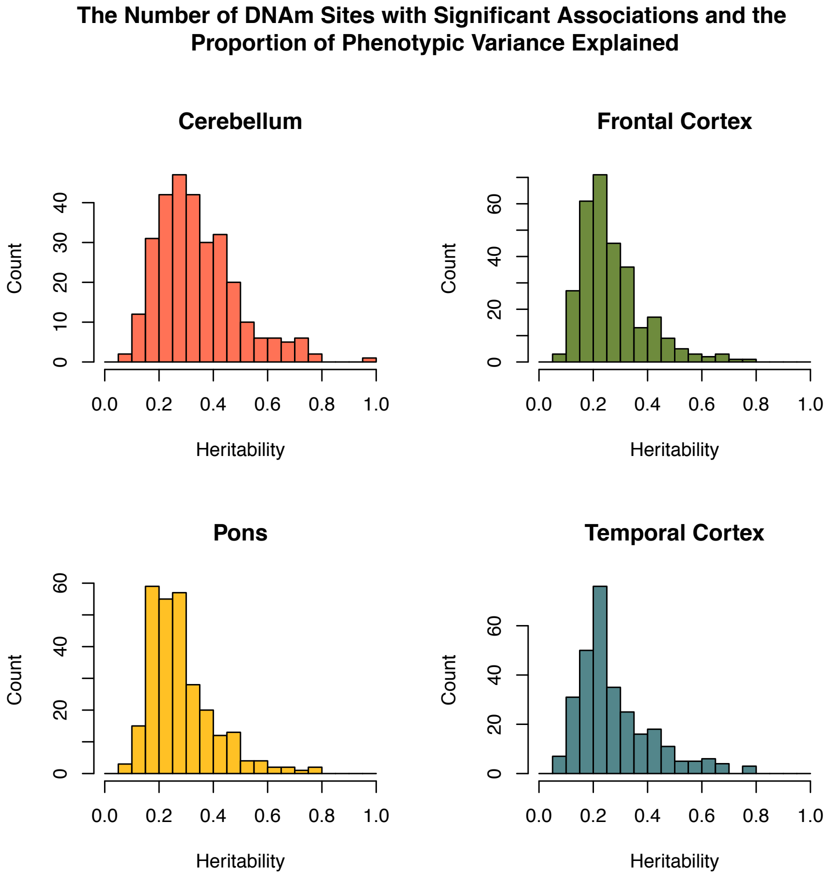
Using the RH approach described in 2.2.6 we calculated $\hat{h}^2_{g,r}$ for +/- 1MB regions surrounding the DNAm sites that were autosomal and +/- 1MB of a SNP associated with at least one of the seven brain related disorders/diseases listed in 2.2.4. We found that at between 9% - 9.7% of the DNAm sites, DNAm level was significantly associated with the local genomic region (Table 7). Individual significant local genomic regions explained between 6.6% and 97.5% of the phenotypic variation at the associated DNAm site and average $\hat{h}^2_{g,r}$ was 0.27-0.34 (Figure 7, Table 7).

Table 7 Regional Heritability Analysis for DNAm level

	DNAm Sites Analyzed (n)	DNAm Sites (n) with significant $\hat{h}^2_{g,r}$	DNAm Sites (%) with significant $\hat{h}^2_{g,r}$	Average $\hat{h}^2_{g,r}$ for Significant Associations
CRBL	3049	294	9.6	0.34
FCTX	3064	297	9.7	0.27
PONS	3062	277	9.0	0.28
TCTX	3065	292	9.5	0.28
Total	12240	1160	9.5	0.29

Figure 7 Significant estimates of regional heritability

The distribution of $\hat{h}^2_{g,r}$ for genomic regions significantly associated with DNAm level.



2.3.3 Evidence for Genetic Correlation of DNA Methylation Level Across Brain Regions

We have obtained $\hat{h}^2_{g,r}$ for 3049 DNAm sites common to each of the four BRs. However, testing the association of DNAm level measured in each of the four BRs with local genomic variation does not constitute four independent experiments. This is because tissue was extracted from the four BRs from a common set of individuals. However, assuming four independent experiments one would expect by chance 0.019 of the DNAm sites to have a significant $\hat{h}^2_{g,r}$ in all four BRs at P < 0.05 (3049*0.05*0.05*0.05*0.05). In fact, we found that $\hat{h}^2_{g,r}$ is significant for 42 of the DNAm sites in all four BRs (Figure 8) which is 2210 times larger than expected if the measurement of DNAm level in the four BRs was independent. To determine the extent of the similarity of the genetic effects across the BRs we investigated the correlation of $\hat{h}^2_{g,r}$ across the DNAm sites for the four BRs, including estimates of $\hat{h}^2_{g,r}$ at all levels of significance (ie. we did not discard non-significant results). The overall correlation was between 0.25 and 0.45 depending on the two BRs being compared (Table 8).

Figure 8 Heritable DNAm Sites in the Four Brain Regions

The figure shows the number of DNAm sites with a significant $\hat{h}^2_{g,r}$ across tissues as well as the count of tissue specific associations.

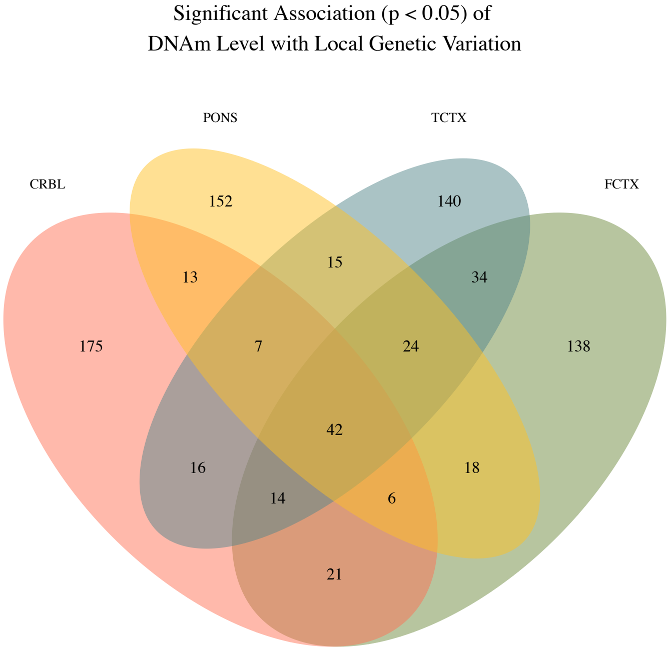


Table 8 Correlations of RH Estimates across Brain Regions for the 3049 DNAm Sites Analysed

	CRBL	FCTX	PONS	TCTX
CRBL				
FCTX	0.26			
PONS	0.19	0.34		
TCTX	0.23	0.45	0.37	

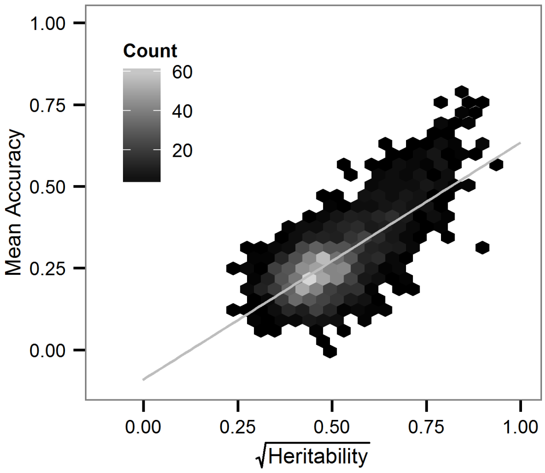
2.3.4 DNA Methylation Level can be Accurately Predicted From Local Genetic Variation

DNAm level at 1826 autosomal DNAm sites that were heritable in the TCTX was predicted from the local SNPs using the statistical methodology outlined in 2.2.6.1. This analysis was conducted using the overlapping set of 314434 SNPs genotyped for a control dataset for Parkinson's disease. The set of overlapping SNPs were used to facilitate additional analysis that was then conducted by others. Samples used to obtain the SNP effects were not used to predict the phenotype and we conducted five fold cross validation, with each fold containing 25 samples. Four of the five folds (n=100) were used to estimate the SNP effects and one fold (n=25) was then used to obtain a prediction of the phenotype. The correlation of the predicted and observed phenotype was computed. We used the residual phenotype on the M-value scale to estimate SNP effects; therefore, we compared the predicted phenotype to the residual phenotype on the M-value scale. The process was repeated so that a correlation of the predicted and observed phenotypes was obtained for each of the five folds. The mean accuracy for a DNAm site was calculated as the average of the five correlations, one from each fold. To facilitate across trait comparison the mean accuracy was normalized by dividing by the theoretical upper limit, which is the square root of $\hat{h}^2_{g,r}$. The normalized mean accuracies for the 1826 DNAm sites ranged between

-0.0012 and 1.115 with a mean of 0.5409 and approximated a normal distribution (Shapiro-Wilk normality test; P-value for the deviation from normality = 0.2402). Moreover, the accuracy scaled with the theoretical upper limit (Figure 9). We calculated the standard error of the sample variance for each DNAm site using the conventional variance among means formula [81] using the accuracy obtained within each of five folds. We found that the resulting mean accuracy below zero and greater than one, for one and three DNAm sites respectively fell within two standard errors of the boundary (0 or 1). Therefore, the mean accuracies below zero and greater than one were consistent with sampling variation. Additionally, we found that the number of SNPs used to capture the regional effects did not explain a substantial proportion of the variance in the estimates of the normalized accuracy ($r^2 = 0.0580$).

Figure 9 Accuracy of Five Fold Cross Validations

Mean accuracy of five fold cross validations as a function of $\hat{h}^2_{g,r}$



2.3.5 Regional Genetic Variation Affects both Fluid Intelligence and Local DNA Methylation Levels

Rowe et al. (unpublished at the time of this research) used the RH approach to test the association of overlapping genomic regions with measurements of cognitive ability. Fluid intelligence (FI) is a measure of the ability to solve novel problems through reasoning and it can be thought of as on the spot thinking [82]. Rowe et al. (unpublished at the time of this research) found that the genomic region most significantly associated with FI spanned Chromosome 5 at 126711782-127335370 ($\hat{h}^2_{g,r}$= 0.02, SE = 0.009). We refer to the dataset used for the aforementioned analysis as the FI dataset. The FI dataset contained genotypic information and a measurement of FI for 1804 samples. Collaborating with Rowe et al. we investigated if heritable variation in the level of DNAm or gene expression within Chromosome 5 126711782-127335370 and measured in brain tissue could affect FI.
To this end we followed a three-step process. Firstly, we identified a set of 86 SNPs common to both the FI dataset and brain dataset and located within Chromosome 5 126711782-127335370. We used these SNPs to obtain $\hat{h}^2_{g,r}$ for
DNAm sites and gene expression probes within Chromosome 5 126711782-127335370 and measured in the CRBL, FCTX, PONS or TCTX. Subsequently, we identified the CPs with a significant $\hat{h}^2_{g,r}$ because it is only reasonable to predict a phenotype from the genotype if evidence suggests that the phenotype is indeed a function of the genotype. Secondly, following 2.2.6.2 in the brain dataset the individual effects of the 86 SNPs on the heritable CPs were determined. Thirdly, following we 2.2.6.2 used the estimated SNP effects to predict the CPs in the FI dataset using the FI genotypes.

In the brain dataset, two DNAm sites, cg04431054 and cg15851800 and two mRNA probes, ILMN_1652306 and ILMN_1685140 were located within the region on

Chromosome 5 (Table 9). Cg04431054 and cg15851800 are located 381 basepairs apart, 277 basepairs upstream and 104 basepairs downstream of the 5` start of PRRC1 that spans chromosome 5 at basepair location 126,853,301-126,890,781 (which is encoded on the forward strand). ILMN_1685140 is designed to target transcripts of PRRC1 and ILMN_1652306 transcripts of MEGF10, Chromosome 5: 126,626,523-126,801,429.

Table 9 Cellular Phenotypes Located in a Genomic Region Associated with Fluid Intelligence

The cellular phenotype (DNAm site or mRNA probe) is listed by the Illumina identification and the location is given in the format of chromosome then basepair position in genome build 37.

Cellular Phenotype	Location
cg04431054	5:126853024
cg15851800	5:126853405
ILMN_1685140	5:126886130
ILMN_1652306	5:126789385

The estimates, $\hat{h}^2_{g,r}$ were obtained for cg04431054, cg15851800, ILMN_1652306 and ILMN_1685140 for measurements of DNAm level from each of the four BRs, except for ILMN_1652306, which did not pass the quality control procedure for the CRBL.

We found that the region on chromosome 5:126711782-127335370 explained a significant ($p < 0.0001$) proportion of the heritability for cg04431054 measured in each of the four BR (Table 10). The estimated heritability and standard error of cg04431054 measured in the CRBL, FCTX, PONS and TCTX was 0.463 se 0.124, 0.237 se 0.104, 0.278 se 0.111 and 0.326 se 0.110 respectively (Table 10). However, phenotypic variation in cg15851800, ILMN_1652306 or ILMN_1685140 was not significantly associated with genetic variation at chromosome 5:126711782-127335370 (Table 10).

Table 10 Heritability of Cellular Phenotypes Located in a Genomic Region Associated with Fluid Intelligence

The $\hat{h}^2_{g,r}$for each of the DNAm sites and gene expression probes analysed

Cellular Phenotype	Tissue	$\hat{h}^2_{g,r}$	P
cg04431054	CRBL	0.463	$1.370*10^{-8}$
cg15851800	CRBL	0.000	0.500
cg04431054	FCTX	0.237	$1.190*10^{-5}$
cg15851800	FCTX	0.020	0.325
cg04431054	PONS	0.278	$1.270*10^{-5}$
cg15851800	PONS	0.003	0.477
cg04431054	TCTX	0.326	$1.020*10^{-8}$
cg15851800	TCTX	0.082	0.063
ILMN_1685140	CRBL	0.025	0.315
ILMN_1652306	FCTX	0.000	0.500
ILMN_1685140	FCTX	0.000	0.500
ILMN_1652306	PONS	0.000	0.500
ILMN_1685140	PONS	0.046	0.075
ILMN_1652306	TCTX	0.000	0.500
ILMN_1685140	TCTX	0.000	0.500

Subsequently, we estimated the individual SNP effects on cg04431054 in the four BRs and predicted the DNAm level in the FI dataset. The predicted DNAm level for the TCTX was significantly associated with FI and explained 0.5% of the variation in FI (Table 11). The regression coefficient was positive (0.295, se =0.004) indicating that an increase in DNAm level was associated with increase in FI. A significant association between predicted DNAm level in the other three BR and FI was not observed (Table 11)

Table 11 Association of Predicted cg04431054 with Fluid Intelligence.

Tissue indicates the region of brain in which DNAm at cg04431054 was measured. Tp is the significance of the association of predicted DNAm level with FI and r^2 is the proportion of the variance in FI explained by predicted DNAm level.

Tissue	Tp	r^2
CRBL	0.779	0
FCTX	0.092	0
PONS	0.861	0
TCTX	0.004	0.005

2.4 Discussion

Using simulation experiments, we have shown that effects of causative variation within a 2MB region can be accurately and simultaneously captured using tagging variation within the same genomic region. This result was true for different heritabilities simulated from different numbers of causative SNPs. The most likely explanation for this result is that the tagging variation was in sufficient LD with the causative variation so that the effects of the causative variation could be detected. In contrast, in our simulation experiments we found that variation in *trans* with the causative variation, or genome-wide variation could not robustly capture causative effects within a 2MB region. The *trans* SNPs were outwith the 2MB causative region and thus it is likely that they were not in LD with the causative variation, which results in the inability of the *trans* variation to capture the causative SNP effects. Although genome-wide variation spans the causative 2MB region, the majority of the genome-wide variation was outwith the causative 2MB region. In this case it is likely that SNPs extraneous to the causative region that were not in LD with the causative variation added noise to the estimate of the effects accurately captured within the causative region. A similar phenomenon is likely to be occurring when we unsuccessfully attempted to capture the effects of 100 causative SNP distributed across the genome with genome-wide variation. In this last case, the 100 causative SNPs were used to simulate a heritability of either 0.2 or 0.9. A simulated heritability of 0.9 allowed for some of the (simulated) causative SNPs to have an effect within the range exhibited by discovered *trans* acting effects on DNAm level (r^2 of 0.19-0.78) [2]. Since we could not detect simulated causative genome-wide effects that were similar to (non-simulated) true causative genome-wide effects detected by others [2], it was unlikely that we could detect true causative genome-wide effects in our dataset. Therefore, in our analyses we fitted only a regional variance component, which is unlike some other studies using RH analyses [64,65]. These other studies [64,65] have fitted both the effects of regional and genome-wide variation using two variance components. Therefore, our estimates for the effect of local genetic variation on DNAm level were relative to a null model of no effect of local genetic variation on DNAm. In studies fitting both a regional and genome-wide variance component [64,65] the regional genetic effects on a trait can be compared against a null model of the genome-wide effects. The implication for our study is that we did not have an estimate of the background genetic effects on DNAm level to which we could compare the estimate of local genetic effects.

We have shown that genetic variation in risk regions for brain related disorders/diseases was significantly associated with the DNAm level of a nearby

DNAm site in non-diseased brain tissue. Additionally, we have produced estimates of the heritability that simultaneously captured the effects of genetic variation +/- 1MB of a DNAm site on DNAm level at the DNAm site. Our results indicated that it is possible that genetic variation may lead to disease by first influencing variation in DNAm level in a tissue related to disease.

During the undertaking of this research a paper was published [83] which utilized the brain dataset and the methodological framework outlined in 2.2.6 to obtain $\hat{h}^2_{g,r}$ for individual DNAm sites. While this study [83] was similar to our own and fit one variance component, it differed in that $\hat{h}^2_{g,r}$ was obtained for a 50KB window centred on a DNAm site. Additionally, DNAm sites genome-wide were analysed rather than those selected for location within a risk region for a brain related disorder. Moreover, we cannot directly compare the number of significant associations ourselves and the authors [83] obtained because we both used different significance thresholds. However, generally, both our study and that of Quon et al.[83] were in accordance, which highlights the utility of the RH approach to estimate the effects of local genetic variation on DNAm level. Despite the difficulties in comparing the exact results across our study and that from Quon et al. [83] (as described in the paragraph above) several similar trends emerged. Both Quon et al. [83] and ourselves found that the percentage of heritable DNAm sites was similar for all four BRs. Quon et al. [83] found that 3.0%, 3.1%, 2.9%, and 3.9% of DNAm sites were heritable for the CRBL, FCTX, PONS and TCTX respectively while we found that 9.6%, 9.7%, 9.0% and 9.5% of DNAm sites were heritable for the CRBL, FCTX, PONS and TCTX respectively (Table 7). Additionally, the modal $\hat{h}^2_{g,r}$ provided by Quon et al. [83] for all DNAm sites deemed significantly heritable was ~0.2, which is similar to our finding (Figure 7). Quon et al. [83] also examined the correlation between $\hat{h}^2_{g,r}$ across the four BRs. The results the authors [83] obtained were in accordance with our own. The TCTX and FCTX were the most highly correlated with respect to $\hat{h}^2_{g,r}$ and $\hat{h}^2_{g,r}$ from the CRBL were least correlated with the other three BRs. There are several explanations for this result. Firstly, the power to detect genetic effects could be different for the different BRs. Power can be affected by the sample size and the sample size is different for each of the BRs. The standard error of the RH estimate will be greater for a smaller sample size and smaller for a larger sample size. In turn the magnitude of the correlation between the genetic effects across two BRs could be related to the standard error of the RH estimates. Indeed, we found that the two BRs with the greatest number of samples had the highest correlation of genetic effects. Additionally, the two BRs with the lowest number of samples had the lowest correlation of genetic effects. In fact, generally the correlation of genetic effects decreased as the number of samples within the two BRs being compared decreased. Secondly, the result could be explained by an increase in shared regulatory control of the TCTX and FCTX compared to that between the CRBL and the other three BRs. For example, genes not expressed in a tissue can be associated with densely packed chromatin and extensive DNAm at the promoter region making them inaccessible to the transcription machinery [84]. In this case the genes are tightly regulated so that they are not expressed and no association between DNAm level at the gene and local genetic variation will be observed. However, if the gene is expressed the chromatin structure is open to facilitate transcription and in this case local genetic variation has an opportunity to affect DNAm level. Given that the majority of SNPs would be shared across BRs, an increase in shared open chromatin regions between two BRs would lead to an increase in the potential for genetic variation to affect DNAm level in both BRs. It is conceivable that the TCTX and FCTX share more open chromatin regions than the CRBL does with the other three BRs.

Finally, we have shown that with 100 unrelated individuals the effects of local SNPs on DNAm level can be estimated and used to accurately predict the DNAm level of the DNAm site in a separate population sample. Although we conducted our cross validations using DNAm level as the phenotype we expect the result to hold for other phenotypes that have similar genetic architecture, such as gene expression level. An implication of this result is that DNAm level, and possibly other CPs, can be imputed into currently available GWAS datasets which contain genotype data and data related to a disease. In this situation, following imputation, DNAm level (predicted from

SNP effects on DNAm level in non-diseased individuals) can be tested for association with the disease. A significant association obtained could suggest a causal role for the DNAm site in the aetiology of the disease. This is because the phenotypic variation in DNAm level at the DNAm site was present before the onset of disease. We have used this ideology and shown that phenotypic variation in DNAm level at cg04431054 measured in the TCTX was associated with FI [1]. As cg04431054 is located 277 basepairs upstream of PRRC1, it is also possible that gene expression of PRRC1 in TCTX could affect FI.

2.5 Contributions

Konrad Rawlik provided a script to conduct the quality control of the mRNA data and provided the information in 2.2.1.3 outlining the procedure.

Suzanne Rowe and colleagues [1] provided the information that the region on chromosome 5 was associated with FI and the FI dataset.

Gibbs et al. [2], the Division of Aging Biology, the Division of Geriatrics and Clinical Gerontology (NIA) were responsible for collection and initial research on the brain dataset. The brain dataset was downloaded from the NCBI Data Repositories dbGaP (Accession Number: phs000249.v1.p1) and GEO (GSE15745).

Both my supervisors provided comments on drafts of this chapter.

Chapter 3 Investigation of Trans Regulation of DNA Methylation Level

3.1 Introduction

Genetic variation is associated with DNAm level at some DNAm sites across the genome [2,54-62,83]. GWASs, have tested the effect of SNPs with DNAm level at individual DNAm sites. These studies have revealed a substantial number of associations between SNP and a local DNAm site [2,54,55,61].

In practice, for hypothesis free GWASs, long-range genetic effects on DNAm level are particularly challenging to detect. The increased number of association analyses needed to test SNPs genome-wide compared to test only SNPs local to a DNAm site requires an increased significance threshold. Assuming a given spectrum of SNP effects, an increased significance threshold will lead to a reduction in the proportion of significant associations. Therefore, in comparison to a *cis* analysis, a *trans* analysis will suffer from lower power to discover SNP associations.

Despite the lower power of a *trans* analysis compared to a *cis* analysis, GWASs have detected some long range effects on DNAm level [2,54,55] (Table 4). In addition, a GWAS showed an enrichment of observed *trans* acting effects on DNAm level beyond chance expectation. This result was visible as inflation above the 1:1 line in a uniform qqplot of genome-wide SNP effects on DNAm level at individual DNAm sites across the genome [55]. Taken together, GWASs provide evidence that some SNPs associate with DNAm level in *trans*.

There are numerous mechanisms by which genetic variation could affect DNAm level in *trans*. An example is that genetic variation affects the quality of local gene expression that in turn directly influences the DNAm level at distal DNAm sites.

This has been observed for mutations within genes that play a major role in facilitating DNA methylation, such as DNMTs [85,86] and the enzyme, Methylenetetrahydrofolate reductase (MTHFR) [87], which is involved with synthesizing methyl donors (Table 12). Genetic variation within a gene that may interact with DNMTs has also been associated with changes in global DNAm levels. Rs10876043 within DIP2B was associated with the first principal component of the autosomal DNAm levels [55]. DIP2B contains a DMAP1 domain, DMAP1 has been found to form a complex with DMNT1 and the chromatin-modifying enzyme: HDAC (reviewed in [88]). Therefore, it seems plausible that variation in DIP2B could affect the enzymatic activity of the DMNT1 complex and global DNAm level. Additionally, genetic variation could affect DNAm level in *trans* in a subtler manner than by decreasing the enzymatic function of genes that interact with or are themselves responsible for ensuring faithful conservation of DNAm. In this case, given the complex interaction between gene expression and epigenetic modifications, there are in theory multiple scenarios whereby genetic variation perturbing one variable in *cis* could lead to distal changes in DNAm level. Knowledge of genetic variation acting in *trans* to influence DNAm level is a step towards unravelling 1) long-range interactions, 2) the interplay between genetic variation, epigenetic modifications and gene expression, and 3) regulatory networks.

In this study, we tested genetic variation across the genome with DNAm level measured in the CRBL. Following on from our previous study, Chapter 2, we used the RH approach [64] to test the association of 2MB genomic regions across the genome with DNAm level at individual DNAm sites. We conducted this analysis for a subset of DNAm sites located in a locus associated with a brain related disorder/disease (see 2.2.4). In the majority of cases, when compared to a GWAS, the RH approach is thought to be a more powerful approach for detecting genetic effects (see 2.1). The increased power of the RH approach is thought to come from two main sources. Firstly, the RH approach can capture the combined effect of variants with small effect size within a genomic region. Secondly, in comparison to a

GWAS approach using the RH approach reduces the number of statistical tests. This leads to a reduction of the stringency of the multiple testing threshold.

Table 12 SNP with Distal and Widespread Effects on DNAm Level

The table indicates SNP(s) within genes that play a major role in facilitating DNA methylation. The SNP(s) within these genes are associated with distal and widespread effects on DNAm level.

Gene	Position of SNP	Frequency of SNP	Mechanism	Effect	Ref.
DNMT3B	Multiple reported in C-terminal position	Rare, ICF type 1 and usually die before adulthood	Decreased Enzymatic activity of DNMT3B	Centromeric and pericentromeric hypomethylation with classical satellite repeats	[85]
DNMT3L	Arginine to glutamine substitution in exon 10 within C-terminal position	Rare	Affects interaction with DNMT3A	Telomeric and sub telomeric hypomethylation	[86]
MTHFR	C to T substitution at position 677	Common	Decreased enzymatic activity of MTHFR which depends on folate for synthesizing methyl donors	Hypomethylation when folate levels are low	[87]

3.2 Materials and Methods

3.2.1 Defining 2MB Genomic Regions Across the Genome for Association with DNAm Level

We defined 2MB regions across the genome to test for the association with 3049 DNAm sites that were measured in the CRBL and located within a locus associated with a brain related disorder/disease (see 2.2.4). To segment the genome, the length of each chromosome was downloaded from the UCSC Genome Browser Database [89]. Due to the substantial number of statistical tests to be performed and in order to minimize our multiple testing correction threshold, we excluded genomic regions where there may have been reduced power to detect an association between the genomic region and DNAm level. To this end genomic regions with minimal genetic variants were excluded, these regions were those that covered the centromeres or contained a low number of SNPs. Centromeric regions were omitted because these regions contain satellite repeats that have not been well characterized in genome assemblies; therefore, these regions were unlikely to be represented by many SNPs on the array (reviewed in [90]). To define the 2MB regions, each chromosome was divided into 2MB regions from the start of the chromosome until the centromere and from the centromere until the end of the chromosome. A total of 56 2MB genomic regions were removed because they contained 0 SNPs. An additional count of 2, 2, and 1 genomic regions were removed because they contained 2, 3 and 6 SNPs respectively. This is because given the average extent of LD in the human genome [91] it was likely that the sparse SNPs in these regions would insufficiently tag casual polymorphism in the 2MB region. This resulted in a total of 1310 genomic regions to be tested for the association with DNAm level using the RH method outlined in 2.2.6.1. The number of SNPs within the 1310 genomic regions ranged from 11 to 1272 with a mean and median of 394 and 380 respectively.

3.2.2 Fine Mapping Methylation QTL within a 2MB Genomic Region

The RH approach and a single SNP association approach were used to fine map QTL within a 2MB region. To implement the RH approach for fine mapping, the 2MB genomic region was divided into smaller window sizes defined by a number of SNPs. We used two different window sizes, both 20 and 50 SNPs. The window sizes of 20 SNPs and 50SNPs were used for all windows within the 2MB region except the final windows, which we allowed to contain greater than the specified number of SNPs if necessary. Consecutive windows overlapped by one SNP so that all casual variation was located within a window. The RH analysis was conducted as described in 2.2.6.1.The single SNP association analysis was carried out in plink v1.06 [66] for all SNPs within the 2MB region and on the residuals of the rank transformed (Beta-value) phenotype.

3.2.3 Calculation of Lambda for each Genomic Region

The genomic inflation factor is a measure of the average extent to which test statistics from a GWAS deviate from the expected null distribution of no association between SNPs and a trait (reviewed in [92]). As illustrated by Bell et al. [55] in the context of multiple phenotypes, genomic inflation can be used as a measure of the enrichment of true positive associations. Similar to the aforementioned concept, for the association of each 2MB genomic region with 3049 DNAm sites we measured the deviation of the P-values from the distribution expected under the hypothesis of no association, the uniform distribution. To calculate lambda, we transformed the P-values to the –log10 scale and then calculated the expected quantiles from the uniform distribution. Subsequently, for each genomic region we applied a linear model to the 850 highest (-log10) P-values and calculated the regression slope. We used the 850 highest P-values because under the null hypothesis the test statistic is distributed as a 50:50 mixture of chi-squared distributions with zero and one degrees of freedom [93] , which results in a P value of 0.5 when the LRT test statistic is zero. This truncation of observed P-values at 0.5 means they are not uniformly distributed between 0 and 1 as expected under the null hypothesis. Therefore, given a number of points from which to calculate the expected quantiles of the uniform distribution (n = 3049), a subset of the data points, the most non-significant P-values, will be more significant than the expectation assuming a range from 0-1. We observed that the maximum number of P-values at the truncation point (P = 0.5) for a genomic region was 2169. Therefore, by choosing to use only the 850 (3049 – 2169 = 880 > 850) most significant P-values we do not let the inflation of P-values at the truncation point influence the calculation of lambda.

3.2.4 Permutation Analysis to Set an Empirical Significance Threshold

To determine an empirical distribution of P-values and lambda values under the null hypothesis of no association between regional genetic variation and DNAm level, we conducted 100 permutations of the phenotype data with respect to the genotype data. To parallel our original analyses, we conserved the phenotypic correlations among the observed data and the effects of the explanatory variables on the DNAm sites. To do this we permuted the DNAm level at all DNAm sites and the measured explanatory variables together for a sample. After each permutation each genomic region was tested for association with each of the 3049 DNAm sites using the RH method as outlined in 2.2.6.1. Subsequently, the empirical threshold for the association of a DNAm site with a genomic region was determined by selecting the most significant P-value from each permutation. The 95th percentile of these P-values was used to specify the significance threshold. In a similar manner, to determine the empirical threshold for lambda we used the 95th percentile of the distribution of the highest lambda from each permutation.

3.2.5 Investigating Gene Function

We queried the University of Santa Cruz Table Bowser [89] with defined genome coordinates to obtain a list of genes and gene descriptives (Table 13) for the genome coordinates.

Table 13 Information Queried From the UCSC Table Browser

The table indicates the fields that were queried using the UCSC Table Browser and explains how the fields can be interpreted.

Field	Explanation
Gene Symbol	Gene Name
Broad Phenotype	Phenotype (Disease) associated with genetic variation within the Gene
Disease Class	Category of Broad Phenotype
Narrow Phenotype	Molecular phenotype associated with genetic variation within the Gene
Publication	Name of report for the association of genetic variation within the Gene and Broad or Narrow Phenotype
Gene Description	Gene name in full
Kegg Entrez	Kegg Pathway ID and Entrez ID for the Gene
Event ID	Reactome event ID
Event Description	Reactome event explanation

3.3 Results

3.3.1 *Trans* Association of Genomic Regions with DNAm Level

We found no significant association between DNAm level and a genomic region at a Bonferroni threshold adjusted for the total number of tests conducted (n = 3994190,

$P < 1.15*10^{-8}$) or at our empirical threshold determined by 100 permutation tests ($P < 1.50*10^{-8}$). Despite the lack of statistical significance we explored if there was biological evidence to support an association between DNAm level and a genomic region in *trans*. To do this we selected the association most likely not to result from a chance correlation, our most significant result, and we investigated if the result seemed biologically plausible. The most significant association (P = 5.39*10-6) was found between chr8:110100000-112099999 and the DNAm site, cg10002103, on chr12:46766730. Cg10002103 is located in the transcription start site of solute carrier family 38 member 2 (SLC38A2) which is a gene involved with amino acid transport. The genomic region at chr8:110100000-112099999 contained several genes (Figure 10).We localized the signal to a 213,665 basepair region at 10460488-110674152 that contained three genes (Figure 11). We used the UCSC database to investigate the function of the three genes (3.2.5). However, we did not find an obvious biological role for the interaction of any of the genes with DNAm level at SLC38A2 (Table 14).

Figure 10 Chromosome 8:110100000-112099999

The region most significantly associated with DNAm level at a DNAm site (cg10002103) in trans using the RH approach. The first track (SNPS) depicts the location of SNPs present within our dataset and the second track depicts known transcripts (UCSC database). Additional tracks depict the location of common variation within the region, SNPs associated with clinical traits, enhancers, the histone modification H3K27Ac and transcription factor interactions. There are several genes within the region but no obvious connection between regulation of the genes and DNAm level at cg10002103.

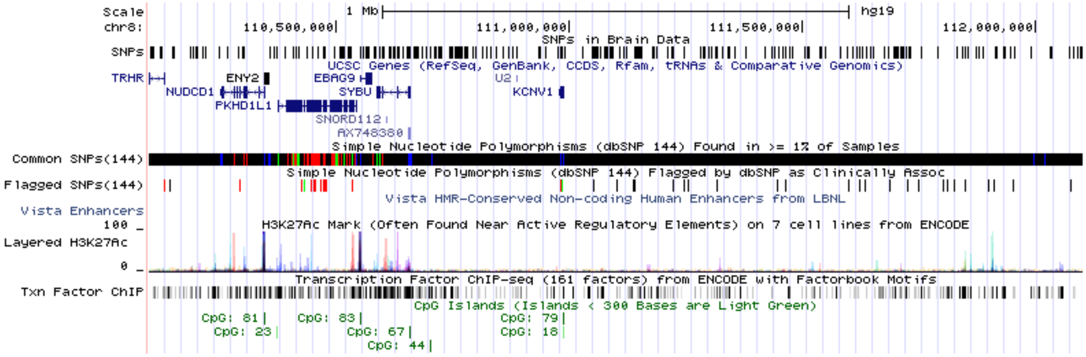

Scale
1 Mb
hg19
chr8:
110,500,000
111,000,000
111,500,000
112,000,000
SNPs in Brain Data
SNPs
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
TRHR
ENY2
EBAG9
U2
NUDCD1
SYBU
KCNV1
PKHD1L1
SNORD112
AX748380
Simple Nucleotide Polymorphisms (dbSNP 144) Found in >= 1% of Samples
Common SNPs(144)
Simple Nucleotide Polymorphisms (dbSNP 144) Flagged by dbSNP as Clinically Assoc
Flagged SNPs(144)
Vista HMR-Conserved Non-coding Human Enhancers from LBNL
Vista Enhancers
100 _
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
0 _
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Txn Factor ChIP
CpG Islands (Islands < 300 Bases are Light Green)
CpG: 81
CpG: 83
CpG: 79
CpG: 23
CpG: 67
CpG: 18
CpG: 44

Figure 11 Fine Mapping Association Signal within Chromosome 8:110100000-112099999

The association of genetic variation within chr8:110100000-112099999 with cg10002103 at chr12:46766730. The x-axis indicates the location of the genetic variation tested and the y-axis the significance of the association. Three analyses are shown in different windows, the GWAS fitted the effect of SNPs individually whereas RH 20SNPs and RH 50SNPs simultaneously fitted the effects of 20 and 50 SNPs respectively using the RH method.

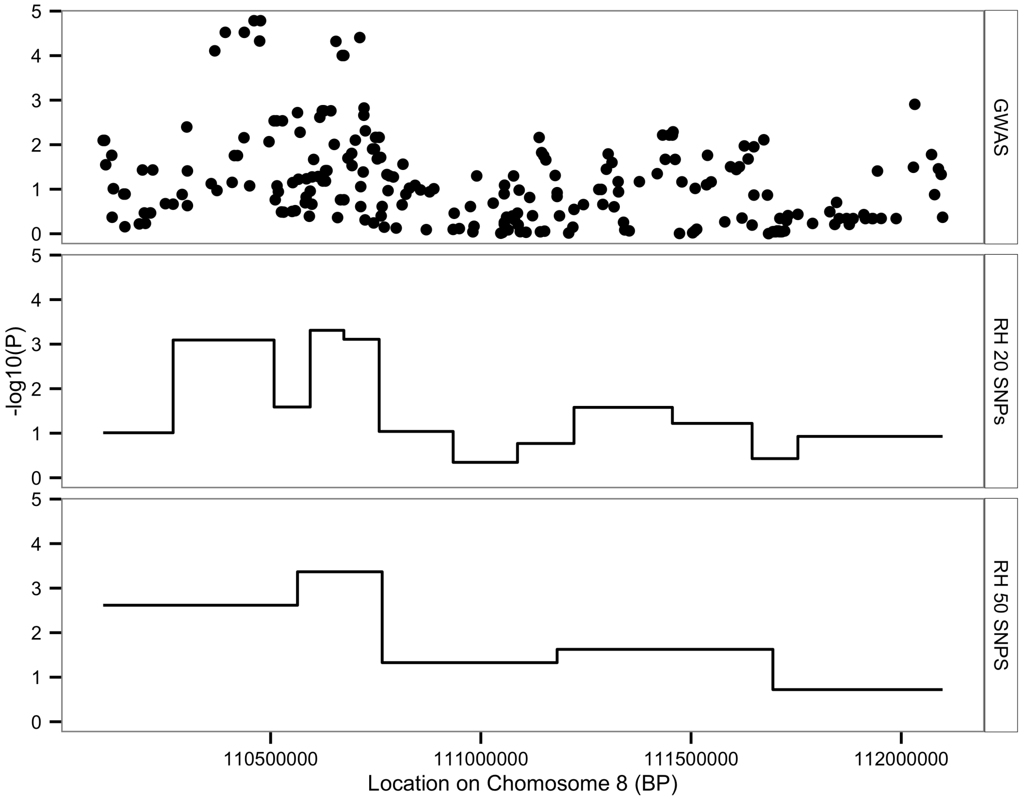


Table 14 Genes that Localize with Genetic Variation within Chr8110100000-112099999 Most Strongly Associated with Cg10002103

Genes that localize to the region of chromosome 8:110100000-112099999 that had the strongest signal for association with cg10002103. The gene information was retrieved using the UCSC table browser. There is no obvious biological connection between DNAm level at cg10002103 and the three genes based on their description.

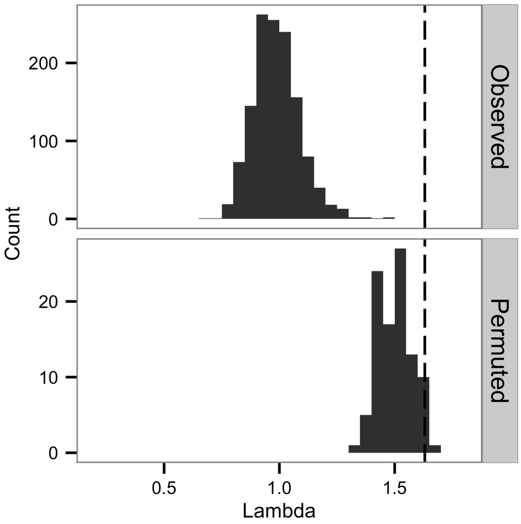
Gene	Transcribed Region	Description	Broad Phenotype
PKHD1L1	110374705:110543500	Polycystic kidney and hepatic disease 1 (autosomal recessive)-like 1	Tobacco Use Disorder
EBAG9	110551928:110577391	Estrogen receptor binding site associated, antigen, 9	Early-stage breast cancers
SYBU	110586404:110661132	Syntabulin (syntaxin-interacting) (SYBU), transcript variant 15	Cholesterol, LDL, Fibrinogen,

3.3.2 Inflation of Expected P-values for Association of DNAm Sites with a Genomic Region

Subsequently, we tested for enrichment of associations between DNAm level at DNAm sites genome-wide and an individual 2MB genomic region by calculating the variable, lambda (see 3.2.3). We hypothesized that a genomic region with a high lambda could be regulating DNAm level in *trans* at a number of DNAm sites across the genome. The distribution of observed lambda ranged from 0.69 to 1.5 with a mean of 0.99 (Figure 12). However, an empirically determined threshold of P < 0.05 set by 100 permutation tests revealed that the observed lambda statistics were not significantly different from those obtained by chance association between DNAm level and a genomic region (Figure 12).

Figure 12 The Observed and Permuted Distribution of Lambda

The "Observed" window shows the observed distribution (n=1310) of lambda values. The highest lambda value from each of the 100 permutations is plotted in the window labelled "Permuted". The dashed line shows the 0.05 threshold at Lambda = 1.63 as determined from the 100 permutations. None of the observed values for Lambda meet the empirical threshold specified by the permutation analysis.



3.4 Discussion

We did not find evidence to support the hypothesis that DNAm level is regulated by distal genetic variation. This hypothesis was tested using two different significance thresholds, firstly by applying a Bonferroni correction and secondly, by an empirical threshold determined by permutation. We used the Bonferroni method to adjust for the number of DNAm sites and genomic regions tested; however, we anticipated that this threshold would not account for phenotypic correlation among the DNAm sites. Phenotypic correlation among the DNAm sites violates the assumption of the

Bonferroni method that the statistical tests are independent. A Bonferroni threshold that assumes more tests were independent than the reality would be conservative. DNAm sites could be phenotypically correlated if they were regulated in a similar manner and DNAm sites up to 1KB exhibit similar profiles of DNAm level [27]. Given that the analysed DNAm sites were chosen based on location within defined genomic regions, risk regions for a brain related disorder/disease (see 2.2.4), it seemed plausible that some of the DNAm sites could be phenotypically correlated. Therefore, we permutated the labels of the phenotypic data maintaining the correlation structure among the DNAm sites. The significance threshold set from the permutation analyses accounted for the phenotypic correlation among the DNAm sites. The fact that the empirical threshold was only marginally more liberal than the Bonferroni threshold ($P < 1.50*10^{-8}$ and $P < 1.15*10^{-8}$ respectively) suggested that the phenotypic correlation among the DNAm sites was minimal and that in the majority of cases the statistical tests were independent.

Our result of no association between DNAm level and genetic variants in *trans* contrasts the significant *trans* acting associations found by previous authors' [2] using the same dataset and a GWAS approach. A direct comparison between our study and that of Gibbs et al. [2] based on the probability of test statistics is challenging given the use of a different statistical methodology and significance threshold. Furthermore, we note that we have used only a small subset (~11%) of the total number of DNAm sites analysed by Gibbs et al. [2]. The DNAm sites we chose to analyse were enriched for location near genetic variation that confers risk to a brain related diosorder/disease and as we showed in 2.3.2 for significant *cis* $\hat{h}^2_{r,g}$. If indeed the subset of DNAm sites we analyzed were regulated by local genetic variation then it is not wholly unsurprising that distal effects were small and went undetected. This is because the heritability cannot exceed one. Therefore, partitioning the heritability into *cis* and *trans* acting genetic effects necessitates that as the magnitude of either one of the components increases the other must decrease.

We did not find any evidence that indicated that a genomic region regulated DNAm level at a number of DNAm sites across the genome. Lambda calculated from our observed data was distributed around that expected. Initially, this was an encouraging result as it suggested that a proportion of the genomic regions exhibited enrichment for association with DNAm level. In conjunction, this conformed to our biological hypothesis that only a minority and not all genomic regions could act as *trans* regulatory hubs for DNAm level. However, permutation experiments revealed that the observed lambda likely resulted from spurious association between DNAm level and the genetic variation within a genomic region.

3.5 Contributions

Gibbs et al. [2], the Division of Aging Biology, the Division of Geriatrics and Clinical Gerontology (NIA) were responsible for collection and initial research on the brain dataset. The brain dataset was downloaded from the NCBI Data Repositories dbGaP (Accession Number: phs000249.v1.p1) and GEO (GSE15745).

Both my supervisors provided comments on drafts of this chapter.

Chapter 4 The Heritability and Patterns of DNA Methylation in Normal Human Colorectum

4.1 Introduction

Studies have shown that DNAm level is linked to gene expression level [30,31]. At promoter regions, within a population of individuals average DNAm level and average gene expression level of an associated gene were negatively correlated across genes [30]. Additionally, DNAm level has been found to associate with sex and age [31,59,94] and several environmental factors including early life socioeconomic status and stress [31]. Furthermore, results from GWASs and twin studies indicate that the genotype can affect level of DNAm [2,54-58] (Table 3, Table 4) and that *cis* acting genetic variation can explain a substantial proportion of the phenotypic variation for some DNAm sites [2,54,55].

Recently, studies have indicated that the extent to which genetic variation affects variation in DNAm level may differ within tissue depending on the functional genomic context of the DNAm sites. In peripheral blood lymphocytes $\hat{h}^2_{ped}$ of DNAm level for DNAm sites located in high and low CpG density regions of the genome was assessed [58]. This study revealed that estimates in regions of high CpG density were on average 0.127 or 0.158, whereas in regions of low CpG density estimates were greater, on average 0.235 or 0.223, depending on which probe type (Infinium I or Infinium II) was used to assay DNAm level. In human brain tissue, RH analysis revealed an increased proportion of heritable DNAm sites in regions of the genome with low CpG density compared to high CpG density [83]. In addition, a decreased proportion of heritable DNAm sites local to genes upregulated in a tissue specific manner compared to genes expressed ubiquitously across tissues was observed [83]. While the aforementioned studies have begun to explore the extent of heritability for DNAm sites located in different genomic contexts they have been conducted in a minority of tissues and have considered a limited selection of functional subgroups for DNAm sites.

To build on the work of others who have investigated the control of DNAm in different genomic contexts, we assayed 196081 DNAm sites with the HM450K in healthy colorectum tissue collected from 132 unrelated Colombian subjects who attended Colonoscopy examination and with diagnosis of hyperplasic polyp, adenoma or carcinoma. We grouped the DNAm sites based on location in relation to CpG density, expression status and functional regions of genes. We refer to these groups collectively as contextual groups. Within each contextual group we assessed the profile of DNAm level calculated across all the 132 samples at individual DNAm sites. We estimated the combined effect of local genetic variation +/ 1MB of a DNAm site on DNAm level at the DNAm site using the RH approach (see 2.2.6.1). We also studied the phenotypic profiles and $\hat{h}^2_{r,g}$ for DNAm level of genes expressed in epithelial cells obtained from laser capture microdissection (LCM) and whole colon biopsy (WCB). In addition we contrasted the distribution of $\hat{h}^2_{r,g}$ for DNAm sites within and outwith known susceptibility loci for Colorectal Cancer (CRC, OMIM #114500).

4.2 Methods

4.2.1 Ethical Approval

The study had ethical approval from the Ethics Board of The National Cancer Institute of Colombia and from The ethics committee of the General University Hospital of Elche. All participants gave informed written consent.

4.2.2 Phenotype QC

Within each colorectal tissue sample the two intensity values that correspond to the number of methylated and unmethylated copies of a DNAm were corrected for any variation that arose from non-specific binding. This background correction was applied by subtracting the median fluorescence measured by the control probes from the intensity values treating intensities measured in the two colour channels separately and using the Bioconductor package, ‘lumi’ [95]. Subsequently, samples for which the assaying process failed were identified as those with a low average intensity value (below 2500) measured in either or both of the colour channels and were removed. We examined the percentage of probes that were not detected above background levels of variation ($P = 0.01$) for each sample and found that no samples exceeded our threshold of 5% for exclusion. Samples where the recorded sex of the individual did not match the sex estimated from the levels of DNAm measured on the X chromosome were removed. DNAm probes were removed if they contained a SNP within the target sequence or at the site of single base extension or if they were cross-reactive [96]. Additionally, probes were removed if they were not detected above background levels of variation for greater than 5% of samples ($P<0.01$). Out of 144 samples and 486428 autosomal probes this procedure resulted in 132 samples and 196081 autosomal DNAm probes left for downstream analysis. Colour bias was taken into account by comparing for all samples the within sample distribution of total intensities measured by the type I probes in the green channel to those measured in the red channel. A quantile normalization adjustment was applied within the Bioconductor package, ‘lumi’ [95], so that the intensity values measured in the two colour channels followed a similar distribution across and within individuals. We also applied a correction to account for technical variation due to the probe design type using the BMIQ algorithm [97].

We conducted analyses and reported levels of DNAm using the M-value scale. This scale reduces the dependence between the variance and mean of site-specific DNAm level that is observed on the Beta-value scale [9]. M-values are a logit transformation of the Beta-values and an M-value of 0 equates to a 50% level of DNAm whereas a positive and a negative M-value relates to a greater and less than 50% DNAm level respectively.

4.2.3 Genotype QC

We followed a standard quality control procedure (reviewed in [68]) using Plink [66] and genotype data obtained from the Illumina HumanOmniExpress Exome Chip. Samples for which greater than 5% of SNPs did not genotype were excluded. Based on the application of three successive filters, SNPs were removed if 1) they failed to type in greater than 5% of samples or 2) if they were out of Hardy Weinberg Equilibrium (P<0.0001) or 3) if they had a MAF less than 0.01. Additionally, the HBD for each sample was calculated from SNPs along the X chromosome and the estimate was checked against the recorded sex to identify samples to remove because of a discrepancy between the recorded and observed identity. Out of an initial 464 samples and 944534 SNPs, 451 samples and 682945 remained for downstream analysis.

4.2.4 Testing the Effect of Explanatory Variables on DNAm Level

The explanatory variables sex and age were included in all subsequent analyses because they have previously been found to associate with level of DNAm [31,59,94]. The 132 samples with quality genotype and phenotype information consisted of 65 females and 67 males. The samples ranged in age from 30 to 84 years, with a mean and median age of 60 and 61 respectively.

We calculated the principal components from the post quality control (autosomal) DNAm levels for the 132 samples with genotype and phenotype information. We now refer to the principal components of the DNAm levels as eigenprobes and they represent a measure of global covariation in DNAm level among the samples. We then tested 5 explanatory variables: biopsy location (9 levels), diagnosis (5 levels), city of recruitment (6 levels), genotyping plate (16 levels) and position on the genotyping plate (12 levels), to determine if they could explain a significant proportion of the variance in the first 20 eigenprobes. In the majority of cases these explanatory variables were not found to significantly correlate with global patterns of DNAm level as measured by the eigenprobes (Table 15). This result and the fact that fitting these explanatory variables would lead to a substantial loss of degrees of freedom in our study meant that we chose not to include them in our downstream analysis.

Table 15 Association of Potential Explanatory Variables with Global DNAm Level

The P-values from an ANOVA testing the variance explained by the additional variables on global DNAm level measured by the principal components of the DNAm level. Location: biological location from where the sample biopsy was taken, diagnosis: stage of malignancy, city: city of recruitment, plate: genotyping chip and position: position on the genotyping chip. Italic values indicate those tests that were significant at the unadjusted level (P < 0.05) and bold values indicate tests significant at the level adjusted for the number of tests conducted for a explanatory variable (P < 0.0025).

		Explanatory Variables				
		Location	Diagnosis	City	Plate	Position
Phenotype Principal Components	C1	0.163	0.978	0.451	0.882	0.218
	C2	*0.038*	*0.044*	0.277	0.346	*0.030*
	C3	0.388	0.891	0.568	0.688	0.552
	C4	0.678	0.429	0.231	0.607	**<0.001**
	C5	*0.049*	0.215	0.837	0.863	0.053
	C6	**<0.001**	**<0.001**	0.100	0.125	*0.006*
	C7	*0.012*	0.748	0.483	0.324	0.113
	C8	*0.047*	**0.002**	0.810	0.463	0.621
	C9	0.829	0.712	*0.014*	0.676	0.190
	C10	*0.031*	*0.006*	0.706	**<0.001**	0.857
	C11	*0.048*	0.067	0.899	0.072	0.731
	C12	0.301	0.764	0.534	0.272	0.268
	C13	0.791	0.815	0.782	0.120	0.377
	C14	0.760	0.475	0.561	0.900	*0.043*
	C15	0.407	0.191	0.168	0.157	0.173
	C16	0.758	0.668	0.316	0.739	0.692
	C17	0.359	0.940	0.286	0.401	*0.011*
	C18	0.181	0.985	0.783	0.236	0.507
	C19	0.234	0.299	0.804	0.759	0.817
	C20	0.064	0.447	0.435	*0.007*	0.238

We tested if population admixture, quantified by the genotypic principal components, had a significant relationship with the variation in global DNAm levels as measured by the eigenprobes. The genotypic principal components for 132 samples and 343838 autosomal SNPs pruned for linkage disequilibrium (one of a pair of adjacent SNPs are removed if $r^2 > 0.5$ in a window of 50 SNPs, the window being shifted 5bp along and the process repeated; see plink --indep-pairwise) were calculated. We assessed if the proportion of variance in the eigenprobes explained by the genotypic PC was significant by comparison of each of three different full models (Equation 10) to one reduced model (Equation 11) for each of the first twenty eigenprobes. The three full models (Equation 10) differed only in the number of genotypic principal components included either 2,10 or 20 (*n*). In both the full and reduced model y is a vector representing the sample loading.

Equation 10

$$y = u + \beta_{sex} + \beta_{age} + \sum\nolimits_{1}^{n} \beta_{i} + \varepsilon$$

Equation 11

$$y = u + \beta_{sex} + \beta_{age} + \varepsilon$$

Our analysis revealed that neither of the three formulations of the genotypic PC explained a significant proportion of the variation in global DNAm levels (Table 16). Additionally, we found that if we included the first two principal components the estimates of RH obtained were highly similar to those obtained when the two genotype principal components were not included (r = 0.965, slope of the regression line = 0.992 for y = RH estimate without PC included and x = RH estimate with PC included). Therefore, we did not include the genomic PC in our analyses.

Table 16 Association of Genotype Principal Components with Variation in Global DNAm Level

Three combinations of genotypic principal components were tested for association with each of the first 20 DNAm principal components (eigenprobes). For instance, PC1:2 indicates that the first 2 genotype principal components were tested for association with each of the first 20 DNAm principal components, denoted by C1 through C20. The value in each cell is the P-value from the significance test. Values in italic indicate those which were significant at the unadjusted threshold (P < 0.05). Values in bold indicate a test was significant at a threshold adjusted for the number of tests conducted a genotype principal component combination (P < 0.0025).

		Genotype Principal Components		
		PC1:2	PC1:10	PC1:20
Phenotype Principal Components	C1	0.185	0.202	0.595
	C2	0.601	0.993	0.940
	C3	0.918	0.605	0.756
	C4	*0.014*	*0.003*	0.054
	C5	0.701	*0.006*	**0.001**
	C6	0.141	0.071	0.247
	C7	0.485	0.273	0.711
	C8	*0.036*	0.063	0.064
	C9	0.515	0.878	0.988
	C10	0.864	0.994	0.998
	C11	0.267	0.728	0.929
	C12	0.679	0.965	0.978
	C13	0.645	0.509	0.833
	C14	0.937	*0.013*	*0.023*
	C15	0.282	*0.003*	**0.001**
	C16	0.278	*0.004*	*0.045*
	C17	*0.010*	**<0.001**	**<0.001**
	C18	0.709	*0.024*	*0.045*
	C19	0.976	0.211	*0.033*
	C20	0.157	0.071	0.108

4.2.5 Identification of Genes Expressed in Colon Tissue

Genes expressed in general colon tissue and specifically in colon epithelial cells were identified based on the analysis of normal tissue from biopsies of 12 people undergoing colonoscopic examination at General University Hospital of Elche (Spain). In order to separate epithelial specific expression, tissue samples were sliced with alternate slices assigned to the *whole tissue* and *epithelial* conditions. In the whole tissue condition combined slices for each individual were assayed for gene expression. In the epithelial condition we pooled epithelial cells, isolated using Laser Capture Microdisection (MMI CellCutPlus), from each slice for each individual and mRNA from the slices was amplified prior to being assayed for gene expression. The gene expression assay on the 24 samples was performed using the HumanHT-12 Expression BeadChip. Quality control indicated failure of four samples (one in the whole tissue and three in the epithelial condition) which were removed from subsequent analysis. We then identified for each condition mRNA probes which were detected above background ($P<0.01$) in more than 80% of samples, i.e., 9 or more of 11 samples and 8 or more of 9 samples for the whole tissue and epithelial conditions respectively. This yielded 9223 probes in the whole tissue and 4071 probes in the epithelial conditions. Probes were mapped to genes using the Illumina provided manifest file for the HumanHT-12 Expression BeadChip platform, yielding a list of 8114 genes expressed in general colon tissue and 3754 genes expressed in epithelium. As expected a majority of genes identified in the epithelial condition were also detected in the whole tissue which contains both the epithelial and other cells, with only 10 specific to the epithelial condition.

4.2.6 Location of a DNAm site with Respect to a Gene

A DNAm site was considered to be located within the transcription start site (TSS) of a gene expressed in colon tissue if in the HM450K manifest file the site was recorded as being located within 200bp upstream of the TSS (TSS200) or, within 200-1500bp upstream of the TSS (TSS1500) of a gene in our list of expressed genes. All other DNAm sites located within the TSS200 or TSS1500 region of a gene were considered as being located within the TSS of a gene not expressed in colon tissue. Intragenic DNAm sites were those documented in the HM450K manifest file as located within the 5'UTR, 1st exon, gene body or 3'UTR of a gene. Intergenic DNAm sites were those not documented as residing within a gene. We applied successive filters in the aforementioned order so that each DNAm site fit into one of the four mutually exclusive categories.

4.2.7 Significance Testing of Proportions

To test if two proportions were significantly different from one another we used the prop.test function in R [75]. In brief, this function assumes that the two sample sizes are sufficiently large so that the distribution of the first proportion minus the second proportion is Gaussian. We applied a two-tailed test because we did not have a prior expectation of the relative magnitudes of the two proportions being tested.

4.3 Results

4.3.1 Average DNAm Level and Relationship to CpG Density, Genic Location and Gene Expression

In the manifest file for the HM450K array each DNAm site is annotated as being located either within a CpG island (island), within 2kb upstream or downstream of a island (north shore and south shore respectively), within 2-4kb upstream or downstream of an island (north shelf and south shelf respectively) or none of the aforementioned categories which we term sea. An island was defined as being composed of one or more adjacent sections of the genome each 500bp in length with a C and G density greater than 50% and an observed to expected ratio of CpG dinucleotides greater than 0.60 [98]. We grouped our 196081 DNAm sites based on aforementioned HM450K array annotation (Table 17)

Table 17 DNAm Sites Grouped by Physical Distance from Islands

The number of DNAm sites within each of six contextual groups defined by distance from CpG islands. As described in the body of the thesis, north and south shores encompassed up to 2KB upstream and downstream of islands respectively. Regions 2-4KB upstream and downstream of islands were defined respectively as north and south shelves [98]. Sea is any DNAm site not annotated as being located within an island, shelf or shore in the HM450K manifest file.

Genomic Context with Relation to CpG Density	Island	North Shore	South Shore	North Shelf	South Shelf	Sea	Total
Number of DNAm Sites	74272	27405	21158	8323	7504	57419	196081

Using DNAm level adjusted for gender and age we found a substantial difference in the distribution of average DNAm level across contextual groups of varying CpG density (Figure 13). The average DNAm level of DNAm sites in islands tended to be much lower than that of DNAm sites located in the sea (mean and median M-value was -2.67 and -3.51, and 2.01 and 1.57 for islands and sea, respectively). Additionally, our results showed that the distribution of average DNAm level for DNAm sites in the north and south shores were similar to one another and more similar to the distribution of average DNAm level for DNAm sites in islands rather than DNAm sites in the sea (Figure 13). Conversely, the distribution of average DNAm level for DNAm sites located in the north and south shelves were similar to one another and were more similar to the distribution observed for DNAm sites located within the sea rather than within islands (Figure 13). Additionally, we found that within shores the mean DNAm level was a function of the distance from the edge of the island. Mean DNAm level increased with distance from the edge of the island in a non-linear fashion (Figure 14). However, this relationship is not observed for DNAm sites located within shelves (Figure 14).

Figure 13 Distribution of Mean DNAm level with Respect to CpG Density

Methylation level was measured on the M-value scale where a DNAm level of 0 can be interpreted as a 50% DNAm level. A DNAm level < 0 and a DNAm level greater than > 0 can be interpreted as lower and greater than 50% DNAm respectively. The majority of DNAm sites in islands exhibited a low average DNAm level, whereas the majority of DNAm sites in low density CG regions (sea) exhibited a high average DNAm level.

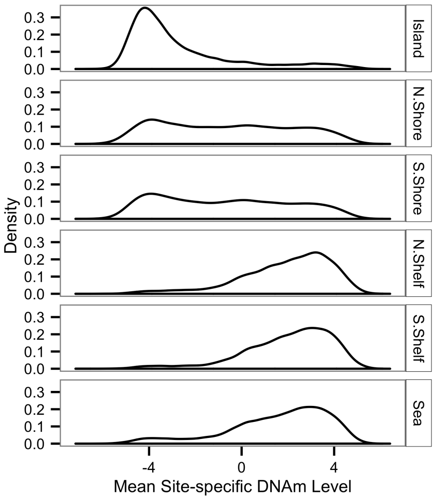
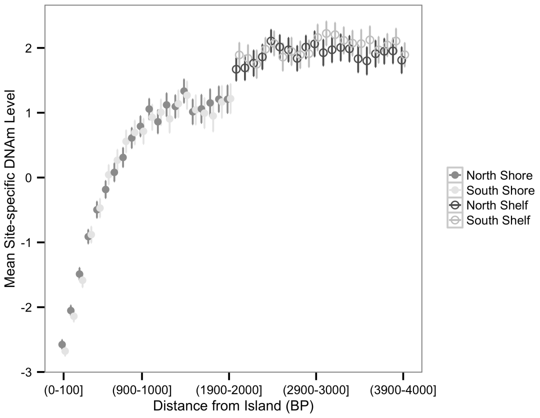


Figure 14 Mean DNAm level as a Function of Distance from the Edge of the Island

The 4000 BP region upstream (north) and downstream (south) of islands was divided into bins of 100 BP. The average of the mean DNAm levels for DNAm sites residing within each bin is shown as a white circle enclosed by a line indicating +/- 2 standard errors of the mean. A shore is up to 2000 BP from an island and a shelf is between 2000 and 4000BP from an island.



Unless otherwise specified, the following analyses were based on the two most extreme cases of CpG density: high CpG density regions (islands) and low CpG density regions (sea). We tested if the DNAm level of DNAm sites located in the TSS of genes expressed in WCB was different to those located in the TSS of genes expressed in cells from the colon epithelium collected using LCM. The genes that were expressed in the LCM and the WCB were excluded from the WCB group for this analysis. We found that DNAm sites in the TSS of genes expressed in WCB had an M-value that was on average 0.16 greater than DNAm sites located in the TSS of genes expressed in LCM (mean WCB = -3.30, mean LCM = -3.46, T-test P= $2.337*10^{-7}$).

Subsequently, we grouped DNAm sites located within the sea or an island into four mutually exclusive sets based on location a) in a transcription start site (TSS) of a gene expressed in WCB b) in a TSS of a gene that is not expressed in WCB c) in intragenic DNA, where we do not distinguish between genes expressed or not expressed in colon because the methylation level of intragenic DNAm sites has not been correlated with the expression of the surrounding gene or d) intergenic DNA (Table 18).

Table 18 Count of DNAm Sites within Each of Eight Contextual Groups

Genomic Context	Island	Sea
TSS Expressed	13838	2074
TSS Not Expressed	19051	7603
Intragenic	31355	30108
Intergenic	10028	17634
Total	74272	57419

Additionally, we choose to use the full set of genes expressed and not expressed in WCB rather than exclusively in colonic epithelial cells because DNAm level was assayed from WCB. We refer to each of the eight contextual groups individually as: island TSS expressed (within an island and a TSS of a gene expressed in colon), island TSS not expressed (within an island and in a TSS of a gene not expressed in colon), island intragenic (within an island and intragenic), island intergenic (within an island and intergenic), sea TSS expressed (within the sea and the TSS of a gene expressed in colon), sea TSS not expressed (within the sea and the TSS of a gene not expressed in colon), sea intragenic (within the sea and intragenic) and sea intergenic (within the sea and intergenic). Within each of the eight contextual groups we investigated the distribution of mean DNAm level.

We found a significant difference in the distribution of mean DNAm level between island TSS expressed and island TSS not expressed (Kolmogorov-Smirnov test; $P<2.16*10^{-16}$) and between sea TSS expressed and sea TSS not expressed (Kolmogorov-Smirnov test; $P<2.16*10^{-16}$) (Figure 15). We compared the mean of the sea TSS expressed to that of the sea TSS not expressed and we compared the mean of the island TSS expressed to that of island TSS not expressed. These two comparisons were both statistically significant (T-test, $P<2.16*10^{-16}$, $P<2.16*10^{-16}$) and in both cases being located in the TSS of genes not expressed in colon led to an overall greater mean DNAm level. Additionally, we found that the average of the distribution of mean DNAm level for DNAm sites located in intragenic and intergenic regions was greater than for DNAm sites located in a TSS of a gene (Figure 16).

Figure 15 Distribution of Mean DNAm level for Eight Contextual Groups

There are clear differences in distribution of mean DNAm level between DNAm sites located in different genomic contexts, in particular between DNAm sites located in the context of high and low CpG density.

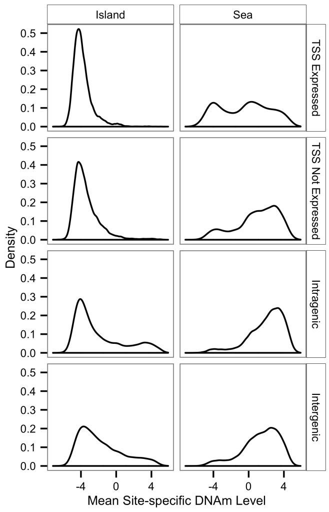
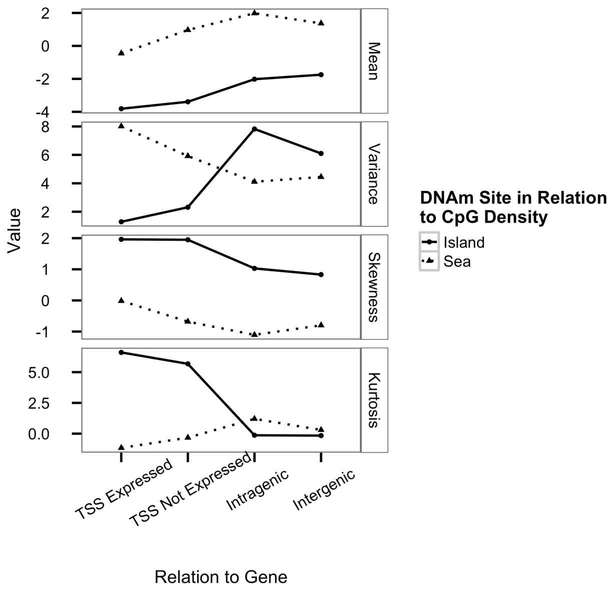


Figure 16 Moments of the Distributions of Mean DNAm level for Eight Contextual Groups

Mean DNAm level is on average higher at DNAm sites located in the TSS of a gene not expressed in WCB than at DNAm sites in the TSS of a gene expressed in WCB. Additionally, DNAm sites located intragenically or intergenically had a higher Average mean DNAm level than DNAm sites located within the TSS of a gene (expressed/not expressed) in WCB.



To assess if the phenotypic variation was adequate for downstream RH analysis, we investigated the extent to which DNAm level varied across the 132 samples for the full set of 196081 DNAm sites that passed our quality control procedure (Figure 17). The coefficient of variation (CV) ranged between 0.032 and 14889. Subsequently, we conducted exploratory enrichment analysis for the CV both within contextual groups relating to CpG density and genic location (Figure 18, Figure 19). We specified 5 different thresholds for the minimum CV (1, 1.5, 2, 2.5 and 3) to determine how sensitive our results were to the magnitude of the chosen threshold. We found that for each minimum specified CV there was enrichment for DNAm sites located in the sea compared to DNAm sites located in an island (Figure 18). Specifically and for example, at a minimum CV threshold of one, the proportion of DNAm sites located in the sea was 3.19 fold that of the proportion of DNAm sites located in an island and the difference in the two proportions was statistically significant ($P<2.16*10^{-16}$). Additionally, at each threshold within the island location we found enrichment for DNAm sites located in intragenic and intergenic regions when compared to DNAm sites located in the TSS of a gene (Figure 19). This result was not found within the sea location (Figure 19). Within the island contextual groups and at a minimum CV of one there was a 1.60 fold enrichment of intergenic over intragenic DNAm sites ($P<2.16*10^{-16}$), a 2.63 fold enrichment of intragenic to TSS Not Expressed DNAm sites ($P<2.16*10^{-16}$) and a 1.18 fold enrichment of TSS Not Expressed to TSS Expressed DNAm sites ($P = 1.50*10^{-15}$). Our analysis revealed that DNAm sites located in the sea were more variable than DNAm sites located in islands; moreover, within islands the DNAm sites within intergenic and intragenic regions were more variable than DNAm sites located in TSS regions and this result is not observed for DNAm sites located in the sea.

Figure 17 Distribution of the Coefficient of Variation

The coefficient of variation was calculated for all DNAm sites used in analysis (n=196081). The x-axis is binned in 0.02 increments and is limited to a maximum value of 3 for aesthetic purposes. Two percent of DNAm sites have a coefficient of variation greater than 3.

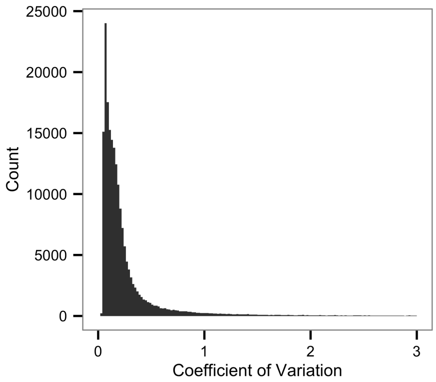


Figure 18 Enrichment Analysis for Variability of DNAm level in Islands and Sea

Five different thresholds were specified to determine how sensitive the results were to the magnitude of the threshold. The y-axis is the proportion of DNAm sites with a coefficient of variation greater or equal to the specified minimum coefficient of variation within each genomic context (island and sea).

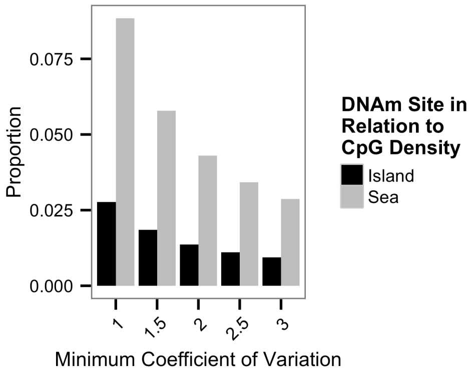
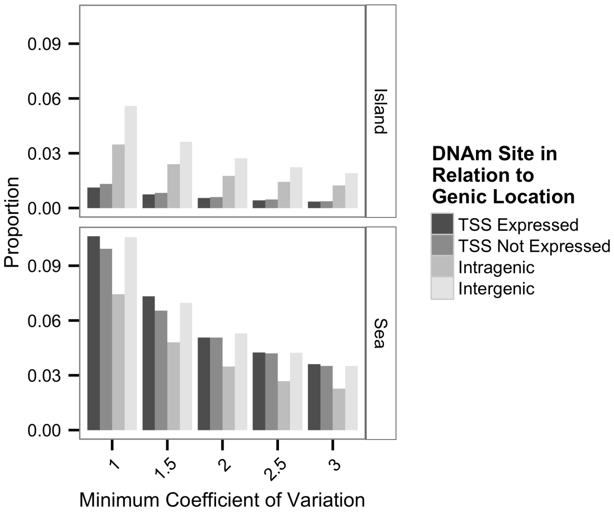


Figure 19 Enrichment Analysis for Variability of DNAm level with Respect to Eight Genomic Contexts

The proportion (y-axis) of DNAm sites within each genic location and island or sea with a coefficient of variation greater than the threshold defined on the x-axis

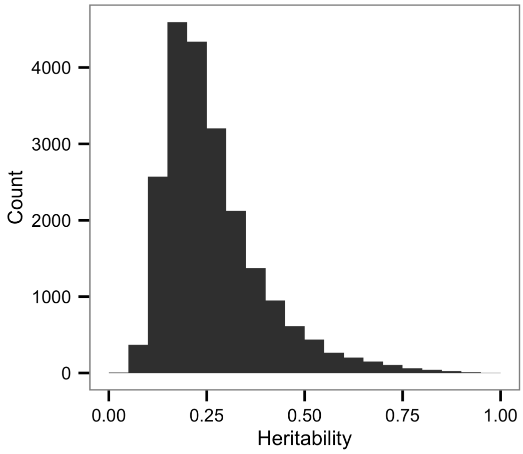


4.3.2 Heritabilities of DNAm Level at Individual DNAm Sites

We found that at 21447 DNAm sites, 10.94% of the 196081 tested, SNPs within 1MB explained a significant proportion of the variation in DNAm level (nominal P < 0.05). The percentage of heritable DNAm sites exceeded that expected from a false positive rate of 0.05 under the null hypothesis that DNAm level is not associated with local genetic variation. For significantly heritable loci, the proportion of the phenotypic variance in DNAm level under local genetic control ranged between 0.04 and 0.99 with a mean of 0.27 and median of 0.23 (Figure 20).

Figure 20 Estimated Regional Heritability for Heritable DNAm Sites.

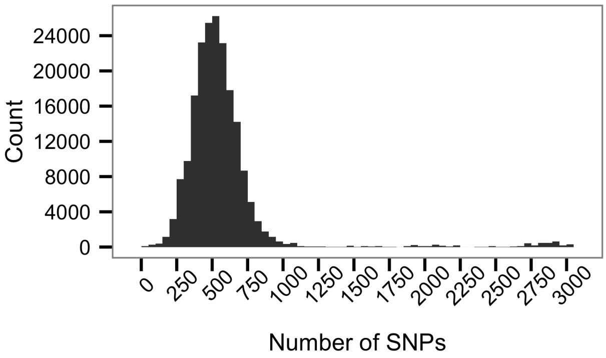
The distribution of local RH estimates for the DNAm sites with a significant RH (P <0.05). Each bar represents a range of 0.05.



We found that the number of SNPs in the local genomic region (Figure 21) explained a minute but significant proportion of the variance in RH estimates for the DNAm sites with a significant RH (Univariate Linear Regression: $R^2 = 0.004$, slope = $2.17*10^{-5}$, $P < 2.2*10^{-16}$). For instance, considering the range of the number of SNPs within a region (1 to 3037), at the first decile (304 SNPs) and ninth decile (2733 SNPs) a respective $6.60*10^{-3}$ and $5.94*10^{-2}$ RH for the DNAm sites found to be significantly associated with local genetic variation is expected. We did not find that the variance of DNAm level at each DNAm site, as measured by the CV, was related to the variance in RH estimates (Univariate Linear Regression: P =0.068).

Figure 21 SNPs in the Local Genomic Region.

Distribution of the count of SNPs +/- 1MB surrounding the DNAm sites analysed.

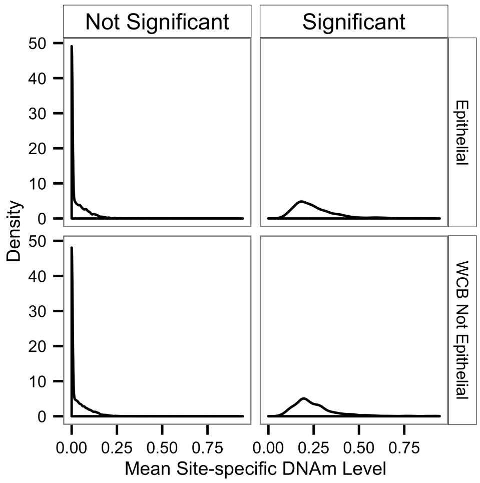


4.3.3 Heritability of DNAm Level at genes expressed in Whole Colorectal Biopsies and Colorectal Epithelial

We investigated the heritability of DNAm level at genes expressed in LCM and in WCB excluding those expressed in LCM (Figure 22). The difference between the average of the distribution of mean DNAm level for the significantly heritable DNAm sites within the two groups was not significant (mean LCM = 0.251, mean WCB = 0.255, T-test P =0.638). Additionally, the proportion of significantly heritable DNAm sites in the LCM and WCB group was 0.0768 and 0.0770 respectively and was not significantly different from one another (P = 0.986).

Figure 22 Heritabilities of DNAm Level at Genes Expressed in Whole Colorectal Biopsies and Colon Epithelial Cells

DNAm sites were grouped based on the significance of the RH estimate and their location with respect to a gene expressed in whole colorectal biopsy or a gene expressed in epithelial cells by laser capture microdissection. DNAm were significantly heritable if P<0.05. Genes expressed in both the epithelial and whole colorectal biopsy were removed from the whole colorectal biopsy group for this analysis.



4.3.4 Heritability of DNAm Level in Whole Colorectal Biopsies by Genomic Context

The proportion of sites with a significant heritability was higher in the sea than in islands ($P < 2.20*10^{-16}$, Table 19). This result was driven by the difference between DNAm sites located within the TSS of a gene or in intragenic regions. The proportion of heritable sites is 1.55 times higher for DNAm located in the TSS of the sea than in the TSS of an island ($P < 2.2*10^{-16}$); additionally, the proportion of heritable DNAm sites is 1.16 times higher for DNAm sites located in the sea intragenic than the island intragenic contextual group ($P = 7.27*10^{-10}$). There was no significant difference in the proportion of heritable DNAm sites located in intergenic regions when comparing between the sea and island contextual groups (P=0.068).

The proportion of heritable DNAm sites was significantly higher for sea intergenic than sea intragenic (P = 3.19*10-10) and was significantly higher for sea TSS not expressed than sea intragenic ($P = 1.11*10^{-3}$). There was no significant difference in the proportion of heritable DNAm sites for the remaining comparisons made within the sea contextual groups. The proportion of heritable DNAm sites was significantly different for all comparisons made within the island contextual groups. The P-value was $< 2.2*10^{-16}$ for all comparisons, except the Island TSS Expressed and Island TSS Not Expressed comparison where the P-value was = $4.37*10^{-7}$, and the island TSS expressed and island intragenic comparison where the P-value was $4.87*10^{-4}$. Within the island contextual groups the proportion of heritable DNAm sites was as follows: intergenic > intragenic > TSS Not expressed > TSS Expressed.

Table 19 Proportion of Heritable DNAm Sites and the Corresponding Mean RH Estimate

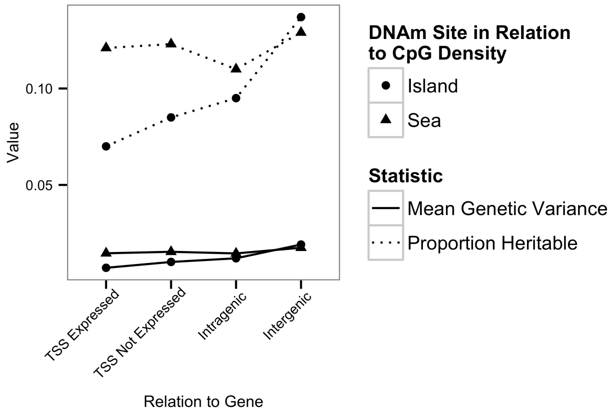
Overall there were a higher proportion of heritable DNAm sites (P < 0.05) located in the sea compared to islands. Additionally, there were a higher proportion of heritable DNAm sites located in intergenic regions compared to regions containing a TSS and intragenic regions. The average RH estimates were similar across the contextual groups.

	Island		Sea	
	Proportion Heritable	Mean $\hat{h}^2_{r,g}$	Proportion Heritable	Mean $\hat{h}^2_{r,g}$
TSS Expressed	0.070	0.250	0.121	0.263
TSS Not Expressed	0.085	0.257	0.123	0.271
Intragenic	0.095	0.260	0.110	0.261
Intergenic	0.137	0.281	0.129	0.277
Total	0.093	0.262	0.118	0.268

The proportion of heritable DNAm sites for each contextual group followed a similar pattern to the mean local genetic variance for each contextual group (Figure 23). Across each four sea and island contextual groups the proportion of heritable DNAm sites was correlated with the mean local genetic variance (Pearson's correlation: r = 0.802 and r = 0.999 for sea and island respectively).

Figure 23 Mean Genetic Variance and the Proportion of Heritable DNAm Sites for Eight Contextual Groups.

The x-axis represents the proportion of heritable DNAm sites (P < 0.05) within each contextual group and the average genetic variance for each contextual group. The proportion of heritable DNAm sites was correlated with the mean local genetic variance.



The average $\hat{h}^2_{r,g}$ for significantly heritable DNAm sites located within each of the genomic contexts was similar (Table 19). DNAm sites significantly associated with local genomic variation and located within an island were on average 0.6% less heritable than DNAm sites located in the sea.

4.3.5 Heritability of DNAm Level in Whole Colorectal Biopsies with Respect to Loci Associated with Colorectal Cancer

An extensive number of genetic variants have been found to associate with complex disease including CRC [99]. However, in the majority of cases how the identified genetic risk variants act to increase disease susceptibility is unknown. Genetic variation could increase risk to disease by mediating changes in DNAm level in healthy tissue. If an association between a susceptibility SNP and DNAm level in healthy tissue is obtained, one possibility is that the variation in DNAm level interacts with additional variables, such as environmental factors, to subsequently lead to disease. Therefore, we sought to determine if variation in DNAm level in healthy colon tissue was associated with genetic variation that incurs susceptibility to CRC. To this end a total of 83 unique autosomal SNPs identified as associating with CRC were downloaded from the NHGRI GWAS catalogue [99]. We defined a region of +/- 1MB surrounding each SNP associated with CRC as a risk region. A total of 10469 DNAm sites were located within a defined risk region and due to our definition of a risk region the calculation of $\hat{h}^2_{r,g}$ included the effects at the position of the risk SNP. Indeed, we found that 10.6% of DNAm sites located within risk regions were heritable ($P < 0.05$) with an average $\hat{h}^2_{r,g}$ of 0.262. Outwith a risk region, 11.0% of DNAm sites were significantly heritable ($P < 0.05$) with an average $\hat{h}^2_{r,g}$ of 0.267. The proportion of DNAm sites that were significantly heritable within and outwith a risk region was not significantly different ($P=0.27$) and the average heritability of DNAm sites located within and outwith risk regions was similar. In conjunction, a recent study [100] found that DNAm levels of two DNAm sites measured in healthy colorectal tissue, cg15193198 and cg24112000, were associated ($FDR < 0.05$) with the local CRC risk variant rs4925386 located on chromosome 20 at 60921044 basepairs. In our study, both cg15193198 and cg24112000 were significantly heritable with respective $\hat{h}^2_{r,g}$ estimates of 0.34 ($P = 2.05*10^{-4}$) and 0.57 ($P = 4.14*10^{-12}$) when including rs4925386 in the calculation of the genetic relationships and 0.34 ($P = 2.64*10^{-4}$) and 0.57 ($P = 6.74*10^{-12}$) when excluding rs4925386. We determined that rs4925386 explained 77% and 53% of the estimate, $\hat{h}^2_{r,g}$, for cg15193198 and cg24112000 respectively and $\hat{h}^2_{r,g}$ was still significant for cg24112000 when fitting rs4925386 as a fixed effect (Table 20).

Table 20 Effects of rs4925386 and Local Genetic Variation on cg15193198 and cg24112000

The estimate for the full model ($\hat{h}^2_{r,g}$ Full) and reduced model ($\hat{h}^2_{r,g}$ Reduced) were calculated from all SNPs +/-1MB of the DNAm site excluding rs4925386. The full model included fitting the genotypes at rs4925386 as a fixed effect. The effect of rs4925386 on DNAm level is reported as the addition of a single copy of the minor allele, Adenine.

	$\hat{h}^2_{r,g}$ Full	$\hat{h}^2_{r,g}$ Full P-value	$\hat{h}^2_{r,g}$ Reduced	$\hat{h}^2_{r,g}$ Reduced P-value	SNP Effect	SNP Effect SE
cg15193198	0.075	0.211	0.333	$2.64 *10^{-4}$	-0.423	0.083
cg24112000	0.269	$3.64*10^{-3}$	0.569	$6.74*10^{-12}$	-0.560	0.088

4.4 Discussion

DNAm sites located in different genomic contexts with respect to CpG density and genic location exhibited unique phenotypic profiles in the human colorectum. We have shown that average DNAm level is related to CpG density. This result has been observed for DNAm level measured at promoters in peripheral blood mononuclear cells and fibroblasts [31,101] and concurred with that found by a recent study profiling DNAm level in 17 somatic tissues [102]. In addition, we found that specifically within shores that there is a shift in average DNAm level from predominantly unmethylated to methylated as distance increases from the edge of the CpG dense islands. This change in DNAm level is suggestive of a transitional zone at the edge of islands captured by the definition of shore. Overall, the lower and less variable average DNAm level at individual DNAm sites located in islands compared to sea is consistent with the traditional view that CpG dense regions of the genome persist due to low methylation and a reduced rate of spontaneous deamination and transition that is typically higher for methylated CpG dinucleotides [103]. Moreover, within islands we observed enrichment for DNAm sites located in TSS with low coefficient of variation (and conversely enrichment for intergenic and intragenic DNAm sites with high coefficient of variation). The low coefficient of variation and low level of DNAm for DNAm sites located in CpG dense regions of the genome and in TSS supports the idea that these DNAm sites target housekeeping genes [101,104]. Housekeeping genes are essential for normal cell maintenance and thus expression of these genes may be tightly regulated and this could be reflected by the low level and low variation of DNAm level at individual DNAm sites in these regions. Additionally, DNAm sites located in the TSS of a gene not expressed in WCB were on average more methylated than DNAm sites located in the TSS of a gene that was expressed in WCB. This result is suggestive of an overall inverse correlation between mean gene expression and mean DNAm level. Finally, our examination of the relationship of DNAm level with genic location revealed that DNAm level was lower in the TSS than within intergenic or intragenic regions. This result concurred with what has been observed in H1 embyronic cells where DNAm level has been shown to decrease between the promoter and 5'UTR region before increasing through the gene body and into the 3'UTR [26]. The level of DNAm has also been shown to be greater in intragenic and intergenic regions compared to promoter regions in human brain frontal cortex grey matter [105].

We have assessed, on a genome-wide scale, the local heritability of DNAm level at individual DNAm sites in normal WCB using unrelated Colombian individuals. A total of 10.94% of DNAm sites in WCB were significantly affected by local genetic variation. The mean $\hat{h}^2_{r,g}$ for the heritable sites was 0.27 but the estimates varied substantially with some DNAm sites exhibiting a low heritability and some DNAm sites exhibiting heritability close to one. The implication is that DNAm level can be inherited through the germ-line. These results were consistent with previous estimates of the number of DNAm sites across the genome affected by local genetic variation and with the wide range of heritability estimates reported for DNAm [83]. Indeed, we found that that there were an increased proportion of significantly heritable DNAm sites located in the sea compared to islands. Overall, our finding that heritable DNAm sites were enriched for location outside of islands is in accordance with what is observed in human brain tissue [83].We hypothesize that the substantial difference in the mean estimates of $\hat{h}^2_{ped}$ for DNAm sites located in islands and in sea obtained in peripheral blood lymphocytes [58] and outlined in the background section of this paper may have resulted from a) the inclusion of all DNAm sites rather than just those with a significant heritability estimate b) the use of the pedigree to estimate the contribution of the whole genome to phenotypic variance and/or c) bias due to un-modelled sources of environmental variation. Additionally, the highest proportions of significantly heritable DNAm sites were located in intergenic regions as opposed to within the TSS of a gene or intragenic regions. The estimated heritability is not related to the phenotypic variability of the DNAm site as measured by the CV and is not solely a function of the number of the SNPs within the genomic region. However, the proportion of heritable DNAm sites was correlated with the average estimated genetic variance. Therefore, the lower proportion of heritable DNAm sites observed for a contextual group(s) such as islands compared to other contextual group(s) such as the sea can in part be explained by a reduction in the measured genetic variance. Lower genetic variation could result in lower power to capture the true causative loci or it could be indicative of lower causal variation due to selective constraints.

DNAm level was slightly higher at the TSS of genes expressed in WCB and not LCM compared to those expressed in LCM. This result is consistent with the WCB samples being enriched for epithelial cells and a negative correlation between gene expression and DNAm level. However, the overall RH of DNAm level at the TSS of genes expressed in WCB and not LCM compared to those expressed in LCM was similar. One possible explanation for these results is that in healthy colonic tissue the genes expressed in the colon epithelium are regulated by DNAm level in a similar fashion to the genes expressed in the WCB.

Finally, we have shown that genetic variants in genomic risk regions for CRC can affect DNAm level in healthy colon tissue and that overall DNAm sites within a risk region have similar overall RH to DNAm sites outwith an identified risk region. In conjunction, we have replicated the previous finding that the CRC risk SNP rs4925386 effects DNAm level at cg15193198 and cg24112000.
We showed that when rs4925386 is excluded the regional genetic variation sufficiently captured the causal variation in DNAm level tagged by rs4925386. Moreover, rs4925386 alone did not capture all the genetic variance contributing to variation of cg24112000 and cg15193198 that is captured by the RH approach. This final result highlights the advantage of the RH approach to capture the genetic effects on the phenotype, in this case DNAm level.

There are several shortcomings of our study. Firstly, we did not attempt to quantify the effects of genetic polymorphism outwith +/- 1MB of a DNAm site on variation in DNAm level. These long range effects were likely to exist, and studies of larger sample size that can overcome the burden of multiple testing are necessary to detect their effects. Secondly, our study has utilized healthy colon tissue from multiple locations within the colon and from Colombian subjects attending colonoscopy examination and with diagnosis of hyperplasic polyp, adenoma or carcinoma. We have shown that diagnosis and location of biopsy had effects on genome wide profiles of DNAm level (principal components of DNAm levels); however, the reduced sample size would not allow us to estimate heritability for each of the diagnoses and biopsy location accurately.

We have identified a subset of DNAm sites genome-wide and measured in healthy colon tissue that are influenced by the local genetic variation. Therefore, we have contributed to understanding healthy genetically influenced phenotypic variation in DNAm level in colon tissue. A number of the DNAm sites which we report as heritable are located within CRC risk loci and thus have the potential to mediate genetic susceptibility to CRC. We expect further studies will focus on exploring a role for these DNAm sites in disease aetiology.

4.5 Contributions

Gustavo Hernandez-Sanchez, Maria Carolina Sanabria and Martha Serrano-Lopez collected the Colombian cohort and prepared the biological samples. Konrad Rawlik performed the analysis of the gene expression data and wrote 4.2.5 which describes that procedure. Jose Luis Soto, Adela Castillejo, Cristina Alenda and Eva Hernandez-Illan collected the samples from Spain and performed the LCM and sample preparation. James Prendergast, in addition to the aforementioned collaborators and my supervisors, read a draft of a manuscript pertaining to the work in this chapter.

Chapter 5 Across Tissue Comparison of Genetic Effects on DNA Methylation Level

5.1 Introduction

Studies utilizing DNAm level as a quantitative trait have investigated the contribution of factors effecting variation in DNAm level. Research assessing the genetic basis of inter-individual variation in DNAm level has revealed that 1) the genomic heritability and the family-based estimates of heritability are tissue and site-specific and 2) a substantial proportion of the variation in DNAm level can be explained by local genetic factors [2,54-58,83]. In addition, intra-individual variation in DNAm level between tissues has been observed. In fact, intra-individual variation between tissues appears to be greater that inter-individual variation within a tissue and samples clearly cluster by tissue rather than by individual [102,106,107]. Moreover, human embryonic stem cells (ES) and ES cell derivatives, primary cells, long-term cultured cells and diseased cells can be identified from principal component analysis of genome-wide DNAm levels [29]. In conjunction, the similarity of DNAm level at DNAm sites across the genome is related to the function of the tissues being compared. For example, tissues taken from the same organ system typically have more similar DNAm levels across the genome than tissues taken from different organ systems [108]. In addition to identifying broad differences in the phenotypic profile of tissues, multiple DNAm sites in a region of the genome typically less than 500bp in length [29,107,108] or containing up to 50 DNAm probes [102] that exhibit a DNAm profile distinctive to a type of tissue (tissue-specific differentially methylated regions, TS-DMRs) have been identified. Moreover, individual DNAm sites that vary across tissues have been identified. For instance, one study that comprised 17 different tissues taken from 4 samples found that 17.1% of DNAm sites were either constitutively hypermethylated (Beta-value > 0.9) or hypomethylated (Beta-value < 0.1) across all tissues and samples and that the remaining 82.9% of DNAm sites exhibited variability [102]. A second study identified that 15.4% of DNAm sites had a difference in DNAm level of at least 0.3 on the Beta-value scale in at least one of 18 tissues [108]. The difficulties in obtaining a tissue sample from many vital human tissues have meant that studies of DNAm level tend to be small and to consider a minority of tissues. However, taken together, existing research shows that DNAm level at some DNAm sites varies across tissues.

A locus with a different genetic effect across tissues can be thought of as exhibiting a genetic interaction with the environment where the tissue represents the environment. Genetic by environment (tissue) interactions across tissues have not been well characterized for DNAm level. Recently a study [108] found evidence for difference in allele-specific DNAm (ASM) across fat, gastric, psoas, small bowel and spleen tissue. This study [108] was limited to three samples for each tissue but the results suggested that genetic effects may be tissue specific.

In this study we contrasted local genetic effects on DNAm level between healthy colon tissue and tissue taken from four regions of the human brain using two population samples of unrelated individuals. One population sample, defined as the colon dataset, consisted of 132 Colombian individuals assayed for healthy colon tissue (see Chapter 4). The second population sample, the brain dataset, consisted of 148 individuals of CEU ancestry assayed for DNAm in the cerebellum (CRBL), frontal cortex (FCTX), pons (PONS) and temporal cortex (TCTX) (see Chapter 2). We aimed to determine if there is evidence that genetic variants have a different effect on DNAm level measured in colon and brain tissue.

5.2 Materials and Methods

5.2.1 Data Quality Control

Colon data

We used the processed colon dataset data as outlined in Chapter 4.

Brain data

We used the brain dataset processed as outlined in Chapter 2, however, we used the full set of DNAm sites that pass the quality control procedure rather than a subset located within a risk region for a brain related disorder.

We used DNAm level measured on the M-value scale in both the colon dataset and brain dataset. This scale (M-value), as mentioned in 4.1 is a logit transformation of the proportion of probes that are methylated on the microarray that target a DNAm site [9]

5.2.2 Estimation of the Regional Genomic Heritability

The $\hat{h}^2_{g,r}$ for each DNAm site was estimated using the RH approach as outlined in 2.2.6.1

As previously described within the brain dataset we adjusted for the explanatory variables sex, age, post mortem interval, study, and assay plate. Within the Colon dataset we adjusted for sex, age, and the first two genotype principal components. We fitted the first two genotype principal components in the RH of DNAm level in colon tissue in this chapter merely because it is conventional to use the genotype principal components to adjust for population structure and we have an admixed population sample. Although, in 4.2.4 we have shown that we do not expect the genotype principal components to explain a significant proportion of the variation in

DNAm level in the colon dataset. Therefore, although we fitted the first two genotype principal components in this chapter and not in Chapter 4, we do not expect that the adjustment for these two covariates will have substantially affected the magnitude of $\hat{h}^2_{g,r}$.

5.2.3 Meta Analysis of the Regional Genomic Heritability Across Tissues

To test the hypothesis that genetic variation within a genomic region has a significant effect on DNAm level at an individual DNAm site across multiple tissues we used Fisher's method [109] for combining independent P-values. The probability of rejecting the null hypothesis that DNAm level is heritable in multiple tissues is related to the probability of DNAm level being heritable within each of the individual tissues. Assuming that each tissue represents an independent test of the heritability of DNAm level at a given DNAm site, the P-value from Fisher's method test statistic is distributed chi-squared with degrees of freedom equal to 2 multiplied by the number of P-values being combined. However, if the assumption of independent tests does not hold then the logic of Fisher's method does not apply. Repeated sampling from up to four brain tissues from one individual present in the brain dataset could result in covariance in DNAm level among multiple brain tissues. Therefore, given that the test of heritability of DNAm level may not be independent across the brain tissues we do not combine the P-values from all five tissues together. Instead, we conducted an analysis using Fisher's method on the significance of the RH result for DNAm sites in colon tissue and one brain tissue at a time. To reiterate, we applied Fisher's method to the four pairs of datasets, colon and CRBL, colon and FCTX, colon and PONS, colon and TCTX. In each case, the –log10(P) of the RH estimate of the DNAm site measured in the two tissues was used to calculate the natural log(P) and Fisher's method was applied with four degrees of freedom. This procedure is a meta-analysis for the RH of DNAm level across tissues assuming that DNAm level at a DNAm site is the same trait measured in the two tissues.

5.2.4 Identity of DNAm Sites with Respect to CpG Density and Genic Location

We used the manifest file accompanying the HM450K DNAm array to determine location of a DNAm site with respect to CpG density and genic location. A description of the information provided in the HM450K manifest file was provided in 4.2.6. We determined that out of the 12355 DNAm sites measured in colon and brain tissue 3914 mapped to multiple genic locations. As described in the results section, we conducted analyses excluding these DNAm sites and without excluding these DNAm sites. We did this to determine if DNAm sites that mapped to multiple genic locations affected the enrichment analysis that we conducted. We found the same trend of enrichment with and without the exclusion of these DNAm sites.

5.2.5 Testing Regional SNPs for Association with Local DNAm Level

We used the Wald Test implemented in the software Plink [66] to test the effect of individual SNPs on DNAm level within a single tissue. DNAm level is regressed on the SNP genotype and the Wald test is used to test the null hypothesis that the regression coefficient is not significantly different from zero. Rejection of the null hypothesis implies that the alternative hypothesis is true and that the SNP has an effect on DNAm Level. We specified the significance thresholds we used in the results section.

5.2.6 Determining if Genetic Effects were Different Across Tissues

To determine if SNP effects were different across two tissues we compared the regression coefficient at the SNP across the two tissues using Equation 12 [110].

Testing this against the standard normal distribution provided an asymptotic P-value for the significance of the difference in the SNP effect across the two datasets.

Equation 12

$$Z = \frac{b_1 - b_2}{\sqrt{{SEb_1}^2 + {SEb_2}^2}}$$

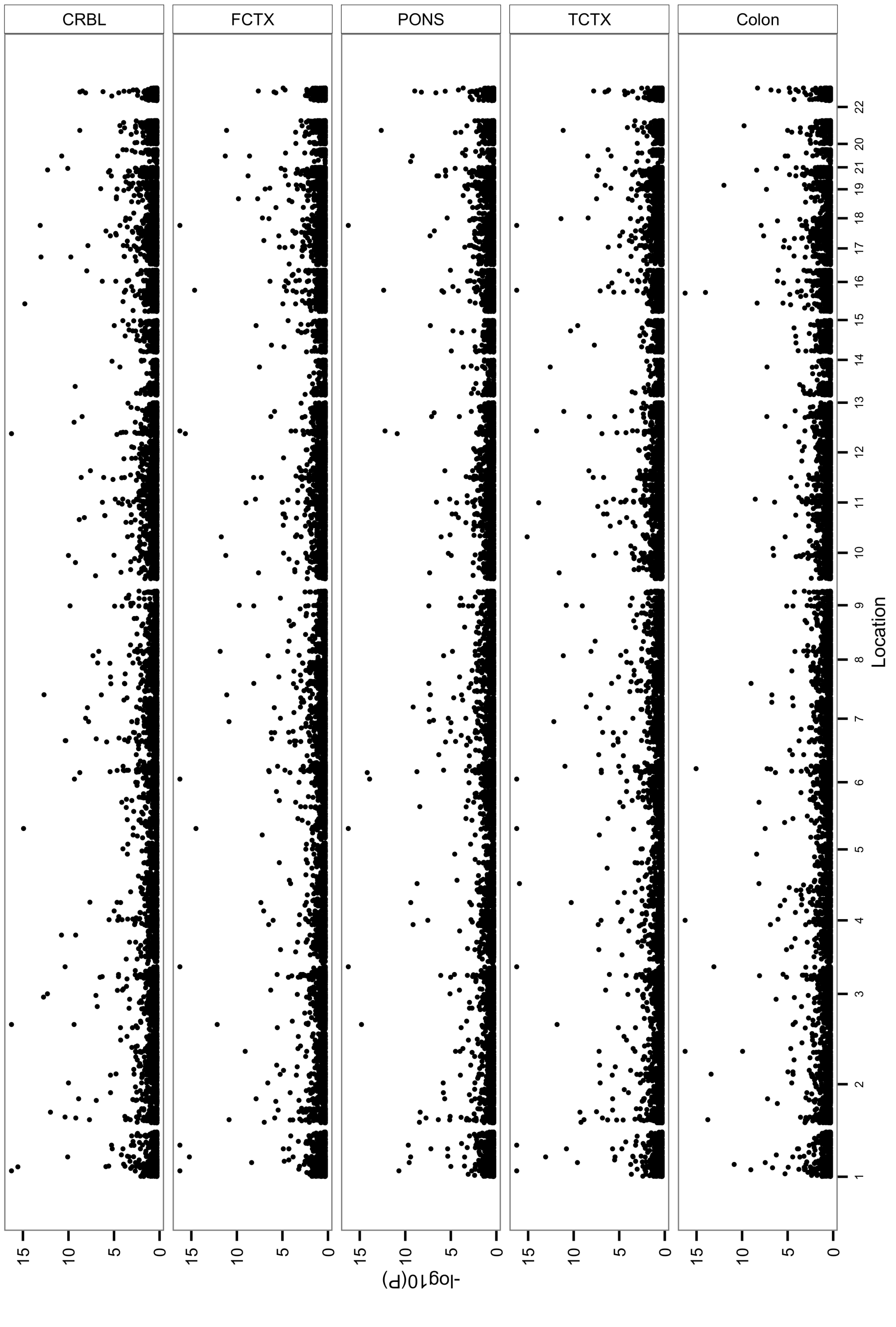
5.3 Results

5.3.1 Regional Genetic Effects on DNAm Level Shared Across Colon and Brain Tissue

A total of 12355 DNAm sites passed quality control procedures in all 5 tissues. Within each tissue and for each of the 12355 DNAm sites we took a RH approach and partitioned the phenotypic variance into that attributable to local (+/- 1MB) genetic effects surrounding the DNAm site (Figure 24).

Figure 24 Significance of Regional Genomic Heritability for 5 Tissues

The figure shows the –log10(P) from the cis RH analysis for a total of 12355 DNAm sites across the genome and common to the five tissues. The –log10(P) for each of the five tissues is plotted separately.



The correlation of the significance of $\hat{h}^2_{g,r}$ (Table 21) ranged between 0.20 and 0.71 indicating that genetic effects on some DNAm sites may be shared across tissues. Additionally, we found that the correlation of the –log10(P) of the estimates of $\hat{h}^2_{g,r}$ were much lower between colon and a brain tissue than between two brain tissues (Table 21). Therefore, less genetic effects may be shared across colon and a brain tissue than across two brain tissues. Shared genetic effects across tissues could reflect common biological regulation of DNAm level. It seems plausible that the regulation of DNAm level is more similar within tissues of similar function, such as brain tissue, than across tissues with different function, such as brain and colon tissue. Alternatively, there may appear to be increased sharing of genetic effects across brain tissues compared to across brain and colon tissue due to shared technical variation within the brain dataset. The shared technical variation within the brain dataset could have resulted from using the same set of individuals to obtain a sample from each brain tissue or from using similar experimental protocols to obtain a measurement of DNAm level.

Table 21 Pearson correlation of the significance (-log10(P)) of the regional genomic heritability between tissues

The correlation of the significance of the RH estimates is lower between colon and a brain tissue than between two brain tissues

	CRBL	FCTX	PONS	TCTX	COLON
CRBL					
FCTX	0.45				
PONS	0.45	0.61			
TCTX	0.43	0.71	0.63		
COLON	0.20	0.30	0.30	0.32	

To identify DNAm sites heritable in colon or a brain tissue, or both colon and a brain tissue we used Fisher's method and combined the P-values from the RH analysis for the colon and CRBL tissue, colon and FCTX tissue, colon and PONS tissue and colon and TCTX tissue. We did not combine the four brain tissues together with colon tissue because as stated in section 5.2.3 the assumption of independent tests may be violated due to repeated sampling from multiple brain tissues for an individual. We found that depending on the two tissues being combined between 11-12 % of DNAm sites were heritable (Table 22). The average of $\hat{h}^2_{g,r}$ was comparable at $P < 0.05$ for colon and each of the brain tissues.

Table 22 Genomic Regions Heritable in Colon and a Brain Tissue

The table shows the results from Fisher's method conducted on colon and each brain tissue separately. A total of 12355 DNAm sites were tested for association with the local genetic variation in each of the tissues and the significance was combined across colon and each of the brain tissues using Fisher's method. The table gives the count and proportion of DNAm sites significant after application of Fisher's method using an unadjusted for multiple testing threshold and a threshold adjusted for the number of traits tested (n=12355).

		P<=0.05	Adj. P<=0.05
Colon and CRBL	Count (Proportion) Regions	1467 (0.12)	127 (0.01)
	Average RH Estimate colon, CRBL	0.20, 0.23	0.32, 0.51
Colon and FCTX	Count (Proportion) Regions	1460 (0.12)	115 (0.01)
	Average RH Estimate colon, FCTX	0.20, 0.18	0.36, 0.43
Colon and PONS	Count (Proportion) Regions	1380 (0.11)	114 (0.01)
	Average RH Estimate colon, PONS	0.21, 0.18	0.37, 0.45
Colon and TCTX	Count (Proportion) Regions	1459 (0.12)	143 (0.01)
	Average RH Estimate colon, TCTX	0.20, 0.19	0.34, 0.46

Subsequently, we focused on using the colon and FCTX tissue for all downstream analyses. We selected the FCTX brain tissue as we have the largest sample size for this brain tissue; therefore, heritability estimates may be the most precise for this brain tissue.

Next we investigated the genic location and CpG density at DNAm sites that were and were not heritable using $\hat{h}^2_{g,r}$ from the Fisher's method analysis with the colon and FCTX tissue. At our stringent threshold (corrected for the number of traits analysed) we found that the proportion of heritable DNAm sites were depleted within CpG islands compared to DNAm sites that were not heritable ($P = 1.56*10^{-11}$, Figure 25). Conversely, the proportion of heritable DNAm sites were 1.96 fold enriched for location at the edge of CpG Islands (Shores) and 1.06 fold enriched for location in areas of low CpG density (Sea) compared to DNAm sites not found to be heritable. At a nominal significance ($P < 0.05$) threshold we found a very similar trend of enrichment (results not shown). When we excluded DNAm sites that mapped to multiple genic locations and when we used a stringent significance threshold to adjust for the number of traits tested, we found that the proportion of heritable DNAsites were 1.50 fold enriched 200-1500bp upstream of a TSS (TSS1500) of a gene compared to DNAm sites not significant in colon and FCTX tissue ($P = 3.04*10^{-4}$) (Figure 26). Again, we found a similar trend of enrichment at a nominal significance threshold ($P < 0.05$) and when we did not exclude DNAm sites that mapped to multiple genetic locations (results not shown). Thus, DNAm sites with a significant and non-significant RH localize to different genomic regions.

Figure 25 CpG Density and RH Significance for DNAm Level measured in Colon and FCTX Tissue.

The proportion of DNAm sites significantly (Adj. P < 0.05) or not significantly associated with local genetic variation across colon and FCTX tissue with respect to CpG Density. Heritable DNAm sites across colon and FCTX tissue are depleted within the CpG dense islands.

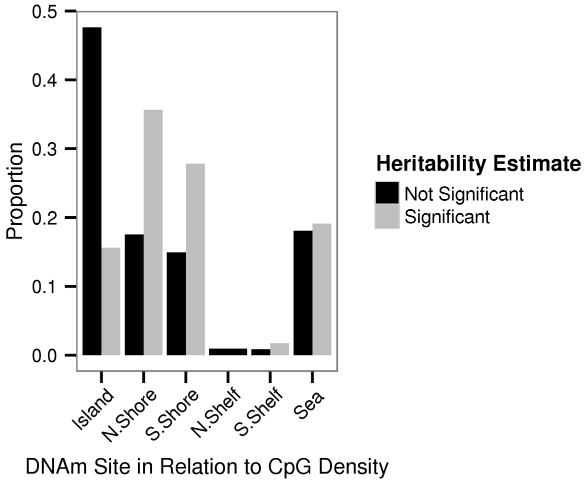
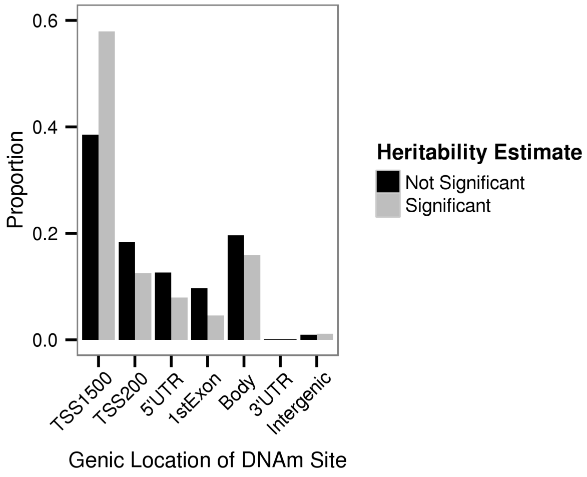


Figure 26 Genic Location and RH significance for DNAm level measured in the colon and FCTX tissue.

The proportion of DNAm sites significantly (Adj. P < 0.05) or not significantly associated with local genetic variation in colon and FCTX tissue with respect to genic location. DNAm sites that mapped to multiple locations were removed. DNAm sites that are heritable across colon and FCTX tissue are enriched within 200-1500 basepairs upstream of a TSS.



5.3.2 Location, Significance and Magnitude of SNP effects within Heritable Regions Across Colon and FCTX Tissue

We have identified genomic regions where genetic variation affected local DNAm level across colon and FCTX tissue. Subsequently, within the heritable genomic regions we wanted to narrow down the associated genetic variation to individual SNPs. Therefore we assessed the association of each SNP within the regions that were heritable (regions were deemed heritable at a significance threshold adjusted for the number of DNAm sites tested Adj. $P < 0.05$, n=115, Table 22) with the corresponding DNAm site. Between 43-691 SNPs (mean = 207 median = 200) per genomic region were common to both the datasets and were tested for association with a DNAm site. A total of 23828 SNPs were tested in each of the two tissues. In the initial analysis below we used the residual DNAm level as the phenotype. We found that SNP effects closer to the DNAm site were more significant with a steep rise of the –log10(P) starting at approximately 500KB from the DNAm site (Figure 27, Figure 28). We also observed a decrease in the effect size as distance from the DNAm site decreases, particularly within the first 500KB from the DNAm site (Figure 29, Figure 30).

Figure 27 Significance of SNP Effect and Distance from DNAm site

SNPs within +/-1MB of a DNAm site with a significant $\hat{h}^2_{g,r}$ (Adj. P < 0.05) across colon and FCTX tissue were tested for association with the DNAm site. The –log10(P) from the association tests are shown as a function of the distance of the SNPs from the DNAm sites. Results from the colon and FCTX tissue are plotted on the graph. A negative and positive distance indicates a distance upstream and downstream of the DNAm site respectively. SNP effects on DNAm level are higher when the SNP is closer to the DNAm site.

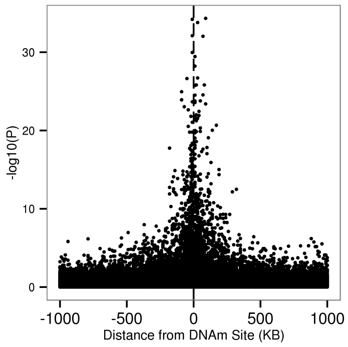


Figure 28 Average Significance of SNP Effects and Distance from DNAm site

The average –log10(P) from the SNP association analyses is calculated within bins of 10KB from the DNAm site. SNP effects on DNAm level in colon are on average less significant than SNP effects on DNAm level measured in FCTX tissue.

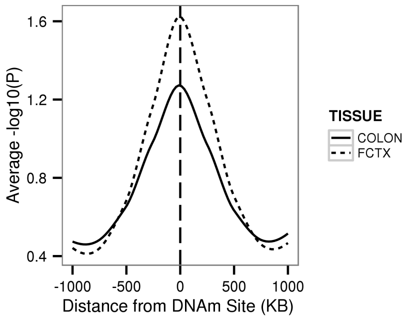


Figure 29 SNP Effect and Distance from the DNAm site

The absolute value of the estimated SNP effects from the SNP association analyses is shown as a function of the distance of the SNP from the DNAm site. There is an increase in higher effects as the distance between the SNP and the DNAm site decreases.

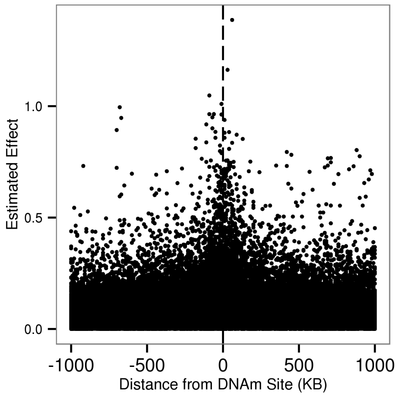
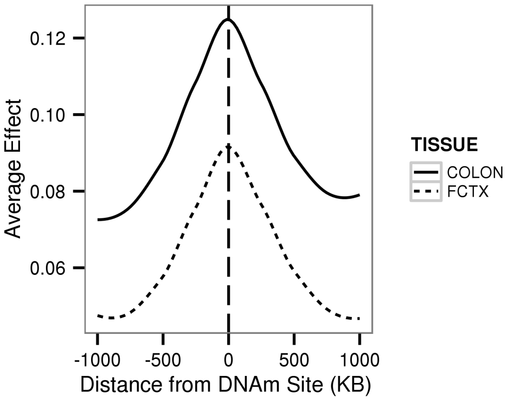


Figure 30 Average SNP effects and Distance from the DNAm site

The average SNP effect is measured within each bin of 10KB from the DNAm site and shown as a function of the distance between the SNP and the DNAm site. The average SNP effect on DNAm level is consistently different for DNAm level measured in the two tissues.



We observed that the average SNP effect within 10KB bins +/- 1MB of the DNAm site, was consistently larger in colon tissue than in brain tissue (Figure 30). The systematic difference in the average SNP effect on DNAm level between colon and FCTX tissue could reflect a true biological difference of the effect of SNPs in the two tissues. For instance, the environmental variance may be lower for DNAm level measured in the colon tissue than in the FCTX tissue. An alternative explanation for the systematic difference in the average the of SNP effects on DNAm level measured in colon and FCTX tissue is that DNAm level is measured on a different scale in the two tissues. Although, the measurements of DNAm level were on the M-value scale in the colon and brain datasets, DNAm level was measured with a different array in the two datasets (HM27K and HM450K). The two arrays differed in the chemistry used to obtain a measurement of DNAm level. In conjunction, different quality control procedures were conducted to reduce technical variation within the two datasets. Therefore, there was a risk that the systematic difference in average SNP effects on DNAm level between colon and FCTX tissue resulted from a difference in scale of the phenotype across the two tissues. Therefore, we investigated the distribution of residual DNAm level (DNAm level after the adjustment for the explanatory variables named in 5.2.2) at the individual DNAm sites analysed (n=115) within colon and FCTX tissue. We found that the mean of the distribution of mean residual DNAm level was smaller for colon tissue (-0.79) than for FCTX tissue (-0.66). Additionally, we found that the mean of the distribution of the standard deviation of residual DNAm level was greater for colon tissue (0.54) than for FCTX tissue (0.34). The difference in the residual phenotype across the two tissues could be indicative of DNAm level being on a different measurement scale in the two tissues.

Figure 31 Mean Residual DNAm Level in colon and FCTX Tissue

Distribution of mean residual DNAm level across samples for each DNAm site. The average of each distribution is different, smaller for colon than for FCTX tissue.

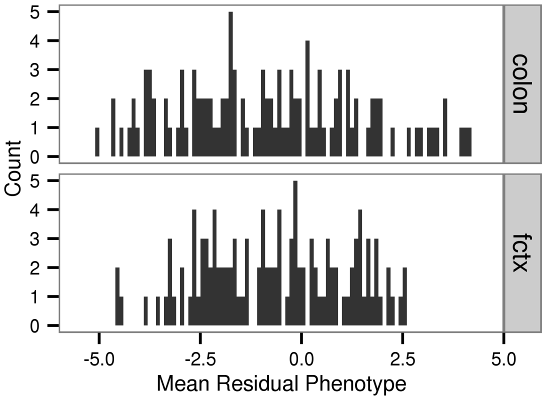
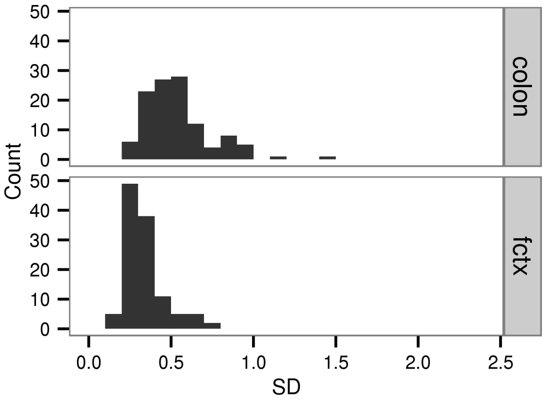


Figure 32 Standard Deviation of Residual DNAm Level in colon and FCTX Tissue

Distribution of the standard deviation of residual DNAm level across samples for individual DNAm sites. DNAm level measured in the colon tissue is on average more variable than DNAm level measured in the FCTX tissue.



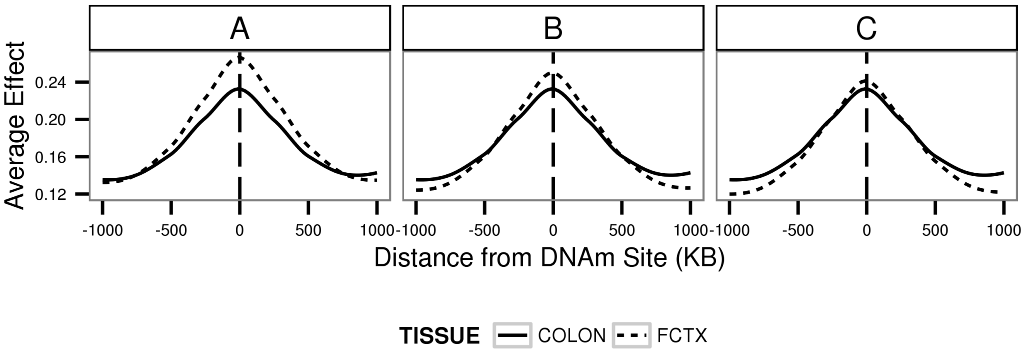
We aimed to compare individual SNP effects on DNAm level measured in colon and brain tissue. It did not make sense to compare individual SNP effects across colon and the FCTX tissue given the systematic difference in average SNP effects between the two tissues. The systematic difference in average SNP effects would likely lead to a widespread difference in individual SNP effects across the two tissues. Therefore, to minimize the systematic difference we rank normal transformed the residual DNAm levels at a DNAm site within each tissue. After rank normal transformation residual DNAm level at each DNAm site was distributed with a mean of zero and variance of one.

After rank normal transformation of the residual DNAm level at a DNAm site within each tissue we tested the effect of the SNPs on the rank normal transformed DNAm levels. We observed that a small systematic difference in the average estimated SNP effects in the two tissues remained (Figure 33A). This difference is likely due to a systematic difference in the magnitude of the phenotypic variance prior to rank transformation. A different phenotypic variance prior to rank normal transformation would lead to the estimated effect being rescaled by a different factor.

So that we could meaningfully compare individual SNP effects on DNAm level measured in colon and FCTX tissue we applied a second normalization procedure to further reduce the difference in average estimated SNP effect size across the two tissues (Figure 33). In this case, we rescaled the SNP effects so that average SNP effect in the two tissues was the same. To do this we adjusted the SNP effects in the FCTX by a constant factor that was the ratio of the average absolute value of the SNP effects in colon to the average absolute value of the SNP effects in the FCTX. We tried two alternative procedures for this adjustment and selected the procedure that preformed the best, that is the procedure that reduced the difference in average estimated SNP effect size across the two tissues the most. In the first case we included all the SNP effects and in the second case we included only the effects of SNPs located within +/- 500KB of the DNAm site. We used only SNPs located within +/- 500KB of the DNAm site because we observed that this region surrounding the DNAm site exhibited the largest systematic difference in average SNP effect between the two tissues (Figure 33). Moreover, there were substantially more significant SNP effects within +/- 500KB of the DNAm site than outside +/- 500KB of the DNAm site (Figure 27); therefore, this region is of particular interest for studying a difference in genetic effect across tissues. We found that the systematic difference in average SNP effect was reduced the most within +/-500KB of the DNAm site when we used the second of the two aforementioned procedures, that is when we used only SNP effects within +/-500KB of the DNAm site to rescale the SNP effects in the FCTX (Figure 33). Therefore, we used the second procedure to rescale the SNP effects and the standard error of the SNP effects in the FCTX (for SNPs within the 115 genomic regions significantly heritable across colon and FCTX tissue at our adjusted threshold $P < 0.05/12355$). This rescaling of the SNP effects and standard error of the SNP effects in FCTX tissue reduced the systematic difference in average SNP effect on DNAm level measured in colon and FCTX tissue. Now we could meaningfully make pairwise comparisons of the effects of individual SNPs across colon and FCTX tissue.

Figure 33 Recalibration of SNP effects

We rank normal transformed residual DNAm level at each DNAm site and in colon and FCTX tissue to ensure DNAm level was distributed equally in the two tissues. After rank normal transformation of residual DNAm level in colon and FCTX tissue a systematic difference in average SNP effects between the two tissues remained (Panel A). To minimize this difference the SNP effects in the FCTX tissue were rescaled by a constant factor. We used all SNPs within +/- 1MB (Panel B) or within +/- 500KB (Panel C) for the recalibration. The second recalibration procedure (Panel C) preformed the best at reducing the systematic difference in average SNP effects between the two tissues.



5.3.3 Comparison of within Tissue SNP Effects on DNAm Level Across Colon and FCTX Tissue

We aimed to compare the effect of individual SNPs across colon and FCTX tissue. For this analysis, we used the SNP effects and the standard error of the SNP effects that had been rescaled in the FCTX tissue. Regression analysis of the SNP effects in the FCTX tissue on the SNP effects in colon tissue revealed a highly significant ($P < 2*10^{-16}$) positive association between the SNP effects (+/- 500KB of the DNAm site) in the two tissues ($R^2 = 0.18$). This suggested that some SNP effects were shared across the two tissues. Next we wanted to contrast individual SNP effects across the two tissues to establish if there was evidence of tissue specific SNP effects. To this end, we first selected the SNPs significantly associated with DNAm level in either or both of the tissues. Secondly, we tested if the effect of those SNPs was significantly different in the two tissues as described in 5.2.6. In doing the first step, we found that 546 SNPs were significantly associated with DNAm level in at least one of the two tissues at a stringent Bonferroni significance threshold that adjusted for the total number of SNPs examined within the 115 +/- 1MB regions ($P < 0.05/23828$). In the second step, we tested if the effect of 546 SNPs was significantly different across the two tissues using a stringent threshold that adjusted for the total number of tests conducted ($P < 0.05/546$). We found that 17 SNPs had a significantly different effect in colon and FCTX tissue (Table 23,Table 24). These SNPs were associated with nine DNAm sites that represented eight loci across six chromosomes (Table 23). Within each of the eight loci the SNPs were consistently associated with DNAm level in either colon or FCTX tissue but not both tissues (Table 24). This means that within a tissue and a locus all the SNPs were either, positively, negatively or not associated with DNAm level at the local DNAm site. Additionally, at each locus the within tissue association of the SNPs with DNAm level was significant only for DNAm level measured in one of the two tissues.

Table 23 Genomic Location of Tissue Specific SNP Effects on DNAm level

The table shows location of the DNAm site and SNP pairs where the SNP was found to exhibitg a significantly (Adj. P < 0.05) different effect across colon and FCTX tissue. Eight loci across six chromosomes are represented. The distance between the SNP and the DNAm site is also provided with a negative and positive value indicating the SNP was upstream and downstream of the DNAm site.

Locus	CHR	DNAm Site	DNAm Site BP	SNP	SNP BP	DIST (BP)
1	1	cg14356550	18808102	rs3007733	18800911	-7191
	1	cg14356550	18808102	rs2992753	18808292	190
	1	cg14356550	18808102	rs10907313	18814565	6463
2	1	cg27535305	53392650	rs11206043	53402552	9902
3	1	cg08380539	85039702	rs11163981	84953415	-86287
	1	cg08380539	85039702	rs2945152	85048641	8939
	1	cg08380539	85039702	rs2945145	85062615	22913
	1	cg08380539	85039702	rs2911597	85065769	26067
	1	cg08380539	85039702	rs11164023	85089402	49700
4	6	cg09548084	8436218	rs12210654	8747817	311599
5	12	cg25229172	96336121	rs6538691	96353507	17386
6	14	cg24603941	39703302	rs2038281	39915197	211895
7	15	cg01789267	45545111	rs1706767	45569099	23988
8	18	cg09052983	52495605	rs12457718	52499626	4021
	18	cg22105582	52496070	rs4477812	52357470	-138600
	18	cg22105582	52496070	rs9964078	52384590	-111480
	18	cg22105582	52496070	rs12457718	52499626	3556

Table 24 Strength of Effect of Tissue Specific SNP Effects on DNAm level

Values in bold indicate a significant within tissue SNP effect on DNAm level. SNPs within each of the eight loci are either significantly associated with colon or FCTX tissue but not both tissues.

	DNAm Site	SNP	Effect Colon	SE Colon	P-value Colon	Effect FCTX	SE FCTX	P-value FCTX	Z Score GXE	P-value GXE
1	cg14356550	rs3007733	1.12	0.10	**2.20E-20**	-0.11	0.12	3.69E-01	-7.82	5.37E-15
	cg14356550	rs2992753	1.14	0.10	**2.33E-21**	-0.11	0.12	3.56E-01	-7.99	1.38E-15
	cg14356550	rs10907313	0.88	0.15	**3.63E-08**	-0.15	0.13	2.48E-01	-5.21	1.85E-07
2	cg27535305	rs11206043	0.29	0.12	1.40E-02	1.00	0.08	**8.10E-26**	5.09	3.65E-07
3	cg08380539	rs11163981	0.07	0.12	5.40E-01	0.97	0.08	**9.76E-24**	6.25	4.03E-10
	cg08380539	rs2945152	0.15	0.13	2.43E-01	1.01	0.07	**4.62E-27**	5.95	2.70E-09
	cg08380539	rs2945145	0.14	0.13	2.81E-01	1.02	0.08	**3.63E-25**	5.90	3.67E-09
	cg08380539	rs2911597	0.16	0.13	2.04E-01	1.02	0.08	**3.63E-25**	5.68	1.36E-08
	cg08380539	rs11164023	0.20	0.13	1.32E-01	1.06	0.09	**2.08E-20**	5.37	7.80E-08
4	cg09548084	rs12210654	-0.13	0.13	2.91E-01	-0.93	0.11	**4.55E-14**	-4.76	1.96E-06
5	cg25229172	rs6538691	-0.24	0.12	5.18E-02	0.60	0.11	**2.86E-07**	5.10	3.41E-07
6	cg24603941	rs2038281	0.29	0.14	3.89E-02	-0.67	0.10	**6.54E-10**	-5.61	1.99E-08
7	cg01789267	rs1706767	-0.78	0.10	**4.40E-12**	0.03	0.13	7.87E-01	4.96	6.95E-07
8	cg09052983	rs12457718	0.09	0.13	4.92E-01	0.95	0.10	**3.49E-16**	5.19	2.13E-07
	cg22105582	rs4477812	-0.15	0.13	2.61E-01	0.82	0.13	**5.45E-09**	5.26	1.47E-07
	cg22105582	rs9964078	-0.08	0.13	5.38E-01	0.85	0.13	**2.06E-09**	5.05	4.48E-07
	cg22105582	rs12457718	-0.09	0.13	4.83E-01	1.00	0.10	**2.48E-18**	6.69	2.31E-11

5.3.4 Bioinformatics Investigation of methQTL with Tissue Specific Effects

We used online databases to investigate the function of the eight loci that showed statistical evidence of tissue specific genetic effects on DNAm level. We aimed to determine if there was a biological reason why the eight loci appeared to be regulated differently across colon and FCTX tissue and to validate our findings. We identified the gene closest to the DNAm site and information pertaining to the gene using the NCBI Gene database (http://www.ncbi.nlm.nih.gov/gene/) (Table 25,Table 26). Using the location of our methQTLs, we also identified the closest published SNP in the NHGRI-EMBI Catalgoue of Published Genome-wide Association Studies and the associated trait (Table 27) ([99]). Subsequently, we discuss the loci where a biological interpretation of the statistical evidence for tissue specific genetic effects on DNAm level were most evident, based on the information obtained from querying the online databases.

Within Locus 2 both the DNAm site and associated SNP, rs11206043 were located within SCP2. SCP2 is a lipid transfer protein (Table 25). Individuals deficient in SCP2 may be categorized as having the neurological condition: Leukoencephalopathy with Dystonia and Motor Neuropathy (OMIM #613724) (Table 25). Additionally, A SNP 180KB downstream of rs11206043 has been found to associate with the neurological condition: Hippocampal Atrophy (Table 27). We found that the addition of the minor allele of rs11206043 significantly increased DNAm level at SCP2 in non-diseased FCTX tissue but not in colon tissue (Table 24). Assuming that disease arises from variation in DNAm level measured in a tissue relevant to disease, our result support the idea that this locus is involved in the aetiology of brain related traits. Additionally, rs11206043 has been found to associate with gene expression of SCP2 in Lymphoblastoid cells; however, strength of this association was only suggestive as it was below the standard genome-wide association study significance threshold (Table 26). However, the aforementioned finding suggests that the influence of rs11206043 on cellular phenotypes may not be limited to brain tissue.

Table 25 Genes Local to DNAm Sites with Tissue Specific QTL

The gene processes inferred from Gene Ontology [111] is given in column 3, for instance if multiple related processes were found for a gene the over arching process was reported. Relationship between the gene and a trait of clinical importance as listed in the NCBI Genetic Testing Registry [112] or inferred from publications is listed in column 4.

Loci	Gene Symbol	Gene Name	Gene Process	Gene and Trait Relationships
1	KLHDC7A	Kelch domain containing 7A	Protein ubiquitination	-
2	SCP2	Non-specific lipid-transfer protein	Metabolic Processes and Transport	Deficiency of SCP2 is Leukoencephalopathy with dystonia and motor neuropathy
3	CTBS	chitobiase, di-N-acetyl-	Chitin and Oligosaccharide Catabolic Processes	SNP in CTBS linked to multiple myeloma
4	SLC35B3	Solute carrier family 35, member B3	Metabolic processes, Catabolic Processes and Transmembrane transport	Predominately expressed in colon cells [113] and have a function in cancer proliferation [114]
5	CCDC38 AMDHD1	Coiled-coil domain containing 38 Amidohydrolase domain containing 1	Not Found Cellular nitrogen compound metabolic process and Histidine catabolic process	-
6	MIA2	Melanoma inhibitory activity 2	Cholesterol and Triglyceride Homeostatsis	-
7	SLC28A2	Solute carrier family 28	Nucleobase-containing compound metabolic process, Transmembrane transport	-
8	RAB27B	Member RAS oncogene family	Melansome and protein transport Signal Transduction, Positive regulation of Exocytosis	Expression of RAB27B is marker for gastrointestinal stromal tumors [115] and breast cancer [116]

Table 26 Published eQTL Local to Our Tissue Specific methQTL

Published eQTL results were found within 5 of the 8 tissue specific methQTL. In addition to being a methQTL in our study, one SNP (bold), was found to have a marginally significant effect on gene expression within the same locus and in Lymphoblastoid cells.

Locus	SNP ID	SNP CHR:BP	Probe CHR:BP	P-value	Tissue	Ref
1	rs12144656	1:18792299	1:18807424	$2.26*10^{-6}$	Liver	[117]
2	**rs11206043**	1:53402552	1:53392948	$4.10*10^{-5}$	Lymphoblastoid Cells	[118]
3	rs17110590	1:84978431	1:85020222	$4.75*10^{-5}$	Liver	[117]
5	rs10735337 rs7307113	12:96329933 12:95959233	12:96260827 12:96337071	$3.87*10^{-12}$ $1.29*10^{-3}$	Liver Liver	[117]

Table 27 Significant GWAS Results Local to SNP Associated with Tissue Specific DNAm Level

Distance between the GWAS SNP and the closest SNP associated with tissue specific DNAm level in our study is shown. A negative and positive distance indicates the GWAS hit was upstream and downstream of our SNP, respectively.

	GWAS Hit				
Locus	Distance (BP)	SNP ID	Trait	P-value	Ref
1	- 5656	rs3007729	Diabetic Retinopathy	$5*10^{-6}$	[119]
2	+179118	rs3820201	Hippocampal Atrophy	$1*10^{-6}$	[120]
3	-84454	rs604708	IgG Glyosylation	$8*10^{-6}$	[121]
4	+250994	rs2064197	Glucose Homeostasis Traits	$7*10^{-6}$	[122]
5	-82079	rs1036429	Pulmonary Function	$1*10^{-7}$	[123]
6	-894692	rs10498345	Coronary Spasm	$9*10^{-7}$	[124]
7	-894692 -69052	rs2453533 rs765787	Kidney Diseases Uric Acid	$5*10^{-22}$ $3*10^{-6}$	[125] [126]
8	+253074 +252391	rs4801131 rs12966547	Schizophrenia Combined Neurological and Psychiatric Traits	$1*10^{-8}$ $3*10^{-10}$	[127] [128]

In locus 4, rs12210654 was associated with level of cg09548084 in colon but not brain tissue (Table 24). Cg09548084 is located within SLC35B3, which is preferentially expressed in colon tissue (Table 25). Our result can be explained by assuming that DNAm level is linked to gene expression and chromatin structure and that only genetic variation can affect DNAm level and gene expression when the chromatin structure is open. For example, open chromatin facilities gene expression and can provide an opportunity for genetic variation to perturb level of gene expression. In contrast, closed chromatin inhibits the molecules necessary for transcription and translation to access the DNA, thus genetic variation in these regions may not influence the level of local gene expression. Given these circumstances, an expressed gene would more likely be affected by genetic variation than a gene that is switched off. Low or no expression of SLC35B3 in tissues other than colon could indicate that this gene is switched off in these tissues. In comparison, expression of SLC35B3 in colon tissue follows from an open chromatin structure that is susceptible to genetic variation altering DNAm level at the locus.

5.4 Discussion

5.4.1 Regional Heritability Across Colon and A Brain Tissue

Based on the RH analysis of each of the five tissues individually, (see 2.3.2 and 4.3.2) we found that more DNAm sites were heritable across colon and a brain tissue than expected if the heritability of DNAm level at a DNAm site is independent across tissues. For instance, we find that ~10% of DNAm sites have a significant RH in colon tissue and in each of the four brain tissues; therefore, by chance we might expect DNAm level at 1% of DNAm sites to be significantly heritable in both colon and any one of the brain tissues when in fact DNAm level at the DNAm site is not heritable in both of the two tissues. However, when we used Fisher's method and combined the significance of the RH estimates across colon and a brain tissue we found that between 11-12% of DNAm sites tested were heritable. This result is likely to, in part, reflect the procedure used to determine DNAm sites significantly affected by local genetic variation across colon and a brain tissue. Fisher's method will identify DNAm sites with significant RH in both tissues, some DNAm sites with a significant RH in one tissue and a small RH in the second tissue and possibly some DNAm sites with a small RH in both tissues. The small effect may not have been identified in RH analysis on the tissue alone. Therefore, by using Fisher's method we have increased the power to detect effects across tissues compared to if we selected DNAm sites with a significant RH in each of the two tissues independently.

Previously, we have shown that CpG dense regions of the genome contain less genetic variation than that observed for the remainder of the genome (see 4.3.1). Additionally, we have shown that in colon tissue DNAm sites located in CpG dense regions of the genome are less likely to be heritable (see 4.3.1). Given these two previous findings, it is not surprising that loci with significant $\hat{h}^2_{g,r}$ across colon and FCTX tissue were depleted for location within CpG dense regions of the genome compared to regions of low CpG density. However, given our previous finding in colon tissue that DNAm sites located within the TSS of a gene were less likely to be heritable than intragenic and intergenic DNAm sites it was surprising that DNAm sites with significant $\hat{h}^2_{g,r}$ across colon and FCTX tissue were enriched for location within the TSS of a gene. One possible explanation is that genetic effects on DNAm sites located within TSS are smaller than the genetic effects on DNAm sites located in intragenic and intergenic regions. In conjunction, the genetic effects on DNAm sites in intragenic and intergenic regions are more tissue specific than the genetic effects on DNAm sites located with TSS. In this case, when combining information across tissues the genetic effects on DNAm sites located in TSS become significant where as the tissue specific genetic effects on DNAm sites located in intragenic and intergenic regions do not become significant.

5.4.2 SNP Effects on DNAm Level Measured in Colon and FCTX Tissue

We have shown that that there is an inverse relationship between the distance of a SNP from a DNAm site and both the strength of association and the corresponding effect size. This relationship has been observed for DNAm level measured in brain tissue [2] and in studies of gene expression level. Our results indicated that the strength of association and effect size of a SNP declined rapidly within 500KB of the DNAm site. In 2.3.1 and 2.4 we have respectively shown and discussed that when estimating the RH including extraneous SNPs that do not tag the causal variation can lower the power to detect the causal variation. The results from our SNP association analyses suggest that in our colon and brain datasets to optimally capture local genetic effects using the RH approach a maximum window size of +/- 500KB surrounding the DNAm site would be appropriate. To extend this insight to additional datasets one would need to consider the extent of LD within the dataset and the number of SNPs per proposed region. On average there was 207 SNPs in the 2MB regions we used. Therefore, a maximum window size in a human dataset, or dataset with similar average LD, may have a considerably lower number of SNPs. Indeed, others' that have conducted RH analysis on the brain dataset have found that a window size of 50KB centred on the DNAm site provided the highest count of significant RH estimates [83]. In the aforementioned study [83] the 50KB region contained between 0 and 40 SNPs with a mode of 6. Thus, reducing the region size from 2MB and the number of SNPs would likely increase the power to detect heritable genomic regions. In turn, this may uncover additional genomic regions where the combined genetic effects are not currently significant enough to surpass our cut-off threshold.

We found a systematic difference between the SNP effects estimated from DNAm level measured in colon and FCTX tissue. This difference could reflect a true biological difference between the regulation of DNAm level in the two tissues. For instance, in our initial analysis we found that the average SNP effects were lower for

DNAm level measured in colon tissue than in FCTX tissue which could mean that overall the environment plays a decreased role in regulating DNAm level in the colon tissue compared to the FCTX tissue. However, a different phenotypic measurement scale in the two tissues could also lead to a systematic difference in SNP effects between the two tissues. In both datasets, DNAm level at a DNAm site was measured as the logit transformation of the proportion of methylated oligonucleotides (M-value scale) ([9]). Despite using the M-value measurement scale for DNAm level we observed a difference in the standard deviation of the residual phenotype between the two tissues. On average, the standard deviation of the residual phenotype was greater for DNAm level measured in the colon tissue compared to in the FCTX tissue. There are several reasons why this result could occur. Firstly, it is possible that DNAm level at a DNAm site is more variable in colon tissue than in FCTX tissue. Secondly, there may be a different extent of random noise in the measurement of DNAm level between the two datasets. Thirdly, DNAm level could be measured on a different scale in the two datasets. Different arrays were used to measure DNAm level in the two datasets, in the brain dataset the HM27K array was used and in the colon dataset the HM450K array was used. Similar to the HM27K the use of the HM450K involves bisulfite treatment and amplification of genomic DNA, leading to the conversion of an unmethylated cytosine to thymine. The proportion of targets containing a thymine to those containing a cytosine is quantified from the intensity measures of fluorescently labelled nucleotides added to the array. Adenine and thymine are labelled with a red fluorescent molecule and guanine and cytosine are labelled with a green fluorescent molecule [98]. However, the HM450K differs from the HM27K in that it employs two different probe designs to assess methylation level. Each DNAm site on the HM450K array is assayed by either the Infinium I design, the design used on the HM27K array or the Infinium II design. The majority (90%) of DNAm sites assayed on the HM27K array by the Infinium I design are assayed by the Infinium II design on the HM450K array [98]. The Infinium II design differs from the Infinium I design in the number of bead types and colour channels used to interrogate a DNAm site and the location where the fluorescent nucleotide anneals to the target DNA. The Infinium I design uses two probe types to assay DNAm level at a DNAm site, one probe to bind to the unmethylayed target and one probe to bind to the methylated target. The extent of target binding to the two probes is measured in the same colour channel, either red or green. The colour channel used depends on the nucleotide base in the probe that is next to the DNAm site being assayed. In contrast the Infinium II design uses one probe to assay both the methylated and unmethylated target by single base extension complementary to the interrogated cytosine [98]. This design results in the incorporation of an adenine complementary to a methylated cytosine and a guanine complementary to an unmethylated cytosine; therefore, resulting in the use of two colour channels to measure the proportion of methylation at a single DNAm site. An additional difference between the Infinium I and Infinium II design is that unlike the Infinium I design the Infinium II design can incorporate up to three degenerate R bases without compromising the ability of the target to anneal to the probe. Therefore, in contrast to the Infinium I, the Infinium II design does not assume that DNAm sites within the probe target are of the same phase (methylated or unmethylated) to the DNAm site being assayed. However, a limitation of the Infinium II probe design being able to incorporate degenerate R bases is that the probes are restricted to assaying DNAm level in lower CpG density regions of the genome than the Infiunium I design. For instance, there is an enrichment for Infinium II probes over Infinium I probes in CpG shores and regions distant to CpG islands, but not within CpG islands, reflecting this constraint of the Infinium II design [129]. Incorporation of two probes designs, (Infinium I and Infinium II) leads to additional quality control steps for the data from the HM450k array (colon dataset) compared to the data obtained from the HM27K array (brain dataset). For instance, the use of two colour channels to measure DNAm level at a single DNAm site with the Infinium II design means that it is necessary to account for colour bias (we used quantile normalization see 4.2.2) [130,131]. Additionally, the distribution of beta values produced by probes of the Infinium I and II design are different to one another [97,98,129,132-135]. This difference is thought to be in part due to technical factors and was corrected for when using data from the HM450K array (we used the BMIQ algorithm [97] see 4.2.2). The differences in how the arrays obtained the measurement of DNAm level and consequentially different quality control procedures for the two arrays could result in DNAm level being measured on a different scale across the two arrays. The correlation between DNAm level at DNAm sites assayed by the two arrays (25,978 Infinium I on HM27K and 25,978 Infinium I and Infinium II on HM450K) is high (r=0.98) [98]. However DNAm level assayed by these two arrays does not appear to be on average equal (the regression coefficient is not provided in the reference but observation of the graph indicates that for the most part the values do not lie on the 1:1 line) [98]. At most of the DNAm sites, DNAm level appears to be higher when measured on the 450K array compared to the 27K array [98]. In the preceding analysis [98] only one sample was assayed so the difference in the variance of DNAm level at a DNAm site across the two arrays was not reported. However, the difference in mean could reflect a difference in measurement scale across the two arrays. Our samples have been measured on either of the two arrays and we have no overlap of samples measured on both arrays. Therefore, it is challenging to determine if the difference in phenotypic standard deviation and average estimated SNP effect across the two datasets represents a true biological difference in the regulation of DNAm level or if it results from the use of a different scale of measurement across the two arrays. The risk that measurement of DNAm level was scaled differently in the two tissues led us to rank normal transform the residual phenotypes. We did this because scale of measurement could affect the estimate of the SNP effects and we wanted to meaningfully compare the individual SNP effects across the two tissues.

This transformation of the residual phenotypes reduced the systematic difference in average estimated SNP effects between colon and FCTX tissue. The small systematic difference that remained after rank normal transformation could be explained as a real biological difference between the two datasets or by a systematic difference in the magnitude of noise between the two arrays. Rank normal transformation first standardizes the residual DNAm level at a DNAm site (z-score) and then ranks the standardized residuals so that they are normally distributed and have a mean of zero and variance of 1. If the SNP effect on DNAm level is the same in two tissues but the mean level of noise is not, then the residual DNAm level and the SNP effect will be rescaled by a different factor. The result is that the effect in the two tissues will appear to be different after rank normal transformation. We minimized the systematic difference in average SNP effect across the two tissues that may have occurred due to a different level of noise in the two datasets by rescaling the SNP effects in the FCTX tissue. This allowed us to directly compare the magnitude of the SNP effects across the two tissues and identify tissue specific genetic effects.

In addition, for identifying SNP with a different effect across the two datasets we analysed SNP within genomic regions found to have a significant $\hat{h}^2_{g,r}$ across colon and FCTX tissue at our Bonferroni threshold adjusted for the number of traits analysed. Compared to the use of a nominal threshold, the use of a stringent threshold increased our confidence that the significant regions represented true positive associations of genetic variation with DNAm level. In turn, we used a conservative Bonferroni adjusted threshold to determine individual SNPs within the genomic regions that were associated with DNAm level. Overall, this meant that we could be confident that the within tissue associations of SNP with DNAm level were true positive associations and not false positive associations. Therefore, when we contrasted the estimated SNP effects across the two tissues we were likely not contrasting the effect of SNPs on DNAm level that resulted from chance (ie. false positive associations). When we compared the SNP effects across colon and FCTX we also used a conservative Bonferroni threshold to identify SNPs with a different effect in the two tissues. This meant that we could be confident that our result; 17 SNPs have a different genetic effect in the two datasets, reflected the true biological situation.

SNPs with a different genetic effect in different tissues can be thought of as exhibiting a genetic by environment interaction where the tissue is the environment. A genetic by environment interaction (tissue) for DNAm level does not necessarily mean that different biological processes are regulating DNAm level at the DNAm site. Assume the following biological model where a SNP affects transcription factor (TF) binding and gene expression level that in turn has an inverse affect on DNAm level. In this case, the SNP affects DNAm level in all tissues in which the gene is transcribed. However, the abundance of circulating TF in a tissue could also affect the extent to which the gene is transcribed and consequentially DNAm level. A genetic by environment interaction could occur if the SNP affects DNAm level in the tissue proportionally to the amount of TF abundance (and TF abundance is different in the two tissues). Similarly, a SNP that affects DNAm level in a tissue proportionally to the amount of DNMT or methyl donors present in the tissue would have a different effect size across the tissues if abundance of the DNMT or methyl donors were different. Even a different direction of a SNP effect across tissues does not necessitate that the associated DNAm site be regulated differently in the tissues. For instance, if a SNP leads to variation in DNAm level in two tissues and in one tissue a mechanism takes effect so as to stabilize DNAm level, whereas in a second tissue this mechanism does not take effect, then a difference in the direction of SNP effect could be seen across the two tissues. The aforementioned examples could be viewed as genetic by environment interactions driven by tissue specific stochastic variation where the core underlying biological process regulating DNAm level remains similar across tissues. However, it is also possible that SNP effects on DNAm level differ across tissues due to different processes regulating DNAm level in the tissues. For instance, in different tissues different proteins may bind to the DNA to regulate DNAm level. Different proteins may be affected by a SNP to a varying degree and this would lead to an overall different estimated SNP effect on DNAm level across the tissues.

The 17 SNPs that we found to have a different effect in colon and brain tissue represented 8 loci. We queried online data repositories to establish a biological explanation for the tissue specific genetic effects on DNAm level; however, this proved to be challenging given the sparse knowledge of the function of genes in different tissues. However, we provided general plausible explanations for several of our results. We expect increased research assessing the relationship between genetic variations, cellular phenotypes and diseases/traits to enhance the functional annotation of the genome. This is turn will help biological interpretation of future studies similar to our own which has investigated tissue specific genetic effects on DNAm level.

We have identified SNP that have a different effect in colon tissue collected from Colombian individuals and FCTX tissue obtained from individuals genetically similar to the HapMap CEU population (as determined by cluster analysis, see 2.2.1.1.3). Therefore, in our analyses tissue is confounded with population. This means that we cannot be sure that a genetic by environment interaction results from a difference in SNP effect across tissues (colon and FCTX) and it may result from a difference across populations (south and north American). Nevertheless, our procedure has revealed that genetic by environment interactions for DNAm level do exist and we expect larger scale studies of this nature will follow.

5.5 Contributions

The colon dataset was collected and prepared by Gustavo Hernandez-Sanchez, Maria Carolina Sanabria and Martha Serrano-Lopez, Jose Luis Soto, Adela Castillejo, Cristina Alenda and Eva Hernandez-Illan

Gibbs et al. [2], the Division of Aging Biology, the Division of Geriatrics and Clinical Gerontology (NIA) were responsible for collection and initial research on the brain dataset. The brain dataset was downloaded from the NCBI Data Repositories dbGaP (Accession Number: phs000249.v1.p1) and GEO (GSE15745).

Both my supervisors provided comments on drafts of this chapter.

Chapter 6 Final Discussion and Conclusion

6.1 Findings

Overall, our RH study suggests that variation in DNAm level at some DNAm sites is at least partially controlled by local nuclear genetic variation (Table 28). For instance, even at a conservative Bonferroni threshold ($P < 2.4*10^{-7}$) that adjusted for the total number of *cis* association tests conducted across all five tissues, DNAm level was heritable at a number of DNAm sites within each tissue (Table 28).

Table 28 Number of Associations between DNAm Level and the Local Genomic Region at Different Significance Thresholds.

Each column heading indicates the Bonferroni adjustment conducted and the significance threshold. The number within each cell is the count of DNAm sites significant at the specified threshold.

Tissue	Nominal $P < 0.05$	Traits within Tissue CRBL $P < 1.64*10^{-05}$ FCTX $P < 1.63*10^{-05}$ PONS $P < 1.63*10^{-05}$ TCTX $P < 1.63*10^{-05}$ Colon $P < 2.5*10^{-07}$	Traits within Tissue Set Brain $P < 4.1*10^{-06}$ Colon $P < 2.5*10^{-07}$	All Tests Conducted $P < 2.4*10^{-07}$
CRBL	294	22	15	8
FCTX	297	33	23	16
PONS	277	23	19	12
TCTX	292	32	28	22
Colon	21471	920	920	917

In conjunction, we used the RH framework and found that DNAm level could be predicted from the local sequence variants with an accuracy that scaled with the estimated $\hat{h}^2_{r,g}$. Furthermore, the RH approach was able to capture additional effects of *cis* acting genetic variation beyond that detected by the CRC risk SNP, rs4925386. In contrast to the *cis* effects detected, using our study design we found no evidence that suggested that DNAm level was associated with regional genetic variants in *trans*.

We examined the local $\hat{h}^2_{r,g}$ of DNAm sites located within and outwith a disease susceptibility region and measured in a relevant tissue and found that the two sets exhibited a similar overall pattern of estimated $\hat{h}^2_{r,g}$.

Additionally the propensity for DNAm level to be associated with the local sequence variation differed with respect to CpG dinucleotide density and genic location. Most notably, DNAm sites located in CpG dense regions of the genome were less likely to be heritable than DNAm sites located in CpG sparse regions of the genome. Additionally, within both CpG dense and CpG sparse regions of the genome intergenic DNAm sites were more likely to be heritable than intragenic DNAm sites.

Finally, we found evidence that genetic effects on DNAm level differed between DNAm level measured in colon tissue from South American individuals and DNAm level measured in FCTX tissue from North American individuals.

6.2 Implications

6.2.1 Regional Heritability is a Useful Approach for Estimating Genetic Effects on Cellular Phenotypes such as DNAm Level.

The successful application of the RH method to DNAm level using a small sample of nominally unrelated individuals has implications for estimating the genetic effects on other cellular phenotypes. There are several characteristics of the phenotypic and genotypic data and RH approach which provide power for $\hat{h}^2_{r,g}$. Firstly, while unrelated individuals share a small proportion of their whole genome in common, regionally, the proportion shared between two individuals can vary substantially in the population (and from genomic region to genomic region). Secondly, simultaneously testing all SNPs within a genomic region compared to testing SNPs individually for association with a trait allows for small effects to be combined to a measureable estimate. Moreover, the use of regions rather than single SNPs can reduce the number of statistical tests conducted and allow for a less stringent significance threshold. Thirdly, genetic effects on DNAm level are relatively large in comparison to those observed for ultimate phenotypes (UPs) and they are enriched locally to the DNAm site. Given this last point, we expect the RH method will be of utility in estimating the genomic contribution to cellular phenotypes (CPs) with a similar overall genetic architecture to DNAm level, such as level of gene expression.

Moreover, the observation that risk SNP contributed to a proportion of the local $\hat{h}^2_{r,g}$ for DNAm level in healthy tissue has implications for predictive genetic models. This result highlighted that in comparison to use of single SNP, the RH approach can pick up additional genetic effects in a relevant healthy tissue which may mediate susceptibility to an UP. Therefore, genetic models that aim to predict an UP or DNAm level could benefit from using regional genetic variation rather than single SNP.

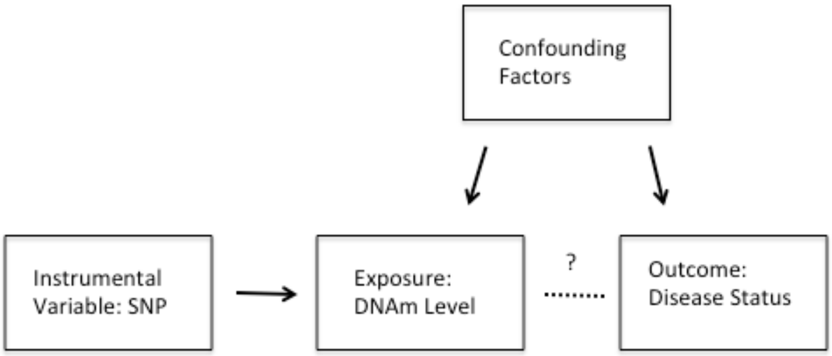
6.2.2 Genetic Variation and DNAm Level at an Associated DNAm site can be used to Determine if DNAm level Causes Disease

One implication of DNAm level being heritable is that genetic variation associated with DNAm level can be used to determine if DNAm level causes disease using the Mendelian randomization framework (Figure 34). Correlation between DNAm level and a disease, such as that which may be obtained from EWAS, can arise from several scenarios. For instance, changes in DNAm level may cause changes in disease status or vice-versa. Additionally, the change in DNAm level and disease status could independently be caused by an additional factor (confounding). An assumption of Mendelian randomization is that the SNP is not correlated with unmeasured factors associated with an exposure (DNAm level) and an outcome (disease). Therefore, if the SNP is associated with the exposure (DNAm level) it can be used as a proxy for the exposure (DNAm level), which negates the effects of the confounding variable(s). A second assumption of Mendelian randomization is that the SNP is independent of the outcome (disease) given the exposure (DNAm level) [136]. Given these assumptions, then an association between the SNP and the outcome (disease) indicates that the exposure (DNAm level) causes disease. Our research has indicated that *cis* rather than *trans* acting genetic effects on DNAm level predominate. The substantial *cis* acting associations uncovered reveal that *cis* acting genetic variation is a promising pool of genetic variation to use in Mendelian randomization experiments.

Mendelian randomization studies are of particular utility for identifying targets for therapy. Unlike studies that uncover association between DNAm level and disease but provide no insight into the direction of causality, MR can identify potential exposures that cause disease. Understanding the cause and effect relationship between DNAm level and disease is of interest to the medical community because DNAm level is a modifiable exposure that can be targeted for therapy.

Figure 34 Mendelian Randomization Framework

A SNP associated with DNAm level can be used to determine the causal relationship between DNAm level and disease status. It is assumed that the SNP is not associated with the confounding factors, that is it associated with DNAm level and that the only association with disease is through DNAm level. In this case, association of the SNP with disease status suggests that change in disease status is caused by change in DNAm level.



6.2.3 Methods for Increasing the Power to Detect Genetic Effects on Ultimate Phenotypes

Our research has implications when considering QTL to prioritize for association with an UP. The location of a DNAm site with respect to susceptibility regions was independent of the likelihood of DNAm level being heritable and the heritability estimate (see 4.3.5). Therefore, using QTL associated with DNAm level (methQTL) to prioritize genomic regions for association with an UP would not provide enrichment for significant associations beyond that which would be obtained from a random set of genomic regions. Prioritization by this method would be an inefficient strategy for reducing the significance threshold as a means for increasing the power to detect the effects of genetic variation in an association study. Interestingly, this finding contrasts that found in studies of gene expression. GWAS results at P $<10^{-5}$ were enriched for eQTLs in blood [137] and schizophrenia susceptibility SNPs were enriched for eQTLs in human brain [138]. Additional studies [52,139] have indicated that susceptibility SNPs are enriched for eQTLs and they have used eQTL results to prioritize sub significant QTL from GWAS for replication. One explanation for the difference in enrichment of eQTLs compared to methQTLs in susceptibility QTL could be related to the functional consequence of QTL. For instance, QTL may more often influence an UP by acting directly on level of gene expression rather than on level of DNAm. Further studies could provide evidence for this hypothesis and validate our finding that QTLs (for UPs) are not enriched for methQTLs. Our study has focused on using regional genetic variation local to DNAm sites within and outwith susceptibility loci, where a susceptibility locus is defined as +/- 1MB of susceptibility SNP. The aforementioned eQTL research has been conducted using a single SNP approach. Enrichment analysis using the RH method with gene expression level and the single SNP approach with methylation level could help resolve if the difference is an artefact of the use of different experimental design.

6.2.4 DNAm Level does not Always Need to be Assayed to be used in an Analysis

An implication of accurate prediction of DNAm level in healthy tissue is that it is not always necessary for DNAm level to be directly assayed. DNAm level could be accurately predicted in the currently abundant large GWAS datasets. Predicted DNAm level could then be tested for association with an UP. This would provide a set of DNAm sites putatively associated with the UP in healthy tissue for replication at a reduced significance threshold. Moreover, because the predicted DNAm level is based on genetic effects estimated in healthy tissue relevant to disease an association between predicted DNAm level and a disease will reflect variation in DNAm level which exists prior to disease. Hence, unlike EWAS, this type of study can provide highly suggestive evidence for variation in DNAm level causing disease. However, in order to predict DNAm level in a disparate dataset the regional SNP effects must first be estimated. We have shown that a small number of samples can be used to obtain accurate estimates for SNPs effects within a 2MB region and local to a DNAm site. While local SNP effects on DNAm level are thought to be tissue specific, provided sufficient power, the effects need only be estimated once within each tissue to facilitate prediction of DNAm level into a large number of datasets. Therefore, it would be of utility to the research community to store regional SNP effects on DNAm level in the public domain. The data repository, GRASP [140], currently contains results from association of SNP with CPs. However, this database is not optimally designed to hold regional SNP effects and overall it contains limited information regarding methQTL. For instance only one single SNP methQTL study conducted in one tissue is represented in GRASP [140] and the direction of the effect of SNP on DNAm level is not noted. A database containing the location and size of the region analysed, $\hat{h}^2_{r,g}$, tissue, population and effect of SNPs within the genomic region would be of benefit to the research community.

6.2.5 DNAm Sites in Different Genomic Contexts may have Different Functional Roles

Finding that the local RH of DNAm Level is related to the location of the DNAm site with respect to defined genomic contexts has implications for understanding the function of the human genome. Functional elements of the genome can be determined by several different scientific approaches and each approach will necessitate a slightly different interpretation of what it means for an element to be 'functional' (reviewed in [141]). Firstly, evolutionary methods that measure the extent of neutral or positive selection can define functional elements in the genome. Assays that measure biochemical activity at regions of the genome provide a second measure of functionality. Thirdly, genetic variation can be defined as functional if it is associated with a phenotype (reviewed in [141]). Our study can primarily inform on functional genetic variation based on this third definition. For instance, our RH analyses determined that genetic variation can have a functional effect on DNAm level for DNAm sites located in the eight genomic contexts we studied (see Chapter 4 in particular 4.3.4). However, we also showed that the probability of genetic variation having a functional effect on DNAm level is related to the genomic context of the DNAm site (see 4.3.4). The propensity of a genomic context to be functional may be related to the biological role of DNAm level at DNAm sites located within the genomic context. In this view, DNAm level is related to fitness (by affecting an UP) and to a different extent within different genomic contexts. In conjunction, the lower probability of DNAm level being heritable reflects lower genetic variation due to negative selection acting to purge SNP with effects that decrease fitness.

6.3 Further Research

Broadly, further research will focus on using healthy tissue to understand the genetic and environmental component of DNAm level, the molecular role of DNAm level and how these factors and variation in DNAm level relate to disease and trait aetiology.

6.3.1 Estimating Genetic Effects on DNAm Level

To date, studies exploring the genetic architecture of DNAm level have concentrated on the analysis of *cis* acting genetic variation. This is primarily due to the determination that a substantial proportion of the phenotypic variation in DNAm level can be explained by local genetic variation and that these effects are relatively easy to quantify. In contrast, less attention has been paid to understanding the contribution of *trans* acting genetic variation on DNAm level. This is likely a consequence of needing to adopt a more stringent significance threshold to test genetic variation genome-wide compared to when testing local genetic variation. With a given number of samples, adopting a more stringent significance threshold leads to decreased power to detect genetic effects and an increased minimum detectable effect size. Therefore, if the distributions of effect sizes are similar for *trans* and *cis* acting genetic variation, then the proportion of *trans* acting effects detected will be smaller than the proportion of *cis* effects detected. Indeed, this is what we have observed in our studies (see 2.3.2, Chapter 3 and 4.3.2) and what others' [2,54,55,61] who have assessed the effects of both *cis* and *trans* acting variation on DNAm level have shown.

Currently, research assaying DNAm level has typically been limited to using a small set of samples. This is because the ability to quantify DNAm level at individual sites genome-wide is relatively new and samples from many tissues are challenging to obtain and can take years to accumulate. As studies get larger in sample size, providing the multiple testing threshold does not dramatically change, we expect there to be an increase in the number of *trans* acting variants discovered both with the single SNP and RH approach. Moreover, within the RH framework a larger sample size would facilitate the fitting of a second variance component to estimate the total effect of *trans* acting variation on DNAm level.

However, there is scope for studies of small sample size to investigate *trans* acting effects on DNAm level; although based on our work and the publish literature we suggest developing a novel approach. For instance, we would consider using the RH approach with different window sizes. Little information regarding the genetic architecture of *trans* acting variation makes it challenging to determine the optimal window size for analysis. For instance and as previously discussed, unlike a single SNP approach the RH approach can capture the combined effect of multiple small effect variants acting within a region. However, inclusion of extraneous genetic variants can induce noise and reduce the power of the RH approach. In conjunction, in the case of one or few causal genetic variants acting on a phenotype, the power of the RH approach may be decreased in comparison to a single SNP approach. Therefore, based on the underlying genetic architecture of the causal variation the size of the genomic region tested will affect the power to detect an association. An additional approach that could be undertaken either within the RH framework or the single SNP framework is targeted analysis. Targeted analysis would involve testing the effect of a subset of genetic variation that is a priori thought to affect DNAm level in *trans*. If one hypothesizes that 1) genetic variation that affects gene expression will also affect DNAm level at the gene and 2) genes that are co-expressed will be to some extent affected by the same genetic variation, then gene networks could be used to prioritize genetic variation. For instance, genes involved with a process related to the tissue in which DNAm level is measured could be used to select genomic regions of interest. Within these regions, genetic variation could be tested for association with DNAm level. Finally, if one assumes only the first of the two hypotheses above then genetic variation found to have a *trans* effect on gene expression level could be prioritized for association with DNAm level measured in the same tissue. This last point necessitates that *trans* effects on gene expression level have been tested and reported in the tissue of interest. In the case of our brain dataset this has occurred; therefore, it would be relatively straightforward for us to conduct the outlined experiment in human brain tissue.

Our finding that in most cases DNAm level is not solely determined by genetic variation raises the question of what environmental factors influence variation in DNAm level. Often datasets are limited in the number of environmental factors that are measured, measuring only those with a probable effect or those that are simple to measure. This makes examination of putative environmental effects in the primary study challenging. One environmental factor not analysed in our study of the colon dataset presented in Chapter 4 is the location of biopsy with respect to what is clinically defined as proximal and distal colon. There are several differences between the proximal and distal colon (reviewed in [142]); therefore, they could be thought of as different cellular environments. We hypothesize that these two environments lead to variation in DNAm level. Following the investigation of the RH of DNAm level in colon tissue presented in Chapter 4, proximal and distal colon has been classified and is being analysed for association with DNAm level by Konrad Rawlik.

In addition, further studies that investigate if genetic effects on DNAm level differ across tissues will aid in our understanding of the regulation of DNAm level. Based on our work (see Chapter 5) and that of others' [2] we would suggest focusing on investigating the effects of SNPs within +/- 500KB of a DNAm site on the DNAm level. To conduct the meta-analysis, one could use summary statistics from within tissue association analyses or one could follow a new method for meta-analysing correlated traits [143]. This new method [143] could be used to meta-analyse DNAm level measured in the brain dataset and our colon dataset, using the measurements from the four brain regions simultaneously.

6.3.2 Understanding the Molecular Role of DNAm Level

We have shown that the genetic variation local to a DNAm site could explain a proportion of the phenotypic variation in DNAm level. This result leads naturally to the question of the molecular role of DNAm level in cellular processes and how this role is perturbed by genetic variation. For instance, does genetic variation alter DNAm level directly or does it first alter other CPs such as gene expression that consequentially leads to a change in DNAm level? Ideally, to investigate these questions the more types of CPs measured (gene expression, different chromatin marks) the better and preferably in the same individuals. Consider the underlying assumption that DNAm level will phenotypically co-vary with a CP that it works in concert with to regulate cellular processes. In this case, genetic variation that affects DNAm level will also affect the CP with which DNAm level co-varies. Therefore, examining genetic effects on DNAm level and co-varying CPs could provide an indication of the molecular role of DNAm level. An example of a particular study design that could be used is a bivariate or multivariate analysis with the RH approach. If the dataset is not of adequate power to facilitate the use of a multivariate analysis then one could conduct separate genetic association analysis for each CP and examine the correlation of the resulting P-values. While these preceding two approaches may provide an indication of CPs that interact, they do not inform on the cause and effect relationship between the CPs. However, A two-step MR process

[144] could help resolve causal pathways. These lines of research could be explored now in the brain dataset using gene expression level and DNAm level.

An important consideration is a ‘dynamic’ role of DNAm level where DNAm level functions differently with respect to other CPs under certain circumstances, such as location of the DNAm site in the genome. Indeed, we showed a different propensity for DNAm level to be heritable across different genomic contexts. It would be interesting to determine if the probability of other CPs being heritable, such as gene expression level, varies accordingly. A difference in the extent of genetic co-variance in different contextual groups could be indicative of a different molecular role for DNAm level across those groups. However, the extent of genetic variation could affect the power to detect genetic effects and also the estimate of genetic co-variation. Overlaying information of evolutionary conservation and positive selection would help determine if estimates of the genetic effects were influenced by the extent of genetic variation. Moreover, understanding the evolution of the DNA sequence with respect to the heritability of DNAm level (and the genetic co-variance of DNAm level and CPs) could provide valuable insight into the functional role of DNAm level.

6.3.3 Determining if DNAm Level is Causal for Disease

Our study (see 4.3.2) has provided a list of DNAm sites where DNAm level measured in healthy colon tissue is associated with local genetic variation. There are several different methods that could be used to try and discern which, if any, of these heritable DNAm sites are causal for disease of the colon such as CRC. Firstly, as described in 2.2.6.2 the *cis* acting SNP effects on DNAm level could be estimated. Subsequently, the estimated SNP effects could be used in a disparate dataset containing genotypic and case control status for CRC to predict DNAm level. Predicted DNAm level could be tested for association with disease outcome to provide a set of candidate DNAm sites causal for CRC. Secondly, a GWAS could be used to identify the SNP within the local region most strongly associated with DNAm level. The SNP could then be used in a one step mendelian randomization process in a dataset containing genotypic information, measurements of DNAm level and CRC status. Thirdly, longitudinal data could provide an indication of changes in DNAm level occurring prior to disease onset. This last study is clearly challenging to undertake. It would take years to acquire the required data or it would necessitate that the relevant tissue have been measured and the disease status recorded in a pre-existing longitudinal study. However, the preceding first two lines of experimentation are quite easily facilitated and could be conducted using our colon dataset. This is because in the colon dataset we have used only a subset of the samples for which we have extracted tissue.

6.4 Final Words

Understanding the genetic basis of disease is fundamental to advancement in personalized medicine. Currently, many human diseases are proving challenging to predict from the genotype alone. Understanding how phenotypic variation in DNAm level relates to genetic variation and disease susceptibility is a promising avenue of research to progress personalized medicine. For instance, DNAm level correlated with disease and under genetic control could be used to suggest individuals to target for prospective treatment. In addition, DNAm level which causes disease can itself be used as a target for therapy. With the extensive datasets and computational resources now available, this era is likely to bring many pivotal discoveries in human genetics.

6.5 Contributions

Albert Tenesa provided comments on drafts of this Chapter

Chapter 7 Appendix: Copy of Publication

OPEN ACCESS Freely available online

PLOS | ONE

Complex Variation in Measures of General Intelligence and Cognitive Change

Suzanne J. Rowe[1], Amy Rowlatt[1], Gail Davies[2], Sarah E. Harris[2,3], David J. Porteous[2,3], David C. Liewald[2], Geraldine McNeill[4], John M. Starr[2,5], Ian J. Deary[2,6]☯, Albert Tenesa[1,7]*☯

1 The Roslin Institute, The University of Edinburgh, Roslin, Scotland, United Kingdom, **2** Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, Scotland, United Kingdom, **3** Medical Genetics Section, The University of Edinburgh, Edinburgh, Scotland, United Kingdom, **4** Institute of Applied Health Sciences, University of Aberdeen, Aberdeen, Scotland, United Kingdom, **5** Alzheimer Scotland Dementia Research Centre, The University of Edinburgh, Edinburgh, Scotland, United Kingdom, **6** Department of Psychology, University of Edinburgh, Edinburgh, Scotland, United Kingdom, **7** Medical Research Council Human Genetics Unit at the Medical Research Council Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, Scotland, United Kingdom

Abstract

Combining information from multiple SNPs may capture a greater amount of genetic variation than from the sum of individual SNP effects and help identifying missing heritability. Regions may capture variation from multiple common variants of small effect, multiple rare variants or a combination of both. We describe regional heritability mapping of human cognition. Measures of crystallised (g_C) and fluid intelligence (g_f) in late adulthood (64–79 years) were available for 1806 individuals genotyped for 549,692 autosomal single nucleotide polymorphisms (SNPs). The same individuals were tested at age 11, enabling us the rare opportunity to measure cognitive change across most of their lifespan. 547,750 SNPs ranked by position are divided into 10, 908 overlapping regions of 101 SNPs to estimate the genetic variance each region explains, an approach that resembles classical linkage methods. We also estimate the genetic variation explained by individual autosomes and by SNPs within genes. Empirical significance thresholds are estimated separately for each trait from whole genome scans of 500 permutated data sets. The 5% significance threshold for the likelihood ratio test of a single region ranged from 17–17.5 for the three traits. This is the equivalent to nominal significance under the expectation of a chi-squared distribution (between 1df and 0) of $P < 1.44 \times 10^{-5}$. These thresholds indicate that the distribution of the likelihood ratio test from this type of variance component analysis should be estimated empirically. Furthermore, we show that estimates of variation explained by these regions can be grossly overestimated. After applying permutation thresholds, a region for g_f on chromosome 5 spanning the *PRRC1* gene is significant at a genome-wide 10% empirical threshold. Analysis of gene methylation on the temporal cortex provides support for the association of PRRC1 and fluid intelligence (P = 0.004), and provides a prime candidate gene for high throughput sequencing of these uniquely informative cohorts.

Citation: Rowe SJ, Rowlatt A, Davies G, Harris SE, Porteous DJ, et al. (2013) Complex Variation in Measures of General Intelligence and Cognitive Change. PLoS ONE 8(12): e81189. doi:10.1371/journal.pone.0081189

Editor: Huiping Zhang, Yale University, United States of America

Received February 13, 2013; **Accepted** October 20, 2013; **Published** December 12, 2013

Funding: Genotyping of the ABC1936, LBC1921, and LBC1936 cohorts and the analyses conducted here were supported by the UK's Biotechnology and Biological Sciences Research Council (BBSRC). Phenotype collection in the Lothian Birth Cohort 1921 was supported by the BBSRC, The Royal Society, and The Chief Scientist Office of the Scottish Government. Phenotype collection in the Lothian Birth Cohort 1936 was supported by Research Into Ageing (continues as part of Age UK's The Disconnected Mind project). Phenotype collection in the Aberdeen Birth Cohort 1936 was supported by the BBSRC, the Wellcome Trust, and the Alzheimer's Research Trust. SJR, AR and AT are funded by the BBSRC through the Roslin Institute's strategic programme grant and project grant BB/K000195/1. The Brain data was provided by the Division of Aging Biology and the Division of Geriatrics and Clinical Gerontology (NIA) through the NIH GWAS Data Repository (dbGaP Accession Number: phs000249.v1.p1) and funded as part of the Intramural Research Program, NIA. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: Albert.Tenesa@ed.ac.uk

☯ These authors contributed equally to this work.

Introduction

Loss of cognitive function is one of the most feared aspects of growing old. Intelligence and the rate of age related cognitive change vary widely in healthy individuals and have been associated with health status, longevity and quality of life [1,2,3,4,5,6]. As the general population ages, cognitive health is of paramount importance, and understanding the underlying mechanisms of general intelligence and age-related decline has wide-ranging social and economic implications. Although pathological cognitive decline has been studied in diseases such as Alzheimer's [7], available phenotypic measures for lifetime changes in cognitive abilities of healthy individuals are rare. An important part of the variation in human general intelligence and in non-pathological, age-associated cognitive decline [8,9] can be attributed to heritable genetic variation. Identifying the genes and loci that contribute to the estimated genetic variance would offer new biological insight, with opportunities to develop tailored interventions and to inform policy makers.

Here we analyse the genetic contributions to complex variation in three measures of intelligence: (i) crystallised intelligence; (ii) fluid general intelligence; and (iii) lifetime change in intelligence. We use three Scottish birth cohorts whose intelligence was measured in childhood (age 11 years) and again in late adulthood (age 65 to 79 years). Crystallised intelligence (g_c) is typically assessed using vocabulary and knowledge-based tests, and tends to remain stable with age. Fluid intelligence (g_f) is assessed using tests that require on-the-spot thinking — often with abstract materials and under time pressure — and tends to peak in early adulthood and decline thereafter [10,11]. Here, cognitive change was measured as fluid intelligence in old age adjusted for intelligence measured at age 11 as described in Deary et al. [9], who showed, using the same data, that the lower bound estimate for the proportion of variation in lifetime change of intelligence explained by genetic factors was 0.24.

To date, as is seen in many complex traits, and despite moderate-to-high heritability estimates, genomic studies have yielded little knowledge of the underlying genetic factors affecting cognitive traits. Although studies for other complex traits have been successful at garnering the few common genetic variants that explain a sizeable amount of variation, genome wide association studies (GWAS) have generally failed to capture a large proportion of the genetic variation in complex traits [12,13,14]. A recent GWAS for crystallised and fluid intelligence did not result in any replicable genome-wide significant association despite moderately high heritability estimates of 0.4 (s.e. 0.11) and 0.51(s.e. 0.11) for g_c and g_f respectively for the population under study [8]. To address this gap, we have applied a recently proposed analytical approach [15] that captures the combined effect of multiple genetic variants at a region of the genome, thereby identifying some of the heritability missing when applying standard 'one at a time' SNP analyses [16,17]. This approach has the potential to overcome stringent multiple testing penalties and has been shown to be more powerful than the 'one at a time' SNP approach in simulated and real data [15]. We hypothesise that combinations of common and rare variants, that are not in complete LD with common tagging SNPs, may account for a substantial part of the missing heritability and that these will be best captured by estimating the genetic variation from an entire 'region' or geographically co-located set of SNPs. The trade-off comes between capturing as much variation as possible, whilst having the resolution to locate causal effects. Here we divide the genome in two ways (regionally and functionally): firstly, into overlapping regions of 101 SNPs; and secondly by chromosome, separating SNPs that lie within genes and SNPs that map outside a 5 kb boundary of genes. We examine the genetic variation explained by each region or chromosome for crystallised and fluid intelligence and for the lifetime change in fluid intelligence, and we compare that to the most significant results obtained from the 'one SNP at a time' association approach.

Materials and Methods

Phenotypic Data

Ethical approval for all the projects was obtained from the Lothian Research Ethics Committee. Data were gathered from three longitudinal studies of relatively healthy older individuals with detailed cognitive phenotypes: the Lothian Birth Cohorts of 1921 (LBC1921, N = 550) and 1936 (LBC1936, N = 1091), and the Aberdeen Birth Cohort of 1936 (ABC1936, N = 548). The years 1921 and 1936 refer to the participant's year of birth. Participants took a validated intelligence test at a mean age of 11 years: the Moray House Test No. 12 (MHT), which is a test of general intelligence [18,19] and detailed follow-up assessments at a mean age (sd) of 79.1 (0.6), 69.5 (0.8) and 64.6 (0.9) for LBC1921, LBC1936 and ABC1936, respectively. Cognitive test scores from age 11 and old age were available.

Construction of phenotypes

Selection of individuals, ethical consent, and full details of the assessments have been described in previous studies [8,9,18,19,20,21]. In brief, for each cohort, cognitive phenotypes of fluid-type and crystallized-type intelligence were constructed [19,20]. The final measure of lifetime cognitive change was constructed by adjusting fluid intelligence in old age for prior cognitive ability providing a quantitative measure of cognitive change from age 11 to old age. Phenotypes were adjusted within cohort for age and standardised within gender, and are further defined in Appendix 1.

Genotypic data

Following informed consent, venesected whole blood was collected for DNA extraction. A total of 599,011 single nucleotide polymorphisms (SNPs) were genotyped using the Illumina610-Quadv1 chip as described previously [8]. Quality control (QC) procedures were performed per SNP and per sample. Individuals were excluded from further analysis if genetic and reported gender did not agree. Samples with a call rate ≤0.95, and those showing evidence of non-European descent by multidimensional scaling analysis, were also removed. SNPs were included in the analyses if they met the following conditions: call rate ≥0.98, minor allele frequency ≥0.01, and Hardy-Weinberg equilibrium test with p≥0.001. To avoid bias from hidden family structure, if a pair of individuals shared more than 2.5% of the genome in common, one individual was omitted from the analysis. After QC, 1804 individuals (ABC1936, N = 376; LBC1921, N = 484; LBC1936, N = 944), and 547,750 autosomal SNPs were included in the analysis.

Estimation of regional and functional genetic contribution

In a population of unrelated individuals, SNP genotypes can be used to estimate shared co-ancestry or identity by state between individuals with rare SNPs weighted more heavily. Under certain assumptions it can be shown that a region that is shown to be identical by state will also be identical by descent [22]. The $n \times n$ genomic relationship matrix (GRM) of relatedness at a population level between n individuals gives the covariance structure for the phenotype based on the premise that the more related two individuals are, or the greater the amount of the genome they share in common, the greater the expectation of phenotypic similarity.

Using theory adapted from standard variance components or pedigree based linkage analysis [23,24,25] and further developed for genomic prediction [26,27,28], a GRM containing information from the genotypes of m SNPs can be used to solve a linear mixed model [Model 1] and partition the phenotypic variance into estimates of the genetic and environmental variance [15,29]. To avoid confusion with the well-known family-based estimates of heritability [30] we define the amount of phenotypic variance captured by the genotypes of unrelated individuals as population-sense heritability (h^2_{ps}). The linear mixed model (LMM) is:

$$Y = X\beta + Iu + e \qquad \text{(Model 1)}$$

Where $\mathbf{Y}$ is an $n \times 1$ vector of phenotypes for n individuals; $\mathbf{X}_{n \times 21}$ is the incidence matrix relating the regression coefficients for 20 principal components and gender to the n individuals; $\boldsymbol{\beta}$ is a 21×1 vector of fixed effects; $\mathbf{u}$ is a $n \times 1$ vector of the additive genomic random effects where $u \sim N(0, G\sigma^2_u)$, $\mathbf{G}$ is an $n \times n$ genomic relationship matrix estimated from the SNP genotypes and $\boldsymbol{\sigma}^2_{\mathbf{u}}$

is the genetic variance captured by the SNPs used to estimate the relationships among the n individuals; **I** is an $n \times n$ identity matrix; and **e** is an $n \times 1$ vector of individual residual effects. The variance of **Y** is $var\ (Y) = G\sigma^{2}_{u} + I\sigma^{2}_{e}$. **G** is calculated following Van Raden (2008) [28]. In short, an $n \times m$ matrix, **W**, is constructed where m is the number of SNPs available. The elements of **W**, w_{ij}, are defined as $w_{ij} = (x_{ij}\text{-}2p_j)/\sqrt{2p_j(1\text{-}p_j)}$ with x_{ij} being 0, 1 or 2 for the three possible SNP genotypes for the j-th SNP of the i-th individual and p_j being the allele frequency of the j-th SNP. **G** is calculated as WW'/m.

An extension of this to a bivariate analysis [Model 2] was used to estimate phenotypic and genetic covariances amongst measures of intelligence.

$$Y_1 = X_1\beta_1 + Iu_1 + e_1$$

$$Y_2 = X_2\beta_2 + Iu_2 + e_2 \qquad \text{(Model 2)}$$

Where 1 and 2 refer to trait 1 and trait 2, $\mathbf{u}_1$ and $\mathbf{u}_2$ are $n \times 1$ vectors of additive genomic random effects. **G** is the genomic relationship matrix between all individuals as described above. The additive genetic covariance of $\mathbf{Y_1}$ and $\mathbf{Y_2}$ - cov($\mathbf{u}_1$, $\mathbf{u}_2$) = σ^{2}_{u12} and the environmental covariance cov ($\mathbf{e_1}$, $\mathbf{e_2}$) is σ^{2}_{e12}. The additive genetic correlation of Y_1 and Y_2 is $\sigma^{2}_{u12}/\sigma_{u1}\ \sigma_{u2}$, and the variance-covariance matrix for $\mathbf{Y = [Y_1, Y_2]}$ is $V = \begin{pmatrix} G\sigma^2_{u1} + I\sigma^2_{e1} & G\sigma^2_{u12} \\ G\sigma^2_{u12} & G\sigma^2_{u2} + I\sigma^2_{e2} \end{pmatrix}$. A full derivation of the estimation of the genetic covariance is given in [31].

Regional population-sense heritability

Yang et al. [32] implement the linear mixed model [Model 1] in the software package GCTA and have shown that the method can be used to partition the genetic variation across chromosomes and functional regions of the genome such as genes [15].

By combining information on multiple SNPs within a genomic region we aim to capture a substantial part of the heritability missed by traditional 'one SNP at a time' approaches. Identifying those regions of the genome that capture most variation is an efficient way of selecting candidate regions for high throughput sequencing that could complement whole-exome sequencing experiments until whole genome sequencing is feasible for large numbers of samples. Here, autosomal SNPs were ranked by genomic location and divided into regions spanning 101 consecutive SNPs. Regions were overlapping to allow for the possibility that genetic variation is distributed among two or more windows, with a shared region between two consecutive regions spanning 50 SNPs, resulting in 10,908 overlapping regions from 547,750 SNPs. Each region was fitted individually in the linear mixed model [Model 3].

$$Y = X\beta + Iu_R + e \qquad \text{(Model 3)}$$

Where R is the genomic region. $\mathbf{u}_R$ is a vector of n additive genomic random effects from the region, n is the number of individuals and **I** is the identity matrix as described above. $Var(Y) = G_R\sigma^2_{uR} + I\sigma^2_e$; where $\mathbf{G_R}$ is a GRM derived only from SNPs within the defined region.

Functional population-sense heritability

Genes are the most important functional units of the genome. In order to investigate their contribution to variation in cognition we partitioned, for each of the autosomes, the genetic variance captured by SNPs located inside and outside genes. SNPs mapping to each autosome were separated into those that mapped within 5 kb of the transcription start and end sites of a gene (i.e. within genes) and those that mapped outside these limits. Genome build 37 was used to identify genes and gene limits. A linear mixed model was used to fit forty-four variance components simultaneously, capturing SNPs within genes and SNPs outside genes on each of the 22 human autosomes [Model 4].

$$Y = X\beta + \sum_{c=1}^{22} Iu_c^{in} + \sum_{c=1}^{22} Iu_c^{out} + e \qquad \text{(Model 4)}$$

Where $\mathbf{u_c}^{\boldsymbol{in}}$ is the vector of additive genomic random effects which for each chromosome is solved using a GRM derived from SNPs which lie within genes or within a 5 kb boundary of a gene on that chromosome c; $\mathbf{u_c}^{\boldsymbol{out}}$ is a vector of additive genomic random effects solved using a GRM derived from SNPs which lie outside genes on that chromosome c.

For comparison we grouped SNPs by chromosome and the population-sense heritability was estimated for individual chromosomes [Model 5]. This approach was used previously in a meta-analysis of five cohorts including those described here for adult fluid and crystallised intelligence [8] but not for cognitive change.

$$Y = X\beta + \sum_{c=1}^{22} Iu_c + e \qquad \text{(Model 5)}$$

Where $\mathbf{u}_c$ is the vector of additive genomic random effects on chromosome c solved for each chromosome using a GRM derived from SNPs which lie on that chromosome c.

Model fitting

Initially all SNPs were fitted in the model to estimate the genetic variance and overall heritability for the three cognitive traits in the population. Bivariate analyses to estimate covariances amongst the three cognitive measures were performed using ASReml 2 software [33]. To avoid confounding of genetic variation of the trait and potential variation due to population stratification, eigenvectors were estimated from the genetic relationship matrix and the first 20 principal components were fitted as covariates in the linear mixed model. Sex was also fitted into the model. Analyses were subsequently carried out fitting the regions defined above to estimate regional and functional population-sense heritability.

GCTA/ACTA [34] solves the LMM and obtains estimates of genetic and residual variances by restricted maximum likelihood (REML) using the average information (AI) algorithm.

Test statistics were obtained using a standard likelihood ratio test (LRT) statistic calculated as twice the difference between the log likelihoods of the full model and a null or reduced model that did not fit a genetic component. For a single test, the expectation of the LRT for testing one extra variance component is a 50:50 mixture of a point mass of 0 and a chi square distribution with 1df [35]. This is so because under the null hypothesis the true value of the variance components is on the boundary of the parameter space defined by the alternative hypothesis.

Results from the 10,908 regions were ranked by likelihood ratio test statistic. The top ten non –overlapping or approximately 0.1% of regions were fitted back into a linear model with an eleventh 'polygenic' variance component comprising all the available autosomal SNPs. This model was tested against a null model containing only the polygenic variance component under the expectation that the likelihood ratio test is distributed as a chi-square with ten degrees of freedom. We repeated the analyses without the 'polygenic' variance component and obtained virtually the same results.

Finally, the contribution of the identified top ten regions for each of the traits were analysed for putative pleiotropic effects across cognitive phenotypes.

Permutation analysis

To date there is little evidence for the empirical distribution of a suitable threshold for the LRT statistic when testing multiple genomic regions. Rowe et al. [36] showed that for variance components based QTL mapping methods, the test statistic and the variance explained can be hugely inflated if multiple testing and the underlying genetic architecture are not properly accounted for. Given that over 10 000 tests were performed, many of which were highly correlated due to the overlap of regions, and the novelty of the approach, we derived the empirical distribution of the test statistic using ACTA [34] to perform 100 permutations for each of the traits resulting in empirical thresholds for individual tests ranging from 17.6 for g_f to 18.8 for g_c for a type 1 error rate of 5%. As 100 permutations is not sufficient to ensure a stable estimate of the threshold, but testing 10,908 regions for three traits hundreds of times is computationally intensive, we repeated the analyses using non-overlapping windows and carried out a further 500 permutations. A permutation involved randomly permuting the phenotypic and genotypic data and testing 5454 alternate or non-over-lapping regions on the permuted data set. For each set of permuted data; i) regional population-sense heritabilities were estimated for all regions (each spanning 101 SNPs) and ii) The top ten regions ranked by LRT test statistic from each permutation were simultaneously fitted into a linear model to determine their combined contribution. These were fitted with and without a 'polygenic' component. This gave the empirical distribution of the test statistic under the null hypothesis for fitting a single region and for when the ten top ranking regions are fitted simultaneously.

'One SNP at a time' genome-wide association analysis

The software package PLINK [37] was used to carry out single SNP association tests to assess whether the SNPs of greatest significance were associated with the regions from [4] that explained the greatest amount of genetic variation.

Results

Variance captured by all autosomal SNPs or population-sense heritability

For simplicity we define the proportion of phenotypic variance captured by SNP genotypes in unrelated individuals as population-sense heritability (h^2_{ps}) to distinguish it from the often used narrow and broad sense heritability [29]. Heritabilities, phenotypic and genetic correlations are given in Table 1. Population-sense heritability estimates for cognitive traits ranged from 0.19 (s.e. 0.2) to 0.37 (s.e. 0.19). Estimates for crystallised intelligence are similar to those from the larger previous study [8]. Fluid intelligence estimates differ slightly due to differences in sample size, study design and population demographics. Fluid intelligence was highly genetically correlated to both cognitive change $r_A = 0.95$ (s.e. 0.25), and to crystallised intelligence $r_A = 0.66$ (s.e. 0.34) (i.e. the amount of correlation emerging from pleiotropy is high). There was little genetic correlation between crystallised intelligence and cognitive change $r_A = 0.008$ (s.e. 0.53).

Regional population-sense heritability

The distributions of regional population-sense heritability estimates for the three traits are similar. Most regions explain variance close to zero with 1.7 to 2.5% explaining greater than 1% of variation, 0.07 to 0.18% explaining greater than 2%, and only 0.02% explaining greater than 3%.

The likelihood ratio test statistic for the regional heritability scan across the genome and the most significant hits from the genome wide association analyses ($-\log_{10}$P-value>2.7) are given in Figure 1. Table 2 gives details of the top ten regions for each trait ranked by LRT and appendix 2 gives the known genes for each of these regions and pathway analysis. The top ten single SNP associations for the three traits were all within regions with $h^2_{ps}>1\%$ (Table S1 in File S1). The correlation between the greatest $-\log_{10}$ (*P*-value) for SNP association in each region and $-\log_{10}$ *P*-value from the LRT test for each region was 0.52 (Figure 2). When regions were ranked by LRT a region on chromosome 6 ranking 3rd and 4th for cognitive change and fluid intelligence respectively also contained the top SNP in the GWAS for cognitive change. For fluid intelligence, the top ranking region on chromosome 5 spanned the third ranking single SNP association ($P<3.41$E-06). This region on chromosome 5 associated with fluid intelligence was the only region for all three traits to exceed genome-wide significance at the $P<0.10$ threshold. When the top ten regions (Table 2) from each trait were fitted together in a LMM they explained 13% ($P_{perm} = 0.58$), 15% ($P_{perm} = 0.11$) and 18% ($P_{perm} = 0.43$) of the phenotypic variation for crystallised intelligence, fluid intelligence, and cognitive change respectively. Table 3 shows regions that explained greater than 1% of phenotypic variation in more than 1 trait including regions on chromosome 9 and 11 that potentially have pleiotropic effects on all three traits.

Regions were defined by number of SNPs; hence there was variation in physical length of regions across the genome, with the average region spanning 534 kb. No relationship was found between the physical length of a region and its significance or the amount of additive genetic variation explained (Figure S1 in File S1).

Permutation analyses

To estimate empirical thresholds, phenotypic data for each of the three traits were permuted 500 times to attain an estimate of the null distribution when genotype and phenotype were randomly assorted. We performed 5,454 REML analyses across the genome for each of the permuted data sets resulting in over 8.2 million single tests. The results were ranked by log likelihood and compared to a null model using an LRT. The resulting genome-wide significance thresholds for the LRT ($P<0.05$) were 17.2 for crystallised intelligence, 17.5 for fluid intelligence and 17.08 for cognitive change Figure 3 shows that the distributions of the test statistic for the three traits were very similar and that they were highly inflated when compared to the expectation of the null distribution for a single test. Thresholds were close to those for the 10,908 tests but less conservative than a Bonferroni correction for 5,454 independent tests which would result in a 5% threshold of 19.7. Table 4 shows that the genome-wide threshold values were stable after 300 permutations indicating that 500 permutations was sufficient to estimate 5 and 10% genome-wide thresholds.

The distributions from the permutation analysis (Figure 3) show that by chance in 5% of cases the variance explained by a region exceeded 3.8, 3.8 and 4.0% for g_c, g_f and cognitive change respectively.

For each permutation the top ten regions were identified, i.e. those with the greatest likelihoods and fitted simultaneously into a

Table 1. Population-sense heritability (diagonal), phenotypic (upper diagonal) and genetic (lower diagonal) correlations for measures of general intelligence and cognitive decline estimated from relationship matrices based on 547,750 SNP genotypes.

Trait	Crystallised Intelligence	Fluid Intelligence	Cognitive change
Crystallised intelligence (n = 1791)	0.36 (0.19)	0.59 (0.01)	0.22 (0.02)
Fluid intelligence (n = 1706)	0.66 (0.34)	0.19(0.20)	0.78 (0.009)
Cognitive change (n = 1602)	0.0084(0.53)	0.95 (0.25)	0.26(0.22)

Heritabilities on diagonal, genetic correlations below diagonal, phenotypic correlations above diagonal and standard errors given in brackets.
doi:10.1371/journal.pone.0081189.t001

LMM. An LRT was calculated as twice the difference between the log likelihood of a model fitting ten regions and a null model without a genetic effect, and we did not fit a polygenic model when testing the top ten regions. The 95th percentile was used to estimate a 5% genome-wide threshold for significance of the LRT between a model fitting the top ten regions of the genome; and a null model. The polygenic component was omitted as the original genetic structure was removed by the permutation of genotypes and phenotypes. The 5% genome-wide threshold was P<3.3E-24 for crystallised intelligence, P<1.42E-24 for fluid intelligence and P<1.03 E-24.

Functional population-sense heritability

Figure 4 shows estimates of population-sense heritability for each of the 22 autosomes, and for h^2_{ps} estimates using information from SNPs inside genes and estimates using information from SNPs outside genes for each chromosome and trait. For crystallised intelligence heritability estimates from SNPs on autosomes 3, 5, 11, 15 and 19 were significantly different from zero. When divided further chromosomes 9, 15 and 19 had significant estimates for h^2_{ps} within genes. For fluid intelligence, estimates of h^2_{ps} on chromosomes 3, 9 and 10 were significant, explaining 6, 5, and 8% phenotypic variance, respectively.

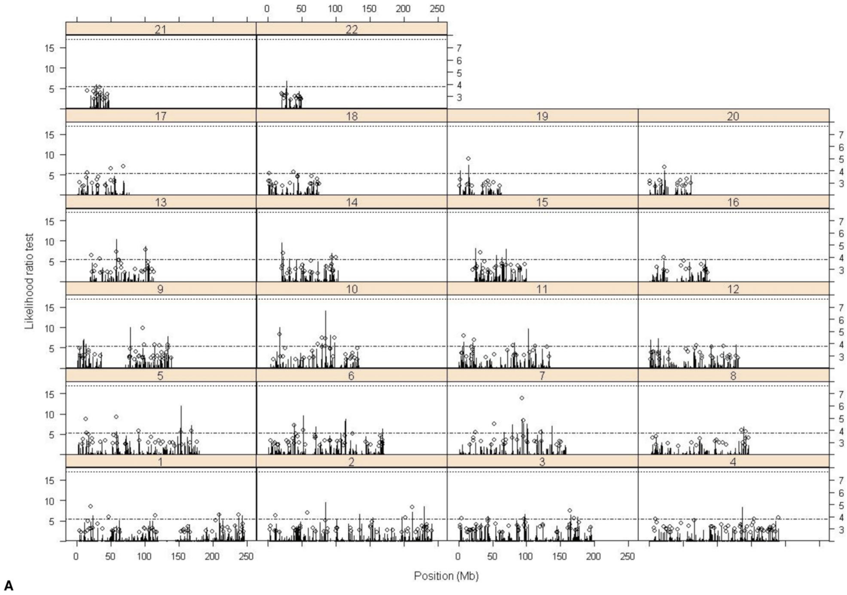


Figure 1. Plot of likelihood ratio test for phenotypic variance explained by each of 10,908 regions (groups of 101 consecutive SNPS) (bars) and $-\log_{10}$ P-values>2.7 for single SNP association (circles). Dashed line is 1% nominal significance threshold for LRT for individual regions, dotted line is 5% genome-wide significance threshold for individual regions obtained by permutation analysis. **A** crystallised intelligence n = 1791, **B** fluid intelligence n = 1706 , and **C** cognitive change n = 1602.
doi:10.1371/journal.pone.0081189.g001

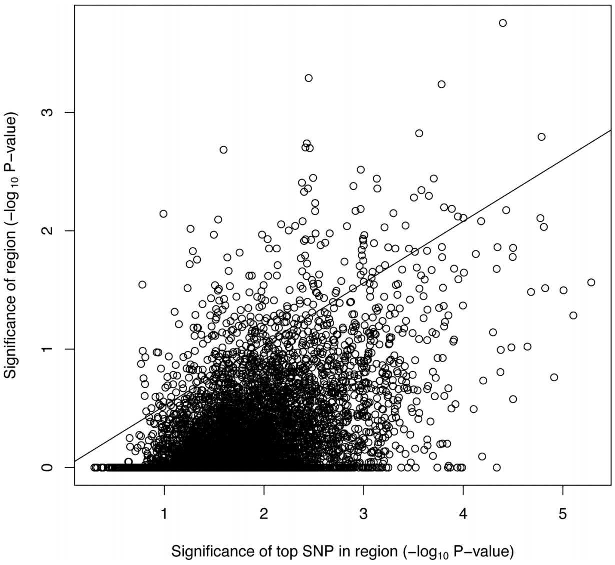


Figure 2. Comparison of significance of region and top SNP within region. Scatter plot of $-\log_{10}$ P-values for single SNP association of greatest significance in region and significance of LRT test for variance explained by entire region (each region contains 101 SNPs). Correlation coefficient is 0.52.
doi:10.1371/journal.pone.0081189.g002

Autosomal h^2_{ps} within genes was significant for chromosomes 9, 14 and 15 and outside genes for chromosomes 3, 16 and 22. For cognitive change chromosomes 4 and 10 had significant estimates of h^2_{ps} with chromosome 6 significant for h^2_{ps} outside genes.

Genetic variation of the traits differed across autosomes and for SNPs within or outside genes. SNPs within genes explained 48, 64 and 38% of the total genetic variation for g_c, g_f and cognitive change respectively. There was no correlation between estimates of autosomal heritability and the number of SNPs used to estimate each genetic relationship matrix (Table S2 in File S1). Distributions of allele frequencies for SNPs inside and outside genes did not differ $P<0.99$).

Brain-related intermediate traits

The top region associated with g_f was genome-wide significant at the 10%, however, even if the locus was truly associated with g_f we were not expecting a high level of statistical support due to the small sample size of the study. In order to gather further independent evidence that could support or reject the association of the locus with g_f we assembled previously published data of brain-measured intermediate phenotypes [38]. Within the chromosome 5 region we found two DNAm sites, cg04431054 and cg15851800 and two mRNA probes ILMN_1652306 and ILMN_1685140. DNAm sites cg04431054 and cg15851800 are located 381 base-pairs apart, cg04431054 is 277 base-pairs upstream of PRRC1, and cg15851800 is 104 base-pairs downstream of the transcription start site of PRRC1, which spans chromosome 5 at base-pair location 126,853,301–126,890,781. ILMN_1685140 targets transcripts of PRRC1 (Proline-Rich Coiled-Coil 1) and ILMN_1652306 transcripts of MEGF10, a receptor for amyloid beta uptake, located between position 126,626,523 and 126,801,429. All four intermediate phenotypes were measured on tissue from the Cerebellum (CRBL), Frontal Cortex (FCTX), the Pons (PONS) and the Temporal Cortex (TCTX). ILMN_1652306 did not pass our quality control procedure for the CRBL and thus was excluded from further analyses. Regional genetic relationships were estimated from 86 available SNPs located within the top 101 SNPs region associated with g_f. h^2_{ps} was estimated with ACTA [34].

The 86 SNPs located on chromosome 5 between 126711782–127335370 base-pairs explain a significant ($P<0.0001$) proportion of the phenotypic variation of cg04431054 for each of the four brain tissues (Table 5). h^2_{ps} of cg04431054 measured in the CRBL, FCTX, PONS and TCTX brain regions was 0.46, 0.24, 0.28 and 0.33, respectively.

So far, we have shown that the 623 kb region of chromosome 5 associated with g_f is associated with cg04431054 levels in the CRBL, FCTX, PONS and TCTX brain regions. However, we have not yet shown a direct link between cg04431054 levels and g_f. To do that, we estimate the effect of the 86 SNPs on the brain phenotypes and construct a genetic score [39] for each individual with g_f phenotypes. A significant regression of genetic score for cg04431054 with g_f would indicate a link between the levels of

Table 2. Variance explained for top ten regions ranked by significance or LRT for crystallised and fluid intelligence and cognitive decline.

Chr	region start (bp)	region end (bp)	h^2_{ps} region	s.e	h^2_{ps} full model[a]	s.e.	LRT	Greatest single SNP association in region $-LOG_{10}$ (P)	SNP Var (r^2)
Crystallised Intelligence									
10	84493034	84943238	0.01	0.008	0.011[b]	0.007	14.08	4.4	0.009
5	153024650	153532086	0.02	0.01	0.017	0.009	12.07	2.45	0.003
10	84323605	84670475	0.02	0.012	-	-	11.85	3.78	0.008
13	57449351	58113705	0.01	0.008	0.01	0.006	10.48	4.45	0.008
9	78430995	78767837	0.01	0.008	0.008	0.006	10.08	3.56	0.008
10	17430161	17790975	0.01	0.008	0.015	0.008	9.95	4.79	0.008
11	102565882	102978790	0.02	0.01	0.015	0.008	9.72	2.43	0.005
14	20640453	21072443	0.03	0.014	0.02	0.012	9.57	2.42	0.004
6	51858157	52238923	0.01	0.008	0.007	0.006	9.55	2.46	0.005
2	84702898	85301342	0.01	0.008	0.013	0.008	9.49	1.59	0.003
13	100772901	101089435	0.02	0.009	0.014	0.008	8.78	2.97	0.002
Fluid Intelligence									
5	126711782	127335370	0.02	0.009	0.013	0.008	16.00	5.47	0.013
6	39140691	39378453	0.03	0.013	0.016	0.012	14.10	3.74	0.009
13	65117143	65633593	0.02	0.01	0.015	0.01	14.07	4.15	0.009
6	740414	1013400	0.02	0.009	0.013	0.008	12.36	4.74	0.011
6	39236400	39493104	0.04	0.018	-	-	12.34	3.52	0.008
11	102565882	102978790	0.02	0.009	0.015	0.008	11.55	4.42	0.010
9	78430995	78767837	0.01	0.009	0.01	0.007	11.07	3.25	0.007
11	102824059	103220693	0.01	0.007	-	-	10.91	3.34	0.007
3	101162780	101999012	0.02	0.011	0.02	0.011	10.55	5.04	0.012
5	33703559	34034521	0.02	0.012	0.016	0.01	9.44	0	0.005
2	151358558	151655394	0.02	0.01	0.012	0.008	9.37	3.43	0.008
5	127010643	127650653	0.01	0.009	0.015	0.008	9.33	0.91	0.007
Cognitive Change									
4	53606097	54158143	0.02	0.009	0.01	0.008	10.44	4.4	0.011
15	90960003	91404141	0.02	0.011	0.017	0.011	10.15	4.83	0.012
6	740414	1013400	0.02	0.009	0.014	0.009	10.08	5.57	0.014
4	62441864	63300488	0.03	0.014	0.024	0.012	9.50	2.57	0.006
6	891665	1138987	0.02	0.009	-	-	8.77	1.56	0.003
6	12418779	12930959	0.02	0.009	0.014	0.009	8.70	4.02	0.010
2	237734083	238123037	0.02	0.011	0.016	0.006	8.30	3.6	0.009
13	98189341	98677491	0.04	0.022	0.035	0.018	8.12	2.59	0.006
14	64270578	64666246	0.02	0.011	0.016	0.01	8.10	2.77	0.006
6	88043140	88678348	0.01	0.007	0.008	0.007	8.08	3.37	0.008
4	148617678	149254898	0.02	0.01	0.023	0.013	7.92	2.71	0.006

[a] heritability of region when full model fitting 11 variance components first ten independent (i.e. non overlapping) regions and rest of genome.

[b] Only the best supported of multiple overlapping regions was fitted.

doi:10.1371/journal.pone.0081189.t002

cg04431054 and g_f. Only one of the four brain regions (TCTX) showed a significant association with g_f ($P = 0.004$), and explained 0.5% of the phenotypic variance. The regression coefficient was positive (0.295, se = 0.004) indicating a positive correlation between methylation levels and g_f. Hence, our analyses of brain-related intermediate phenotypes provides supporting evidence of the region being truly associated with g_f, uncovers the likely target region of the brain associated with g_f and identifies PRRC1 as a candidate gene for g_f.

Discussion

We implemented a recently proposed method of genome scanning by expanding single SNP analysis to the estimation of genetic variance explained by regions spanning 101 co-located SNPs. After deriving empirical thresholds by permutation analysis we show that stringent thresholds close to that of a bonferroni correction are necessary for evaluating the likelihood ratio test statistic and that the distribution of multiple tests is highly inflated

Table 3. Pleiotropic regions affecting multiple traits.

Chr	region start (bp)	region end (bp)	h^2 Crystallised	s.e.	h^2 Fluid	s.e.	h^2 Cog change	s.e.
6	740414	1013400	0.00	0.01	0.02	0.01	0.01	0.01
14	64270578	64666246	0.00	0.00	0.02	0.01	0.02	0.01
9	78430995	78767837	0.01	0.01	0.01	0.01	0.02	0.00
10	17430161	17790975	0.01	0.01	0.02	0.00	0.00	0.01
11	102565882	102978790	0.02	0.01	0.02	0.01	0.01	0.01
11	102824059	103220693	0.02	0.01	0.02	0.01	0.01	0.01

doi:10.1371/journal.pone.0081189.t003

when compared to the null distribution for a single test. This is also true for estimates of heritability (h^2_{ps}). Table 2 shows that within the top ten regions ranked by LRT, only a region on chromosome 6 for fluid intelligence and a region on chromosome 13 for cognitive change explained a greater proportion of the genetic variance (h^2_{ps}) than 95% of the ranked permutation analyses. Despite this the LRT for the comparison of the linear models did not achieve genome wide significance for either of these regions.

We did find a genome wide significant region ($P<0.10$) for the LRT statistic on chromosome 5 associated with fluid intelligence. The region spans the *CTXN3* gene (*cortexin 3*) (Figure 5), a brain (including foetal brain) and kidney specific integral membrane protein, highly enriched in cortex and located on 5q23. This gene has been previously identified as a candidate for schizophrenia and measures of cognitive change [40]. In the GWAS, the third highest ranking SNP *rs790837* ($P<10^{-6}$) is located at position 127004506, 10 kb away from this gene.

The *CTXN3-SLC12A2* region is a strong candidate region and has been linked to brain function and schizophrenia in multiple studies [41,42,43]. The relationship between pre-morbid measures of intelligence and the risk of schizophrenia is also documented as greater than with many other psychoses [44,45]. Although the overlapping region containing *SLC12A2* ranked within the top ten regions with an LRT of 9.33 (Table 2), here the region containing *cortexin 3* (LRT = 16) was the only region to achieve genome wide significance ($P<0.10$). Nonetheless, the strongest evidence suggests that variation of methylation levels at the promoter region of PRRC1 are mediating variation if g_f. The function of the Golgi-associated PRRC1 gene in the brain is unknown and will require follow-up functional studies.

The population-sense heritabilities for fluid intelligence are lower than those previously reported possibly due to an older demographic. Family based (narrow-sense) estimates of heritability for IQ related traits have been shown to decline somewhat with age [46]. This, in part, will be due to an increase in environmental variance.

Autosomal heritability

In general the estimates of genomic heritability for chromosomes reflected the analysis of smaller regions in that the regions with the highest test statistics are located on chromosomes explaining the greatest variance. The sum of heritability estimates for individual chromosomes was inflated by 20–50% compared to estimating the heritability for the entire genome. When heritabilities were estimated from SNPs inside and outside genes (i.e. fitting 44 variance components) heritability for fluid intelligence was doubled when compared to fitting the 22 autosomes (Table S3 in

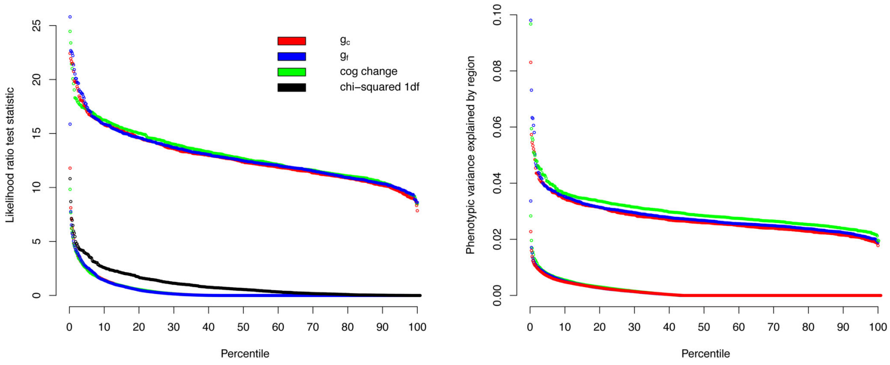


Figure 3. Distribution of the likelihood ratio test and variance explained under the null hypothesis. Comparison of the distribution of likelihood ratio test and variance explained for 5454 regions spanning 101 SNPs for fluid intelligence, crystallised intelligence and cognitive change. Lower set of distributions for each plot are from the real data, upper set are the 5% genome-wide significance threshold from each of 500 permuted data sets i.e. empirical null distribution.
doi:10.1371/journal.pone.0081189.g003

Table 4. Genome wide thresholds for the Likelihood Ratio Test (LRT) derived from N permutations.

Genome-wide threshold for LRT

	Fluid intelligence		**Crystallised intelligence**		**Cognitive Change**	
N	***P*<0.05**	***P*<0.10**	***P*<0.05**	***P*<0.10**	***P*<0.05**	***P*<0.10**
100	19.0	16.5	19.4	16.3	18.0	16.6
200	17.8	15.8	18.1	16.3	17.6	16.5
300	17.5	15.8	17.6	16	17.0	16.1
400	17.4	15.8	17.1	15.8	17.3	16.3
500	17.5	15.8	17.1	15.9	17.2	16.2

doi:10.1371/journal.pone.0081189.t004

File S1). This could be due to fitting so many correlated variance components simultaneously; however, estimates for crystallised intelligence remained stable. It is also possible that this is due to the lack of independence of SNPs within chromosomes inflating estimates, although fitting all 44 variance components simultaneously should account for this. It is probable that more information from a greater number of individuals would enable more precise estimates of covariances and therefore more accurate estimation and partition of variance components. It is also possible that crystallised intelligence is a more polygenic trait with some genetic variance contributed from most chromosomes whereas fluid intelligence and cognitive change show variation around many autosomal estimates which are truly zero.

Pleiotropy

Only 2.5% of regions show an h^2_{ps} greater than 1%. Despite this there is much overlap between the three traits with top regions affecting multiple traits (Table 3). This suggests that the three traits are likely to be affected by the same genes and biological pathways. However, the direction of the effects in these regions will tend to be different for traits such as cognitive change and crystallised intelligence that show a genetic correlation close to zero (Table 1). A single region might also contain linked QTL alleles or regulatory factors in coupling or cis.

It is also feasible that regional significance is biased by other factors making a region more or less likely to explain variation in one or multiple traits. We found no relationship between physical length of region and test statistic. Yang et al. [29] proposed that the genetic variation explained by a region was proportional to the total length of genes. We did not find this in the current study. The unadjusted r^2 values for the relationship between heritability of autosomes and total length of known genes on each chromosome was 0.14, 0.02, and 0.01 for crystallised intelligence, fluid intelligence, and cognitive change with corresponding p-values of 0.07, 0.54, and 0.65. It is possible that this is dependent on the heritability and the genetic architecture of the trait, i.e. the more

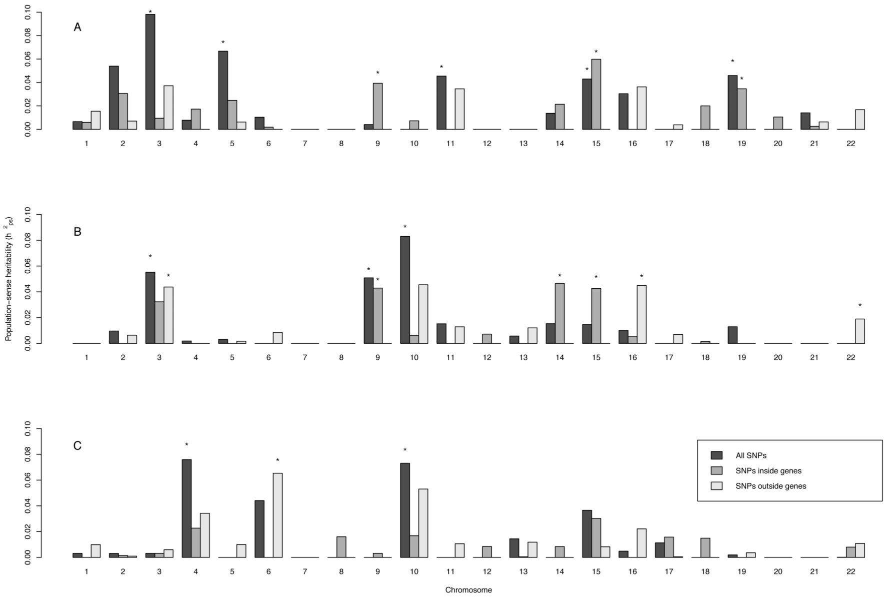


Figure 4. Distribution of population sense-heritability inside and outside genes. Distribution of heritability estimated from all SNPs, SNPs inside genes and SNPs outside genes by chromosome for crystallised intelligence, fluid intelligence and cognitive decline.
doi:10.1371/journal.pone.0081189.g004

Table 5. Population-sense regional heritability for each brain-measured intermediate phenotype within the top g_f associated region on chromosome 5.

		Regional heritability of brain-measured intermediate traits		
Intermediate Phenotype	**Tissue**	**h^2_{ps}**	**SE**	**P**
cg04431054	*CRBL*	*0.463*	*0.124*	*1.370E-08*
cg15851800	CRBL	0.000	0.075	0.500
cg04431054	*FCTX*	*0.237*	*0.104*	*1.190E-05*
cg15851800	FCTX	0.020	0.050	0.325
cg04431054	*PONS*	*0.278*	*0.111*	*1.270E-05*
cg15851800	PONS	0.003	0.046	0.477
cg04431054	*TCTX*	*0.326*	*0.110*	*1.020E-08*
cg15851800	TCTX	0.082	0.078	0.063
ILMN_1685140	CRBL	0.025	0.053	0.315
ILMN_1652306	FCTX	0.000	0.049	0.500
ILMN_1685140	FCTX	0.000	0.041	0.500
ILMN_1652306	PONS	0.000	0.033	0.500
ILMN_1685140	PONS	0.046	0.051	0.075
ILMN_1652306	TCTX	0.000	0.079	0.500
ILMN_1685140	TCTX	0.000	0.045	0.500

Tissue: brain region, h^2_{ps}: estimated regional population-sense heritability, SE: estimated standard error of the regional population-sense heritability. P:p-value from the LRT test testing the significance of the genetic variance component.
doi:10.1371/journal.pone.0081189.t005

polygenic the trait the higher the correlation between the amount of heritable genetic material on each chromosome and the estimate of heritability. This is reflected in Yang et al.'s report where, although height and BMI were highly correlated with the length of genes, there was variation amongst traits with an r^2 value of only 0.02 for von Willebrand factor.

Distributions of the regional heritability test statistic ($-\log_{10}$ P-value) were compared across traits and gender using a Kolmogorov-Smirnov test. Cognitive change differed from fluid and crystallised intelligence ($P<2.2E\text{-}16$ and $P<5.0E\text{-}11$, respectively). Differences between crystallised and fluid intelligence were less marked ($P<0.01$). Interestingly, we found some evidence that the distribution of heritability across the genome for cognitive change differs in males and females. Genetic variation was higher in females and the Kolmogorov-Smirnov test of the distributions of heritabilities for the 10,908 regions in males (n = 871) and females (n = 933) was suggestive at *P*-value of 0.06, although the test does not account for the correlation of the regions and is likely to be inflated. A previous study showed higher variation within males for a measure of general intelligence [47]. It is possible that the increased environmental variance attributable to old age happens sooner in males than females.

It is not clear from this study whether there is utility in a method which expands single SNP analyses to encompass genomic regions

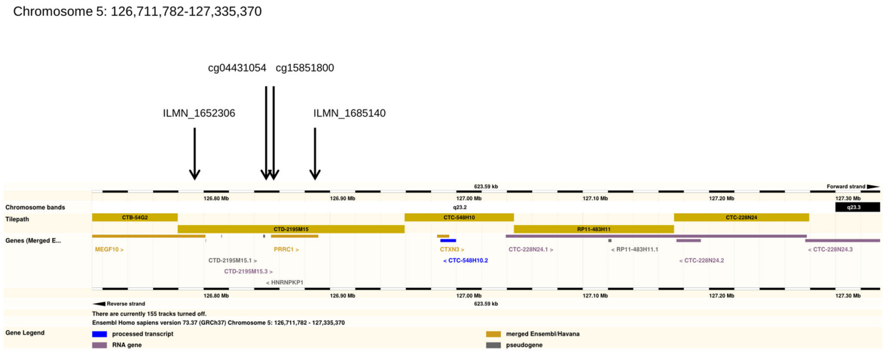


Figure 5. Region on chromosome 5 significantly associated with fluid intelligence. Annotation from Ensembl genome browser.
doi:10.1371/journal.pone.0081189.g005

and that it is able to capture complex local genetic architectures. We acknowledge the limitations of our analysis. Statistical power and accuracy of estimation of variance components is most certainly an issue. Fluid intelligence and cognitive change are important traits and to date lifetime measurements are rare. This limits our ability to increase the sample size. We have shown that the heritability of a region or autosome is not merely a function of its length or the number of genes contained therein. It will be desirable to test the methodology with much larger data sets. It would be interesting to assess whether the regions of greatest significance are enriched for psychiatric genes in comparison to randomly selected regions. Gene set enrichment analyses developed for microarray analysis could be a useful tool for this.

Conclusions

Using a recently proposed population-based linkage scan of the genome we have conducted a search for regions significantly associated with measures of cognition and age related cognitive change. Permutation analysis shows that test statistics and variance explained by a single window were highly inflated when compared to the assumption of a chi square distribution for a single test. We found a significant region on chromosome 5 associated with fluid intelligence explaining 2% of phenotypic variation.

Although single SNP and regional analysis have similar profiles, the ranking of the top regions differ. The regions with the highest test statistic although not genome-wide significant did affect multiple traits and encompass biologically plausible and interesting putative candidate genes. These regions indicate areas of the genome where re-sequencing efforts could be focused to disentangle the fine scale contribution of linked genes and pathways. Although our methodology would benefit from larger sample sizes and increased power, the results give new insights into the study of general intelligence and the underlying mechanisms of cognitive change.

Supporting Information

File S1 This file contains Figure S1 and Tables S1–S3. Figure S1, Plot of relationship between length of region and LRT. **Table S1**, Top Results for single SNP association. **Table S2**, Correlations between autosomal estimates of h^2_{ps} for Crystallised intelligence, Fluid intelligence and Cognitive change with the number of SNPs used to estimate GRM, the number of genes and total length of genes and heritability of autosomes. **Table S3**, Heritability estimates for the 22 autosomes.
(DOCX)

Acknowledgments

We thank the cohort participants who contributed to these studies. We thank Chris Haley for helpful comments and discussion, Andy Law for bioinformatics expertise and Konrad Rawlik for performing the QC of expression data and Oriol Canela-Xandri for his help with the figures. The work was carried out at the Roslin Institute.

Author Contributions

Conceived and designed the experiments: AT. Performed the experiments: AT ID SR JS. Analyzed the data: SR AT AR. Contributed reagents/materials/analysis tools: GD SH DP DL GM JS ID AT. Wrote the paper: SR AT. Critical read of manuscript: SR GD SH DP DL GM JS ID AT AR.

References

1. Batty GD, Deary IJ, Gottfredson LS (2007) Premorbid (early life) IQ and later mortality risk: systematic review. Annals of epidemiology 17: 278–288.
2. Batterham PJ, Christensen H, Mackinnon AJ (2011) Comparison of Age and Time-to-Death in the Dedifferentiation of Late-Life Cognitive Abilities. Psychology and Aging 26: 844–851.
3. Batterham PJ, Mackinnon AJ, Christensen H (2012) The Association Between Change in Cognitive Ability and Cause-Specific Mortality in a Community Sample of Older Adults. Psychology and Aging 27: 229–236.
4. Batty GD, Deary IJ, Benzeval M, Der G (2010) Does IQ predict cardiovascular disease mortality as strongly as established risk factors? Comparison of effect estimates using the West of Scotland Twenty-07 cohort study. European Journal of Cardiovascular Prevention & Rehabilitation 17: 24–27.
5. Batty GD, Kivimaki M, Deary IJ (2010) Intelligence, education, and mortality. British Medical Journal 340.
6. Shipley BA, Der G, Taylor MD, Deary IJ (2007) Association between mortality and cognitive change over 7 years in a large representative sample of UK residents. Psychosomatic Medicine 69: 640–650.
7. Perry RJ, Hodges JR (1999) Attention and executive deficits in Alzheimer's disease. A critical review. Brain : a journal of neurology 122 (Pt 3): 383–404.
8. Davies G, Tenesa A, Payton A, Yang J, Harris SE, et al. (2011) Genome-wide association studies establish that human intelligence is highly heritable and polygenic. Molecular Psychiatry 16: 996–1005.
9. Deary IJ, Yang J, Davies G, Harris SE, Tenesa A, et al. (2012) Genetic contributions to stability and change in intelligence from childhood to old age. Nature 482: 212–215.
10. Cattell RB (1941) Some theoretical issues in adult intelligence testing. Psychological Bulletin 38: 592.
11. Horn JL, Cattell RB (1966) Refinement and test of the theory of fluid and crystallized general intelligences. Journal of educational psychology 57: 253–270.
12. Maher B (2008) Personal genomes: The case of the missing heritability. Nature 456: 18–21.
13. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747–753.
14. Lander ES (2011) Initial impact of the sequencing of the human genome. Nature 470: 187–197.
15. Nagamine Y, Pong-Wong R, Navarro P, Vitart V, Hayward C, et al. (2012) Localising loci underlying complex trait variation using Regional Genomic Relationship Mapping. PloS one 7: e46501.
16. Lee SH, Wray NR, Goddard ME, Visscher PM (2011) Estimating missing heritability for disease from genome-wide association studies. American journal of human genetics 88: 294–305.
17. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. Nature genetics 42: 565–569.
18. Scottish-Council-Research-Education (1933) The intelligence of Scottish children: a national survey of an age-group. Scottish Council for Research in Education.
19. Scottish-Council-Research-Education (1949) The Trend of Scottish Intelligence. Scottish Council for Research in Education.
20. Deary IJ, Gow AJ, Taylor MD, Corley J, Brett C, et al. (2007) The Lothian Birth Cohort 1936: a study to examine influences on cognitive ageing from age 11 to age 70 and beyond. BMC geriatrics 7: 28.
21. Deary IJ, Corley J, Gow AJ, Harris SE, Houlihan LM, et al. (2009) Age-associated cognitive decline. British medical bulletin 92: 135–152.
22. Powell JE, Visscher PM, Goddard ME (2010) Reconciling the analysis of IBD and IBS in complex trait studies. Nature reviews Genetics 11: 800–805.
23. Fernando RL, Grossman M (1989) Marker Assisted Selection Using Best Linear Unbiased Prediction. Genetics Selection Evolution 21: 467–477.
24. Almasy L, Blangero J (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. American journal of human genetics 62: 1198–1211.
25. Goldgar DE (1990) Multipoint analysis of human quantitative genetic variation. American journal of human genetics 47: 957–967.
26. Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.
27. Nejati-Javaremi A, Smith C, Gibson JP (1997) Effect of total allelic relationship on accuracy of evaluation and response to selection. Journal of animal science 75: 1738–1745.
28. VanRaden PM (2008) Efficient methods to compute genomic predictions. Journal of dairy science 91: 4414–4423.
29. Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, et al. (2011) Genome partitioning of genetic variation for complex traits using common SNPs. Nature genetics 43: 519–U544.
30. Falconer DS, Mackay TFC (1996) Introduction to Quantitative Genetics: Longman.
31. Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR (2012) Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics 28: 2540–2542.

32. Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. American journal of human genetics 88: 76–82.
33. Gilmour AR, Thompson R, Cullis BR (1995) Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. Biometrics 51: 1440–1450.
34. Gray A, Stewart I, Tenesa A (2012) Advanced complex trait analysis. Bioinformatics 28: 3134–3136.
35. Self SG, Liang KY (1987) Asymptotic Properties of Maximum-Likelihood Estimators and Likelihood Ratio Tests under Nonstandard Conditions. Journal of the American Statistical Association 82: 605–610.
36. Rowe SJ, Pong-Wong R, Haley CS, Knott SA, De Koning DJ (2008) Detecting dominant QTL with variance component analysis in simulated pedigrees. Genet Res (Camb) 90: 363–374.
37. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. American journal of human genetics 81: 559–575.
38. Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, et al. (2010) Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS genetics 6: e1000952.
39. Wray NR, Goddard ME, Visscher PM (2007) Prediction of individual genetic risk to disease from genome-wide association studies. Genome Research 17: 1520–1528.
40. Potkin SG, Turner JA, Guffanti G, Lakatos A, Fallon JH, et al. (2009) A Genome-Wide Association Study of Schizophrenia Using Brain Activation as a Quantitative Phenotype. Schizophrenia Bulletin 35: 96–108.
41. Panichareon B, Nakayama K, Iwamoto S, Thurakitwannakarn W, Sukhumsirichart W (2012) Association of CTXN3-SLC12A2 polymorphisms and schizophrenia in a Thai population. Behav Brain Funct 8: 27.
42. Lewis CM, Levinson DF, Wise LH, DeLisi LE, Straub RE, et al. (2003) Genome scan meta-analysis of schizophrenia and bipolar disorder, part II: Schizophrenia. Am J Hum Genet 73: 34–48.
43. Almasy L, Gur RC, Haack K, Cole SA, Calkins ME, et al. (2008) A genome screen for quantitative trait loci influencing schizophrenia and neurocognitive phenotypes. Am J Psychiatry 165: 1185–1192.
44. Schulz J, Sundin J, Leask S, Done DJ (2012) Risk of Adult Schizophrenia and Its Relationship to Childhood IQ in the 1958 British Birth Cohort. Schizophr Bull.
45. Gunnell D, Harrison G, Rasmussen F, Fouskakis D, Tynelius P (2002) Associations between premorbid intellectual performance, early-life exposures and early-onset schizophrenia. Cohort study. Br J Psychiatry 181: 298–305.
46. Finkel D, Pedersen NL, McGue M, McClearn GE (1995) Heritability of cognitive abilities in adult twins: comparison of Minnesota and Swedish data. Behavior genetics 25: 421–431.
47. Deary IJ, Irwing P, Der G, Bates TC (2007) Brother-sister differences in the g factor in intelligence: Analysis of full, opposite-sex siblings from the NLSY 1979. Intelligence 35: 451–456.

1. Rowe SJ, Rowlatt A, Davies G, Harris SE, Porteous DJ, et al. (2013) Complex variation in measures of general intelligence and cognitive change. PLoS One 8: e81189.
2. Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, et al. (2010) Abundant Quantitative Trait Loci Exist for DNA Methylation and Gene Expression in Human Brain. PLoS Genet 6: e1000952.
3. Wheeler H, Johnson T (1904) Researches on Pyrimidine Derivatives: 5-Methycytosine. American Chemical Journal 31: 591.
4. Johnson TB, Coghill RD (1925) Researches on Pyrimidines. C111. The Discovery of 5-methy-cytosine in Tuberculinic Acid, The Nucleic Acid of the Tubercle Bacillus1. Journal of the American Chemical Society 47: 2838-2844.
5. Harrison A, Parle-McDermott A (2011) DNA methylation: a timeline of methods and applications. Frontiers in Genetics 2: 74.
6. Laird PW (2010) Principles and challenges of genome-wide DNA methylation analysis. Nat Rev Genet 11: 191-203.
7. Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, et al. (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci U S A 89: 1827-1831.
8. Smith ZD, Chan MM, Mikkelsen TS, Gu H, Gnirke A, et al. (2012) A unique regulatory phase of DNA methylation in the early mammalian embryo. Nature 484: 339-344.
9. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, et al. (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics 11: 587.
10. Bestor TH (2000) The DNA methyltransferases of mammals. Human Molecular Genetics 9: 2395-2402.
11. Bestor TH, Ingram VM (1983) Two DNA methyltransferases from murine erythroleukemia cells: purification, sequence specificity, and mode of interaction with DNA. Proceedings of the National Academy of Sciences 80: 5559-5563.
12. Yoder JA, Soman NS, Verdine GL, Bestor TH (1997) DNA (cytosine-5)-methyltransferases in mouse cells and tissues. studies with a mechanism-based probe1. Journal of Molecular Biology 270: 385-395.
13. Leonhardt H, Page AW, Weier HU, Bestor TH (1992) A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. Cell 71: 865-873.
14. Brero A, Leonhardt H, Cardoso MC (2006) Replication and translation of epigenetic information. Curr Top Microbiol Immunol 301: 21-44.
15. Okano M, Bell DW, Haber DA, Li E (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell 99: 247-257.

16. Kaneda M, Okano M, Hata K, Sado T, Tsujimoto N, et al. (2004) Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. Nature 429: 900-903.
17. Hata K, Okano M, Lei H, Li E (2002) Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. Development 129: 1983-1993.
18. Jia D, Jurkowska RZ, Zhang X, Jeltsch A, Cheng X (2007) Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. Nature 449: 248-251.
19. Cantone I, Fisher AG (2013) Epigenetic programming and reprogramming during development. Nature Structural Molecular Biology 20: 282-289.
20. Maupetit-Méhouas S, Nury D, Arnaud P (2013) Epigenetic Reprogramming in the Mammalian Germline. In: Naumova AK, Greenwood CMT, editors. Epigenetics and Complex Traits: Springer New York. pp. 3-34.
21. Holiday R, Pugh JE (1975) DNA Modification Mechanisms and Gene Activity during Developement Science 187: 226-232.
22. Sandovici I (2013) Establishment of Tissue-Specific Epigenetic States During Development. In: Naumova AK, Greenwood CMT, editors. Epigenetics and Complex Traits: Springer New York. pp. 35-62.
23. Weaver JR, Susiarjo M, Bartolomei MS (2009) Imprinting and epigenetic changes in the early embryo. Mammalian Genome 20: 532-543.
24. Hiura H, Obata Y, Komiyama J, Shirai M, Kono T (2006) Oocyte growth-dependent progression of maternal imprinting in mice. Genes to Cells 11: 353-361.
25. Li JY, Lees-Murdock DJ, Xu GL, Walsh CP (2004) Timing of establishment of paternal methylation imprints in the mouse. Genomics 84: 952-960.
26. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462: 315-322.
27. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, et al. (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. Nature Genetics 38: 1378-1385.
28. Rakyan VK, Hildmann T, Novik KL, Lewin J, Tost J, et al. (2004) DNA Methylation Profiling of the Human Major Histocompatibility Complex: A Pilot Study for the Human Epigenome Project. PLoS Biol 2: e405.
29. Ziller MJ, Gu H, Muller F, Donaghey J, Tsai LT, et al. (2013) Charting a dynamic DNA methylation landscape of the human genome. Nature 500: 477-481.
30. Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, Buil A, Ongen H, et al. (2013) Passive and active DNA methylation and the interplay with genetic variation in gene regulation. eLife 2: e00523.
31. Lam LL, Emberly E, Fraser HB, Neumann SM, Chen E, et al. (2012) Factors underlying variable DNA methylation in a human community cohort. Proceedings of the National Academy of Sciences of the United States of America 109: 17253-17260.

32. Chiba H, Hiura H, Okae H, Miyauchi N, Sato F, et al. (2013) DNA methylation errors in imprinting disorders and assisted reproductive technology. Pediatr Int 55: 542-549.
33. Robertson KD (2005) DNA methylation and human disease. Nat Rev Genet 6: 597-610.
34. Laird PW (2003) The power and the promise of DNA methylation markers. Nat Rev Cancer 3: 253-266.
35. Bell C (2013) Epigenome-Wide Association Studies: Potential Insights into Human Disease. In: Naumova AK, Greenwood CMT, editors. Epigenetics and Complex Traits: Springer New York. pp. 287-317.
36. Lynch M, Walsh B (1997) Genetics and Analysis of Quantitative Traits. Sunderland Massachusetts: Sinauer Associates.
37. Falconer DS, Mackay TFC (1996) Introduction to Quantitative Genetics.
38. Fisher RA (1918) The Correlation between Relatives on the Supposition of Mendelian Inheritance. Transactions of the Royal Society of Edinburgh 52: 399-433.
39. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.
40. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431: 931-945.
41. The International HapMap Consortium (2003) The International HapMap Project. Nature 426: 789-796.
42. The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.
43. The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861.
44. Black JR, Clark SJ (2015) Age-related macular degeneration: genome-wide association studies to translation.
45. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, et al. (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet 46: 1173-1186.
46. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747-753.
47. Visscher PM, Medland SE, Ferreira MAR, Morley KI, Zhu G, et al. (2006) Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings. PLoS Genet 2: e41.
48. Manolio TA, Brooks LD, Collins FA (2008) A HapMap harvest of insights into the genetics of common disease. Journal of Clinical Medicine 118: 1590-1605.
49. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. (2010) Common SNPs explain a large proportion of heritability for human height. Nature genetics 42: 565-569.
50. Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, et al. (2007) Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature 448: 470-473.

51. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, et al. (2008) Genetics of gene expression and its effect on disease. Nature 452: 423-428.

52. Fransen K, Visschedijk MC, van Sommeren S, Fu JY, Franke L, et al. (2010) Analysis of SNPs with an effect on gene expression identifies UBE2L3 and BCL3 as potential new risk genes for Crohn's disease. Hum Mol Genet 19: 3482-3488.

53. Dermitzakis ET (2008) From gene expression to disease risk. Nat Genet 40: 492-493.

54. Zhang D, Cheng L, Badner JA, Chen C, Chen Q, et al. (2010) Genetic Control of Individual Differences in Gene-Specific Methylation in Human Brain. The American Journal of Human Genetics 86: 411-419.

55. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, et al. (2011) DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biology 12: 405-418.

56. Gervin K, Hammerø M, Akselsen HE, Moe R, Nygård H, et al. (2011) Extensive variation and low heritability of DNA methylation identified in a twin study. Genome Research: 1813-1821.

57. Kaminsky ZA, Tang T, Wang S-C, Ptak C, Oh GHT, et al. (2009) DNA methylation profiles in monozygotic and dizygotic twins. Nature Genetics 41: 240-245.

58. McRae A, Powell J, Henders A, Bowdler L, Hemani G, et al. (2014) Contribution of genetic variation to transgenerational inheritance of DNA methylation. Genome Biology 15: R73.

59. Boks MP, Derks EM, Weisenberger DJ, Strengman E, Janson E, et al. (2009) The Relationship of DNA Methylation with Age, Gender and Genotype in Twins and Healthy Controls. PLoS ONE 4: e6767.

60. Gordon L, Joo JE, Powell JE, Ollikainen M, Novakovic B, et al. (2012) Neonatal DNA methylation profile in human twins is specified by a complex interplay between intrauterine environmental and genetic factors, subject to tissue-specific influence. Genome Res 22: 1395-1406.

61. van Eijk K, de Jong S, Boks M, Langeveld T, Colas F, et al. (2012) Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. BMC Genomics 13: 636.

62. Hellman A, Chess A (2010) Extensive sequence-influenced DNA methylation polymorphism in the human genome. Epigenetics & Chromatin 3: 11.

63. Drong AW, Nicholson G, Hedman AK, Meduri E, Grundberg E, et al. (2013) The presence of methylation quantitative trait loci indicates a direct genetic influence on the level of DNA methylation in adipose tissue. PLoS One 8: e55923.

64. Nagamine Y, Pong-Wong R, Navarro P, Vitart V, Hayward C, et al. (2012) Localising Loci underlying Complex Trait Variation Using Regional Genomic Relationship Mapping. PLoS ONE 7: e46501.

65. Uemoto Y, Pong-Wong R, Navarro P, Vitart V, Hayward C, et al. (2013) The power of regional heritability analysis for rare and common variant detection:

simulations and application to eye biometrical traits. Frontiers in Genetics 4: 232.

66. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics 81: 559-575.
67. LaFramboise T (2009) Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. Nucleic Acids Research 37: 4181-4193.
68. Weale M (2010) Quality control for genome-wide association studies. Methods Molecular Biology 628: 341-372.
69. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, et al. (2010) Data quality control in genetic case-control association studies. Nat Protocols 5: 1564-1573.
70. Kuan PF, Wang S, Zhou X, Chu H (2010) A statistical framework for Illumina DNA methylation arrays. Bioinformatics.
71. Bibikova M, Lin Z, Zhou L, Chudin E, Garcia EW, et al. (2006) High-throughput DNA methylation profiling using universal bead arrays. Genome Research 16: 383-393.
72. Smyth GK (2005) limma: Linear Models for Microarray Data. In: Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S, editors. Bioinformatics and Computational Biology Solutions Using R and Bioconductor: Springer New York. pp. 397-420.
73. Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biology 5: R80.
74. Kauffmann A, Gentleman R, Huber W (2009) arrayQualityMetrics—a bioconductor package for quality assessment of microarray data. Bioinformatics 25: 415-416.
75. R Development Core Team (2011) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
76. Sokal RR, Rohlf FJ (1995) Linear Regression. Biometry. 3rd ed. United States of America: Fourth Printing. pp. 455-457.
77. Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: A Tool for Genome-wide Complex Trait Analysis. The American Journal of Human Genetics 88: 76-82.
78. Cebamanos L, Gray A, Stewart I, Tenesa A (2014) Regional heritability advanced complex trait analysis for GPU and traditional parallel architectures. Bioinformatics.
79. VanRaden PM (2008) Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science 91: 4414-4423.
80. Stram DO, Lee JW (1994) Variance components testing in the longitudinal mixed effects model. Biometrics 50: 1171-1177.
81. Sokal RR, Rohlf FJ (1995) Introduction to Analysis of Variance. Biometry. United States of America: Fourth Printing. pp. 180-184.

82. Gray JR, Thompson PM (2004) Neurobiology of intelligence: science and ethics. Nat Rev Neurosci 5: 471-482.
83. Quon G, Lippert C, Heckerman D, Listgarten J (2013) Patterns of methylation heritability in a genome-wide analysis of four brain regions. Nucleic Acids Research 41: 2095-2104.
84. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, et al. (2012) The accessible chromatin landscape of the human genome. Nature 489: 75-82.
85. Jiang YL, Rigolet M, Bourc'his D, Nigon F, Bokesoy I, et al. (2005) DNMT3B mutations and DNA methylation defect define two types of ICF syndrome. Hum Mutat 25: 56-63.
86. El-Maarri O, Kareta MS, Mikeska T, Becker T, Diaz-Lacava A, et al. (2009) A systematic search for DNA methyltransferase polymorphisms reveals a rare DNMT3L variant associated with subtelomeric hypomethylation. Hum Mol Genet 18: 1755-1768.
87. Friso S, Choi S-W, Girelli D, Mason JB, Dolnikowski GG, et al. (2002) A common mutation in the 5,10-methylenetetrahydrofolate reductase gene affects genomic DNA methylation through an interaction with folate status. Proceedings of the National Academy of Sciences 99: 5606-5611.
88. Winnepenninckx B, Debacker K, Ramsay J, Smeets D, Smits A, et al. (2007) CGG-repeat expansion in the DIP2B gene is associated with the fragile site FRA12A on chromosome 12q13.1. Am J Hum Genet 80: 221-231.
89. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The human genome browser at UCSC. Genome Res 12: 996-1006.
90. Alkan C, Cardone MF, Catacchio CR, Antonacci F, O'Brien SJ, et al. (2011) Genome-wide characterization of centromeric satellites from multiple mammalian genomes. Genome Res 21: 137-145.
91. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, et al. (2001) Linkage disequilibrium in the human genome. Nature 411: 199-204.
92. Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, et al. (2011) Genomic inflation factors under polygenic inheritance. Eur J Hum Genet 19: 807-812.
93. Visscher PM (2006) A note on the asymptotic distribution of likelihood ratio tests to test variance components. Twin Res Hum Genet 9: 490-495.
94. Tapp HS, Commane DM, Bradburn DM, Arasaradnam R, Mathers JC, et al. (2013) Nutritional factors and gender influence age-related DNA methylation in the human rectal mucosa. Aging Cell 12: 148-155.
95. Du P, Kibbe WA, Lin SM (2008) lumi: a pipeline for processing Illumina microarray. Bioinformatics 24: 1547-1548. doi: 1510.1093/bioinformatics/btn1224. Epub 2008 May 1548.
96. Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, et al. (2013) Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics 8: 203-209.
97. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, et al. (2013) A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics 29: 189-196.

98. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, et al. (2011) High density DNA methylation array with single CpG site resolution. Genomics 98: 288-295.

99. Hindorff LA, MacArthur J, Morales J, A. JH, Hall PN, et al. A Catalog of Published Genome-wide Association Studies. Available at: http://www.genome.gov/gwastudies.

100. Heyn H, Sayols S, Moutinho C, Vidal E, Sanchez-Mut JV, et al. (2014) Linkage of DNA methylation quantitative trait loci to human cancer risk. Cell Reports 7: 331-338.

101. Weber M, Hellmann I, Stadler MB, Ramos L, Paabo S, et al. (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nature Genetics 39: 457-466.

102. Lokk K, Modhukur V, Rajashekar B, Martens K, Magi R, et al. (2014) DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. Genome Biology 15: r54.

103. Cohen NM, Kenigsberg E, Tanay A (2011) Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. Cell 145: 773-786.

104. Saxonov S, Berg P, Brutlag DL (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proceedings of the National Academy of Sciences of the United States of America 103: 1412-1417.

105. Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D/'Souza C, et al. (2010) Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature 466: 253-257.

106. Ladd-Acosta C, Pevsner J, Sabunciyan S, Yolken RH, Webster MJ, et al. (2007) DNA Methylation Signatures within the Human Brain. The American Journal of Human Genetics 81: 1304-1315.

107. Davies MN, Volta M, Pidsley R, Lunnon K, Dixit A, et al. (2012) Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood. Genome Biol 13: R43.

108. Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, et al. (2015) Human body epigenome maps reveal noncanonical DNA methylation variation. Nature 523: 212-216.

109. Sokal RR, Rohlf FJ (1995) Combining Probabilities From Tests of Significance. Biometry. 3rd ed. United States of America: Fourth Printing. pp. 794-797.

110. Paternoster R, Brame R, Mazerolle P, Piquero A (1998) Using the Correct Statistical Test for the Equality of Regression Coefficients. Criminology 36: 859-866.

111. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29.

112. Rubinstein WS, Maglott DR, Lee JM, Kattman BL, Malheiro AJ, et al. (2012) The NIH genetic testing registry: a new, centralized database of genetic tests to enable access to comprehensive information and improve transparency. Nucleic Acids Research.

113. Kamiyama S, Sasaki N, Goda E, Ui-Tei K, Saigo K, et al. (2006) Molecular cloning and characterization of a novel 3'-phosphoadenosine 5'-phosphosulfate transporter, PAPST2. J Biol Chem 281: 10945-10953.
114. Kamiyama S, Ichimiya T, Ikehara Y, Takase T, Fujimoto I, et al. (2011) Expression and the role of 3'-phosphoadenosine 5'-phosphosulfate transporters in human colorectal carcinoma. Glycobiology 21: 235-246.
115. Wang W, Ni Q, Wang H, Zhang S, Zhu H (2014) Prognostic value of Rab27B nuclear expression in gastrointestinal stromal tumors. Dis Markers 2014: 942181.
116. Zhang JX, Huang XX, Cai MB, Tong ZT, Chen JW, et al. (2012) Overexpression of the secretory small GTPase Rab27B in human breast cancer correlates closely with lymph node metastasis and predicts poor prognosis. J Transl Med 10: 242.
117. Schadt EE, Molony C, Chudin E, Hao K, Yang X, et al. (2008) Mapping the genetic architecture of gene expression in human liver. PLoS Biol 6: e107.
118. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, et al. (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. Nature 464: 773-777.
119. Grassi MA, Tikhomirov A, Ramalingam S, Below JE, Cox NJ, et al. (2011) Genome-wide meta-analysis for severe diabetic retinopathy. Hum Mol Genet 20: 2472-2481.
120. Melville SA, Buros J, Parrado AR, Vardarajan B, Logue MW, et al. (2012) Multiple loci influencing hippocampal degeneration identified by genome scan. Ann Neurol 72: 65-75.
121. Lauc G, Huffman JE, Pucic M, Zgaga L, Adamczyk B, et al. (2013) Loci associated with N-glycosylation of human immunoglobulin G show pleiotropy with autoimmune diseases and haematological cancers. PLoS Genet 9: e1003225.
122. Palmer ND, Goodarzi MO, Langefeld CD, Wang N, Guo X, et al. (2015) Genetic Variants Associated With Quantitative Glucose Homeostasis Traits Translate to Type 2 Diabetes in Mexican Americans: The GUARDIAN (Genetics Underlying Diabetes in Hispanics) Consortium. Diabetes 64: 1853-1866.
123. Hancock DB, Artigas MS, Gharib SA, Henry A, Manichaikul A, et al. (2012) Genome-wide joint meta-analysis of SNP and SNP-by-smoking interaction identifies novel loci for pulmonary function. PLoS Genet 8: e1003098.
124. Suzuki S, Yoshimura M, Nakayama M, Abe K, Yamamuro M, et al. (2007) A novel genetic marker for coronary spasm in women from a genome-wide single nucleotide polymorphism analysis. Pharmacogenet Genomics 17: 919-930.
125. Kottgen A, Pattaro C, Boger CA, Fuchsberger C, Olden M, et al. (2010) New loci associated with kidney function and chronic kidney disease. Nat Genet 42: 376-384.
126. Karns R, Zhang G, Sun G, Rao Indugula S, Cheng H, et al. (2012) Genome-wide association of serum uric acid concentration: replication of sequence

variants in an island population of the Adriatic coast of Croatia. Ann Hum Genet 76: 121-127.

127. Ripke S, O'Dushlaine C, Chambert K, Moran JL, Kahler AK, et al. (2013) Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat Genet 45: 1150-1159.

128. Consortium C-DGoPD (2013) Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. Lancet 381: 1371-1379.

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

130. Triche TJ, Jr., Weisenberger DJ, Van Den Berg D, Laird PW, Siegmund KD (2013) Low-level processing of Illumina Infinium DNA Methylation BeadArrays. Nucleic Acids Res 41: e90. doi: 10.1093/nar/gkt1090. Epub 2013 Mar 1099.

131. Siegmund KD (2011) Statistical approaches for the analysis of DNA methylation microarray data. Hum Genet 129: 585-595. doi: 510.1007/s00439-00011-00993-x. Epub 02011 Apr 00426.

132. Pidsley R, CC YW, Volta M, Lunnon K, Mill J, et al. (2013) A data-driven approach to preprocessing Illumina 450K methylation array data. BMC Genomics 14:293.: 10.1186/1471-2164-1114-1293.

133. Marabita F, Almgren M, Lindholm ME, Ruhrmann S, Fagerstrom-Billai F, et al. (2013) An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. Epigenetics 8: 333-346.

134. Touleimat N, Tost J (2012) Complete pipeline for Infinium® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. Epigenomics 4: 325-341.

135. Maksimovic J, Gordon L, Oshlack A (2012) SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. Genome Biol 13: R44. doi: 10.1186/gb-2012-1113-1186-r1144.

136. Sheehan N, Didelez V, Burton PR, Tobin MD (2008) Mendelian Randomization and Causal Inference in Observational Epidemiology. PLoS Medicine 5: 1205-1210.

137. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, et al. (2010) Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. PLoS Genet 6: e1000888.

138. Richards AL, Jones L, Moskvina V, Kirov G, Gejman PV, et al. (2012) Schizophrenia susceptibility alleles are enriched for alleles that affect gene expression in adult human brain. Mol Psychiatry 17: 193-201.

139. Schulte EC, Schramm K, Schurmann C, Lichtner P, Herder C, et al. (2014) Blood <italic>cis</italic>-eQTL Analysis Fails to Identify Novel Association Signals among Sub-Threshold Candidates from Genome-Wide Association Studies in Restless Legs Syndrome. PLoS ONE 9: e98092.

140. Leslie R, O'Donnell CJ, Johnson AD (2014) GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. Bioinformatics 30: i185-194.
141. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, et al. (2014) Defining functional DNA elements in the human genome. Proceedings of the National Academy of Sciences 111: 6131-6138.
142. Iacopetta B (2002) Are there two sides to colorectal cancer? Int J Cancer 101: 403-408.
143. Shen X, Wang X, Ning Z, Tsepilov Y, Shirali M, et al. (2015) Simple multi-trait analysis identifies novel loci associated with growth and obesity measures.
144. Relton CL, Davey Smith G (2012) Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. Int J Epidemiol 41: 161-176.