DNA Methylation as a Biomarker for Age-Related Cognitive Impairment

Laura Michelle Barrett

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Institute of Genetic Medicine

September 2014

Abstract

Due to the ageing population, the number of patients diagnosed with age-related diseases such as stroke and Parkinson's disease are on the rise. In both post-stroke dementia (PSD) and mild cognitive impairment in Parkinson's disease (PD-MCI), the mechanisms resulting in cognitive decline are unknown. This project aims to identify a biomarker which could predict those patients most at risk of developing cognitive decline, which would subsequently assist healthcare professionals in recommending early treatment and care.

Epigenetics is an emerging field in which biomarkers have previously been useful in prognostication of cancers and prediction of cardiovascular disease. In this study, 30 patients from a PSD cohort (COGFAST) and 48 patients from a PD-MCI cohort (ICICLE) were analysed using the Illumina HumanMethylation450 BeadChip to identify differentially methylated positions which could predict patients who would later develop cognitive decline. Top hits were validated using Pyrosequencing to confirm DNA methylation differences in a replication cohort.

Individual CpG sites within *APOB* and *NGF* were identified as potential blood-based biomarkers for PSD and one CpG site within *CHCHD5* was highlighted as a potential blood-based biomarker for PD-MCI. In addition, methylation at one CpG site within *NGF* and a CpG site (cg18837178) within a non-coding RNA, were found to be associated with Braak staging (degree of brain pathology) using DNA from two brain regions. *NGF* deregulation has previously been associated with Alzheimer's disease, and this finding indicates it may also have a role in the development of PSD.

These novel findings represent the first steps towards the identification of blood-based biomarkers to assist with diagnosis of PSD and PD-MCI, but require further validation in a larger independent cohort. The differentially methylated genes identified may also give insight into some of the mechanisms involved in these complex diseases, potentially leading to the future development of targeted preventative treatments.

Acknowledgements

Firstly, I would like to thank my supervisors, Professor Caroline Relton, Professor John Mathers and Professor Raj Kalaria for their valuable support and guidance throughout the project. I would also like to thank Dr Louise Allan and Dr Alison Yarnall for their help and advice regarding the cohorts used in this thesis. Special thanks go to Dr Catherine Potter and Dr Hannah Elliott who provided much needed assistance in using R and analysing the HM450 data. Thanks are also given to Polly Usher for her help with the ICICLE Pyrosequencing work. Polly performed all primer optimisations, assay validations and the Pyrosequencing using the ICICLE samples, enabling me to spend more time completing other work. The University of Bristol allowed me to use their Illumina iScan to perform the HM450 arrays. Thank you to Dr Wendy McArdle for her help with the HM450 sample preparation and running the arrays. Thanks also go to the staff at the Newcastle Brain Tissue Resource for providing me with the tissue and information required to complete this work and all participants of COGFAST and ICICLE without whom this project would not have been possible. I would also like to thank my family and friends who all provided invaluable support and encouragement throughout my PhD. Finally, I would like to thank the Medical Research Council and Newcastle University for providing the funding which made this project possible.

Statement of Work Undertaken

Study cohorts used in this project were established prior to the start of this project and all phenotypic data had already been collected. In ICICLE, all DNA had been extracted from whole blood samples prior to the start of this project. In COGFAST, the brain regions were sectioned by the Newcastle Brain Tissue Resource and I performed all subsequent DNA, RNA and protein extractions using both the baseline blood samples and two brain regions. All DNA samples (COGFAST and ICICLE) were then processed by myself and sent to the University of Bristol where Dr Wendy McArdle performed the HM450 BeadChip and sent the raw data back. I then performed all HM450 normalisation, data analysis and selection of top hits. Using the COGFAST samples, I performed all the lab work for the Pyrosequencing and data analysis. Using the ICICLE samples, I designed the primers for the selected assays and all primer optimisations, validations and Pyrosequencing were performed by Polly Usher from Newcastle University. I then analysed the resulting Pyrosequencing data. All other lab work and data analysis were performed by myself.

Table of Contents

List of Tables

List of Figures

Abbreviations

5-aza	5-aza-2’-deoxycytidine
5-hmC	5-hydroxy-methylcytosine
5-mC	5-methylcytosine
AAD	Age acceleration difference
AAR	Age acceleration residual
Aβ	Beta amyloid
AD	Alzheimer’s disease
APOB	Apolipoprotein B
APOE4	Apolipoprotein E4
APP	Amyloid precursor protein
BLAT	BLAST-like alignment tool
BM	Bisulphite modified
bp	Base pairs
BSA	Bovine serum albumin
CA	Healthy controls
CADASIL	Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy
CAMCOG	Cambridge cognitive examination
CANTAB	Cambridge neuropsychological test automated battery
CDR	Cognitive drug research battery
CHCHD5	Coiled-coil-helix-coiled-coil-helix domain containing 5
CN	Cognitively normal
COGFAST	Cognitive function after stroke
CpH	Non-CpG
CRP	C-reactive protein
CSF	Cerebrospinal fluid

CVD	Cardiovascular disease
D	Demented
DBS	Deep brain stimulation
ddNTPs	Dideoxynucleotide triphosphates
DLB	Dementia with Lewy bodies
DLPFC	Dorsolateral prefrontal cortex
DMP	Differentially methylated positions
DMR	Differentially methylated regions
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferases
DS	Alzheimer’s disease, no stroke
DSM	Diagnostic and statistical manual of mental disorders
EDTA	Ethylenediaminetetraacetic acid
EWAS	Epigenome-wide association study
FACS	Fluorescence activated cell sorting
FTD	Frontotemporal dementia
FTD-ALS	Frontotemporal dementia-amyotrophic lateral sclerosis
GDS-15	Geriatric depression scale
GEoCoDE	Genomic and epigenomic complex disease epidemiology
GWAS	Genome-wide association studies
HM450	HumanMethylation450
HPLP	High performance liquid chromatography
HRP	Horseradish peroxidase
ICD	International classification of diseases
ICICLE	Incidence of cognitive impairments in cohorts with longitudinal evaluation
IHD	Ischaemic heart disease

IL-8	Interleukin-8
IQ	Intelligence quotient
IQR	Interquartile range
LACS	Lacunar stroke
LCM	Laser capture microdissection
LDL	Low-density lipoprotein
LINE-1	Long interspersed nucleotide element 1
MAO-B	Monoamine oxidase B
MCI	Mild cognitive impairment
MDS-UPDRS	Movement disorder society-unified Parkinson’s disease rating scale
MiRNA	Micro ribonucleic acid
mmHg	millimetres of mercury
MMSE	Mini mental state examination
MoCA	Montreal cognitive assessment
MR	Mendelian randomisation
MRI	Magnetic resonance imaging
mRNA	Messenger ribonucleic acid
MTHFR	Methylenetetrahydrofolate reductase
MZ	Monozygotic
NBTR	Newcastle brain tissue resource
NGF	Nerve growth factor
NMS	Non-motor symptoms
OCSP	Oxfordshire community stroke project
PACS	Partial anterior circulation stroke
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PD	Parkinson’s disease

PDD	Parkinson’s disease dementia
PD-MCI	Mild cognitive impairment in Parkinson’s disease
PM	Post-mortem
POCS	Posterior circulation stroke
PROGRESS	Perindopril protection against recurrent stroke study
PSD	Post-stroke dementia
QC	Quality control
RCF	Red cell folate
RNA	Ribonucleic acid
SAM	S-adenosylmethionine
SCU	Stroke care unit
SD	Standard deviation
SES	Socio-economic status
SMART	Simple modular architecture research tool
SNCA	Alpha-synuclein
SNP	Single nucleotide polymorphism
SQN	Subset quantile normalisation
ssDNA	Single stranded DNA
TACS	Total anterior circulation stroke
TBE	Tris-borate-ethylenediaminetetraacetic acid
TBS	Tris buffered saline
TBS-T	Tris buffered saline-tween
TIA	Transient ischaemic attack
TNF	Tumour necrosis factor
UTR	Untranslated region
VaD	Vascular dementia
WHO	World health organisation

Xist	X-inactive specific transcript

Glossary of Variables

Table 1. COGFAST Exposures

Variable	Abbreviation	Description
Oxfordshire Community Stroke Project	OCSP	Classification of stroke type (TACS, LACS, PACS, TOCS) using clinical criteria.
Side of body affected by stroke	Side of body	The side of the body (left/right), if any (none), affected by the stroke using clinical data.
Degree of weakness in arm		The classification of the degree of weakness in the limb (no weakness, some weakness, no movement) on presentation of stroke using clinical notes.
Degree of weakness in leg		The classification of degree of weakness in the limb (no weakness, some weakness, no movement) on presentation of stroke using clinical notes.
Dysphasia		The presence or absence of dysphasia at stroke presentation as recorded by clinical notes.
Number of CVD risk factors	No CVD risk factors	Number of Cardiovascular risk factors (Smoking, Diabetes, Hyperlipidaemia, Peripheral vascular disease, IHD, Atrial fibrillation, Hypertension).
Hypertension		Present or absent, defined by blood pressure >140/90 mmHg.
Atrial fibrillation		Present or absent, defined by WHO criteria.
Ischaemic heart disease	IHD	Present or absent, defined by WHO criteria.
Myocardial Infarction		Present or absent, defined by WHO criteria.
Angina		Present or absent, defined by WHO criteria.
Cardiac failure		Present or absent, defined by WHO criteria.
Hypercholesterolemia		Present or absent, defined by WHO criteria.
Diabetes		Present or absent, defined by WHO criteria.
Intermittent claudication		Present or absent, defined by WHO criteria.
Smoking		Smoking status (current, ex or never).

Other COGFAST exposures included in this study were baseline CAMCOG scores. Definitions of these can be found in Table 2. COGFAST Outcomes.

Table 2. COGFAST Outcomes

Variable	Abbreviation	Description
Diagnosis		Diagnosis of post-stroke dementia based on neuropsychometric testing, agreement with DSM IIIR and IV criteria for dementia and post-mortem neuropathological examination.
Braak staging		Degree of pathology (I-VI) assessed by post-mortem neuropathological examination (Braak and Braak, 1995).
Mini Mental State Examination	MMSE	Questionnaire test assessing various cognitive domains. ≤24/30 indicates some degree of cognitive impairment (Folstein *et al.*, 1975).
Orientation*		Score comprised of 10 items from the MMSE.
Language comprehension*		Score out of 9 based on verbal and non-verbal responses to both spoken and written questions.
Language expression*		A combination of naming, repetition, fluency and definitions. Scored out of 21.
Memory remote*		Ability to recall famous historic events and people.
Memory recent*		Ability to recall current news and recent events.
Memory learning*		Ability to recall and recognise pictures and patterns.
Memory total*		A combined score of the remote, recent and learning memory tests. Scored out of 27.
Attention*		Assessed by counting backwards individually and in multiples of 7. Scored out of 7.
Praxis*		Ability to copy, draw and write. Scored out of 12.
Calculation*		The ability to perform addition and subtraction. Scored out of 2.
Abstract thinking*		The ability to form links and similarities between two objects, e.g. apple/banana. Scored out of 8.
Perception*		The ability to identify familiar objects and people. Scored out of 11.
Executive function		An independent assessment of executive function assessing visual reasoning and verbal fluency. Scored out of 28 (Leeds *et al.*, 2001).
Total CAMCOG score*		The total combined score of all CAMCOG tests. Maximum score = 107.

* represents a CAMCOG item (de Koning *et al.*, 1998).

Table 3. ICICLE Exposures

Variable	Abbreviation	Description
Education		Number of years spent in education.
National Adult Reading Test	NART	NART score used to predict premorbid IQ.
Height		Height (m) at recruitment.
Weight		Weight (kg) at recruitment.
Body Mass Index	BMI	BMI (kg/m^2) at recruitment.
Ischaemic heart disease	IHD	Present or absent, defined by WHO criteria.
Diabetes		Present or absent, defined by WHO criteria.
Hypertension		Present or absent, defined by blood pressure >140/90 mmHg.
Hypercholesterolaemia		Present or absent, defined by WHO criteria.
Levodopa daily dose	LEDD	Daily dose of Levodopa or equivalent medication (mg/day).
Geriatric Depression Scale	GDS	Geriatric Depression Scale score. A score ≥5 indicates depression (Yesavage *et al.*, 1982).
Red cell folate	RCF	Red cell folate measurement (µg/L) at recruitment.
Vitamin B12	B12	Vitamin B12 measurement (ng/L) at recruitment.
Homocysteine		Homocysteine measurement (µmol/L) at recruitment.
Alcohol		Average units of alcohol per week.
Smoking		Smoking status (current, ex or never).

Table 4. ICICLE Outcomes

Variable	Abbreviation	Description
COGNITIVE		
Montreal Cognitive Assessment	MoCA	Test assessing global cognitive domains. <26/30 indicates some degree of cognitive impairment (Nasreddine *et al.*, 2005).
Mini Mental State Examination	MMSE	Questionnaire test assessing global cognitive domains. ≤24/30 indicates significant degree of cognitive impairment (Folstein *et al.*, 1975).
Total FAS		Test of phonemic fluency (frontal/executive function) where participants have to list as many words beginning with F, A and S as they can in 60 seconds.
Animals		Test of semantic fluency where participants have to list as many animals as they can in 90 seconds.
Non-motor symptoms questionnaire – memory	NMSQ memory	Self-completed questionnaire assessing participant's memory disturbance as present or absent (Romenets *et al.*, 2012).
Non-motor symptoms questionnaire - concentration	NMSQ concentration	Self-completed questionnaire assessing participant's concentration disturbance as present or absent (Romenets *et al.*, 2012).
Total number of non-motor symptoms	Total no NMS	Total number of all non-motor symptoms experienced.
Power of attention		A sum of reaction times (and measure of attention) measured using the Cognitive Drug Research (CDR) battery (Wesnes *et al.*, 2002). A higher score indicates more impairment.
Digit vigilance accuracy		A domain of the CDR battery measuring attention (Wesnes *et al.*, 2002).
Pattern recognition memory	PRM	Part of the CANTAB and a test of visual memory where participants are tested on their ability to recognise previously seen patterns (Sahakian *et al.*, 1988) (maximum score of 24).
Spatial recognition memory	SRM	Part of the CANTAB and a test of visual memory where participants are tested on their ability to remember the spatial positioning of squares on the screen (Sahakian *et al.*, 1988) (maximum score of 20).
Paired associates learning	PAL	Part of the CANTAB and test of visual memory where participants are tested on their ability to remember the location of patterns (Sahakian *et al.*, 1988) (greater score indicates more impairment).

One touch stockings	OTS	Part of the CANTAB assessing spatial planning (executive function) where participants are tested on their ability to plan the movement of balls to match the pattern displayed on screen (Sahakian *et al.*, 1988) (maximum score of 20).
Pentagon copying		Part of the MMSE testing visuospatial function where participants have to copy pentagons (Folstein *et al.*, 1975).
Naming		Part of the MoCA where participants have to name pictured animals (Nasreddine *et al.*, 2005).
Language		Part of the MoCA where participants have to verbally repeat sentences (Nasreddine *et al.*, 2005).
Language total		The combined score of the naming and language tests above (Nasreddine *et al.*, 2005).
Cognitive complaint		Presence or absence of a self-reported cognitive deficit.
Mild cognitive impairment	MCI	Diagnosis of MCI was given if participants showed impairment (1-2 standard deviations below the normative values) on at least 2 neuropsychological tests.
MOTOR		
Hoehn and Yahr		Part 4 of the MDS-UPDRS. A broad scale assessing the level of disability ranging from 1 (unilateral involvement) to 5 (bedridden/wheelchair bound) (Hoehn and Yahr, 1967).
Movement disorder society - unified Parkinson's disease rating scale part II	MDS-UPDRS II	A self-reported evaluation of the activities of daily life such as cutting food and handwriting (Goetz *et al.*, 2008) with a greater score indicating more perceived impairment in activities of daily living.
Movement disorder society - unified Parkinson's disease rating scale part II	MDS-UPDRS III	An evaluation of all aspects of motor ability assessed by a clinician (Goetz *et al.*, 2008) with a greater score indicating more severe motor disease.
Tremor dominant phenotype		A summed score for the assessment of tremor using the MDS-UPDRS III (Goetz *et al.*, 2008) giving a specific phenotype.
Postural instability gait difficulty phenotype	PIGD	A summed score for the assessment of gait using the MDS-UPDRS III (Goetz *et al.*, 2008), giving a specific phenotype.

Chapter 1. Introduction

1.1 Introduction to molecular epidemiology

Molecular epidemiology is a term referring to the study of the effect of potential exposures (both environmental and genetic), at a molecular level, on the aetiology, prediction or prevention of disease across populations. Molecular epidemiology can be useful in the identification of mechanisms culminating in disease but also in the identification of biomarkers of either exposure or disease which can be considered as indirect mechanisms or may even be unrelated to mechanisms. A biomarker has been defined as “a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention” (Biomarkers Definitions Working Group, 2001).

1.1.1 Use of biomarkers in disease prediction and prognosis

The term biomarker refers to a biological marker that can be quantified in a medical setting and is either associated with an exposure or disease phenotype. There are many different types of biomarkers and many ways in which they may be identified. Some possible techniques used in the identification of biomarkers include; genetic screening, imaging or immunohistochemistry techniques, metabolomics, proteomics, transcriptomics and more recently epigenomics. There are a number of ways in which biomarkers can be useful in disease prediction and prognosis.

These include:

- measuring exposures,
- detecting early stages of disease/predicting individuals at-risk of disease,
- aiding diagnosis of disease,
- disease stratification/staging, and
- treatment monitoring.

Biomarkers used in disease prediction and prognosis are most useful when measured in a relatively non-invasive manner using a readily accessible tissue such as blood. However, this makes it less likely that the biomarker will be mechanistically involved in the pathogenesis of disease since many diseases are localised to specific tissues.

1.1.2 Use of intermediate phenotypes to inform about disease mechanisms

Whilst biomarkers used in the prediction and to some extent the prognosis of disease do not need to be mechanistically linked to disease, some biomarkers can inform researchers as to the mechanisms involved in disease pathogenesis. These are particularly useful in the development of targeted treatments which target the cause of the disease rather than symptoms alone. However, the robust identification of biomarkers to inform about disease mechanisms is more complex than identifying biomarkers to predict disease. The main reason for this is that to reliably detect changes in biological measurements that are mechanistically linked to disease, the diseased tissue is most likely required (Foley *et al.*, 2009; Talens *et al.*, 2010). In diseases such as Alzheimer's disease (AD) and Parkinson's disease (PD) where the diseased tissue is the brain, prospective studies are impossible since the brain can only be sampled post-mortem. It is therefore impossible to study and monitor the effects of treatment on these biological markers. However, for diseases such as autoimmune disease, which affect the blood, this can be an extremely useful function of mechanistic biomarkers.

1.1.3 DNA methylation as a biomarker

A current approach to identifying biomarkers is epigenetics, in particular deoxyribonucleic acid (DNA) methylation, both of which are explained in more detail in Section 1.3. DNA methylation has been described as a useful biomarker in several diseases. If DNA methylation changes are detected prior to disease onset they can be useful in informing disease prevention. Two examples of when DNA methylation changes have been detected prior to disease onset and are potential biomarkers of subsequent disease are outlined below.

Kim *et al.,* (2010) used a population-based prospective cohort to identify DNA methylation differences between participants with cardiovascular disease (CVD) (myocardial infarction and stroke) or a predisposing condition (diabetes and hypertension) and healthy individuals. Global methylation levels were measured and compared in three groups of participants; 1) in participants with CVD related traits at baseline, 2) in participants who were free of CVD related disease at baseline but developed a condition prior to follow up, and 3) in participants who remained free of CVD related disease at follow up. The authors found that participants who developed a CVD related disease during the study had elevated global methylation levels compared to those who remained disease-free. However, those who presented with a disease at baseline had even higher global methylation levels. As methylation differences could be detected prior to the onset of disease this study suggests that global methylation levels could be used to detect those individuals who are at risk of developing CVD and related diseases (Kim *et al.*, 2010). The findings would need to be replicated in an independent study before being evaluated for their clinical utility.

DNA methylation has recently been identified as a possible predictive biomarker of preeclampsia. In a study by Anderson *et al.* (2013), maternal blood DNA was collected during the first trimester of pregnancy and mothers were followed up until birth. A genome-wide methylation array analysing methylation levels at over 480,000 CpG sites identified 207 loci which displayed differential methylation between women who developed preeclampsia compared with those with a normotensive pregnancy. In addition, they also discovered that many of the methylation differences were also present in placental DNA, highlighting the possible transgenerational effects in DNA methylation (Anderson *et al.*, 2013). The study design used here does not allow one to draw conclusions about causality, as pre-existing hypertensive disorders may have been present in early pregnancy that could alter DNA methylation, rather than DNA methylation differences actually inducing preeclampsia.

Nevertheless, despite limitations in study design, these two examples show that it is possible to identify epigenetic biomarkers to enable the early prediction of disease and provides an opportunity to intervene in disease prevention. The use of biomarkers may offer great benefits to the prediction of a range of diseases including neurodegenerative disorders.

1.2 Age-related diseases involving the loss of cognitive function

1.2.1 The ageing brain

There are a number of changes that occur in the human brain as part of the normal ageing process. The ageing brain undergoes structural, chemical and neuropsychological changes. Several normal brain changes are described below.

Gross brain volume has been reported to decrease annually in old age (over 60 years of age), although changes are not uniform across all brain regions. The hippocampus has been widely studied and an average of 0.8 to 2% reduction in brain volume each year has been reported. Vast reductions in cortical volumes have been reported, especially in the frontal and temporal lobes. Brain atrophy is also a key feature of Alzheimer's disease (AD), however studies have identified that atrophy rates are approximately 1% higher in AD affected brains (Fjell *et al.*, 2009).

A number of studies have reported changes in neuronal numbers, however due to the difficulty involved in performing these counts and the possibility that neurons shrink rather than disappear, many conflicting reports have been published. It is thought however, that in certain regions of the brain, such as the hippocampus, there is some neuronal loss with age (Anderton, 2002).

The brain also experiences chemical changes with age, including levels of a number of neurotransmitters found within the brain. From early adulthood, levels of dopamine, which is a neurotransmitter with key roles in motor control and cognition, are found to reduce by 10% per decade. Levels of serotonin, which is an important neurotransmitter known to regulate learning and memory, are also seen to reduce in the adult ageing brain (Peters, 2006). Both dopamine and serotonin are associated with age-related disorders such as Parkinson's disease (Morgan *et al*., 1987; Xu *et al*., 2012).

Some degree of cognitive decline is also expected during normal ageing with the most commonly affected areas of cognition being attention and memory. The accumulation of oxidative damage caused by reactive oxygen species over the life course is also thought to contribute to this later life cognitive decline (Lovell and Markesbery, 2007).

Many of the brain changes described above, that occur as part of the normal ageing process could be used to explain why cognitive decline can be experienced in otherwise unaffected brains. However, the brain changes described above are also symptoms of many age-related diseases affecting the brain. The use of symptoms alone in the diagnosis of such diseases would therefore be inaccurate. The use of markers, such as molecular biomarkers, not seen in the normal ageing process, would therefore be beneficial in distinguishing between a normal aged and a diseased brain.

A number of age-related diseases exist which affect cognitive function including AD and other dementias and Parkinson's disease. This thesis will be restricted to consideration of post-stroke dementia and Parkinson's disease.

1.2.2 Stroke

1.2.2.1 *Incidence*

Approximately fifteen million incident strokes occur worldwide annually resulting in five million deaths and an additional five million survivors permanently unable to live independently (Redon *et al.*, 2011). This accounts for a loss of 44 million disability-adjusted life years (Mukherjee and Patil, 2011). The incidence rate for men is 1.25 times greater than for women, however more women die each year as a result of stroke, probably due to their extended lifespan (Sacco *et al.*, 1997). Worryingly, due to the increasingly ageing population and the increasing prevalence of the major risk factors for stroke such as hypertension, the burden of stroke is set to rise over the coming decades.

A stroke occurs due to a disruption of the cerebral blood supply resulting in a focal neurological deficit (Markus, 2011). The World Health Organisation (WHO) defines a stroke as "*rapidly developing clinical signs of focal (at times global) disturbance of cerebral function, lasting more than 24 hours or leading to death with no apparent cause other than of vascular origin*" (Sudlow and Warlow, 1996). Symptoms include sudden onset of impaired speech, loss of vision and paralysis, often only affecting one side of the body (Moskowitz *et al.*, 2010).

1.2.2.2 *Types of stroke*

Approximately 80% of all strokes are ischaemic whilst the remaining 20% are haemorrhagic (Figure 1). Ischaemic strokes occur when cerebral blood flow is obstructed. There are a number of causes for this. The most common cause, being responsible for approximately half of all ischaemic strokes, is atherothrombotic disease of the large extracranial or occasionally intracranial arteries supplying the brain. Lacunar infarcts, caused by the occlusion of a small, deep, perforating cerebral artery within the white and deep grey matter of the brain are responsible for about 25% of ischaemic strokes. 20% of ischaemic strokes arise when a clot (which usually forms in the heart) breaks off and embolises to the brain. This is known as cardioembolism. Atrial fibrillation is a common cause of embolism or blockage of an artery to the heart. The remaining 5% of ischaemic strokes have much rarer causes, including vasculitis, the inflammatory destruction of blood vessels (Sudlow and Warlow, 1996; Markus, 2011). Haemorrhagic strokes can be split into two subtypes. 75% are primary intracerebral haemorrhages where a blood vessel inside the brain ruptures and 25% are subarachnoid haemorrhages with the burst vessel occurring in the subarachnoid space (Warlow *et al.*, 2003). Some patients may also suffer from a mini stroke or a transient ischaemic attack (TIA). This is very similar to a stroke but has a better prognosis as symptoms dissipate within 24 hours (Warlow *et al.*, 2003).

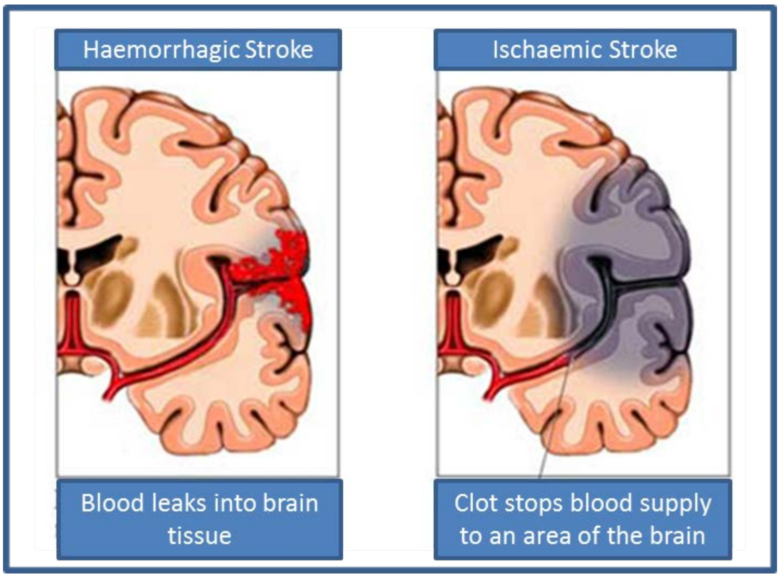

Haemorrhagic Stroke
Ischaemic Stroke
Blood leaks into brain tissue
Clot stops blood supply to an area of the brain

Figure 1: Types of stroke. Adapted from http://www.webmd.com/stroke/ischemic-versus-hemorrhagic-stroke.

1.2.2.3 *Risk factors*

There are many risk factors for stroke, many of which are related to lifestyle and therefore could be modifiable. With the successful management of risk factors, it has been estimated that up to 75% of strokes can be prevented (Cumming and Brodtmann, 2011). Table 1 shows a list of modifiable and non-modifiable risk factors. The key risk factors are explained in more detail below.

Modifiable	**Potentially modifiable**	**Non-modifiable**
Hypertension	Diabetes mellitus	Age
Atrial fibrillation	Hyperhomocysteinaemia	Sex
Infective endocarditis	Left ventricular hypertrophy	Hereditary/familial factors
Mitral stenosis		Race/ethnicity
Recent large myocardial infarction		Geographic location
Cigarette smoking		Transient ischaemic attack
Sickle cell disease		
Asymptomatic carotid stenosis		

Table 1: A list of well-documented risk factors. Adapted from Sacco *et al.* (1997).

As we age, the risk of developing many diseases increases and this is true for stroke. Age is the single most important risk factor for stroke. It has been found that the risk of stroke doubles for every successive decade after the age of 55 in both men and women (Sacco *et al.*, 1997). The brain changes with age. Studies using post-mortem brain tissue have suggested that between the ages of 20 and 60 years, brain weight decreases by approximately 0.1% each year, with the hippocampus and cerebral cortex being the most affected regions. Similar studies have also found brain volume to decrease between the ages of 30 and 50 by 0.1-0.2% per year and this decrease in brain volume increases to 0.3-0.5% each year in the over 70s. In addition to brain weight and brain volume changes with increasing age, there are also certain changes which occur in the brain which may also increase the risk of stroke. There are a number of changes which occur in the white matter that are thought to increase the susceptibility of axons to ischaemia (Chen *et al.*, 2010). Rodent studies have been important in this area; most notably a study by Scavone *et al.* (2005), which found that white matter in ageing rats was more vulnerable to ischaemia, due to a reduced ability of ageing axons to maintain membrane properties. This compromised ability was found to be caused by reduced Na+-K+-ATPase performance found in the brains of ageing rats (Scavone *et al.*, 2005). Another study suggesting ischaemia is more likely in aged brains, showed that in old mice there is a two-fold increase in the glial glutamate transporter, highlighting an increase in glutamatergic signalling (Baltan *et al.*, 2008). A consequence of high glutamate levels is an influx of calcium ions into cells leading to cell death (Brustovetsky *et al.*, 2009).

It has been suggested that genetic predisposition may also contribute to some cases of stroke. The heritability estimate for ischaemic stroke has been reported to be 37.9%, however there are variations between stroke subtypes (Bevan *et al.*, 2012). Recent studies have identified several genes which may predispose an individual to stroke, including the area of Chromosome 9p21 surrounding the genes *CDKN2A* and *CDKN2B* (Gschwendtner *et al.*, 2009) and genetic variants in *PITX2* and *ZFHX3* on Chromosome 4q25 (Gudbjartsson *et al.*, 2007; Gretarsdottir *et al.*, 2008). Other inherited causes of stroke include disorders of imprinted genomic loci such as Prader-Willi syndrome and the X-linked disorder Fabry disease (Qureshi and Mehler, 2010). It is also thought that geographic location and ethnicity can increase the risk of a stroke although these may be due to underlying genetic differences.

Another non-modifiable risk factor for a stroke is a previous stroke or TIA. A TIA often acts as a warning sign that the patient is at a high risk of stroke. As would be expected, the risk of recurrent stroke is much higher in the first week after TIA and reduces as time goes on. Within the first month after minor stroke or TIA, the risk of a recurrent stroke is as high as 30% in some subgroups (Donnan *et al.*, 2008), this then reduces to 17% after 90 days (Furie *et al.*, 2011).

Hypertension is the single most important modifiable risk factor for stroke. Hypertension is thought to play a key role in 54% of strokes worldwide. Hypertension is defined as a systolic blood pressure ≥140mmHg or diastolic blood pressure ≥90 mmHg. Increasing levels of hypertension are found throughout the world (Chobanian, 2009), but increases have been more notable in India and China, possibly as a result of the recent demographic change (Mukherjee and Patil, 2011). This seems to have been translated into an increase in stroke incidence by over 100% in low-middle income countries reported by a systematic review of publications reporting stroke incidence between 1970 and 2008. This is likely due to increasing lifespan, westernised diets and decreasing levels of physical inactivity in developing countries, as well as more effective healthcare and preventative treatments in high income countries, which have seen a 42% decrease in stroke incidence between 1970 and 2008 (Feigin *et al.*, 2009).

Atrial fibrillation is the most common cause of cardiac embolism and is most prevalent among the elderly (Iwasaki *et al.*, 2011). It has been estimated that atrial fibrillation is present in almost half of all cardioembolic strokes (Furie *et al.*, 2011). It therefore carries a high relative risk of stroke (Warlow *et al.*, 2003). An early study into atrial fibrillation as a risk factor for stroke found that in subjects aged 50-59, 1.5% of strokes were attributable to atrial fibrillation compared to 23.5% in subjects aged 80-89. This shows that atrial fibrillation is much more common in the elderly with almost a quarter of strokes in the over 80s being attributable to atrial fibrillation (Furie *et al.*, 2011).

Other risk factors include those related to metabolism such as diabetes and hypercholesterolaemia. 15-33% of stroke patients in the USA have diabetes compared to just 8% of the overall adult population highlighting that diabetes is an important risk factor for stroke. This may be because of the increases in incidence of atherosclerosis, hypertension and obesity. Hypercholesterolaemia can also increase the risks of atherosclerosis and heart disease resulting in subsequent increases in risk for stroke (Furie *et al.*, 2011).

Other potential risk factors such as homocysteine levels are less well studied but recent studies have suggested there may be a causal relationship (Casas *et al.*, 2005). However, data from several randomised controlled trials have shown no effect (Toole *et al.*, 2004; Lee *et al.*, 2010). Other lifestyle choices which result in poor health are also risk factors for stroke including smoking, alcohol consumption, obesity and a sedentary lifestyle (Furie *et al*., 2011). In Table 1, geographic location is listed as a non-modifiable risk factor, however the term ‘geography’ encompasses a number of risk factors. Different geographic locations are likely to have different genetic profiles and possibly even different ethnicities which may affect the risk of stroke. Geography may also reflect socio-economic status (SES) which encompasses factors such as smoking, education and profession, some of which are modifiable risk factors of stroke. Using geography to assess risk of stroke should be used with caution and should assess how (whether by genetics, SES or some other factor) geography has an effect on the risk of stroke.

Many of the established risk factors described above have been studied further to try and understand how they increase the risk of stroke. As mentioned above, many are associated with atherosclerosis or the stiffening of the arteries as well as the narrowing and thickening of arterioles and capillaries. In the brain, these changes in the vascular structure translate into a reduction in the cerebral blood flow meaning the brain is not adequately perfused. Risk factors which contribute to this failure of cerebral blood flow regulation include ageing, diabetes, hypertension and hypercholesterolaemia (Moskowitz *et al.*, 2010).

1.2.2.4 ***Molecular mechanisms***

Due to the wealth of information regarding risk factors it is thought that multiple systems contribute to stroke aetiology. Research into molecular mechanisms is now hoped to identify biomarkers of disease and also shed light into the molecular mechanisms involved.

Expression studies including a recent publication by Zhang *et al.* (2014) have highlighted genes which are differentially expressed in ischaemic stroke cases compared to controls. Zhang *et al.* (2014) used a gene expression array to compare the expression profiles in peripheral blood mononuclear cells of 20 ischaemic stroke cases with 20 controls and analysis revealed 37 differentially expressed genes. By producing an interaction network, they highlighted two genes – tumour necrosis factor (*TNF*) and interleukin 8 (*IL-8*) as having the highest level of interaction, suggesting they may have a critical role in ischaemic stroke. These results are very interesting as both genes are known to be involved in the inflammation process, a mechanism widely implicated in the aetiology of ischaemic stroke (Zhang *et al.*, 2014b).

Proteomic studies have also been fruitful in the search for biomarkers. Over a decade ago, C-reactive protein (CRP), an indicator of inflammation, was highlighted as a potential biomarker in the prediction of future stroke. Individuals in the highest quartile for CRP had between a 2-2.7 risk of ischaemic stroke, compared to those in the lowest quartile for CRP (Rost *et al.*, 2001). More recently, using mass-spectrometry techniques Garcia-Berrocoso *et al.* (2013) highlighted 51 proteins that were altered in the area of the brain with an infarct. Five of these were analysed in 60 peripheral blood samples taken at stroke admission. Three of these proteins were also found to be differentially translated in the blood of stroke patients early after onset. Increased levels of gelsolin, cystatin A and reduced levels of dihydropyrimidinase-related protein 2 were associated with a poor prognosis. This study highlighted three proteins which may be potential biomarkers for outcomes after stroke (Garcia-Berrocoso *et al.*, 2013).

Research into biomarkers for stroke is still a long way behind diseases such as diabetes and immunological disease, with limited studies particularly in the field of metabolomics, but recent research has been promising (Kim *et al.*, 2013).

1.2.2.5 ***Stroke prevention and treatment***

As a TIA often acts as a warning sign to patients that they are at high risk of suffering a stroke, this can encourage people to take action and change their lifestyles in order to prevent a stroke. Secondary preventative strategies include the ceasing of hormone replacement therapy in postmenopausal women (Warlow *et al.*, 2003) and carotid endarterectomy which is effective in stroke patients with stenosis (narrowing of the symptomatic carotid artery) of at least 70% if carried out within twelve weeks of TIA or first stroke (Donnan *et al.*, 2008). Other ways to prevent stroke either following a TIA/first stroke or if at high risk of a first stroke include lowering blood pressure, treatment with antiplatelet or anticoagulant drugs, reducing blood cholesterol levels, all of which are described in Section 1.2.3.6.

It is vital that stroke patients receive treatment early after stroke symptoms appear. In addition to aspirin being an effective preventative treatment at daily doses of 75-150mg, if doses of 160-300mg are administered within 48 hours of stroke it can reduce the risk of early recurrent stroke and mortality within 14 days and increases the chance of disability-free survival (Warlow *et al.*, 2003). Treatment with aspirin is attractive due to its low cost, ease of administration and low risk of side effects. However, the benefits of taking aspirin as a treatment are quite small with less than 1% of patients saved from death or disability (Donnan *et al.*, 2008).

A more effective method to reduce the risk of disability after stroke is early clot lysis with intravenous recombinant tissue-plasminogen activator (Chen *et al.*, 2010). However, few patients are suitable to receive thrombolysis treatment due to the short therapeutic time window of just three hours post-stroke and the risk of symptomatic intracerebral haemorrhage experienced by 6-7% of treated patients (Donnan *et al.*, 2008). It is thought that this method of treatment is effective across all age groups however; there has been a persistent underrepresentation of the oldest old in large-scale trials (Chen *et al.*, 2010).

Other potentially effective stroke treatments include intra-arterial fibrinolysis, fibrinogen-depleting agents and inhibitors of glycoprotein IIb/IIIa. Many drugs are being investigated for use in neuroprotection, however very few drugs make it through to clinical trials (Warlow *et al.*, 2003; Chen *et al.*, 2010). Due to the effectiveness of therapeutic hypothermia in cardiac arrest and neonatal ischaemia patients, investigations are underway into the translation of this technique for use in adult stroke victims (Moskowitz *et al.*, 2010).

Arguably the most successful advancement in recent decades is the introduction of stroke care units (SCUs) into routine practice. SCUs are suitable for all stroke patients regardless of age or subtype and involve specialised assessments by a multidisciplinary team. Patients admitted to SCUs have a 20% reduced risk of mortality and a 20% reduced risk of long-term disability (Donnan *et al.*, 2008). Due to the effectiveness and widespread availability and accessibility of stroke care units, they are much more favourable than thrombolysis in the vast majority of cases (Warlow *et al.*, 2003; Chen *et al.*, 2010). It is unclear what makes SCUs so successful in reducing risk of recurrence, as well as mortality and dependency. It is most probably the combination of rehabilitating techniques including early mobilization and effective blood pressure monitoring, as well a general adherence to best practice (Donnan *et al.*, 2008; Chen *et al.*, 2010). Due to their effectiveness, the widespread introduction of SCUs across developing countries is recognised to be a priority to reduce stroke mortality (Warlow *et al.*, 2003).

1.2.2.6 ***Consequences of stroke***

Accountable for approximately 9.7% of all deaths worldwide (Mukherjee and Patil, 2011), stroke is the third leading cause of death around the world and one of the leading causes of adult neurological disability. Though treatment and preventative measures do exist for stroke, about 30% of patients die within a year of a stroke and of those that survive, almost half of them are left dependent on others. Many of those who are able to live independently, however, are not free of residual physical or cognitive deficits, so some are unable to continue with their normal daily family and working life. One consequence of stroke is the increased risk of late-onset epilepsy (Bevan and Markus, 2011). There is approximately an 11.5% risk of at least one seizure in the five years following a stroke; however this may depend on the severity and location of the stroke (Burn *et al.*, 1997). As well as age being a major risk factor for stroke, the elderly also recover much slower than younger stroke sufferers, meaning that the consequences of the stroke are often more severe in the elderly patients. This is partly due to other co-morbidities increasing disability and needs (Chen *et al.*, 2010). The main cause of dependency after a stroke is dementia.

1.2.3 Post-stroke dementia

1.2.3.1 ***Definition***

Post-stroke dementia (PSD) refers to any type of dementia present after a stroke, irrespective of its cause (Leys *et al.*, 2005). Although vascular dementia (VaD) is a direct result of vascular changes such as cerebral infarcts and haemorrhages, not all patients who suffer from a stroke have VaD. PSD is most commonly diagnosed as VaD, Alzheimer’s disease (AD) or mixed dementia, as it may be the result of vascular lesions, Alzheimer pathology, white matter changes or a mixture of these (Leys *et al.*, 2005).

1.2.3.2 ***Prevalence of post-stroke dementia***

The prevalence of PSD is difficult to calculate accurately, as results vary depending on which classification system is used. A study by Pohjasvaara *et al.* (1998) utilised the Diagnostic and Statistical Manual of Mental Disorders, Third Edition (DSM-III) definition of dementia, in addition to a comprehensive neuropsychological examination, to determine the prevalence of PSD in a cohort of 337 stroke survivors. It was reported to be 31.8%. The authors observed that PSD patients were older, more frequent current smokers with a lower level of education and a previous history of cerebrovascular disease and stroke (Pohjasvaara *et al.*, 1998). Similarly, Inzitari *et al.* (1998) adopted the International Classification of Diseases, 10th Revision (ICD-10) criteria to determine prevalence of PSD in a cohort of 339 stroke survivors. The ICD-10 has a lower sensitivity to dementia than the DSM-III and a prevalence of 16.8% was reported. They also found that those with PSD were older and more likely to suffer from atrial fibrillation (Inzitari *et al.*, 1998). The prevalence of PSD is likely to increase over the next few decades due to the ageing demographic of the population and decreasing mortality following a stroke (Leys *et al.*, 2005). Though studies are not in agreement about the prevalence of dementia they do agree that stroke doubles the risk of dementia. Many stroke survivors who are not classed as having PSD do however have some degree of cognitive impairment. Previous studies using the Cognitive Function After Stroke (COGFAST) cohort (Section 2.1.1), which recruited cognitively normal stroke survivors three months post-stroke, found that 9.4% of participants developed dementia within between three months and fifteen months post-stroke, this increased to 21.4% at three years post-stroke and after seven years, the figure had reached 39.5% (Allan *et al.*, 2011). These figures highlight that although some individuals regain cognitive function three months post-stroke, the risk of developing PSD remain high and increases with age.

1.2.3.3 ***Risk factors***

Over and above the risk factors for stroke itself described in detail in previous sections, age remains a major risk factor for PSD. It is a disorder that most commonly affects the elderly, those with a low level of education and those who were dependent (unable to live independently) or cognitively impaired before the stroke (Leys *et al.*, 2005). The Perindopril Protection Against Recurrent Stroke Study (PROGRESS) has identified high blood pressure as a possible risk factor for dementia. In this trial 6105 stroke patients were randomly assigned to either the active treatment (perindopril and indapamide) which reduces blood pressure or a placebo group. It was found that those taking the treatment were at a lower risk of developing dementia. 6.3% of the treatment group developed dementia compared with 7.1% in the placebo group. The risk of cognitive decline was also reduced by the treatment; 9.1% experienced cognitive decline following a stroke in the treatment group compared with 11.0% in the placebo group (Tzourio *et al.*, 2003).

It is now accepted that well-known vascular risk factors including diabetes, atrial fibrillation, smoking and depression not only increase the risk of cerebrovascular disease but also increase the likelihood of dementia (Leys *et al.*, 2005). Vascular risk factors have been associated with changes in brain volume. Studies comparing brain volume and cognition in hypertensive patients have identified that those with higher blood pressure have significantly smaller whole brain, hippocampal and thalamic volumes, more white matter hyperintensities and perform less well in memory and language tests than participants with normal blood pressure (Cumming and Brodtmann, 2010).

Neuroimaging following a stroke may also indicate which patients are at risk of PSD. The presence of silent infarcts, global cerebral and medial-temporal-lobe atrophy and white matter changes are all suggested risk factors. In addition to these factors, the location, cause and indeed the severity of the presenting stroke may affect the risk of PSD. A previous stroke and stroke recurrence are also likely to increase the risk of PSD (Leys *et al.*, 2005).

There can also be genetic contributors to post-stroke dementia. Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL) is the most common form of hereditary stroke disorder which leads to the degeneration of smooth muscle cells in blood vessels and affects approximately 50 families in the UK. It occurs as a result of at least 60 mutations in the *Notch3* gene and patients progress to dementia (Kalaria *et al.*, 2004).

The apolipoprotein E4 (APOE4) allele is a well-established genetic risk factor for AD but may also be considered a risk factor for PSD. Presence of APOE4 increases the risk of vascular disease through its effects on cholesterol transport and metabolism which accelerate atherogenesis. It is now thought that the APOE risk genotype is related to the progression of ischaemic white matter lesion load in addition to post-stroke cognitive decline (Cumming and Brodtmann, 2010).

1.2.3.4 *Consequences of post-stroke dementia*

Several studies have shown that patients with PSD have an increased mortality compared with stroke patients without dementia. This is true for all types of dementia which occur independently of stroke. It has been reported that only 39% of stroke patients with dementia will survive for five years post-stroke compared with 75% of cognitively normal stroke patients (Kalaria and Ballard, 2001). A reason for this may be that dementia patients may be less compliant with treatments to prevent new vascular events. Dementia may also worsen co-morbidities thus increasing the risk of mortality

(Leys *et al.*, 2005). It has also been suggested that dementia could increase the risk of stroke recurrence by three times (Moroney *et al.*, 1997).

The brain is the major tissue affected by both stroke and dementia. As previously mentioned, the brain is the site of damage after both an ischaemic and haemorrhagic stroke. The brain also changes following the onset of dementia, one of the most common and obvious pathologies is brain atrophy which can be detected by magnetic resonance imaging. Studies into the pathophysiology of AD have revealed that the hippocampus and cortex are the most affected areas in this slow and progressive disease. It is characterised by senile plaques (extracellular aggregates of β-amyloid (Aβ) proteins) and neurofibrillary tangles (intracellular hyperphosphorylated Tau proteins) (Iraola-Guzman *et al.*, 2011). Dementia with Lewy bodies (DLB) is characterised by the accumulation of Lewy bodies which are neuronal inclusions of alpha-synuclein that arise from neurofilaments and paired helical filament epitopes. A patient often presents with DLB before or with Parkinsonism. Whilst AD and DLB do have some distinctive characteristics, there is a considerable amount of overlap, making a definitive diagnosis difficult. One way to distinguish between AD and DLB is to perform a dopamine transporter scan (DaTscan) to look at the level of dopamine transporter uptake in the basal ganglia. A DLB patient would have a much lower uptake rate than an AD patient (McKeith, 2002; Cummings *et al.*, 2011). As the name suggests, frontotemporal dementia (FTD) is characterised by the neurodegeneration and atrophy of the frontal and/or temporal lobe. Protein inclusions may also present in FTD. These can either be Tau, the microtubule-associated protein found in AD or TAR DNA-binding protein 43 (Cardarelli *et al.*, 2010). Vascular dementia is a term which has caused much debate among clinicians as approximately half of patients with vascular cognitive impairment have dementia. It is characterised by a rapid onset and stepwise progression. Neuroimaging detects infarcts, haemorrhages and/or white matter disease (Cumming and Brodtmann, 2011).

1.2.3.5 ***Mechanisms***

For many years, it has been clear that there is a strong association between stroke and dementia, however, there is no clear evidence of a direct pathway linking the two. It is unknown whether PSD is due to vascular disease directly causing neuropathological changes associated with dementia or whether there is any synergistic action between neuropathology and vascular disease to worsen cognition.

To try and identify mechanisms involved, studies have begun to look into how the brain is affected pathophysiologically by PSD. Sachdev *et al.* (2009) carried out a study on 104 post-stroke patients, 45 of whom had mild cognitive impairment and 59 who were cognitively intact. They found that those with some cognitive impairment had more white matter hyperintensities but did not differ in the number or volume of infarctions, or brain and hippocampal volume. Most of the cognitive deficits were frontal and were most likely a result of the white and grey matter atrophy caused by ischaemia (Sachdev *et al.*, 2009). A similar study by Stebbins *et al.* (2008) which assessed grey matter atrophy found that in stroke patients with cognitive impairment, there was a greater reduction in grey matter volume compared with cognitively normal stroke survivors, predominantly in the thalamus (Stebbins *et al.*, 2008). The thalamus is regarded as an important brain region for long-term memory attention and executive function (Cumming and Brodtmann, 2010).

Previous studies using the Cognitive Function after Stroke (COGFAST) study, which is comprised of a cohort of >75 year old stroke survivors who are cognitively functional three months post-stroke, have found that even those who are cognitively normal following a stroke still often experience deficits in attention and executive function. These deficits then worsen in those patients who develop dementia and a number of other brain function impairments including orientation and memory arise (Stephens *et al.*, 2004). However, improvement in cognition can occur (Ballard *et al.*, 2003). A recent study looked at hippocampal neuronal density and volume using COGFAST participants. Whilst hippocampal neuronal density did not differ between demented and non-demented post-stroke patients, differences were observed in the volumes of hippocampal cornu ammonis 1 (CA1) and CA2 neurones in PSD patients when compared with cognitively intact stroke survivors. Neuronal volume was positively correlated with cognitive function (Gemmell *et al.,* 2012).

Although vascular risk factors have been associated with both stroke and dementia risk, they cannot fully explain the relationship between stroke and subsequent cognitive decline. Studies have still found associations between white matter hyperintensities and increased risk of cognitive decline, independent of vascular risk factors (Cumming and Brodtmann, 2010).

There are a number of suggested mechanistic links. One possible link is that ischaemia may be a catalyst for amyloid deposition. Another suggestion is that there may be a leakage of serum components through the blood-brain barrier as a result of the stroke that could cause neuronal damage. Or the disruption is cerebral blood supply may not only lead to a stroke but also to secondary neuronal degeneration (Cumming and Brodtmann, 2010).

Due to this uncertainty there is an urgent need to elucidate the mechanistic link between stroke and dementia and develop better treatment and preventative measures.

1.2.3.6 ***Prediction, prevention and treatment***

As there is currently no effective cure for dementia, research has aimed to find ways to prevent dementia onset. As there are many parallels between stroke risk factors and dementia risk factors, many of the ways to prevent stroke mentioned in 1.2.2.5 are also adopted in the prevention of post stroke dementia and are considered below.

As hypertension is one of the main risk factors for stroke, finding strategies to reduce the risk has been a priority in the Western world for many years. Since the 1950s strategies to lower blood pressure have been used to reduce the risk of a first stroke. Lowering blood pressure has also been shown to be effective in reducing the risk of a secondary stroke (Donnan *et al.*, 2008). Meta-analyses of randomised controlled trials found lowering blood pressure accounted for 30-40% reductions in secondary stroke risk across all age groups (Furie *et al.*, 2011).

Aspirin is a widely prescribed antiplatelet drug which reduces the risk of all vascular events, including primary and secondary stroke, by 22% (Donnan *et al.*, 2008). It is recommended to be taken by healthy women over the age of 65 to reduce risk of ischaemic stroke as well in the very old who have atherosclerosis or other conditions placing them at a high risk of cardiovascular event (Chen *et al.*, 2010).

In patients with atrial fibrillation, warfarin reduces the relative risk of a recurrent stroke by 68% (Chen *et al.*, 2010). However, there are serious risks in taking warfarin – major bleeding including intracerebral haemorrhage may occur (risk of 0.3-0.6% per year). The risks of major bleeding increase with age, hypertension and more severe anticoagulation. Warfarin should not be taken in combination with aspirin as this may increase the risk of suffering a haemorrhage and as warfarin is the most biologically effective secondary preventative treatment (Donnan *et al.*, 2008), warfarin should be prescribed over aspirin (Warlow *et al.*, 2003).

Although there is no strong evidence that high blood cholesterol levels increase the risk of stroke, treatment with statins to reduce the concentration of cholesterol in plasma have proven effective in reducing stroke risk in both patients with or without coronary disease. Meta-analyses have found a 22% reduction in relative risk in patients younger than 65 years of age and a 19% reduction in those over 65 years of age (Chen *et al.*, 2010).

Due to the mechanisms resulting in post-stroke dementia being currently unknown, the ability to predict stroke survivors who are likely to develop dementia would be extremely valuable in the targeting of preventative treatments. A biomarker identifying these patients at high risk of developing dementia may help to identify mechanisms involved and lead to more suitable treatments as well as helping to target which patients and their families require more support following the stroke.

Finding methods to treat PSD is difficult as not only does the clinician have to treat the stroke but they also have to treat the dementia. AD and VaD patients have shown improvements when treated with cholinesterase inhibitors so these provide potential treatment options for PSD however they have not been trialled in PSD cases (Leys *et al.*, 2005). Similar to the treatment of strokes, rehabilitation such as physiotherapy and speech therapy are currently used in the treatment of PSD.

1.2.4 Parkinson’s disease

1.2.4.1 *Prevalence*

Idiopathic Parkinson’s disease (PD) is the second most common neurodegenerative disorder affecting approximately 1% of the over-65 population (Saracchi *et al.*, 2014).

The mean age of onset is 62 years of age, although young-onset PD can affect individuals under 45 although these only account for a small proportion (<10%) of PD cases. The disorder is more common in males than females. Parkinson's disease is characterised by a series of motor symptoms although symptoms can differ between individuals. Clinical criteria state that bradykinesia plus at least one other motor symptom (tremor, rigidity, postural instability) should be present for diagnosis (Hughes *et al.*, 1992). A number of non-motor symptoms such as cognitive impairment, gastrointestinal disturbance and disturbed sleep are also found in Parkinson's disease which often can precede motor symptoms, making the diagnosis of early stage PD difficult. The diagnosis of PD, as already described, is currently clinically derived and can only be confirmed by neuropathological examination post-mortem (Saracchi *et al.*, 2014).

1.2.4.2 *Risk factors*

There are several known risk factors for PD, both genetic and environmental. As PD is an age-related disease, age is the greatest risk factor for PD. The prevalence increases from 1% in a population of 60 year olds to 3% in those over 80. Less than 10% of PD cases are caused by genetic factors. There are around 28 chromosomal locations that have been implicated with PD. The most studied genes related to PD include parkin, α-synuclein (*SNCA*) and *LRRK2*, which is the commonest genetic abnormality in Western populations (Gilks *et al.*, 2005). In addition to family history, PD is also more common in males and those who work in the agriculture sector or live in rural areas, a link that has been tentatively linked to exposure to pesticides (Gillies *et al.*, 2014). It has been suggested that several factors are protective against PD including tobacco smoking and caffeine intake (Schapira and Jenner, 2011). Other factors associated with an increased risk of PD include head trauma (Marques *et al.*, 2011).

1.2.4.3 *Molecular mechanisms*

As diagnosis is based on clinical symptoms, the earliest of which are often non-motor making the early diagnosis of PD difficult, research into the molecular mechanisms of disease and identification of biomarkers to aid diagnosis and prediction of disease is important. There have been a number of "omics" techniques employed to try to achieve this.

Gene expression profiles have been performed in numerous PD cohorts although many of the findings are inconsistent. A seemingly promising study highlighted 22 genes with differential expression in PD cases. The most interesting gene highlighted was that of *ST13*, the expression of which was significantly reduced in PD cases; this was an interesting finding since *ST13* is involved in the misfolding of SNCA, which may be of particular relevance to PD (Scherzer *et al.*, 2007). However, a study targeting *ST13* using real-time polymerase chain reaction (PCR) in a cohort of early-stage PD cases did not support these findings and it concluded that *ST13* was not a suitable biomarker for early PD (Shadrina *et al.*, 2010).

Protein studies have been performed to quantify the protein content of a biological sample. Using this approach a number of proteins including APOE and IL-8 have been identified as potential biomarkers for diagnosis and disease progression with altered levels found in PD patients. The choice of tissue for biomarker identification, cerebrospinal fluid (CSF) or blood, has been the focus of some debate. The proteome of the substantia nigra has also recently been characterised (Saracchi *et al.*, 2014).

Compared to stroke, many more metabolomics studies have been performed in PD research. A recent study that compared the metabolomics profiles of PD cases and controls identified different levels of uric acid and glutathione, both of which are antioxidants, in the plasma of PD cases compared to healthy controls (Bogdanov *et al.*, 2008). A study has also highlighted increased pyruvate levels in PD cases again using plasma (Ahmed *et al.*, 2009). Metabolomic studies have therefore proved fruitful in the study of biomarkers for PD however, replication is required using an alternative platform and larger cohorts.

In conclusion, further research is required in this field to identify biomarkers which are able to detect at-risk individuals to offer the best help possible whether that be disease-modifying therapeutic strategies or just psychological support (Saracchi *et al.*, 2014).

1.2.4.4 ***Prevention and treatment of Parkinson's disease***

Due to the key risk factors for PD being linked to demographics it is a disease with little scope for onset prevention. Research into the molecular mechanisms that cause PD will hopefully shed some light on possible prediction and prevention targets in coming years.

There are a number of different treatments for PD. A range of pharmacotherapies are available although all focus on the symptomatic features of disease, rather than preventing disease progression. In PD, levels of dopamine, a neurotransmitter usually found in abundance in the central nervous system, are significantly reduced compared to controls. Levodopa, which has the ability to cross the blood brain barrier and is converted into dopamine, is currently the gold standard, however there are a number of side effects associated with long-term administration. The most significant limitation of Levodopa is that after each dose there is an attenuation effect where symptoms of PD are no longer under control. Dyskinesia (involuntary muscle movements) is another common complication associated with long-term Levodopa usage. To help combat the attenuation effect Levodopa can be administered with catechol-O-methyltransferase inhibitors, which acts by extending the time Levodopa is available in the brain. Other treatments currently used to reduce PD symptoms include dopamine agonists which act directly on dopamine receptors within the central nervous system and monoamine oxidase B (MAO-B) inhibitors, which block the degradation of dopamine (Tarazi *et al.*, 2014).

Since all the current treatments are associated with significant side-effects, new treatments currently in development are hoped to overcome some of the problems described above. The dyskinesias which may result from extended Levodopa use may be helped by treatment with glutamate receptor antagonists which are currently in development. Another type of treatment currently undergoing research are adenosine A_{2A} receptor antagonists; the early data on which suggests they may improve the mobility of PD patients. Due to the hypothesis that oxidative stress plays an important role in neurodegeneration, it has been suggested that antioxidants may have instrumental therapeutic effects. It is hoped that the removal of free radicals by antioxidants will prevent apoptosis and neuronal degeneration (Tarazi *et al.*, 2014).

In addition to pharmacotherapies, surgery provides a number of treatment options. Deep brain stimulation (DBS) is a possible option for patients whose motor symptoms persist and/or in those with refractory dyskinesias despite receiving medical treatment. A combination of DBS and medical treatments can greatly increase the quality of life of patients compared with pharmacotherapies alone. However, due to the invasive nature of the surgery there is the risk of a number of potentially serious complications including infection, intracranial haemorrhage and seizures. Problems with cognition and psychological symptoms such as anxiety and depression may also arise (Tarazi *et al.*, 2014).

1.2.4.5 ***Consequences of Parkinson's disease***

Following onset of disease, PD patients often experience a decrease in quality of life which can be attributed to increased risk of depression, falls, gait instability and cognitive impairment (Schrag *et al.*, 2000). Following PD onset, approximately 18.3% of patients are unable to live independently and require some level of care (Riedel *et al.*, 2012).

Cognitive impairment is a common complication of Parkinson's disease. Parkinson's disease dementia (PDD) affects almost 80% of PD patients when followed longitudinally in community studies (Hely *et al.*, 2008). In addition, mild cognitive impairment in Parkinson's disease (PD-MCI) is increasingly recognised. This definition has largely been borne out of the literature on Alzheimer's disease and is thought to represent a transitional state between normal cognition and dementia (Petersen, 2004). Although not everyone with PD-MCI will develop PDD, it is a recognised risk factor for the disorder. MCI is diagnosed in a PD patient who suffers from cognitive deficits not solely attributed to age but is still able to live independently (Litvan *et al.*, 2012). Cross-sectionally, approximately one in four PD patients without dementia are diagnosed as having PD-MCI (Aarsland and Kurz, 2010). A longitudinal study by Janvin *et al.,* (2006) found that after four years, 62% of PD-MCI patients had developed PDD compared with 20% of those who were cognitively normal prior to dementia diagnosis (Janvin *et al.,* 2006). This suggests that PD-MCI may be an early manifestation of PDD (Mufson *et al.*, 2012a). However, since not all PD-MCI cases develop dementia there are likely to be multiple mechanisms involved in the progression to dementia. Research into biomarkers would help elucidate mechanisms of cognitive impairment in PD. A number of risk factors have been described for PDD including older age, lower education level, a faster progression of motor symptoms, depression and sleep disorders. In addition, there have been a few genetic risk factors defined including carrying the *APOE* allele, the HI haplotype in the *MAPT* gene and mutations in the *SNCA* gene although some studies have reported inconsistent findings (Pagonabarraga and Kulisevsky, 2012).

There are several suggested mechanisms for the development of PDD. It is thought that most of these mechanisms may apply to PD-MCI although to a lesser degree and evidence is severely limited. Widespread cortical Lewy body deposition was one of the earliest recognised neuropathological correlates. Additional potential mechanisms include abnormal β-amyloid (Aβ) deposition, a result of abnormal amyloid precursor protein (APP) processing. PD-MCI and PDD patients have been found to have reduced levels of Aβ-42 (a marker of amyloid aggregation) in the CSF compared to cognitively normal cases. Aβ levels have also been associated with memory function in early PD patients (Yarnall *et al.*, 2013). However, there is very little evidence of any biomarkers that are able to predict PD-MCI or PDD in a cohort of PD-MCI cases highlighting that research in this area is still in its infancy. Much larger studies, especially involving post-mortem information, are required for further validation.

Prevention and treatment options for PD-MCI are very limited. Current treatments focus on the symptoms of PD. There have been no randomised control trials in PD-MCI to date (Yarnall *et al.*, 2013). Since the mechanisms resulting in PD-MCI are largely not understood, this makes finding suitable treatment options challenging. Research into possible biomarkers would be beneficial to help predict at-risk individuals, monitor disease progression and help find treatment targets.

1.3 Epigenetics

Each cell in a multicellular organism is genotypically identical. However, the various cell types are phenotypically different. These differences are due to differences in gene expression between the various cell types. Differences in gene expression are maintained by epigenetic control (Reik, 2007). Phenotypic differences are not only observed at the single-cell level. Monozygotic twins have the same genotype yet are phenotypically different, with variable differences in their appearances as well as disease discordance (Fraga *et al.*, 2005). The term epigenetics refers to the mechanisms which cause mitotically heritable but reversible changes to gene expression and phenotype, without modifying the DNA sequence.

Epigenetic mechanisms are important regulators of all biological processes throughout the life course from conception to death. Although epigenetic marks are established early in life, they are dynamic and receptive to internal and environmental stimuli, which may increase disease risk in later life (Delcuve *et al.*, 2009). Epigenetics is a very rapidly expanding field, with new discoveries aiding our understanding of the mechanisms and involvement in both development and disease. Histone modifications, microRNAs (miRNAs) and DNA methylation are all common types of epigenetic mechanisms and are described in further detail below.

1.3.1 Histones and chromatin

In eukaryotic organisms, genomic DNA is condensed as chromatin in the nucleus. Each unit of chromatin consists of DNA, at a length of 140-150 base pairs wrapped around the nucleosome, which is the core of chromatin, composed of an octamer of four different histones (H2A, H2B, H3 and H4). Another component of chromatin is histone H1 which keeps the DNA in place (Figure 2). All histones are composed of a globular C-terminal domain and are flanked by a highly variable N-terminal tail (Kim *et al.*, 2009). Histones, particularly the N-terminal tails of H3 and H4, can be modified by a number of post-translational mechanisms. Histone modifications act by changing the structure of chromatin into either euchromatin (“active”); in which the DNA is accessible for transcription or heterochromatin (“inactive”); where the DNA is inaccessible for transcription. Histones can be modified at different sites simultaneously and each histone can have several modifications instigating cross-talk between the different marks. There are currently many types of known modifications including phosphorylation, methylation and acetylation (Lorenzen *et al.*, 2012).

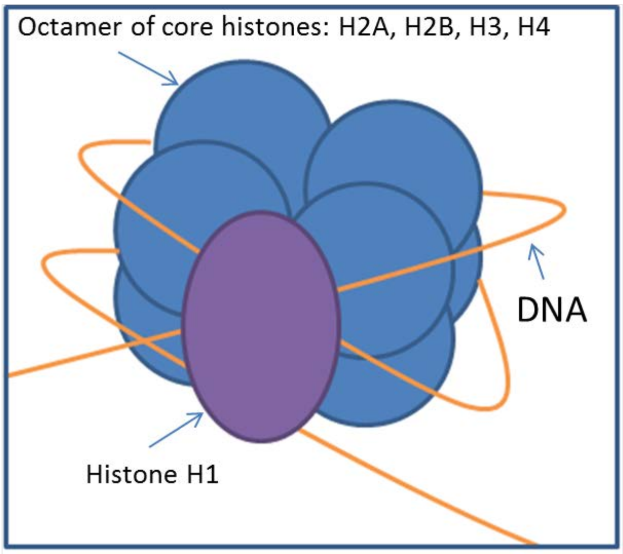


Figure 2: One unit of chromatin composed of DNA wrapped around an octamer of histones. Histone H1 keeps the structure in place.

1.3.2 MiRNAs

Recent high-throughput analysis of the transcriptome has revealed that the majority of DNA is transcribed as non-coding ribonucleic acid (RNA). A total of 90% of the genome is transcribed with only 1-2% transcribed as proteins – this shows that the vast majority of transcribed DNA is non-coding RNA. In recent years it has become apparent that they may play a role in epigenetic control. To date, over 800 different miRNAs have been identified in humans. miRNAs are small (22 nucleotides in length) non-coding RNAs which can bind to target mRNA transcripts that are complimentary in sequence. Many miRNAs have gene regulatory functions and are important in the control of mRNA stability. Due to having only been discovered in humans in 2001 (Sato *et al.*, 2011), they have been less widely studied than other epigenetic mechanisms and the majority of studies have focussed on cancer. However in recent years, it has been shown that they are involved in the post-transcriptional degradation of mRNA leading to the repression of translation and silencing of genes (Lorenzen *et al.*, 2012). In tumours, many miRNAs are aberrantly expressed suggesting that they may have a role in the development of cancer (Portela and Esteller, 2010; Kasinski and Slack, 2011).

1.3.3 DNA methylation

DNA methylation is by far the most commonly studied epigenetic mechanism. There are a number of reasons for this including it being found at high levels across the genome (Shenker and Flanagan, 2012), having key roles in a number of important developmental functions (Huh *et al.*, 2013) and, more importantly, being the most stable epigenetic modification (Talens *et al.*, 2010). DNA methylation involves the addition of a methyl group to the 5-carbon of the pyrimidine ring of a cytosine residue to form 5-methylcytosine (5-mC) (Figure 3A). In mammalian genomes, this cytosine residue is usually 5' adjacent to a guanine residue, forming a CpG site (Figure 3B) (Pearce *et al.*, 2012).

1.3.3.1 ***Types of DNA modification***

Whilst most DNA methylation occurs within CpG sites and most commonly forms 5-mC, there are some exceptions. Cytosine residues can also be converted to 5-hydroxy-methylcytosine (5-hmC), formed by the addition of a hydroxyl group to 5-mC by the ten-eleven translocation (TET) enzymes (Figure 3A) (Booth *et al.*, 2012). 5-hmC has been implicated in embryonic stem cell differentiation. Recently, non-CpG methylation has been described. This refers to methylation occurring at cytosine residues not within CpG sites, known as non-CpG (CpH) sites. These occurrences are rare, found only in embryonic stem cells and are thought to be involved in the maintenance of the pluripotent state (Portela and Esteller, 2010).

1.3.3.2 ***Genomic content***

CpG dinucleotides are found within the mammalian genome at a much lower frequency than expected by random assortment, accounting for 1% of the genome (Kim *et al.*, 2009). This CpG site deficiency is thought to be due to the high mutation rate or deamination of methylated cytosine residues to thymine (Duret and Galtier, 2000). Furthermore, CpG sites are not distributed evenly throughout the genome but instead commonly found in dense clusters referred to as CpG islands (Kim *et al.*, 2009). CpG islands are usually located within gene promoters (approximately 70%) and are most commonly unmethylated or display very low levels of methylation (Bell and Spector, 2011). CpG islands are defined as regions of 200-300 base pairs in length which have a CG content of approximately 50% (Lorenzen *et al.*, 2012). In addition to CpG islands, CpG shores, shelves and open seas have now been described. Shores are defined as regions up to 2kb from CpG islands, shelves are considered regions 2-4kb from CpG islands and open sea regions are isolated CpG sites with no specific designation (Figure 4). CpG islands are the most CpG dense region within the genome with decreasing numbers found in shores, shelves and open seas (Rechache *et al.*, 2012).

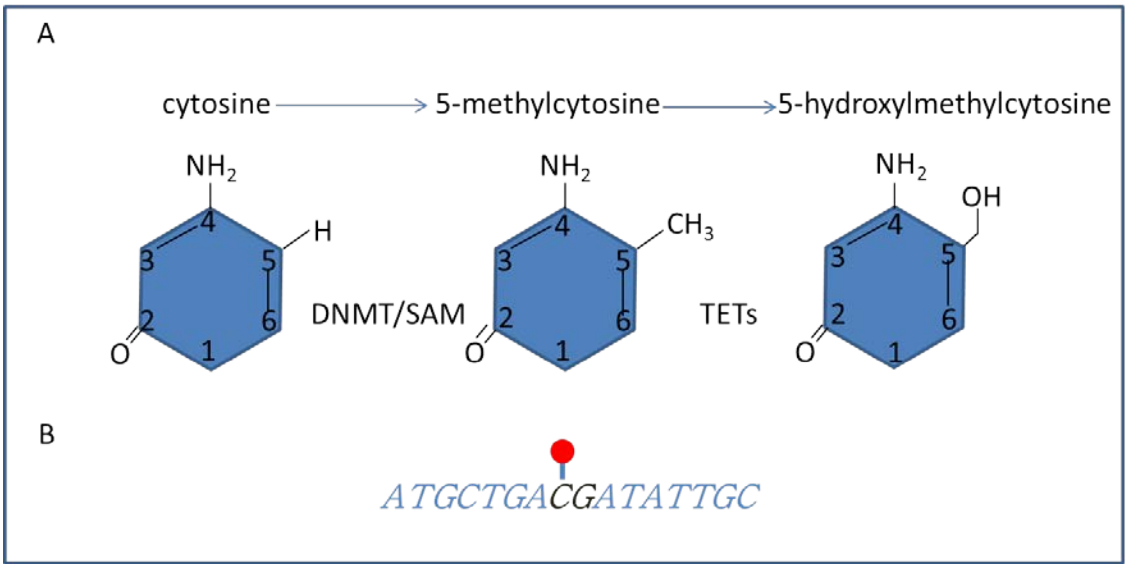


Figure 3: DNA methylation. A. Biochemical structure of cytosine and 5-methylcytosine. B. CpG site; in humans DNA methylation usually occurs at cytosine residues located 5’ adjacent to guanine residues forming the commonly denoted CpG site. Red lollipop highlights presence of methyl group.

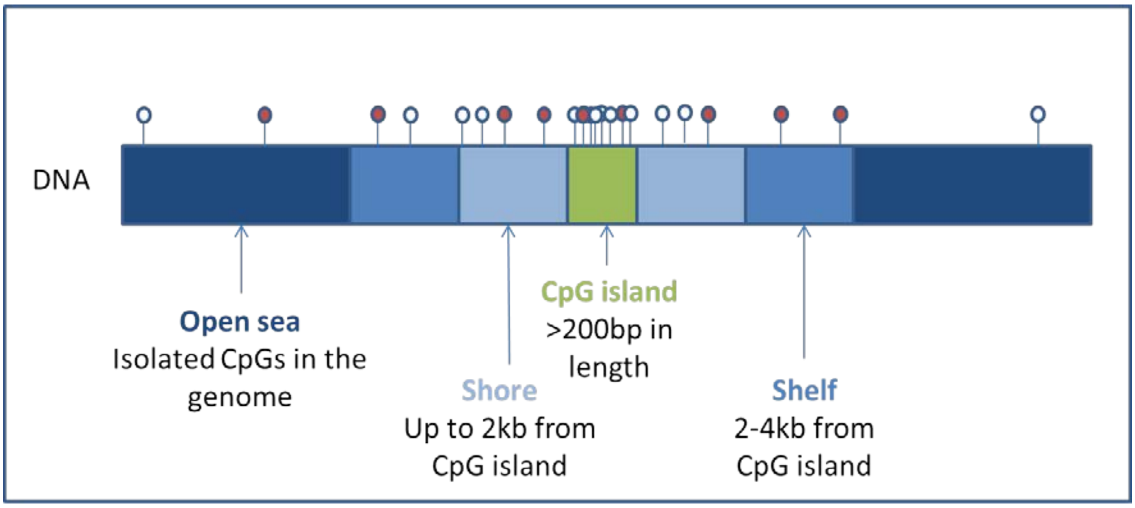


Figure 4: CpG islands, shores, shelves and open sea. Lollipops indicate CpG density within each region. Red lollipops show methylated CpG sites and white lollipops show unmethylated CpG sites.

For many years it was believed that the most functionally important DNA methylation occurred in promoters and that most DNA methylation changes occur in CpG islands. Promoter methylation has been associated with an inverse relationship with gene expression in the majority of cases, although there are several exceptions. Methylation has been found to be more strongly related to gene expression in the CpG island shores. This suggests that methylation in the CpG island shores may have more functional importance than the CpG islands (Irizarry *et al.*, 2009; Bell and Spector, 2011).

1.3.3.3 ***Key functions***

1.3.3.3.1 Transcription

DNA methylation has often been associated with an inhibition of transcription and gene silencing. There are two main mechanisms by which DNA methylation may inhibit transcription and subsequent gene expression. DNA methylation may repress transcription by impeding transcription factors binding to their target sites (Boyes and Bird, 1991). More recently, methyl-CpG binding proteins which bind to methylation marks and repress transcription by recruiting histone-modifying proteins have been discovered (Iraola-Guzman *et al.*, 2011; Lorenzen *et al.*, 2012). Whilst an inverse relationship between DNA methylation and gene expression has been reported in many cases, it is now thought that this relationship is not strictly true. This inverse relationship is still expected at promoter regions, however, the opposite has been found within gene bodies with increased gene body methylation being associated with increased transcription (Wagner *et al.,* 2014). DNA methylation has also been associated with the regulation of splicing without necessarily affecting the overall transcript levels of the given gene (Gelfman *et al.,* 2013*).*

1.3.3.3.2 X-inactivation

DNA methylation has a key role in X inactivation, the silencing of one X chromosome to ensure equal sex-linked gene dosage between males (XY) and females (XX). X inactivation genes are silenced by a non-coding RNA gene called X-inactive specific transcript (*Xist*) which coats the entire length of the X chromosome. *Xist* is only expressed on the inactive X chromosome. The *Xist* gene is hypermethylated and therefore silenced on the active X chromosome whilst the active *Xist* gene is hypomethylated on the silenced X chromosome. This suggests that DNA methylation may control the expression of *Xist* and therefore have an important role in X inactivation (Panning and Jaenisch, 1996).

1.3.3.3.3 Genomic imprinting

DNA methylation has an important role in genomic imprinting. Genomic imprinting occurs in only 1% of autosomal genes (Jirtle and Skinner, 2007) and involves one parental allele of the gene (from either parent) being hypermethylated affecting the expression of that allele resulting in monoallelic expression (Delcuve *et al.*, 2009). Imprinted genes show both parent-of-origin specific differential methylation and parent-of-origin specific differential expression. This means that genes are differentially methylated and expressed depending on which parent they originated from (Bartolomei, 2009). There are several developmental disorders that are associated with deregulation of genomic imprinting including Prader-Willi and Beckwith-Wiedermann syndrome (Jirtle and Skinner, 2007).

1.3.3.3.4 Genomic stability

DNA methylation is also important in preventing the translocation of repetitive and transposable elements thereby preventing insertional mutagenesis. Over 99% of transposable elements are heavily methylated and therefore transcriptionally silent highlighting that DNA methylation plays a key role in maintaining genomic stability (Delcuve *et al.*, 2009). Repetitive elements account for 45% of the human genome and it is thought that the loss of methylation in these repetitive elements may account for a large portion of the characteristic global hypomethylation that is seen in a number of cancers (Weisenberger *et al.*, 2005).

1.3.3.4 ***Regulation***

1.3.3.4.1 DNA methyltransferases

The enzymatic transfer of methyl groups to DNA is catalysed by DNA methyltransferases (DNMTs). DNMT3a and DNMT3b are important in establishing methylation profiles in early development but are expressed throughout the life course (Siegmund *et al.*, 2007) and are responsible for *de novo* methylation whereas DNMT1 is involved in the maintenance of methylation patterns during replication (Qureshi and Mehler, 2010). The DNMTs transfer a methyl group from the methyl donor SAM, to a cytosine in the DNA sequence to produce 5-mC (Figure 3) (Gravina and Vijg, 2010). DNMTs are vital in the establishment of methylation profiles in early development. Knockout experiments have identified DNMTs as fundamental proteins for a successful mammalian development. The homozygous deletion of *Dnmt1* or *Dnmt3b* has been found to result in embryonic lethality (Wu and Zhang, 2011). *DNMT* expression is abundant in the brain, implicating these enzymes in brain development and neurodegeneration (Endres *et al.*, 2000). Until recent years, DNA methylation patterns in post-mitotic tissue such as the brain were believed to be fixed. However, the recent discovery of DNA methyltransferases in brain tissue have pointed to the possibility that methylation patterns are labile and responsive to endogenous and exogenous stimuli (Miller and Sweatt, 2007).

1.3.3.4.2 Environmental factors

Due to the labile nature of epigenetics, environmental factors can cause epigenetic modifications which may contribute to the development of abnormal diseased phenotypes. Since epigenetic modifications present a plausible link between environmental exposures and effects on gene expression, research has focussed on identifying the effects of the environment on DNA methylation to help identify possible mechanisms in diseases with an established environmental component (Jaenisch and Bird, 2003; Jirtle and Skinner, 2007).

Research has identified many associations between environmental exposures and changes in DNA methylation. Some of these environmental exposures include nutritional (folate intake (Jaenisch and Bird, 2003)), chemical (tobacco smoking (Breitling *et al.*, 2011) and lead exposure (Bakulski *et al.*, 2012b)) and physical factors (age – see Sections 1.3.4.1)

Folate and other dietary supplements such as vitamins can influence disease susceptibility and have been shown to have marked effects in the incidence of various cancers such as of the colon and liver. This is due to the effect folate and other vitamins have on the availability of methyl groups and the subsequent effect on DNA methylation levels. Low folate levels have been associated with genomic instability and genomic hypomethylation (Jaenisch and Bird, 2003) as well as neural tube defects (Crider *et al.*, 2012). This hypomethylation which occurs as a result of low folate can also affect expression levels of oncogenes, increasing the susceptibility to colon and liver cancers (Jaenisch and Bird, 2003). Lead exposure has been linked with altered DNA methylation which may influence cognition. This is described in more detail in Section 1.3.5.2.2.

In recent years, smoking has been a very well-researched environmental exposure. Tobacco smoking has long been known to be harmful to health and it is possible that the effects of smoking are mediated through DNA methylation. Smoking has previously been associated with changes in global methylation and altered methylation of several cancer-related genes. Breitling *et al.* (2011) performed one of the first genome-wide methylation arrays analysing methylation levels in DNA from peripheral whole blood in smokers and non-smokers aged between 50 and 60 years. Of 27,000 CpG sites interrogated, only one CpG site was significantly associated with smoking and replicable in a larger, non-overlapping cohort. The *F2RL3* locus was significantly less methylated in smokers than non-smokers (Breitling *et al.*, 2011). This finding has been supported by other studies including the study by Shenker *et al.* (2013) which performed a HumanMethylation450 (HM450) array to identify loci differentially methylated in smoking. This study also identified other differentially methylated loci including aryl-hydrocarbon receptor repressor (*AHRR*) (Shenker *et al.*, 2013). Both of these genes have been implicated in disease; *AHRR* has been associated with breast and colon cancer whilst *F2RL3* is involved in inflammation pointing to a possible role in cardiovascular related diseases (Breitling *et al.*, 2011). Recent research has pointed to a possible mechanistic link between smoking and premature death through the altered methylation of *F2RL3*. This association between *F2RL3* is strongest in men (Zhang *et al.*, 2014a).

1.3.3.4.3 Genetic factors

In addition to being mitotically heritable (i.e. passed on during cell division) epigenetic marks have also been shown to be phenotypically heritable to some degree (i.e. regulated by genetic variants passed on from parent to offspring). There are a number of ways a single nucleotide polymorphism (SNP) can influence DNA methylation at specific CpG sites. These are shown in Figure 5.

Firstly (Figure 5A), if either the cytosine or guanine residue of the CpG site itself is a SNP this will have a direct impact on the methylation status of that site. For instance, if the alternative allele of either residue is inherited the CpG site will be disrupted and methylation will not occur. Secondly, (Figure 5B) SNPs or other genetic variants may influence the methylation status of local (cis) or distant (trans) CpG sites. Finally, (Figure 5C) the presence of a SNP at one CpG site may also alter co-methylation at neighbouring CpG sites. A combination of these various scenarios may also be observed.

The proportion of locus-specific methylation variance caused by genetic variation is known as DNA methylation heritability. Twin studies have been valuable in genetic heritability studies due to their shared genetic variants and almost identical environmental exposures (Bell and Spector, 2011). An example of DNA methylation heritability can be seen in the H19/IGF2 region whereby SNPs in the H19/IGF2 region have been associated with methylation status of both *IGF2* and *H19* indicating that this methylation change may be under genetic control (Zhang *et al.*, 2010).

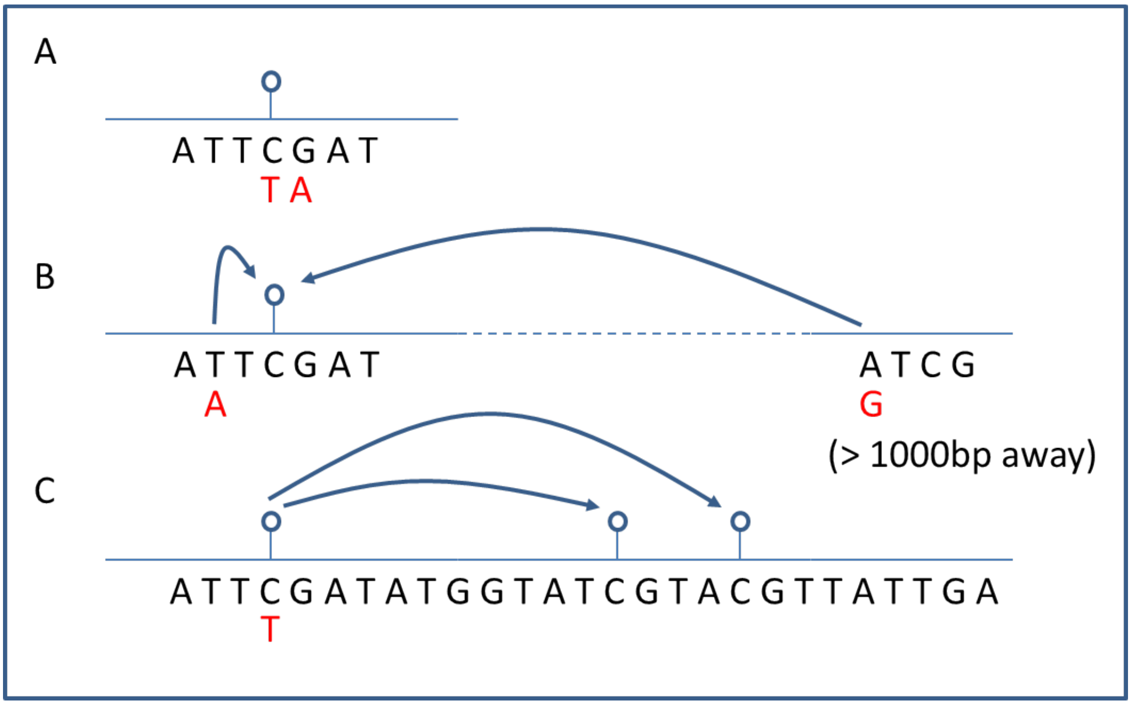


Figure 5: How can genetic control influence DNA methylation? A. CpG site is a SNP. B. SNPs may influence methylation status of cis or trans CpG sites. C. A SNP may result in co-methylation of neighbouring SNPs. Red bases indicate SNP.

1.3.4 Epigenetics across the life course

Epigenetic mechanisms are important regulators of all biological processes throughout the life course from conception to death. Although epigenetic marks are established early in life, they are dynamic and receptive to internal and environmental stimuli, which may increase disease risk in later life (Delcuve *et al.*, 2009). The Developmental Origins of Health and Disease hypothesis postulates that early life exposures can influence health in later life (Barker, 2006). Studies have found exposures *in utero* can have effects on DNA methylation later in life. The Dutch Hunger Winter of 1944-45 is a well-documented example of how diet *in utero* can affect methylation decades later. Individuals who were exposed to the famine prenatally had a much lower methylation at the imprinted *IGF2* locus compared with their same-sex unexposed siblings (Heijmans *et al.*, 2008). Studies like this indicate how exposures in early life can affect the epigenome which can have an effect on health on later life.

1.3.4.1 *Epigenetics in ageing and age-related disease*

Epigenetic mechanisms are known to have an important role in development (Reik *et al.*, 2001) but in recent years it has been suggested that they may also be a feature of the ageing process. Epigenetic patterns change throughout the ageing process. Twin studies have identified epigenetic drift occurring as individuals age with greater differences seen in twins in later life than in their early years (Fraga *et al.*, 2005). This could be due to the normal ageing process or it could be due to environmental exposures experienced throughout the life course or potentially as a result of disease (Bakulski and Fallin, 2014).

A healthy development relies on the successful regulation of DNA methylation. Many age-related diseases such as cancer, result from a fault in the epigenetic regulation of gene expression (Yang *et al.*, 2004). DNA methylation usually acts to silence gene expression. In general, DNA methylation globally (across the genome) decreases with age resulting in the expression of genes which should usually be silenced by methylation (although this straightforward relationship is far from clear in many instances). This may explain why many diseases are more commonly found among elderly populations. However, several genes are hypermethylated during ageing, an example of which is the oestrogen receptor (Kahn and Fraga, 2009) which has previously been associated with increased risk of atherosclerosis (Post *et al.*, 1999). However, it is difficult to distinguish between normal changes in methylation in line with ageing and changes causally linked with disease.

1.3.5 The role of DNA methylation in the ageing brain

1.3.5.1 *Normal development*

Genome-wide methylation analysis has been carried out in mammalian brains at multiple stages of life from early development to adult brains, the results of which suggest a key role for DNA methylation in the function and development of mammalian brains. Expression of *DNMT3A* was found to be critical for establishing a normal brain DNA methylation pattern and a subsequent normal brain development (Nguyen *et al.*, 2007). Another finding was that methylation at CpH sites was found to accumulate in neurons, but not glia, through early childhood to adolescence (Lister *et al.*, 2013). Furthermore, an inverse relationship between DNA methylation and gene function was observed. Genes known to be upregulated during development and associated with neuronal function were found to be associated with an increased promoter methylation in glia and a hypomethylation in neurons. The reverse was seen in genes downregulated during brain development, with decreased methylation levels in glia, but hypermethylation in neurons. These highlight the functional role of DNA methylation in brain development (Lister *et al.*, 2013).

Epigenetic research has also identified differences in DNA methylation between neuronal and non-neuronal nuclei obtained from human prefrontal cortex. The study by Iwamoto *et al.* (2011) showed that promoters of neuronal nuclei had lower global DNA methylation and also showed higher inter-individual variation when compared to non-neuronal nuclei. These findings suggest that neuronal cells have a greater ability to alter their epigenetic status in response to developmental and environmental conditions (Iwamoto *et al.*, 2011).

Studies have investigated the effect of ageing on DNA methylation across brain regions. A key study in this field is by Hernandez *et al.* (2011) in which DNA methylation was analysed at over 27,000 CpG sites throughout the human genome in multiple brain regions; pons, frontal cortex, temporal cortex and cerebellum, extracted from individuals aged between 1 and 102 years. 589 CpG sites were found to be associated with age in at least one brain region. Increased methylation at ten CpG sites was associated with age in all four brain regions. Further investigation of the associated sites revealed that many were in close proximity to genes associated with DNA binding and regulated transcription, suggesting that specific age-related changes in DNA methylation might be important in the maintenance of transcription and gene expression in ageing tissues (Hernandez *et al.*, 2011).

A similar study looked at genome-wide DNA methylation signatures in the human prefrontal cortex across the lifespan with samples taken from the second trimester of gestation to old age. This study by Numata *et al.* (2012) identified the vast temporal patterns in DNA methylation across the life course. The prenatal period displayed the fastest changes in DNA methylation with the rate slowing during childhood and later life. The ten genes identified as being positively correlated with age in Hernandez *et al.* (2011) were also confirmed in this study. The age-related changes were predominantly found in CpG island shores, rather than in the islands themselves highlighting that CpG island shores may have a more important functional role in ageing and disease than previously thought (Numata *et al.*, 2012).

It is largely known that the formation of long-term memory is dependent on gene expression and a number of studies have now highlighted a pivotal role for epigenetics in learning. Histone modifications have previously been associated with learning and memory formation and more recent research points to a role for DNA methylation (Lubin *et al*., 2011). This evidence includes; the finding that DNA methylation changes are triggered by learning in the adult hippocampus; DNA methylation changes have been found to occur at both memory-permissive and memory-suppressive gene promoters when treated with DNMT inhibitors, suggesting a critical role for DNA methylation in memory formation (Lubin *et al.*, 2011). Furthermore, a study investigating the effect of DNMT inhibitors on memory found that when hippocampal slices were treated with DNMT inhibitors, there was a reduction in long-term potentiation, the cellular mechanism behind long term memory formation (Nelson and Monteggia, 2011). The potential role of DNA methylation in memory formation suggests epigenetics may also be involved in disorders affecting memory and cognitive decline including PSD and PD.

1.3.5.2 ***DNA methylation in diseased brains***

There is growing evidence to suggest that epigenetic mechanisms are involved in many diseases affecting the brain in old age. Epigenetic modifications have been implicated in several neurological, neurodegenerative and psychiatric disorders. Epigenetic mechanisms are known to be involved in the aetiologies of many neurological disorders. Mutations in the methyl-CpG-binding protein 2 gene on the X chromosome which codes for a protein involved in both DNA methylation and histone acetylation causes Rett syndrome; a neurodevelopmental disorder of the grey matter of the brain. Fragile X syndrome results from the hypermethylation of a CGG trinucleotide repeat on the X chromosome which results in the failure to express the fragile X mental retardation protein (FMRP), which is required for normal neural development. DNA methylation has also been associated with neurodegenerative disorders such as Alzheimer’s disease and Huntington’s disease and psychiatric disorders such as depression and schizophrenia (Nelson and Monteggia, 2011). Due to so many associations between epigenetic regulation and neurological disease, it is hypothesised that DNA methylation may have a role in the regulation of neuronal function and cognition, both during development and in the mature brain (Nelson and Monteggia, 2011). Many epigenetic studies have looked into DNA methylation patterns in the brain throughout the lifespan to try and identify the function of epigenetics in both brain development and disease. Here the discussion is limited to stroke, PSD and PD.

1.3.5.2.1 DNA methylation and stroke

Recent literature reports a number of associations between aberrant DNA methylation and several diseases including obesity, cancer and stroke. A number of studies have been performed in recent years to establish the relationship between DNA methylation and stroke. There is a delay between stroke and neuronal death and it is thought that this delay is due to transcriptional changes causing cell death (Hwang *et al.*, 2013).

It is recognised that DNA hypomethylation is associated with cardiovascular related diseases such as atherosclerosis and stroke. Baccarelli *et al.* (2010) used DNA samples extracted from the blood of 712 elderly male participants of the Normative Aging Study to measure DNA methylation levels of Long Interspersed Nucleotide Element 1 (LINE-1) which, being a repetitive element found throughout the genome is representative of global genomic methylation levels. Medical records and physician examinations were used to diagnose pre-existing CVD. The remaining participants who were free of CVD at baseline were followed up for approximately three years and the onset of any cardiovascular related disease was noted. Baccarelli *et al.* (2010) found that those with ischaemic heart disease (IHD) and/or stroke had lower levels of LINE-1 methylation and also reported that low LINE-1 methylation levels are associated with an increased risk of IHD and stroke. In addition, 24 participants who were disease-free at baseline but developed either IHD or stroke during follow-up had a second blood sample taken after IHD/stroke diagnosis. No difference in methylation levels were detected between the two samples. These results suggest that the decrease in LINE-1 methylation seen in the diseased participants precedes disease onset and might be useful in predicting disease risk (Baccarelli *et al.*, 2010). Another study by Castillo-Diaz *et al.* (2010) also reported DNA hypomethylation in patients at risk of stroke. They conducted array-based analysis of DNA methylation in both atherosclerotic and control arteries and found decreased levels of DNA methylation in the CpG islands of diseased arteries. Genes showing differential methylation were found to be in close proximity to known genes involved in atherogenesis suggesting DNA methylation may increase the risk of stroke by promoting atherogenesis (Castillo-Diaz *et al.*, 2010).

It is hypothesised that these methylation changes may be due to environmental factors such as diet, smoking behaviour and environmental pollutants (Mathers *et al.*, 2010) which then increase the risk of diseases such as stroke.

Using a mouse model for focal cerebral ischaemia, Endres *et al.* (2000) found that DNA methylation levels increased following a brain ischaemic injury. In addition, Endres *et al.* (2000) found that both treatment with 5-aza-2’-deoxycytidine (5-aza), an inhibitor of DNA methylation, and decreased expression of *Dnmt1* significantly reduced the extent of the ischaemic injury (Endres *et al.*, 2000; Qureshi and Mehler, 2010).

DNA methylation may play a role in the regulation of certain genes such as thrombospondin 1 in response to cerebral ischaemia (Qureshi and Mehler, 2010). The effect of DNA methylation on the vulnerability of the brain to ischaemic injury has also been suggested to be a likely candidate for the mechanism by which methylenetetrahydrofolate reductase (MTHFR) deficiency causes an increased risk of stroke, but this has not yet been proven (Qureshi and Mehler, 2010). MTHFR is the catalyst for the synthesis of 5-methyltetrahydrofolate, the precursor of SAM, and is therefore an important factor for DNA methylation (Stern *et al.*, 2000). Both MTHFR deficiency and certain *MTHFR* gene polymorphisms, notably C677T, have been implicated in hyperhomocysteinaemia and consequently in an increased risk of several diseases including stroke. Similarly, it was found that patients with high levels of plasma homocysteine, a common biomarker for vascular disease, had a significant decrease in leukocyte global DNA methylation when compared to a control group. It is thought that this alteration in DNA methylation may promote the vulnerability of the brain to ischaemic injury (Castro *et al.*, 2003).

The epigenetic effects of a stroke are not only detectable in peripheral blood and the affected brain tissue. It is thought that other somatic tissues may be affected by the insult through an epigenetic bystander-effect. Bystander-effects have previously been documented in radiation and chemotherapy exposure, where neighbouring cells can be affected. Kovalchuk *et al.* (2012) investigated levels of DNMTs in heart, liver and kidney tissues taken from rodents with an ischaemic insult. They reported slight changes in DNMT1 levels in the heart ($p<0.1$) but significant increases ($p<0.05$) in DNMT3a in the kidneys of stroke rodents. They also investigated levels of histone modifications in each of the tissues and found significant increases of 16% in the levels of H3K4 trimethylation, a 15% increase in acetylated H3K9 and 15% decrease in methylated H3K9 in the kidney tissue from stroke rats only. Gene expression analysis highlighted 22 genes which were up-regulated in the kidney tissue of the rats with stroke. These subtle yet significant findings are of great interest in the investigation into the role of epigenetics in stroke. As kidney damage and acute kidney failure are well-described post-stroke complications it is interesting to note the sensitivity of kidney tissue to ischaemic insult at the molecular level. The results of this study suggest that DNA methylation may play a role in the aetiology of stroke as well as having a role in the consequences post-stroke (Kovalchuk *et al.*, 2012). Antecedents of stroke may also be relevant to post-stroke dementia.

1.3.5.2.2 DNA methylation and dementia

Environmental and nutritional factors along with their actions through epigenetic mechanisms may contribute to the high prevalence of dementia in older populations. Reduced folate and vitamin B6 and B12 levels lead to increased levels of homocysteine, which is not only a known risk factor for AD but also, reduced DNA methylation (Leszek *et al.*, 2012). There is now a wealth of evidence to suggest DNA methylation is associated with dementia. The vast majority of studies have focussed on AD; however, there are a slow growing number of publications investigating the role of epigenetics in other dementias.

Most studies have compared methylation and expression levels of AD cases and controls. A global hypomethylation has been reported in the majority of studies with increasing age; however gene-specific DNA methylation analysis has revealed both hypomethylation and hypermethylation. Some studies have used easily accessible peripheral blood (Bollati *et al.*, 2011) to identify methylation and gene expression differences whilst others have used the most relevant target tissue; the post-mortem brain (Wang *et al.*, 2008; Bakulski *et al.*, 2012a; Chouliaras *et al.*, 2013). Methylation is tissue-specific; therefore the most suitable tissue to interrogate DNA methylation in dementia patients is the brain. This topic is explored in more detail in Section 1.4.3.1.

An interesting study by Chouliaras *et al.* (2013) compared levels of 5-mC and 5-hmC in the hippocampus of AD patients and age-matched controls. Both 5-mC and 5-hmC were found at reduced levels in the AD hippocampal tissues. Neuropathological investigations of the hippocampus revealed a negative correlation between levels of 5-mC and 5-hmC and amyloid plaque load, suggesting that DNA methylation is indeed associated with AD pathology (Chouliaras *et al.*, 2013).

Several genes implicated in early onset AD, including *PSEN1* and *APP* have been found to be hypomethylated in the brains of AD patients. This hypomethylation results in the aberrant upregulation of these genes which could be causal in the accumulation of β-amyloid (Aβ) (Sung *et al.*, 2011). In other genes, including *APOE* and *MTHFR* hypermethylation has been identified in the prefrontal cortex of AD brains and peripheral lymphocytes (Wang *et al.*, 2008). This highlights that the pathogenesis of AD may be caused by both the hypomethylation and hypermethylation of certain genes (Sung *et al.*, 2011). Further studies have highlighted that both hypomethylation and hypermethylation may be involved in disease pathogenesis. These include the finding that *CREB5* is hypomethylated in AD patients when compared to controls (Graff and Mansuy, 2009; Chouliaras *et al.*, 2010) and the study by Siegmund *et al.* (2007) which reported higher methylation of the *SORBS3* locus in the cerebral cortex of AD cases than in controls. *SORBS3* has previously been detected in the neuronal synapse indicating a possible function for *SORBS3* in AD progression (Siegmund *et al.*, 2007; Urdinguio *et al.*, 2009).

Recent DNA methylation studies have identified a number of loci differentially methylated in AD, some of which, e.g. *ANK1* have previously been associated with AD (De Jager *et al.*, 2014; Lunnon *et al.*, 2014). This indicates that DNA methylation analysis may highlight potential mechanisms of disease aetiology. Another recent study has highlighted differences in global DNA methylation using peripheral blood mononuclear cells between Alzheimer's disease cases and healthy controls, with global hypermethylation identified in the Alzheimer's cases. This hypermethylation was also associated with the presence of the APOE4 allele. This study suggests that global DNA methylation could be a useful biomarker of Alzheimer's disease (Di Francesco *et al.*, 2014). Another recent study by Yu *et al.* (2015) looked at genome-wide DNA methylation levels in the dorsolateral prefrontal cortex and identified a number of loci associated with AD pathology further implicating DNA methylation in the pathophysiology of AD (Yu *et al.*, 2015).

One of the larger more recent studies comparing methylation in brain tissue between AD cases and controls was conducted by Bakulski *et al.* (2012). Genome-wide analysis quantifying methylation at over 27,000 CpG sites across the genome was performed on twelve AD and twelve age and sex matched controls. They found 948 CpG sites that were associated with AD. The mean methylation difference was only subtle at 2.9%. Interestingly, one of those CpG sites, located in the promoter of *TMEM59* was 7.9% hypomethylated in cases when compared to controls. They extended their investigations to include gene expression and reported a negative correlation between DNA methylation and expression of *TMEM59*. As this gene has previously been implicated in the post-translational processing of APP, this study strengthens the suggestion that DNA methylation has a role in AD pathology (Bakulski *et al.*, 2012a).

Studies have also looked at early life exposures and their risk of AD in later life. One recent mouse study has identified that early life exposure to lead disrupts the normal methylation and expression patterns of AD-related genes in later life (Alashwal *et al.*, 2012). Studies of lead exposure in rats has identified dose-dependent decreases in the methylation of the promoter region of amyloid precursor protein (APP), indicating a possible link between lead exposure and AD (Li *et al.*, 2012). Lead exposure has also been well studied in humans. As it is a known neurotoxicant in children, associated with an increased risk of reduced intelligence quotient (IQ) and behavioural problems, it was hypothesised that lead exposure may influence cognitive function in later life. Lead exposure has also been suggested as a possible risk factor for AD and other dementias in epidemiological human studies. Cumulative lifetime lead exposure has been associated with accelerated cognitive decline and lead exposure has been associated with increased levels of homocysteine further supporting the evidence that lead exposure increases risk of AD by disrupting regulatory epigenetic mechanisms (Bakulski *et al.*, 2012b).

Desplats *et al.* (2011) studied the role of epigenetics in dementia with Lewy bodies and identified depleted levels of DNMT1 protein in the cell nuclei of diseased brains. This infers that DNA methylation may be involved in the mechanisms resulting in disease progression (Desplats *et al.*, 2011).

All reported changes in DNA methylation have been relatively small when compared to age-matched controls. However, it could be that these subtle changes in DNA methylation, when in combination with other predisposing factors such as *APOE* genotype, could augment the susceptibility to neurodegeneration (Leszek *et al.*, 2012).

1.3.5.2.3 DNA methylation and post-stroke dementia

Whilst there have been methylation studies in both stroke and dementia cohorts DNA methylation studies have never been performed in a post-stroke dementia cohort. Although little is known about methylation patterns in PSD it is likely that since methylation is involved in both stroke and dementia, as described in 1.3.5.2.1 and 1.3.5.2.2, DNA methylation may also have a role in PSD. Antecedents of stroke may also be relevant to post stroke dementia. Literature searches in September 2014 using the search terms “DNA methylation” AND “post-stroke dementia” and “DNA methylation” AND “stroke” AND “cognitive impairment” yielded no publications in PubMed (NCBI) highlighting a gap in current research.

1.3.5.2.4 DNA methylation and Parkinson’s disease

The role of DNA methylation in PD is not as well studied as in AD. However, there are a limited number of studies which suggest DNA methylation may have a mechanistic role in PD progression.

Arguably the main finding to date is that methylation is reduced in the first intron of α-synuclein (*SNCA*) and is associated with increased expression in the substantia nigra of PD cases (Jowaed *et al.*, 2010). Methylation of *SNCA* has also been studied in white blood cell DNA but no differences between PD cases and controls have been identified (Richter *et al.*, 2012), possibly due to the localisation of *SNCA* to neurons (Lu *et al.*, 2013). The finding that *SNCA* methylation levels differ in PD cases compared to controls is of great interest since α-synuclein is a major component of Lewy bodies, a feature of PD. Levels of DNMT1 have also been found to be reduced in PD brains indicating a possible cause of decreased *SNCA* methylation levels (Lu *et al.*, 2013).

Other studies have focussed on inflammation related genes such as tumour necrosis factor α (*TNFα*) which is hypothesised to promote dopaminergic cell death, possibly through the action of DNA methylation. TNFα is detected at much higher levels in the CSF of PD cases which suggests that DNA methylation may be reduced thereby causing *TNF-α* to be over expressed in PD. This, however, is just speculation and requires further study (Lu *et al.*, 2013).

The methylation status of cytochrome P450 2E1 (*CYP2E1*) has also been analysed. *CYP2E1* was found to be hypomethylated in the brains of PD cases which may suggest a role for detoxifying enzymes in PD pathobiology, however more research is required (Lu et al 2013).

An increased plasma homocysteine concentration has been associated with PD. Mouse models have shown that increased levels of homocysteine are detrimental to dopaminergic neurons (Lu *et al.*, 2013). Increased levels of homocysteine are also associated with hypomethylation further indicating that methylation may have a role in PD.

Very few studies have looked at the methylation profiles of PD-MCI cases compared to controls. One study that has found associations between methylation and cognitive function measured levels of a number of biochemical markers in the plasma of PD patients. They identified those with increased levels of SAM and Vitamin B6 were less likely to be cognitively impaired. Increased levels of SAM were also associated with lower levels of APP and SNCA. It has been speculated that Vitamin B6 may influence the degradation of APP which in turn affects α-synuclein levels, however the effect of Vitamin B6 on α-synuclein needs to be investigated further before any conclusions can be drawn (Obeid *et al.*, 2009).

1.3.5.3 ***DNA methylation as a candidate for prediction and prognosis in the ageing brain***

Epigenetics is one such possible mechanism that has not yet been researched in a post-stroke dementia cohort. Previous studies have successfully identified differential methylation in both stroke patients and dementia patients when compared to healthy age-matched controls. As epigenetic marks are affected in both of these diseases it is hypothesised that DNA methylation changes will be detected in PSD patients. As epigenetics is important in both stroke and dementia, it may also be a key factor in contributing to the risk of developing PSD. As described in Section 1.3.5.2.1, LINE-1 hypomethylation was detected prior to stroke suggesting that this epigenetic marker could be used as a biomarker for stroke. Like PSD, DNA methylation studies have not been carried out in a PD-MCI cohort. The possibility of using DNA methylation as a biomarker for PSD would be extremely beneficial to both patients and health care services. The early detection of patients at-risk of developing PD-MCI would be of great benefit to PD patients. The earlier detection may allow more specific disease-modifying treatments to be administered earlier which could prevent the development of disease. In both diseases the identification of biomarkers could be useful in a number of ways. A biomarker can aid the early diagnosis of disease and repeated measures can help monitor disease progression. Biomarkers could also be used to monitor treatment success and to indicate exposure status. Finally, a biomarker detected in the affected tissue (in PSD and PD-MCI this is the brain) can be used to help understand the mechanisms involved in the disease pathway. These mechanistic markers could then help to target treatment options such as the inhibition of the involved pathways. Therefore, in both PSD and PD-MCI, the identification of biomarkers can have great utility for a variety of appplications.

1.4 Studying DNA methylation in the brain

1.4.1 Epidemiological approaches

Epigenetic epidemiology is the term used to describe the study of epigenetic variation within populations by combining well established epidemiological approaches with epigenomic technologies (Mill and Heijmans, 2013). Large-scale population-based studies are becoming increasingly common in studies investigating the associations between epigenetics and disease phenotype, due to the recent advances in genomic technologies (Ng *et al.*, 2012). Epigenetic epidemiology aims to add to the understanding of mechanisms involved in the associations between exposures and human disease phenotypes (Heijmans and Mill, 2012). A number of considerations need to be explored when designing suitable epidemiological studies for the effective interrogation of epigenetic mechanisms. A few of these are described below.

1.4.1.1 ***Study designs***

There are many different study designs available for use in an epigenetic epidemiology context each with their own advantages and disadvantages. It is important to select the correct type of study design to use in each study; this depends on the exposures and outcomes of interest and the effect of any confounding. The types of study design used in this thesis are described below.

Cross-sectional study

A cross-sectional study measures all exposures and phenotypic measures at one time-point only. They are suitable for studies wishing to determine the prevalence rate of a phenotype or disease. They can also be used to assess differences between different sample sets, such as different sexes or ethnicities at that one time-point. Cross-sectional studies take a snapshot of the epigenome at the specific point in time, and when combined with data on exposures and phenotypic measures, can be used to assess the relationship between exposure and the epigenome and/or the epigenome and disease. Cross-sectional studies are advantageous as they are cheap to carry out and maintain as they do not require too much effort on the study participants behalf. However, results are only representative of the study sample and due to only one time-point used it is impossible to assess the temporal nature of epigenetics or investigate causality.

Cohort studies

In a cohort study, participants are usually recruited for a specific time period, however many cohort studies are open-ended and participants are followed up throughout their life. Studies including prospectively collected data and biological samples at multiple time points are known as longitudinal studies. Prospectively collected data is of more value than retrospectively collected data as it is more accurate due to the risk of recall bias arising in many retrospective studies. These studies are much more valuable for epigenetic studies hoping to establish temporal relationships and biological relevance of epigenetic changes associated with exposures and/or disease. At recruitment, baseline measurements and phenotypic information including detailed exposure information is collected and at each follow-up throughout the study further biological and phenotypic measurements are collected. This allows any changes in DNA methylation or disease state to be assessed over time. A cohort study is usually made up of hundreds or even thousands of participants. As clinical and epigenetic measurements are collected before disease onset, cohort studies can be used to infer temporal relationships and strengthen causality as well as remove the problem of reverse causation. Although being more expensive, studies utilising multiple time points are increasingly being recognised as the most suitable study design to investigate the epigenetics of common complex diseases. Several research collaborations involving cohort studies such as Genomic and Epigenomic Complex Disease Epidemiology (GEoCoDE) (http://www.bristol.ac.uk/caite/geocode/) have been formed which increase sample sizes and power in the investigations of disease risk factors throughout the life course. Collaborations such as GEoCoDE have also enabled findings from individual studies to be replicated in a much larger cohort (Ng *et al.*, 2012). Nested case-control studies can also be created from existing cohort studies. Participants who develop the disease phenotype of interest during the cohort study period are recruited into the nested case-control study as cases and matched individuals free of disease are taken from the cohort to be used as controls.

A summary of the different study designs used in epidemiological studies and their applications is found in Table 2.

Study Design	Application
Cross-sectional study	Prevalence of an epigenetic mark in a well-defined population subgroup
Retrospective case-control study	Permanent epigenetic marks among individuals with and without disease
Cohort study	Epigenetic mechanisms underlying a risk factor-disease association
Nested (prospective) case-control study	Epigenetic marks predisposing to disease Biomarkers for early disease detection Biomarkers for disease risk
Intervention study	Effect of intervention on epigenetic pattern Effect of epigenetic therapies on disease
Family-based study	Transgenerational inheritance of epigenetic traits
Birth cohort	Influence of preconceptional and prenatal factors on establishment of the genome

Table 2: Study designs used in epigenetic epidemiology and their applications. (Taken from Michels, 2012).

1.4.1.2 ***The value of twin studies***

Cases of discordance in monozygotic (MZ) twins, where one twin developed a disease whilst the other did not, have been extremely valuable in the study of epigenetics and the ageing process. Due to MZ twins having the same genetic profiles, DNA methylation changes over the life course appear to be a credible explanation for this discordance between MZ twin pairs (Gilbert, 2009). More divergence in DNA methylation patterns has been observed between elderly twin pairs when compared to younger twin pairs. This suggests that epigenetic control may be more stringent during early life (Fraga *et al.*, 2005). These findings were replicated in a larger study by Talens *et al.* (2012) which assessed methylation in MZ twins of a wider age range. This study found more discordance in methylation between the older twin-pairs compared with younger twin-pairs in a number of common disease associated genes (Talens *et al.*, 2012).

Significant differences in DNA methylation have been identified in twins discordant for breast cancer. The Infinium HumanMethylation450 (HM450) BeadChip (Illumina, USA) has been performed on peripheral blood taken from fifteen twin pairs; one of whom had been diagnosed with the breast cancer whilst the other remained healthy. 403 CpG sites were found to be differentially methylated, 97% of which were hypomethylated in the twin with cancer. In addition to identifying differentially methylated genes after cancer diagnosis, blood samples taken prior to breast cancer diagnosis were also analysed and this analysis identified potential biomarkers for breast cancer. *DOK7* was identified as the most likely candidate for a biomarker as it displayed the most significant change in DNA methylation and these changes were detectable three years prior to breast cancer diagnosis. This study shows the value of using MZ twins in epigenetic studies. They are the most efficient study design due to factors which would usually add variability such as age, genetic factors and other environmental factors being controlled for (Heyn *et al.*, 2013).

Twin studies have also been important in the study of epigenetics and neurodegenerative disorders. Mastroeni *et al.* (2009) analysed DNA methylation levels in the cortical neurons of one pair of MZ twins discordant for AD. Both twins were male and died between ages 76-79. The neuropathology of the twins differed greatly with the AD twin having Braak stage VI pathology whilst the unaffected twin had Braak stage II pathology. Immunochemistry of the temporal cortex revealed lower levels of 5-methylcytosine in the AD twin compared to the unaffected twin. When the test was repeated using tissue from the cerebellum (largely unaffected by AD) no difference in levels of 5-methylcytosine were identified. This indicates that the temporal cortex differences in 5-methylcytosine were not due to tissue handling differences or other human error, but may suggest a role for epigenetic mechanisms in the pathophysiology of AD (Mastroeni *et al.*, 2009).

1.4.2 Methods to quantify DNA methylation

Quantification of DNA methylation can be either global or gene-specific. The different methods of quantifying DNA methylation are explored below.

1.4.2.1 ***Global DNA methylation***

The quantification of DNA methylation on a global scale takes advantage of repetitive elements that are found throughout the genome. Repetitive elements are transposable elements of DNA found repeated throughout the genome. Quantifying the methylation levels of these repetitive elements measures the average methylation level found throughout the genome, making it a good surrogate for global methylation analysis. Repetitive elements can be categorised into several categories, the most commonly used being long interspersed nucleotide elements and short interspersed nucleotide elements. Sat2 and Alu are also commonly used measures of global DNA methylation levels (Weisenberger *et al.*, 2005). Measuring DNA methylation levels of repetitive elements usually requires PCR primers specific to a particular repeated sequence such as LINE-1 (Yang *et al.*, 2004). A more sensitive way of assessing global methylation uses High Performance Liquid Chromatography (HPLC) which separates nucleotides according to size and quantifies unmethylated and methylated cytosine residues. Whilst this technique is more sensitive, it requires larger amounts of DNA compared to the PCR-based techniques using repetitive elements and therefore limits its use (Lisanti *et al.*, 2013). Quantification of global DNA methylation levels was not used in this thesis but many important studies in the field of epigenetics and ageing have performed global methylation analysis.

1.4.2.2 ***Targeted DNA methylation***

Whilst global methylation is an average measure of methylation levels throughout the genome, DNA methylation can also be measured at the level of individual CpG sites using either genome-wide or candidate gene (locus-specific) approaches.

1.4.2.2.1 Genome-wide methylation

Studies analysing genome-wide methylation are becoming increasingly popular as technological advances are increasing the number of CpG sites that can be interrogated per sample, simultaneously. Microarray techniques are the most commonly used method to measure genome-wide methylation (although only a small fraction of CpG sites are included in the microarrays). The HM450 has, in recent years, replaced the 27K as the most commonly used BeadChip. The HM450 BeadChip assesses the methylation status at over 480,000 CpG sites which are spread throughout the genome. CpG sites included in the array are not only found in CpG islands, but CpG shores, gene bodies and 3’-and 5’-untranslated regions (UTRs), to name a few (Bibikova *et al.*, 2011). Microarray platforms use BeadChips which have probes able to discriminate between methylated and unmethylated DNA, so when bisulphite modified DNA (see Sections 1.4.2.2.2 and 2.3) is hybridised to the BeadChips the methylation status at each CpG site captured on the array can be analysed (Beck and Rakyan, 2008). The HM450 array was used in this thesis and is described in more detail in Section 2.4.

1.4.2.2.2 Locus-specific methylation

For many years, the gold standard for measuring DNA methylation at a specific locus has been bisulphite sequencing (Beck and Rakyan, 2008). This involves treating genomic DNA with sodium bisulphite which deaminates unmethylated cytosine residues into uracil. As methylated cytosines are resistant to this modification, they remain as cytosine residues. On PCR amplification the uracil residues are amplified as thymine bases and subsequent sequencing of DNA calculates DNA methylation levels of each CpG site by comparing the relative abundance of cytosine to thymine residues (Yang *et al.*, 2004).

A relatively recent development and a more cost effective approach is Pyrosequencing, which was initially developed for SNP analysis of DNA, but when preceded by bisulphite modification, can quantify the methylation of specific CpG sites in the DNA sequence. Pyrosequencing adopts a “sequence by synthesis” strategy whereby one nucleotide is incorporated into the single stranded sequence at a time. When a nucleotide complementary to the target template is added, pyrophosphate is released which initiates an enzyme cascade during which luciferase converts luciferin into oxyluciferin generating visible light which can be detected by a camera. The amount of light generated when each nucleotide is added determines the DNA sequence. When a bisulphite modified DNA sequence is run on the Pyrosequencer the methylation status of each CpG site can be read as a C/T SNP. The amount of light generated when cytosine and thymine residues are incorporated is compared to quantify DNA methylation (Dupont *et al.*, 2004; Reed *et al.*, 2010). In essence, Pyrosequencing is a form of targeted bisulphite sequencing which allows short fragments of DNA to be analysed rather than the whole genome. A limitation of bisulphite sequencing techniques is the inability to distinguish between 5-mC and 5-hmC. Using these techniques, all methylation marks (5-mC and 5-hmC) are presumed to be 5-mC. Studies using these techniques are therefore unable to detect whether differences in DNA methylation between groups are due to differences in 5-mC or 5-hmC (Huang *et al.*, 2010). Pyrosequencing has been used in this thesis and is described in Section 2.6.6.

1.4.3 Considerations when designing an epigenetic study

There are a number of factors that need to be considered when designing an epigenetic study. The two issues pertinent to this thesis are described below.

1.4.3.1 *Tissue specificity*

In multicellular organisms, different cells have different gene expression profiles which determine the function of each cell. Each organ in the human body, therefore, is made of a number of different cells; each cell type with a different expression, and therefore a different epigenetic profile. DNA methylation patterns are highly tissue-specific with greater variation found between tissues within individuals than between individuals themselves (Byun *et al.*, 2009). For this reason, it is important to use the affected/diseased tissue as the target tissue in epigenetic studies. However, this can cause problems for epigenetic studies especially since, in the majority of cases, to extract a sample of affected tissue from a living cohort requires an invasive procedure such as a biopsy. Animal models are useful in the study of 'hard to reach' tissues as these can be obtained post-mortem. Additionally, studies of the human brain are only possible using post-mortem tissue. For this reason, many studies resort to using easily accessible such as buccal cells or most commonly peripheral blood as a source of DNA. However, it is not well established whether or not blood is a useful surrogate for each human tissue – investigations into this are still on-going (Davies *et al.*, 2012; Masliah *et al.*, 2013). In the search for a clinical biomarker however, DNA methylation changes that are mechanistic are not required, making blood an ideal, easily accessible source of DNA. Cell heterogeneity is also an important consideration. Blood and other tissues such as brain are made up of different cellular compositions which could bias epigenetic measurements (Ng *et al.*, 2012).

1.4.3.2 ***Temporality of epigenetic change***

Epigenetic changes are dynamic. Studies have shown epigenetic profiles to alter with age (Fraga *et al.*, 2005) and other environmental stimuli such as smoking (Breitling *et al.*, 2011). Some epigenetic changes may be due to the disease in question whereas others may be due to certain exposures experienced either before or after disease onset. To establish whether epigenetic changes are causal in disease onset or progression the timing of epigenetic changes needs to be known. This requires epigenetic signatures to be profiled immediately before and after any exposures of interest to be able to determine whether any later life changes in phenotype such as disease onset are as a result of the exposure and not just a result of reverse causation. This is often not practical in human populations due to not knowing when certain exposures will be experienced in advance. Studies investigating exposures which result in a rapid and temporary change in epigenetic profiles may be best suited to experimental approaches.

1.4.4 Approaches to studying the role of DNA methylation in the aetiology of post-stroke dementia and mild cognitive impairment in Parkinson’s disease

A longitudinal study design which measures DNA methylation at two or more time points is particularly useful in epigenetic studies when causality is to be established. Longitudinal studies are able to track methylation changes over time enabling the temporal relationship (Section 1.4.3.2) between exposure and disease to be established. In contrast, cross-sectional studies which measure methylation at only one time point are unable to capture the dynamic nature of epigenetic mechanisms and, therefore cannot be used to establish causality (Ng *et al.*, 2012).

In the study of both PSD and PD-MCI, the most affected organ is the brain which makes longitudinal studies difficult, as the brain can only be extracted following death. The use of peripheral blood as a surrogate for the target tissue must therefore be used as it is an easily accessible DNA source, however, tissue-specific methylation differences, as mentioned in Section 1.4.3.1, do exist (Yang *et al.*, 2010; McKay *et al.*, 2011). DNA methylation in peripheral blood has been a successful predictor of other diseases (Brennan *et al.*, 2012) so the possibility that peripheral blood could be used to identify patients at risk of PSD and PD-MCI is feasible and exciting. Brain tissue can be used to assess the methylation status at death but to identify at-risk individuals a DNA sample collected before the onset of dementia is required. Peripheral blood may therefore be a suitable choice.

1.5 Exploring the functional importance of observational associations detected in epigenetic epidemiological studies

Observational studies which highlight an association between an epigenetic alteration and phenotype are not able to establish either causality or functional relevance. Establishing causality and functional relevance is an important part of epigenetic studies. Identifying differences in DNA methylation does not mean that epigenetic mechanisms are mechanistically involved in disease; DNA methylation may instead be a consequence of disease or may be entirely unrelated. Identifying functional pathways offers the greatest chance for disease prevention. There are a number of ways we can elucidate functional pathways and disease mechanisms involved in disease pathogenesis in DNA methylation studies and these will be outlined below.

1.5.1 Biological function

The simplest approach that may be taken to assess whether an alteration in DNA methylation may have functional importance is to look at the biological function of the gene in question.

Since the human genome has been fully sequenced we can predict the likely protein structure that will be produced by gene transcripts. One of the resources available for predicting protein structure is Simple Modular Architecture Research Tool (SMART) (http://smart.embl-heidelberg.de/). SMART looks for conserved domains that have been previously characterised in a range of proteins. An example of this would be if a protein contains a transmembrane domain, it is likely to have a function at the cell membrane such as in a cell signalling pathway.

Gene ontology aggregates known gene attributes into a single database (http://www.geneontology.org/). The data collected comes from a range of organisms so data from one organism can be used to infer function of a homologous gene in another organism. The database contains information on biological processes, molecular function and cellular component. This is especially useful when looking at multiple genes to see if they act on the same pathways or have a shared function. By using these resources it allows us to prioritise only the most functionally relevant hits for more detailed analysis.

1.5.2 Gene expression

DNA methylation is commonly inversely correlated with gene expression. When CpG sites, particularly those close to the transcription start site, are methylated this impedes transcription factor binding thereby repressing gene expression (Boyes and Bird, 1991). This relationship, as previously mentioned, is not expected within gene bodies where an increase in methylation has been associated with an increase in transcription levels (Wagner *et al.,* 2014). It is possible to test whether DNA methylation impacts upon transcription, gene expression analyses by interrogating publicly available data or by experimental studies where RNA samples are available. To assess the expression levels of a number of genes an expression microarray or RNA sequencing could be performed. For validation of array or RNA sequencing data, or if analysis is targeted to just a few genes, quantitative PCR can be used. Using cell culture, the effects of DNA methylation on gene expression can be directly monitored using de-methylating agents such as 5-aza (Hitchins *et al.*, 2011).

In addition, quantification of protein levels could also determine whether differential gene expression is also detectable at the protein level. This could be achieved using either a proteomic (array or 2 dimensional gel electrophoresis) approach or the more targeted method of Western blotting or enzyme-linked immunosorbent assay. If methylation differences were found to alter gene expression and protein levels then this would provide compelling evidence that the alteration in DNA methylation is exerting a biological effect and could be involved in disease causation.

1.6 Study objectives, aims and hypotheses

The main aims of this thesis were to identify epigenetic biomarkers present in the blood of stroke survivors and Parkinson’s disease patients which could be used to predict those patients at risk of developing post-stroke dementia and mild cognitive impairment in Parkinson’s disease.

The objectives of this work were:

1. To identify blood-based differentially methylated positions (DMPs) which could be potential biomarkers of post-stroke dementia;
2. To identify DMPs in the brain that are associated with impaired cognitive function in stroke patients;
3. To identify blood-based DMPs which could be potential biomarkers of mild cognitive impairment in Parkinson’s disease;
4. To assess whether DNA methylation can mediate the relationship between exposure and outcome;
5. To assess whether DNA methylation can be used as an indicator of exposure status;
6. To assess whether biological age predicted from DNA methylation data can identify individuals at risk of cognitive decline.

The following hypotheses were tested:

1. Specific changes in blood DNA methylation measured prior to onset of dementia predict later cognitive impairment;
2. Patients suffering from post-stroke dementia exhibit different DNA methylation signatures in their blood and brain tissue when compared with cognitively normal post-stroke patients;
3. Patients suffering from PD-MCI exhibit different DNA methylation signatures in peripheral blood when compared with cognitively normal Parkinson’s cases;
4. Specific changes in DNA methylation measured at diagnosis of Parkinson’s disease predict later cognitive and motor impairment;
5. DNA methylation may mediate the relationship between exposure and outcome;
6. DNA methylation can be used as an indicator of exposure status;
7. Biological age predicted from DNA methylation data may predict participants at increased risk of cognitive decline.

1.7 Summary of thesis chapters

This thesis is structured as described below. Chapter 2 contains a summary of the methods and equipment used throughout this thesis. Chapters 3-7 contain the results of this thesis. In more detail, Chapter 3 addresses the possibility of using DNA methylation data to identify a biomarker of post-stroke dementia. Chapter 4 investigates whether DNA methylation data can inform about disease mechanisms involved in post-stroke dementia. Chapter 5 addresses the possibility of using DNA methylation data to identify a biomarker of mild cognitive impairment in Parkinson’s disease. Chapter 6 investigates the relationship between exposures and DNA methylation and whether DNA methylation data can be used as a surrogate for exposure status. Chapter 7 explores the epigenetic clock method and assesses whether age predicted using DNA methylation data can predict those more likely to suffer from age-related cognitive impairment. Finally, Chapter 8 discusses the main findings of this thesis, the strengths and limitations of this work and explores possible next steps in this field.

Chapter 2. Methods

2.1 Study cohorts

Two cohorts were utilised throughout the course of this project to investigate the role of DNA methylation in post-stroke dementia and Parkinson's disease and to assess the utility of epigenetic biomarkers in predicting disease outcome. The cohorts are described in the following sections.

2.1.1 COGFAST

The **Cog**nitive **F**unction **A**fter **St**roke (COGFAST) cohort (Allan et al., 2011) is a prospective observational study initiated in 1999 with the aim of understanding the risk factors and mechanisms that contribute to post-stroke dementia in elderly stroke survivors.

Recruitment: Hospital-based stroke registers were utilised to recruit cognitively normal (non-demented) stroke survivors three months post-stroke. Patients were excluded from the study if they; i) were under the age of 75 years; ii) had a significant physical illness or disability; iii) had a Mini-Mental State Examination (MMSE) (Folstein *et al.*, 1975) score of <24 and iv) were diagnosed with dementia according to the DSM-IIIR. Of the 706 older stroke survivors originally screened consecutively from six hospital-based stroke registers in Tyneside and Wearside in the North East of England, a total of 355 individuals were eligible to be recruited at baseline for the follow up COGFAST study.

Clinical assessment: At baseline (three months after stroke) all potential participants had their cognitive function assessed by a series of neuropsychological assessments including the Cambridge Cognitive Examination (CAMCOG) (de Koning *et al.*, 1998); a battery of tests which aids dementia diagnosis by assessing a range of cognitive functions affected by dementia and is able to detect mild cognitive impairment, and the MMSE; a series of questions and tests with a total score of 30, used to assess a person's language, attention and memory, useful for diagnosis of dementia as well as progression and severity. Participants' medical history and information relating to lifestyle was also recorded. Information relating to the index stroke was also collected at baseline.

Follow-up: The neuropsychological assessments were repeated each year to assess whether the participant's cognitive function had deteriorated and whether they were clinically demented.

*Biological samplin*g*:* At baseline, a peripheral whole blood sample was collected and stored in vacutainer tubes at -80°C. On death, the brain was extracted and the neuropathology of the brain was assessed. Each brain was given a Braak Staging score to inform in the final diagnosis of dementia. Braak staging is a method used to determine the degree of AD pathology through the assessment of the location of neurofibrillary tangles (Braak and Braak, 1995). It is commonly used in both Parkinson's disease and Alzheimer's disease. A final diagnosis of dementia as well as the Braak staging of the brain was given at autopsy and brains were frozen at -80°C to preserve tissue integrity (Allan *et al.*, 2011).

Ethical approval: Approval for use of samples and data for research was obtained from the Joint Ethics Committee of Newcastle and North Tyneside Health Authority, the University of Newcastle upon Tyne and the University of Northumbria at Newcastle. Participants gave written informed consent in accordance with the Declaration of Helsinki.

2.1.1.1 ***Sample selection***

All COGFAST study members were eligible for inclusion in the epigenetic studies reported in this thesis. Participants were selected based upon the availability of biological samples. Participant samples were included if they had donated both a baseline blood sample and a post-mortem brain sample. As both the hippocampus and dorsolateral prefrontal cortex (DLPFC) are regions of the brain associated with ageing and dementia, these regions were selected for epigenetic analysis. The DLPFC has a functional role in working memory and executive function. Studies assessing the effects of DLPFC damage have reported a worsening of spatial and verbal reasoning (Barbey *et al*., 2013). The hippocampus is a widely studied region of the brain and has been strongly implicated in long-term, and to a lesser extent, short-term memory formation. Damage to the hippocampus has been found to affect the ability to form new memories (Abrahams *et al.*, 1997; King *et al*., 2004). Specific regions were requested from COGFAST samples collected by the Newcastle Brain Tissue Resource (NBTR). These regions were: hippocampus, Level 18-19 according to the Newcastle Brain Map and dorsolateral prefrontal cortex, Levels 5-7 according to the Newcastle Brain Map (BA 9+46) (Perry and Oakley, 1993). Figure 6 shows the locations of the regions of interest in relation to the brain.

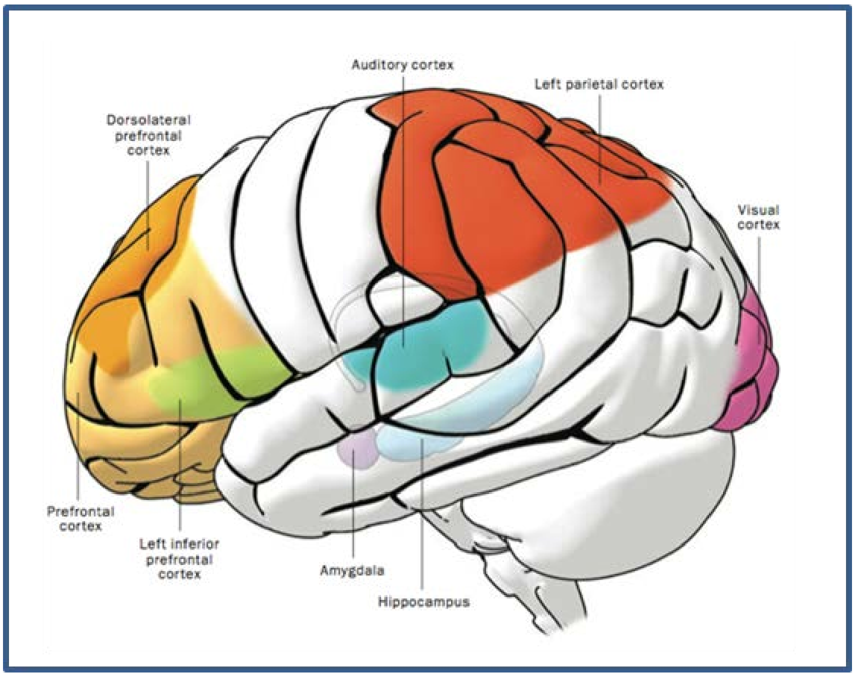


Figure 6: Anatomy of the brain. (Taken from Wickelgren, 2012).

2.1.1.2 ***Variables used in the analysis***

Hippocampus and DLPFC tissue and corresponding baseline blood samples were made available from the NBTR for the current project. A range of variables were used as outcome measures, exposure measures and covariates. Definitions for all variables can be found in the Glossary (Tables 1 and 2).

2.1.1.2.1 Outcome measures

The following variables were used as outcome measures in the analysis:

- Cognitive scores - MMSE (Folstein *et al.*, 1975) and a number of CAMCOG (de Koning *et al.*, 1998) scores (abstract thinking, attention, calculation, executive function, language comprehension, language expression, memory learning, memory recent, memory remote, memory total, orientation, perception, praxis, total CAMCOG score).
- Neuropathological information – Braak Staging, diagnosis.

2.1.1.2.2 Exposure measures

The following variables were used as exposure measures in the analysis:

- Medical history - hypertension, angina, atrial fibrillation, cardiac failure, intermittent claudication and ischaemic heart disease.

- Stroke-related factors – degree of weakness in arm/leg, side of the body affected, dysphasia and the Oxfordshire Community Stroke Project (OCSP) class (Bamford *et al.*, 1991).
- Risk factors – Smoking status and total number of cardiovascular risk factors.
- Baseline cognition scores – MMSE, abstract thinking, attention, calculation, executive function, language comprehension, language expression, memory learning, memory recent, memory remote, memory total, orientation, perception, praxis, total CAMCOG score.

2.1.1.2.3 Covariates

The following variables were used as covariates in the analysis: Baseline age, Sex and Chip on which the samples were scanned.

2.1.2 ICICLE

The **I**ncidence of **C**ognitive **I**mpairments in **C**ohorts with **L**ongitudinal **E**valuation or ICICLE study (Yarnall *et al.*, 2014) aimed to recruit all newly diagnosed Parkinson's disease (PD) patients in the Newcastle and Gateshead area.

Recruitment: The recruitment period ran from June 2009 to December 2011 and all primary care practices as well as secondary care health workers such as neurologists, geriatricians and PD nurse specialists were encouraged to refer all patients with suspected Parkinsonism. Participants were also excluded from the study if they i) were diagnosed with PD prior to the study recruitment window; ii) lacked the English skills required to complete questionnaires or assessments; and iii) showed signs of significant cognitive impairment scoring <24 on the MMSE. Age-matched healthy control subjects were recruited through local advertising and other community sources ensuring they were representative of the North East population. In total, 158 cases and 99 controls were recruited onto the project.

Clinical assessment: All patients recruited onto the study were diagnosed with idiopathic Parkinson's disease by a movement disorder specialist. Patients diagnosed with vascular, drug-induced or atypical forms of Parkinsonism were excluded from the ICICLE study. After consenting to participate in the study all PD cases underwent a clinical assessment by a physician. This assessment included education level, biometric measurements and other lifestyle factors such as smoking status. Symptom history, other co morbid diseases and medication use were also recorded. A number of cognitive tests were utilised including the MMSE (Folstein *et al.*, 1975), Montreal Cognitive Assessment (MoCA) (Nasreddine *et al.*, 2005), CDR (Wesnes *et al.*, 2002) and CANTAB (Sahakian *et al.*, 1988) to assess global cognitive function, the Geriatric Depression Scale (GDS-15) (Yesavage *et al.*, 1982) to assess depression, Hoehn and Yahr (Hoehn and Yahr, 1967) to rate the severity of disease (score 1-5; with higher score indicating more severe disease) and The Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS) parts II and III (Goetz *et al.*, 2008) to assess motor disability. A diagnosis of MCI was given if participants showed impairment (1-2 standard deviations below normative values) on at least two neuropsychological tests.

Follow-up: All cases and controls were invited to take part in a follow up at 18 months. All assessments were repeated at follow up to check for any alterations in disease phenotype (Khoo *et al.*, 2013).

Biological sampling: A peripheral blood sample was taken from both cases and controls at baseline. The blood sample was collected in a vacutainer containing ethylenediaminetetraacetic acid (EDTA) and stored at -80°C. DNA extracted from baseline blood samples was provided for epigenetic analysis.

Ethical approval: Approval for use of samples and data for research was obtained from the Newcastle and North Tyneside Research Ethics Committee (REC:08/H0906/147) and the study was performed according to the Declaration of Helsinki. All subjects provided written informed consent.

2.1.2.1 ***Sample selection***

150 PD cases and 90 controls donated a baseline blood sample. From these baseline blood DNA samples, 48 were selected for inclusion in the discovery phase of this project. 24 samples were from Parkinson's disease participants with mild cognitive impairment (MCI) and 24 Parkinson's disease participants without MCI. Initial sample selection was based on phenotype, with the selected PD-MCI cases displaying the most impaired cognition (at 2SD below normative levels). All other samples (with normal cognition) selected for inclusion in the discovery phase were age and sex matched. All other ICICLE DNA samples were available for use in the validation phase of this study.

2.1.2.2 ***Variables used in the analysis***

A range of variables were used as outcome measures, exposure measures and covariates. Definitions for all variables can be found in the Glossary (Tables 3 and 4).

2.1.2.2.1 Outcome variables

The following variables were used as outcome measures in the analysis:

- Cognition- MCI, animals, cognitive complaint, language, language total, MMSE, MoCA, naming, NMSQ concentration, NMSQ memory, one touch stockings, paired associates learning, pentagon copying, power of attention, pattern recognition memory, spatial recognition memory and total FAS.
- Motor: digit vigilance accuracy, Hoehn and Yahr, MDS UPDRS Parts II and III, PIGD phenotype and tremor dominant phenotype.

2.1.2.2.2 Exposure variables

The following variables were used as exposure measures in the analysis:

- Lifestyle factors – alcohol, smoking status, weight, BMI, height.
- Risk factors – B12, homocysteine, red cell folate (RCF), education, NART.
- Medical history – IHD, hypertension, hypercholesterolaemia, diabetes, depression.

2.1.2.2.3 Covariates

Age, sex and chip (on which the samples were scanned) were used as covariates in all analyses.

2.2 DNA extraction and quantification

2.2.1 Tissue DNA extraction

Sections from the DLPFC and hippocampus were provided by the NBTR for DNA extraction. To extract DNA from the brain tissue the E.Z.N.A.® Tissue DNA Kit (OMEGA bio-tek, USA) was used and the standard Tissue DNA Spin protocol was followed. The extraction protocol was performed on DLPFC pieces weighing approximately 30mg and hippocampus sections weighing approximately 10mg. Each tissue section (30mg/10mg) was added to 200µl TL Buffer and 25µl OB protease. Tubes were then vortexed and incubated at 55°C on a thermoshaker (Grant Instruments, UK) until all tissue had lysed. Samples were centrifuged for 5 minutes at 13000xg and the supernatant was transferred to a new sterile tube. 220µl BL Buffer was added, vortexed and incubated at 70°C for 10 minutes on a thermoshaker (Grant Instruments, UK). Following the addition of 220µl absolute ethanol, the entire sample was transferred to a HiBind DNA Mini Column and centrifuged at 10000xg for 30 seconds. In a new collection tube, 500µl HB Buffer was added to the column and centrifuged at 10000xg for 30 seconds. Two wash steps with 700µl Wash Buffer followed, each punctuated with a centrifugation step. A final centrifugation step was performed at 13000xg for 2 minutes to completely dry the column. The HiBind DNA Mini Column was inserted into a 1.5ml microcentrifuge tube and 100µl of preheated (70°C) Elution Buffer was added, left to stand at room temperature for 2 minutes and centrifuged at 13000xg for 1 minute to elute the DNA. Following elution, all DLPFC DNA extracted from the same brain section were pooled. Similarly, all hippocampal DNA samples extracted from the same individual were pooled. All DNA samples were stored at -70°C.

2.2.2 Blood DNA extraction

DNA was extracted from frozen blood using the standard Blood and Body Fluid Protocol as outlined in the E.Z.N.A.® Blood DNA Kit (OMEGA bio-tek, USA). This protocol involved the addition of 25µl reconstituted OB protease and 250µl Buffer BL to 250µl whole blood. After vortexing for 15 seconds, the tubes were incubated at 65°C for 10 minutes on a thermoshaker (Grant Instruments, UK) and 260µl absolute ethanol was added. The lysate was transferred to a HiBind DNA column, previously treated with 100µl Equilibration Buffer, and was centrifuged at 10000xg for 1 minute to bind the DNA. In a new collection tube, 500µl HB Buffer was added to the column and centrifuged as above. Two wash steps with 700µl Wash Buffer then followed each punctuated by a centrifugation step. Columns were then centrifuged at 13000xg for 2 minutes to completely dry the column. DNA samples were then eluted in 100µl preheated (65°C) Elution Buffer and left to stand at room temperature for 5 minutes before being centrifuged at 13000xg for 1 minute. DNA samples extracted from the same blood sample were then pooled and DNA samples were stored at -70°C.

2.2.3 Quantification of DNA

DNA samples were quantified using a ND1000 Spectrophotometer (Labtech International Ltd, UK). The ND1000 was first initialised with 1.5µl of distilled H_2O and a blank measurement was taken using 1.5µl of Elution buffer. 1.5µl of each sample was then loaded onto the ND1000 and was measured. Extracted DNA samples with a concentration lower than 50ng/µl were precipitated following the method described below.

2.2.4 Precipitation of DNA

Where low concentrations of extracted DNA were obtained, DNA was precipitated to give a higher concentration. This involved the addition of 2x volume of ice cold 100% ethanol, 1/8x volume of 5M NaCl and 1µl glycogen. Samples were incubated overnight at -70°C. Samples were then centrifuged at 14000xg for 30 minutes and 200µl of ice cold 70% ethanol was added to the supernatant. A 5 minute spin step then followed and after the supernatant was removed, the pellet was left to air dry for 30 minutes. The samples were then re-suspended in 30µl DEPC-treated water and were quantified on the ND1000 Spectrophotometer (Labtech International Ltd, UK).

2.3 Bisulphite modification of DNA

Two distinct methods were utilised within this project to quantify DNA methylation, namely Illumina HumanMethylation450 BeadChip (Section 2.4) and Pyrosequencing (Section 2.5). Both methods require a prior modification of the DNA samples to enable the distinction of methylated from unmethylated cytosine bases, which was achieved by bisulphite conversion.

The Zymo EZ DNA Methylation Gold TM Kit (Cambridge Biosciences, UK) which uses sodium bisulphite to convert unmethylated cytosine residues to uracil and subsequently allows the detection of methylated cytosines was used to bisulphite modify the DNA samples. 500ng of DNA was added to 130µl CT reagent and the following temperature steps were run on a thermocycler (SensoQuest GmnH, Germany): 98°C for 10 minutes, 64°C for 2.5 hours and held at 4°C for up to 20 hours. The DNA samples were added to Zymo-Spin TM IC columns (Cambridge Biosciences, UK) each containing 600µl of M-binding buffer. After a centrifugation step, 100µl of M-Wash Buffer was added to the columns followed by 200µl of M-Desulphonation Buffer, and 2x 200µl M-Wash Buffer. Each addition was punctuated with a centrifugation step. Bisulphite modified DNA was eluted in 15µl of M-Elution Buffer.

2.3.1 Quality control

To check that the bisulphite modification had proceeded to completion i.e. all unmethylated cytosines had been converted to uracil, a PCR specific for converted DNA and a PCR which amplifies unconverted DNA was performed.

The following primers were used:

Unconverted: *B2M* - Forward: CACTGAAAAAGATGAGTATGCC
Reverse: AACATTCCCTGACAATCCC
Converted: *IGSF9* - Forward: 5Biosg/TTAAGGTAGTTTTGGTA
Reverse: CATAAATCCAATAAATAATCTT

The PCR mastermixes and PCR conditions used were as follows:

The *B2M* mastermix was made as follows: 12.5µl 2x Hotstar TAQ Mastermix (Qiagen, UK), 5.5µl dH_2O, 0.5µl of 100pmol/µl forward primer, 0.5µl of 100pmol/µl reverse primer and 1.0µl BM DNA. The following temperature steps were performed on a thermocycler (SensoQuest GmbH, Germany): 95°C for 15 minutes, then 30 cycles of

95°C for 15 seconds, 62°C for 30 seconds and 72°C for 15 seconds, followed by 72°C for 5 minutes and held at 6°C.

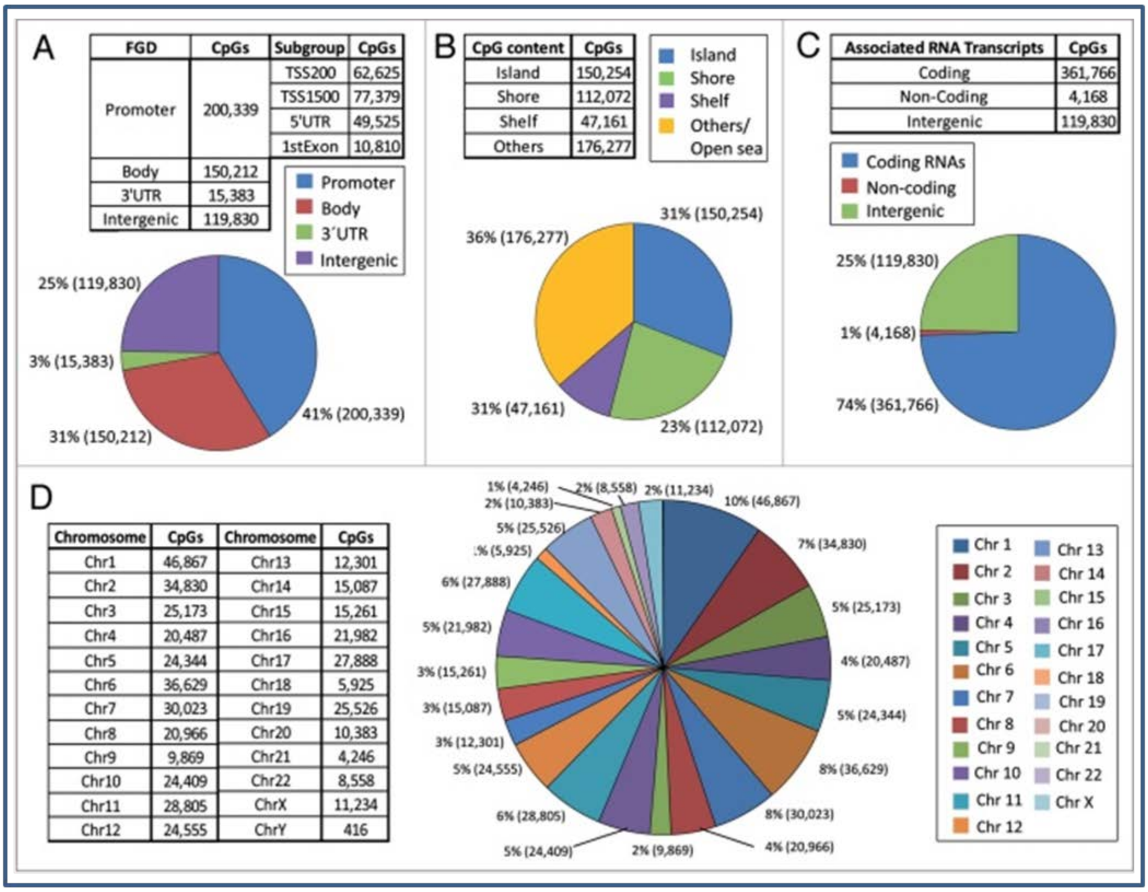
The *IGSF9* mastermix was made as follows: 12.5µl 2x Hotstar TAQ Mastermix (Qiagen, UK), 3.0µl of 25mM $MgCl_2$ (Qiagen, UK), 1.5µl dH_2O, 1.0µl of 100pmol/µl forward primer (biotin labelled) and 1.0µl of 100pmol/µl reverse primer and 1.0µl BM DNA. The following temperature steps were performed on a thermocycler (SensoQuest GmbH, Germany): 95°C for 15 minutes, then 45 cycles of 95°C for 15 seconds, 53°C for 30 seconds and 72°C for 15 seconds, followed by 72°C for 5 minutes and held at 6°C.

The PCR products were then run on a 1% agarose gel to confirm the bisulphite modification was complete in all samples. In the *B2M* reaction, only the unconverted DNA should be amplified whereas in the *IGSF9* reaction, only the converted DNA should be amplified.

2.4 Illumina HumanMethylation450 BeadChip sample preparation

Epigenome wide discovery analyses were performed in the first instance using the Illumina HumanMethylation450 (HM450) BeadChip. This microarray system simultaneously quantifies DNA methylation levels at 485,577 single base (cytosine) positions across the genome, covering 98.9% of UCSC RefGenes with multiple probes per gene (Bibikova *et al.*, 2011). Even though such a large number of CpG sites are included in the array this equates to only 1.7% of all CpG sites within the human genome (Kim *et al.*, 2012b). The vast majority (99.3%) of these cytosine loci are in CpG sites. 200,339 CpGs are located within gene promoters. 150,254 of these CpG sites are found within CpG islands. An even greater proportion of CpG sites are found within the open sea. 361,766 CpGs correspond to coding messenger RNA genes, only 4,168 CpG sites included in the array are non-coding and 119,830 and intergenic. The HM450 includes CpG sites from all 22 autosomal chromosomes as well as the two sex chromosomes. Chromosome 1 has the most CpG sites included in the array with 46,867 and unsurprisingly the Y chromosome has the fewest with only 416. Figure 7 shows an overview of CpG sites in relation to functional region, genomic location, associated RNA transcripts and chromosome location (Sandoval *et al.*, 2011).

The HM450 analyses twelve human DNA samples on each BeadChip. Each BeadChip consists of a number of beads with long target-specific probes to interrogate individual CpG sites. The HM450 uses two different Infinium probe chemistries. 135,501 Infinium I probes are present alongside 350,076 Infinium II probes. Infinium I probes have two separate beads per CpG locus; one which binds and extends along methylated DNA and the other which binds and extends along unmethylated DNA. Hence, they are tagged by and scanned using the same fluorescent dyes. The methylation status is deduced by the ratio of the signal intensities between them. Infinium II probes require just one bead type which can bind and extend along either methylated or unmethylated DNA. The distinction between them is made by different fluorescent dye tags. For instance, when bound to methylated DNA the primer is extended with the use of a green labelled dideoxynucleotide triphosphates (ddNTP) compared to a red labelled ddNTP for unmethylated DNA. Infinium II probes, due to the reduced space they require enable the HM450 platform to interrogate more CpG sites, however the use of two colour channels gives rise to the possibility of dye bias which needs to be adjusted for in data normalisation (Section 2.5.2).



FGD	CpGs	Subgroup	CpGs
Promoter	200,339	TSS200	62,625
		TSS1500	77,379
		5'UTR	49,525
		1stExon	10,810
Body	150,212		
3'UTR	15,383		
Intergenic	119,830		

CpG content	CpGs
Island	150,254
Shore	112,072
Shelf	47,161
Others	176,277

Associated RNA Transcripts	CpGs
Coding	361,766
Non-Coding	4,168
Intergenic	119,830

Chromosome	CpGs	Chromosome	CpGs
Chr1	46,867	Chr13	12,301
Chr2	34,830	Chr14	15,087
Chr3	25,173	Chr15	15,261
Chr4	20,487	Chr16	21,982
Chr5	24,344	Chr17	27,888
Chr6	36,629	Chr18	5,925
Chr7	30,023	Chr19	25,526
Chr8	20,966	Chr20	10,383
Chr9	9,869	Chr21	4,246
Chr10	24,409	Chr22	8,558
Chr11	28,805	ChrX	11,234
Chr12	24,555	ChrY	416

Figure 7: Description of CpG sites included on the HM450 BeadChip. A. Functional genomic distribution (FGD) and number of CpGs within each group. B. CpG content within genomic location. C. Associated RNA transcripts and number of CpGs. D. Number of CpGs within each chromosome. (Figure taken from Sandoval *et al.*, 2011).

Aliquots of 5µl bisulphite modified DNA were sent to the University of Bristol for processing and analysis using the Illumina HM450 BeadChip. The samples were first treated with a whole genome amplification step followed by the enzymatic fragmentation. The fragmented and re-suspended DNA samples were loaded onto the BeadChips which were then incubated at 48°C in the Illumina Hybridization Oven for 16-24 hours. Any unhybridised fragments were washed away and a TEM reagent used to extend the primers hybridised to DNA on the BeadChip, which was then stained. The BeadChips were then imaged on an Illumina iScan; a confocal laser scanning system (Illumina, USA). Data were output as raw intensity (.idat) files. Each DNA sample has two .idat files – one red and one green. Data were then taken forward for normalisation and analysis.

2.5 HM450 data pre-processing

2.5.1 Sample quality control

Data were exported from the iScan System in .sdf and .idat files and uploaded into the manufacturer's GenomeStudio software for quality control (QC) assessment. The HM450 BeadChip uses several types of controls for quality assurances including both sample-independent and sample-dependent controls. Sample-independent controls evaluate the quality of specific steps in the process. These include staining controls, extension controls, target removal controls and hybridisation controls. Sample-dependent controls allow the performance across samples to be evaluated. These include bisulphite conversion, specificity controls, non-polymorphic controls and negative controls. These controls are described in more detail below (Illumina User Guide v1.8).

The most basic are the staining and extension controls. Staining controls are used to examine the efficiency of the staining step, independent of the hybridisation and extension step. Extension controls are used to test the extension efficiency of A, T, C and G nucleotides from a hairpin probe.

Synthetic targets are present in the hybridisation buffer and complement the sequence on the array perfectly, thereby allowing the probe to extend on the synthetic target producing a signal. The synthetic targets are present at various concentrations – high, medium and low and therefore the beads should give off signals at various intensities. Signal intensity should increase as the synthetic target concentration increases. The hybridisation controls use these synthetic targets to test the overall performance of the entire assay.

The hybridisation buffer also contains target removal probes which test the efficiency of the stripping step after the extension reaction. Compared to the hybridisation controls, the signal of the target removal probes should be much lower, indicating that the targets were successfully removed following extension.

Two sets of controls are present to assess the efficiency of the bisulphite conversion step; one based on the Infinium I probe design and the other based on the Infinium II probe design. For the Infinium I probe design there are two probes. One binds to unconverted DNA and the other to converted DNA. Both are subsequently extended and signals detected using the same colour channel. For Infinium II, the converted and unconverted probes are extended and detected by distinct signals i.e. red and green. Ultimately, extension of either type of unconverted probe indicates the presence of unconverted DNA and an incomplete bisulphite conversion step.

There are also two types of specificity controls. The first, designed for Infinium I probes. Each chip includes a number of randomly permutated sequences or negative controls that should not hybridise to the DNA template. They are used to define system background.

Non-polymorphic controls are also included to test the overall performance of the assay. Assay performance is assessed across different samples with a particular base being queried in a non-polymorphic region of the bisulphite modified genome.

These controls can be visualised in the Controls Dashboard in GenomeStudio and those sample positions that do not meet the quality standard, as recommended by the manufacturer, were excluded from any further analysis.

Two data files were exported from GenomeStudio. The first included the methylation data as raw signal intensities (red and green) with no background correction or data normalisation, along with other measures relating to QC and probe design. The variables included are defined in Table 3. The second data file included the control data.

Variable	Definition
Detection PVal	The chance that the target sequence signal was distinguishable from the negative control
Avg_NBEADS	Average number of beads per probe for gene
TargetID	Probe name identifier
CHR	Chromosome on which target is found
Color_Channel	Colour channel
INFINIUM_DESIGN_TYPE	Assay type – Infinium I or Infinium II
Probe_SNPs	Assays with SNPs present within probe >10bp from query site
Probe_SNPs_10	Assays with SNPs present within probe<10bp from query site
UCSC_Refgene_Name	UCSC gene name
UCSC_Refgene_Accession	UCSC accession number
USCS_Refgene_Group	UCSC gene region category
Relation_to_UCSC_CPG_Island	Relationship to CpG island: shore or shelf
Mapinfo	Probe position for hg19
COORDINATE_36	Probe position for hg18
STRAND	Design strand

Table 3: Defined variables exported from GenomeStudio. (Adapted from Illumina User Guide v1.8).

2.5.2 Probe quality control and normalisation

In addition to assessing the quality of data at the sample level, HM450 data QC was also assessed at the probe level. In each of the following probe filtering steps, the default parameters set by Touleimat and Tost (2012) were used. Firstly, a detection p value threshold of 0.01 was set and individual samples were dropped if more than 20% of probes reported detection p values greater than this. Secondly, it is expected that on average fifteen beads per probe will be hybridised to the BeadChip in any given experiment. The greater the number, the more reliable the data becomes, as it is taken as an average across the beads. Consequently, if for any reason only a small number of beads hybridise to the chip, the resulting data could be considered less robust. For this reason, probes with less than three beads were dropped from the analysis. Finally, a number of probes were dropped due to analytical concerns. These include; allosomal probes, which due to sex biases require separate analysis (Touleimat and Tost, 2012); 65 probes that were included on the array to genotype SNPs rather than measure methylation but can be used for sample QC and mix-up detection; probes spanning genetic variants (minor allele frequency>5% in the Caucasian population) that may influence probe hybridization and thus methylation levels detected (Touleimat and Tost, 2012). Figure 8 reveals the number of probes dropped from analysis during the filtering steps.

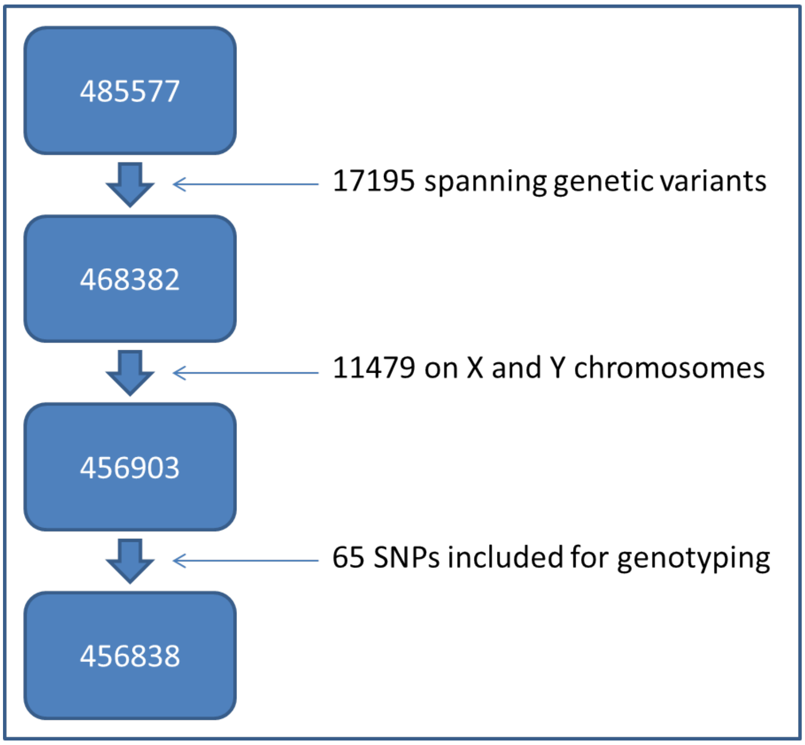

485577
17195 spanning genetic variants
468382
11479 on X and Y chromosomes
456903
65 SNPs included for genotyping
456838

Figure 8: Number of probes dropped at each stage during probe filtering.

As with all microarray data, data resulting from the HM450 BeadChip are susceptible to technical artefacts and batch effects. Technical variation refers to variation resulting from technical factors rather than biological differences between samples. A number of technical factors as well as batch effects (differences seen between groups of samples processed at different times or in separate sets) can affect HM450 data. Some of these effects may arise from differences in the handling of samples, such as DNA extraction or bisulphite modifications done in batches. Similarly, differences arising in the processing of DNA samples prior to being scanned, such as labelling and hybridisation can cause a batch effect. The location of the samples on the chip may also affect the level of staining experienced by each sample. Background noise can vary between different scans as well as other technical variation that arises due to the two probe chemistries included in the array and the use of two colour channels giving rise to probe type bias and colour dye bias, respectively (Sun *et al.*, 2011). Hence, a three step normalisation process was performed on the data following filtering and QC. Firstly, the signal intensities measured by negative controls were used to correct for background noise. Secondly, as the HM450 uses two colour channels, the intensities measured in the two colour channels might be imbalanced due to biases in labelling and scanning so colour correction was required to make these data comparable. Finally, an adjustment for probe type bias was performed, which is described below.

A number of normalisation packages have been developed for use with the HM450 data. Some of these methods only adjust for colour, others only for probe type and most do not enable any pre-filtering steps to be carried out. However, the HM450 analysis pipeline proposed by Touleimat and Tost (2012) which encompassed all of these steps was applied to the data analysed in this thesis. Following sample and probe filtering (as described above), background and colour correction, this package performs a probe type correction based on subset quantile normalisation (SQN). SQN is a method used to adjust for the two different probe types being used and enables data generated from them to be more comparable. There are two reasons SQN is performed. The first reason is that the probes capture CpG sites located randomly across the genome and, for biological reasons (1.3.3.2), sites mapping to similar genomic regions demonstrate distinct levels of methylation. For example, sites in CpG islands are generally unmethylated; sites in non-CpG islands are generally highly methylated. Secondly, the numbers of type I and type II probes mapping to these distinct regions vary greatly. This specific SQN process takes these genomic regions into account by performing quantile normalisation on subsets of probes based on their genomic location (e.g. CpG island or shore) rather than across the full 450K probes as a whole. Infinium I probes are used as anchor probes to estimate the reference quartiles to which the Infinium II probes are normalised. There are a group of anchor probes for each type of genomic region (Touleimat and Tost, 2012).

As a final step of pre-processing, any data points with a corresponding detection p value greater than 0.01 and not removed via sample filtering were dropped (i.e. converted to missing) in the data set. Any probes in which more than 90% of samples had missing data were dropped.

2.5.3 Methods used for data summary, visualisation and diagnostics

Graphs showing data before and after normalisation were plotted in R (version 2.15.0).

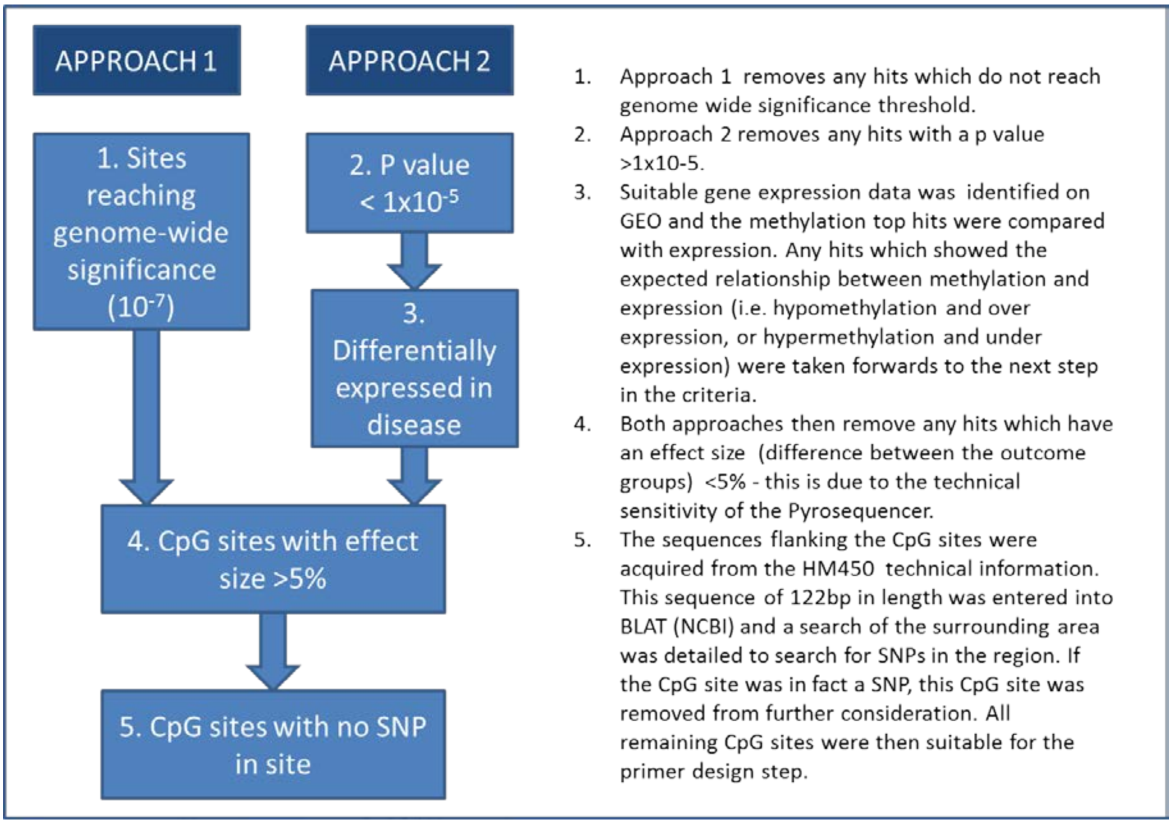
Density plots are similar to histograms but are presented as a continuous line rather than adjacent bars. These show the proportion of probes on the y-axis with a methylation value measured at the corresponding point on the x-axis. The shape of the line gives a representation of the methylation values across all >450K probes for each individual. Each sample is represented on the graph with a separate line. By overlaying the lines in this way it is easier to visualise and compare the methylation distributions across samples. If any samples stand out particularly, it suggests their distribution is somewhat different to the others. There may be a biological reason for this but this approach can also help to detect poor assays. Density plots were plotted in R.

Hierarchical cluster analysis was performed using the complete-linkage method which compares the most distant samples within each group to find and cluster with the closest group. The complete-linkage method was chosen over single-linkage due to the number of samples used in the analysis and the shape of the dendrogram produced. Cluster analysis was performed to identify natural groupings (either phenotypic or technical) amongst the samples based on the methylation data. All cluster analysis was performed in R.

2.6 Pyrosequencing

2.6.1 Selection of top hits

To select which CpG sites from the HM450 analysis to follow up, a ranking system was put in place to prioritise the sites most likely to be real associations and which could be selected for follow up. CpG sites showing differential methylation between the outcome measures were ordered by p value with the most statistically significant difference being ranked number one. Figure 9 outlines two different approaches taken to select top hits for further interrogation.


APPROACH 1
APPROACH 2
1. Sites reaching genome-wide significance (10^{-7})
2. P value < $1x10^{-5}$
3. Differentially expressed in disease
4. CpG sites with effect size >5%
5. CpG sites with no SNP in site

1. Approach 1 removes any hits which do not reach genome wide significance threshold.
2. Approach 2 removes any hits with a p value >1x10-5.
3. Suitable gene expression data was identified on GEO and the methylation top hits were compared with expression. Any hits which showed the expected relationship between methylation and expression (i.e. hypomethylation and over expression, or hypermethylation and under expression) were taken forwards to the next step in the criteria.
4. Both approaches then remove any hits which have an effect size (difference between the outcome groups) <5% - this is due to the technical sensitivity of the Pyrosequencer.
5. The sequences flanking the CpG sites were acquired from the HM450 technical information. This sequence of 122bp in length was entered into BLAT (NCBI) and a search of the surrounding area was detailed to search for SNPs in the region. If the CpG site was in fact a SNP, this CpG site was removed from further consideration. All remaining CpG sites were then suitable for the primer design step.

Figure 9: Two approaches used to select top hits for Pyrosequencing. Step 3 refers to the expected relationship between DNA methylation and gene expression in promoter regions (i.e. an inverse relationship) however it is recognised that this relationship is not true within gene bodies. Although this step has its limitations it was successful at reducing the number of top hits suitable for Pyrosequencing.

2.6.2 Primer design

Using the “View DNA” feature on the BLAST-like alignment tool (BLAT) genome browser, the 400bp upstream of the CpG site’s sequence and the 400bp downstream of the CpG site’s sequence were acquired. This 922bp genomic sequence was then copied and pasted into Microsoft Word to undergo *in silico* bisulphite modification. In many laboratory methods assessing DNA methylation, bisulphite modification is a key step. This treatment results in the deamination of unmethylated cytosine residues to uracil, whilst 5-methylcytosines remain unchanged. This effectively induces a SNP at unmethylated cytosines thereby fixing the methylation pattern allowing detection. For this reason, most downstream methylation assays require bisulphite modified DNA as the input material. The *in silico* bisulphite modification step is a simple but important step of the process as it is bisulphite modified DNA which is required to design primers for subsequent PCRs. The aim of this step is to replace all unmethylated cytosines with uracil but leave methylated cytosines unchanged. By merely looking at a genomic DNA sequence it is impossible to know which CpG sites are methylated so the first step is to protect all potentially methylated CpG sites from undergoing deamination by replacing the CG in the sequence with XG. As methylation only occurs at CpG sites, all other cytosine residues in the DNA sequence are deaminated, meaning all C residues are replaced with T, depicting the real base change occurring when genomic DNA is treated with sodium bisulphite. The final step is to replace all XG residues (the CpG sites) with C/TG. This change represents the two possible alleles (C or T) which may be present depending on the methylation status of each CpG site in the sequence. These bisulphite modified sequences which encompass the CpG site of interest were then analysed using the PSQ Assay Design software (Qiagen, UK). The entire sequence was inputted into the software and the target region was manually set as the CpG site included on the HM450 array. The software then tried to find suitable forward, reverse and sequencing primers for each CpG site. Wherever possible, primers were designed not only for the CpG site included on the array but for at least one additional CpG site either side of the target CpG site. All default parameters were used. The software gives a score ranging from 0 to 100 for each primer set. A score of 0 means that no primers were available for the selected CpG site/s. A score close to 100 means that the primers are very likely to be successful in amplifying the DNA and providing a robust Pyrosequencing assay. In this study, no primers were taken forward that had a score of less than 70. The success rate in finding suitable primers was around 50%. All suitable primers generated by the assay design software were ordered for synthesis by Integrated DNA Technologies (IDT, USA).

2.6.3 Primer optimisation

To test for the optimum PCR conditions, two Pyrosequencing PCR Mastermixes were made up for each assay; one containing $MgCl_2$ and one excluding $MgCl_2$. The Mastermix including $MgCl_2$ was made as follows: 6.25µl Hotstar TAQ Mastermix

(Qiagen, UK), 1.5µl of 25mM $MgCl_2$ (Qiagen, UK), 2.75µl dH_2O, 0.5µl of 100pmol/µl forward primer and 0.5µl of 100pmol/µl reverse primer (one of which is biotin labelled). The Mastermix excluding $MgCl_2$ was made as follows: 12.5µl Hotstar TAQ Mastermix (Qiagen, UK), 4.25µl dH_2O, 0.5µl forward primer and 0.5µl reverse primer (one of which is biotin labelled). 1µl of pooled BM DNA was added to 11.5µl of Mastermix in each well. A gradient PCR spanning 50-60°C was also run to determine the optimum annealing temperature (Tm). The PCR was run on a thermocycler (SensoQuest GmbH, Germany) under the following conditions: 95°C for 15 minutes, then 45 cycles of 95°C for 15 seconds, gradient for 30 seconds and 72°C for 15 seconds, followed by 72°C for 5 minutes and held at 6°C. The composition (with/without MgCl2) and temperature which produced the strongest band on a 1% agarose gel was selected for subsequent PCRs.

2.6.4 Agarose gel

0.7g agarose was dissolved in 70ml Tris-Borate-EDTA (TBE) buffer to make a 1% agarose gel. 4µl of SafeView (NBS Biologicals, UK) was added and poured into a gel tray, combs inserted and left to set. 5µl loading buffer and 5µl of BM DNA was loaded into each well, with 4µl of Gene Ruler^{TM} 100bp DNA Ladder (Fermentas, International) loaded at either side of the gel.

2.6.5 Assay validation

Epitect PCR controls (Qiagen, UK) containing a 100% methylated DNA sample and a 0% methylated DNA sample were used to make a range of reference methylation levels ranging from 0% to 100%. To provide an accurate validation and confirmation of no preferential amplification of either methylated or unmethylated DNA, samples were mixed both pre-PCR and post-PCR. The following tables show how the samples were produced for the pre-PCR mix (Table 4) and post-PCR mix (Table 5). All samples were run in duplicate.

Methylation (%)	**Amount of methylated (100%) Epitect control (µl)**	**Amount of unmethylated (0%) Epitect control (µl)**
95	34.9 of 100%	1.9
90	26.8 of 95%	1.5
75	18.3 of 90%	3.7
50	12 of 75%	6
25	8 of 50%	8
10	6 of 25%	9
5	5 of 10%	5

Table 4: Pre-PCR mix.

The same conditions were used for the PCR mastermix and Pyrosequencing PCR as previously described.

Methylation (%)	**Amount of methylated (100%) Epitect control (µl)**	**Amount of unmethylated (0%) Epitect control (µl)**
95	87.4 of 100%	4.6
90	67 of 95%	3.75
75	45.75 of 90%	9.25
50	30 of 75%	15
25	20 of 50%	20
10	15 of 25%	22.5
5	12.5 of 10%	12.5

Table 5: Post-PCR mix.

2.6.6 Pyrosequencing

In each well of a PCR plate, 2µl of Streptavidin Sepharose beads (GE Healthcare Life Sciences, UK), 38µl of binding buffer (Qiagen, UK) and 20µl of H_2O were added along with 10µl of BM PCR product. The plate was mixed on a plate shaker (Grant Instruments, UK) vigorously for at least 5 minutes. To each well on the Pyrosequencing plate, 0.5µl 10uM sequencing primer and 11.5µl annealing buffer were added. Samples were then cleaned up to single stranded DNA using the Vacuum Prep Workstation (Qiagen, UK). The vacuum block tool was used to remove the PCR product and bead mix (biotin labelled primer bound to bead) from the PCR plate, before a rinse with ethanol to remove any unwanted soluble PCR components. Double stranded DNA was then denatured to single stranded DNA (ssDNA) by aspirating denature buffer and the probe block was then placed in a wash buffer to clean the ssDNA. The vacuum was then switched off and DNA bound to beads was then deposited in the Pyrosequencing plate containing the annealing buffer and sequencing primer. The Pyrosequencing plate was then incubated at 80°C for 2 minutes to allow the primers to anneal and then placed in the Pyrosequencer, which had previously been equipped with appropriate amounts of an enzyme mix, a mixture of substrates and deoxynucleotide triphosphates. Components of each reagent are listed in Table 6. By using a series of algorithms to calculate the most efficient use of bases to complete the sequence, the Pyrosequencing software generates a nucleotide dispensation sequence. This dispensation order includes a number of 'spike' nucleotides which are randomly placed to ensure the sequence being assayed conforms to the input sequence. The samples were then run in duplicate on a Pyromark MD Pyrosequencer (Qiagen, UK). A negative control (containing no DNA) along with a 100% and 0% methylation control were included in every run. Methylation values are displayed as percentage methylation. The Pyrosequencer cannot accurately detect methylation differences less than 5% (Mikeska *et al*., 2011), therefore the duplicate pairs must be within 5% of one another. Any pairs discordant by 5% were repeated. The dispensation order for each validated assay is provided in Appendix A.

Pyrosequencing Reagents	Components
Binding buffer	10mM Tris-HCL;2M NaCl, 1mM EDTA, 0.1% Tween^{TM} 20
Denaturation solution	0.2M NaOH
Wash buffer	10mM Tris-Acetate
Annealing buffer	20mM Tris-Acetate, 2mM Mg-Acetate
Enzyme Mixture	DNA polymerase, sulfurylase, luciferase, apyrase
Substrate Mixture	Adenosine 5’phosphosulfate, luciferin

Table 6: Components of each Pyrosequencing reagent.

2.6.7 Pyrosequencing quality control

In addition to running samples in duplicate, repeating any samples that exceeded the 5% discordance threshold and running a negative control (containing no DNA) as already mentioned, an additional QC step was taken. As Pyrosequencing measures DNA methylation using bisulphite modified DNA sequences, it was essential to ensure that the bisulphite conversion step proceeded to completion. Where possible, each assay included at least one bisulphite treatment control in the dispensation sequence. As cytosines which are not part of a CpG site should be converted to thymine residues following bisulphite treatment, there should be no non-CpG site cytosine residues in the dispensation sequence. By inserting a cytosine residue into the dispensation sequence, it is possible to check whether any unconverted cytosines are still present in the bisulphite modified DNA as a result of an unsuccessful/incomplete bisulphite conversion.

2.7 Other laboratory techniques

A number of experiments were performed to assess the functional significance of any top hits successfully validated via Pyrosequencing. These include quantification of RNA and protein and subsequent western blots.

2.7.1 Tissue RNA extraction

Sections of DLPFC and hippocampus were provided from the NBTR for RNA extraction. To extract RNA from the brain tissue the E.Z.N.A.$^{®}$ Tissue RNA Kit (OMEGA bio-tek, USA) was used and the standard Tissue RNA Spin protocol was followed. The extraction protocol was performed on DLPFC pieces weighing approximately 30mg and hippocampus sections weighing approximately 15mg. In a containment level 2 tissue culture facility, each tissue section (30mg/15mg) was added to 300µl TRK Lysis Buffer which contained 20µl 2-mercaptoethanol per 1ml TRK. Samples were mixed on a thermoshaker (Grant Instruments, UK) until all tissue had lysed. Following lysis, 590µl RNase-free water and 10µl OB protease was added and samples were incubated at 55°C for 10 minutes on a thermoshaker (Grant Instruments, UK). Following a 5 minute centrifugation step at 14000xg, the supernatant was transferred to a clean 1.5ml microcentrifuge tube and was mixed with 450µl absolute ethanol. 700µl sample was transferred into a HiBind RNA Spin Column and was centrifuged at 10000xg for 1 minute. This centrifugation step was repeated until the entire sample had been transferred to the column. A series of three wash steps then followed; 1x 700µl of RNA wash buffer I, 2x 500µl RNA wash buffer II. Each wash step was punctuated by a centrifugation step. The column was then centrifuged at 13000xg for 2 minutes to completely dry the membrane. Samples were eluted in 40µl DEPC-treated water and stored at -70°C.

2.7.2 Blood RNA extraction

RNA was extracted from frozen blood using the standard Isolation of Total RNA from Blood Protocol as outlined in the E.Z.N.A.$^{®}$ Blood RNA Kit (OMEGA bio-tek, USA). To 150µl of whole blood, 750µl ERL buffer was added and incubated on ice for 15 minutes or until lysis of red blood cells was complete and the solution became translucent. Samples were then centrifuged at 400xg at 4°C for 10 minutes to pellet the leukocytes and the supernatant was removed and discarded. Cells were then re-suspended in 300µl ERL buffer. Samples then underwent a second centrifugation step at

400xg at 4°C for 10 minutes and again the supernatant was discarded. White blood cells were lysed in 400µl NTL lysis buffer, containing 20µl 2-mercaptoethanol per 1ml NTL and the entire sample was transferred to a Homogeniser column. After a centrifugation step at 13000xg for 2 minutes, the filtrate was added to 400µl 70% ethanol. 700µl of sample was transferred to a HiBind RNA column and centrifuged at 10000xg for 30 seconds. This centrifugation step was repeated until the entire sample had been transferred to the column. A series of wash steps followed. First, 500µl RWF wash buffer and 2x 700µl of Wash buffer II, each step punctuated by a centrifugation step at 10000xg for 30 seconds. Columns were then centrifuged at 13000xg for 2 minutes to completely dry the column membrane. Samples were eluted in 50µl DEPC-treated water and stored at -70°C.

2.7.3 Quantification of RNA

RNA samples were quantified using a ND1000 Spectrophotometer (Labtech International Ltd, UK). The ND1000 was first initialised with 1.5µl of distilled H_2O and a blank measurement was taken using 1.5µl of DEPC-treated water. RNA samples were also analysed using an Agilent 2100 Bioanalyser. 36 samples were analysed across 3 chips. The RNA samples were heat denatured for 2 minutes at 70°C and immediately snap-frozen along with the RNA ladder. 1µl of dye concentrate was added to 65µl of filtered gel matrix. After decontaminating the electrodes with both RNaseZap and RNase-free water, the chip was loaded into the chip priming station. 9µl of gel dye mix was loaded into the well marked $\boxed{G}$. The plunger was then compressed and held for 30 seconds. After releasing the plunger, 9µl of the gel-dye mix was loaded into wells. 5µl of the NanoMarker was added into each of the 12 sample wells as well as the ladder well. In addition, 1µl of ladder was loaded into the ladder well and 1µl of each sample was loaded into each of the samples wells. After vortexing the chip to ensure thorough mixing, the chip was loaded into the Agilent 2100 Bioanalyser and the run was initiated. This process was repeated for each chip and was cleaned with RNaseZap and distilled water after the last run was complete.

2.7.4 Protein extraction

100µg DLPFC and 4x100µM sections of hippocampus were homogenised in 400µl extraction buffer (50mM Tris-HCl, 5mM EGTA, 10mM EDTA, pH 7.4) (Kirvell *et al.*, 2010), containing a protease and phosphatase inhibitor (Thermo Scientific, USA), in a

Precellys 24 homogeniser (Bertin Technologies, France). After 3 x 20 second cycles in the Precellys 24, samples were then spun in a centrifuge at 13000rpm for 30 minutes and the supernatant was carefully removed and stored at -20°C.

2.7.5 Protein quantification

Proteins were quantified using the Thermo Scientific™ Pierce™ BCA™ Protein Assay (Thermo Scientific, USA). Standards (n=9) were made by mixing Bovine Serum Albumin (BSA) of a known concentration with Phosphate Buffered Saline (PBS) as shown in Table 7.

Vial	Volume of diluent (µL)	Volume and source of BSA (µL)	Final BSA Concentration (µg/mL)
1	0	30 of stock	2000
2	12.5	37.5 of stock	1500
3	32.5	32.5 of stock	1000
4	17.5	17.5 of vial 2	750
5	32.5	32.5 of vial 3	500
6	32.5	32.5 of vial 5	250
7	32.5	32.5 of vial 6	125
8	40.0	10 of vial 7	25
9	40.0	0	0

Table 7: Dilution factors for standards to be used in protein quantification.

25µl of each standard and 5µl of protein sample were placed into a well of a clear 96 well microwell plate (Beckman Coulter, USA). 200µl of WR (50 parts BCA Reagent A: 1 part BCA Reagent B) was added into each well containing standard or protein sample. In addition 200µl of WR was added into a blank well containing 25µl of PBS. The plate was mixed for 30 seconds, covered and incubated at 37°C for 30 minutes. Once cooled, the plate was placed in a FLUOstar Omega plate reader (BMG Labtech, Germany) and absorbance was measured at 562nm. The absorbance of the blank well was subtracted from all wells and a standard curve was calculated which was then used to calculate the concentrations of the unknown protein samples using the Omega Mars software (BMG Labtech, Germany). Protein samples were then standardised to 25µg and made up to 15µl using the extraction buffer described in Section 2.7.4. An equal mix of all DLPFC samples and hippocampal samples were made to be used as standards in each Western blot.

2.7.6 Western blots

2.7.6.1 *Antibody optimisation*

To assess the optimal blotting conditions for the primary antibodies, four different conditions were tested shown in Table 8.

Conditions tested	**Blocking Agent**	**Primary Antibody Diluent**
1	5% milk	5% milk
2	5% milk	1% milk
3	5% BSA	5% BSA
4	5% BSA	1% BSA

Table 8: Four different primary antibody conditions.

The conditions which resulted in a strong band with little background were selected as the optimum conditions for the primary antibody.

2.7.6.2 *Western blot*

Firstly, all the buffers required for the Western blot were made, as outlined in Table 9.

Buffer	**Ingredients**
1L 10X Running Buffer	30.3g Tris base, 144g glycine and 10g SDS
1L 10X Transfer Buffer	144g glycine and 30.3g Tris base
1L 10X TBS Tween 20 (TBST)	12.11g Tris base, 29.22g NaCl and 10ml Tween 20

Table 9: Ingredients for all buffers used in Western blotting.

2.7.6.2.1 Acrylamide gel

90ml 1X running buffer was added to each Amersham ECL Gel Box System (GE Healthcare Life Sciences, UK) and the precast Amersham ECL 12% gel (GE Healthcare Life Sciences, UK) was rinsed, inserted into the gel tank and pre-run at 160V for 12 minutes. Meanwhile, 15µl protein sample was mixed with 15µl Laemmli sample buffer (Sigma-Aldrich, USA) and heated to 96°C for 5 minutes using a thermocycler (SensoQuest GmbH, Germany). Once the pre-run was complete, the well comb was removed and 6ml 1x Running buffer was added into the wells. 25µl of the protein and loading buffer mix was then added into each well. On each gel, 10µl SpectraBr ladder (Fermentas, International) was added in one of the wells along with 25µl of the two standards described in Section 2.7.5. The gel was then run at 160V for 60 minutes.

2.7.6.2.2 Activation of the membrane

The membrane was cut to size and covered in 100% methanol for 1 minute. The membrane was then equilibrated in 1X Transfer buffer until required.

2.7.6.2.3 Blotting

All equipment (transfer tank, cassette, sponges and filter paper) used in the protein transfer were soaked in 1X transfer buffer. Once the acrylamide gel was run, the blotting cassette was opened and a sponge placed on one side. Blotting paper was then added onto this sponge. The gel was removed from the cast and the stacking gel was cut off. The gel was then added onto the filter paper and the membrane placed on top of the gel. An additional piece of filter paper and a sponge was added onto the transfer sandwich and a glass rod was used to remove any air bubbles that may impede the transfer of proteins from the gel onto the membrane. The cassette was then closed and added into the transfer tank. The tank was filled with 1X transfer buffer and covered in ice to ensure the transfer buffer was kept as cold as possible. The transfer was run at 100V for 1 hour.

2.7.6.2.4 Blocking with 5% milk

5% milk (low-fat skimmed milk powder) was made up in 1X TBS-tween and the membrane was incubated in the 5% milk solution for 1 hour.

2.7.6.2.5 Blotting with antibodies

The membrane was then incubated with anti-NGF antibody (1:200 in 5% milk, rabbit polyclonal IgG) (SantaCruz Biotechnology) for 1 hour. The membrane was then washed in TBST 3x 5 minutes. The membrane was then incubated with anti-rabbit IgG antibody conjugated with horseradish peroxidase (HRP) (1:1000 in 1% milk) (Cell Signalling Technology, USA) for 1 hour, followed by 3x 5 minute washes in TBST.

2.7.6.2.6 Imaging of membrane

Equal parts of A and B taken from the Amersham ECL Prime Western Blotting Detection Reagent (GE Healthcare Life Sciences, UK) were mixed and washed over the membrane. The membrane was then covered in a thin plastic sheet and imaged using a Syngene G:Box camera (Syngene, UK) and Genesys software (Syngene, UK).

2.7.6.2.7 Loading standard

To strip the membrane of the anti-NGF antibody the membrane was covered with Restore Western Blot Stripping Buffer (Thermo Scientific, USA) for 15 minutes. The membrane was then blocked with 5% milk and left at 4°C overnight. The membrane was then incubated with anti-alpha-tubulin (α-tubulin) antibody (which is used as a loading standard) (1:1000 in 1% milk, mouse monoclonal IgG) (Sigma-Aldrich, USA) for 1 hour. 3x 5 minute washes in TBST then followed. The membrane was then incubated in the secondary anti-mouse IgG antibody conjugated with HRP (1:1000 in 1% milk) (Dako, Denmark) for 1 hour and washed 3 times for 5 minutes each in TBST. Again equal parts of A and B from the Amersham ECL Prime Western Blotting Detection Reagent (GE Healthcare Life Sciences, UK) were mixed and used to cover the membrane. A plastic film was used to cover the membrane and imaged using a Syngene G:Box camera (Syngene, UK) and Genesys software (Syngene, UK).

2.8 Statistical methods

2.8.1 DNA methylation measures

2.8.1.1 *Methylation as a beta value*

Following normalisation of the Illumina HM450 data, a measure of methylation was given as a beta value, ranging from 0 to 1. The beta value is the ratio of methylated probe intensity over the sum of the methylated and unmethylated probe intensities. The equation is shown below.

$$Beta\ value = \frac{methylation\ probe\ intensity}{methylation\ probe\ intensity + unmethylated\ probe\ intensity}$$

A beta value of 0 suggests that every copy of the CpG site is unmethylated (this can be interpreted as 0% methylated) and a value of 1 suggests that every copy of the CpG site is methylated (often interpreted as 100% methylation) (Du et al 2010).

2.8.1.2 *Methylation as a percentage*

The Pyrosequencer generates methylation values as percentage methylation ranging from 0% to 100%. 0% indicates that none of the copies of the CpG site are methylated whilst a value of 100% indicates that all copies of the CpG site are methylated. A value of 50% indicates that half of all copies of the CpG site are methylated whilst the other half of copies are unmethylated.

2.8.2 HM450 data analysis

The CpGassoc package (Barfield *et al*., 2012) which performs multiple linear regression analyses with continuous predictor variables and ANOVA for categorical predictor variables was implemented in R (version 2.15.0). For all analyses, age, sex and chip (i.e. the microarray chip upon which the samples were scanned) were included in the analysis model as covariates. The outcome variables for COGFAST and ICICLE are outlined in Glossary Tables 2 and 4, respectively.

2.8.2.1 *Cell type adjustment*

2.8.2.1.1 Blood

HM450 data were adjusted for cell composition using an algorithm described by Houseman *et al.* (2012). This algorithm is able to estimate the proportion of immune cells in unfractionated whole blood samples without the need to directly count proportions within fresh samples. The authors first identified the DNA profiles of each of the principal immune components of blood (B cells, granulocytes, monocytes, NK cells, T cell subtypes), then developed a *regression calibration* tool, which can be used to estimate cell distributions and act as a surrogate for cell distribution in subsequent analyses (Houseman *et al.*, 2012). HM450 data analyses (i.e. CpGassoc) were repeated after adjustment for cell type to assess whether the DNA methylation differences were immunologically mediated.

2.8.2.1.2 Brain

HM450 data were adjusted for cell composition using CETS, an R package described by Guintivano *et al.* (2013). This package is able to estimate the proportion of neurons and glia in bulk brain tissue without the need to directly count proportions within fresh samples using Fluorescence Activated Cell Sorting (FACS) or laser capture microdissection. The authors first identified the DNA profiles of neurons and glia using prefrontal cortical tissue and then developed an algorithm, which can be used to estimate cell distributions and act as a surrogate for cell distribution in subsequent analyses (Guintivano *et al.*, 2013). To best match the COGFAST cohort, the CETS package was modified to only include cell proportion data from healthy Caucasians. HM450 data analyses (i.e. CpGassoc) were repeated after adjustment for cell type to assess whether the DNA methylation differences were associated with cellular composition.

2.8.2.2 *Calculating methylation scores to predict smoking status*

Methylation values published by Zeilinger *et al.* (2013) were used as a reference data set (Zeilinger *et al.*, 2013). Methylation at 183 CpG sites was used to calculate weighted methylation scores for each DNA sample. Median methylation values of never smokers identified by Zeilinger *et al.* (2013) were used as the reference values and the associated effect sizes were used as weights. The difference between the ICICLE beta value and reference beta value was calculated for each CpG site. The sum of all CpG site scores was calculated to give a final weighted score. The threshold score able to distinguish current from never smokers was calculated using random forests (Breiman, 2001) in R (Version 3.1.0).

2.8.2.3 *Predicting age using DNA methylation*

The age of each blood and brain sample was predicted using the algorithm described by Horvath (2013). This script uses the methylation values at 353 CpG sites (measured on the HM450) to predict the age of each sample. In addition to predicted age, the algorithm also provides the age acceleration difference (difference between DNA methylation age and chronological age) and the age acceleration residual (residual from regressing DNA methylation age on chronological age, i.e. DNA methylation age corrected for chronological age).

2.8.3 Pyrosequencing data analysis

For all data passing QC, a mean methylation value for the duplicates was calculated and was used for all Pyrosequencing analyses. Pyrosequencing data were analysed using ANOVA with age and sex as covariates. All analysis was performed in Stata (Version 11.2).

2.8.4 Western blot analysis

Band densitometry was utilised to calculate relative protein levels across multiple blots. The ImageJ software (National Institutes of Health, USA) is able to generate values for band intensity in gel lanes. For each gel, the NGF and α-tubulin blots were aligned and opened in ImageJ software (National Institutes of Health, USA). The rectangular selections tool was used to define lanes and these were opened in the profile plot analysis. Background was eliminated using the straight line tool and the wand tool was used to calculate the area under curve. The results were exported into Microsoft Excel. The values calculated for each antibody exposure were normalised to the mean of the standards loaded onto the gel.

The relative value for NGF was calculated using the following:

$$Relative\ NGF = \frac{Normalised\ NGF}{Normalised\ \alpha{-}tubulin}$$

Chapter 3. Blood-based methylation signatures as a predictor of post-stroke dementia

3.1 Introduction

Post-stroke dementia (PSD) is a disabling consequence of a stroke, affecting approximately a third of stroke survivors (Pohjasvaara *et al.*, 1998; Allan *et al.*, 2011). Due to the increasingly ageing population and consequent rising prevalence of PSD, the associated healthcare costs are becoming a substantial burden to public healthcare services. In 2010, dementia cost the UK National Health Service over €22 billion and this figure is likely to be higher today (Fineberg *et al.*, 2013). A key hurdle in the treatment of PSD is that we do not yet understand the mechanisms that follow stroke and which culminate in this disease. It is unclear why some stroke survivors develop dementia whilst others remain cognitively normal. It would therefore be beneficial to develop a method of predicting which stroke survivors are likely to develop dementia, so that early preventative treatments could be implemented for those individuals at highest risk.

DNA methylation is an epigenetic mark closely involved in gene regulation and variation in DNA methylation is known to be associated with an array of complex phenotypic traits and common human diseases including cancer (Teschendorff *et al.*, 2009), cardiovascular disease (Baccarelli *et al.*, 2010) and metabolic disorders (Grundberg *et al.*, 2013). Epigenetic mechanisms, including altered DNA methylation patterns, have also been implicated in increased risk of both stroke (Endres *et al.*, 2000; Castillo-Diaz *et al.*, 2010) and dementia (Wang *et al.*, 2008; Bollati *et al.*, 2011; Bakulski *et al.*, 2012a; Chouliaras *et al.*, 2013). Although there are several reliable associations between DNA methylation and disease, a causal role of DNA methylation has not been demonstrated in many instances. It is conceivable that DNA methylation patterns, either pre-existing or resulting from the stroke itself, may contribute towards the long term prognosis in stroke survivors, including the variable and unpredictable pathogenesis of PSD in some survivors but not others. This mechanistic link requires the analysis of the diseased tissue (in this case the brain) (see Chapter 4 for further exploration of this issue). An alternative, complementary potential use of DNA methylation information is as a robust biomarker to predict disease or to aid prognosis. This does not require a causal relationship to be established, nor does it require the target disease tissue to be studied. A surrogate, more readily accessible source of DNA, such as peripheral blood, can be used for this application. This is explored within this Chapter.

DNA methylation marks, measured in both a global and gene-specific manner are beginning to emerge as biomarkers of disease risk and prognosis, particularly within the field of cancer. For instance, distinct DNA methylation profiles in tissue-specific and peripheral blood samples have been shown to be predictive of bladder (Marsit *et al.*, 2011), ovarian (Teschendorff *et al.*, 2009) and lung cancer (Dietrich *et al.*, 2012) as well as disease subtype in patients with medulloblastoma (Schwalbe *et al.*, 2013). DNA methylation patterns have not only been associated with risk of disease development but also have been found to have prognostic value in cancer. In oesophageal adenocarcinoma patients, increased methylation of seven cancer-related genes was associated with reduced survival and tumour recurrence (Brock *et al.*, 2003). The hypomethylation of one CpG site within *ZAP-70* has been identified as a reproducible biomarker for poor prognosis in chronic lymphocytic leukaemia (Claus *et al.*, 2012).

This approach has yet to be applied in PSD. In this Chapter I have postulated that DNA methylation signatures, detectable in peripheral blood drawn after a stroke, will be predictive of subsequent cognitive decline and possibly overt dementia. To test this hypothesis, the prospectively collected COGFAST study is used (Section 2.1.1). This is a prospective study composed of individuals who have suffered a stroke, had biological samples collected and stored at this time and that have been followed up over several years with cognitive assessments undertaken at multiple intervals.

3.2 Experimental design

General methods are described in Chapter 2, with additional details relevant to this investigation described in detail here.

3.2.1 Study cohort

The COGFAST study recruited 355 stroke survivors. A subset of 46 blood samples from the cohort (Section 2.1.1), all of whom were stroke survivors over the age of 75 at the time of stroke were utilised for this study. A diagnosis of dementia was made based on neuropsychometric testing and agreement with DSM IIIR and IV criteria for dementia, and confirmed by post-mortem (PM) neuropathological examination. To date, 31 COGFAST subjects who came to autopsy have been diagnosed with dementia, with a similar number of individuals who were not demented at death. Data relating to clinical history, including details of stroke severity as well as stroke risk factors were available for analysis. Variables relating to the stroke event such as Oxfordshire Community Stroke Project score (OCSP), side of body affected and degree of weakness were collected at the time of stroke. Age at death, PM delay, Braak staging and years between stroke and death were collected at death. All other variables were collected at baseline (recruitment into study) and are described in the Glossary (Table 2). Samples used for the discovery analysis were selected from deceased participants, to enable neuropathological information to be utilised in the analysis. Blood samples were collected three months post-stroke, when all participants were cognitively normal, with a mini mental state examination (MMSE) score >24. Samples were divided into a ‘discovery’ sample (n=30) and ‘Pyrosequencing’ sample (n=46), the latter consisting of the discovery sample and an additional sixteen samples all drawn from the same cohort.

3.2.2 Blood DNA extraction, quantification and precipitation

Blood samples were held in long-term storage at -80°C and were selected and removed from storage for the purposes of this study. DNA was extracted from 3ml blood using the standard Blood and Body Fluid Protocol as outlined in the EZNA Blood DNA Kit (OMEGA bio-tek, USA) (Section 2.2.2). DNA samples were quantified using a ND1000 Spectrophotometer (Labtech International Ltd, UK) (Section 2.2.3). Samples with a concentration lower than 50ng/µl were ethanol precipitated and re-suspended in 30µl DEPC-treated water (Section 2.2.4) before requantification.

3.2.3 Illumina HumanMethylation450 BeadChip analysis

Epigenome wide discovery analysis was performed using the Illumina HumanMethylation450 (HM450) BeadChip. Thirty individuals with >1µg blood derived DNA were selected for discovery phase DNA methylation analysis. These 30 samples included 17 participants diagnosed with dementia (D) and 13 cognitively normal (CN) participants, who were matched on sex and age within three years. 500ng of DNA from each sample was treated with sodium bisulphite, to convert unmethylated cytosine residues to uracil and subsequently allow the detection of methylated cytosines, using the Zymo EZ DNA Methylation Gold^{TM} Kit (Cambridge Biosciences, UK). Bisulphite modified (BM) DNA was eluted in 15µl of M-Elution Buffer (Section 2.3). To check that the bisulphite modification had been successful, a PCR specific for BM DNA and a PCR which amplifies unconverted DNA was performed (Section 2.3.1). The PCR products were run on a 1% agarose gel to confirm that the BM conversion was successful. Aliquots of 5µl BM-DNA were analysed using the Illumina HM450 BeadChip (Section 2.4) at the MRC Integrative Epidemiology Unit, University of Bristol. Data were returned as .idat files containing raw signal intensity values from the array. These files were uploaded into GenomeStudio (Illumina, USA) for sample QC. A number of controls included on each BeadChip were used to assess the quality of the samples (Section 2.5.1). Methylation data were extracted as beta values, with no background correction or data normalisation. Beta values are methylation scores for each CpG, ranging from 0 (unmethylated) to 1 (methylated) and calculated as the ratio of methylated probe intensity over the sum of the methylated and unmethylated probe intensities.

The analysis pipeline proposed by Touleimat and Tost (Touleimat and Tost, 2012) was implemented in R (version 2.15.0) to filter probes and samples and to normalise the beta values. The details of the normalisation are provided in Section 2.5.2.

3.2.4 Analysis of HM450 data

A total of 455,173 probes and 29 samples were included in the final dataset. Exploratory analyses were performed across the methylation data, by considering distribution density plots, mean-sd plots and hierarchical clustering with heat plots. The CpGassoc package (Barfield *et al.*, 2012), which performs multiple linear regression analyses with a continuous predictor variable and ANOVA for categorical predictor variables, was implemented in R (version 2.15.0). CpGassoc was performed to test for associations between methylation and diagnosis (D/CN), the last MMSE and CAMCOG scores before death and Braak staging as PSD outcome variables. For all analyses, baseline age, sex and chip (i.e. the microarray chip upon which the samples were scanned) were included in the analysis model as covariates. Data generated provided a beta value difference between the comparison groups (for categorical variables) and a beta value per unit change in the predictor variable (for ordinal and continuous variables). The beta values had an associated p-value denoting the strength of statistical significance with standard errors denoting the confidence around the statistical estimate of association. To test for sensitivity to cellular composition, HM450 data were adjusted for cellular composition using the method described by Houseman *et al.* (2012). Statistical analyses were then repeated to assess whether cell composition accounted for any significance observed between methylation and the outcome variable.

3.2.5 Selection of top HM450 hits

To select which CpG sites were the most promising candidates to validate, two approaches were taken. CpG sites showing differential methylation between the outcome measures were ordered by p-value with the most significant differences being ranked first. In the first approach, p values above the cut off imposed for multiple test correction (0.05/455,173) were not considered. The second approach included all CpG sites with a p value $<1x10^{-5}$ (to include more hits) and compared methylation data with publicly available expression data. The Gene Expression Omnibus (GEO) website (NCBI) (http://www.ncbi.nlm.nih.gov/geo/) was searched for an appropriate expression dataset with which to compare the methylation data generated in this project. The closest expression dataset was GSE4229, a study which compared expression levels in blood between AD patients and healthy controls (Maes *et al.*, 2009). Each of the methylation hits was categorised into either hypomethylated or hypermethylated in PSD and hits were searched within the expression dataset. Any locus differentially expressed in AD was categorised into either over or under expressed in AD. Those hits which showed the expected inverse relationship between methylation and expression (i.e. hypermethylated and under expressed, hypomethylated and over expressed) were taken forward to the next step. This inverse relationship between methylation and expression is only expected in promoter regions, it is less clear that it is always the case in other regions, such as gene bodies. The next criterion in both approaches was that the effect size (difference between the outcome groups) was more than 5%. This criterion is based upon the technical limitations of validating any observed differences using Pyrosequencing, which is suggested to be unable to robustly detect differences of less than 5% (Mikeska *et al.*, 2011). Finally, a BLAT search was performed to test for the presence of a SNP within the CpG site. This step was implemented as the list of CpGs used in the filtering step was found to be incomplete. If the CpG site was indeed a SNP, or if there was a SNP in the probe sequence, this site was removed from consideration. This step removed any CpG sites likely to be directly under genetic influence.

The HM450 data were processed and analysed as described to generate a priority list of target differentially methylated positions (DMPs). The Illumina gene ID was used to assign a gene name to the probe; where this was not possible, the HM450 probe ID was retained. These DMPs were then taken forward for validation using Pyrosequencing.

3.2.6 Pyrosequencing

Primers for amplicons covering DMPs were designed using the PSQ Assay Design software (Section 2.6.2) and synthesised by Integrated DNA Technologies (IDT, USA). Where possible, primers were designed, not only to cover the target DMP, but also any neighbouring CpG sites that could be captured within a reasonable amplicon size (50-250bp). Primers were first optimised (Section 2.6.3) and then validated on the Pyrosequencer using Epitect PCR controls (Qiagen, UK) (Section 2.6.5), prior to being run on COGFAST samples. PCR controls containing a 100% methylated DNA sample and a 0% methylated DNA sample were used to make reference methylation levels ranging from 0% to 100%. These validation mixes were created both pre-PCR and post-PCR. Ten of the 22 DMP primer sets were successfully validated. All COGFAST samples were run in duplicate and repeated if duplicate discordance was >5%.

Pyrosequencing assays (n=10) were performed on the “discovery” sample (n=29) along with blood derived DNA samples from an additional sixteen individuals drawn from the same cohort. 500ng of DNA was first bisulphite modified, amplified using the optimised Pyrosequencing PCR conditions in Table 10 and finally sequenced on a PyroMark MD Pyrosequencer (Qiagen, UK) following the standard protocol (Section 2.6.6). Details of all Pyrosequencing primers are provided in Table 11.

Gene/DMP	HM450 probe ID	Tm(°C)	$MgCl_2$ +/-	Size of product(bp)	No. CpGs	Successfully validated using Epitect controls
LIMK2	cg02055988	52.7	+	113	1	Yes
CTRL	cg02126424	50.9	-	155	3	
CTRC	cg03096785	50.9	-	144	1	
cg20583640	cg20583640	54.5	-	168	6	
cg18837178	cg18837178	54.5	+	143	2	Yes
C17orf101	cg02973735	50.9	-	272	6	Yes
PCDHGA1/2	cg06742719	52.7	+	102	2	
cg11349123	cg11349123	54.5	+	129	2	
APOB	cg23603877	50.9	+	105	2	Yes
cg21463981	cg21463981	52.7	+	126	1	Yes
EIF4E3	cg01228342	50.9	-	118	1	Yes
FAXC	cg05621516	56.4	+	345	3	
HSPB3	cg11391732	50.9	-	171	3	Yes
CISD1	cg17324161	53.6	-	330	3	Yes
NGF	cg00794813	53.6	-	305	4	Yes
DSCAM	cg19477942	53.6	-	398	3	
NNMT	cg09632136	53.6	-	346	2	
COX16	cg23562388	53.6	-	378	2	
cg24902278	cg24902278	50.9	-	375	7	
cg08758568	cg08758568	53.6	-	416	2	
cg02925831	cg02925831	50.7	-	291	2	
cg02490189	cg02490189	50.4	-	373	5	Yes

Table 10: Optimal PCR conditions for each assay. Tm = Annealing temperature.

Gene	Forward primer (5’>3’)	Reverse primer (5’>3’)	Sequencing primer (5’>3’)
LIMK2	5Biosg/GGAAAAATTTGAATATTTATA	CATTTCCCTTATATTTACTATC	CTTCCTATTTATCATTTTTA
cg18837178	5Biosg/TATGGTGATTTGTGATTAG	CCATCTTCTCAAATTACT	AAATAAACCTACTTTCTTCC
C17orf101	GTTTTAGTAATGGTGAGTTG	5Biosg/CCCCAATCCTACTACTAT	TTTTTTGTAGAGGATATAAT
APOB	TGGAGAAATTAGGTATGT	5Biosg/AATACACTATTCCAATTATC	TTGTTTTTGGGAATATATAG
cg21463981	GATGTTAGTGGGTTTAAT	5Biosg/AACCCAAATTAACTAAATAC	TTTATTAGAAGGAATTGAGA
HSPB3	AGGTTGGTTGTTGATAAA	5Biosg/CCAACTAAATCCTTTACTTC	GTTGTTGATAAAAGTATAAT
EIF4E3	TGTTTATAGGGTGTGATATT	5Biosg/ACACTAAAAACTACCATCTAAT	TGTTTTTTATGTGGGAAT
NGF	GGAATTATATTTAGAGAGTAA	5Biosg/TACTCCTATAAATCCTATTAA	ATTTTAGGTTGTTTAAAAAG
CISD1	5Biosg/TTTTAGTATTATTGGAGGTTAT	CCAACTACTCAAAAATCTAA	ATTTAAAATACAAATATCCC
cg02490189	GGGTATTTTGAAATATAGTA	5Biosg/CACATCATCCATATCTATA	TGTTGGATTTGTAGGAGA

Table 11: Forward, reverse and sequencing primers for each successfully validated DMP Pyrosequencing assay. 5Biosg = Biotin label.

3.2.7 Statistical analyses

Summary statistics of sub-samples used from the COGFAST cohort were calculated using Chi squared tests, t-tests or Wilcoxon rank sum tests. Within each cohort, differences between D and CN participants were investigated. Following correction for multiple testing, a p value $<3.79\text{x}10^{-4}$ was considered significant. Tests were also performed to look for differences between the discovery cohort and entire COGFAST cohort, as well as the Pyrosequencing cohort and the entire COGFAST cohort. This was to investigate whether the sample subsets were representative of the entire COGFAST cohort. To reduce the risk of false positives, the p value was adjusted for the number of independent tests performed and a p value $<2.84\text{x}10^{-4}$ was considered significant.

Associations between methylation and the outcome variables (diagnosis (D/CN), last MMSE, Braak staging, abstract thinking, attention, calculation, executive function, language comprehension, language expression, memory learning, memory recent, memory remote, memory total, orientation, perception, praxis and total CAMCOG score) were first assessed in the initial "discovery" set (n=29) using regression analyses, adjusted for baseline age and sex. Subsequently, association analyses were repeated including data from the additional samples ("Pyrosequencing" cohort). Given the non-normally distributed nature of methylation data, sensitivity analyses were also performed by comparing differences in the distribution of methylation between sample groups using non-parametric Wilcoxon rank-sum or Kruskal-Wallis tests. In addition to assessing individual DMPs for methylation levels using Pyrosequencing, for those loci containing multiple CpG sites, regional methylation levels were also investigated using the average methylation value across all CpG sites. These validation analyses were performed in Stata.

Correlations between estimates of methylation at each CpG site measured by the HM450 array and by Pyrosequencing were assessed using Spearman's rank in Stata, due to the non-normal distribution of the methylation data.

3.3 Results

Summary statistics for clinical variables of COGFAST participants and sub-groups utilised in this study are summarised in Table 12, Table 13 and Table 14. A number of participants had a large amount of missing data relating to a number of variables. For each variable, the total number of data points used is equal to the number of participants with the relevant information recorded.

Within group comparisons between D and CN demonstrate very similar characteristics relating to stroke and general characteristic variables in each of the 3 cohorts used (Table 12). Participants with dementia were found to be older at both time of stroke and time of death in the Pyrosequencing cohort however these differences were not significant. Both the discovery and Pyrosequencing cohort were considered representative of the entire COGFAST cohort as no significant differences were observed between cohorts in any variable.

No differences were observed between D and CN in any cohort group in relation to medical history or known stroke risk factors (Table 13). No differences were observed in either D or CN participants between any of the cohort groups, indicating that the discovery and Pyrosequencing cohorts were representative of the entire COGFAST cohort.

The within group comparisons between D and CN demonstrate robust differences in the majority of cognitive tests as expected (Table 14). In the entire COGFAST cohort, D and CN had significantly different results in all cognitive tests. There were no significant differences in cognitive test scores in either D or CN participants between the discovery/Pyrosequencing cohort and the entire COGFAST cohort.

	Entire COGFAST			**Discovery Cohort**			**Pyrosequencing Cohort**		
Variables	D	CN	p value	D	CN	p value	D	CN	p value
Baseline Age (Yrs) mean(SD)	81.18 (4.71)	80.36 (4.19)	0.314[‡]	82.53 (5.25)	80.62 (4.35)	0.296[§]	83.08 (5.32)	80.04 (3.79)	0.024[§]
Death Age (Yrs) mean(SD)	85.03 (5.31)	83.81 (4.01)	0.060[‡]	88.18 (5.94)	86.23 (5.15)	0.355[§]	88.42 (5.26)	85.32 (4.59)	0.030[§]
Sex (%males)	54.41	60.16	0.391[†]	47.06	46.15	0.961[†]	50.00	56.00	0.781[†]
PM delay (Hrs) mean(SD)	50.64 (27.20)	54.53 (22.70)	0.940[§]	40.82 (23.11)	32.46 (15.41)	0.335[‡]	40.82 (23.11)	38.93 (22.26)	0.762[‡]
Survival post stroke (Yrs) mean(SD)	4.01 (2.49)	3.46 (2.20)	0.169[‡]	6.0 (2.87)	6.00 (2.66)	0.976[§]	5.335 (2.68)	5.28 (2.94)	0.933[§]
OCSP class n(%)			0.638[†]			0.632[†]			0.571[†]
LACS	20 (43.48)	42 (38.18)		4 (30.77)	3 (23.08)		7 (31.82)	7 (28.00)	
PACS	20 (43.48)	47 (42.73)		5 (38.46)	6 (46.15)		10 (45.45)	12 (48.00)	
TACS	3 (6.52)	6 (5.45)		3 (23.08)	1 (7.69)		3 (13.64)	1 (4.00)	
POCS	3 (6.52)	15 (13.64)		1 (7.69)	3 (23.08)		2 (9.09)	5 (20.00)	
Side of body n(%)			0.647[†]			1.00[†]			0.517[†]
No lateralising signs	8 (17.02)	16 (13.91)		2 (13.33)	1 (7.69)		2 (8.33)	2 (8.00)	
Right	16 (34.04)	48 (41.74)		8 (53.33)	7 (53.85)		16 (66.67)	13 (52.00)	
Left	23 (48.94)	51 (44.35)		5 (33.33)	5 (38.46)		6 (25.00)	10 (40.00)	
Deg weakness arm n(%)			1.00[†]			0.505[†]			0.456[†]
No deficit	13 (27.66)	33 (28.70)		5 (33.33)	3 (23.08)		10 (41.67)	7 (28.00)	
Weakness	33 (70.21)	78 (67.83)		8 (53.33)	10 (76.92)		12 (41.38)	17 (68.00)	
No movement	1 (2.13)	4 (3.48)		2 (13.33)	0		2 (8.33)	1 (4.00)	
Deg weakness leg n(%)			0.429[†]			0.607[†]			0.401[†]
No deficit	20 (42.55)	46 (40.00)		7 (46.67)	6 (46.15)		10 (41.67)	13 (52.00)	
Weakness	27 (57.45)	65 (56.52)		6 (40.00)	7 (53.85)		12 (50.00)	12 (48.00)	
No movement	0	4 (3.48)		2 (13.33)	0		2 (8.33)	0	
Dysphasia n(%)	8/47 (17.02)	23/115 (20.00)	0.662[†]	6/15 (40.00)	3 (23.08)	0.435[†]	9 (37.50)	7/25 (28.00)	0.478[†]

Table 12: Summary statistics for variables relating to the stroke event. Data are presented for the entire COGFAST cohort, the initial discovery sample set and the Pyrosequencing cohort. D = Participants with dementia. CN = Cognitively normal. [†] = chi squared, [‡] = Wilcoxon rank sum, [§]= t test. LACS=lacunar stroke, PACS=partial anterior circulation stroke, TACS=total anterior circulation stroke, POCS=posterior circulation stroke. Deg=degree.

	Entire COGFAST			Discovery Cohort			Pyrosequencing Cohort		
Variables	D	CN	P value	D	CN	p value	D	CN	p value
Antihypertensive therapy	23/46 (50.00)	61/114 (53.51)	0.687†	6/15 (40.00)	7/13 (53.85)	0.464†	10/24 (41.67)	13/25 (52.00)	0.469†
Antiplatelet therapy	25/46 (54.35)	57/114 (50.00)	0.619†	7/15 (46.67)	5/13 (38.46)	0.662†	10/24 (41.67)	12/25 (48.00)	0.656†
Anticoagulant therapy	1/46 (2.17)	12/114 (10.53)	0.111†	2/15 (13.33)	0	0.484†	3/24 (12.50)	0	0.110†
Lipid lowering therapy	1/46 (2.17)	10/114 (8.77)	0.180†	0	1/13 (7.69)	0.464†	0	1/25 (4.00)	1.00†
Heart failure therapy loop diuretics	7/46 (15.22)	21/114 (18.42)	0.629†	4/15 (26.67)	3/13 (23.08)	1.00†	5/24 (20.83)	7/25 (28.00)	0.742†
Antidepressants	8/46 (17.39)	12/114 (10.53)	0.235†	3/15 (20.00)	1/13 (7.69)	0.60†	5/24 (20.83)	2/25 (8.00)	0.247†
No of CVD risk factors (mode(range))	2 (0-4)	1 (0-6)	0.440‡	2 (0-4)	1 (0-4)	0.314§	2 (0-4)	1 (0-4)	0.394§
Hypertension	31/48 (64.58)	63/115 (54.78)	0.248†	8/15 (53.33)	7/13 (53.85)	0.978†	13/24 (54.17)	15/25 (56.00)	1.00†
Atrial fibrillation	4/48 (8.33)	18/113 (15.93)	0.315†	5/15 (33.33)	3/13 (23.08)	0.686†	6/24 (25.00)	5/25 (20.00)	0.742†
IHD	22/48 (45.83)	39/114 (34.21)	0.163†	4/15 (26.67)	2/13 (15.38)	0.655†	8/24 (33.33)	5/25 (20.00)	0.345†
Angina	21/48 (43.75)	35/114 (30.70)	0.111†	4/15 (26.67)	2/13 (15.38)	0.655†	7/24 (48.98)	5/25 (20.00)	0.52†
Cardiac failure	7/48(14.58)	21/115 (18.26)	0.570†	3/15 (20.00)	3/13 (23.08)	1.00†	4/24 (16.67)	8/25 (32.00)	0.321†
Hyperchol	2/47 (4.26)	16/115 (13.91)	0.099†	1/15 (6.67)	0	1.00†	2/24 (8.33)	1/25 (4.00)	0.609†
Diabetes	6/48 (12.50)	8/115 (6.96)	0.250†	1/15 (6.67)	1/13 (7.69)	1.00†	2/24 (8.33)	1/25 (4.00)	0.609†
Intermittent claudication	4/48 (8.33)	11/115 (9.57)	1.00†	2/15 (13.33)	1/13 (7.69)	1.00†	2/24 (8.33)	4/25 (16.00)	0.667†
Smoking			0.563†			0.292†			0.300†
Current	6 (9.23)	16 (13.91)		3 (20.00)	0		4/23 (17.39)	1/23 (4.35)	
Ex-smoker	3 (4.62)	8 (6.96)		1 (6.67)	1 (7.69)		1/23 (4.35)	3/23 (13.04)	
Non-smoker	56 (86.15)	91 (79.13)		11 (73.33)	12 (92.31)		18/23 (78.26)	19/23 (82.61)	

Table 13: Summary statistics for variables relating to stroke/dementia risk factors and medical history. Data are presented for the entire COGFAST cohort, the initial discovery sample set and the Pyrosequencing cohort. D = Participants with dementia. CN = Cognitively normal. † = chi squared, ‡ = Wilcoxon rank sum, §= t test. Hyperchol=hypercholesterolaemia. Unless stated, data for n(%) are displayed.

	Entire COGFAST			**Discovery Cohort**			**Pyrosequencing Cohort**		
Variables	D	CN	p value	D	CN	p value	D	CN	p value
Braak Staging n(%)			0.586†			0.987†			0.814†
I	4 (12.9)	6 (20.00)		2 (11.76)	2 (15.38)		2 (9.09)	3 (15.79)	
II	9 (29.03)	9 (30.00)		3 (17.65)	4 (30.77)		6 (27.27)	8 (42.11)	
III	9 (29.03)	9 (30.00)		6 (35.29)	4 (30.77)		8 (36.36)	4 (21.05)	
IV	5 (16.13)	4 (13.33)		3 (17.65)	2 (15.38)		3 (13.64)	2 (10.53)	
V	3 (9.68)	2 (6.67)		2 (11.76)	1 (7.69)		2 (9.09)	2 (10.53)	
VI	1 (3.23)	0		1 (5.88)	0		1 (4.55)	0	
Last MMSE before death	16.15 (6.29)	25.84 (2.59)	1.65E-23‡	15.82 (5.49)	26.38 (3.07)	1.89E-05‡	17.88 (6.16)	23.32 (5.99)	7.00E-04‡
Orientation	5.61 (2.81)	9.25 (1.04)	1.80E-20‡	4.65 (2.60)	9.15 (1.21)	2.163E-05‡	5.89 (3.00)	8.4 (1.85)	0.001‡
Language comprehension	7.00 (1.83)	8.22 (1.25)	2.05E-08‡	6.65 (2.32)	8.69 (0.48)	0.0005§	6.96 (2.01)	8.24 (0.88)	0.007‡
Language expression	13.19 (3.48)	16.85 (1.94)	2.26E-15‡	12.59 (3.52)	17.54 (1.13)	0.0001§	13.42 (3.31)	16.36 (2.29)	0.001‡
Memory remote	3.69 (1.66)	5.14 (1.01)	7.79E-10‡	3.88 (1.58)	5.00 (0.58)	0.036‡	4.00 (1.39)	4.64 (1.22)	0.058‡
Memory recent	2.00 (1.40)	3.67 (0.67)	9.50E-18‡	2.06 (1.34)	3.38 (0.77)	0.004‡	2.35 (1.35)	3.08 (1.00)	0.048‡
Memory learning	7.31 (3.92)	12.50 (2.62)	4.13E-17‡	7.53 (3.74)	13.23 (1.83)	0.0001‡	8.5 (4.19)	11.24 (3.55)	0.014§
Attention	2.96 (2.16)	5.68 (1.48)	5.92E-15‡	2.88 (1.96)	6.31 (0.85)	4.077E-05‡	3.35 (2.23)	5.28 (1.81)	0.002‡
Praxis	6.88 (2.23)	9.69 (1.97)	9.93E-14‡	6.94 (2.38)	10.69 (1.03)	2.214E-05‡	7.31 (2.78)	9.08 (2.61)	0.013‡
Calculation	0.88 (0.75)	1.76 (0.48)	1.18E-15‡	0.88 (0.86)	1.69 (0.48)	0.009‡	1.08 (0.89)	1.48 (0.59)	0.111‡
Abstract thinking	4.12 (2.30)	5.94 (1.64)	9.24E-08‡	4.12 (2.74)	6.23 (1.64)	0.030§	4.62 (2.38)	5.64 (1.96)	0.117§
Perception	5.33 (2.08)	6.77 (1.61)	1.29E-06‡	4.94 (2.73)	7.38 (1.56)	0.009§	5.42 (2.76)	7.12 (1.56)	0.020§
Total CAMCOG	58.78 (17.52)	85.45 (8.61)	1.48E-23‡	57.76 (18.68)	89.31 (6.18)	4.54E-06§	63.31 (19.68)	80.56 (14.15)	0.001‡
Executive function	9.61 (4.10)	15.42 (4.34)	1.07E-14§	9.29 (5.03)	17.38 (3.04)	0.0001§	10.92 (5.16)	14.88 (4.6)	0.007§
Memory total	12.90 (5.86)	21.30 (3.19)	1.44E-19‡	13.41 (5.80)	21.46 (2.90)	0.0001§	14.81 (6.03)	18.6 (5.28)	0.014‡

Table 14: Summary statistics for variables relating to both pathology and cognitive-based outcomes. Data are presented for the entire COGAST cohort, the initial discovery sample set and the Pyrosequencing cohort. D = Participants with dementia. CN = Cognitively normal. † = chi squared, ‡ = Wilcoxon rank sum, §=t test. Unless stated, mean (SD) are presented for each variable.

3.3.1 Discovery analysis using HM450 data

Of the 30 blood samples that underwent HM450 analysis, 29 samples passed the quality control assessment of the inbuilt microarray controls assessed in GenomeStudio. One sample (from the dementia group) failed due to incomplete bisulphite modification and was dropped from further analysis.

Low failure rates were observed on both a sample and probe by probe basis, assessed using detection p values generated from the array scanner. As described in Section 2.5.2, samples were only removed if ≥20% of probes had a detection p value >0.01. Probes were removed if ≥10% of samples had a detection p value >0.01. No samples were removed from analysis using these criteria. 1665 probes were removed. Summaries of the detection p values observed for samples and probes are shown in Figure 10A and 10B.

Following data filtering and normalisation, as described in Section 2.5.2, 455,173 CpG sites were taken forward for analysis measured in baseline blood samples from 29 individuals; 16 D vs 13 CN.

The small sample size of this study limited the power to detect methylation differences. This study had 2.6% power to detect a mean difference of 5% methylation between groups, with a standard deviation of 3 and alpha=1.1×10^{-7}.

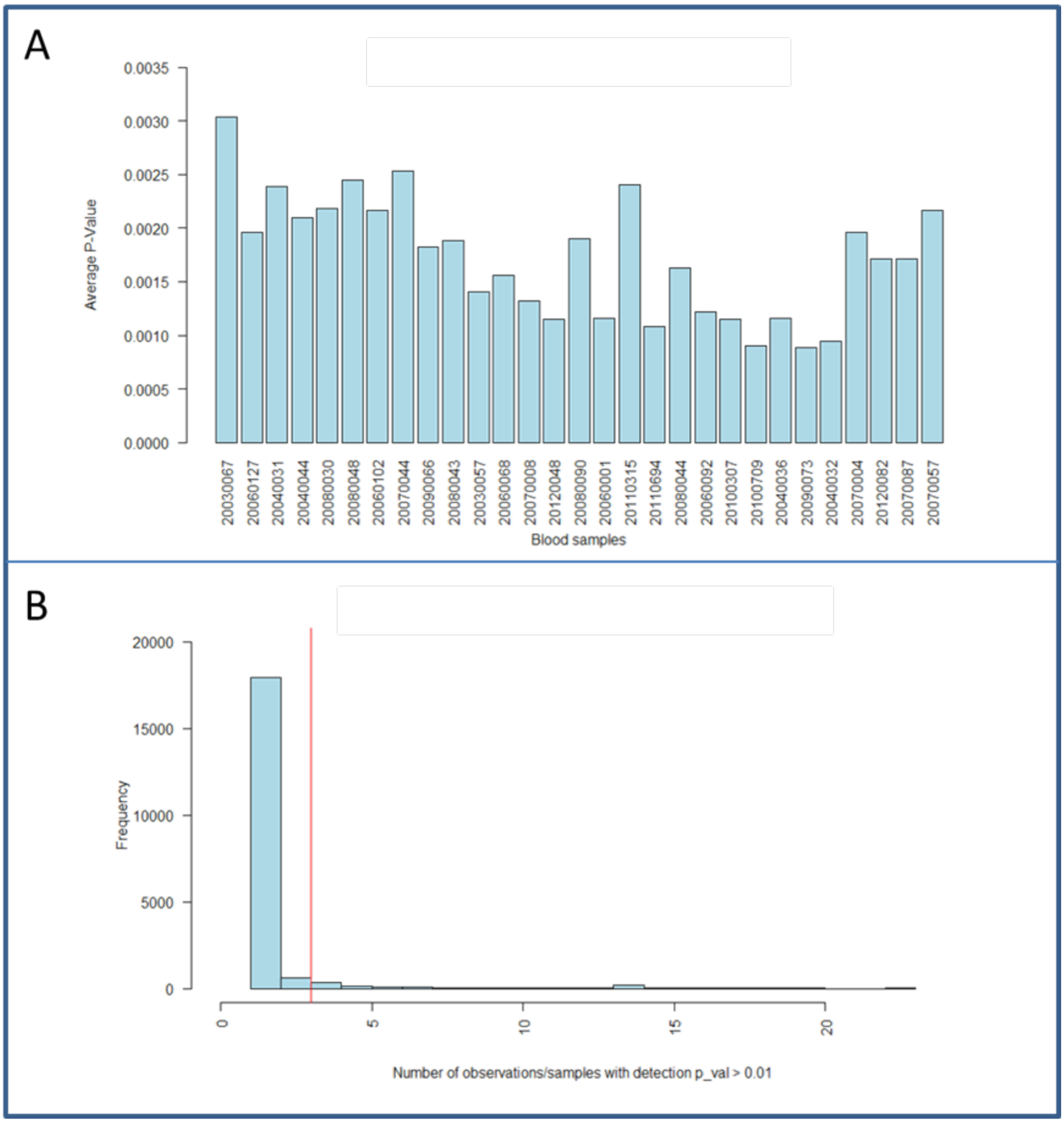


Figure 10A: Average detection p values in each blood sample. For each sample the average detection p value is very low and samples are of approximately equal quality.

Figure 10B: Histogram to show the number of probes against the corresponding number of samples with p values greater than 0.01. The red line indicates 10%. 465939 probes had no missing data and for space reasons, this data is not shown on the histogram.

3.3.1.1 *Distribution of genome-wide methylation values*

Figure 11 (A-F) shows the distribution of methylation values using both raw (A-C) and normalised (D-F) data. Figure 11A is a density plot showing the distribution of raw methylation beta values for each blood sample. This shows a commonly observed bimodal distribution; the majority of CpG sites are very lowly or very highly methylated with much fewer CpG sites exhibiting intermediate methylation levels. The distribution for each individual can be more easily seen in a box plot (Figure 11B). This shows that there is some variability in the methylation levels of each blood sample, particularly with regard to the median methylation and interquartile range (IQR). Each sample has a

minimum methylation level of 0% and a maximum of 100% with a median around 60-70% methylated. Some of this variation could be due to batch effects and sample processing. The position of the sample on the chip is also thought to have an effect on methylation values (Dedeurwaerder *et al.*, 2013) but this variation should be removed after normalisation. Mean and standard deviation (SD) methylation beta values across all probes, prior to normalisation and QC, are depicted in Figure 11C to show heteroskedasticity. Heteroskedasticity refers to the phenomenon whereby the variance in a measure such as methylation beta values is not constant across all levels of that measure. As can be seen, there is a wide range of variance across the probes. In particular, probes with a mid-range methylation value between 30-80% show greater variability compared to those at the extreme ends of the methylation spectrum.
Figure 11 (D-F) shows the distribution of methylation after normalisation. All plots using normalised data show that there is much less variation between samples. All samples have a median methylation value of around 50%. This suggests that most of the variation was removed during normalisation and is therefore likely to be due to technical variation.

3.3.1.2 *Cluster analysis*

Hierarchical cluster analysis, using the complete-linkage method, was performed on the normalised data to look for natural groupings (either phenotypic or technical) amongst the samples based on the methylation data. No clear clusters were observed for diagnosis of dementia, indicating that global DNA methylation signatures are not sufficiently distinct in blood samples drawn from stroke patients who subsequently go on to develop dementia (Figure 12).

However, as shown in Figure 13, even after normalisation and removal of probes mapping to the sex chromosomes, there remained some subtle groupings amongst samples based on sex. Due to these results, sex was included as a covariate in subsequent analyses of single-point CpG sites. In contrast, there were no apparent clusters based on experimental chip (Figure 14), suggesting that any batch effects were removed. Nonetheless, to fully account for any remaining, albeit subtle, batch effects, chip was also included as a covariate in the single-point analysis model. The cluster plots show that the data is clustered into two main groups, however this clustering was not due to diagnosis, sex, chip or any other batch effect measured.

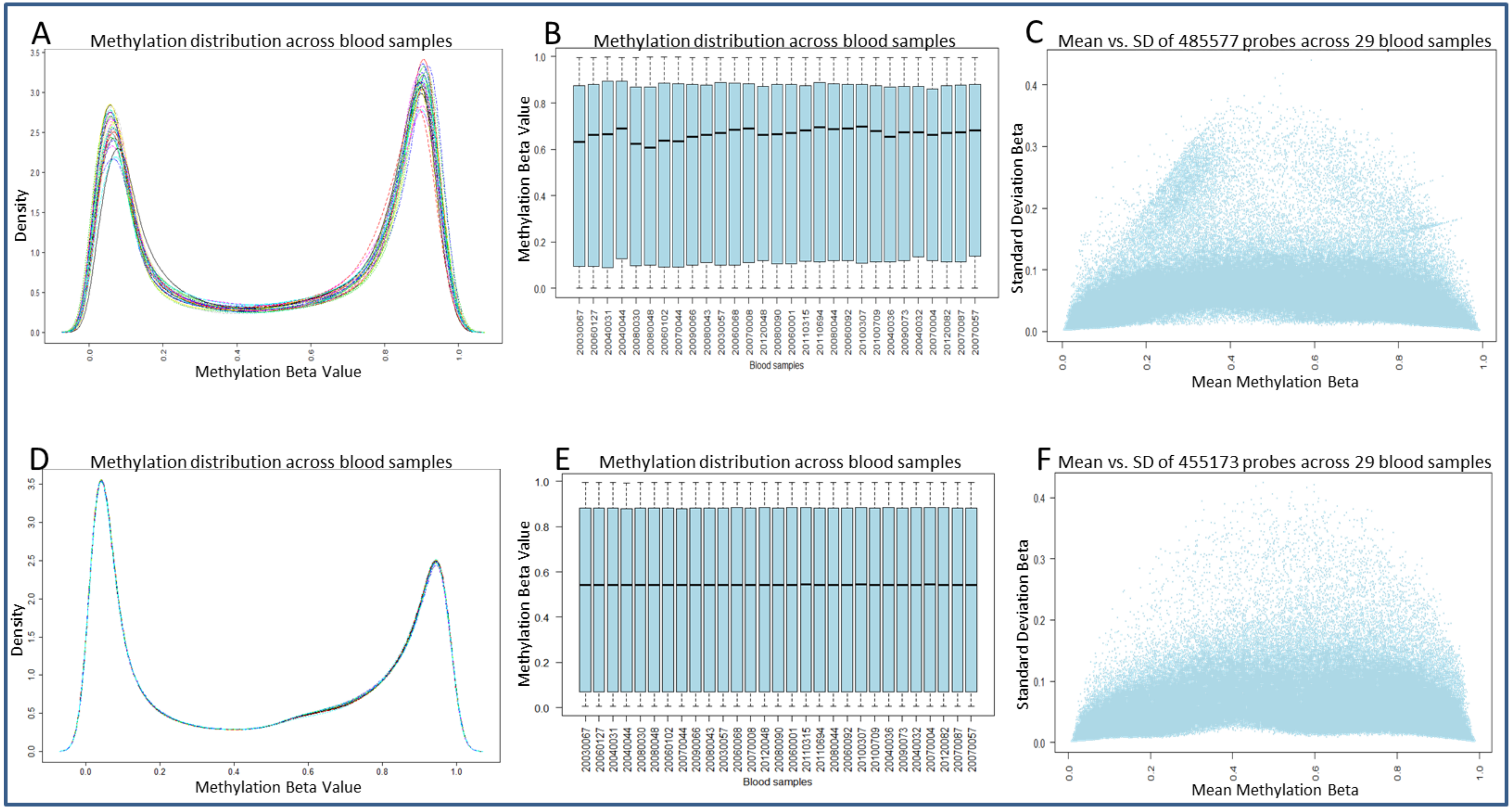


Figure 11: A composite figure showing the distributions of methylation using raw and normalised data. A. Density plot showing methylation distribution in 29 blood samples using raw data. B. Box plot showing mean methylation raw beta values across all blood samples. C. Mean vs SD plot showing heteroskedasticity using raw data. D. Density plot showing methylation distribution in 29 blood samples using normalised data. E. Box plot showing mean methylation normalised beta values across all blood samples. F. Mean vs SD plot showing heteroskedasticity using normalised data.

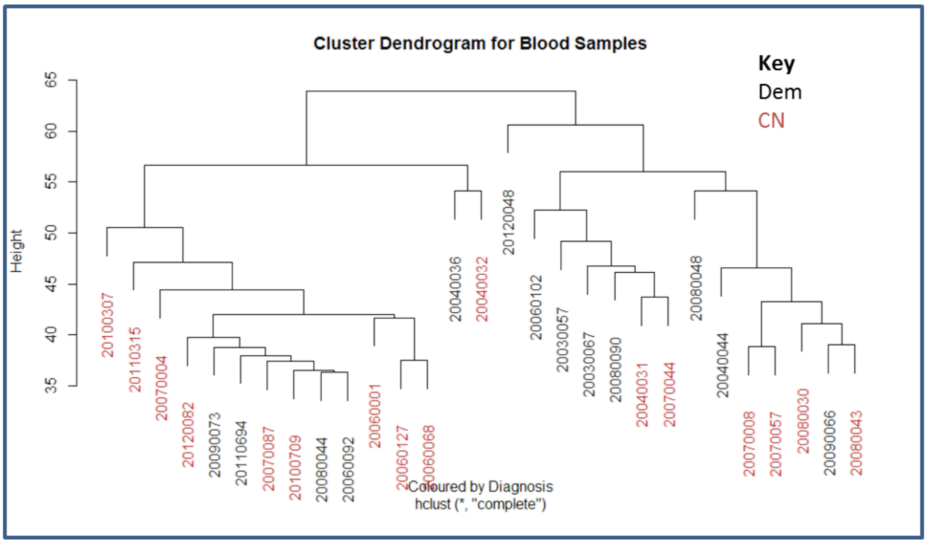


Figure 12: Cluster dendrogram for all 29 blood samples coloured by diagnosis.

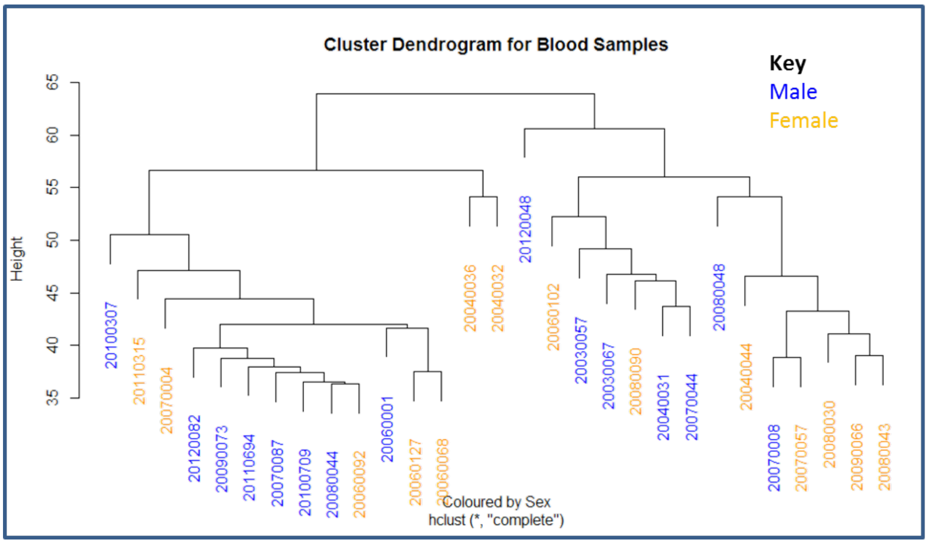


Figure 13: Cluster dendrogram for all 29 blood samples coloured by sex.

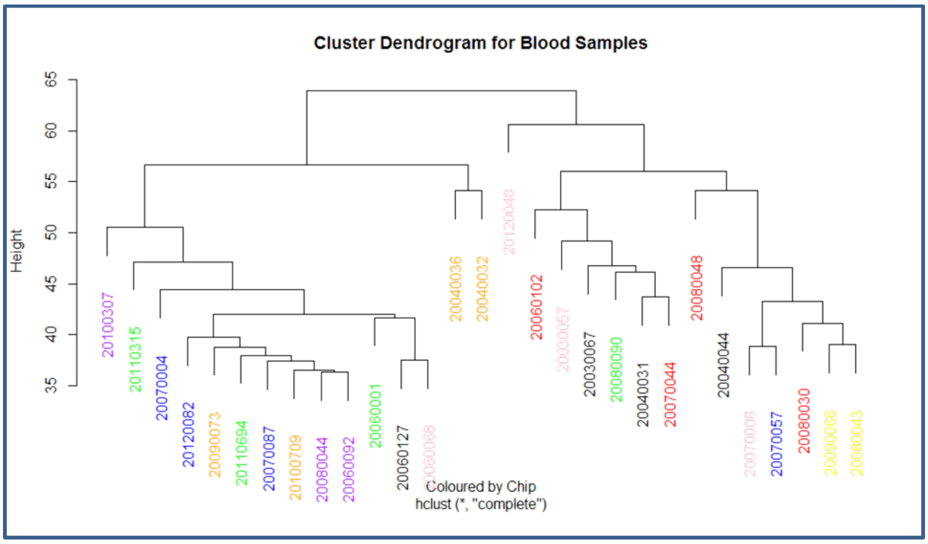


Figure 14: Cluster dendrogram for all 29 blood samples coloured by chip.

3.3.1.3 ***Single-point analysis***

ANOVA was performed on all beta values across 455,173 CpG sites against the following outcome measures: diagnosis (D/CN), last MMSE result before death, Braak staging and a number of CAMCOG scores (abstract thinking, attention, calculation, executive function, language comprehension, language expression, memory learning, memory recent, memory remote, memory total, orientation, perception, praxis and total CAMCOG score). Baseline age, sex and chip were included in the analysis model as covariates. Results from this step were exported as a table in Microsoft Excel and ordered by ascending p value.

Two approaches were employed when selecting top hits.

3.3.1.4 ***Selection of top hits – approach 1***

The first approach selects hits which reach genome wide significance. When testing for association between methylation and both diagnosis and last MMSE, no CpG sites were highlighted as significant at the conservative genome wide correction threshold of $p<1.10x10^{-7}$. Likewise, a number of CAMCOG scores (abstract thinking, attention, calculation, executive function, language comprehension, language expression, memory learning, memory recent, memory total, orientation, perception and praxis) were not associated with methylation at any of the 455,173 CpG sites included in the analysis. However, when Braak staging was included in the model as the outcome variable, methylation was significantly different at 74 CpG sites (DMPs) (Figure 15 and Figure 16). One CpG site was found to be associated with the CAMCOG variable “memory remote” (Figure 17 and Figure 18) and another with the total CAMCOG score (Figure 19 and Figure 20). These 76 DMPs were then considered as targets for validation by Pyrosequencing.

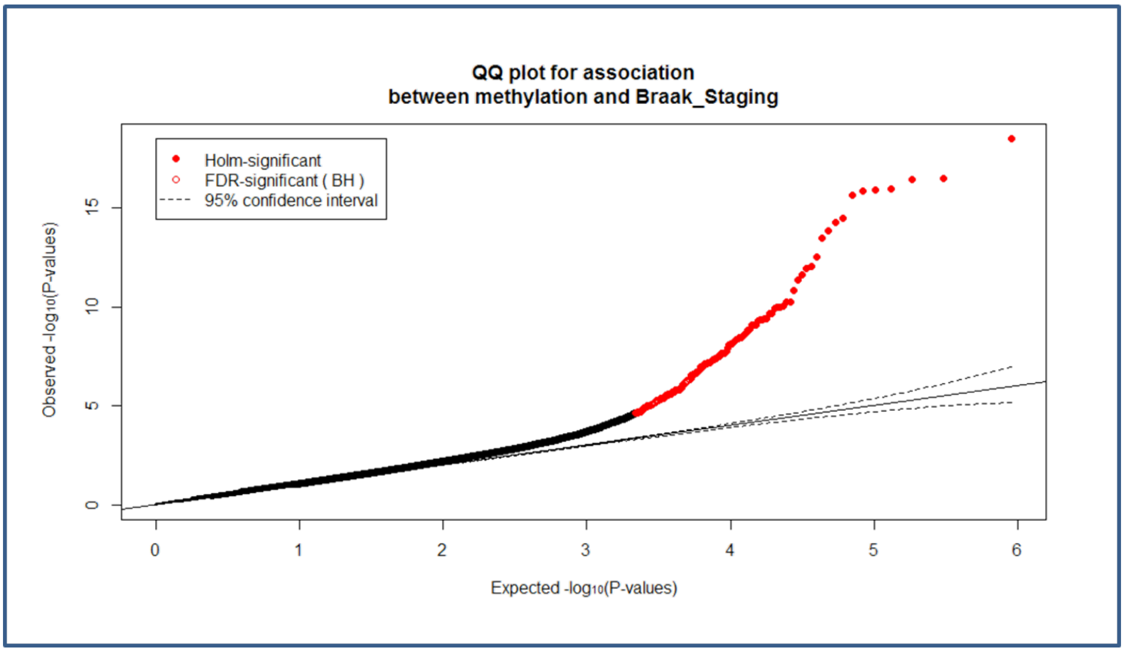


Figure 15: Quantile-quantile (QQ) plot for association between methylation and Braak staging. The QQ plot shows the expected against the observed p values. Data points plotted in red reach the Bonferroni corrected significance threshold.

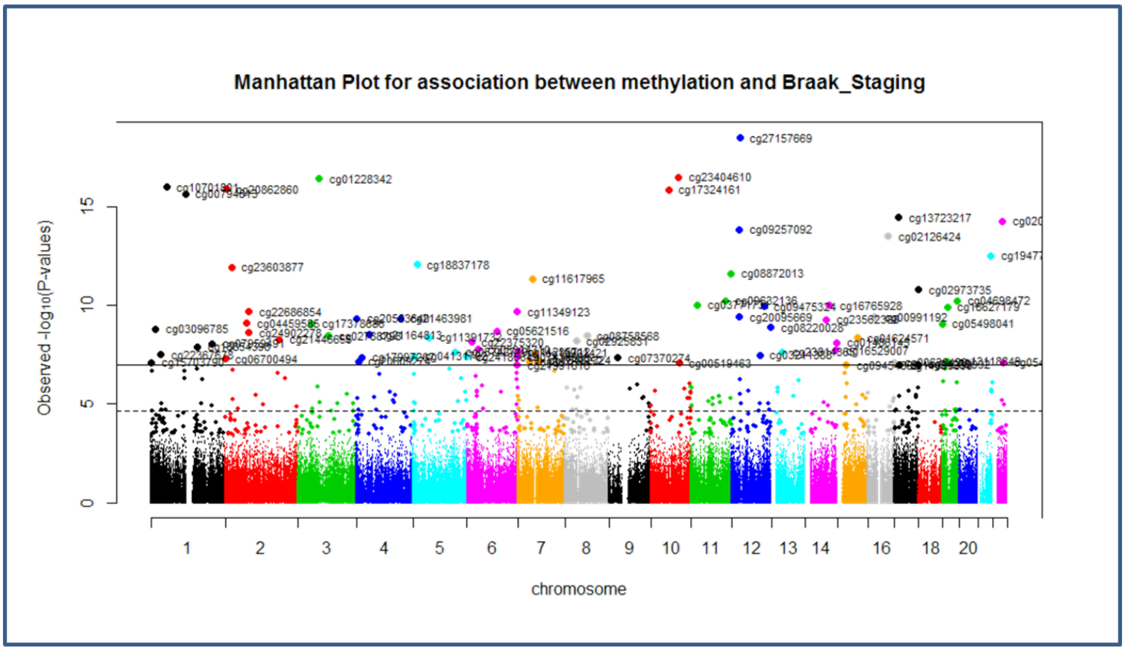


Figure 16: Manhattan plot for association between methylation and Braak staging. Each circle represents an individual CpG site. The continuous line marks the p<1.1x10^{-7} significance threshold. CpG sites with p values below the threshold are labelled with their Illumina ID.

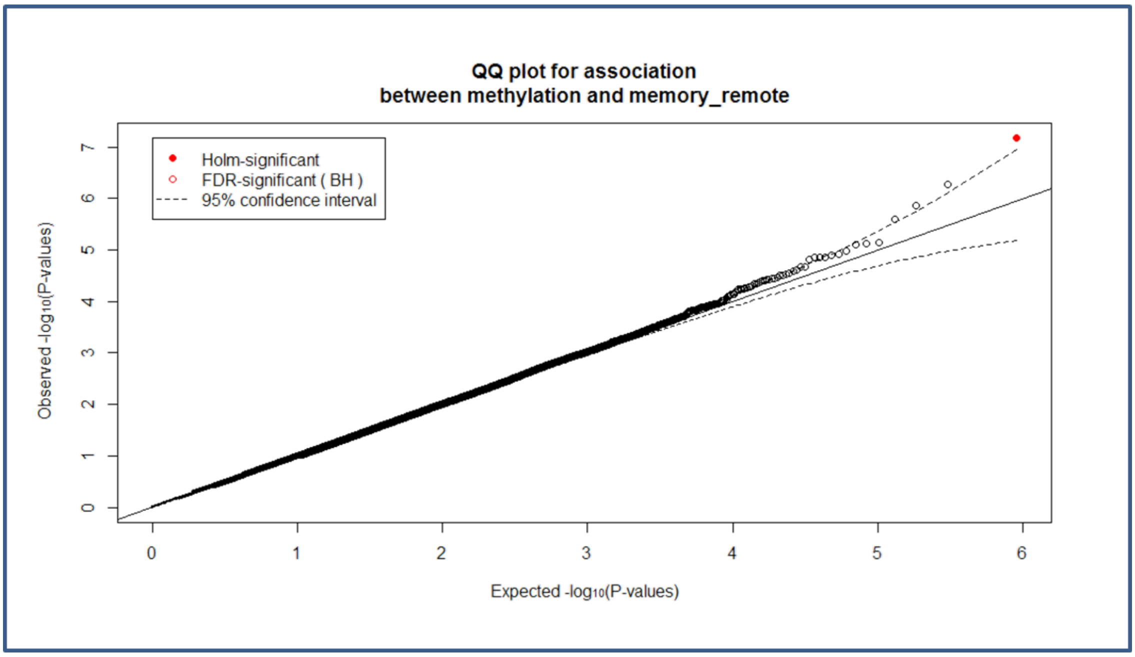


Figure 17: QQ plot for association between methylation and memory remote. The QQ plot shows the expected against the observed p values. Data points plotted in red reach the Bonferroni corrected significance threshold.

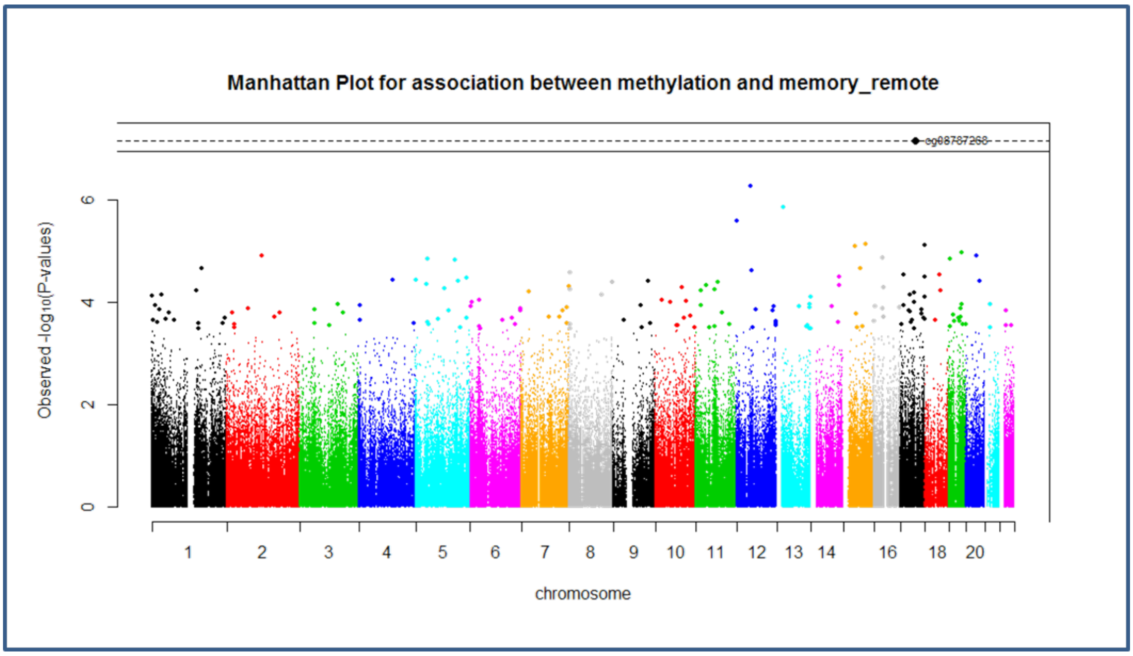


Figure 18: Manhattan plot for association between methylation and memory remote. Each circle represents an individual CpG site. The continuous line marks the $p<1.1x10^{-7}$ significance threshold. CpG sites with p values below the threshold are labelled with their Illumina ID.

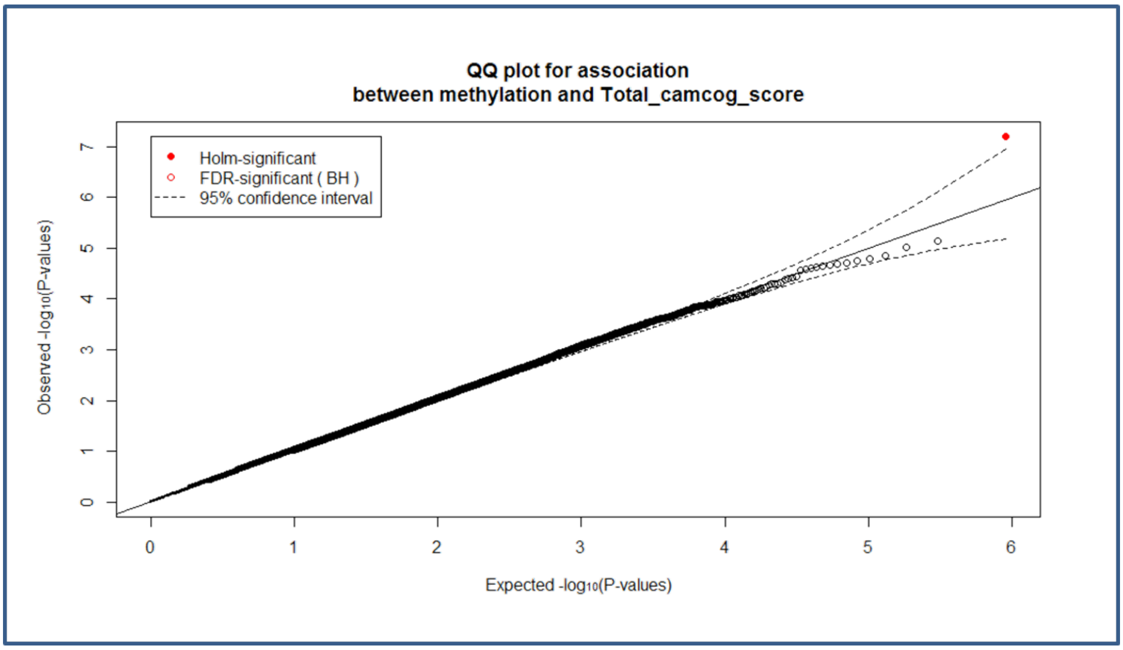


Figure 19: QQ plot for association between methylation and total CAMCOG score. The QQ plot shows the expected against the observed p values. Data points plotted in red reach the Bonferroni corrected significance threshold.

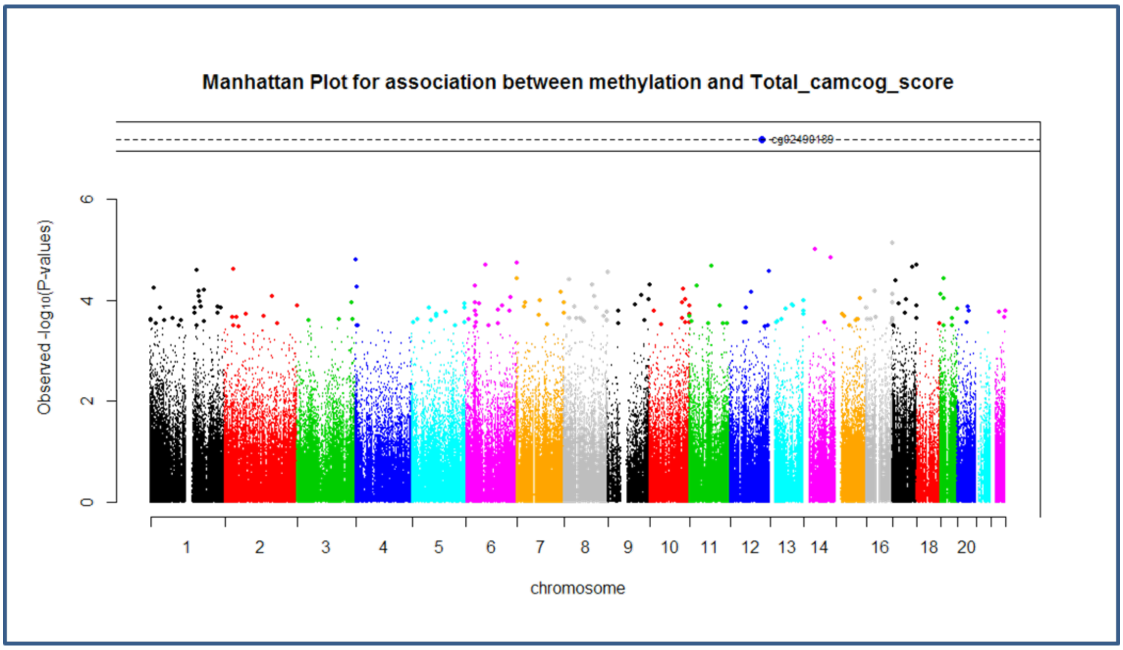
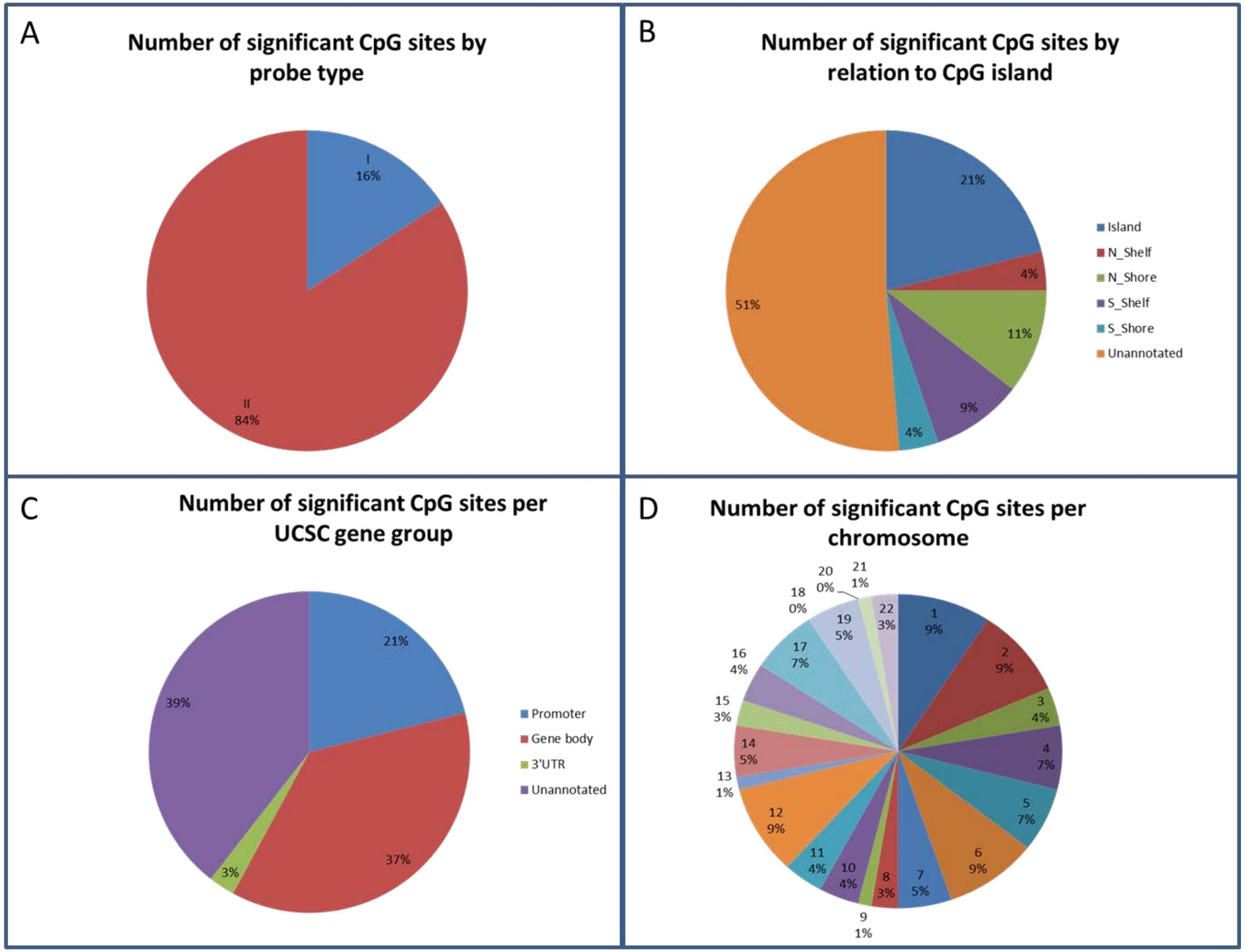


Figure 20: Manhattan plot for association between methylation and total CAMCOG score. Each circle represents an individual CpG site. The continuous line marks the $p<1.1x10^{-7}$ significance threshold. CpG sites with p values below the threshold are labelled with their Illumina ID.

Figure 21 (A-D) shows the distribution of the most significant 76 CpG sites (when tested with Braak staging, memory remote and total CAMCOG score as outcome variables) across the different probe chemistries and genomic region (chromosome, position relative to nearby CpG island and regions within gene).

A chi squared goodness of fit test was performed to test whether the distributions of hits observed were as expected. A significantly lower number of significant CpG sites were measured using probe type I chemistry than expected (16% vs 27%, p=0.01) (Figure 21A). The majority of significant CpG sites are unannotated on the Illumina technical information sheet and have no information on genomic location or relation to CpG islands. 21% of all CpG sites are located within CpG islands (Figure 21B). Differences in the position of the CpG site in relation to CpG islands were also not as expected, with fewer significant CpG sites being observed in the CpG islands ($p=9.12x10^{-4}$). 21% of all significant CpG sites are located within gene promoters (Figure 21C); this indicates that some of these CpG sites may have functional relevance. In addition, differences in functional genomic location (eg. promoter, gene body) were also identified between the number of CpG sites observed and expected in each region (p=0.01). Significant CpG sites do not appear to be more likely to appear on one chromosome than any other as hits are spread across the genome. No differences were observed in the distribution of CpG sites across the chromosomes (p=0.74) (Figure 21D).


Number of significant CpG sites by probe type

Number of significant CpG sites by relation to CpG island

Number of significant CpG sites per UCSC gene group

Number of significant CpG sites per chromosome

Figure 21: Description of statistically significant CpG sites identified in blood samples. A. Distribution of CpG sites measured using different probe types. B. position relative to nearby CpG island. C. Functional genomic location. D. Proportion of statistically significant CpGs located on each chromosome.

Of the 76 top hits, 30 primer pairs were successfully designed (Table 15). Four CpG sites dropped out of the selection due to having small (<5%) effect sizes, i.e. the difference between the lowest and highest measure (e.g. the mean methylation of Braak stage I individuals minus the mean methylation of Braak stage VI individuals). 5% was chosen as the minimum effect size due to the technical sensitivity of the Pyrosequencer and to limit the validation phase to those CpG sites most likely to exhibit a methylation difference. Although a number of SNPs were removed during the HM450 filtering step (Section 2.5.2), an additional SNP search was incorporated into the work flow. A BLAT search (NCBI) of the probe sequence highlighted that 18 of the remaining 72 CpG sites were in fact SNPs and these were removed from the selection. A search on the National Human Genome Research Institute (NHGRI) Genome-Wide Association Studies (GWAS) Catalog (www.genome.gov/gwastudies) revealed that none of these 18 SNPs had previously been associated with post-stroke dementia or any related disease. The remaining 54 loci were taken forward for primer design.

Thirty assays were successfully designed, with PSQ design software quality scores > 70 and no other warnings. Therefore, the limiting factors at this stage of the study design, were the technical requirements for successful assay design using the Pyrosequencing platform. Figure 22 shows the selection of top hits.

Variable	CpG site	P value	Gene	Test statistic	Effect size	SNP	Primers
Braak Staging	cg17324161	1.43E-16	*CISD1*	771.70	0.815	No	Yes
Braak Staging	cg03771731	9.63E-11	*SLC6A5*	110.62	0.604	No	No
Braak Staging	cg05498041	8.53E-10		80.10	0.599	No	No
Braak Staging	cg00794813	2.29E-16	*NGF*	721.32	0.599	No	Yes
Braak Staging	cg02055988	5.43E-15	*LIMK2*	457.68	0.590	No	Yes
Braak Staging	cg02126424	3.29E-14	*CTRL*	353.07	0.570	No	Yes
Braak Staging	cg03096785	1.56E-09	*CTRC*	73.18	0.550	No	Yes
Braak Staging	cg07438999	3.94E-08		44.92	0.539	No	No
Braak Staging	cg09632136	6.02E-11	*NNMT*	118.54	0.496	No	Yes
Braak Staging	cg18837178	8.63E-13		220.13	0.465	No	Yes
Braak Staging	cg20583640	4.67E-10		87.60	0.457	No	Yes
Braak Staging	cg16765928	1.05E-10	*NRXN3*	109.15	0.449	No	No
Braak Staging	cg02973735	1.58E-11	*C17orf101*	144.25	0.427	No	Yes
Braak Staging	cg19477942	3.08E-13	*DSCAM*	255.57	0.426	No	Yes
Braak Staging	cg02788798	3.49E-09		64.90	0.403	No	No
Braak Staging	cg09454882	1.06E-07	*FAM189A1*	38.56	0.398	No	No
Braak Staging	cg11349123	2.13E-10		98.44	0.365	No	Yes
Braak Staging	cg13118849	7.02E-08	*NCAN*	47.17	0.348	No	No
Braak Staging	cg05475277	7.91E-08	*FOXRED2*	40.34	0.346	No	Yes
Braak Staging	cg02925831	6.28E-09		59.40	0.316	No	Yes
Braak Staging	cg20609274	6.90E-08	*PPP2R2C*	41.21	0.313	No	Yes
Braak Staging	cg08872013	2.51E-12	*OPCML*	188.56	0.311	No	No
Braak Staging	cg23603877	1.20E-12	*APOB*	209.81	0.311	No	Yes
Braak Staging	cg05314310	1.68E-08	*LHFPL5*	51.15	0.295	No	Yes
Braak Staging	cg04131792	3.81E-08		45.15	0.295	No	No
Braak Staging	cg19901421	2.64E-08		47.75	0.276	No	No
Braak Staging	cg21164813	3.06E-09		66.18	0.271	No	No
Braak Staging	cg06700494	5.02E-08		43.27	0.264	No	Yes
Braak Staging	cg11235712	2.15E-08		49.30	0.242	No	Yes
Braak Staging	cg08758568	3.64E-09		64.49	0.241	No	Yes
Braak Staging	cg05621516	2.01E-09	*FAXC*	70.47	0.227	No	Yes
Braak Staging	cg17997207	4.40E-08	*FAM184B*	44.16	0.215	No	No
Braak Staging	cg24239992	9.86E-08		38.99	0.201	No	No
Braak Staging	cg11617965	4.52E-12	*PKD1L1*	173.08	0.184	No	No
Braak Staging	cg21991616	1.09E-07	*HGC6.3*	38.39	0.165	No	No
Braak Staging	cg11391732	4.15E-09	*HSPB3*	63.24	0.159	No	Yes
Braak Staging	cg01486145	7.96E-09		57.31	0.154	No	Yes
Braak Staging	cg03211388	3.39E-08		45.97	0.150	No	Yes
Memory remote	cg08787268	6.73E-08		-8.75	0.059	No	No
Total CAMCOG	cg02490189	6.33E-08	*NUAK1*	8.78	-0.050	No	Yes
Braak Staging	cg00519463	8.40E-08	*KIF11*	39.97	-0.093	No	No
Braak Staging	cg07959491	8.73E-09		56.51	-0.098	No	No
Braak Staging	cg09475324	1.11E-10	*FICD*	108.26	-0.099	No	No
Braak Staging	cg23562388	5.26E-10	*COX16*	103.70	-0.214	No	Yes

Braak Staging	cg09257092	1.43E-14	*ITPR2*	397.96	-0.225	No	No
Braak Staging	cg01228342	3.85E-17		931.54	-0.236	No	Yes
Braak Staging	cg23814365	2.37E-08	*DCLK1*	48.55	-0.242	No	Yes
Braak Staging	cg21463981	4.63E-10		87.71	-0.288	No	Yes
Braak Staging	cg22686854	2.18E-10		98.05	-0.331	No	No
Braak Staging	cg22367678	3.17E-08		46.43	-0.347	No	No
Braak Staging	cg06742719	2.30E-08	*PCDHGA2*	48.78	-0.352	No	Yes
Braak Staging	cg14099398	1.04E-07	*PLD6*	38.67	-0.381	No	No
Braak Staging	cg24902278	2.46E-09		68.39	-0.385	No	Yes
Braak Staging	cg16482324	6.12E-08	*AUTS2*	41.97	-0.435	No	No
Braak Staging	cg21446655	5.43E-09	*UBE2E3*	71.45	0.764	Yes	
Braak Staging	cg23404610	3.27E-17		953.62	0.623	Yes	
Braak Staging	cg27157669	3.30E-19	*KLHD5C*	1841.79	0.575	Yes	
Braak Staging	cg01624571	4.45E-09	*SMAD3*	62.56	0.573	Yes	
Braak Staging	cg17378686	8.49E-10	*PTH1R*	80.16	0.561	Yes	
Braak Staging	cg16529007	1.91E-08		50.19	0.559	Yes	
Braak Staging	cg20862860	1.28E-16	*TSSC1*	784.32	0.537	Yes	
Braak Staging	cg08220028	1.36E-09	*GALNT9*	74.75	0.529	Yes	
Braak Staging	cg20095669	3.92E-10	*PPFIBP1*	108.62	0.524	Yes	
Braak Staging	cg04698472	5.84E-11	*SIGLEC12*	119.05	0.514	Yes	
Braak Staging	cg00991192	3.96E-10		89.78	0.474	Yes	
Braak Staging	cg24185649	4.03E-08	*ZNF354B*	44.76	0.432	Yes	
Braak Staging	cg22375320	7.13E-09		58.27	0.397	Yes	
Braak Staging	cg07370274	4.70E-08		43.71	0.319	Yes	
Braak Staging	cg18854398	1.28E-08	*C1orf66*	53.31	-0.100	Yes	
Braak Staging	cg04459585	7.94E-10		80.97	-0.125	Yes	
Braak Staging	cg18627179	1.20E-10	*RPL18A*	131.04	-0.202	Yes	
Braak Staging	cg13723217	3.57E-15		486.12	-0.359	Yes	
Braak Staging	cg01289352	7.05E-08		41.07	0.048		
Braak Staging	cg10701801	1.06E-16		805.80	0.027		
Braak Staging	cg15703790	7.97E-08		40.29	0.022		
Braak Staging	cg00630420	8.06E-08		40.23	0.018		

Table 15: Details about top 76 CpG sites selected based on genome wide significance.

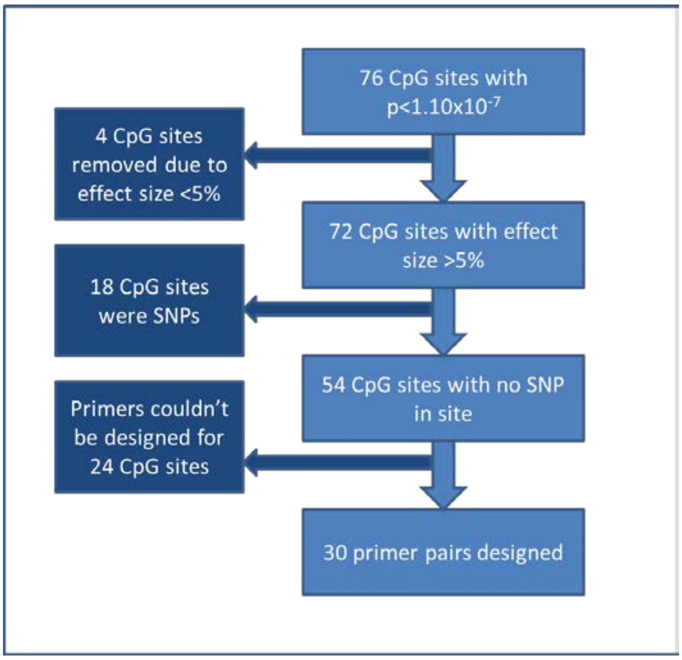

76 CpG sites with $p<1.10\text{x}10^{-7}$
4 CpG sites removed due to effect size <5%
72 CpG sites with effect size >5%
18 CpG sites were SNPs
54 CpG sites with no SNP in site
Primers couldn't be designed for 24 CpG sites
30 primer pairs designed

Figure 22: Flow diagram summarising the selection of top hits that reach genome wide significance.

The HM450 data were then adjusted for cellular composition using the method described by Houseman *et al.* (2012) and the CpGassoc analyses were repeated. The before and after cell adjustment analyses were compared to look for differences. At the majority of CpG sites, cell adjustment did not alter the results at all (p value or F/t statistic). The effect, however, was altered slightly in only three of the 76 selected CpG sites (Table 16).

Variable	**CpG site**	**Pre-adjustment**		**Post-adjustment**	
		F/t statistic	P value	F/t statistic	P value
Braak Staging	cg15703790	40.29 (F)	7.97E-08	23.14 (F)	2.62E-06
Braak Staging	cg00630420	40.23 (F)	8.06E-08	52.39 (F)	1.44E-08
Total CAMCOG	cg02490189	8.78 (t)	6.33E-08	9.24 (t)	2.97E-08

Table 16: CpG sites showing an effect when adjusted for cellular composition.

Since the effects of cellular composition are very small and in some cases strengthen the intensity of the effect, all three sites were still considered to be largely unaffected by cellular composition and were still considered to be potential biomarkers. As seen in Table 15, one of these CpG sites (cg02490189) was taken forward for validation analyses on the Pyrosequencer. These results suggest that differences in cellular composition have very little effect on the methylation differences observed between outcome groups in this Chapter.

3.3.1.5 ***Selection of top hits – approach 2***

The second approach utilised for selecting top hits from the HM450 analysis using ANOVA involved lowering the significance threshold to $p<1x10^{-5}$ and comparing methylation data with publicly available expression data (Maes *et al.*, 2009). The expression data set used is described in Section 3.2.5.

Table 17 shows the number of hits that were selected for Pyrosequencing consideration using the second selection approach. Lowering the significance threshold from $p<10^{-7}$ to $p<10^{-5}$ greatly increased the number of top hits in the selection pool, i.e. the number of significant hits increased from 76 to 224. Of these 224 hits, 44 had expression data available and five of these were differentially expressed in AD. The five remaining hits were then considered in terms of the direction of effect, i.e. whether the CpG sites were hyper or hypo methylated in AD and whether they were up or down regulated in AD. Three CpG sites were found to have the expected methylation-expression relationship (hypermethylated and downregulated in AD or hypomethylated and upregulated in AD).

Variable	Hits significant to <1x10^{-5} (n)	Hits represented on the expression array (n)	Hits differentially expressed (n)	Hits with the expected methylation-expression relationship (n)
Diagnosis	9	4	1	1
Braak staging	147	26	4	2
MMSE	4	0	0	0
Orientation	4	2	0	0
Language comprehension	11	2	0	0
Language expression	3	1	0	0
Memory remote	6	0	0	0
Memory recent	6	2	0	0
Memory learning	5	2	0	0
Memory total	3	1	0	0
Attention	3	1	0	0
Praxis	4	0	0	0
Calculation	4	1	0	0
Abstract thinking	3	1	0	0
Perception	6	1	0	0
Executive function	3	0	0	0
Total CAMCOG score	3	0	0	0
Total number of hits	**224**	**44**	**5**	**3**

Table 17: Number of CpG sites taken forward at each stage in the selection of top hits.

Figure 23 shows the relationship between methylation and expression in each of the significant CpG sites that had both methylation and expression data. Sites (n=3) within the shaded regions showed the expected methylation-expression relationship at the significance threshold imposed. Two loci were identified as being hypomethylated in post-stroke dementia and upregulated in AD. One locus was identified as having the reverse relationship with hypermethylation of a site within the gene in post-stroke dementia and down regulation of the gene in AD. These three sites were taken forwards for further investigation (Table 18).

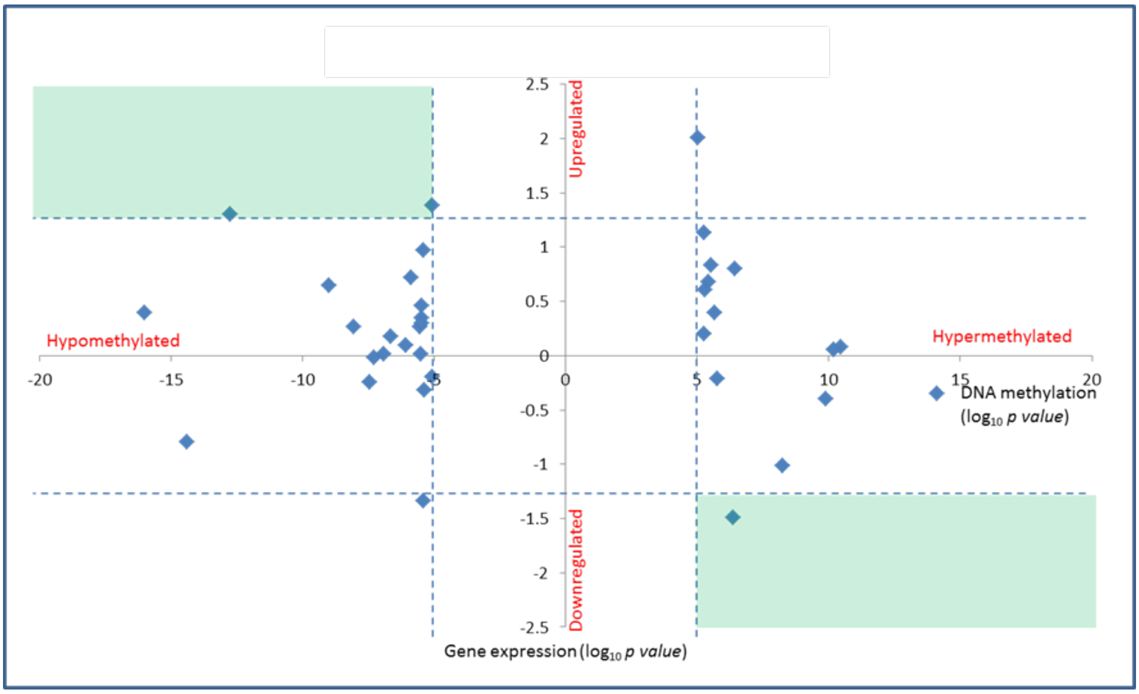


Figure 23: Quadrant plot of DMPs and associated expression levels. DNA methylation (x axis) and gene expression (y axis) are graphically presented as $\log_{10}$ p values. Significance is defined as $<1\text{x}10^{-5}$ for methylation and <1.13 for expression (as indicated by the vertical and horizontal dashed lines). The direction of effect is indicated by the red text. The shaded areas represent sites with the expected inverse relationship between methylation and expression. Sites within these shaded areas were taken forward for further analysis.

Variable	CpG site	Gene	Methylation p value	Methylation effect size	Expression p value	SNP	Primers
Braak staging	cg02055988	*LIMK2*	5.43E-15	0.590	0.049	No	Yes
Braak staging	cg22963863	*SNF8*	4.29E-07	-0.031	0.032		
Diagnosis (D/CN)	cg17633463	*UBTD1*	8.47E-06	-0.013	0.041		

Table 18: Description of 3 CpG sites showing expression differences in AD.

Only one (cg02055988) of the three CpG sites had an effect size >5%. Since cg02055988 reached genome wide significance it has already been selected using the first approach and is included in Table 18. The other two CpG sites (cg17633463 and cg22963863) were dropped due to a small effect size.

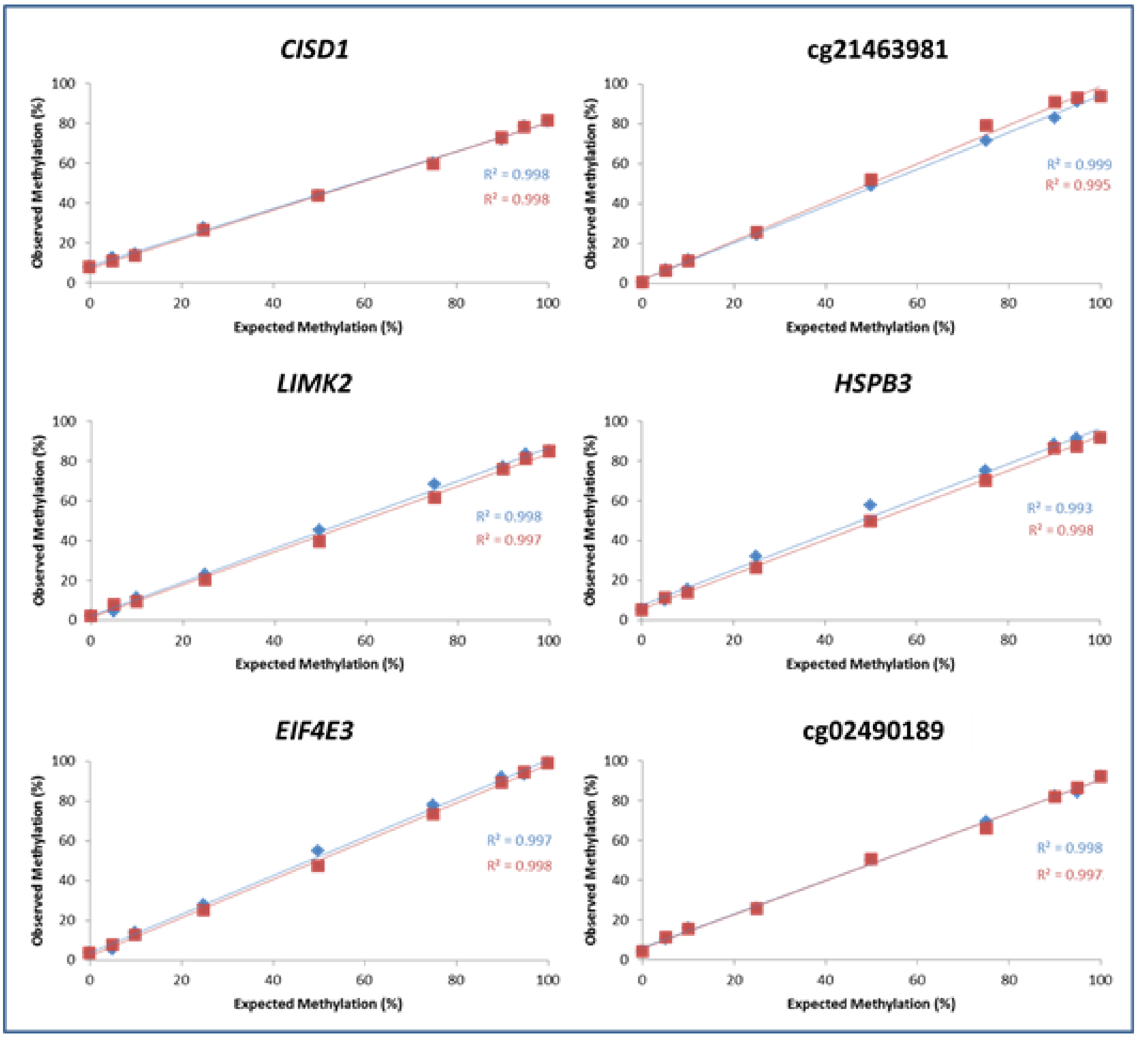
3.3.2 Primer optimisation and validation

Using both approaches to select top hits, primer pairs for 30 loci were successfully designed as in Table 15. All 30 primer pairs targeted at least the index CpG (CpG included on the HM450). Where possible, amplicons were designed to capture neighbouring CpG sites.

Primer pairs for eight loci could not be optimised and were removed from further analysis. The primer pairs for the remaining 22 successfully optimised assays were used to amplify, by PCR, a range of methylation concentrations spanning 0% to 100% to ensure the assay was able to measure the full range of possible methylation values. An assay was considered to be validated if three conditions were reached; 1) the observed methylation value for 100% was above 80%; 2) the observed methylation value for 0% was below 10% and 3) there was a linear relationship with an R^2 value above 0.99. Ten assays were successfully validated. Concordance of observed and reference methylation values are shown in Table 19, using the mean methylation value across all CpG sites in each amplicon. The validation results for both the pre-PCR and post-PCR mixes were plotted as a scatter graph with a trend line and R^2 values presented (Figure 24).

Expected Meth (%)	**Observed Methylation (%)**																			
	C17orf101		*APOB*		cg18837178		*NGF*		*CISD1*		cg21463981		*LIMK2*		*HSPB3*		*EIF4E3*		cg02490189	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
100	92.8	92.8	92.9	92.9	85.7	85.7	93.8	93.8	81.0	81.5	94.0	94.0	84.9	84.9	91.8	91.8	99.3	99.3	92.3	92.3
95	88.5	89.3	89.3	88.0	81.6	81.3	91.1	91.9	78.7	78.2	91.2	93.4	83.5	81.3	91.4	87.3	93.5	94.9	84.2	86.5
90	87.1	83.4	85.1	82.8	75.3	78.0	85.1	87.0	72.3	72.8	83.0	91.0	77.1	76.0	88.4	86.3	92.3	89.5	82.4	81.9
75	77.3	67.7	72.2	68.8	61.5	65.4	75.9	83.0	60.0	59.8	71.8	79.2	68.2	61.7	75.3	70.2	78.0	73.5	69.4	66.2
50	54.6	45.9	54.3	46.9	36.8	43.7	62.0	57.5	43.8	43.9	49.2	52.2	45.2	39.6	57.8	49.8	55.3	47.7	50.2	50.7
25	29.7	23.4	28.2	24.7	26.0	23.6	27.7	29.9	27.6	26.5	24.6	25.7	23.2	20.5	32.1	26.4	28.1	25.2	25.8	25.6
10	11.3	15.6	15.5	11.6	12.7	11.2	15.7	14.3	14.4	13.4	11.7	11.5	11.4	9.4	15.3	13.8	13.9	12.8	16.0	15.5
5	9.1	11.3	9.6	7.8	7.6	7.9	7.5	8.0	12.6	11.0	6.9	6.3	4.4	7.8	10.1	11.2	5.7	7.8	10.7	11.2
0	3.8	3.8	3.4	3.4	3.4	3.4	5.3	5.3	8.0	8.0	0.8	0.8	2.3	2.3	5.4	5.4	3.7	3.7	4.3	4.3

Table 19: Observed mean methylation across all CpGs compared with reference expected methylation level. Pre and post PCR methylation values are presented for each assay. Meth=methylation.


CISD1
Observed Methylation (%)
Expected Methylation (%)
0
20
40
60
80
100
R² = 0.998
R² = 0.998
cg21463981
R² = 0.999
R² = 0.995
LIMK2
R² = 0.998
R² = 0.997
HSPB3
R² = 0.993
R² = 0.998
EIF4E3
R² = 0.997
R² = 0.998
cg02490189
R² = 0.998
R² = 0.997

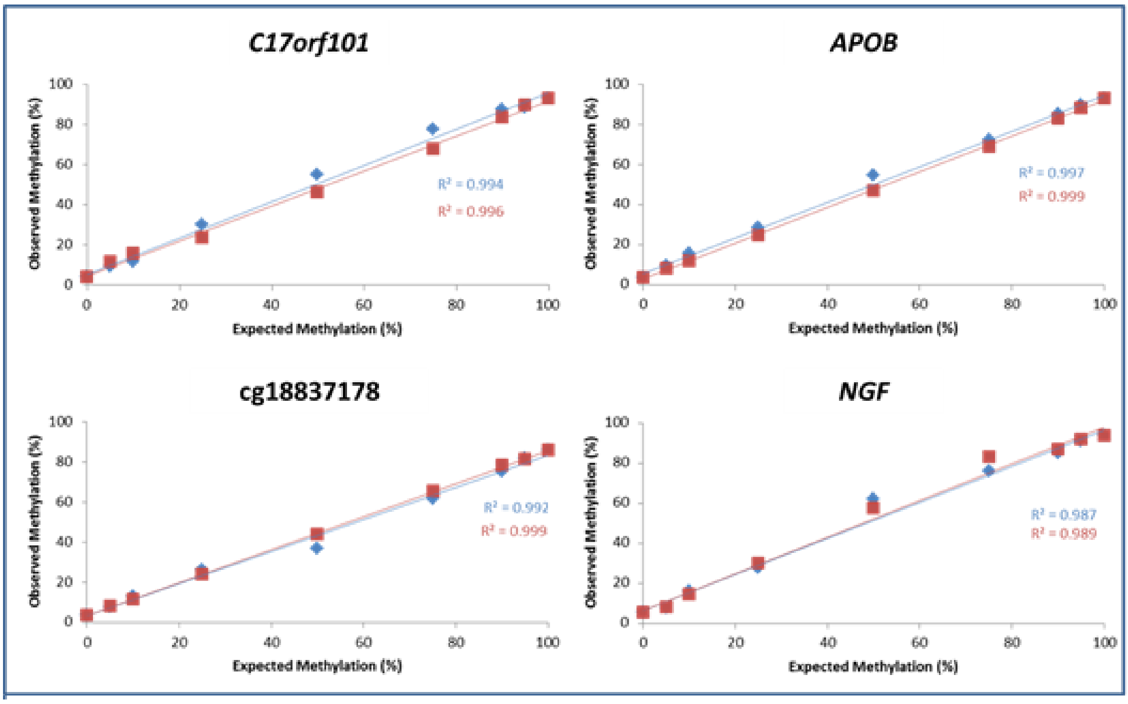


Figure 24: Composite figure of 10 assay validation plots. Pre-PCR methylation values are presented in blue and post-PCR methylation values in red. The R squared value is also displayed in the corresponding colour. An R squared value close to 1 indicates a very close correlation between observed and expected methylation levels.

3.3.3 Pyrosequencing analysis

Methylation was then quantified in duplicate, in all available blood DNA samples across all ten validated assays. Samples were considered to have been assayed successfully, using Pyrosequencing, if the resulting values for duplicates were within 5%. A mean methylation level of these duplicates was calculated for each sample and this mean methylation value was used for all further analyses.

The first stage in the analysis was the comparison of data generated by Pyrosequencing, with the DMP methylation values generated by the HM450 BeadChip, to assess concordance. Then, Pyrosequencing data were analysed to look for association between outcome (adjusted for covariates) and DNA methylation in the “discovery” cohort. If an association was observed in the discovery cohort (samples run on the HM450), data were analysed to look for associations between DNA methylation and outcome in the larger “Pyrosequencing” cohort.

Of the ten technically validated assays, one, *EIF4E3*, a locus significantly associated with Braak staging, did not work on the Pyrosequencer when using the COGFAST blood samples. COGFAST blood DNA was successfully amplified and PCR products were visible on an agarose gel, however the DNA samples repeatedly failed on the Pyrosequencer. Due to the successful amplification of DNA this was likely due to a Pyrosequencing technical error. Results from the remaining nine assays were taken forward for analysis in Stata.

The data from each assay were analysed to test for associations between the outcome variable and methylation at the index CpG site (included on the HM450 BeadChip) in the “discovery” cohort and in a larger “Pyrosequencing” cohort. If multiple CpG sites were included on the Pyrosequencing assay, each neighbouring CpG site was analysed to test for associations with each outcome variable. Finally the mean methylation value across the region (mean of index and neighbouring CpG sites) was included in similar analyses. The results from these analyses will be displayed in the following sections. For seven of the nine assays no significant associations were identified for any CpG site included in the Pyrosequencing assays, indicating that these are likely to be false positive array hits. Six of these assays (*CISD1, LIMK2, HSPB3,* cg21463981, *C17orf101* and cg18837178) had been found, on the HM450, to be associated with Braak staging and one (cg02490189) with total CAMCOG score. These seven loci were thus deemed to be likely false positive array hits. However, two of the validated assays (*APOB* and *NGF*) showed significant associations between methylation at one or more of the CpG sites tested and the selected outcomes and these are described in more detail below.

3.3.3.1 ***APOB and NGF***

3.3.3.1.1 Validation in discovery cohort

All 29 blood DNA samples analysed on the HM450 were run on the Pyrosequencer and results for all duplicate analyses were within 5% of each other for the *APOB* assay and were taken forwards for analysis. Of the 29 samples using the *NGF* assay, there were 28 concordant pairs of data, from which an average methylation value for each sample was taken.

Spearman’s rank correlation was performed in Stata to test the correlation between the two platforms. In neither assay were the methylation values on each platform correlated. Spearman’s correlation gave a coefficient of 0.033 (p=0.865) for *APOB* and 0.028

(p=0.892) for *NGF*. This lack of correlation could be due to the majority of samples having a methylation value around 100% (Figure 25 and Figure 26).

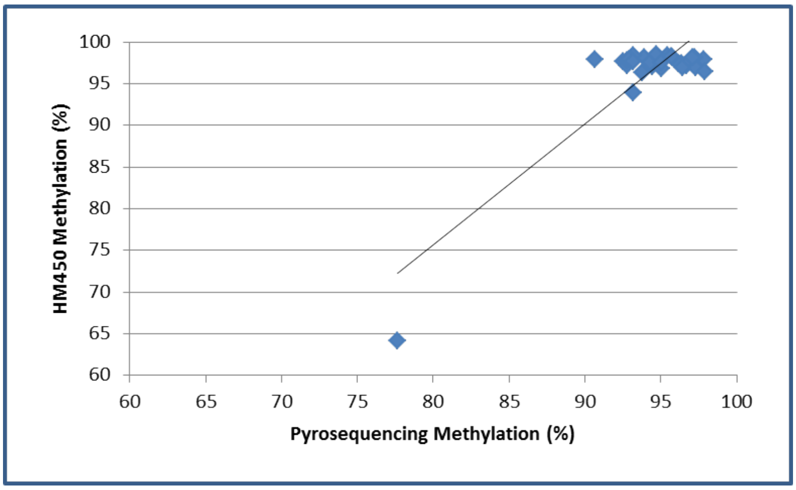


Figure 25: Correlation between methylation measured on the HM450 and Pyrosequencer at the *APOB* locus.

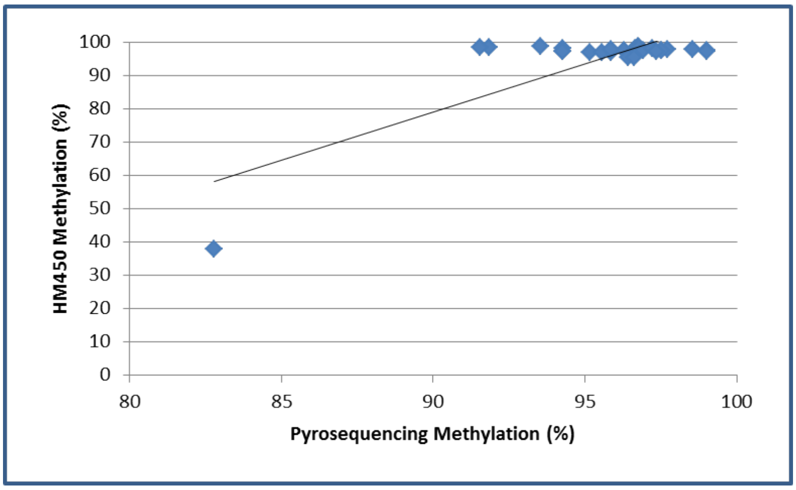


Figure 26: Correlation between methylation measured on the HM450 and Pyrosequencer at the *NGF* locus.

APOB and *NGF* were both highlighted in the HM450 analysis as being associated with Braak staging. When split up into the six Braak stages, the sample size issue becomes more apparent. Table 20 shows the number of samples within each Braak stage category in the discovery sample set. Only one Braak stage VI individual was available in the sample set in the analysis, meaning that any significance differentiating Braak stage VI from others is based on one individual. The HM450 analysis was repeated with this Braak stage VI individual removed and the QQ (Figure 27) and Manhattan (Figure 28) plots revealed that the majority of top hits associated with Braak staging were driven by this one outlier. Only three CpG sites remained significant with this one individual removed.

Braak stage	**N in HM450**	**N in *APOB* discovery cohort**	**N in *APOB* replication cohort**	**N in *NGF* discovery cohort**	**N in *NGF* replication cohort**
I	4	4	5	4	4
II	7	7	14	7	11
III	9	9	11	9	11
IV	5	5	5	4	4
V	3	3	4	3	4
VI	1	1	1	1	1
Total	29	29	40	28	35

Table 20: Number of samples in each cohort by Braak staging.

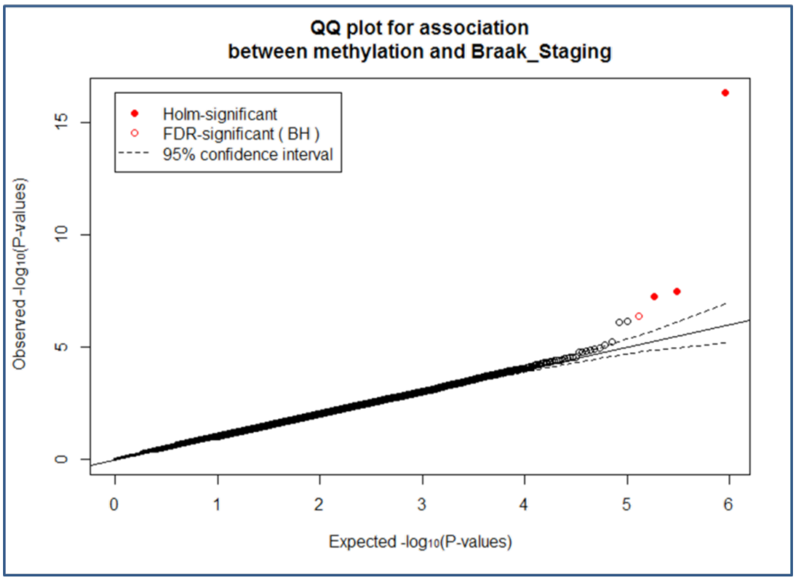


Figure 27: QQ plot for association between methylation and Braak staging with the Braak stage VI individual removed. The QQ plot shows the expected against the observed p values. Data points plotted in red reach the Bonferroni corrected significance threshold.

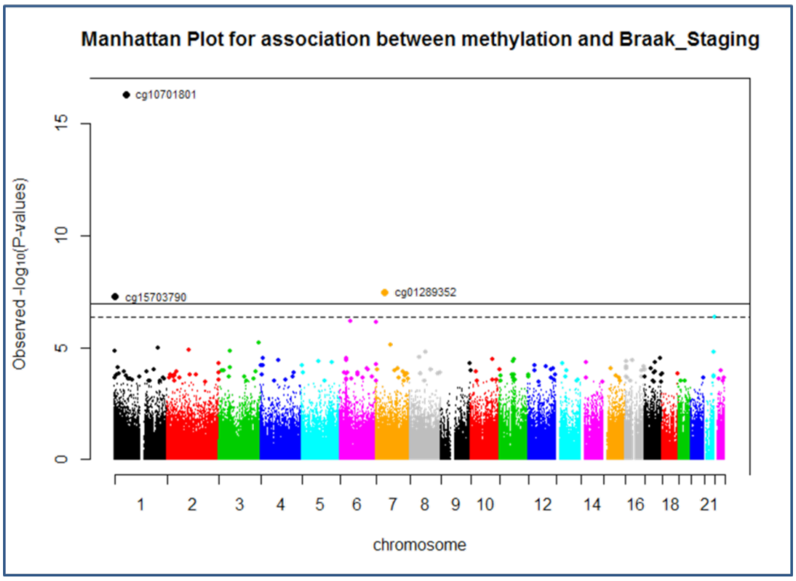


Figure 28: Manhattan plot for association between methylation and Braak staging with the Braak stage VI individual removed. Each circle represents an individual CpG site. The continuous line marks the $p<1.1x10^{-7}$ significance threshold. CpG sites with p values below the threshold are labelled with their Illumina ID.

Analysis of the complete dataset from the HM450 array identified one CpG site within *APOB* and one CpG sites within *NGF* which were differentially methylated in blood samples of stroke patients with varying levels of Braak staging. In both assays, lower methylation values were observed in Braak stage VI when compared to the other Braak stages (I-V).

Figure 29 shows the mean methylation levels for each Braak stage measured at the *APOB* locus and Figure 30 shows the mean methylation levels for each Braak stage measured at the *NGF* locus, on both the HM450 and Pyrosequencer. Both graphs show the mean methylation levels for each Braak stage are similar across the two different platforms in Braak stages I-V. The standard deviation bars indicate that the Pyrosequencing data is slightly more variable than the HM450 in both assays, however on both platforms, the same pattern can be seen in relation to Braak staging, with the Braak stage VI individual having a much lower methylation level at both loci.

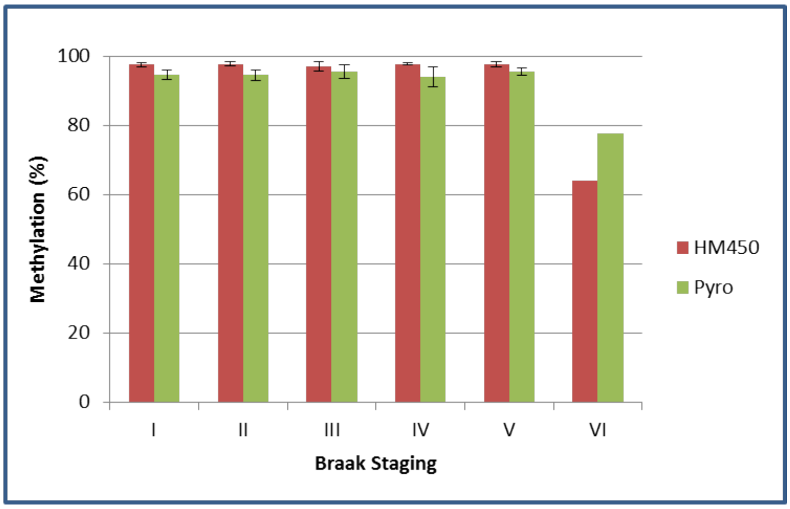


Figure 29: Mean *APOB* locus methylation levels observed for each Braak stage measured on the HM450 and Pyrosequencer.

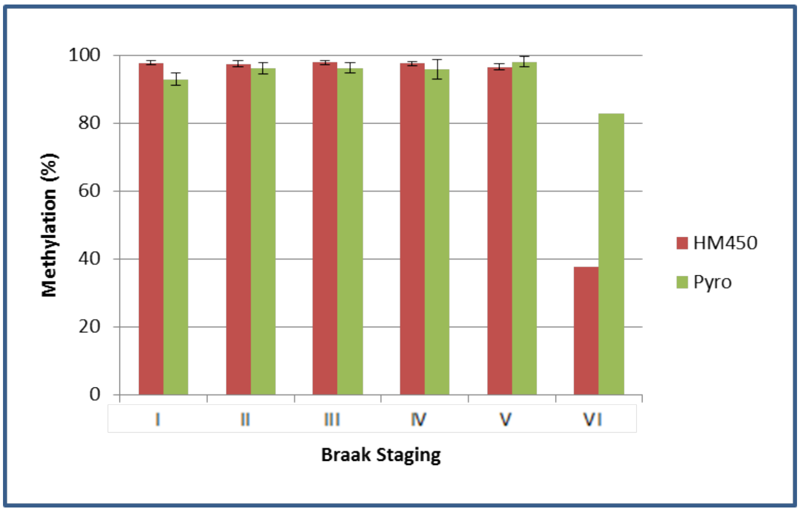


Figure 30: Mean *NGF* locus methylation levels observed for each Braak stage measured on the HM450 and Pyrosequencer.

Figure 31 and Figure 32 show the spread of methylation measured at the index CpG site using both platforms, for *APOB* and *NGF*, respectively. Both graphs show the Braak stage VI individual as an outlier on both platforms; however the difference between this individual and the mean methylation level across all samples is much lower when measured using Pyrosequencing than the HM450.

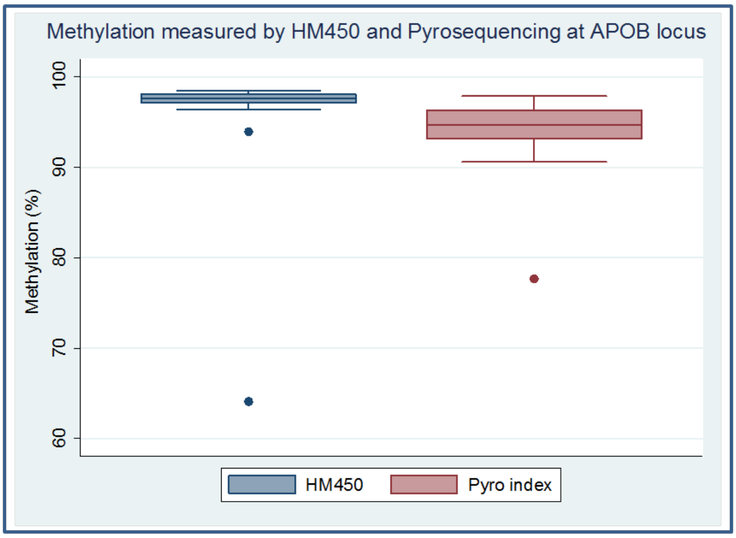


Figure 31: Box plot to show methylation levels measured using the HM450 and Pyrosequencing at the index CpG site in *APOB.*

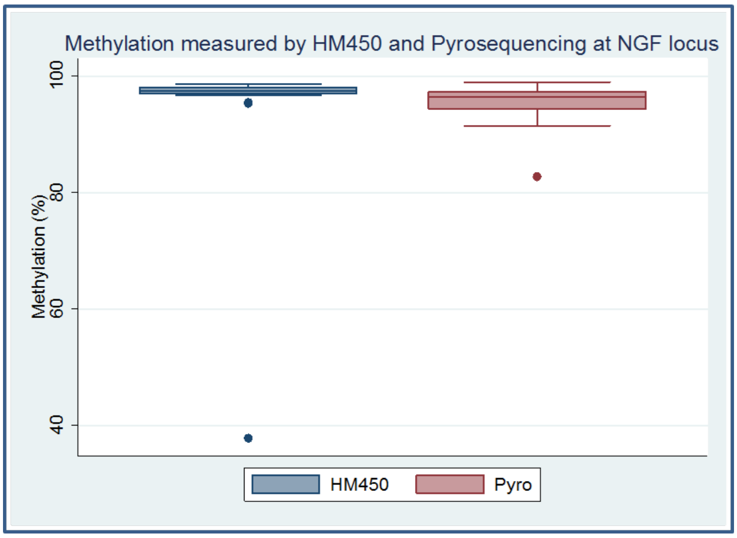


Figure 32: Box plot to show methylation levels measured using the HM450 and Pyrosequencing at the index CpG site in *NGF*.

The analysis model used to assess the relationship between Braak staging and the HM450 data was applied to the Pyrosequencing data. This was to test whether the association between methylation and Braak staging was still present in the Pyrosequencing data. In Stata, ANOVA was performed (as in the analysis of HM450 data) to test for association between methylation at the index site and Braak staging, using sex and baseline age as covariates. 29 blood DNA samples were included in the *APOB* analysis and 28 were included in the analysis of the *NGF* locus. The results of the HM450 were validated using Pyrosequencing at each loci; *APOB* p=0.0003; *NGF* p= 0.008. However, the methylation data were found to have a non-normal distribution using the Shapiro-Wilk normality test. A Kruskal Wallis test was therefore applied to assess whether DNA methylation differed between Braak stages in both the *APOB* and *NGF* assay. In both assays, the results of the Kruskal Wallis suggests there was no difference in DNA methylation between the Braak stages (*APOB* p= 0.338; *NGF* p= 0.392). The Kruskal Wallis test is equivalent to ANOVA but does not assume a normal distribution of the data, which ANOVA does. It compares distributions rather than means of the two groups. The Kruskal Wallis result is therefore likely to be a true reflection of the statistical comparison as the assumptions of ANOVA are violated by the non-normal distribution of the methylation data in this comparison.

When the individual with Braak VI stage pathology was removed from the analysis, there was no significant association with methylation at either locus (*APOB* and *NGF*) and Braak staging using ANOVA (*APOB* p=0.119; *NGF* p=0.334), or Kruskal Wallis

(*APOB* p=0.543; *NGF* p=0.414). This indicates that the results are heavily influenced by this outlier.

3.3.3.1.2 Validation in Pyrosequencing cohort

The Pyrosequencing assay was run across a larger sample set (*APOB*: n=40; *NGF:* n=35) to test whether the association with Braak staging at the index CpG site was still observed in a larger cohort. ANOVA revealed that the association between Braak staging and methylation at the *APOB* locus was more significant using the larger cohort (p=0.0001) than the discovery cohort, whereas Kruskal Wallis highlighted no significant differences in DNA methylation (p=0.264). Likewise, the association between Braak staging and methylation at the *NGF* locus was also more significant when a larger samples size was used in ANOVA analysis (p=0.007) but not with Kruskal Wallis (p=0.505).

3.3.3.1.3 Neighbouring CpG sites

It is also possible to look at the methylation levels of neighbouring CpG sites i.e. CpG sites close to the index CpG site, to assess whether these CpG sites are also associated with Braak staging. The *APOB* assay included the index CpG site along with one neighbouring site. The *NGF* assay included the index CpG site plus three neighbouring sites. The Pyrosequencing data for the neighbouring CpG sites were also analysed using ANOVA and Kruskal Wallis in both the HM450 cohort and the larger Pyrosequencing cohort. Results of ANOVA and Kruskal Wallis analyses are presented in Table 21.

CpG site	P value	
	APOB	*NGF*
Index CpG		
Discovery	0.0003 (0.338)	0.008 (0.392)
Pyrosequencing	0.0001 (0.264)	0.007 (0.505)
Neighbouring site 1		
Discovery	0.677 (0.774)	0.488 (0.362)
Pyrosequencing	0.617 (0.804)	0.633 (0.521)
Neighbouring site 2		
Discovery	NA	0.115 (0.422)
Pyrosequencing	NA	0.075 (0.306)
Neighbouring site 3		
Discovery	NA	0.343 (0.338)
Pyrosequencing	NA	0.821 (0.605)
Region		
Discovery	0.017 (0.533)	0.339 (0.484)
Pyrosequencing	0.007 (0.376)	0.500 (0.694)

Table 21: ANOVA (and Kruskal Wallis) results for *APOB* and *NGF* Pyrosequencing. The methylation of the region is defined as the mean methylation across all CpGs covered in the Pyrosequencing assay. P values are displayed for both the discovery cohort and Pyrosequencing cohort for each CpG site/region.

Although both index CpG sites within *APOB* and *NGF* were found to be significantly associated with Braak staging in both cohorts, no neighbouring CpG within the *APOB* and *NGF* locus, measured by Pyrosequencing, was associated with Braak staging using either the discovery or Pyrosequencing cohort. When the mean of all CpG sites in the amplicon were analysed using ANOVA, the methylation of the *APOB* locus remained significantly associated with Braak staging whereas the *NGF* locus did not. No difference in DNA methylation was detected between Braak stages when the non-parametric Kruskal Wallis test was performed but this is likely due to the small sample size limiting the power to find differences.

These data show that although the data generated on the HM450 and Pyrosequencer were not identical, nor found to be correlated by Spearman’s rank (possibly due to the non-linear relationship with almost all methylation values being close to 100%), the same patterns were seen in relation to Braak staging. The HM450 findings that the Braak VI individual had a significantly lower methylation level compared with the other Braak stages at the index CpG sites within *APOB* and *NGF* was validated using Pyrosequencing.

The major caveat of this study is the small sample size especially when looking at an outcome measure such as Braak staging, which splits samples into several small groups which could indicate why analysis with Kruskal Wallis did not show significant differences in DNA methylation levels between the Braak stages. Only one COGFAST participant was diagnosed with Braak stage VI pathology meaning that the differential methylation seen in this Chapter, at both the *APOB* and *NGF* loci, is in one individual only.

To test whether the observed differential methylation seen in the COGFAST individual was also detected in other demented Braak stage VI individuals, five blood DNA samples were sought from other Braak stage VI individuals from the NBTR and analysed using both the *APOB* and *NGF* Pyrosequencing assays. All five blood samples were from demented patients with Braak stage VI pathology and none of them were known to have had a stroke prior to death.

Figure 33 shows the methylation level of the index CpG within *APOB*, measured by Pyrosequencing, in these five individuals compared with the COGFAST participants. Figure 34 shows the methylation level of the index CpG within *NGF*, measured by Pyrosequencing, in these five individuals compared with the COGFAST participants.

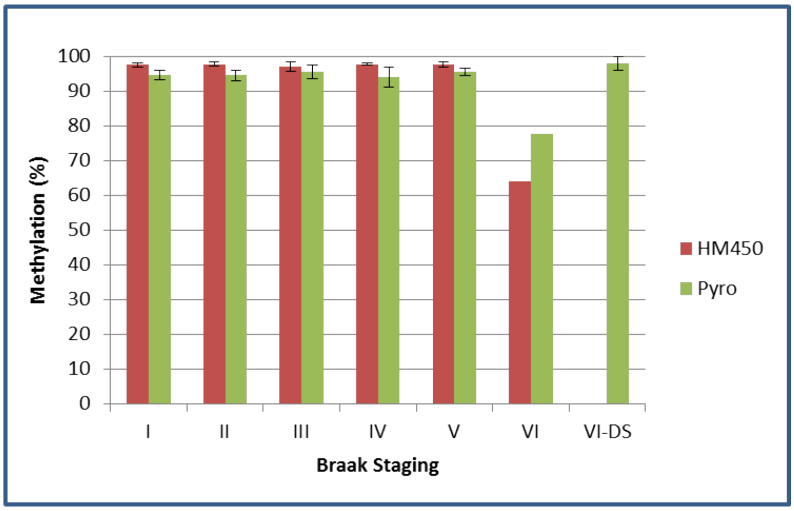


Figure 33: Methylation at *APOB* locus measured in COGFAST and demented Braak stage VI individuals. Braak Stages are displayed on the X axis with demented Braak stage VI denoted as VI-DS.

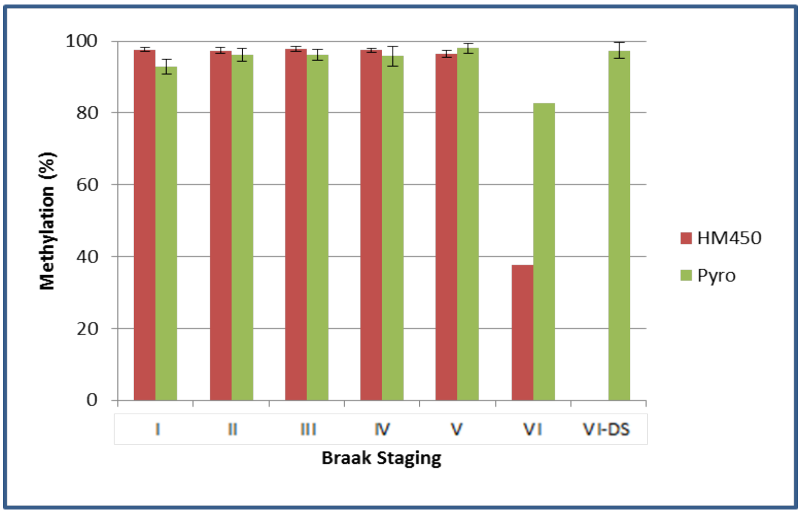


Figure 34: Methylation at *NGF* locus measured in COGFAST and demented Braak stage VI individuals. Braak Stages are displayed on the X axis with demented Braak stage VI denoted as VI-DS.

In both assays, the COGFAST Braak stage VI individual was unique in its methylation patterns, indicating that the results from this Chapter are not representative of all those with Braak stage VI pathology. However, this finding does not necessarily mean that this is a false positive. There are several reasons why this individual may have differential methylation at these loci. Possible explanations are explored in the discussion section below.

3.4 Discussion

The work carried out in this Chapter aimed to identify potential biomarkers of post-stroke dementia using blood samples taken three months post-stroke, when the stroke survivors were cognitively normal. The HM450 array identified 76 potential CpG sites associated with post-stroke dementia, the vast majority of which were found to be associated with Braak staging. Following filtering of top hits, ten assays were run across a larger COGFAST cohort to test whether the association observed between Braak staging and methylation persisted, when measured using Pyrosequencing. Two of these assays (*APOB* and *NGF*) successfully validated the HM450 findings using ANOVA, suggesting that the observations were confirmed using an alternative technology and that they are potential biomarkers for post-stroke dementia. These novel results were promising since both of these genes have previously been implicated in cardiovascular (Gigante *et al.*, 2012) and dementia related diseases (Zhang *et al.*, 2013).

Apolipoprotein B (APOB) is present in two forms in humans; APOB-100 and APOB-48 (Benn, 2009), and is the protein component of lipoproteins (Benn *et al.*, 2007), which have a role in transporting cholesterol and triglycerides in the blood. The role of APOB is vital, since an accumulation of excess cholesterol can lead to atherosclerotic plaques (Dallmeier and Koenig, 2014). An increased level of plasma APOB, believed to reflect increased levels of low-density lipoprotein (LDL), has been associated with atherogenesis (Gigante *et al.*, 2012) and is a major risk factor for atherosclerosis and IHD (Sule *et al.*, 2009; Vaverkova *et al.*, 2009). It has been widely studied in many cohorts of complex disease (Sato *et al.*, 1991; Ford *et al.*, 2013). Increased APOB levels have also been reported to be associated with increased arterial stiffness and arterial pulse wave velocity in adults under the age of 45 years (Koivistoinen *et al.*, 2011). Screening for atherosclerosis usually involves measuring levels of LDL cholesterol in the blood; however increasingly it is being suggested that APOB measures could be a better predictor of atherosclerosis risk (Benn, 2009; Sniderman *et al.*, 2012).

Whilst APOB and LDL levels are widely believed to increase the risk of cardiovascular diseases and IHD, it is only in recent years that increased APOB levels have also been associated with an increased risk for ischaemic stroke (Walldius *et al.*, 2006; Benn, 2009). These findings have been supported by studies identifying a reduction in risk of ischaemic stroke by treatment with statins, which can lower APOB levels by 25-45%

(Heart Protection Study Collaborative Group, 2002; Amarenco *et al.*, 2006). High APOB levels have been associated with the thickening of intima-media, which could indicate the mechanistic link between atherosclerosis and stroke (Walldius *et al.*, 2006).

APOB has also been linked to dementia. Aggregation of β-amyloid (Aβ) deposits is regarded as the hallmark of Alzheimer's disease and, in several tissues including the blood, APOB has been found to co-localise with Aβ at an intra-cellular level (Kuo *et al.*, 1998; Galloway *et al.*, 2009; Lam *et al.*, 2011). A study comparing levels of a range of lipid-related measures between AD patients and healthy controls reported elevated levels of APOB in the blood of AD patients (p=0.004) (Caramelli *et al.*, 1999). These studies highlight a possible mechanistic link between APOB and AD; however these mechanisms are currently unknown. The findings presented in this thesis indicate a possible mechanistic link for the aberrant methylation of *APOB* in post-stroke individuals with Braak stage VI pathology. At the time of writing, no papers had been published that reported the methylation patterns of *APOB* in any cardiovascular-related cohorts indicating that this may be a novel finding. Replication in a larger cohort is required to investigate whether this association is seen when more individuals with Braak stage VI pathology are included.

Nerve growth factor (*NGF*) is a member of the neurotropic family and has an essential role in the development and maintenance of neurons, both during development and throughout adulthood. NGF is produced in several regions of the brain including the cortex and hippocampus. Since NGF has been found to have a role in the maintenance and survival of cortical neurons, it has been suggested that NGF may have a potentially protective role in conditions such as ischaemia and Alzheimer's disease (Aloe *et al.*, 2012). It is also thought that NGF could be delivered to the brain to increase neurogenesis thereby enhancing brain recovery (Lee *et al.*, 1998; Greenberg and Jin, 2006).

Many studies have compared the levels of *NGF* expression and protein abundance between AD brains and healthy controls. Several papers report increased levels of *NGF* expression in the CSF (Hock *et al.*, 2000; Mashayekhi and Salehin, 2006) and brain regions of AD patients (Fahnestock *et al.*, 2001). Other papers have reported conflicting findings with increased degradation of NGF in AD brains compared with controls

(Bruno *et al.*, 2009) and reduced *NGF* expression levels in AD affected brains (Higgins and Mufson, 1989). These apparently conflicting findings could be due to the heterogeneity of the brain regions studied or differences in the severity of disease. Hellweg *et al.* (1998) reported differences in NGF proteins levels between preclinical cases, AD cases and healthy controls. Preclinical cases had reduced levels of NGF compared with healthy controls, whereas cases with later stage AD had much higher levels of NGF protein (Hellweg *et al.*, 1998). The reasons for, and mechanisms which result in altered NGF levels, are currently not well understood. Although there is conflicting evidence, there is a general consensus that NGF is likely to be involved in the pathogenesis of AD. NGF is explored in more detail in Chapter 4.

The major caveat of this study is the small sample size resulting in the most noteworthy associations being driven by the methylation profile of one individual. In both loci, the lone COGFAST individual with Braak stage VI pathology had a reduced methylation value when compared to other COGFAST participants with a lower Braak staging. The methylation profile seen in the COGFAST Braak stage VI individual was not seen in non-stroke, demented individuals with Braak stage VI pathology. This individual is clearly different in terms of their methylation profile but it would appear (due to the lack of replication in other high Braak stage individuals) that this is not attributable to Braak stage pathology *per se*. One plausible explanation is that this individual has some particular unmeasured feature that confounds the observed association. It would be extremely interesting to extend this study, using more blood DNA samples taken from stroke survivors who later developed Braak stage VI pathology. The possibility that this individual could have a lower methylation value across the genome was investigated. The mean beta level across all probes on the HM450 was 49.5% for each blood DNA sample indicating that this Braak stage VI individual did not experience a genome-wide methylation difference. COGFAST is a complex cohort with a potentially heterogeneous mix of underlying risk factors and this heterogeneity may contribute to the difficulty in elucidating clear DMPs. Due to the small sample size of COGFAST it was necessary for the Pyrosequencing cohort to include those samples used in the discovery phase (HM450) of the study. For this reason, the validation phase (Pyrosequencing) of this study cannot be considered as a true replication. It was decided that all 46 samples would be included in the Pyrosequencing study rather than just the 16 samples which had not been included on the HM450 array, as a sample of sixteen would not allow for significant differences in DNA methylation to be detected. To be able to replicate the findings, this study would require analysis in a larger independent cohort.

The analysis model used is robust at the population level, however in order to explain differences at an individual level, the clinical characteristics of the COGFAST subject with Braak stage VI pathology has been investigated, to identify the possible reasons for the *NGF* and *APOB* methylation deviations from the norm. It would not be practical, due to time limitations, to look at this level of detail in every participant due to too many possible contributing factors. This level of detail is only useful when investigating difficult to explain results. If the influences on methylation are subtle then the population based approach used in this thesis can be too noisy to identify these differences, highlighting a need for a more in depth assessment of possible contributing clinical factors. Table 22 describes the clinical features for this individual.

As shown in Table 22 this individual was initially diagnosed as having vascular dementia but this diagnosis was questionable. The individual had a high burden of brain vascular lesions as well as Braak stage VI neurofibrillary pathology indicating that a diagnosis of mixed dementia was more appropriate. Due to the fluctuations in cognition identified following the index stroke, there is a possibility that this individual may have had DLB pathology, since fluctuations are a common symptom of DLB (Taylor et al 2013); however, a DaTscan was never performed to investigate this, so the absolute diagnosis of dementia type is somewhat incomplete[1]. This subject differs from other participants in a number of ways. This subject was recruited into COGFAST via a different route. Stroke physicians were the usual means of referral, however this individual was referred in by an Old Age Psychiatrist, meaning that the index stroke was not well characterised. The three strokes experienced prior to the index stroke are also ill-defined with little information available on the location and severity of strokes. COGFAST recruited participants who were referred following their first known stroke. This distinguishes the COGFAST Braak stage VI individual from the others since this case had experienced multiple strokes and was likely to have a higher burden of vascular pathology. The effect these previous strokes may have on DNA methylation and how this links in with disease is explored in Chapter 4.

[1] The same clinician (Louise Allan, personal communication), now more experienced, revisited this case and gave her expert opinion on the diagnosis of this case.

Clinical Feature	Description
Diagnosis	Vascular dementia
Previous stroke history	Three recurrent strokes in 1995 Evidence of a left occipital lobe infarct
Previous stroke symptoms	Left sided weakness Dizziness Confusion Slurred speech Delirium
Recruitment route into study	Old Age Psychiatrist
Index stroke	Right posterior parietal infarct
Cognitive symptoms post-index stroke	Stepwise cognitive decline Accelerated memory loss Disorientation Fluctuations in cognition Aggressive behaviour Assistance required with all personal activities of daily living
Past medical history	Hypertension Atrial fibrillation Hypercholesterolaemia Chronic Obstructive Pulmonary Disease Colonic polypectomy Falls
Drug history	Warfarin Pravastatin Bedroflumethiazide Temazepam Amlodipine Salbutamol Beclomethasone inhalers Cod liver oil

Table 22: Clinical features of COGFAST Braak stage VI individual.

The initial study design of COGFAST had one major limitation. The study design did not allow for recurrent strokes, occurring after the index stroke, to be recorded. Consequently, it is possible that a number of participants had multiple strokes following recruitment into the study, but this information was not recorded. These recurrent strokes could have had an impact on both DNA methylation and cognition. For example, it is possible that those participants who developed dementia following the index stroke had experienced additional strokes, thereby making them more susceptible to cognitive decline (Tatemichi *et al.*, 1993; Leys *et al.*, 2005). This information would be vital for assessing the effect of multiple strokes on cognition.

Although DNA methylation analyses were performed on a total of 46 blood DNA samples, differential DNA methylation was only observed in one individual. However, the findings may still be of interest. Validation in a larger cohort is required before these loci can be considered as possible biomarkers. However, the methylation loci identified appeared to only discriminate the most severe end of the spectrum of Braak stage phenotype, casting doubt on whether they would be a useful sensitive biomarker of prediction.

A number of further steps could be undertaken to give a more thorough analysis of the possible biomarkers for PSD. The HM450 data highlighted a number of CpG sites (n=24) that were differentially methylated between outcome groups which were not amenable to Pyrosequencing. There were several reasons why primers could not be designed for these CpG sites such as the region surrounding the index CpG being a densely populated CpG island. These CpG sites were not followed up and analysed using Pyrosequencing, however these CpG sites could still be potential biomarkers for PSD. These assays could be revisited and analysed using an alternative candidate gene platform such as the Sequenom® Epityper™ (Ehrich *et al.*, 2005) or bisulphite sequencing (Herman *et al.*, 1996). Similarly a number of CpG sites were removed because they were in fact a SNP or there was a SNP present in the probe sequence (n=18). Some of these SNPs had a very low frequency in a European population and were perhaps excluded when they may not have affected findings. These SNPs could therefore also have been potential biomarkers that were excluded from validation analyses. A number of CpG sites also dropped out of the validation phase due to failure of PCR optimisation (n=8). However, these CpG sites could also be potential biomarkers for PSD and the DNA methylation status of these CpG sites could be investigated using an alternative approach. To select CpG sites to validate by Pyrosequencing, two approaches were followed. The first selected hits solely based on p value. The second looked at publicly available expression data. The selection process implemented only considered those sites which showed an inverse relationship between methylation and expression, to help narrow down those sites most likely to be functionally related to PSD. However, this selection criterion may have excluded some potential biomarkers. An inverse relationship between methylation and expression is only expected within promoter regions. It is less clear that this relationship is true in other regions of the gene (Jjingo *et al.*, 2012; Jones, 2012). For this reason it may be possible that any of the genes previously identified as being differentially expressed in AD may be potential biomarkers for PSD, regardless of the direction of effect. In addition, the selection of top hits were based on expression data from only one study.

Taking an average from multiple expression studies would have been a better, more accurate approach. Furthermore, due to the absence of available PSD expression data, the expression dataset used to select the top hits was composed of AD cases so not entirely comparable to PSD cases.

In addition, this work only looked at individual CpG sites and their associations with outcome. It would be interesting to analyse the data in an alternative way and look at DMRs in addition to DMPs. Gene set enrichment could be used to identify correlated networks of loci that differ between groups. Furthermore, given that the epigenome-wide association study (EWAS) approach used in this Chapter highlighted two loci previously implicated in stroke and dementia, a candidate or pathway based approach could be adapted to only analyse “known” loci involved in disease. This would therefore reduce the multiple testing burden caused by EWAS but would eliminate the possibility of uncovering novel biomarkers.

Due to the heterogeneous nature of blood, it is possible that any differences in DNA methylation observed between outcome groups is due to the cellular composition of the blood. Methylation beta values were adjusted for cellular composition to determine whether the differential methylation observed in relation to Braak staging were actually due to differences in cellular composition. Adjusting for cellular composition revealed very little difference in significance indicating that differences in cellular composition have very little effect on the methylation differences observed between outcome groups in this Chapter.

Chapter 4. Epigenetic mechanisms in post-stroke dementia

4.1 Introduction

In addition to identifying biomarkers for clinical use in the prediction and prognosis of post-stroke dementia (PSD), this project aims to identify possible mechanisms involved in the pathogenesis of PSD.

As described in Chapter 3, epigenetic mechanisms including altered DNA methylation status, have been implicated in an increased risk of both stroke (Endres *et al.*, 2000; Castillo-Diaz *et al.*, 2010) and dementia (Wang *et al.*, 2008; Bollati *et al.*, 2011; Bakulski *et al.*, 2012a; Chouliaras *et al.*, 2013). In many instances, although a reliable association is observed between DNA methylation and a disease, a causal role of perturbed DNA methylation has not been demonstrated. To identify differential DNA methylation which may have a mechanistic role in disease, the analysis of the diseased tissue (in this case the brain) is required.

As previously stated, the mechanisms which result in PSD are currently largely unknown (Section 1.2.3.5), but epigenetics could be one such mechanism. There have been several studies highlighting the potential role of DNA methylation in the pathophysiology of a number of diseases including cancer (Stefansson and Esteller, 2013) and Alzheimer's disease (Mastroeni *et al.*, 2009; Chouliaras *et al.*, 2010; Mastroeni *et al.*, 2011), but to date there have been no studies investigating the role of DNA methylation in the pathophysiology of PSD. Epigenetic marks are increasingly being regarded as potential mechanistic links between exposures and disease risk, especially in the field of cancer.

This approach is yet to be applied to PSD. It is postulated that DNA methylation signatures in the brain could indicate possible genes and pathways involved in the pathogenesis of disease. DNA methylation analysis alone is not sufficient to be able to infer causality, other measures such as gene expression and protein levels are required to strengthen the relationship. The COGFAST cohort was used to test whether DNA methylation alterations found in the brains of stroke patients can identify possible mechanisms involved in the pathogenesis of PSD.

4.2 Experimental design

General methods are described in Chapter 2. Many of the methods used in this Chapter overlap with these that have already been described in Section 3.2. Additional details specifically relevant to this investigation are described in detail below.

4.2.1 Study cohort

Brain samples collected from participants of the COGFAST cohort (Section 2.1.1) were utilised in this study. The regions of the brain selected for use in this study were the dorsolateral prefrontal cortex (DLPFC) and the hippocampus. A diagnosis of dementia was given based on neuropathological examination in combination with results of a series of cognitive tests. Data relating to clinical history, including details of stroke severity as well as stroke risk factors, were available to analyse. Age at death, post-mortem (PM) delay, Braak staging and years between stroke and death were collected at death. All other variables were collected at baseline (recruitment into study) and are described in the Glossary (Table 2). Paired (from the same individual) DLPFC and hippocampal samples were divided into a 'discovery' sample (n=30) and 'Pyrosequencing' sample (n=46), the latter consisting of the discovery sample plus an additional sixteen samples drawn from the same cohort.

4.2.2 Tissue DNA extraction, quantification and precipitation

Sections of DLPFC and hippocampus were held in long-term storage at -80°C following sampling and were selected and removed from storage for the purposes of this study. DNA was extracted from 30µg DLPFC and 10µg hippocampus using the Tissue DNA Spin protocol as outlined in the EZNA Tissue DNA Kit (OMEGA bio-tek, USA) (Section 2.2.1). DNA samples were quantified using a ND1000 Spectrophotometer (Labtech International Ltd, UK) (Section 2.2.3). Samples with a concentration lower than 50ng/µl were ethanol precipitated and re-suspended in 30µl DEPC-treated water (Section 2.2.4).

4.2.3 Illumina HumanMethylation450 BeadChip analysis

Epigenome wide discovery analysis was performed using the Illumina HM450 BeadChip. Thirty individuals with >1µg paired DLPFC and hippocampal derived DNA were selected for discovery phase DNA methylation analysis. Section 3.2.3 describes how the DNA samples were processed and how the HM450 data were filtered and normalised.

4.2.4 Analysis of HM450 data

A total of 431,832 probes and 24 samples were included in the final DLPFC dataset. A total of 434,422 probes and 28 samples were included in the final hippocampal dataset. As in Chapter 3, exploratory analyses were performed across the methylation data by considering distribution density plots, mean-sd plots and hierarchical clustering with heat plots. The CpGassoc package (Barfield *et al.*, 2012), which performs multiple linear regression analyses with a continuous predictor variable and ANOVA for categorical predictor variables was implemented in R (version 2.15.0). CpGassoc was performed to test for associations between methylation and; diagnosis (D/CN), the last MMSE and CAMCOG scores before death and Braak staging as predictor variables. For all analyses, age at death, sex and chip (i.e. the microarray chip upon which the samples were scanned) were included in the analysis model as covariates. Data generated provided a beta value as outlined in Section 3.2.4. To test for sensitivity to cellular composition, HM450 data were adjusted for cellular composition using the method described by Guintivano *et al.* (2013). Statistical analyses were then repeated to assess whether cell composition accounted for any significance observed between methylation and the outcome variable.

4.2.5 Selection of top HM450 hits

The selection of top HM450 hits is described in detail in Section 3.2.5. Brain data were treated in the same way as blood data, the only exception being the expression dataset used in the second approach. In this Chapter, GSE36980, a study which compared expression levels in both the frontal cortex and hippocampus between AD patients and healthy controls was used (Hokama *et al.*, 2013). GSE36980 was chosen due to both brain regions of interest being included. However, GSE36980 uses AD cases whereas this study is interested in PSD cases so results are not entirely comparable. In brief, two approaches were taken. In the first approach, p values above the cut off imposed for multiple test correction (DLPFC: 0.05/431,832; hippocampus: 0.05/434,422) were not considered. The second approach included all CpG sites with a p value $<1x10^{-5}$ (to include more hits) and compared methylation data with publicly available expression data. Loci were selected for further analysis if they displayed the expected DNA methylation-expression relationship (i.e. hypermethylated and under expressed, hypomethylated and over expressed). This inverse relationship between methylation and expression is only expected in promoter regions, it is less clear that it is always the case in other regions such as gene bodies, indicating that this may be a potential limitation of the study. The next criterion in both approaches was that the effect size (difference between the outcome groups) was more than 5% due to the suggested technical limitation of the Pyrosequencer (Mikeska *et al.*, 2011). Finally, a BLAT search was performed to test for the presence of a SNP within the CpG site. If the CpG site was indeed a SNP, or if a SNP was identified within the probe sequence, this site was removed from consideration. This step removed any CpG sites likely to be directly under genetic influence.

4.2.6 Pyrosequencing

Details of primer design (Section 2.6.2), primer optimisation (2.6.3) and assay validation (Section 2.6.5) can be found elsewhere (Section 3.2.6).

Pyrosequencing assays (n=2 in DLPFC; n= 5 in hippocampus) were performed on the discovery sample (n=24 in DLPFC; n=28 in hippocampus) along with brain derived DNA samples from an additional sixteen individuals drawn from the same cohort. 500ng of DNA was first bisulphite modified, amplified using the optimised Pyrosequencing PCR conditions in Table 23 and finally sequenced on a PyroMark MD Pyrosequencer (Qiagen, UK) following the standard protocol (Section 2.6.6). Details of all relevant Pyrosequencing primers are provided in Table 24.

Tissue	Gene/DMP	HM450 probe ID	Tm(°C)	$MgCl_2$ +/-	Size of product(bp)	No. CpGs	Successfully validated using Epitect controls
D,H	*CTRC*	cg03096785	50.9	-	144	1	
D,H	cg20583640	cg20583640	54.5	-	168	6	
D,H	cg18837178	cg18837178	54.5	+	143	2	Yes
D,H	*NGF*	cg00794813	53.6	-	305	4	Yes
H	*EIF4E3*	cg01228342	50.9	-	118	1	Yes
H	*LMNA*	cg15447017	53.6	-	298	3	
H	*CTRL*	cg02126424	50.9	-	155	3	
H	cg01063243	cg01063243	56.9	+	394	9	Yes
H	cg12010173	cg12010173	51.5	-	341	4	Yes
H	*KIAA1026*	cg02232840	50.4	-	265	2	

Table 23: Optimal PCR conditions for each assay. Tissue - D=DLPFC, H=Hippocampus. Tm=annealing temperature.

Tissue	Gene	Forward primer (5’>3’)	Reverse primer (5’>3’)	Sequencing primer (5’>3’)
D,H	*NGF*	GGAATTATATTTAGAGAGTAA	5Biosg/TACTCCTATAAATCCTATTAA	ATTTTAGGTTGTTTAAAAAG
D,H	cg18837178	5Biosg/TATGGTGATTTGTGATTAG	CCATCTTCTCAAATTACT	AAATAAACCTACTTTCTTCC
H	cg12010173	5Biosg/GGGATTTTAGTTTATTGTA	ACCCAAATCTACCTATTC	ACCTATATACCCCTCCC
H	*EIF4E3*	TGTTTATAGGGTGTGATATT	5Biosg/ACACTAAAAACTACCATCTAAT	TGTTTTTTATGTGGGAAT
H	cg01063243	5Biosg/TAGTAGGGAGATTATAAAGATAG	ACACTTTCCTTACTCTTCTT	CCCACAACTTCCCAT

Table 24: Forward, reverse and sequencing primer for each successfully validated DMP Pyrosequencing assay. Tissue – D=DLPFC, H=Hippocampus. 5Biosg=biotin label

4.2.7 Statistical analyses

All statistical analyses were carried out in Stata as described in Section 3.2.7. When ANOVA was performed, sex and age at death were included as covariates.

4.2.8 Protein extraction and quantification

Sections of DLPFC and hippocampus were held in long-term storage at -80°C following sampling and were selected and removed from storage for the purposes of this study. Protein was extracted from 100µg DLPFC and 4x100µM sections of hippocampus in 400µl extraction buffer (50mM Tris-HCl, 5mM EGTA, 10mM EDTA, pH 7.4), as outlined in Section 2.7.4.

Total protein levels were quantified using the Thermo Scientific™ Pierce™ BCA™ Protein Assay (Thermo Scientific, USA) (Section 2.7.5). Protein samples were then standardised to 25µg and made up to 15µl using the extraction buffer described in Section 2.7.4.

4.2.9 Western blot optimisation

Four different blotting conditions (Table 8) for anti-NGF were tested to find the optimum conditions. The conditions tested were; blocking in 5% milk followed by dilution of anti-NGF in either 5% milk or 1% milk, and blocking in 5% BSA followed by dilution of anti-NGF in either 5% BSA or 1% BSA. The optimum conditions were selected as the lane which produced the strongest band and least background. The optimum conditions were then used in all future Western blots (Section 2.7.6.1).

4.2.10 Western blot

The detailed protocol for the Western blot can be found in Section 2.7.6.2, but is described in brief below.

15µl of each sample was loaded into a well on a 12% acrylamide gel (Aversham, GE Healthcare Life Sciences, UK) along with 15µl DLPFC standard, 15µl hippocampus standard and 10µl SpectraBr ladder (Fermentas, International) and run at 160V for 1 hour. Following activation of the membrane in 100% methanol, the protein samples were transferred onto the membrane in a transfer tank kept cool using ice at 100V for 1 hour. The membrane was blocked using a 5% milk solution for 1 hour. The membrane was then incubated in 1:200 anti-NGF in 5% milk for 1 hour and washed 3 times for 5 minutes each in TBST to remove any unbound antibody. The membrane was then incubated in 1:1000 anti-rabbit in 1% milk for one hour and washed three times for 5 minutes in TBST. Equal parts of A and B taken from the Amersham ECL Prime Western Blotting Detection Reagent (GE Healthcare Life Sciences, UK) were mixed and washed over the membrane. The membrane was then covered in a thin film and imaged using a Syngene G:Box camera (Syngene, UK) and Genesys software (Syngene, UK).

To check that the total volume of protein loaded into each gel was the same in each lane a loading control was used. First, the membrane was stripped of all bound antibody by washing in Restore Western Blot Stripping Buffer (ThermoScientific, USA) for 15 minutes. The membrane was then blocked with 5% milk and left at 4°C overnight. The membrane was then incubated in α-tubulin antibody for 1 hour. 3 x 5 minute washes in TBST then followed. The membrane was then incubated in the secondary antibody for 1 hour and washed 3 times for 5 minutes each in TBST. Again equal parts of A and B from the Amersham ECL Prime Western Blotting Detection Reagent (GE Healthcare Life Sciences, UK) were mixed and washed over the membrane. A plastic film was used to cover the membrane and imaged using a Syngene G:Box camera (Syngene, UK) and Genesys software (Syngene, UK). ImageJ software (National Institutes of Health, USA) was used to generate values for band density in each gel lane as described in Section 2.8.4 and a relative value for NGF was calculated.

4.3 Results

4.3.1 Discovery analysis

Of the 60 brain (30 DLPFC, 30 Hippocampus) samples that underwent HM450 analysis, 52 samples passed the quality control assessment of the inbuilt microarray controls assessed in GenomeStudio. The remaining eight samples failed due to incomplete bisulphite modification and were dropped from further analysis.

In both tissues, low failure rates were observed on both a sample and probe by probe basis, assessed using detection p values generated by the array scanner. As described in Section 2.5.2, samples were only removed if ≥20% of probes had a detection p value >0.01. Probes were removed if ≥10% of samples had a detection p value >0.01. No samples were removed from analysis using these criteria in both brain regions. 25,006 probes were removed from the DLPFC analysis and 22,416 probes were dropped from the hippocampal analysis. Summaries of the detection p values observed for samples and probes are shown in Figure 35 and Figure 36.

Following data filtering and normalisation, as described in Section 2.5.2, 431,832 CpG sites were taken forward for analysis measured in DLPFC samples from 24 individuals; 13 D vs 11 CN. 434,422 CpG sites were taken forward for analysis measured in hippocampus samples from 28 individuals; 16 D vs 12 CN.

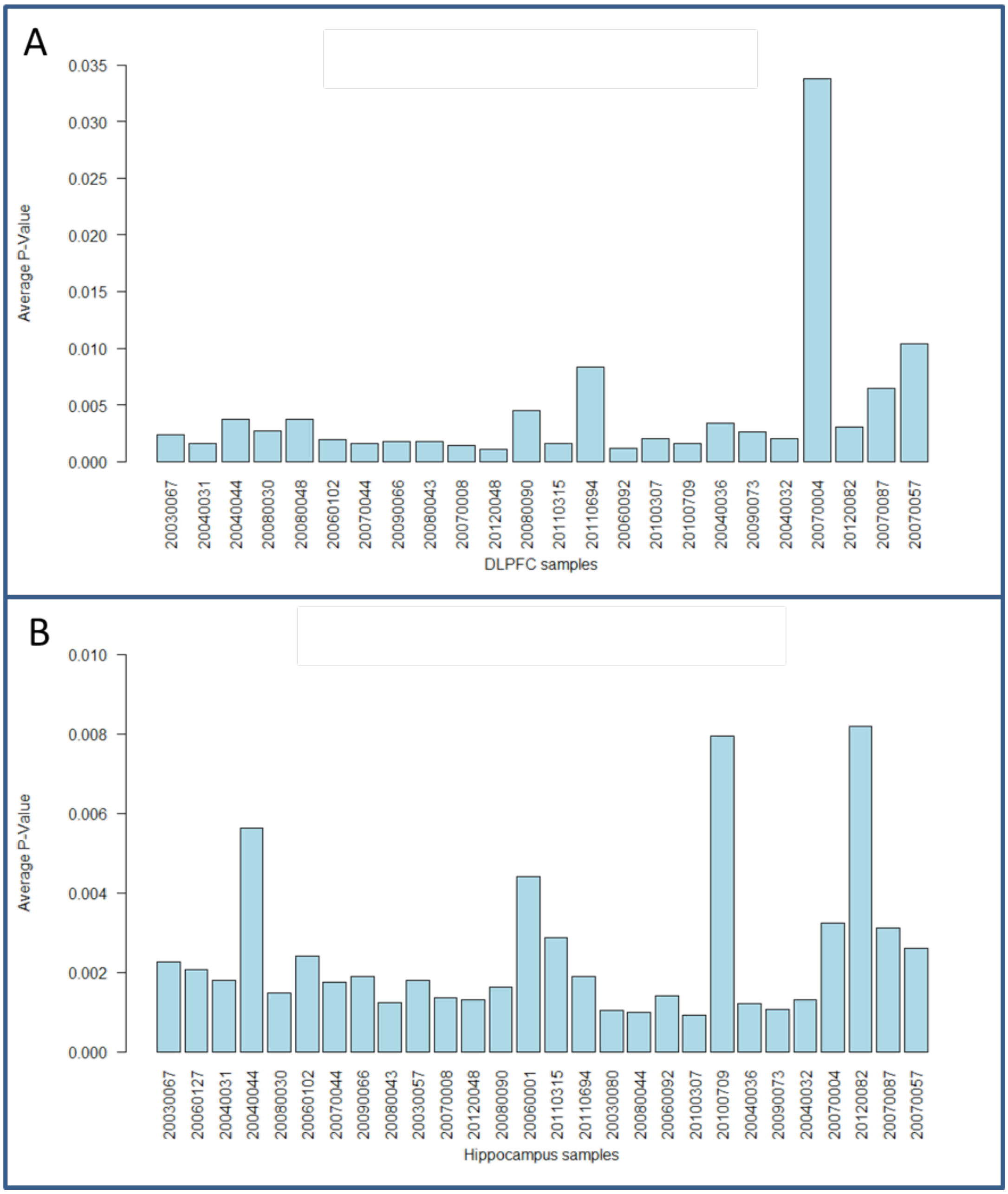


Figure 35A: Average detection p values in each DLPFC sample. For each sample the average detection p value is very low and samples are of approximately equal quality.

Figure 35B: Average detection p values in each hippocampus sample. For each sample the average detection p value is very low and samples are of approximately equal quality.

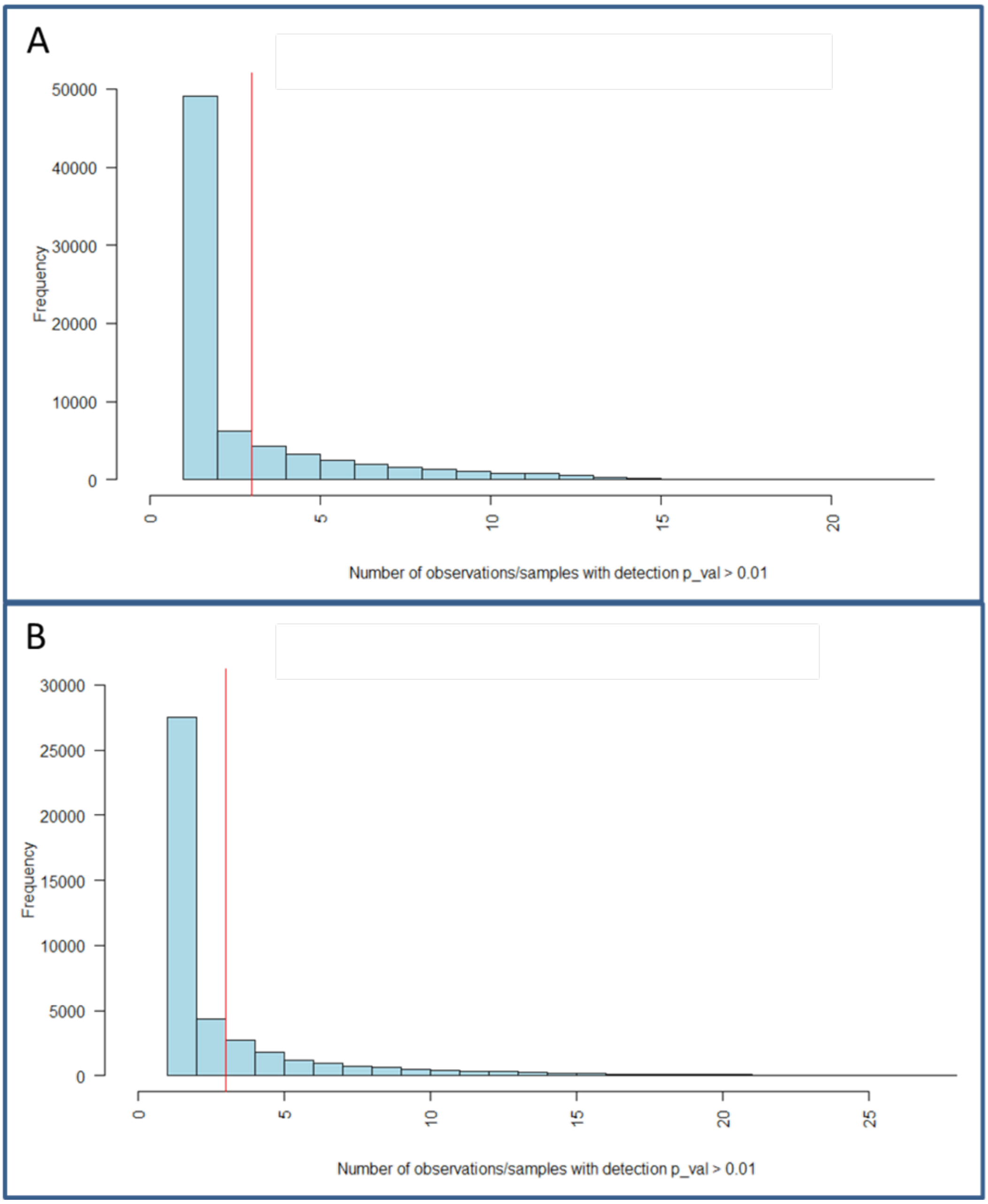


Figure 36A: Histogram to show the number of probes against the corresponding number of samples with p values greater than 0.01 across DLPFC samples. The red line indicates 10%. 411,541 probes had no missing data and for space reasons, this data is not shown on the histogram.

Figure 36B: Histogram to show the number of probes against the corresponding number of samples with p values greater than 0.01 across hippocampus samples. The red line indicates 10%. 442,633 probes had no missing data and for space reasons, this data is not shown on the histogram.

4.3.1.1 *Distribution of genome-wide methylation values*

Figure 37 (A-F) shows the distribution of methylation values using both raw (A-C) and normalised (D-F) DLPFC data. Figure 38 (A-F) shows the distribution of methylation values using both raw (A-C) and normalised (D-F) hippocampus samples. In both tissues, the raw data are much more variable, indicating that much of the variation was technical.

4.3.1.2 *Cluster analysis*

Hierarchical cluster analysis using the complete-linkage method was performed on the normalised data to look for natural groupings (either phenotypic or technical) amongst the samples based on the methylation data. No clear clusters were observed for diagnosis of dementia, indicating that DNA methylation signatures are not sufficiently distinct in either brain region drawn from this cohort (Figure 39 and Figure 42).

However, as shown in Figure 40 and Figure 43, even after normalisation and removal of probes mapping to the sex chromosomes, there remained some subtle groupings amongst samples based on sex. Due to these results, like the analysis in blood, sex was included as a covariate in subsequent analyses of single-point CpG sites. In contrast, there were no apparent clusters based on experimental chip (Figure 41 and Figure 44) suggesting that these batch effects were removed. Nonetheless, to fully account for any remaining, albeit subtle, batch effects, chip was also included as a covariate in the single-point analysis model.

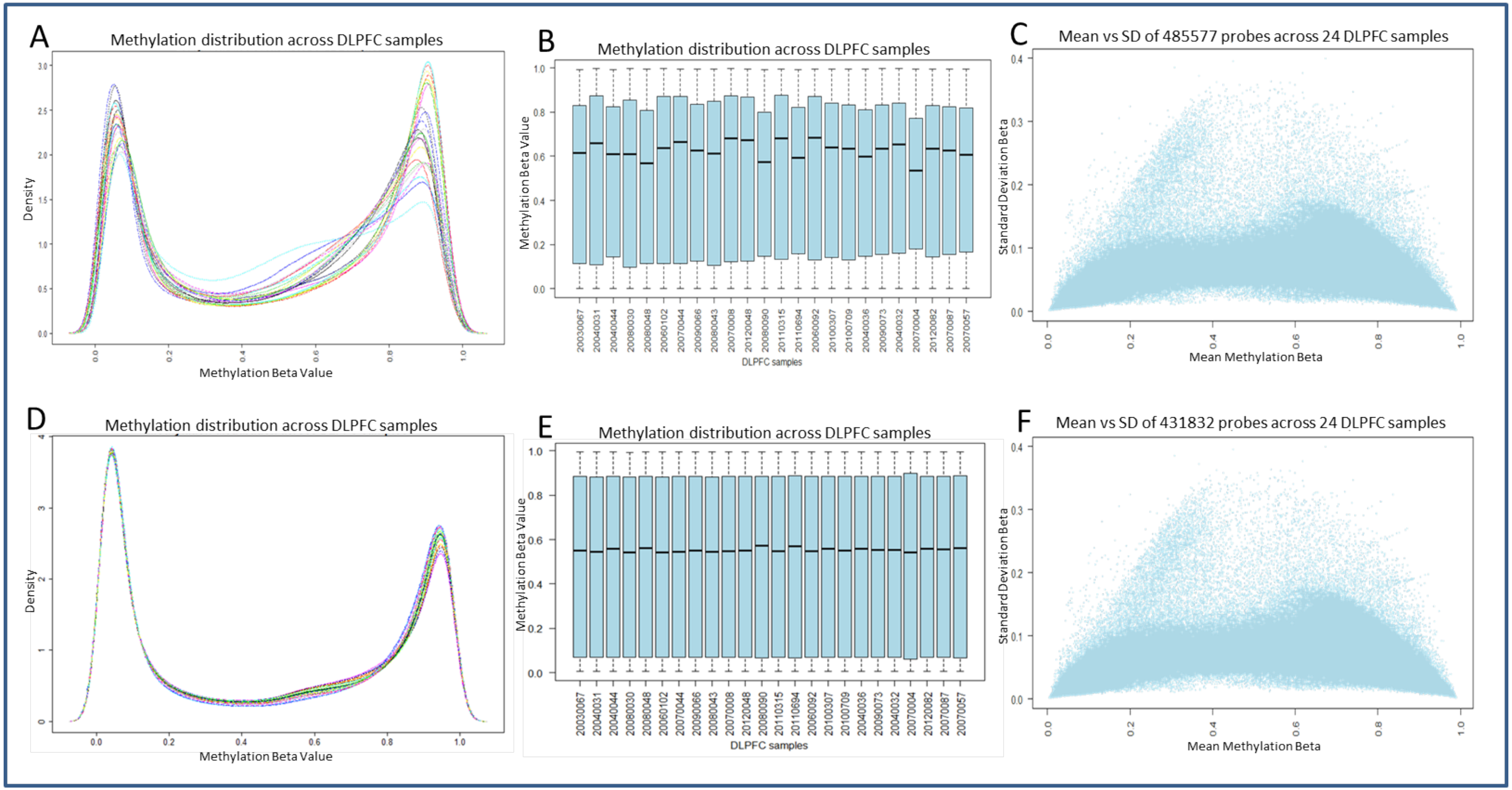


Figure 37: A composite figure showing the distributions of methylation using raw and normalised DLPFC data. A. Density plot showing methylation distribution in 24 DLPFC samples using raw data. B. Box plot showing mean methylation raw beta values across all DLPFC samples. C. Mean vs SD plot showing heteroskedasticity using raw data. D. Density plot showing methylation distribution in 24 DLPFC samples using normalised data. E. Box plot showing mean methylation normalised beta values across all DLPFC samples. F. Mean vs SD plot showing heteroskedasticity using normalised data.

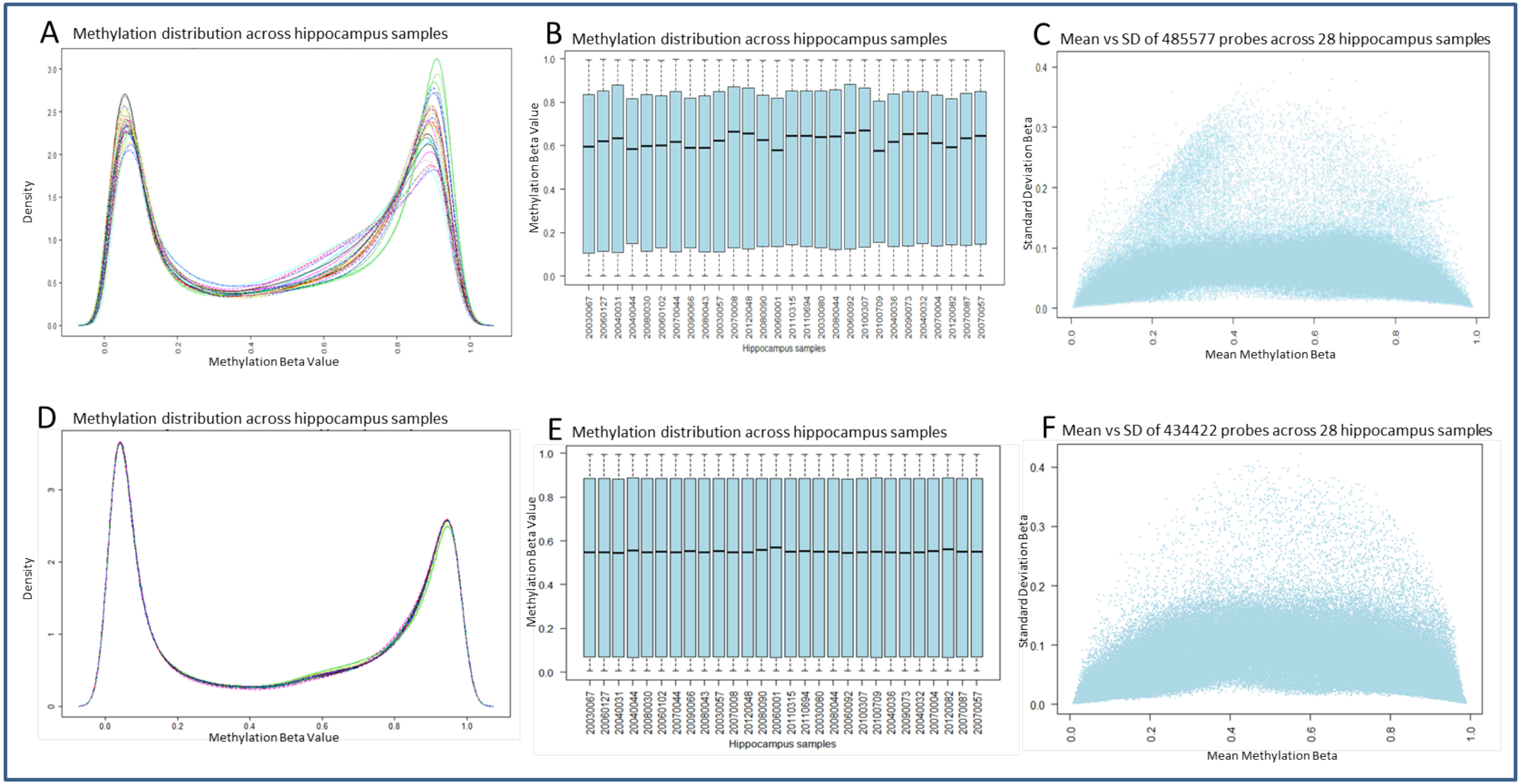


Figure 38: A composite figure showing the distribution of methylation using raw and normalised hippocampus data. A. Density plot showing methylation distribution in 28 hippocampus samples using raw data. B. Box plot showing mean methylation raw beta values across all hippocampus samples. C. Mean vs SD plot showing heteroskedasticity using raw data. D. Density plot showing methylation distribution in 28 hippocampus samples using normalised data. E. Box plot showing mean methylation normalised beta values across all hippocampus samples. F. Mean vs SD plot showing heteroskedasticity using normalised data.

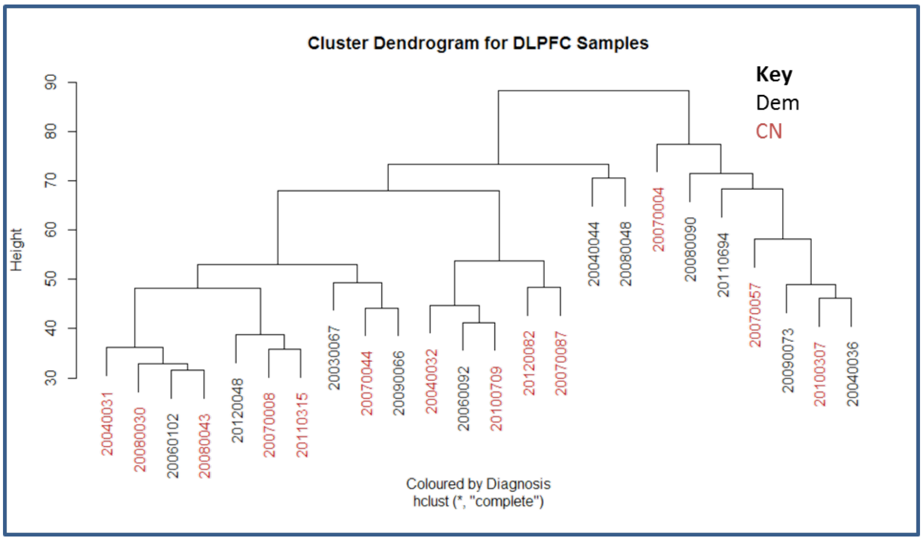


Figure 39: Cluster dendrogram for all 24 DLPFC samples coloured by diagnosis.

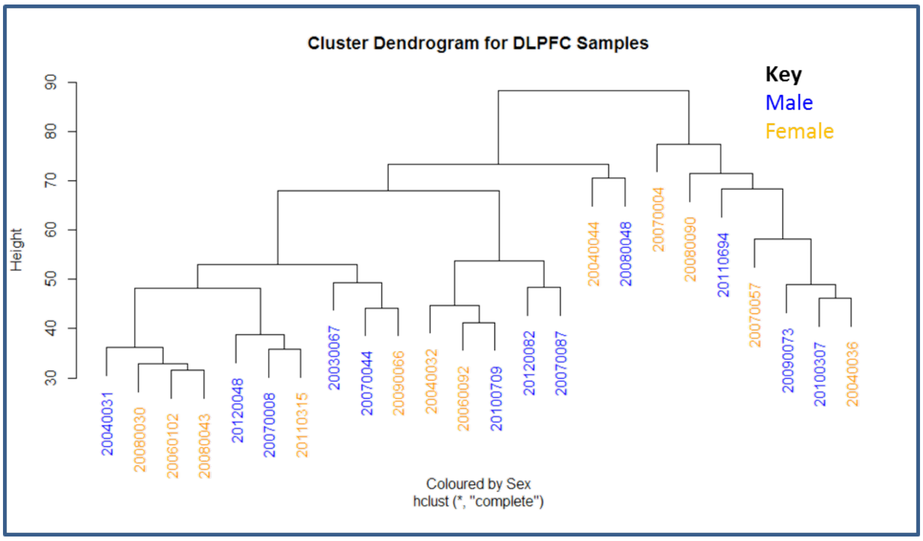


Figure 40: Cluster dendrogram for all 24 DLPFC samples coloured by sex.

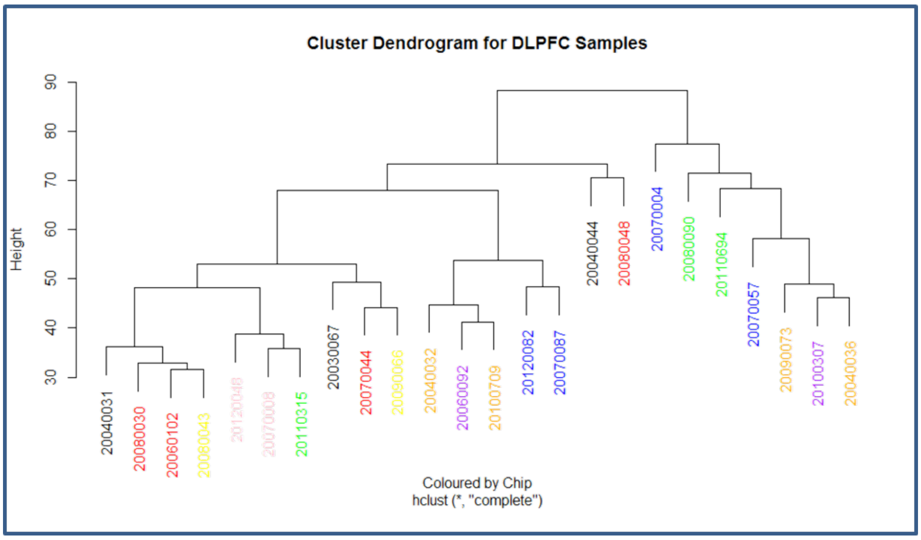


Figure 41: Cluster dendrogram for all 24 DLPFC samples coloured by chip.

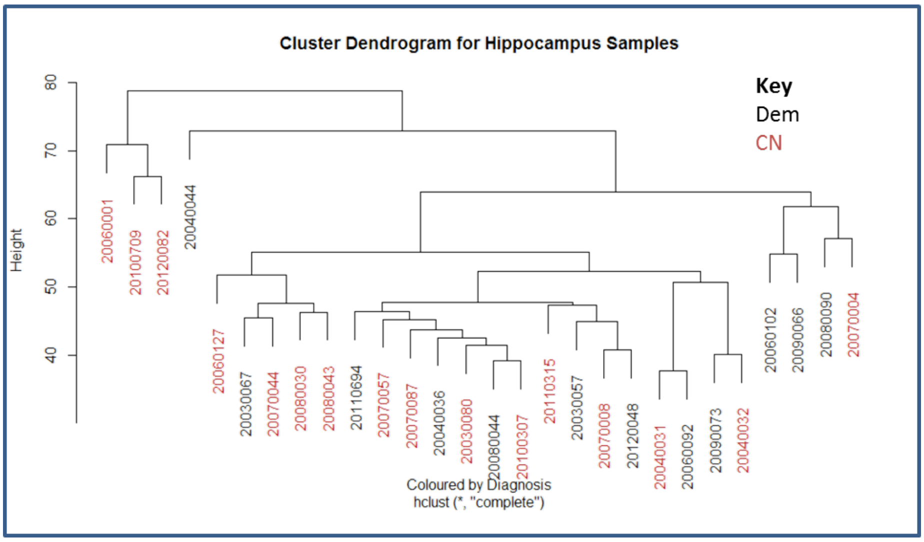


Figure 42: Cluster dendrogram for all 28 hippocampus samples coloured by diagnosis.

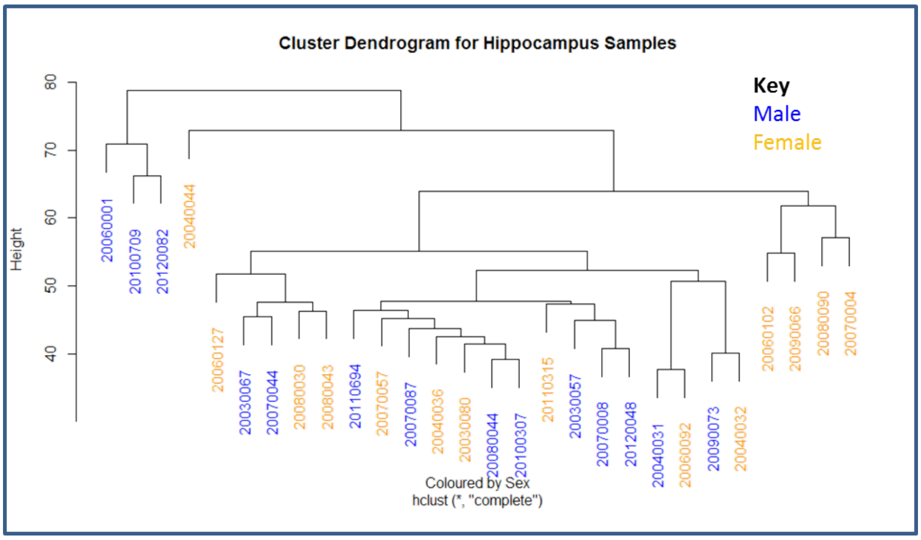


Figure 43: Cluster dendrogram for all 28 hippocampus samples coloured by sex.

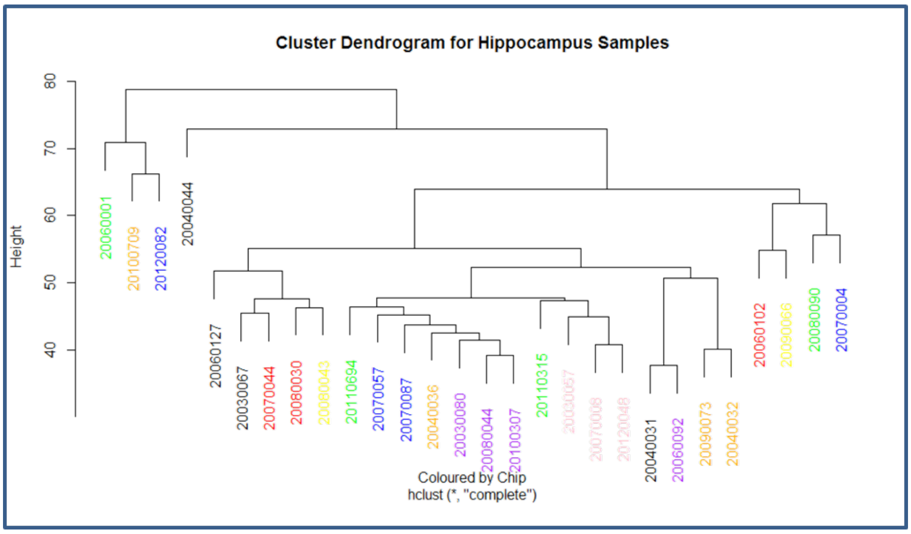


Figure 44: Cluster dendrogram for all 28 hippocampus samples coloured by chip.

4.3.1.3 *Single-point analysis*

ANOVA was performed on all beta values across all CpG sites passing QC (431,822 sites in DLPFC; 434,422 sites in hippocampus) against the following outcome measures: diagnosis (D/CN), last MMSE result before death, Braak staging and a number of CAMCOG scores (abstract thinking, attention, calculation, executive function, language comprehension, language expression, memory learning, memory recent, memory remote, memory total, orientation, perception, praxis, total CAMCOG score). Age at death, sex and chip were included in the analysis model as covariates. Results from this step were exported as a table in Microsoft Excel and ordered by ascending p value.

As for the analysis of blood samples in Chapter 3, two approaches were employed when selecting top hits. These are described below.

4.3.1.4 *Selection of top hits – approach 1*

4.3.1.4.1 Dorsolateral prefrontal cortex

The first approach selects hits which reach genome wide significance. When testing for association between methylation and diagnosis, last MMSE and all CAMCOG variables (abstract thinking, attention, calculation, executive function, language comprehension, language expression, memory learning, memory recent, memory remote, memory total, orientation, perception, praxis and total CAMCOG score), no CpG sites were highlighted as significant at the conservative genome wide corrected threshold of $p<1.16x10^{-7}$. However, when Braak staging was included in the model as the outcome variable, methylation was significantly different at ten CpG sites (DMPs) (Figure 45 and Figure 46). These ten DMPs were then considered as targets for validation by Pyrosequencing.

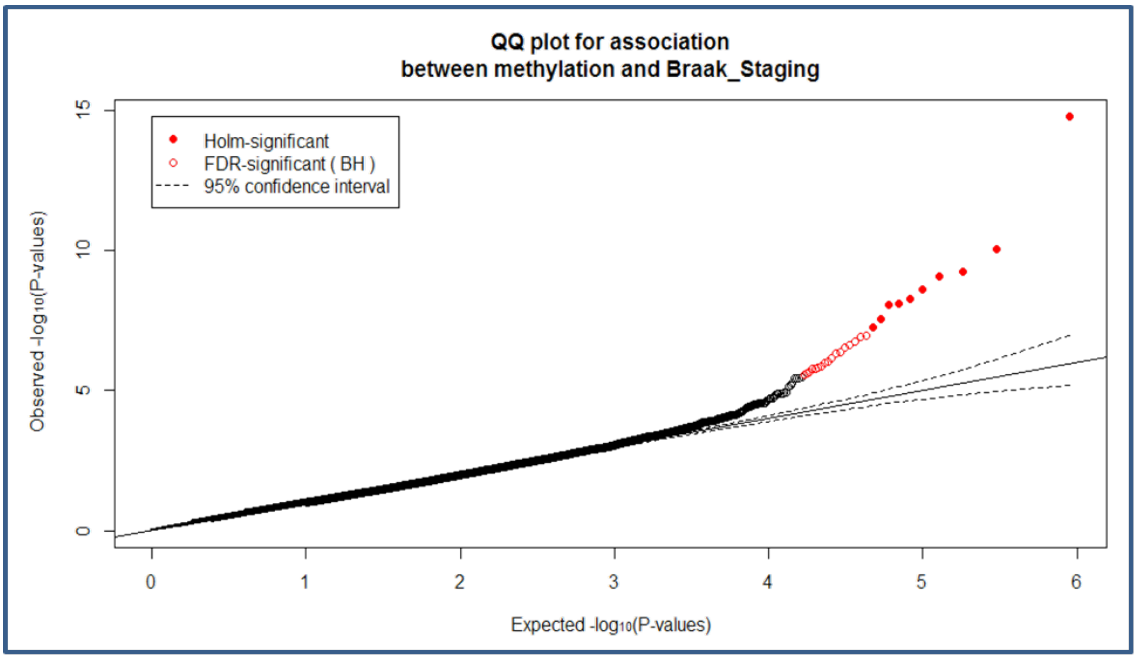


Figure 45: QQ plot for association between methylation and Braak staging in the DLPFC. The QQ plot shows the expected against the observed p values. Data points plotted in red reach the Bonferroni corrected significance threshold.

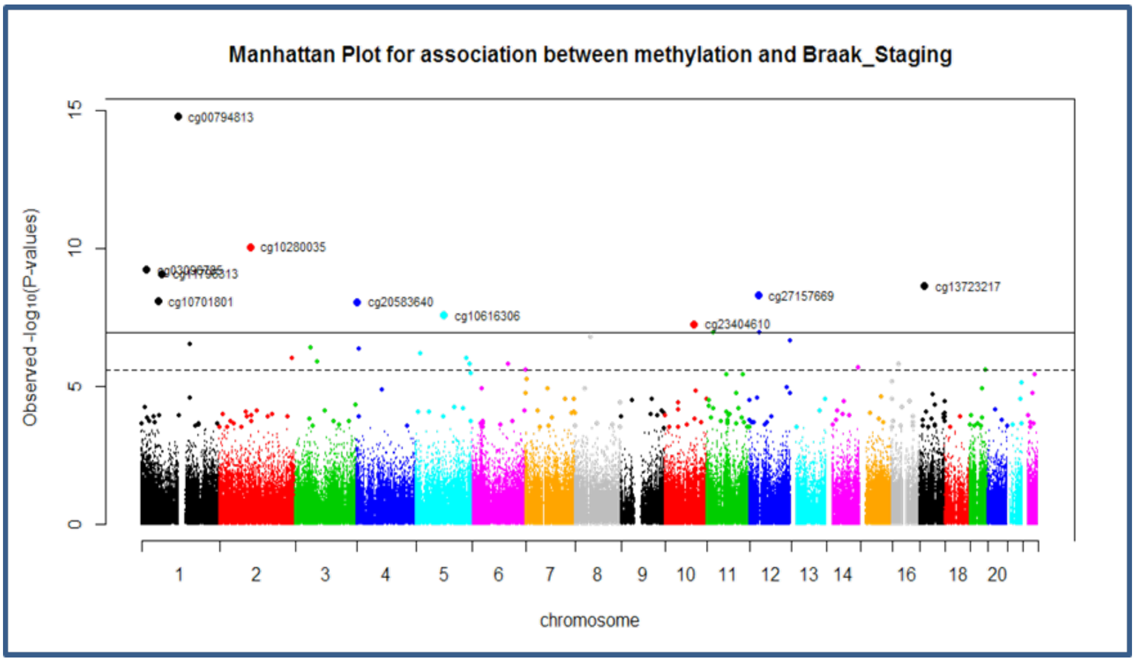


Figure 46: Manhattan plot for association between methylation and Braak staging in the DLPFC. Each circle represents an individual CpG site. The continuous line marks the $p<1.1x10^{-7}$ significance threshold. CpG sites with p values below the threshold are labelled with their Illumina ID.

Of the top ten hits, only four primer pairs were successfully designed (Table 25). Two CpG sites dropped out of the selection due to having a small (<5%) effect size. 5% was chosen as the minimum effect size due to the technical sensitivity of the Pyrosequencer. Although a number of SNPs were removed during the HM450 filtering step (Section 2.5.2), an additional SNP search was incorporated into the work flow. A BLAT search (NCBI) of the probe sequence highlighted that three of the remaining eight CpG sites were in fact SNPs and these were removed from the selection. A search on the NHGRI GWAS Catalog (www.genome.gov/gwastudies) revealed that none of these three SNPs had previously been associated with PSD or any related disease. The remaining five loci were taken forward for primer design. As described in Section 2.6.2, only those sites where primers were able to be designed with a score of 70 and no warnings on the PSQ design software were selected for the validation stage. Figure 47 shows the selection of top hits.

Variable	**CpG site**	**P value**	**Gene**	**F statistic**	**Effect size**	**SNP**	**Primers**
Braak Staging	cg00794813	1.68E-15	*NGF*	5804.12	0.652	No	Yes
Braak Staging	cg03096785	5.91E-10	*CTRC*	337.76	0.606	No	Yes
Braak Staging	cg11796313	8.40E-10		312.23	0.588	No	Yes
Braak Staging	cg20583640	8.90E-09		183.83	0.479	No	Yes
Braak Staging	cg10616306	2.73E-08	*TMEM161B*	142.76	-0.062	No	No
Braak Staging	cg23404610	5.63E-08		121.26	0.549	Yes	
Braak Staging	cg27157669	5.19E-09	*KLDHC5*	207.54	0.525	Yes	
Braak Staging	cg13723217	2.38E-09		247.25	-0.321	Yes	
Braak Staging	cg10701801	8.30E-09		510.15	0.016		
Braak Staging	cg10280035	9.02E-11		514.10	0.011		

Table 25: Details about top 10 CpG sites based on genome wide significance.

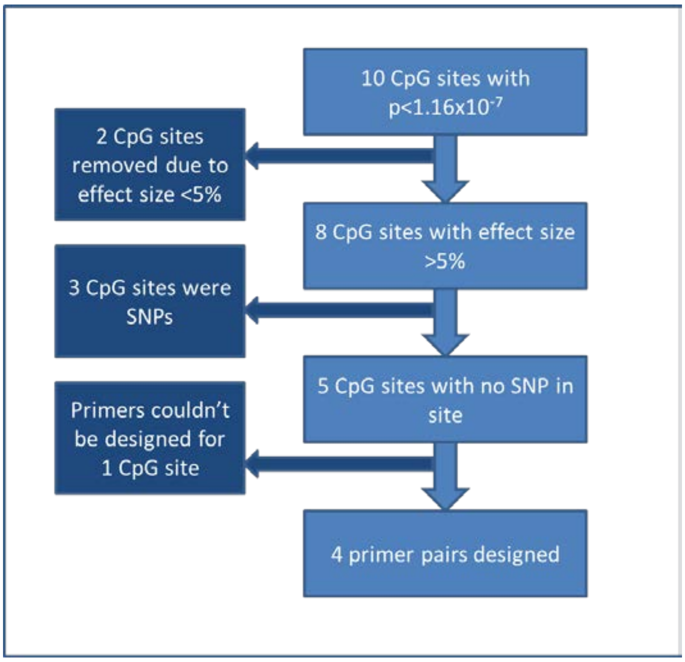

10 CpG sites with p<1.16x10-7
2 CpG sites removed due to effect size <5%
8 CpG sites with effect size >5%
3 CpG sites were SNPs
5 CpG sites with no SNP in site
Primers couldn't be designed for 1 CpG site
4 primer pairs designed

Figure 47: Flow diagram summarising the selection of top hits which reach genome wide significance in the DLPFC.

The HM450 data were then adjusted for cellular composition using the method described by Guintivano *et al.* (2013) and the CpGassoc analyses were repeated. The before and after cell adjustment analyses were compared to look for differences. At the majority of CpG sites, cell adjustment only slightly altered the results (p value and F statistic). The effect, however, was significantly altered in three of the ten selected CpG sites with all three CpG sites losing their significance. The test statistics for before and after cellular adjustment are displayed in Table 26.

Variable	**CpG site**	**Pre-adjustment**		**Post-adjustment**	
		F statistic	P value	F statistic	P value
Braak Staging	cg00794813	5804.12	1.68E-15	6657.13	3.01E-14
Braak Staging	cg10280035	514.10	9.02E-11	582.87	5.05E-10
Braak Staging	cg03096785	337.76	5.91E-10	357.65	3.53E-09
Braak Staging	cg11796313	312.23	8.40E-10	280.23	9.31E-09
Braak Staging	cg13723217	247.25	2.38E-09	229.49	2.06E-08
Braak Staging	cg27157669	207.54	5.19E-09	184.21	4.91E-08
Braak Staging	cg20583640	183.83	8.90E-09	164.70	7.64E-08
Braak Staging	cg23404610	121.26	5.63E-08	126.44	2.17E-07
Braak Staging	cg10616306	142.76	2.73E-08	125.01	2.27E-07
Braak Staging	cg10701801	510.15	8.30E-09	339.54	2.86E-07

Table 26: CpG sites showing an effect when adjusted for cellular composition.

In the DLPFC, three of the ten CpG sites identified as being associated with Braak staging were no longer significant following adjustment for cellular composition. This suggests that cellular composition may be driving the significance identified at these

CpG sites. However, there is only a slight attenuation of significance at all ten CpG sites indicating that cellular composition is not wholly responsible for the significant association between methylation and Braak staging. At the remaining seven CpG sites, methylation is still significantly associated with Braak staging. As seen in Table 25, four of these CpG sites were taken forward for validation analyses on the Pyrosequencer. Significance at all four of these CpG sites remained following cellular adjustment, indicating that the methylation differences identified at these sites are only slightly affected by cellular composition. For this reason, they can still be considered as potential methylation markers which could be involved in the mechanistic pathway.

4.3.1.4.2 Hippocampus

When testing for association between methylation and diagnosis, and methylation and last MMSE, no CpG sites were highlighted as significant at the conservative genome wide corrected threshold of $p<1.16x10^{-7}$. Likewise, a number of CAMCOG scores (abstract thinking, attention, calculation, executive function, language comprehension, language expression, memory learning, memory total, orientation, perception, praxis and total CAMCOG score) were not associated with methylation at any of the 431,832 CpG sites included in the analysis. However, when Braak staging was included in the model as the outcome variable, methylation was significantly different at 30 CpG sites (DMPs) (Figure 48 and Figure 49). Three CpG sites were found to be associated with the CAMCOG variable “memory remote” (Figure 50 and Figure 51) and another with “memory recent” (Figure 52 and Figure 53). These 34 DMPs were then considered as targets for validation by Pyrosequencing.

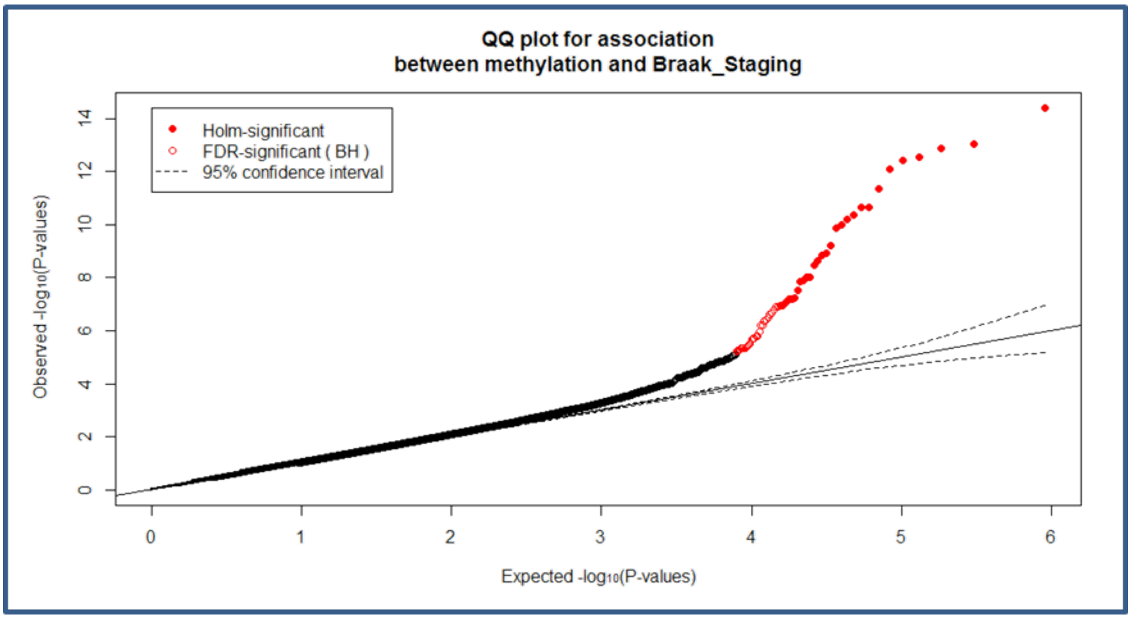


Figure 48: QQ plot for association between methylation and Braak staging in the hippocampus. The QQ plot shows the expected against the observed p values. Data points plotted in red reach the Bonferroni corrected significance threshold.

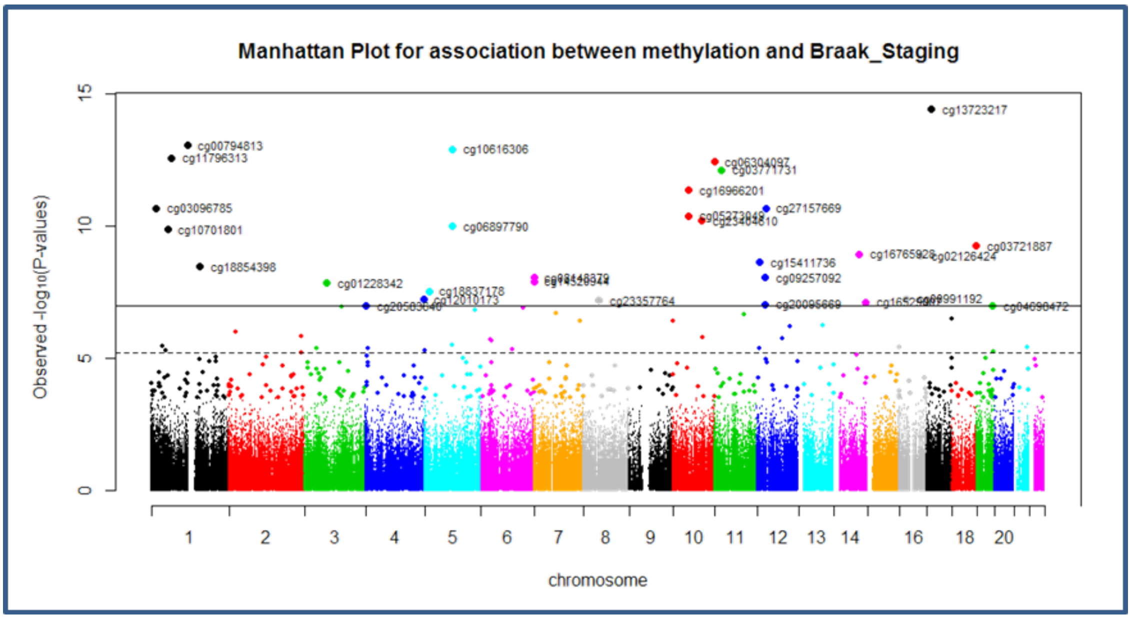


Figure 49: Manhattan plot for association between methylation and Braak staging in the hippocampus. Each circle represents an individual CpG site. The continuous line marks the $p<1.1x10^{-7}$ significance threshold. CpG sites with p values below the threshold are labelled with their Illumina ID.

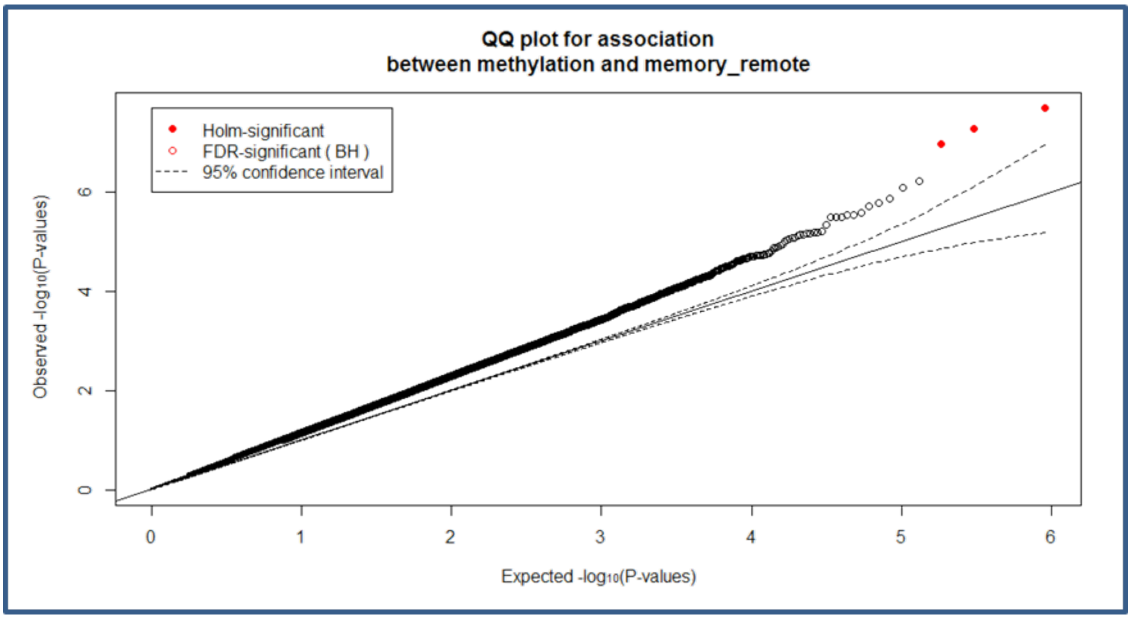


Figure 50: QQ plot for association between methylation and memory remote in the hippocampus. The QQ plot shows the expected against the observed p values. Data points plotted in red reach the Bonferroni corrected significance threshold.

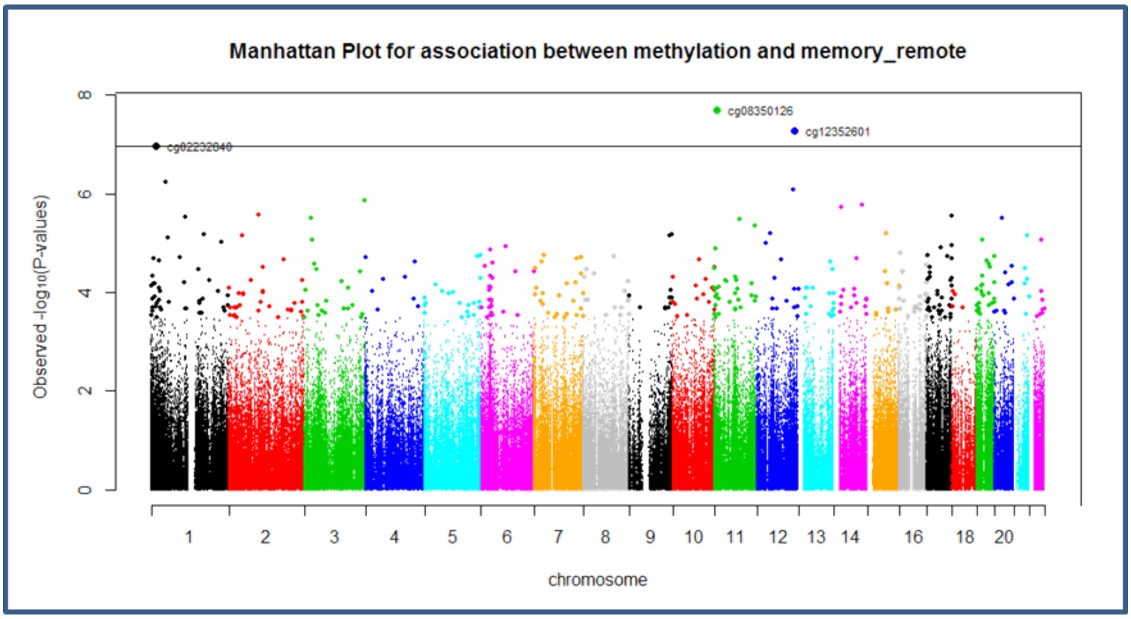


Figure 51: Manhattan plot for association between methylation and memory remote in the hippocampus. Each circle represents an individual CpG site. The continuous line marks the $p<1.1x10^{-7}$ significance threshold. CpG sites with p values below the threshold are labelled with their Illumina ID.

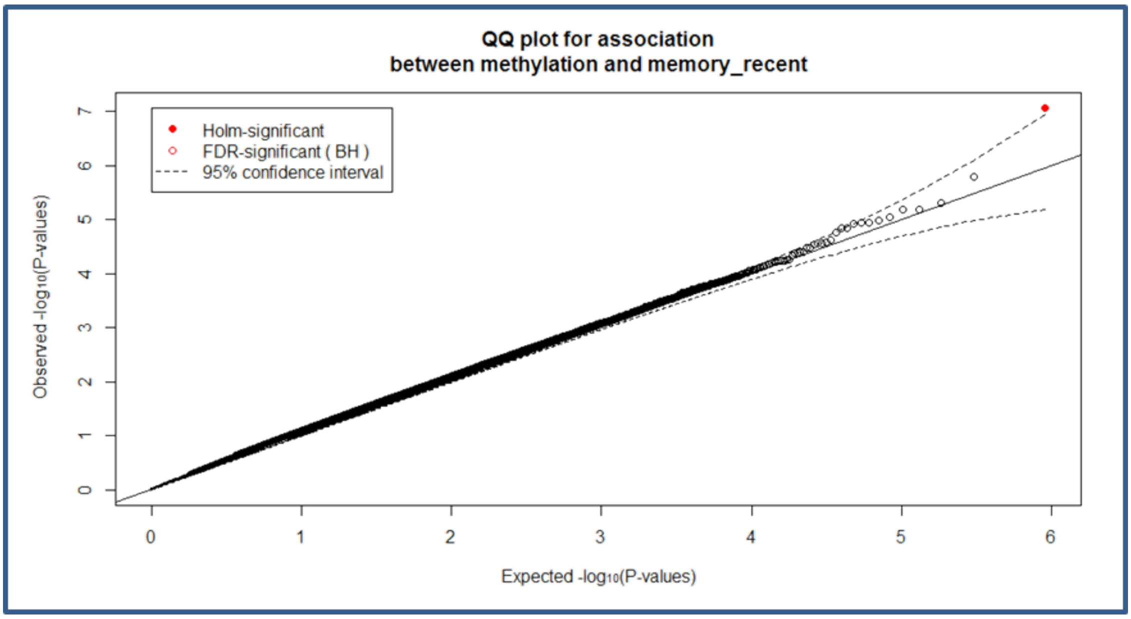


Figure 52: QQ plot for association between methylation and memory recent in the hippocampus. The QQ plot shows the expected against the observed p values. Data points plotted in red reach the Bonferroni corrected significance threshold.

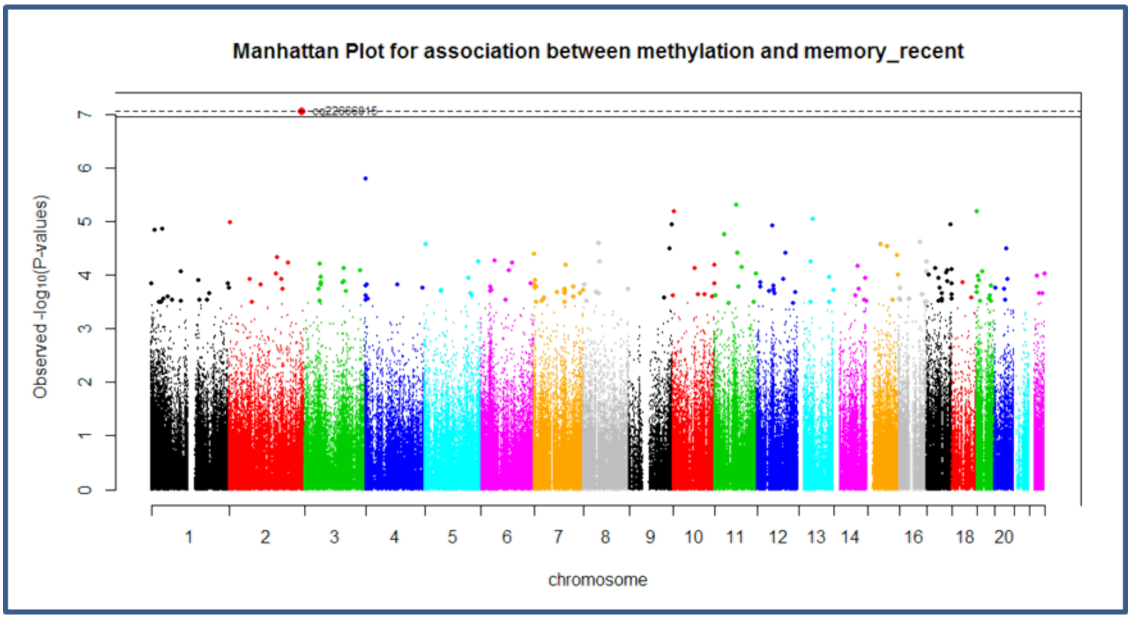


Figure 53: Manhattan plot for association between methylation and memory recent in the hippocampus. Each circle represents an individual CpG site. The continuous line marks the p<1.1x10^{-7} significance threshold. CpG sites with p values below the threshold are labelled with their Illumina ID.

Of the 34 top hits, only nine primer pairs were successfully designed (Table 27). All CpG sites had an effect size >5%. Although a number of SNPs were removed during the HM450 filtering step (Section 2.5.2), an additional SNP search was incorporated into the work flow. A BLAT search (NCBI) of the probe sequence highlighted that 13 of the 34 CpG sites were in fact SNPs and these were removed from the selection. A search on the NHGRI GWAS Catalog (www.genome.gov/gwastudies) revealed that none of these 13 SNPs had previously been associated with PSD or any related disease. The remaining 21 loci were taken forward for primer design. As described in Section 2.6.2, only those sites where primers were able to be designed with a score of 70 and no warnings on the PSQ design software were selected for the validation stage. Figure 54 shows the selection of top hits.

Variable	**CpG site**	**P value**	**Gene**	**Test statistic**	**Effect size**	**SNP**	**Primers**
Braak Staging	cg16966201	4.44E-12		219.57	-0.685	No	No
Braak Staging	cg03771731	7.75E-13	*SLC6A5*	288.12	0.650	No	No
Braak Staging	cg02126424	1.40E-09	*CTRL*	88.76	0.597	No	Yes
Braak Staging	cg11796313	2.76E-13		338.25	0.541	No	Yes
Braak Staging	cg03096785	2.19E-11	*CTRC*	171.10	0.484	No	Yes
Braak Staging	cg18837178	3.06E-08		54.03	0.478	No	Yes
Braak Staging	cg00794813	9.33E-14	*NGF*	400.29	0.464	No	Yes
Braak Staging	cg20583640	1.09E-07		43.91	0.432	No	Yes
Braak Staging	cg16765928	1.17E-09	*NRXN3*	91.33	0.419	No	No
Braak Staging	cg12010173	6.23E-08		48.11	0.293	No	Yes
Braak Staging	cg09257092	9.27E-09	*ITPR2*	65.57	-0.264	No	No
Braak Staging	cg08148379	9.22E-09		65.63	0.235	No	No
Braak Staging	cg06304097	3.71E-13	*TCERG1L*	323.11	-0.177	No	No
Braak Staging	cg14520944	1.28E-08		62.20	0.170	No	No
Memory remote	cg08350126	2.06E-08	*HPX*	9.81	-0.139	No	No
Braak Staging	cg01228342	1.44E-08	*EIF4E3*	61.07	-0.136	No	Yes
Memory remote	cg02232840	1.09E-07	*KIAA1026*	8.73	-0.124	No	Yes
Braak Staging	cg06897790	1.06E-10	*TMEM161B*	133.54	-0.121	No	No
Braak Staging	cg10616306	1.27E-13	*TMEM161B*	381.49	-0.094	No	No
Memory recent	cg22666015	8.60E-08	*INPP5D*	-8.88	0.091	No	No
Memory remote	cg12352601	5.29E-08	*KDM2B*	9.19	-0.077	No	No
Braak Staging	cg16529007	7.94E-08		46.23	0.768	Yes	
Braak Staging	cg23357764	6.65E-08		56.07	0.676	Yes	
Braak Staging	cg20095669	9.30E-08	*PPFIBP1*	52.86	0.675	Yes	
Braak Staging	cg03721887	5.97E-10	*CTDP1*	164.41	0.648	Yes	
Braak Staging	cg04698472	1.09E-07	*SIGLEC12*	43.85	0.608	Yes	
Braak Staging	cg15411736	2.43E-09	*CLECL1*	170.46	0.602	Yes	
Braak Staging	cg23404610	6.40E-11		144.56	0.567	Yes	
Braak Staging	cg05273049	4.29E-11	*ARHGAP22*	153.95	-0.540	Yes	
Braak Staging	cg27157669	2.21E-11	*KLHDC5*	170.87	0.489	Yes	
Braak Staging	cg00991192	5.81E-08		48.66	0.472	Yes	
Braak Staging	cg13723217	4.03E-15		650.96	-0.285	Yes	
Braak Staging	cg18854398	3.47E-09	*C1orf66*	76.77	-0.138	Yes	
Braak Staging	cg10701801	1.36E-10	*OSBPL9*	216.00	0.053	Yes	

Table 27: Details about top 34 CpG sites based on genome wide significance in the hippocampus.

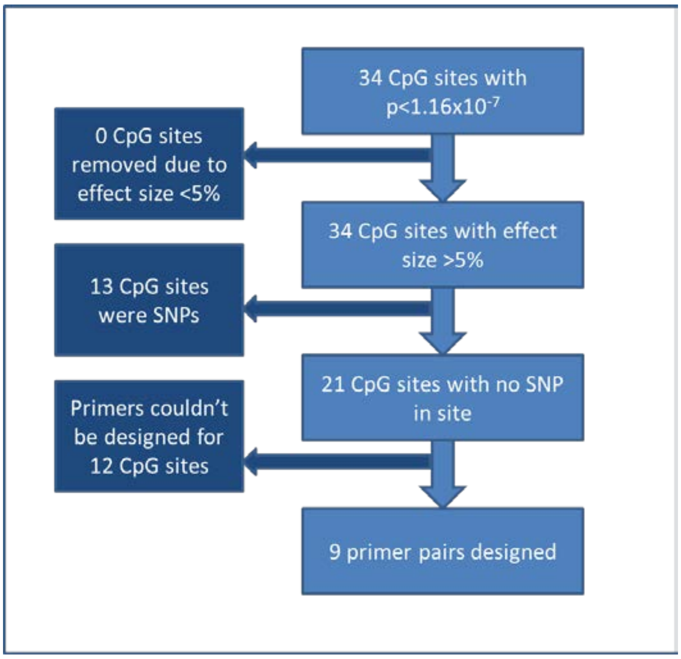

34 CpG sites with p<1.16x10-7
0 CpG sites removed due to effect size <5%
34 CpG sites with effect size >5%
13 CpG sites were SNPs
21 CpG sites with no SNP in site
Primers couldn't be designed for 12 CpG sites
9 primer pairs designed

Figure 54: Flow diagram summarising the selection of top hits which reach genome wide significance in the hippocampus.

The HM450 data were then adjusted for cellular composition using the method described by Guintivano *et al.* (2013) and the CpGassoc analyses were repeated. The before and after cell adjustment analyses were compared to look for differences. At the majority of CpG sites, cell adjustment only slightly altered the results (p value and F/t statistic). The effect, however, was significantly altered in six of the 34 selected CpG sites with all six CpG sites losing their significance. The before and after cellular adjustment test statistics are displayed in Table 28.

Variable	CpG site	Pre-adjustment		Post-adjustment	
		F/t statistic	P value	F/t statistic	P value
Braak Staging	cg13723217	650.96	4.03E-15	699.66	2.33E-14
Braak Staging	cg00794813	400.29	9.33E-14	415.91	5.19E-13
Braak Staging	cg10616306	381.49	1.27E-13	364.04	1.15E-12
Braak Staging	cg11796313	338.25	2.76E-13	304.44	3.32E-12
Braak Staging	cg06304097	323.11	3.71E-13	289.30	4.50E-12
Braak Staging	cg03771731	288.12	7.75E-13	267.98	7.09E-12
Braak Staging	cg16966201	219.57	4.44E-12	231.38	1.69E-11
Braak Staging	cg27157669	170.87	2.21E-11	154.36	1.85E-10
Braak Staging	cg03096785	171.10	2.19E-11	153.65	1.90E-10
Braak Staging	cg23404610	144.56	6.40E-11	146.41	2.53E-10
Braak Staging	cg05273049	153.95	4.29E-11	141.52	3.09E-10
Braak Staging	cg06897790	133.54	1.06E-10	133.16	4.41E-10
Braak Staging	cg10701801	216.00	1.36E-10	197.67	1.17E-09
Braak Staging	cg03721887	164.41	5.97E-10	153.35	4.08E-09
Braak Staging	cg18854398	76.77	3.47E-09	89.16	4.60E-09
Braak Staging	cg08148379	65.63	9.22E-09	88.36	4.85E-09
Braak Staging	cg16765928	91.33	1.17E-09	83.39	6.78E-09
Braak Staging	cg02126424	88.76	1.40E-09	80.05	8.59E-09
Braak Staging	cg15411736	170.46	2.43E-09	148.27	2.31E-08
Memory Remote	cg02232840	8.73	1.09-07	9.95	2.93E-08
Braak Staging	cg23357764	56.07	6.65E-08	75.08	4.01E-08
Braak Staging	cg09257092	65.57	9.27E-09	59.20	4.87E-08
Memory Remote	cg12352601	9.19	5.29E-08	9.58	4.96E-08
Memory Remote	cg08350126	9.81	2.06E-08	9.53	5.37E-08
Braak Staging	cg14520944	62.20	1.28E-08	56.11	6.62E-08
Braak Staging	cg18837178	54.03	3.06E-08	55.52	7.03E-08
Braak Staging	cg01228342	61.07	1.44E-08	54.92	7.48E-08
Braak Staging	cg20583640	43.91	1.09E-07	54.51	7.81E-08
Braak Staging	cg04698472	43.85	1.09E-07	48.73	1.47E-07
Braak Staging	cg20095669	52.86	9.30E-08	57.26	1.67E-07
Memory Recent	cg22666015	-8.88	8.60E-08	-8.66	1.95E-07
Braak Staging	cg00991192	48.66	5.81E-08	44.12	2.58E-07
Braak Staging	cg12010173	48.11	6.23E-08	43.35	2.85E-07
Braak Staging	cg16529007	46.23	7.94E-08	42.33	3.26E-07

Table 28: CpG sites showing an effect when adjusted for cellular composition.

In the hippocampus, six of the 34 CpG sites identified as being associated with an outcome variable (Braak staging: n=5; Memory recent: n=1) were no longer significant following adjustment for cellular composition. This suggests that cellular composition may be driving the significance identified at these CpG sites. However, there is only a slight attenuation of significance at all 34 CpG sites indicating that cellular composition is not wholly responsible for the significant association between methylation and outcome variable. At the remaining 28 CpG sites, methylation is still significantly associated with outcome following adjustment for cellular composition. As seen in Table 27, nine of these CpG sites were taken forward for validation analyses on the Pyrosequencer. Significance at eight of the nine CpG sites remained following cellular adjustment, indicating that the methylation differences identified at these sites are only slightly affected by cellular composition. For this reason, they can still be considered as potential methylation markers which could be involved in the mechanistic pathway. Following cellular adjustment, the significant association was lost between methylation at cg12010173 and Braak staging, however, since the association was only slightly attenuated and only narrowly missed the threshold for genome-wide significance (p=2.85E-07), methylation at cg12010173 was still considered as a potential methylation marker.

4.3.1.5 ***Summary of approach 1***

Single-point analysis highlighted ten hits reaching genome-wide significance in the DLPFC samples and 34 in the hippocampus. Eight hits were significant in both tissues giving a total of 36 different significant hits across the brain regions. Correlation analyses were performed on the top-ranked list of probes identified in both brain regions to assess the correlation between effect sizes in each tissue. The effect sizes were strongly correlated between tissues in both the ten significant DLPFC hits (Rho=0.794, p=0.006) and 34 significant hippocampal hits (Rho=0.862, p=5.69E-11).

The pie charts in Figure 55 (A-D) shows the distribution of the most significant 36 hits across the different probe chemistries and genomic region (chromosome, position relative to nearby CpG island and regions within gene).

A chi squared goodness of fit test was performed to test whether the distributions of hits observed were as expected. A significantly lower number of significant CpG sites were measured using probe type I chemistry than expected (14% vs 27%, p=0.003) (Figure 55A). 14% of the top CpG sites were located within CpG islands (Figure 55B). Differences in the position of the CpG site in relation to CpG island were also not as expected with fewer significant CpG sites being observed in the CpG islands ($p=2.79x10^{-8}$). 31% of all significant CpG sites were located within gene promoters (Figure 55C); this indicates that some of these CpG sites may have functional relevance. No differences in functional genomic location (eg. Promoter, gene body) were identified between the number of CpG sites observed and expected in each region (p=0.109). The number of CpG sites identified on each chromosome is not as one would expect by chance. In particular, chromosomes 1, 10 and 12 had many more significant CpG sites than was expected (p=1.50x10^{-7}) (Figure 55D).

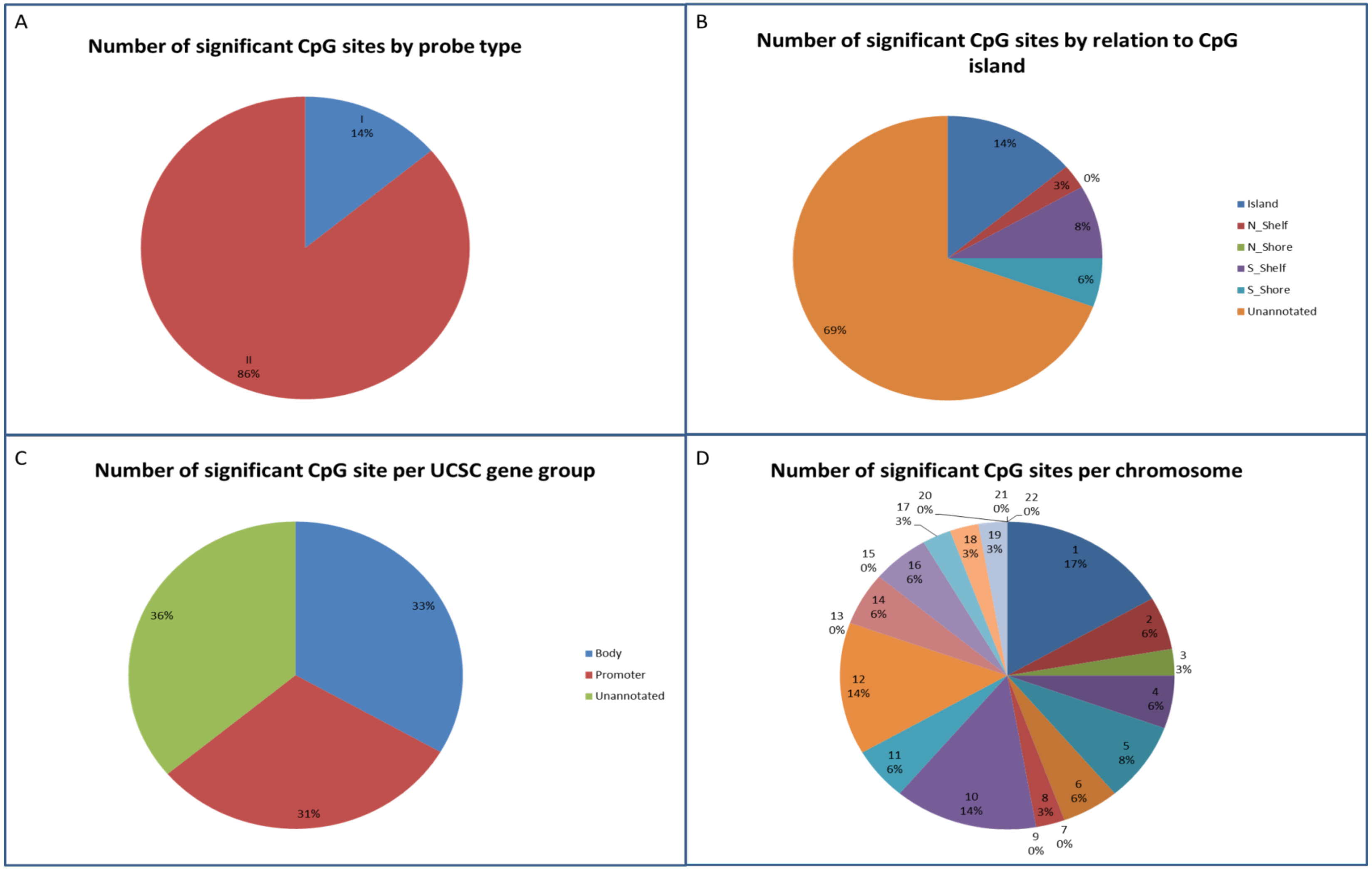


Figure 55: Description of statistically significant CpG sites identified in one or more brain region. A. Distribution of CpG sites measured using different probe types. B. position relative to nearby CpG island. C. Functional genomic location. D. Proportion of statistically significant CpGs located on each chromosome.

4.3.1.6 *Selection of top hits – approach 2*

4.3.1.6.1 Dorsolateral prefrontal cortex

The second approach utilised for selecting top hits from the HM450 analysis using ANOVA involved lowering the significance threshold to $p<1x10^{-5}$ and comparing with publicly available expression data. The expression data set used is described in Section 4.2.5.

Table 29 shows the number of hits that were considered for Pyrosequencing follow-up using the second selection approach. Lowering the threshold from $p<10^{-7}$ to $p<10^{-5}$ greatly increased the number of top hits in the selection pool, the number of significant hits increased from 10 to 128. Of these 128 hits, 114 had expression data available. Four of these were differentially expressed in AD. These four hits were then considered in terms of the direction of effect, i.e. whether the CpG sites were hyper or hypo methylated in PSD and whether they were up or down regulated in AD. Two CpG sites were found to have the expected methylation-expression relationship (hypermethylated and downregulated in AD, hypomethylated and upregulated in AD).

Variable	Hits significant to $<1x10^{-5}$ (n)	Hits represented on the expression array (n)	Hits differentially expressed (n)	Hits with the expected methylation-expression relationship (n)
Diagnosis	1	1		
Braak staging	65	55	2	1
MMSE	4	3		
Orientation	2	2		
Language comprehension	1	1		
Language expression	4	4	1	1
Memory remote	14	13		
Memory recent	5	3		
Memory learning	10	10		
Memory total	4	4		
Attention	6	6	1	
Praxis	3	3		
Abstract thinking	2	2		
Perception	2	2		
Executive function	2	2		
Total CAMCOG score	3	3		
Total number of hits	**128**	**114**	**4**	**2**

Table 29: Number of CpG sites taken forward at each stage in the selection of DLPFC top hits.

Figure 56 shows the relationship between methylation and expression in each of the significant CpG sites that had both methylation and expression data available. Sites (n=2) within the shaded regions showed the expected methylation-expression relationship. Two loci were identified as being hypermethylated in PSD and down regulated in AD. These two sites were taken forwards for further investigation (Table 30).

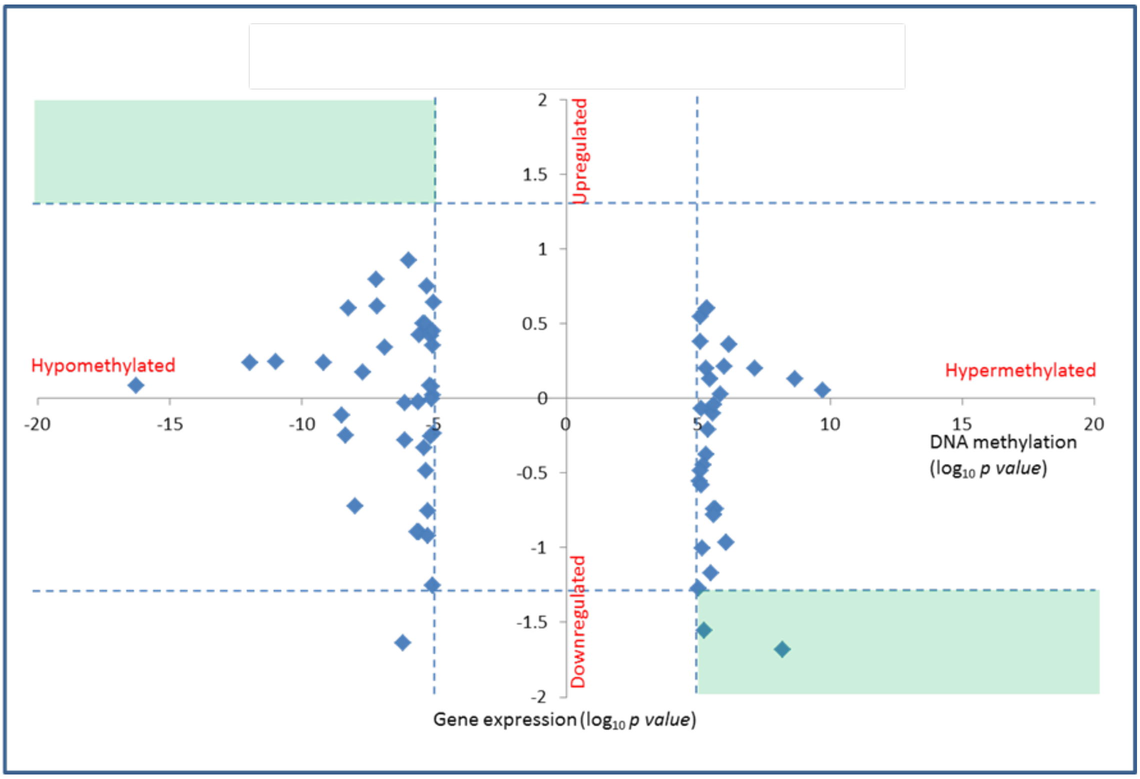


Figure 56: Quadrant plot of DMPs and associated expression levels in the DLPFC. DNA methylation (x axis) and gene expression (y axis) are graphically presented as log10 p values. Significance is defined as p< 1x10^{-5} for methylation and p<1.13 for expression (as indicated by the vertical and horizontal dashed lines). The direction of effect is indicated by the red text. The shaded areas represent sites with the expected inverse relationship between methylation and expression. Sites within these shaded areas were taken forward for further analysis.

Variable	CpG site Gene	Gene	Meth p value	Meth effect size	Exp p value	SNP	Primers
Braak staging	cg10616306	*TMEM161B*	6.18E-09	-0.062	0.021	No	No
Language expression	cg06758670	*GRIN2A*	5.98E-06	0.012	0.028		

Table 30: Description of the 2 CpG sites showing differential expression in AD DLPFC samples. Meth=methylation, Exp = expression.

Only one of the two CpG sites had an effect size >5%. However, primers could not be designed for this CpG site so neither locus was taken forward for Pyrosequencing analysis.

4.3.1.6.2 Hippocampus

Table 31 shows the number of hits that were considered for Pyrosequencing validation using the second selection approach. Lowering the threshold from p<10^{-7} to p<10^{-5} greatly increased the number of top hits in the selection pool. The number of significant hits increased from 36 to 170. Of these 170 hits, 148 had expression data available.

Forty of these were differentially expressed in AD. The 40 remaining hits were then considered in terms of the direction of effect, i.e. whether the CpG sites were hyper or hypo methylated in PSD and whether they were up or down regulated in AD. 21 CpG sites were found to have the expected methylation-expression relationship (hypermethylated and downregulated in AD, hypomethylated and upregulated in AD).

Variable	Hits significant to <1x10^{-5} (n)	Hits represented on the expression array (n)	Hits differentially expressed (n)	Hits with the expected methylation-expression relationship (n)
Diagnosis	4	4	1	1
Braak staging	100	91	26	11
MMSE	4	2		
Orientation	2	2		
Language comprehension	6	5	3	3
Language expression	5	4	1	
Memory remote	12	10	3	1
Memory recent	2	1		
Memory learning	9	8		
Memory total	11	9	3	2
Attention	4	4	1	1
Praxis	1	1	1	1
Abstract thinking	1	1		
Calculation	1	1	1	1
Executive function	1	0		
Total CAMCOG score	7	5		
Total number of hits	**170**	**148**	**40**	**21**

Table 31: Number of CpG sites taken forward at each stage in the selection of hippocampus top hits.

Figure 57 shows the relationship between methylation and expression in each of the significant CpG sites with both methylation and expression data. Sites (n=21) within the shaded regions showed the expected methylation-expression relationship. Twelve loci were identified as being hypomethylated in PSD and upregulated in AD. Nine loci were identified as having the reverse relationship, with hypermethylation of a site within the gene in PSD and down regulation of the gene in AD. These 21 sites were taken forwards for further investigation (Table 32).

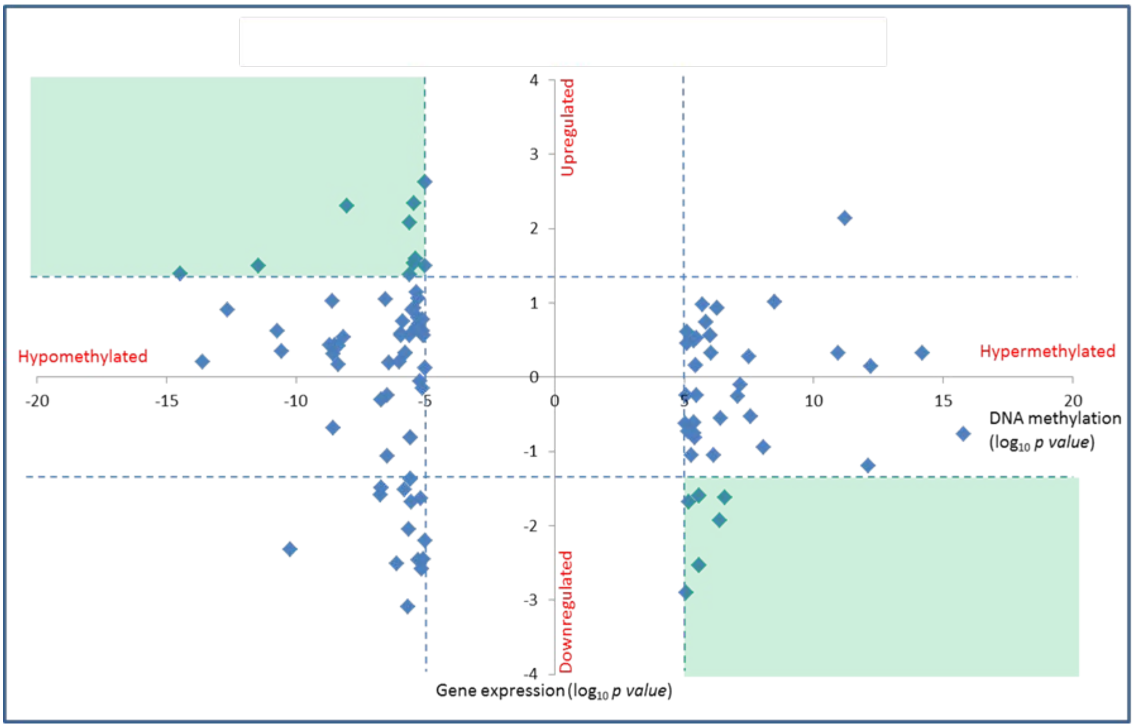


Figure 57: Quadrant plot of DMPs and associated expression levels in the hippocampus. DNA methylation (x axis) and gene expression (y axis) are graphically presented as log10 p values. Significance is defined as p< 1x10^{-5} for methylation and p<1.13 for expression (as indicated by the vertical and horizontal dashed lines). The direction of effect is indicated by the red text. The shaded areas represent sites with the expected inverse relationship between methylation and expression. Sites within these shaded areas were taken forward for further analysis.

Variable	CpG site	Gene	Meth p value	Meth effect size	Exp p value	SNP	Primers
Braak staging	cg00794813	*NGF*	3.51E-15	0.464	0.041	No	Yes
Braak staging	cg20583640	*FAM53A*	9.06E-09	0.432	0.005	No	Yes
Braak staging	cg15447017	*LMNA*	9.84E-06	0.334	0.032	No	Yes
Memory remote	cg06872721	*HOMER3*	3.71E-06	-0.290	0.029	No	No
Braak staging	cg01063243	*MYCBP2*	2.73E-07	-0.284	0.024	No	Yes
Braak staging	cg06872721	*HOMER3*	3.71E-06	0.164	0.029	No	No
Memory total	cg13223682	*OLA1*	2.70E-06	0.142	0.003	No	Yes
Memory total	cg18109033	*NIPAL2*	2.75E-06	0.090	0.025	No	No
Calculation	cg10071848	*ABCD4*	2.48E-06	-0.080	0.042	No	No
Braak staging	cg13223682	*OLA1*	2.70E-06	-0.068	0.003	No	Yes
Diagnosis	cg15549700	*AJAP1*	6.44E-06	0.062	0.021	No	No
Braak staging	cg10633931	*HMGA1*	4.27E-07	-0.061	0.012	No	No
Braak staging	cg23404610	*BTAF1*	3.70E-12	0.567	0.032	Yes	
Braak staging	cg02061626	*APPL2*	2.44E-06	0.239	0.008	Yes	
Language comprehension	cg20125971	*PDLC2*	9.92E-06	-0.174	0.002	Yes	
Language comprehension	cg22303739	*COX4I2*	3.51E-06	-0.038	0.005		
Braak staging	cg18109033	*NIPAL2*	2.75E-06	-0.037	0.025		
Language comprehension	cg01206910	*NARF*	8.36E-06	0.020	0.001		
Praxis	cg02560638	*DYNLRB2*	4.34E-06	-0.014	0.025		
Attention	cg12782834	*PIGX*	8.65E-06	0.011	0.004		
Braak staging	cg20125971	*PDCL2*	9.92E-06	-0.004	0.002		

Table 32: Description of the 21 CpG sites showing expression differences in AD hippocampus samples.

Six of the CpG sites had an effect size <5% and were thus dropped from further analysis. Three of the remaining fifteen CpG sites were in fact SNPs on the HM450 array. Primers could be designed for six of the remaining twelve loci. Two of these (*NGF* and cg20583640) reached genome wide significance and so were already selected using approach 1 (Section 4.3.1.4.2). The other four loci were followed up for Pyrosequencing analysis.

Twelve primer pairs were designed for analysis in at least one brain region. Table 33 shows a summary of loci followed up for validation in the brain.

Brain region	Variable	CpG site	Gene
D,H	Braak staging	cg00794813	***NGF***
D,H	Braak staging	cg03096785	***CTRC***
D,H	Braak staging	cg11796313	
D,H	Braak staging	cg20583640	
H	Braak staging	cg02126424	***CTRL***
H	Braak staging	cg18837178	
H	Braak staging	cg12010173	
H	Braak staging	cg01228342	***EIF4E3***
H	Memory remote	cg02232840	***KIAA1026***
H	Braak staging	cg15447017	***LMNA***
H	Braak staging	cg01063243	***MYCBP2***
H	Braak staging and Memory total	cg13223682	***OLA1***

Table 33: Summary of assays designed for Pyrosequencing analysis in the brain. D=DLPFC, H=Hippocampus.

4.3.2 Primer optimisation and validation

Using both approaches to select top hits, primer pairs for twelve loci were successfully designed as in Table 33. All twelve primer pairs targeted at least the index CpG (CpG included on the HM450). Where possible, amplicons were designed to capture neighbouring CpG sites.

Primer pairs for two loci could not be optimised and were removed from further analysis. The primer pairs for the remaining ten successfully optimised assays were used to amplify, by PCR, a range of methylation concentrations spanning 0% to 100% to ensure the assay was able to measure the full range of possible methylation values. An assay was considered to be validated if three conditions were reached: 1) the observed methylation value for 100% was above 80%; 2) the observed methylation value for 0% was below 10% and 3) there was a linear relationship with an R^2 value above 0.99. Five assays were unable to measure the full range of methylation values and were subsequently dropped from further analysis. Five assays were successfully validated. Concordance of observed and reference methylation values are shown in Table 34, using the mean methylation value across all CpG sites in each amplicon. The validation results for both the pre-PCR and post-PCR mixes were plotted as a scatter graph with a trend line and R^2 values presented (Figure 58).

Expected Methylation (%)	Observed Methylation (%)									
	NGF		cg18837178		cg12010173		*EIF4E3*		cg01063243	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
100	93.8	93.8	85.7	85.7	94.02	95.43	99.3	99.3	95.15	95.13
95	91.1	91.9	81.6	81.3	87.72	89.4	93.5	94.9	89.77	89.64
90	85.1	87.0	75.3	78.0	86.68	88.51	92.3	89.5	89.3	89.68
75	75.9	83.0	61.5	65.4	73.15	73.83	78.0	73.5	84.71	84.37
50	62.0	57.5	36.8	43.7	49.91	54.01	55.3	47.7	64.08	63.39
25	27.7	29.9	26.0	23.6	31.03	32.64	28.1	25.2	32.97	34.54
10	15.7	14.3	12.7	11.2	13.46	12.07	13.9	12.8	19.58	18.07
5	7.5	8.0	7.6	7.9	7.99	6.91	5.7	7.8	12.91	13.03
0	5.3	5.3	3.4	3.4	0.00	3.69	3.7	3.7	8.56	12.07

Table 34: Observed mean methylation across all CpGs significant in brain compared with reference expected methylation level. Pre and post PCR methylation values are presented for each assay.

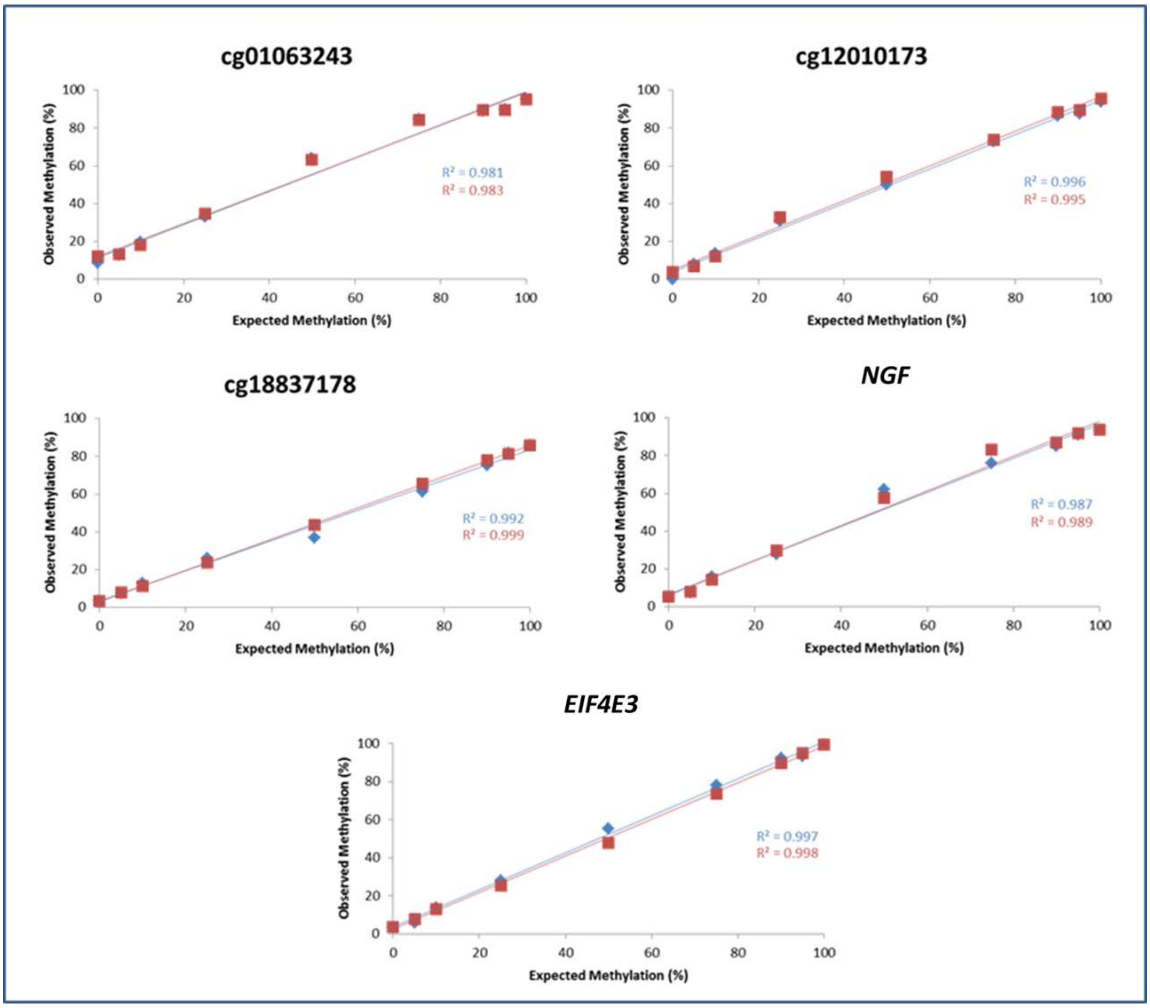


Figure 58: Composite figure of 5 assay validation plots. Pre-PCR methylation values are presented in blue and post-PCR methylation values in red. The R squared value is also displayed in the corresponding colour. An R squared value close to 1 indicates a very close correlation between observed and expected methylation levels.

4.3.3 Pyrosequencing analysis

All available DLPFC and hippocampus samples were then run across the validated assays in duplicate. Samples passed Pyrosequencing QC if duplicates were within 5% of one another. A mean methylation level for these duplicates was calculated for each sample and this mean methylation value was used in all further analyses.

Data generated by Pyrosequencing were first compared to the DMP methylation values generated by the HM450 BeadChip, to assess concordance and then analysed in relation to the outcomes and covariates observed to be associated with methylation in the discovery sample.

NGF was the only assay differentially methylated in DLPFC. All five validated assays were run across hippocampus samples and taken forward for analysis in Stata.

Each assay was analysed to test for an association between the outcome variable and the index CpG site (included on HM450) in the discovery sample set and a larger replication cohort. If multiple CpG sites were included on the Pyrosequencing assay, each neighbouring CpG site was analysed to test for association with outcome and finally the mean methylation value across the region (mean of index and neighbouring CpG sites).

In the DLPFC, the only validated assay associated with Braak staging, *NGF*, failed to work on the Pyrosequencer when using the COGFAST samples. It is likely this was due to a technical problem, possibly due to the DNA source, although faint PCR products were visualised. So at this stage, no assays were run on the Pyrosequencer to validate the HM450 findings in DLPFC tissue.

In the hippocampus, three of the assays which the HM450 identified as being associated with Braak staging (cg1201017, cg01063243 and *EIF4E3*) did not validate the findings of the HM450 and were deemed unlikely methylation markers for PSD. Two of the validated assays (*NGF* and cg18837178) did show significance at some of the tested CpG sites, the results of which are presented below.

4.3.3.1 ***NGF and cg18837178***

4.3.3.1.1 Validation in discovery cohort

All 28 hippocampus samples analysed on the HM450 were run on the Pyrosequencer. Duplicates for 23 of the 28 samples were within 5% of one another for the *NGF* assay and were taken forwards for analysis. A number of samples for cg18837178 either failed on the Pyrosequencer or had duplicates with a difference greater than 5% and were dropped from analysis, leaving data from 17 of the HM450 discovery cohort available for cg18837178 analysis.

Spearman’s rank correlation was performed in Stata to test the correlation between the two platforms. In both assays, the methylation values on each platform showed a weak correlation. Spearman’s correlation gave a coefficient of -0.250 (p=0.333) for cg18837178 and -0.198 (p=0.365) for *NGF* although this could be due to the majority of samples having a methylation score very close to 100% (Figure 59 and Figure 60).

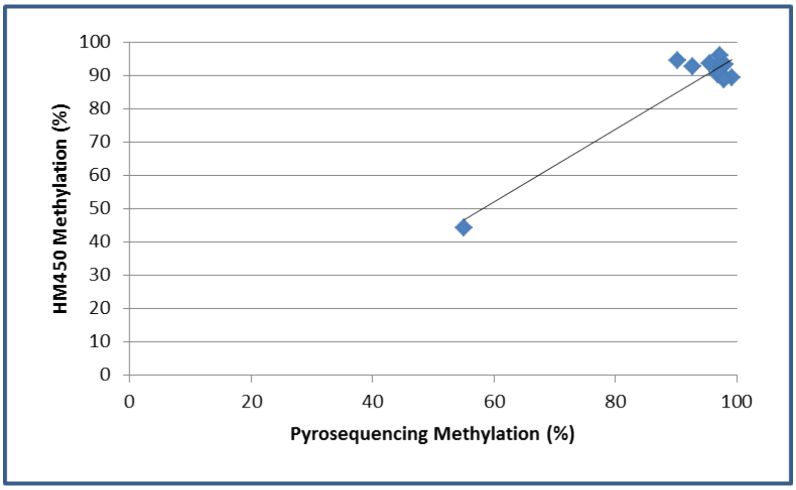


Figure 59: Correlation between methylation measured on the HM450 and Pyrosequencer at the cg18837178 locus.

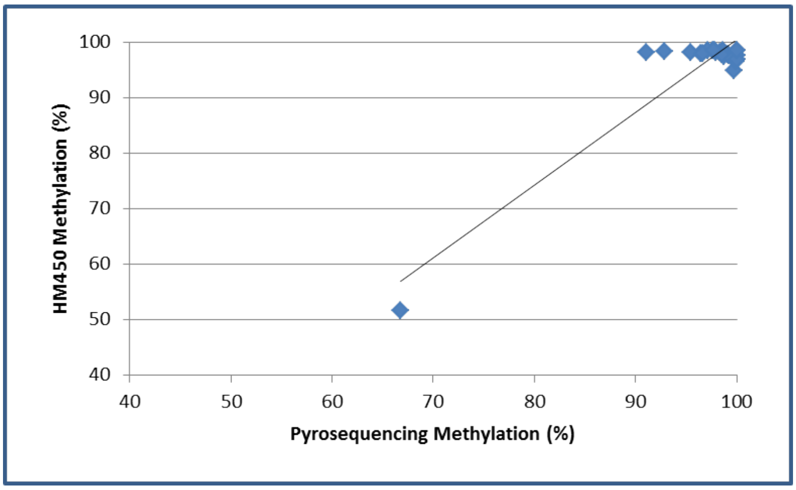


Figure 60: Correlation between methylation measured on the HM450 and Pyrosequencer at the *NGF* locus.

cg18837178 and *NGF* were both highlighted in the HM450 analysis as being associated with Braak staging. When split up into the six Braak stages the sample size issue becomes more apparent. Table 35 shows the number of samples within each Braak stage category in the discovery sample set. Only one Braak stage VI individual is included in the analysis meaning that any significance differentiating Braak stage VI from others is based on one individual. As with the blood samples, the HM450 data analysis was repeated with the Braak stage VI individual removed using the hippocampal and DLPFC data. Fewer hits were found to be significantly associated with Braak staging, as can be seen in the hippocampus plots (Figure 61 and Figure 62) and the DLPFC plots (Figure 63 and Figure 64). This indicates that the majority of CpG sites found to be associated with Braak staging were driven by this one individual.

Braak stage	**N in HM450**	**N in *NGF* discovery cohort**	**N in *NGF* replication cohort**	**N in cg18837178 discovery cohort**	**N in cg18837178 replication cohort**
I	4	4	6	3	4
II	7	6	10	5	9
III	10	8	13	5	10
IV	3	2	4	1	3
V	3	2	2	2	2
VI	1	1	1	1	1
Total	28	23	36	17	29

Table 35: Number of hippocampal samples in each cohort by Braak staging.

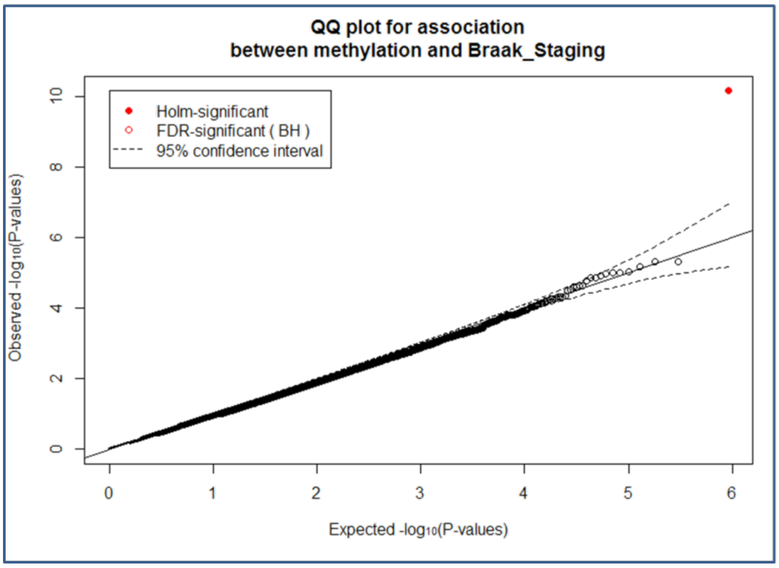


Figure 61: QQ plot for association between hippocampal methylation and Braak staging with the Braak stage VI individual removed. The QQ plot shows the expected against the observed p values. Data points plotted in red reach the Bonferroni corrected significance threshold.

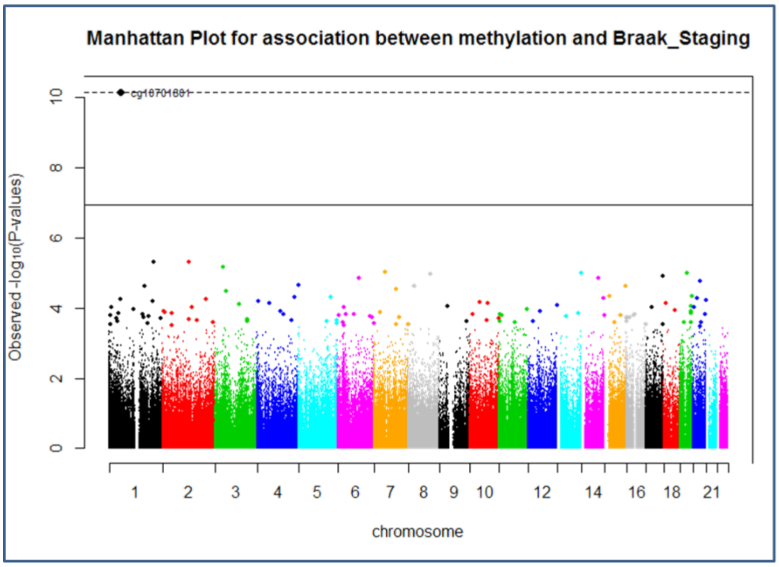


Figure 62: Manhattan plot for association between hippocampal methylation and Braak staging with the Braak stage VI individual removed. Each circle represents an individual CpG site. The continuous line marks the p<1.1x10-7 significance threshold. CpG sites with p values below the threshold are labelled with their Illumina ID.

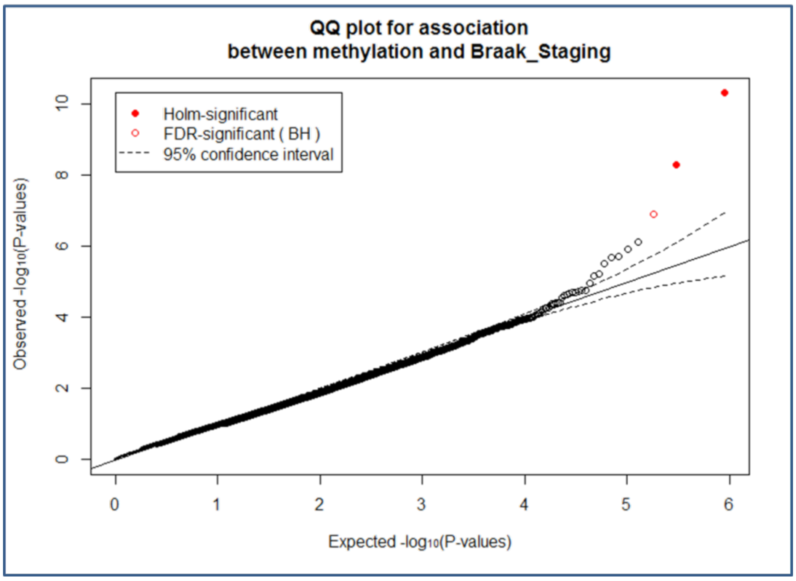


Figure 63: QQ plot for association between DLPFC methylation and Braak staging with the Braak stage VI individual removed. The QQ plot shows the expected against the observed p values. Data points plotted in red reach the Bonferroni corrected significance threshold.

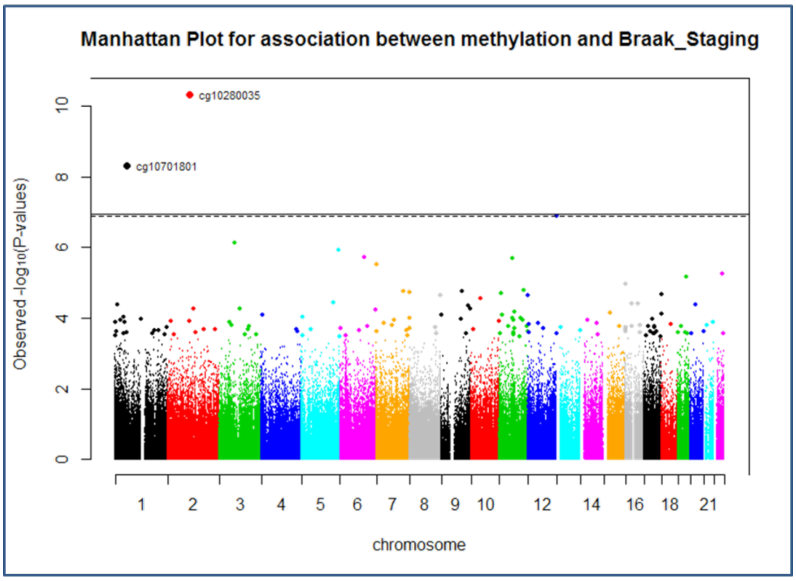


Figure 64: Manhattan plot f or association between DLPFC methylation and Braak staging with the Braak stage VI individual removed. Each circle represents an individual CpG site. The continuous line marks the p<1.1x10-7 significance threshold. CpG sites with p values below the threshold are labelled with their Illumina ID.

The HM450 array identified one CpG site within both cg18837178 and *NGF* to be differentially methylated in hippocampus samples of stroke patients with varying levels of Braak staging. Altered methylation values were found in Braak stage VI when compared to the other Braak stages (I-V).

Figure 65 shows the mean methylation levels for each Braak stage measured at the cg18837178 locus and Figure 66 shows the mean methylation levels for each Braak stage measured at the *NGF* locus on both the HM450 and Pyrosequencer. Both graphs show the mean methylation levels for each Braak stage are similar across the two different platforms in Braak stages I-V. The standard deviation bars indicate that the Pyrosequencing data is slightly more variable than the HM450 in both assays, however on both platforms the same pattern can be seen in relation to Braak staging, with the Braak stage VI individual having a much lower methylation level at both loci.

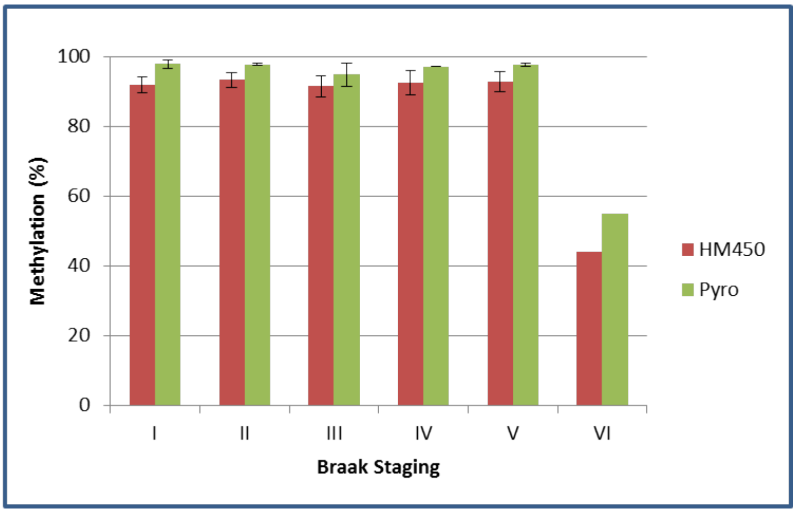


Figure 65: Mean cg18837178 locus methylation levels observed for each Braak stage measured on the HM450 and Pyrosequencer.

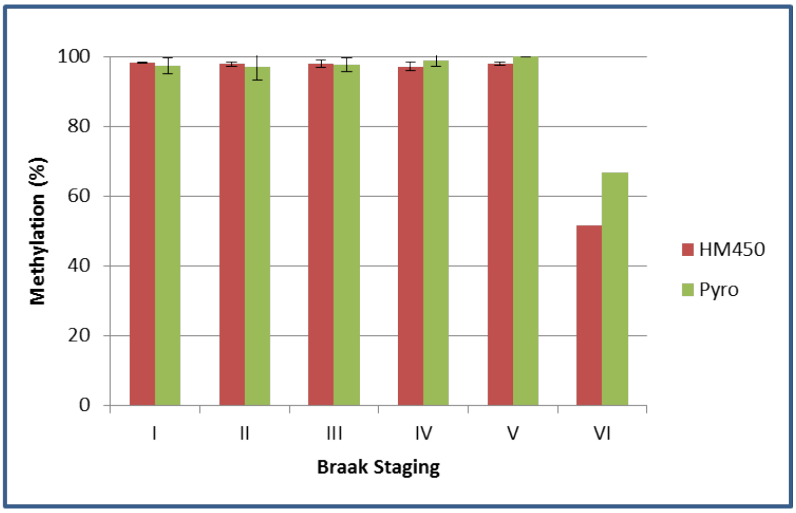


Figure 66: Mean *NGF* locus methylation levels observed for each Braak stage measured on the HM450 and Pyrosequencer.

Figure 67 and Figure 68 show the spread of methylation measured at the index CpG site using both platforms for cg18837178 and *NGF*, respectively. Both graphs show the Braak stage VI individual as an outlier on both platforms; however the difference between this individual and the mean methylation level across all samples is much lower in Pyrosequencing than HM450.

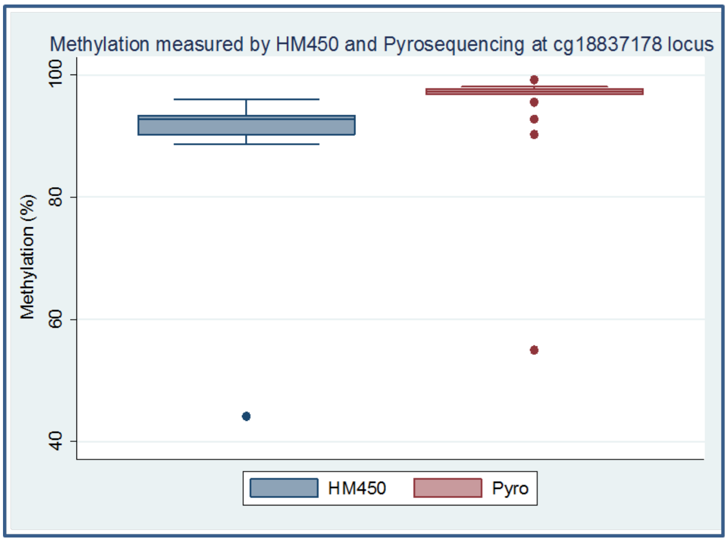


Figure 67: Box plot to show methylation levels measured using the HM450 and Pyrosequencing at the CpG site cg18837178.

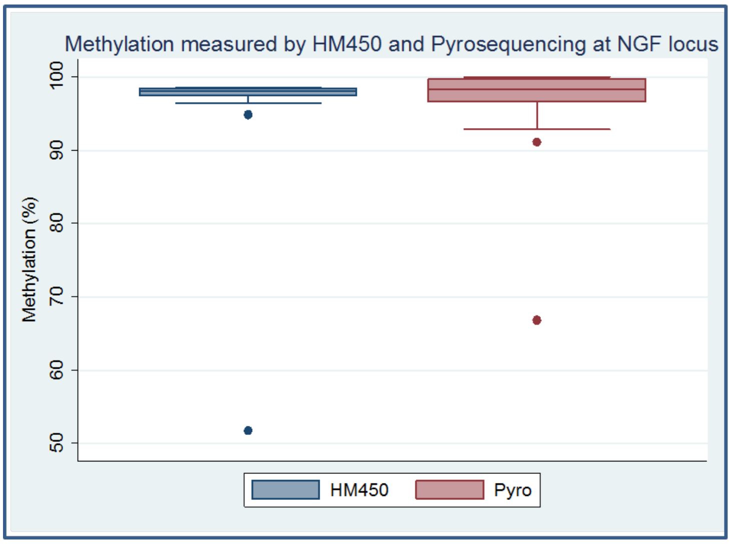


Figure 68: Box plot to show methylation levels measured using the HM450 and Pyrosequencing at the index CpG site in *NGF*.

The analysis model used to assess the relationship between Braak staging and the HM450 data was applied to the Pyrosequencing data. This was to test whether the association between methylation and Braak staging was still present in the Pyrosequencing data. In Stata, ANOVA was performed (as in the analysis of HM450 data) to test for association between methylation at the index site and Braak staging, using sex and age at death as covariates. 17 hippocampal DNA samples were included in the cg18837178 analysis. 23 were included in the analysis of the *NGF* locus. Both loci were validated using Pyrosequencing when analysed using ANOVA; cg18837178 p=0.003; *NGF* p= 0.0004. Due to the non-normally distributed nature of the DNA methylation data (tested using a Shapiro-Wilk normality test), a Kruskal Wallis test was also applied but no significant differences were observed; (cg18837178 p= 0.295; *NGF* p= 0.216).

When the individual with Braak stage VI pathology was removed from the analysis, there was no significant association between methylation at either locus (cg18837178 and *NGF*) and Braak staging using ANOVA (cg18837178 p=0.999; *NGF* p=0.752) or Kruskal Wallis (cg18837178 p=0.832; *NGF* p=0.337). This indicates that the results are heavily influenced by this outlier.

4.3.3.1.2 Validation in Pyrosequencing cohort

The Pyrosequencing assays were run across a larger sample set (cg18837178 - n=29; *NGF* – n=36) to test whether the association with Braak staging at the index CpG site was still observed in a larger cohort. ANOVA revealed that the association between Braak staging and methylation at the cg18837178 locus was indeed even more significant using the larger cohort (p=5.322E-07), but this significance was not replicated when analysed using Kruskal Wallis (p=0.522). Likewise, the association between Braak staging and methylation at the *NGF* locus was also more significant when a larger sample size was used with ANOVA (p=0.0001) but not with Kruskal Wallis analysis (p= 0.1877).

4.3.3.1.3 Neighbouring CpG sites

It is also of interest to look at neighbouring CpG sites to assess whether these CpG sites are also associated with Braak staging. The cg18837178 assay was designed to include the index CpG site along with one neighbouring site. The *NGF* assay was designed to include the index CpG site plus three neighbouring sites. The Pyrosequencing data for the neighbouring CpG site was also analysed using ANOVA and Kruskal Wallis in both the HM450 cohort and the larger Pyrosequencing cohort. Results of ANOVA and Kruskal Wallis analyses are presented in Table 36.

CpG site	**P value**	
	cg18837178	*NGF*
Index CpG		
Discovery	0.003 (0.295)	0.0004 (0.216)
Pyrosequencing	5.32E-07 (0.522)	0.0001 (0.189)
Neighbouring site 1		
Discovery	0.220 (0.730)	0.071 (0.514)
Pyrosequencing	0.516 (0.455)	0.833 (0.218)
Neighbouring site 2		
Discovery	NA	0.550 (0.902)
Pyrosequencing	NA	0.534 (0.497)
Neighbouring site 3		
Discovery	NA	0.326 (0.846)
Pyrosequencing	NA	0.904 (0.791)
Region		
Discovery	0.003 (0.293)	0.294 (0.623)
Pyrosequencing	0.0001 (0.687)	0.268 (0.223)

Table 36: ANOVA (and Kruskal Wallis) results for cg18837178 and *NGF* Pyrosequencing. The methylation of the region is defined as the mean methylation across all CpG sites covered in the Pyrosequencing assay. P values are displayed for both the discovery and Pyrosequencing cohort for each CpG site/region.

Although both index CpG sites within cg18837178 and *NGF* were found to be significantly associated with Braak staging in both cohorts when analysed using ANOVA, no neighbouring CpG within either assay measured by Pyrosequencing was associated with Braak staging using either the discovery or replication cohort. When the mean of all CpG sites in the amplicon were analysed, the cg18837178 locus still remained significantly associated with Braak staging whereas the *NGF* locus did not. All data analysis performed using Kruskal Wallis showed no statistically significant differences in DNA methylation between Braak staging.

These data show that although the data generated on the HM450 and Pyrosequencer were not identical and only weakly correlated by Spearman’s rank, the same patterns were seen in relation to Braak staging. The HM450 findings that the Braak stage VI individual had a significantly different methylation level to other Braak stages at the index CpG site within cg18837178 and *NGF* were validated using Pyrosequencing.

4.3.3.1.4 Further Braak staging analysis

Like the analysis in the blood samples in Chapter 3, additional samples were analysed by Pyrosequencing. Three Braak stage V and seven Braak stage VI (demented, no stroke) hippocampal samples were analysed in both assays.

Figure 69 shows the methylation of the index CpG within cg18837178 measured by Pyrosequencing in these additional Braak stage V and VI individuals compared to the COGFAST participants. Figure 70 shows the methylation of the index CpG within *NGF* measured by Pyrosequencing in Braak stage V and VI individuals compared to the COGFAST participants.

In both assays, the COGFAST Braak stage VI individual was unique in its methylation patterns, indicating that the results in this Chapter are not representative of all Braak stage VI pathology. However, this finding does not necessarily mean that this is a false positive, as described in Chapter 3.

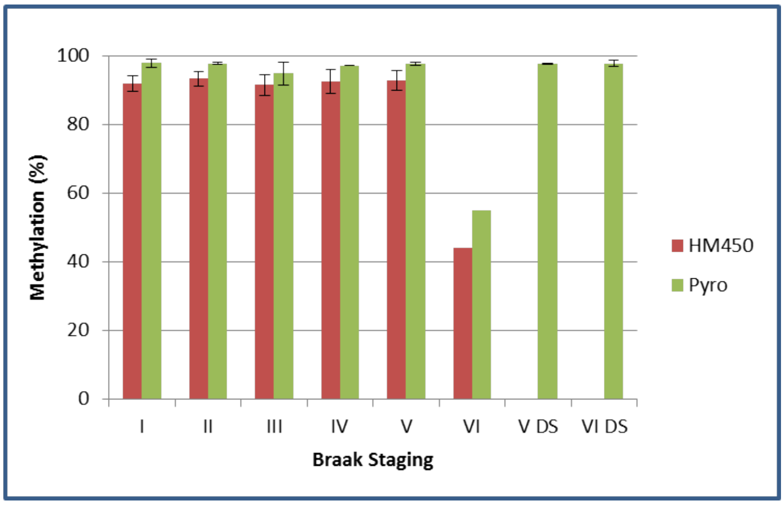


Figure 69: Methylation at cg18837178 locus measured in hippocampal DNA in COGFAST and demented Braak stage V and Braak stage VI individuals. Braak stages are displayed on the X axis with demented Braak stage V and Braak stage VI denoted as V DS and VI DS, respectively.

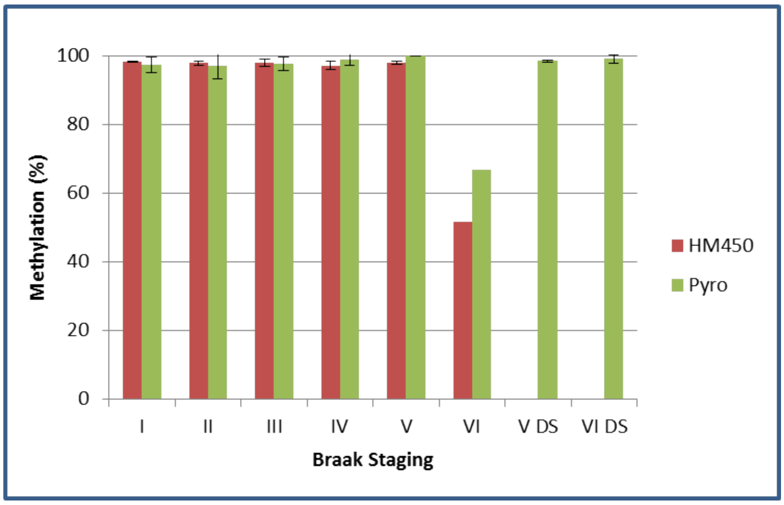


Figure 70: Methylation at *NGF* locus measured in hippocampal DNA in COGFAST and demented Braak stage V and Braak stage VI individuals. Braak stages are displayed on the X axis with demented Braak stage V and Braak stage VI denoted as V DS and VI DS, respectively.

4.3.3.2 *cg18837178 in the dorsolateral prefrontal cortex*

Although cg18837178 was not found to be differentially methylated in the DLPFC, based on the two selection processes used in this Chapter, it was decided to run the cg18837178 assay across DLPFC samples since HM450 results were validated in the hippocampus.

4.3.3.2.1 Validation in discovery cohort

22 of the 24 DLPFC DNA samples analysed on the HM450 were successfully run on the Pyrosequencer, with repeats within 5% of one another.

Spearman's rank correlation was performed in Stata to test the correlation between the two platforms. The coefficient showed a very weak correlation (Rho=-0.098) between the two platforms, possibly due to the non-linear relationship, with almost all methylation values being close to 100%, although this was not significant (p=0.965) (Figure 71).

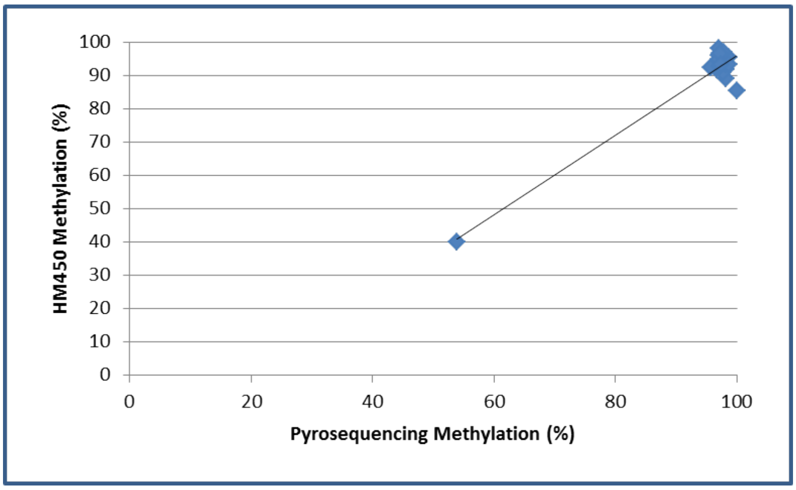


Figure 71: Correlation between methylation measured on the HM450 and Pyrosequencer at the cg18837178 locus.

Although not reaching genome-wide significance, the HM450 did detect differences between Braak stages at the cg18837178 locus in DLPFC samples. Figure 72 shows the mean methylation levels for each Braak stage measured at the cg18837178 locus on both the HM450 and Pyrosequencer. The mean methylation levels for each Braak stage are similar across the two different platforms in Braak stages I-V. The standard deviation bars indicate that the Pyrosequencing data are slightly more variable than the HM450 in both assays, however on both platforms the same pattern can be seen in relation to Braak staging, with the Braak stage VI individual having a much lower methylation level at both loci.

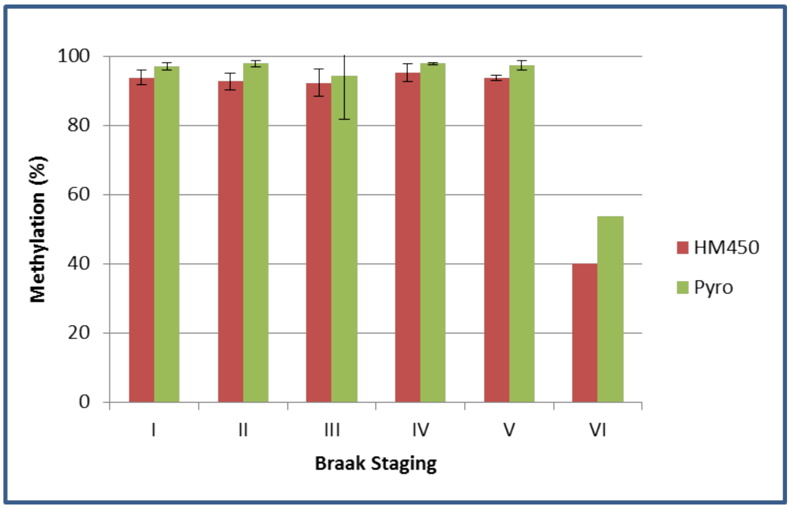


Figure 72: Mean cg18837178 methylation levels observed for each Braak stage measured in the DLPFC on the HM450 and Pyrosequencer.

Figure 73 shows the spread of methylation measured at cg18837178 using both platforms. This shows the Braak stage VI individual as an outlier on both platforms; however the difference between this individual and the mean methylation level across all samples is lower in Pyrosequencing than HM450.

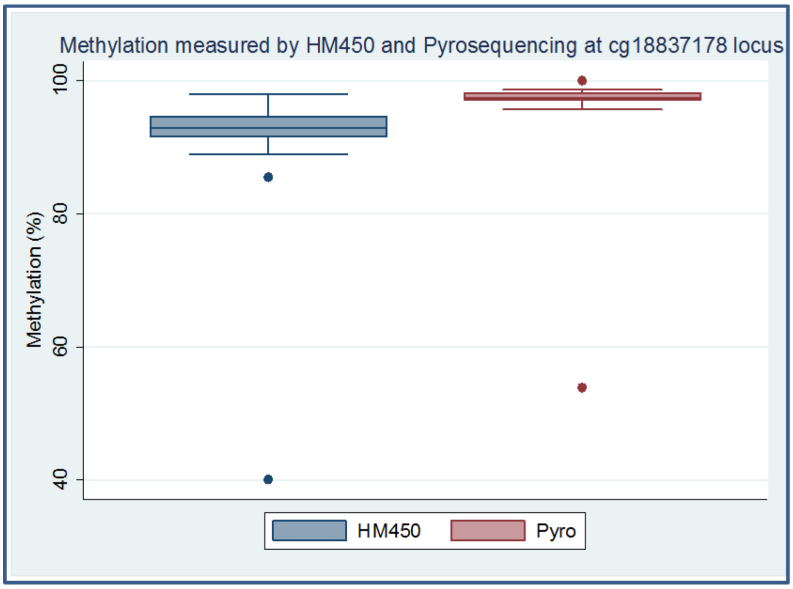


Figure 73: Box plot to show methylation levels measured using the HM450 and Pyrosequencing at the CpG site cg18837178.

The analysis model used to assess the relationship between Braak staging and the HM450 data was applied to the Pyrosequencing data. This was to test whether the association between methylation and Braak staging was still present in the

Pyrosequencing data. In Stata, ANOVA was performed (as in the analysis of HM450 data) to test for association between methylation at the index site and Braak staging, using sex and age at death as covariates. 22 DLPFC DNA samples were included in the analysis and the HM450 findings were validated using ANOVA on the Pyrosequencing data at this CpG site (p= 0.0004). As with all other CpG sites, analysis using Kruskal Wallis did not find significant differences (p=0.498).

When the individual with Braak stage VI pathology was removed from the analysis there was no significant association with methylation and Braak staging using ANOVA (p=0.853) or Kruskal Wallis (p=0.150). This indicates that the results are heavily influenced by this outlier.

4.3.3.2.2 Validation in Pyrosequencing cohort

The Pyrosequencing assay was run across a larger sample size (n=36) to test whether the association with Braak staging at the index CpG site was still observed in a larger cohort. ANOVA revealed that the association between Braak staging and methylation at the cg18837178 locus was not significant using the larger cohort (p=0.210), nor was it significant when analysed using the non-parametric Kruskal Wallis (p=0.482).

4.3.3.2.3 Neighbouring CpG sites

The cg18837178 assay was designed to include the index CpG site along with one neighbouring site. The Pyrosequencing data for the neighbouring CpG site were also analysed using ANOVA and Kruskal Wallis in both the HM450 cohort and the larger Pyrosequencing cohort but no significance was observed using either test. Methylation across the region (mean methylation of both CpG sites targeted in the Pyrosequencing assay) was found to be associated with Braak staging when analysed using ANOVA, but there was no evidence of a statistically significant methylation difference using Kruskal Wallis (Table 37).

CpG site	P value
	cg18837178
Index CpG	
Discovery	0.0004 (0.498)
Pyrosequencing	0.210 (0.482)
Neighbouring site 1	
Discovery	0.144 (0.959)
Pyrosequencing	0.109 (0.814)
Region	
Discovery	0.004 (0.699)
Pyrosequencing	0.297 (0.645)

Table 37: ANOVA (and Kruskal Wallis) results for cg18837178 Pyrosequencing. The methylation of the region is defined as the mean methylation across all CpG sites covered in the Pyrosequencing assay. P values are displayed for both the discovery and Pyrosequencing cohort for each CpG site/region.

These findings show that in the HM450 discovery cohort there were significant differences in methylation levels between Braak stages, however when analysed in a larger cohort, these differences were no longer significant, indicating that cg18837178 is unlikely to be a methylation marker of PSD in the DLPFC.

4.3.4 Protein analysis

In the hippocampus, two Pyrosequencing assays successfully validated the HM450 findings. These were *NGF* and cg18837178. *NGF* is a well described gene whereas cg18837178 is a non-coding RNA. For this reason, NGF was selected to be taken forward for protein analysis by Western blotting, to test whether the changes in methylation seen in the blood and hippocampus were seen at the protein level in the brain. Although the Pyrosequencing assay failed to work using the DLPFC samples, protein levels were measured in both the hippocampus and DLPFC.

4.3.4.1 *NGF optimisation*

Four different conditions were tested to find the optimum blotting conditions for anti-NGF. Figure 74 shows the different conditions tested.

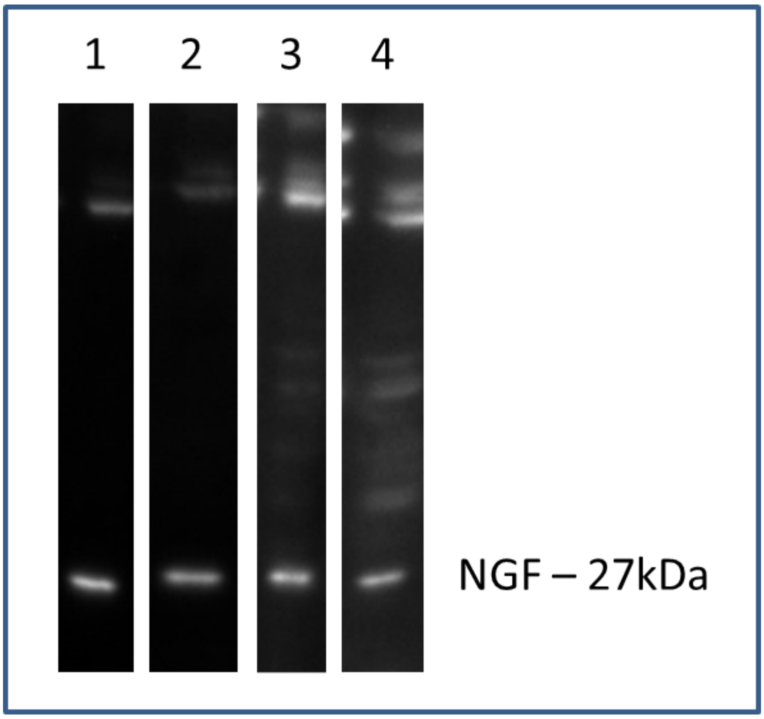


Figure 74: Optimisation of anti-NGF. 1- 5% milk block, 5% milk blot. 2- 5% milk block, 1% milk blot. 3- 5% BSA block, 5% BSA blot. 4- 5% BSA block, 1% BSA blot.

Lane 1 was selected as the optimum blotting conditions for anti-NGF as it gave the brightest band with little background. The optimum conditions for anti-NGF were identified as blocking with 5% milk followed by preparation of anti-NGF antibody in a 5% milk diluent.

4.3.4.2 *Western blots*

Ten COGFAST participants were selected for Western blot analysis. Two participants for each Braak stage, with the exception of Braak stage VI, were included in the Western blot analysis. Unfortunately, insufficient tissue was available for the Braak stage VI sample used in the methylation analysis and could not be included in this experiment. For this reason, it was impossible to investigate whether the methylation differences seen in the COGFAST Braak stage VI individual were translated at the protein level. However, it was still of interest to investigate whether NGF protein levels differed across the various remaining Braak stages, and since it has been implicated in Alzheimer’s disease, it was also of interest to look at NGF levels in AD compared to controls and COGFAST participants. In addition to the ten COGFAST participants already mentioned, six Alzheimer’s disease samples (DS) were included, two had Braak stage V pathology and four had Braak stage VI pathology. In order to test whether there were differences between dementia, stroke and controls, six healthy controls (CA) with no known stroke or dementia history were also included, four of these had Braak stage

II pathology and two had Braak stage III pathology. Numbers of samples per Braak stage in each of the cohort groups used in the Western blot are summarised in Table 38. All Western blots can be found in Appendix B.

Cohort	**Braak stage (n)**					
	I	II	III	IV	V	VI
COGFAST	2	2	2	2	2	
DS					2	4
CA		4	2			

Table 38: Number of samples with each Braak stage pathology in each of the cohort groups. DS refers to Alzheimer's cases with no stroke. CA refers to healthy controls.

First, DLPFC and hippocampal levels of NGF were compared across all samples. Due to much lower alpha-tubulin levels observed in either the DLPFC or hippocampal samples across five individuals, NGF values for both tissues measured in these participants were removed from this analysis. There was a very weak correlation between NGF levels in the DLPFC and hippocampus although this difference was not significant (Spearman's Rho = -0.027, p=0.594) (Figure 75A).

For all further analysis, the DLPFC and hippocampus were considered separately. One DLPFC sample was removed from the analysis due to low alpha-tubulin levels and five hippocampal samples were removed for the same reason.

Next, NGF levels were analysed in regard to cohort group. No differences in relative NGF protein were observed between COGFAST, DS and CA cohort groups in the DLPFC (p=0.083) (Figure 75B) or hippocampus (p=0.194) (Figure 75C).

Braak stages were then grouped across cohorts and NGF levels were compared between Braak stages. No differences were observed between Braak stage in either the DLPFC (p=0.218) (Figure 75D) and hippocampus (p=0.281) (Figure 75E).

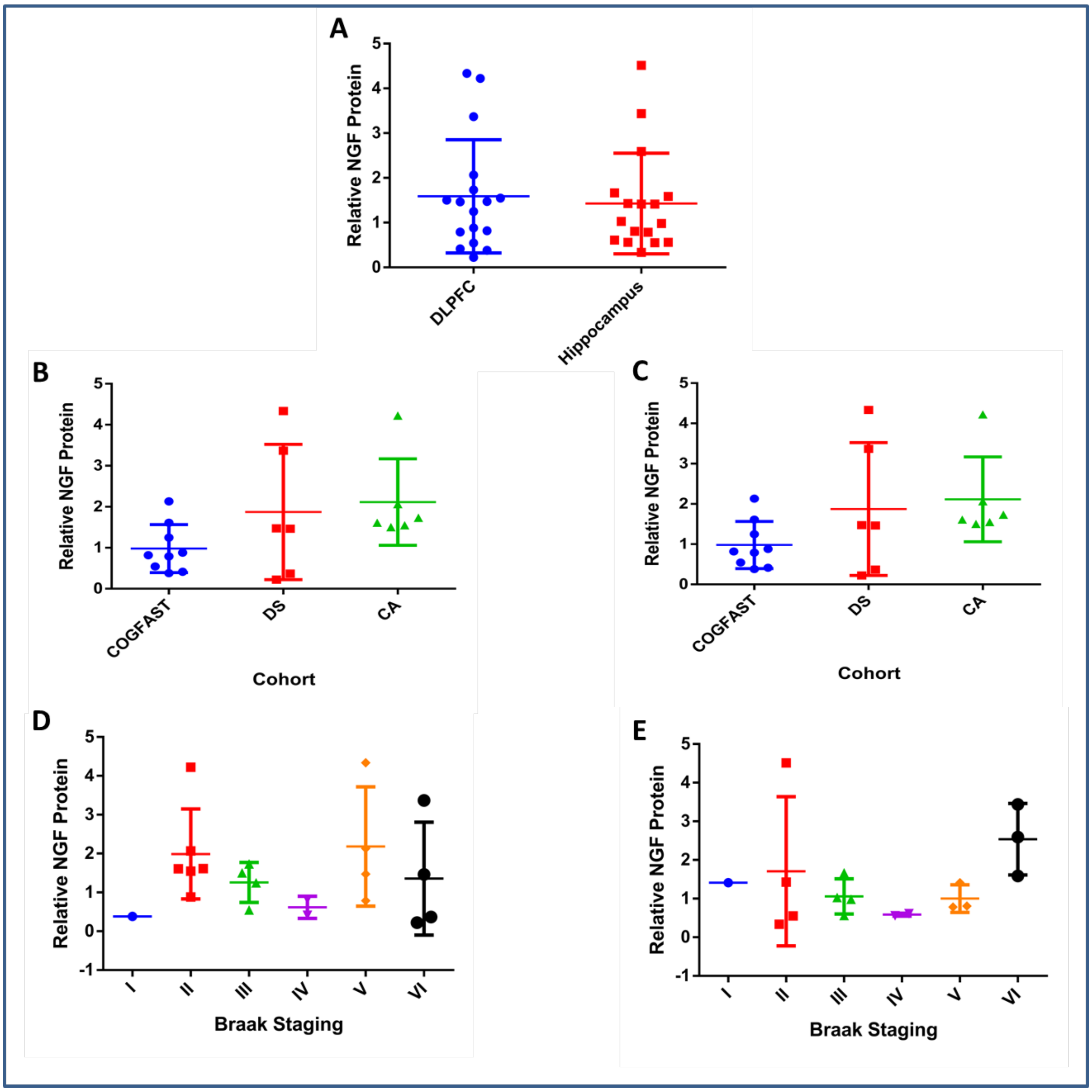


Figure 75: Western blot analysis. A. Comparison between relative NGF levels in the DLPFC and hippocampus. B. Relative NGF levels in the DLPFC across cohort groups. C. Relative NGF levels in the hippocampus across cohort groups. D. Relative NGF levels in the DLPFC across Braak stages. E. Relative NGF levels in the hippocampus across Braak stages. Each dot represents a data point, the centre line represents the mean value and the two outer lines represent 1 standard deviation (SD) from the mean. Any points outside of these lines are more than 1SD away from the mean. DS= demented, no stroke. CA= healthy controls.

Finally, correlation analysis was performed to look for correlation between methylation and protein levels. Since the DLPFC assay did not work on the Pyrosequencer, correlation analysis was only performed on the hippocampal samples. Five of the hippocampal COGFAST samples included in the protein analysis were successfully run using the NGF assay. Spearman’s rank correlation was performed and a strong inverse association was observed (Spearman’s Rho = -0.900, p=0.037) (Figure 76), indicating that increased methylation is associated with reduced protein levels.

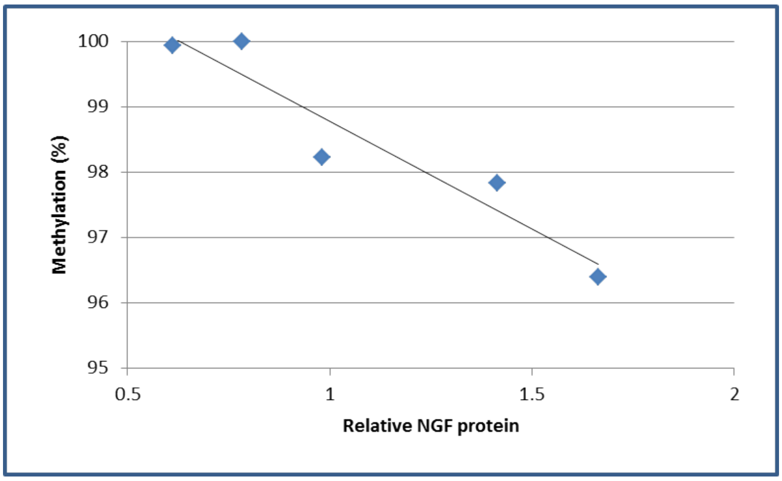


Figure 76: Correlation between hippocampal *NGF* methylation and NGF protein levels.

4.3.5 *APOB* in the brain

Since *APOB* was identified as a possible blood based biomarker, the DLPFC and hippocampal samples were run across the *APOB* assay to see whether the differential methylation in Braak stage VI was also observed in brain tissue. Neither the DLPFC or hippocampal tissues showed the same methylation differences as blood (Figure 77 and Figure 78), nor was methylation at the *APOB* locus associated with Braak staging when analysed using ANOVA (DLPFC p=0.330; hippocampus p= 0.268) and Kruskal Wallis (DLPFC p=0.629; hippocampus p= 0.582).

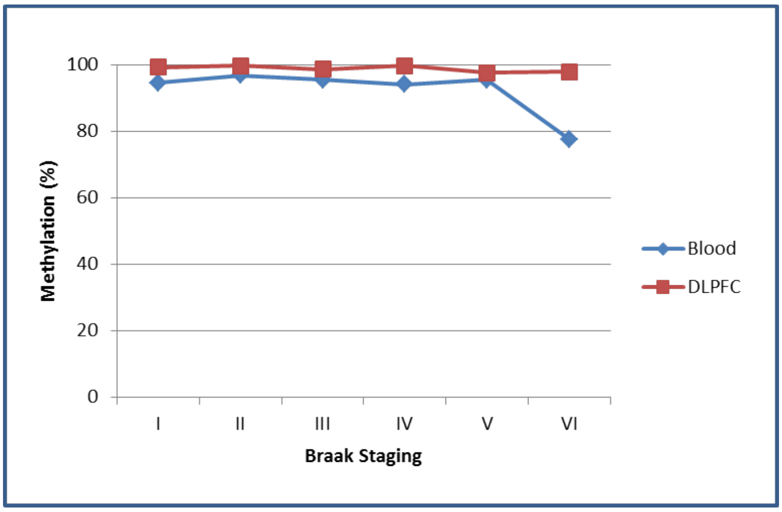


Figure 77: Comparison of mean methylation levels by Braak staging in the blood and DLPFC using Pyrosequencing.

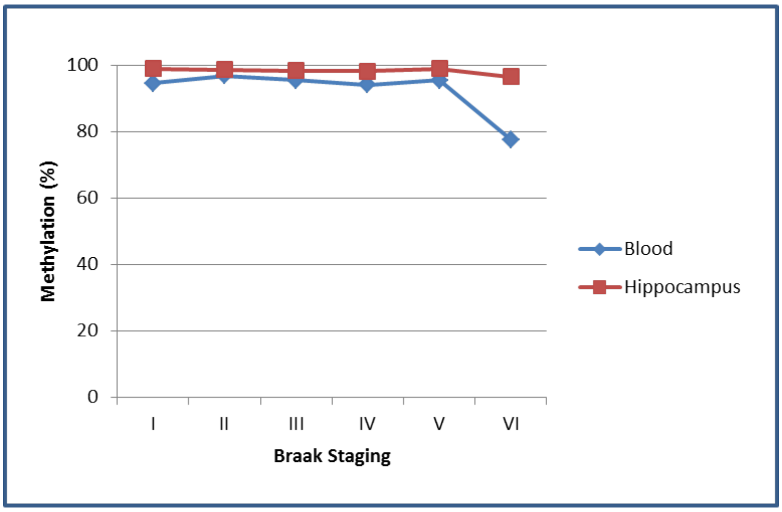


Figure 78: Comparison of mean methylation levels by Braak staging in the blood and hippocampus using Pyrosequencing.

4.4 Discussion

The work carried out in this Chapter aimed to identify DMPs which may be useful in determining the mechanisms involved in PSD using hippocampal and DLPFC brain samples. The HM450 array identified ten CpG sites in the DLPFC and 34 CpG sites in the hippocampus that reached genome-wide significance. An additional 23 sites were identified to be differentially expressed in either the DLPFC or hippocampus in AD. Following the filtering of top hits, two assays were run across DLPFC samples and five assays were run across the hippocampal samples using a larger COGFAST cohort to test whether the association seen between Braak staging and methylation still persisted when measured using Pyrosequencing. When Pyrosequencing data were analysed using ANOVA, two of these assays (cg18837178 and *NGF*) successfully validated the HM450 findings in at least one tissue, suggesting that they may be potential methylation markers mechanistically involved in the pathogenesis of PSD. No significance was observed when data were analysed using Kruskal Wallis, however this is likely to be due to the low sample size limiting the power to detect differences.

As described in Chapter 3, the sample size used in this study is small making analysis of methylation by Braak staging difficult. The results of both this and the previous Chapter are based on the methylation differences observed in one individual. However, the finding that this methylation difference is observed across all three tissues (blood, DLPFC and hippocampus) studied in COGFAST adds confidence to this finding and suggests that the methylation difference unique to this individual is not just an artefact. Again, like in Chapter 3, the methylation differences seen in the COGFAST Braak stage VI individual were not seen in other Braak stage VI individuals with dementia who had not experienced a stroke. This indicates that the methylation signatures are likely due to a combination of stroke and Braak stage VI pathology or some confounding factor. As described in Chapter 3, the Braak stage VI COGFAST case had experienced three recurrent strokes prior to recruitment It is thought that these events may have altered the methylation patterns which were detected here. The clinical characteristics of this case were described in detail in Chapter 3.

Due to cg18837178 being located within a long non-coding RNA (CT49), protein analysis was confined only to NGF. NGF was a very relevant protein to investigate due to previous research indicating that differences in protein levels exist between AD and controls, in addition to the possible role of NGF in disease pathogenesis, explored below. NGF was also particularly interesting since the publicly available expression data set used here revealed differential expression of *NGF* in the hippocampus of AD cases. This suggests that differences may also be seen at the protein level. Although the expression data set used in this study was an AD cohort and therefore not entirely comparable to PSD it is expected that similar pathways are involved in the progression of disease.

β-Amyloid precursor protein (APP) plays a key role in the development of AD as cleavage of APP forms β-Amyloid (Aβ), the component of extracellular senile plaques in the brain which are an important hallmark of AD. APP has also been associated with reduced transportation of NGF in the brain of Alzheimer's cases. It has been found that the overexpression of NGF is associated with reduced transport of this protein. The molecular mechanisms and pathways linking APP with NGF signalling are yet to be identified. This reduction in transport of NGF has been found to increase the vulnerability of cholinergic neurons in AD and lead to Aβ deposition and associated memory deficits (Zhang *et al.*, 2013). These findings indicate that NGF may have a mechanistic role in the progression of AD and possibly other dementias; however, more research is required to elucidate the mechanistic pathways involved. Interestingly, NGF levels have been studied in other neurodegenerative disorders such as Parkinson's disease and NGF is not thought to be involved in disease pathogenesis as no differences in NGF levels have previously been observed between PD cases and controls (Scott *et al.*, 1995).

Unfortunately, due to the omission of the COGFAST participant with Braak stage VI pathology from the protein work described here, it is not within the scope of this thesis to investigate whether the methylation changes observed at the NGF locus in the DLPFC and hippocampus are translated to the protein level. However, although the COGFAST Braak stage VI individual was not included in the protein analysis, it can be predicted, based on the correlation analysis, that the COGFAST Braak stage VI individual would have higher levels of NGF than other COGFAST participants. This theory is supported by previous researchers who have identified higher Braak stages and AD pathology to be associated with increased NGF protein levels in the brain (Scott *et al.*, 1995; Fahnestock *et al.*, 1996; Mufson *et al.*, 2012b). Although the majority of publications support the finding of increased NGF levels in AD cases, some findings within dementia cohorts are inconsistent. AD patients have been shown to exhibit dysfunctional transportation of NGF between brain regions (Zhang *et al.*, 2013). Due to this, depending on which brain region a particular study has sampled, the level of NGF measured may not be comparable to studies that have used an alternative brain region. Some papers have reported reduced levels of NGF in the brains of AD cases when compared to controls (Mufson *et al.*, 1999). In a study investigating serum NGF levels, healthy controls were found to have significantly higher NGF levels than AD cases (Konukoglu *et al.*, 2012). The protein study carried out in this Chapter did not however detect any differences in NGF protein level between AD cases and cognitively normal controls in either the DLPFC or hippocampus.

One of the strengths of this study is that the protein levels were analysed to try and identify an association between DNA methylation and protein levels. However, this approach could not tell us whether methylation had a direct impact on the protein level. An alternative approach to study this would be to use a cell culture approach. Cells could be cultured in the presence of 5-aza in order to cause genome-wide demethylation (Enright *et al.*, 2003). The expression of the target gene could be assessed at a transcript level using quantitative PCR (Yang *et al.*, 2002) and at a protein level using Western blotting (Palii *et al.*, 2008). This would determine whether the expression of the gene in question is affected by DNA methylation. The link between transcription and translation for the target gene could be assessed through the use of small interfering RNAs (siRNA) which would transiently knock down the level of gene transcript. Conversely, expression vectors could be used to drive the expression of the gene of interest (Alekseev *et al.*, 2009). The subsequent effects on protein level could then be assessed by Western blotting.

The finding that *NGF* methylation levels are reduced in a PSD case who had experienced recurrent strokes is very interesting. Although brain protein levels were not measured in this individual it can be tentatively postulated that increased NGF protein levels would be observed. A stroke is known to promote angiogenesis, the formation of new blood vessels in the brain to restore blow flow to affected areas. Angiogenesis requires growth factors and a number of growth factors have been found to be increased following a stroke (Krupinski *et al.*, 1994; Font *et al.*, 2010). In the case of recurrent strokes it can be expected that due to increased ischaemia and therefore increased angiogenesis, levels of NGF protein and other growth factors are found at increased levels. This theory correlates well with the findings of this thesis, which identified reduced DNA methylation levels of *NGF* in the individual who had experienced recurrent strokes. As decreased DNA methylation is often associated with increased gene expression and protein levels it can be hypothesised that the COGFAST Braak stage VI case would have increased NGF protein levels. Given the biological plausibility of this scenario, it would be very interesting to perform this experiment if suitable tissue was available to see if this hypothesis is valid.

CT49 is a potential candidate for future functional investigation. The expression of CT49 has been directly linked to DNA methylation, as demonstrated by Adair and Hogan (2008) who treated ovarian cancer cell lines with 5-aza and observed an increased level of CT49 transcript (Adair and Hogan, 2009). It would be interesting to investigate the role of CT49 in PSD by first quantifying its expression using quantitative PCR in a validation cohort. If an association between DNA methylation and expression was established in these cases, a functional study could be performed. The biological function of CT49 is currently unknown, however it is thought that many long non-coding RNAs have a functional role (Dinger *et al.*, 2009; Mercer *et al.*, 2009), for example, in the regulation of gene-specific transcriptional regulation (Goodrich and Kugel, 2006). After driving the expression of CT49 using an expression construct, an expression array could be utilised to determine which genes, if any, are regulated by CT49. These targets could then be followed up to determine their role in PSD.

In summary, the work described in this Chapter sought to establish whether identification of differential DNA methylation in post-mortem brain tissue from PSD cases could shed light on potential mechanisms in this disease. The study findings were constrained by sample size and the availability of samples with the most severe form of brain pathology (Braak stage VI).

Strengths of the approach included use of the target (brain) tissue rather than reliance on surrogate (blood) tissue. The inclusion of Western blot analysis to quantify differences in protein levels also added strength to the study design, providing an opportunity to interrogate functional links between differential DNA methylation and disease.

Due to the assay design limitations, only a few loci were followed right through to the validation phase by Pyrosequencing. Other techniques such as Sequenom EpiTYPER (Ehrich *et al.*, 2005) or bisulphite sequencing (Herman *et al.*, 1996) could be employed to quantify DNA methylation levels at these loci. There are a number of limitations associated with the selection of top hits which have been outlined in Chapter 3. Briefly these limitations include the selection of only those hits which showed an inverse relationship between DNA methylation and expression, which is unlikely to be true. In addition, the selection of top hits was based on expression data from only one study. Taking an average from multiple expression studies would have been a better, more accurate approach. The study used was chosen due to the brain regions included, however these were AD cases so not entirely comparable to PSD cases.

The HM450 analysis was adjusted for age, sex and chip. PM delay was not found to affect DNA methylation and so was not included as a confounder. It is possible, however, that differential DNA methylation identified here is due to other confounders, not adjusted for in the analysis, rather than statistically significant loci having a mechanistic role in the pathogenesis of disease. Further validation is required in a larger cohort to strengthen the associations identified here.

Due to the heterogeneous nature of brain tissue, it is possible that any differences in DNA methylation observed between outcome groups is due to the cellular composition of the brain. Methylation beta values were adjusted for cellular composition, to determine whether the differential methylation observed in relation to Braak staging was actually due to differences in cellular composition. Adjusting for cellular composition revealed small differences in the significance level, indicating that differences in cellular composition have very little effect on the methylation differences observed between outcome groups in this Chapter.

Chapter 5. Blood-based methylation signatures as a predictor of Parkinson's disease

5.1 Introduction

Parkinson's disease (PD) is the second most common neurodegenerative disorder affecting approximately 1% of the over-65 population (Saracchi *et al.*, 2014). It is characterised by the presence of certain motor deficits and is primarily a motor disorder, however many patients are affected by cognitive impairments. Parkinson's disease dementia (PDD) affects up to 80% of PD patients. Many more PD patients are diagnosed with mild cognitive impairment (PD-MCI) and are still able to live independently (Yarnall *et al.*, 2013). It is unknown why some PD patients develop MCI and even PDD whilst others remain cognitively normal. It is therefore difficult to develop treatments when the mechanisms which result in MCI are not well understood. Often, the cognitive symptoms precede the motor symptoms (Saracchi *et al.*, 2014) so the ability to predict those who are likely to develop PD-MCI would be beneficial, so that preventative treatments could be implemented for those individuals who were at highest risk of progressing to PDD or developing other PD traits.

As previously mentioned in Section 3.1, associations between DNA methylation and many disorders including cancer (Teschendorff *et al.*, 2009), cardiovascular (Baccarelli *et al.*, 2010) and metabolic diseases (Grundberg *et al.*, 2013) have been identified, indicating that DNA methylation may have a role in disease aetiology. Epigenetic mechanisms, including DNA methylation have previously been implicated in PD (Jowaed *et al.*, 2010; Lu *et al.*, 2013). Due to the tissue specific nature of DNA methylation patterns, to assess the mechanistic link between DNA methylation and disease, the diseased tissue (in this case the brain) would ideally be required. However, DNA methylation can also be employed as a biomarker to predict disease or aid prognosis. In this instance, the identification of a biomarker does not require a causal relationship to be identified, nor does it require the target disease tissue to be studied. Peripheral blood DNA provides a more readily accessible source of DNA for this application.

Epigenetic biomarkers, including global and gene-specific DNA methylation markers, are beginning to emerge as biomarkers of disease risk and prognosis, particularly within the field of cancer (Schwalbe *et al.*, 2013). This approach has yet to be applied to PD-MCI. It is postulated that DNA methylation signatures detectable in peripheral blood, drawn following diagnosis of PD, will be predictive of subsequent cognitive decline and possibly overt dementia. In order to pursue this hypothesis, a prospective study is required composed of individuals who have been diagnosed with idiopathic PD, had biological samples collected and stored at this time and that have been followed up over several years with cognitive assessments undertaken at multiple intervals. To test this hypothesis the prospectively collected ICICLE study is used (Section 2.1.2).

5.2 Experimental design

General methods are described in Chapter 2, with additional details relevant to this investigation described in detail here.

5.2.1 Study cohort

Blood DNA samples taken from participants of the ICICLE study (Section 2.1.2), who were newly diagnosed Parkinson’s disease patients in the Newcastle and Gateshead area, were utilised for this study. The ICICLE study recruited 158 PD patients along with 99 controls. Once a diagnosis of idiopathic PD had been confirmed by a specialist neurologist, participants underwent a medical assessment and information on their education level, medical history and lifestyle were recorded. A number of tests and scales were used in the assessment of disease severity including MMSE, MoCA, Hoehn and Yahr and the MDS-UPDRS. All cases and controls were invited to a follow-up at 18 months, where the medical assessments were repeated to check for any disease progression. A peripheral blood sample was collected from 150 PD cases and 90 controls at baseline. Samples were divided into a ‘discovery’ cohort (n=48 PD cases) and a ‘replication’ cohort (n=192; 102 PD cases and 90 controls), the latter consisting of all ICICLE samples not included in the discovery phase.

5.2.2 Blood DNA quantification

Blood DNA samples were held in long-term storage at -80°C following extraction and were selected and removed from storage for the purpose of this study. DNA samples were quantified using a ND1000 Spectrophotometer (Labtech International Ltd, UK) (Section 2.2.3).

5.2.3 Illumina HumanMethylation450 BeadChip analysis

Epigenome wide discovery analysis was performed using the Illumina HumanMethylation450 (HM450) BeadChip. 48 individuals with >1µg blood derived DNA were selected for discovery phase DNA methylation analysis. 24 samples were from PD-MCI participants and 24 from Parkinson’s disease participants without MCI. All samples selected for inclusion in the discovery phase were age and sex matched. The PD-MCI samples selected were those with the most substantially impaired cognition (at 2SD below normative levels). Sections 2.4 and 2.5 describe how samples were processed and data were filtered and normalised.

5.2.4 Analysis of HM450 data

A total of 455,981 probes and 46 PD samples were included in the final dataset. Exploratory analyses were performed across the methylation data by considering distribution density plots, mean-sd plots and hierarchical clustering with heat plots. The CpGassoc package (Barfield *et al*., 2012), which performs multiple linear regression analysis with a continuous predictor variable and ANOVA for categorical predictor variables, was implemented in R (version 2.15.0). CpGassoc was performed to test for associations between methylation and; cognition variables (MCI, animals, cognitive complaint, language, language total, MMSE, MoCA, naming, NMSQ concentration, NMSQ memory, one touch stockings, paired associates learning, pentagon copying, power of attention, pattern recognition memory, spatial recognition memory and total FAS) and motor variables (digit vigilance accuracy, Hoehn and Yahr, MDS-UPDRS Part II and III, PIGD phenotype and tremor dominant phenotype) as outcome variables. For all analyses, age, sex and chip (i.e. the microarray chip upon which the samples were scanned) were included in the analysis model as covariates. Data generated provided a beta value difference between the comparison groups (for categorical variables) and a beta value per unit change in the predictor variable (for ordinal and continuous variables). The beta values had an associated p-value denoting the strength of statistical significance with standard errors (SE) denoting the confidence around the statistical estimate of association. HM450 data were adjusted for cellular composition using the method described by Houseman *et al.* (2012) in a sensitivity analysis, i.e. the CpGassoc analyses were repeated to assess whether cell composition accounted for any significance observed between methylation and the outcome variable.

5.2.5 Selection of top HM450 hits

To select which CpG sites were the most promising candidates to validate and most likely to be functionally relevant, all CpG sites with a p value $<1x10^{-5}$ were compared against publicly available expression data. The GEO website (NCBI) (http://www.ncbi.nlm.nih.gov/geo/) was searched for an appropriate expression dataset with which to compare the methylation data generated in this project. The closest expression dataset found was GSE18838, a study which compared expression levels in blood between PD patients and healthy controls (Shehadeh *et al.*, 2010). Each of the methylation hits were categorised into either hypomethylated or hypermethylated in PD and hits were searched within the expression dataset. Any locus differentially expressed in PD was categorised into either over or under expressed in PD. Those hits which showed the expected methylation-expression relationship (i.e. hypermethylated and under expressed, hypomethylated and over expressed) were taken forward to the next step. The next criterion was that the effect size (difference between the outcome groups) was more than 5%. This criterion is based upon the technical limitations of validating any observed differences using Pyrosequencer (Mikeska *et al.*, 2011). Hence, differences of less than this may simply be due to random technical variation between groups. Finally, a BLAT search was performed to test for the presence of a SNP within the CpG site. If the CpG site was indeed a SNP, or if there was a SNP in the probe sequence, this site was removed from further consideration. This step removed any CpG sites likely to be directly under genetic influence.

The HM450 data were processed and analysed as described to generate a priority list of target differentially methylated positions (DMPs). The Illumina gene ID was used to assign a gene name to the probe; where this was not possible, the HM450 probe ID was retained. These DMPs were then taken forward for validation and replication using Pyrosequencing.

5.2.6 Pyrosequencing

Primers for amplicons covering DMPs were designed using the PSQ Assay Design software (Section 2.6.2) and synthesised by Integrated DNA Technologies (IDT, USA). Where possible, primers were designed not only to cover the target DMP, but also cover any neighbouring CpG sites that could be captured within a technically feasible amplicon size (50-250bp). Primers were first optimised (Section 2.6.3) and then validated on the Pyrosequencer using Epitect PCR controls (Qiagen, UK) (Section 2.6.5), prior to assays being run using ICICLE samples. PCR controls containing a 100% methylated DNA sample and a 0% methylated DNA sample were used to make reference methylation levels ranging from 0% to 100%. These mixes were created both pre-PCR and post-PCR. Thirteen of the twenty DMP primer sets were optimised successfully. All ICICLE samples were run in duplicate and repeated if duplicate discordance was >5%.

Pyrosequencing assays (n=3) were performed on the discovery sample (n=46) along with blood derived DNA samples from an additional 102 PD cases and 90 controls drawn from the same cohort. 500ng of DNA was bisulphite modified, amplified using the optimised Pyrosequencing PCR conditions in Table 39 and finally sequenced on a PyroMark MD Pyrosequencer (Qiagen, UK), following the standard protocol (Section 2.6.6). Details of all Pyrosequencing primers are provided in Table 40. The primer optimisation, assay validation and Pyrosequencing undertaken in this Chapter were performed by Polly Usher, a technician within the Epigenetics Research Group at Newcastle University.

Gene/DMP	HM450 probe ID	Tm (°C)	$MgCl_2$ +/-	Size of product (bp)	No. CpGs	Successfully validated using Epitect controls
CHCHD5	cg21934564	49.8	+	118	1	Yes
TRIM15	cg09769113	52.4	-	338	2	Yes
CUTA	cg09977865	52.4	+	247	1	
CDK14	cg23251279	54.4	+	214	2	
BDH1	cg10728473	49.8	+	134	1	
ARHGEF7	cg14666946	50.0	+	100	3	
ALDHA3A1	cg18957070	49.8	+	139	2	
GSX1	cg25778304	51.1	-	243	4	
PROCA1	cg02206852	49.8	-	310	1	Yes
GPR17	cg20074593	51.1	+	204	1	
FTCD	cg10394047	52.4	+	99	5	
SAT2B	cg26236655	52.4	+	167	1	
NFU1	cg10354495	52.4	+	145	1	

Table 39: Optimal PCR conditions for each assay. Tm=annealing temperature.

Gene	Forward primer (5'>3')	Reverse primer (5'>3')	Sequencing primer (5'>3')
CHCHD5	AATAGGGATGAAGTAATT	5Biosg/CCCTTTAAAACTAATCTAT	GGGTAGGAAGAAGAGGTAG
TRIM15	GTTGGGAGAAATTTATTG	5Biosg/CCAAAAACTACCACACAT	GAAATTTATTGAGAGGAGTA
PROCA1	GGAGGGAGAGAATAGTAGTA	5Biosg/AACAACTAACCAAAATATC	AGTAGAGGGTTAGTTATAGG

Table 40: Forward, reverse and sequencing primers for each successfully validated DMP Pyrosequencing assay. 5Biosg=biotin label.

5.2.7 Statistical analyses

Summary statistics of sub-samples used from the ICICLE cohort were calculated using chi squared tests, t-tests or Wilcoxon rank sum tests depending on whether data followed a normal distribution. Within each cohort, differences between MCI and NCI participants were investigated. Following correction for multiple testing, a p value $<4.27\text{x}10^{-4}$ was considered significant. Tests were also performed to look for differences between the discovery cohort and the entire ICICLE cohort, as well as the replication cohort and the entire ICICLE cohort. This was to investigate whether the sample subsets were representative of the entire ICICLE cohort. To reduce the risk of false positives, the p value was adjusted for the number of independent tests performed and a p value $<3.21\text{x}10^{-4}$ was considered significant.

The association between methylation and the outcome variable of interest was first assessed in the initial HM450 discovery sample set (n=46) using ANOVA analyses, adjusted for baseline age and sex (Pyrosequenced validation samples). Subsequently, association analyses were repeated including data from the additional samples (Pyrosequenced replication samples). Given the non-normally distributed nature of methylation data, sensitivity analyses were also performed by comparing mean methylation levels between sample groups using non-parametric Wilcoxon rank-sum or Kruskal-Wallis tests. In addition to assessing individual DMPs, for loci in which multiple CpG sites were Pyrosequenced, regional methylation levels were assessed by taking the average methylation value across all CpG sites. These validation analyses were performed in Stata.

To check whether there was any correlation between each CpG site measured by the HM450 array and Pyrosequencing, Spearman’s rank was performed in Stata.

5.3 Results

Summary statistics for clinical variables of ICICLE participants and sub-groups utilised in this study are summarised in Table 41, Table 42 and Table 43.

The MCI group were significantly older (75.63 vs 65.87 years, p=5.43E-06), had a lower score on the NART (110.08 vs 117.72, p=3.35E-06) and had a lower number of years in education (11.57 vs 13.40 years, p=0.001) than the no cognitive impairment (NCI) group. The fact that the MCI group had a lower number of years in education and a lower NART score is not surprising as you would expect those who were less academically capable during childhood to also be less academically capable during adulthood and older age. In addition, those with MCI were more likely to have hypertension (44.90% vs 24.72%, p=0.015) and had higher depression scores (3.67 vs 2.57, p=0.026) but had a lower weekly alcohol intake (5.04 vs 9.78 units/week, p=0.012) (Table 41). All cognitive outcomes were distributed in the expected fashion with the MCI group displaying more impaired cognition in all tests (Table 42). Interestingly, the MCI group also displayed worse motor function in all motor outcome variables tested (Table 43).

Both the discovery and replication (Pyrosequencing) cohorts were highly representative of the entire ICICLE cohort, with no significant differences being observed between sample groups for any variable.

	Entire ICICLE			Discovery Cohort			Replication Cohort		
Variables	MCI	NCI	P value	MCI	NCI	P value	MCI	NCI	P value
Age (Yrs) mean(SD)	73.64 (7.93)	65.87 (9.83)	5.43E-06§	70.39 (7.55)	67.45 (5.53)	0.078‡	75.88 (7.51)	65.18 (11.17)	1.63E-05§
Sex (% males)	32.65	33.71	0.9†	40.00	37.03	0.836†	32.26	27.59	0.653†
Education (Yrs) mean(SD)	11.57 (3.37)	13.40 (3.65)	0.001‡	12.65 (3.60)	14.93 (4.50)	0.036‡	10.83 (3.05)	12.73 (3.01)	0.001‡
NART mean(SD)	110.08 (10.37)	117.72 (9.73)	3.35E-06‡	111.68 (10.12)	118.63 (11.76)	0.001‡	109.03 (10.57)	117.32 (8.78)	0.0003‡
Height (m) mean(SD)	1.69 (0.10)	1.71 (0.09)	0.605§	1.69 (0.10)	1.68 (0.08)	0.786§	1.70 (0.09)	1.72 (0.09)	0.373§
Weight (kg) mean(SD)	76.74 (16.21)	79.40 (14.62)	0.327§	77.89 (16.76)	79.32 (12.39)	0.742§	75.94 (16.07)	79.44 (15.56)	0.326§
BMI mean(SD)	26.66 (4.75)	27.26 (4.45)	0.466§	27.10 (4.47)	28.00 (3.75)	0.464§	26.36 (4.99)	26.95 (4.71)	0.588§
IHD n(%)	9 (18.37)	7 (7.87)	0.065†	1 (5.00)	3 (11.11)	0.426†	8 (27.59)	4 (6.45)	0.016†
DM n(%)	6 (12.24)	5 (5.62)	0.197†	2 (10.00)	4 (14.81)	0.488†	4 (13.79)	1 (1.61)	0.034†
HT n(%)	22 (44.90)	22 (24.72)	0.015†	8 (40.00)	7 (25.93)	0.306†	14 (48.28)	15 (24.19)	0.022†
Hyperchol n(%)	8 (16.33)	10 (11.24)	0.395†	2 (10.00)	5 (18.52)	0.352†	6 (20.69)	5 (8.06)	0.098†
Levodopa daily dose (mg) n(SD)	400.2 (222.1)	393.1 (202.2)	0.850§	359.65 (223.46)	362.93 (187.76)	0.958§	428.1 (220.7)	406.2 (208.3)	0.648§
GDS mean (SD)	3.67 (3.17)	2.57 (2.73)	0.026‡	3.1 (3.01)	2.04 (2.03)	0.207‡	4.07 (3.27)	2.79 (2.97)	0.048‡
RCF (µg/L) mean(SD)	278.63 (89.48)	225.92 (107.89)	0.228‡	295.36 (202.72)	269.25 (151.59)	0.549‡	288.1 (162.4)	265.1 (138.9)	0.197‡
B12 (ng/L) mean(SD)	365.70 (126.72)	394.73 (139.91)	0.400‡	379.12 (138.44)	425.17 (174.76)	0.326‡	373.16 (132.04)	415.75 (164.53)	0.180‡
Homocysteine (µmol/L) mean(SD)	13.10 (4.99)	12.33 (3.63)	0.458‡	13.80 (4.36)	11.43 (3.14)	0.015‡	13.50 (4.60)	11.71 (3.30)	0.014‡
Alcohol (units/wk) mean(SD)	5.04 (7.92)	9.78 (12.02)	0.012‡	5.70 (8.86)	12.69 (14.25)	0.063‡	4.59 (7.33)	8.56 (10.86)	0.062‡
Smoking n(%)			0.104†			0.874†			0.110†
Current	1 (2.04)	8 (9.09)		1 (4.76)	1 (4.35)		1 (3.33)	5 (8.20)	
Ex	22 (44.9)	27 (30.68)		9 (42.86)	8 (34.78)		13(43.33)	18 (29.51)	
Never	26 (53.06)	53 (60.23)		11 (52.38)	14 (60.87)		16 (53.33)	38 (62.30)	

Table 41: Summary statistics for exposure variables. Data are presented for the entire ICICLE cohort, the initial discovery sample set and the replication cohort. † = chi squared, ‡ = Wilcoxon rank sum, §= t test. Hyperchol=hypercholesterolaemia.

	Entire ICICLE			Discovery Cohort			Replication Cohort		
Variables	MCI	NCI	P value	MCI	NCI	P value	MCI	NCI	P value
MoCA	23.00 (4.01)	27.25 (2.76)	5.39E-11‡	23.00 (4.24)	27.11 (3.34)	0.0003‡	23.00 (3.92)	27.31 (2.50)	4.91E-08‡
MMSE	26.98 (2.10)	29.02 (1.21)	7.48E-10‡	26.95 (2.28)	29.04 (0.94)	0.0005§	27.00 (2.00)	29.02 (1.31)	5.45E-07‡
FAS	27.88 (8.80)	40.10 (13.25)	1.85E-09§	28.50 (9.28)	45.41 (15.88)	3.99E-05§	27.45 (8.60)	37.75 (11.27)	3.59E-05§
Animals	16.88 (6.23)	23.75 (6.65)	2.42E-08§	18.05 (6.04)	25.81 (6.34)	0.0001§	16.07 (6.33)	22.84 (6.63)	1.46E-05§
NMSQ memory n(%)	31 (63.27)	45 (50.56)	0.151†	13 (65.00)	11 (40.74)	0.100†	18 (62.07)	34 (54.84)	0.516†
NMSQ concentration n(%)	15 (30.61)	26 (29.21)	0.863†	6 (30.00)	5 (18.52)	0.358†	9 (31.03)	21 (33.87)	0.789†
Total number of NMS	9.35 (4.20)	8.55 (4.25)	0.292‡	9.55 (4.96)	7.89 (4.20)	0.221§	9.21 (3.68)	8.84 (4.27)	0.690‡
Power of Attention	1610.4 (256.0)	1317.7 (124.7)	4.45E-10‡	1621.0 (274.9)	1295.3 (96.6)	2.13E-06§	1603.5 (247.5)	1327.9 (134.9)	2.27E-06‡
Digit vigilance accuracy	79.86 (17.88)	95.97 (6.04)	2.15E-12‡	77.08 (20.89)	96.21 (6.49)	6.95E-06‡	81.69 (15.73)	95.85 (5.88)	1.12E-07‡
PRM	17.52 (2.96)	21.03 (2.09)	1.52E-10‡	17.95 (3.03)	21.40 (2.12)	0.0001§	17.21 (2.92)	20.89 (2.07)	4.92E-07‡
PAL	3.12 (1.16)	1.92 (0.62)	1.53E-10‡	3.05 (1.17)	1.94 (0.68)	0.001‡	3.17 (1.16)	1.91 (0.59)	8.87E-08‡
SRM	12.52 (2.81)	15.65 (2.09)	2.65E-09§	12.25 (2.95)	15.84 (2.17)	2.71E-05§	12.71 (2.75)	15.57 (2.07)	4.75E-07§
OTS	10.42 (5.37)	16.05 (3.08)	1.40E-11‡	9.75 (5.91)	16.32 (2.17)	0.0001§	10.89 (5.01)	15.93 (3.39)	4.57E-07‡
Pentagon copying	1.37 (0.78)	1.93 (0.25)	4.87E-08‡	1.35 (0.81)	1.96 (0.19)	0.001‡	1.38 (0.78)	1.92 (0.27)	2.52E-05‡
Naming	2.88 (0.33)	2.92 (0.27)	0.401‡	2.95 (0.22)	2.93 (0.27)	0.741‡	2.83 (0.38)	2.92 (0.27)	0.195‡
Language	1.33 (0.77)	1.74 (0.53)	0.001‡	1.50 (0.76)	1.70 (0.61)	0.317‡	1.21 (0.77)	1.76 (0.50)	0.0002‡
Language total	4.20 (0.84)	4.66 (0.66)	0.0003‡	4.45 (0.76)	4.63 (0.74)	0.307‡	4.03 (0.87)	4.68 (0.62)	0.0001‡
Cognitive complaint n(%)	34 (69.39)	52 (58.43)	0.204†	14 (70.00)	13 (48.15)	0.134†	20 (68.97)	39 (62.90)	0.573†

Table 42: Summary statistics for variables relating to cognitive-based outcomes. Data are presented for the entire ICICLE cohort, the initial discovery sample set and the replication cohort. † = chi squared, ‡ = Wilcoxon rank sum, §= t test. Unless stated, mean (SD) are presented.

	Entire ICICLE			Discovery Cohort			Replication Cohort		
Variables	MCI	NCI	P value	MCI	NCI	P value	MCI	NCI	P value
Hoehn and Yahr n(%)			0.002†			0.214†			0.021†
1	0	8 (8.99)		0	3 (11.11)		0	5 (8.06)	
2	34 (69.39)	69 (77.53)		14 (70.00)	20 (74.07)		20 (68.97)	49 (79.03)	
3	11 (22.45)	12 (13.48)		5 (25.00)	4 (14.81)		6 (20.69)	8 (12.90)	
4	4 (8.16)	0		1 (5.00)	0		3 (10.34)	0	
MDS-UPDRS II mean(SD)	15.02 (5.88)	11.18 (5.96)	0.0004§	13.7 (5.18)	8.44 (5.54)	0.002‡	15.93 (6.24)	12.37 (5.79)	0.009§
MDS-UPDRS III mean(SD)	41.67 (12.37)	30.54 (9.88)	4.82E-08§	41.00 (11.10)	28.63 (10.84)	0.0004§	42.14 (13.35)	31.37 (9.40)	0.0003§
Tremor phenotype mean(SD)	0.90 (0.48)	0.69 (0.37)	0.008§	0.91 (0.54)	0.68 (0.39)	0.089§	0.90 (0.44)	0.69 (0.36)	0.021§
PIGD phenotype mean(SD)	1.04 (0.63)	0.63 (0.44)	0.0001‡	0.96 (0.58)	0.58 (0.45)	0.007‡	1.09 (0.66)	0.65 (0.43)	0.002‡

Table 43: Summary statistics for variables relating to motor-based outcomes. Data are presented for the entire ICICLE cohort, the initial discovery sample set and the replication cohort. † = chi squared, ‡ = Wilcoxon rank sum, §= t test.

5.3.1 Discovery analysis

Of the 48 blood DNA samples that underwent HM450 analysis, 46 samples passed the QC assessment of the inbuilt microarray controls assessed in GenomeStudio. Two samples failed due to incomplete bisulphite modification and were dropped from further analysis.

Low failure rates were observed on both a sample and a probe by probe basis, assessed using detection p values generated from the array scanner. As described in the Methods (Section 2.5.2), samples were only removed if ≥20% of probes had a detection p value >0.01. Probes were removed if ≥10% of samples had a detection p value >0.01. No samples were removed from analysis using these criteria. 857 probes were removed. Summaries of the detection p values observed for samples and probes are shown in Figure 79A and B.

Following data filtering and normalisation, as described in Section 2.5.2, 455,981 CpG sites were taken forward for analysis measured in baseline blood samples from 46 PD individuals; 23 PD-MCI and 23 PD-NCI.

The small sample size of this study limited the power to detect methylation differences. This study had 29.8% power to detect a mean difference of 5% methylation between groups, with a standard deviation of 3 and alpha=$1.1\text{x}10^{-7}$.

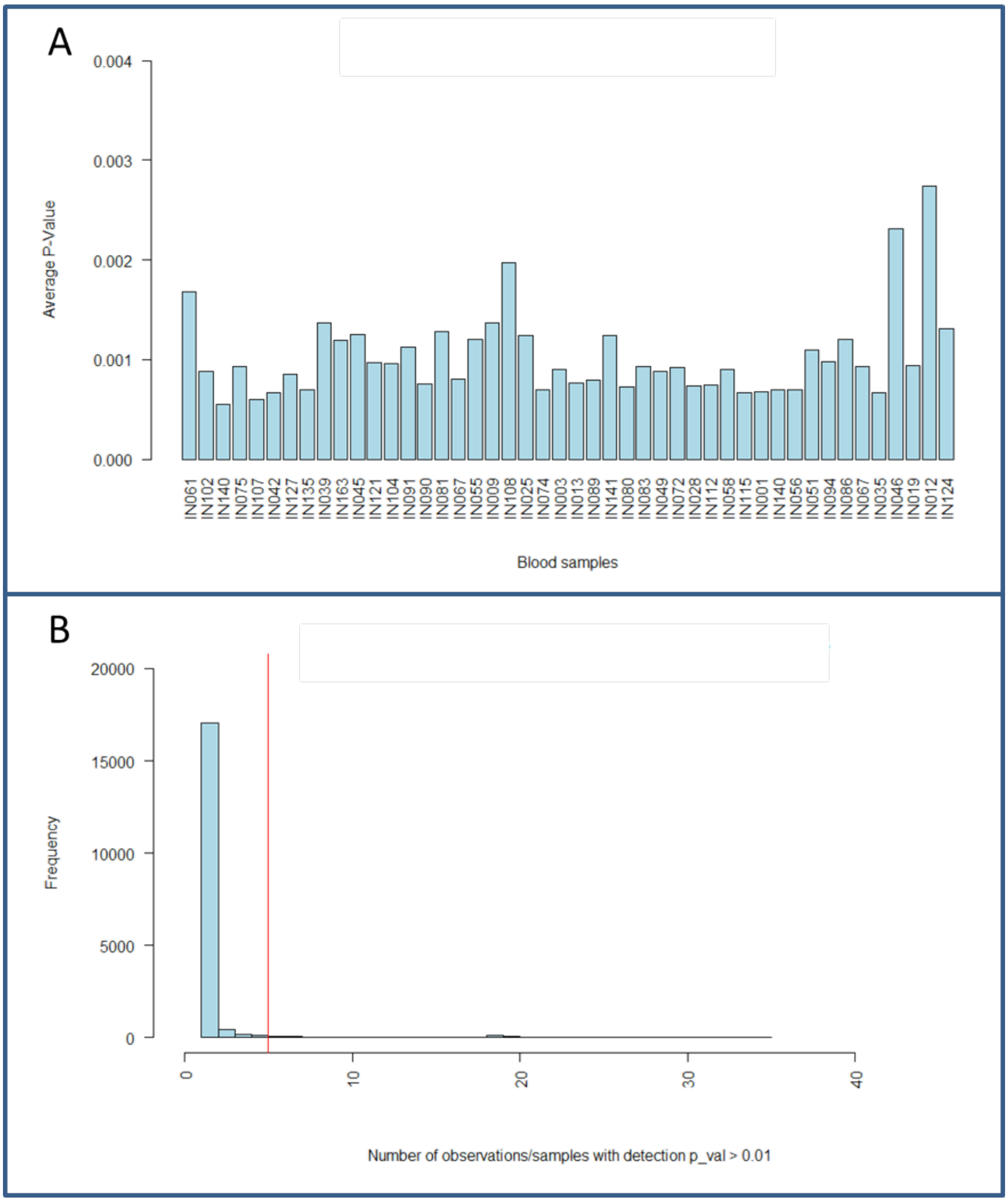


Figure 79A: Average detection p values in each blood sample. For each sample the average detection p value is very low and samples are of approximately equal quality.
Figure 79B: Histogram to show the number of probes against the corresponding number of samples with p values greater than 0.01 across blood. The red line indicates 10%. 467052 probes had no missing data and for space reasons, this data is not shown on the histogram.

5.3.1.1 ***Distribution of genome-wide methylation values***

Figure 80 (A-F) shows the distribution of methylation values using both raw (A-C) and normalised (D-F) data. Figure 80A is a density plot showing the distribution of raw methylation beta values for each blood sample. This shows a commonly observed bimodal distribution; the majority of CpG sites are very lowly or very highly methylated with much fewer CpG sites exhibiting intermediate methylation levels. The distribution for each individual can be more easily seen in a box plot (Figure 80B). This shows that there is some variability in the methylation levels of each blood sample, particularly with regard to the median methylation and IQR. Each sample has a minimum methylation level of 0% and a maximum of 100%, with a median around 60-70% methylated. Some of this variation could be due to batch effects and sample processing. The position of the sample on the chip is also thought to have an effect on the measurement of methylation values (Dedeurwaerder *et al.*, 2013), but this variation should be removed after normalisation. Mean and SD methylation beta values across all probes, prior to normalisation and QC, are depicted in Figure 80C to show heteroskedasticity. Heteroskedasticity refers to the phenomenon whereby the variance in a measure, such as methylation beta values, is not constant across all levels of that measure. As can be seen, there is a wide range of variance across the probes. In particular, probes with a mid-range methylation value between 30-80% show greater variability compared to those at the extreme ends of the methylation spectrum.

Figure 80 (D-F) shows the distribution of methylation after normalisation. All plots using normalised data show that there is much less variation between samples. All samples have a median methylation value of around 50%. This suggests that most of the technical variation was removed during normalisation.

5.3.1.2 ***Cluster analysis***

Hierarchical cluster analysis, using the complete-linkage method, was performed on the normalised data to look for natural groupings (either phenotypic or technical) amongst the samples based on the methylation data. No clear clusters were observed for diagnosis of MCI indicating that DNA methylation signatures are not sufficiently different to distinguish between groups in the blood DNA drawn from this cohort (Figure 81).

However, as shown in Figure 82, even after normalisation and the removal of probes mapping to the sex chromosomes, there remained some subtle groupings amongst samples based on sex. Due to these results, sex was included as a covariate in subsequent analyses of single-point CpG sites. In contrast, there were no apparent clusters based on experimental chip (Figure 83), suggesting that any chip-derived batch effects were removed. Nonetheless, to fully account for any remaining, albeit subtle, batch effects, chip was also included as a covariate in the single-point analysis model.

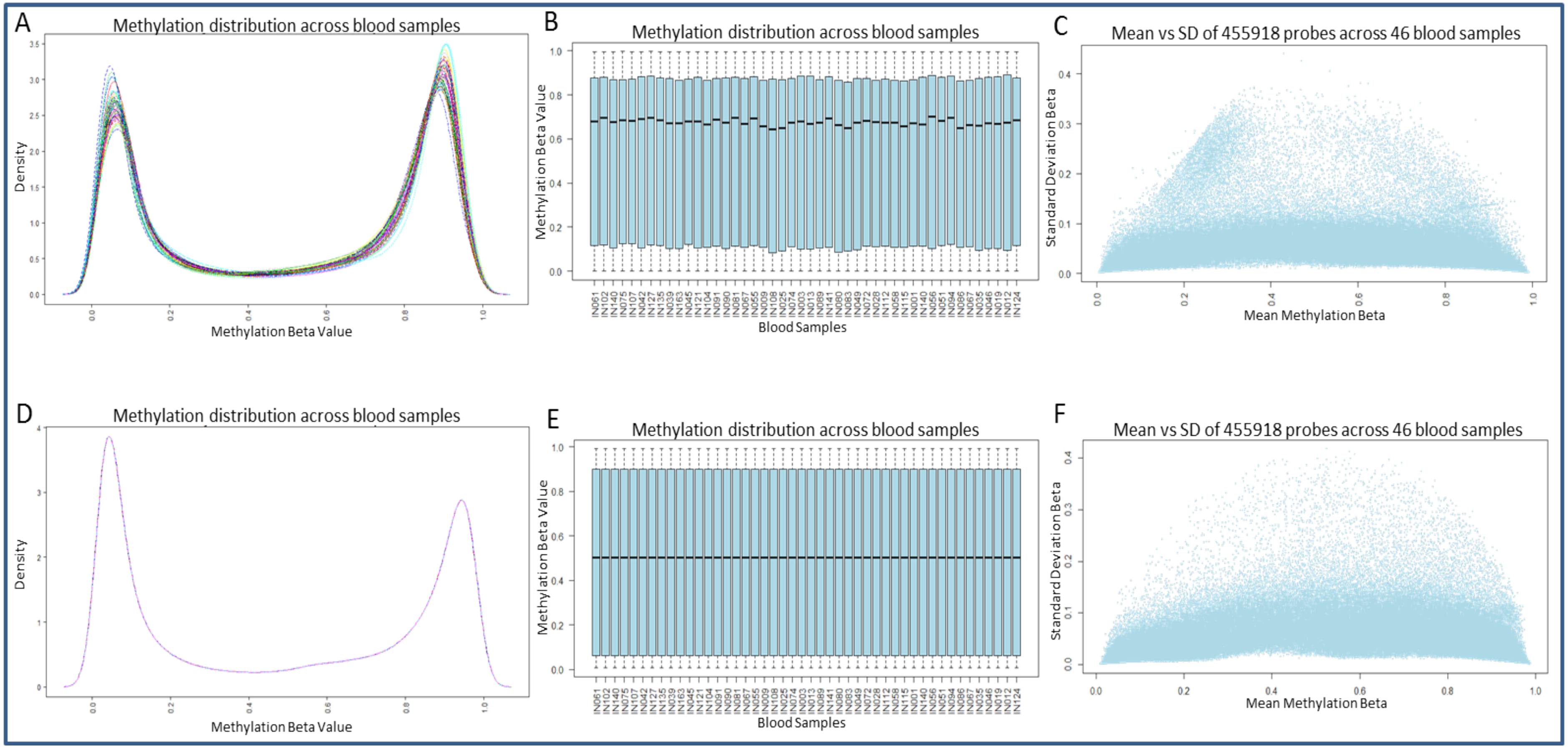


Figure 80: A composite figure showing the distributions of methylation using raw and normalised data. A. Density plot showing methylation distribution in 46 blood samples using raw data. B. Box plot showing mean methylation raw beta values across all blood samples. C. Mean vs SD plot showing heteroskedasticity using raw data. D. Density plot showing methylation distribution in 46 blood samples using normalised data. E. Box plot showing mean methylation normalised beta values across all blood samples. F. Mean vs SD plot showing heteroskedasticity using normalised data.

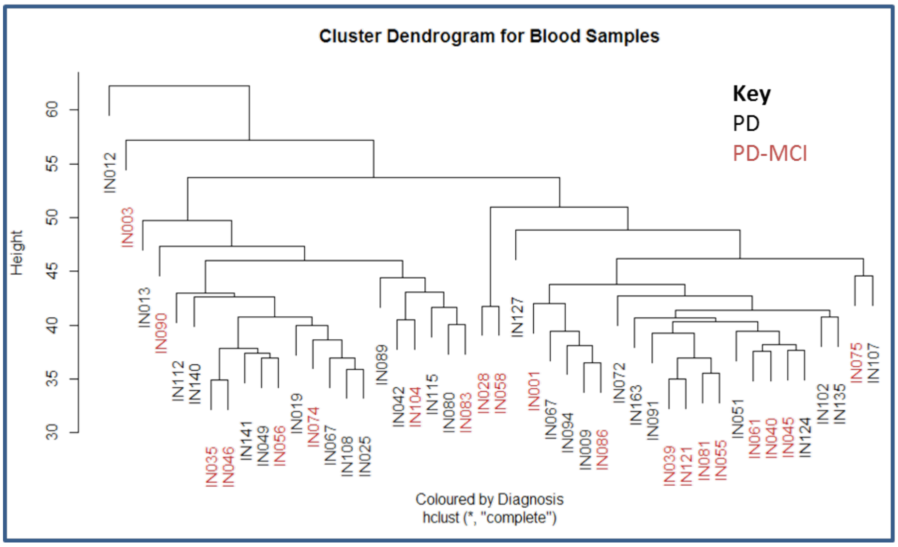


Figure 81: Cluster dendrogram for all 46 blood samples coloured by diagnosis.

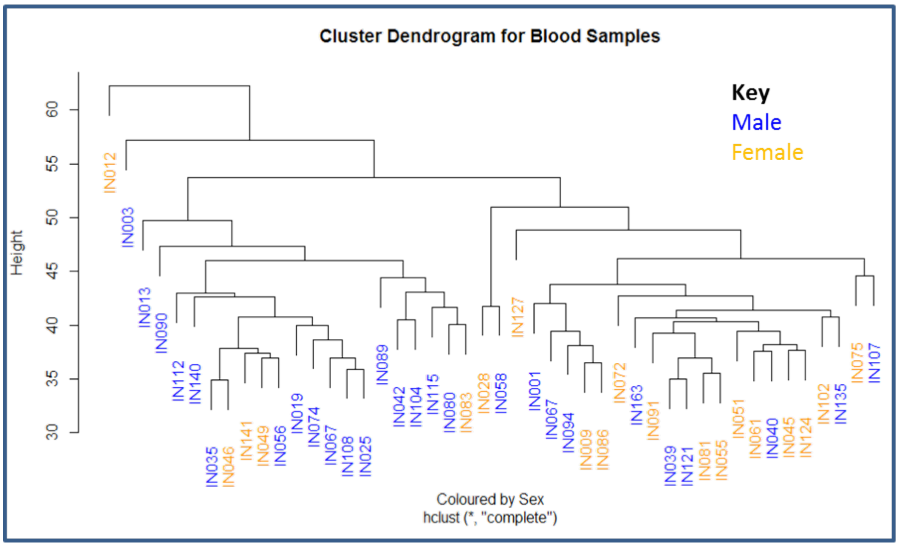


Figure 82: Cluster dendrogram for all 46 blood samples coloured by sex.

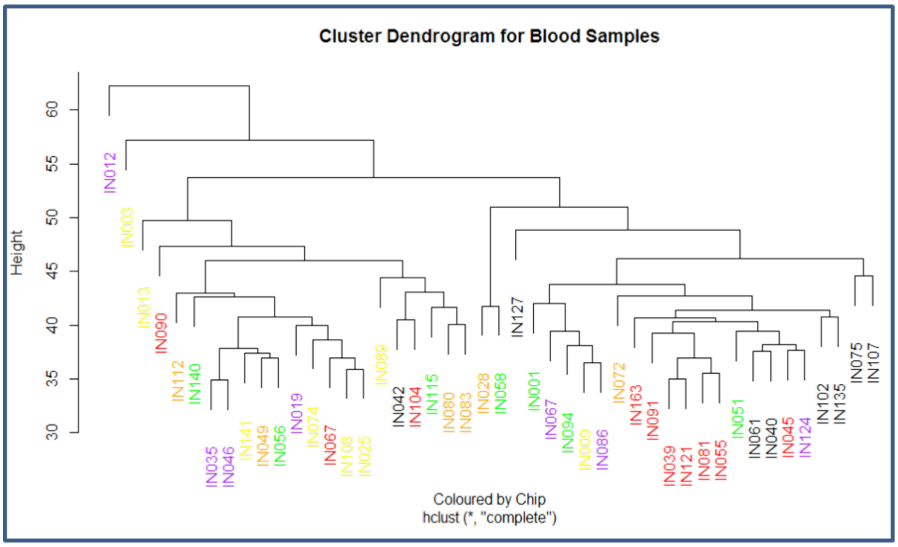


Figure 83: Cluster dendrogram for all 46 blood samples coloured by chip.

5.3.1.3 *Single-point analysis*

ANOVA was performed on all beta values across all CpG sites passing QC (455,981) against the following outcome measures: MCI, animals, cognitive complaint, digit vigilance accuracy, language, language total, MMSE, MoCA, naming, NMSQ concentration, NMSQ memory, one touch stockings, paired associates learning, pentagon copying, power of attention, pattern recognition memory, spatial recognition memory, total FAS, Hoehn and Yahr, MDS-UPDRS Part II and III, PIGD phenotype and tremor dominant phenotype. Age, sex and chip were included in the analysis model as covariates. Results from this step were exported as a table in Microsoft Excel and ordered by ascending p value.

5.3.1.4 *Selection of top hits*

To select the most functionally relevant potential biomarkers, top hits were selected based on p value ($p<1x10^{-5}$) and compared to results of publicly available expression data. The expression data set used is described in Section 5.2.5.

Table 44 shows the number of hits associated with cognition that were considered for Pyrosequencing validation. The number of hits reaching the significance threshold ($p<1x10^{-5}$) across all outcomes was 548. Of these 548 hits, 466 had expression data available and 102 of these were differentially expressed in PD. The 102 hits were then considered in terms of the direction of effect, i.e. whether the CpG sites were hyper or hypo methylated and up or down regulated in PD. 46 CpG sites were found to have the expected methylation-expression relation (hypermethylated and downregulated in PD, hypomethylated and upregulated in PD).

Table 45 shows the number of hits associated with motor outcome that were selected for Pyrosequencing consideration. The number of hits reaching the significance threshold ($p<1x10^{-5}$) across all outcomes was 25. Of these 25 hits, 19 had expression data available and six of these were differentially expressed in PD. The six remaining hits were then considered in terms of the direction of effect, i.e. whether the CpG sites were hyper or hypo methylated and up or down regulated in PD. Three CpG sites were found to have the expected methylation-expression relationship (hypermethylated and downregulated in PD, hypomethylated and upregulated in PD).

Variable	Hits significant to <1x10^{-5} (n)	Hits also on the expression array (n)	Hits differentially expressed (n)	Hits with the expected methylation-expression relationship (n)
MCI	2	1	0	0
Animals	5	4	0	0
Cognitive complaint	9	7	2	2
Language	8	6	1	1
MMSE	39	36	7	4
MoCA	6	5	2	1
Naming	287	251	52	18
NMSQ concentration	13	9	3	2
NMSQ memory	5	5	1	1
OTS	28	24	9	6
PAL	7	7	3	2
Pentagon copying	61	50	10	6
Power of attention	38	31	5	3
PRM	8	4	2	0
SRM	2	2	1	0
Total FAS	3	2	0	0
Digit vigilance accuracy	27	22	4	0
Total hits (n)	**548**	**466**	**102**	**46**

Table 44: Number of CpG sites significantly associated with cognition taken forward at each stage in the selection of top hits.

Variable	Hits significant to <1x10^{-5} (n)	Hits also on the expression array (n)	Hits differentially expressed (n)	Hits with the expected methylation-expression relationship (n)
MDS-UPDRS II	2	1	0	0
MDS-UPDRS III	6	4	2	1
PIGD phenotype	4	4	0	0
Tremor dominant phenotype	13	10	4	2
Total number of hits	**25**	**19**	**6**	**3**

Table 45: Number of CpG sites significantly associated with motor outcome taken forward at each stage in the selection of top hits.

Cognitive Outcome

Figure 84 shows the relationship between methylation and expression in each of the statistically significant CpG sites associated with cognition. Sites (n=46) within the shaded regions showed the expected methylation-expression relationship and at the significance threshold imposed. 26 loci were identified as being hypomethylated in PD and upregulated in PD. 20 loci were identified as having the reverse relationship with hypermethylation of a site within the gene in PD and down regulation of the gene in PD. These 46 sites were taken forward for further investigation (Table 46).

Fifteen CpG sites had an effect size <5% so were removed from further analysis. Two of the remaining sites were in fact SNPs and were also removed from further analysis. This left 29 loci which were inputted into the Pyrosequencing assay design software. Primers were successfully designed for 18 of these loci (cg18957070 features twice on this table, significant with both OTS and pentagon copying). Adjusting for cellular composition (Houseman *et al.*, 2012) had no effect on the vast majority of CpG sites included in the selection in Table 46. Methylation at two CpG sites associated with naming were shown to have a very small effect when adjusted for cellular composition, as shown in Table 47. In both instances, adjusting for cellular composition slightly increased the significance. These results suggest that the methylation differences observed between outcome groups are largely unrelated to cellular composition.

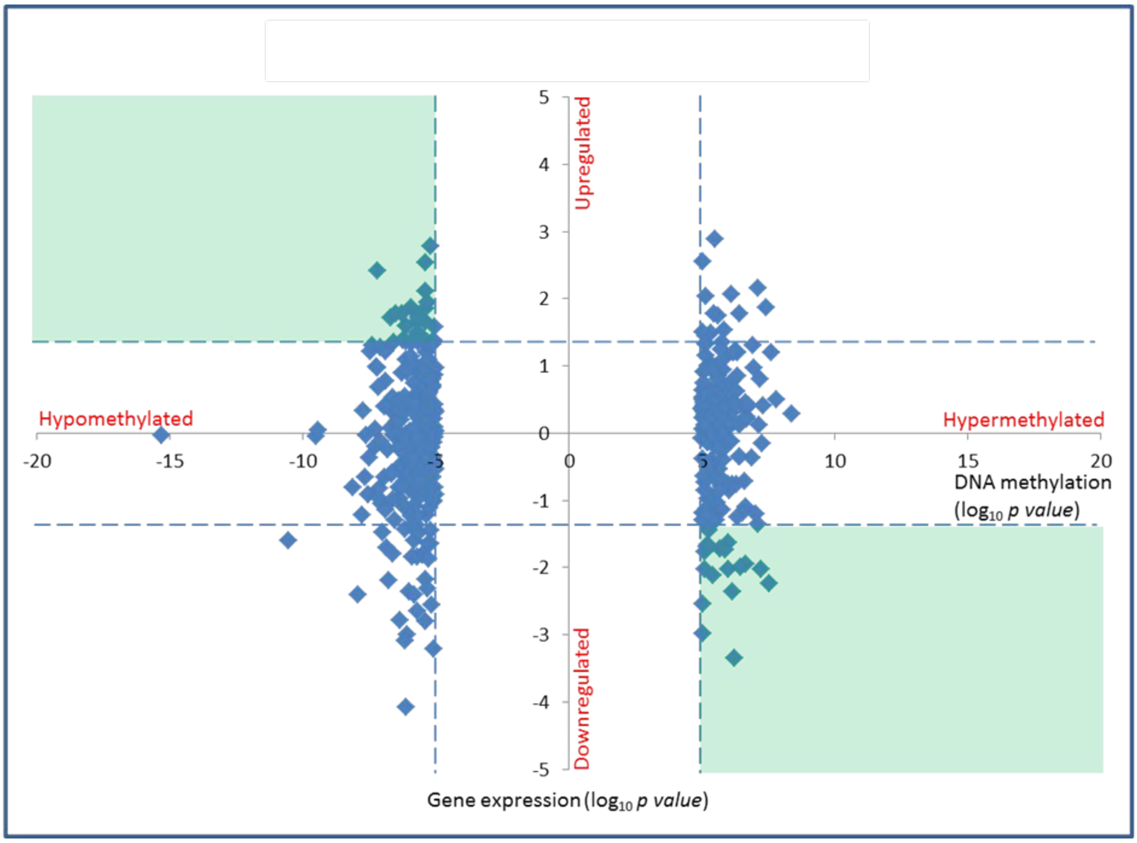


Figure 84: Quadrant plot of DMPs and associated expression levels for cognition outcomes. DNA methylation (x axis) and gene expression (y axis) are graphically presented as $\log_{10}$ p value. Significance is defined as p< 1x10^{-5} for methylation and p<1.13 for expression (as indicated by the vertical and horizontal dashed lines). The direction of effect is indicated by the red text. The shaded areas represent sites with the expected inverse relationship between methylation and expression. Sites within these shaded areas were taken forward for further analysis.

Variable	CpG site	Gene	Meth p value	Meth effect size	Exp p value	SNP	Primers
Naming	cg07719523	*PALM*	2.61E-07	-0.202	0.044	No	No
OTS	cg19350270	*DPP4*	2.90E-06	0.186	0.049	No	No
PoA	cg26236655	*SATB2*	7.44E-06	0.172	0.047	No	Yes
Pentagon copying	cg19026233	*UNC80*	5.73E-07	-0.149	0.041	No	No
MoCA	cg19350270	*DPP4*	9.69E-06	0.136	0.049	No	No
Naming	cg10728473	*BDH1*	9.78E-06	0.128	0.003	No	Yes
Naming	cg23251279	*CDK14*	7.64E-07	-0.103	0.042	No	Yes
Pentagon copying	cg10394047	*FTCD*	3.97E-06	-0.086	0.014	No	Yes
Naming	cg23519637	*C7orf50*	4.08E-06	-0.077	0.003	No	Yes
Naming	cg16252933	*CCDC6*	9.09E-06	0.074	0.001	No	No
Naming	cg01974375	*PI4KB*	7.09E-07	0.062	0.004	No	No
Naming	cg03162143	*PCGF3*	2.32E-06	-0.060	0.034	No	Yes
NMSQ conc	cg05022061	*CARD14*	5.19E-06	0.060	0.046	No	No
Naming	cg09769113	*TRIM15*	7.71E-08	-0.058	0.049	No	Yes
Naming	cg26915799	*IPO9*	2.36E-07	0.057	0.011	No	No
NMSQ conc	cg14666946	*ARHGEF7*	6.15E-06	-0.057	0.022	No	Yes
PoA	cg06351643	*FGF3*	2.22E-06	0.056	0.045	No	No
MMSE	cg10354495	*NFU1*	9.18E-06	-0.055	0.027	No	Yes
Naming	cg09977865	*CUTA*	7.73E-08	0.055	0.044	No	Yes
PAL	cg03768106	*TIMM8B*	5.41E-06	-0.054	0.021	No	Yes
Pentagon copying	cg20074593	*GPR17*	1.20E-06	-0.054	0.047	No	Yes
Pentagon copying	cg02206852	*PROCA1*	6.57E-08	-0.052	0.004	No	Yes
OTS	cg18957070	*ALDH3A1*	2.09E-06	-0.051	0.019	No	Yes
Language	cg03936721	*RUNDC3A*	7.91E-06	0.051	0.017	No	Yes
Naming	cg21934564	*CHCHD5*	2.86E-08	0.051	0.006	No	Yes
MMSE	cg07667522	*INPPL1*	2.15E-06	-0.050	0.019	No	No
MMSE	cg16514838	*C16orf45*	2.12E-06	0.050	0.021	No	Yes
Pentagon copying	cg18957070	*ALDH3A1*	2.31E-06	-0.050	0.019	No	Yes
PAL	cg25778304	*GSX1*	3.96E-06	-0.050	0.008	No	Yes
Naming	cg25246894	*SDS*	6.10E-06	-0.232	0.002	Yes	
Naming	cg16357930	*PLXNB1*	1.93E-06	-0.073	0.036	Yes	
Naming	cg04318855	*DOCK10*	1.12E-07	0.049	0.009		
OTS	cg02208529	*DBN1*	7.50E-07	-0.048	0.025		
Cognitive complaint	cg05410587	*KIAA1530*	5.28E-07	0.044	0.017		
OTS	cg03902642	*ROBLD3*	1.34E-06	-0.041	0.041		
Naming	cg26418433	*VPS52*	3.69E-07	0.041	0.010		
NMSQ memory	cg05410587	*KIAA1530*	3.12E-07	0.038	0.017		
Naming	cg15322876	*KRT5*	2.04E-06	-0.034	0.032		
Naming	cg20249462	*AP1B1*	6.45E-06	0.029	0.020		
OTS	cg16687450	*CHEK1*	8.58E-06	-0.023	0.041		
Naming	cg19925791	*RPS6KA1*	1.07E-06	0.021	0.023		
MMSE	cg23056047	*PFN4*	4.09E-06	0.017	0.008		
PoA	cg05918679	*RPL18*	5.97E-07	-0.011	0.0004		
OTS	cg22734480	*ABHD8*	9.80E-07	0.009	0.009		

Pentagon copying	cg02517488	*RP1L1*	4.89E-06	-0.003	0.012		
Cognitive complaint	cg09213106	*DHX35*	7.99E-06	-0.002	0.009		

Table 46: Description of 46 CpG sites associated with cognitive traits showing expression differences in PD, ranked in order of effect size. Meth =methylation, Exp=expression, OTS=one touch stockings, PoA=power of attention, PAL=paired associates learning.

Variable	**CpG site**	**Pre-adjustment**		**Post-adjustment**	
		t statistic	p value	t statistic	p value
Naming	cg09769113	6.76	7.71E-08	7.08	3.02E-08
Naming	cg04318855	-6.64	1.12E-07	-6.82	6.63E-08

Table 47: CpG sites affected by cellular composition adjustment.

Motor Outcome

Figure 85 shows the relationship between methylation and expression in each of the statistically significant CpG sites with motor outcome. Sites (n=3) within the shaded regions showed the expected methylation-expression relationship and at the significance threshold imposed. Two loci were identified as being hypomethylated in PD and upregulated in PD. One locus was identified as having the reverse relationship with hypermethylation of a site within the gene in PD and down regulation of the gene in PD. These three sites were taken forward for further investigation (Table 48).

All three CpG sites related to a motor variable had an effect size >5% and none of the CpG sites were found to be SNPs. Therefore all loci were taken forwards for Pyrosequencing assay design. Primers were successfully designed for two of these loci. When adjusted for cellular composition using the method outlined in Houseman *et al.* (2012), the significance level of each of these three CpG sites was not attenuated, suggesting that cellular composition was not responsible for these differences in methylation.

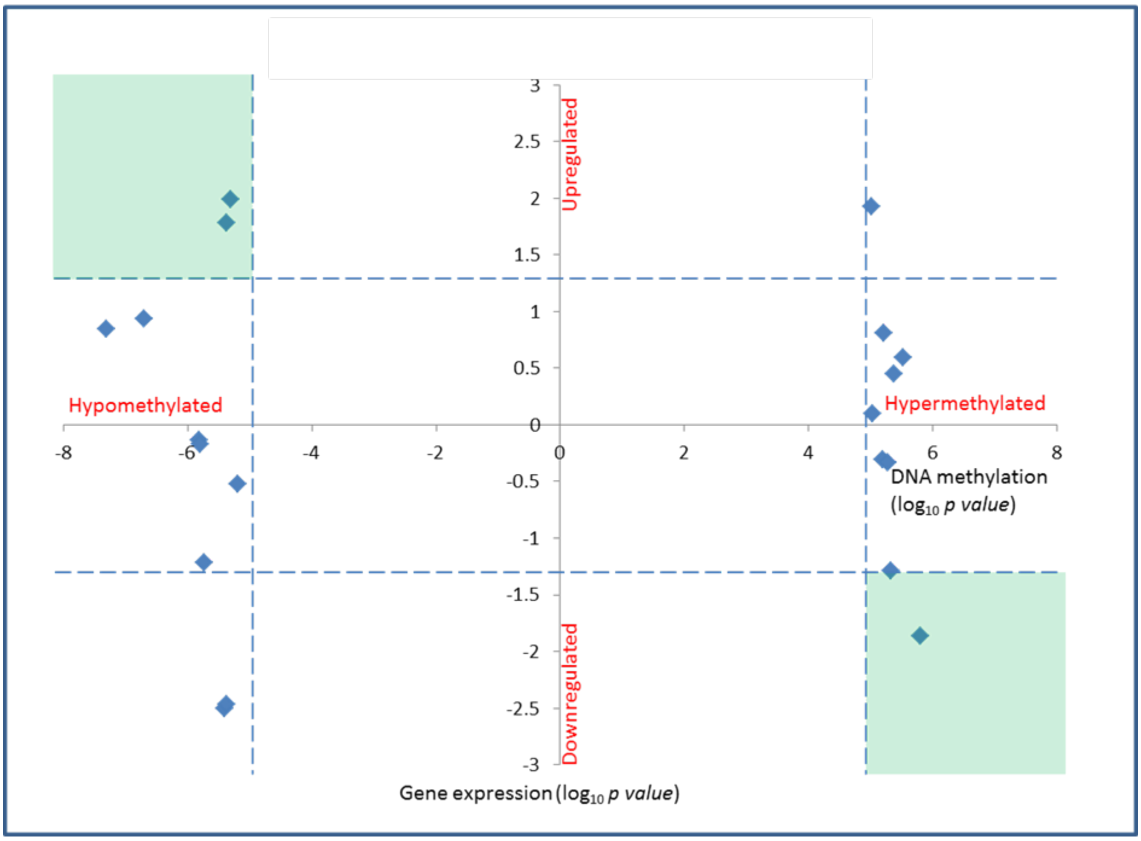


Figure 85: Quadrant plot of DMPs and associated expression levels for motor outcomes. DNA methylation (x axis) and gene expression (y axis) are graphically presented as $\log_{10}$ p value. Significance is defined as p< $1\text{x}10^{-5}$ for methylation and p<1.13 for expression (as indicated by the vertical and horizontal dashed lines). The direction of effect is indicated by the red text. The shaded areas represent sites with the expected inverse relationship between methylation and expression. Sites within these shaded areas were taken forward for further analysis.

Variable	CpG site	Gene	Meth p value	Meth effect size	Exp p value	SNP	Primers
Tremor dominant phenotype	cg16099282	*PGS1*	1.59E-06	-0.070	0.014	No	Yes
Tremor dominant phenotype	cg24560729	*KIAA1530*	4.32E-06	0.057	0.017	No	No
MDS-UPDRS III	cg18067096	*MOGAT1*	4.99E-06	0.050	0.010	No	Yes

Table 48: Description of 3 motor sites showing expression differences in PD. Meth=methylation, Exp=expression.

5.3.2 Primer optimisation and validation

Using the selection criteria described above, primer pairs for twenty loci (eighteen associated with a cognitive variable and two associated with a motor variable) were successfully designed as in Table 46 and Table 48. All twenty primer pairs targeted at least the index CpG (CpG included on the HM450). Where possible, amplicons were designed to capture additional neighbouring CpG sites.

Primer pairs for seven assays could not be optimised and were removed from further analysis. The primer pairs for the remaining thirteen successfully optimised assays were used to amplify, by PCR, a range of methylation concentrations spanning 0% to 100%, to ensure the assay was able to measure the full range of possible methylation values. An assay was considered to be validated if three conditions were reached: 1) the observed methylation value for 100% was above 80%; 2) the observed methylation value for 0% was below 10% and 3) there was a linear relationship with an R^2 value above 0.99. Ten assays were unable to measure the full range of methylation values and were subsequently dropped from further analysis. Three assays were successfully validated. Concordance of observed and reference methylation values are shown in Table 49, using the mean methylation value across all CpG sites in each amplicon. The validation results for both the pre-PCR and post-PCR mixes were plotted on a scatter graph with a trend line and R^2 values presented (Figure 86).

Expected Methylation (%)	**Observed Methylation (%)**					
	PROCA1		*CHCHD5*		*TRIM15*	
	Pre	Post	Pre	Post	Pre	Post
100	88.1	84.40	95.67	95.71	98.21	100.00
95	87.07	84.21	77.53	80.67	91.58	92.96
90	82.76	83.33	70.09	73.31	84.74	87.04
75	65.77	64.41	59.93	59.37	69.25	72.10
50	48.20	48.16	41.06	42.80	54.24	59.44
25	19.66	19.22	23.44	26.02	20.91	19.86
10	10.38	10.66	14.75	13.92	8.17	9.12
5	4.51	4.85	9.68	9.12	7.91	5.57
0	3.16	3.37	2.97	1.97	4.00	3.60

Table 49: Observed mean methylation across all CpGs compared with reference expected methylation level. Pre and post PCR methylation values are presented for each assay.

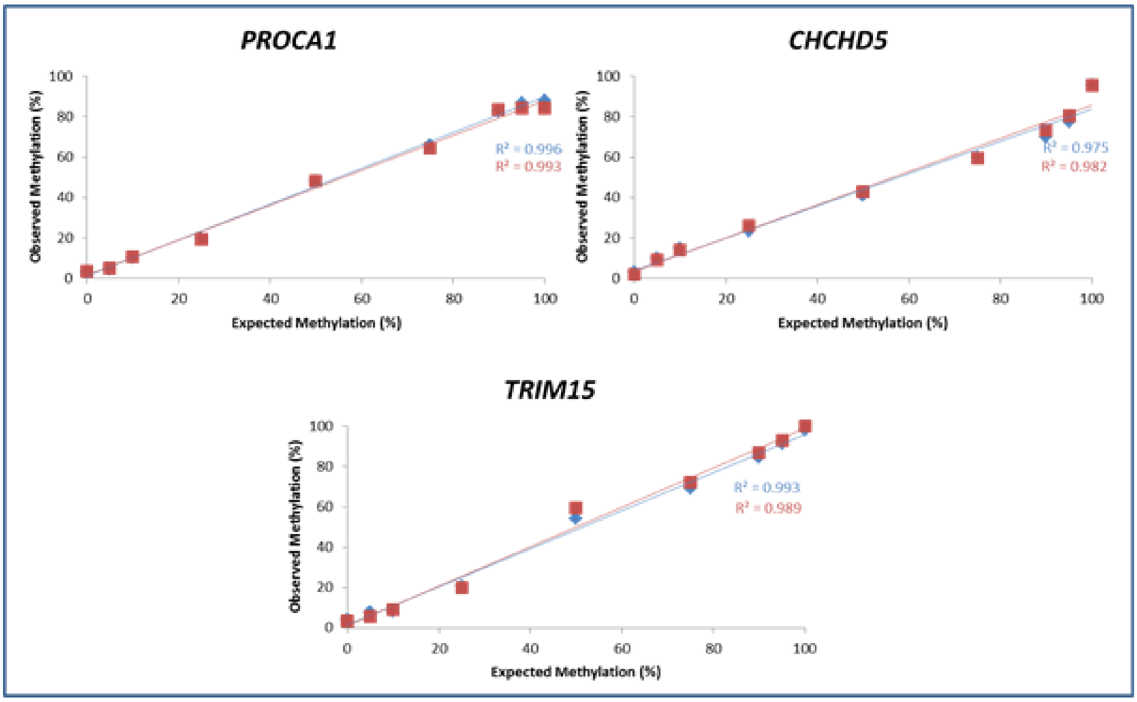


Figure 86: Composite figure of 3 assay validation plots. Pre-PCR methylation values are presented in blue and post-PCR methylation values in red. The R squared value is also displayed in the corresponding colour. An R squared value close to 1 indicates a very close correlation between observed and expected methylation levels.

5.3.3 Pyrosequencing analysis

All available blood DNA samples taken from ICICLE participants with PD were then run across all three validated assays in duplicate, along with all available blood DNA samples from healthy controls recruited into ICICLE. Samples were considered to have been assayed successfully using Pyrosequencing if the resulting values for duplicates were within 5%. A mean methylation level of these duplicates was calculated for each sample and this mean methylation value was used in all further analyses.

The data generated by Pyrosequencing were firstly compared to the DMP methylation values generated by the HM450 BeadChip to assess concordance. DNA methylation levels measured by Pyrosequencing were then analysed to investigate association between outcome (adjusted for covariates) and DNA methylation in the “discovery” cohort (samples run on the HM450). If associations were observed in the “discovery cohort, data were then analysed to look for associations in the larger “replication” cohort.

Of the three technically validated assays, two did not work on the Pyrosequencer when using ICICLE blood DNA samples, these being; *TRIM15* (associated with naming score) and *PROCA1* (associated with pentagon copying score). In both assays, the ICICLE blood DNA samples were successfully amplified with PCR products visible on an agarose gel, however DNA samples repeatedly failed on the Pyrosequencer. Since the PCR was successful, the error therefore occurred at the Pyrosequencing stage of the validation phase. Results of the remaining assay, *CHCHD5*, which was associated with naming score using the HM450 data, were taken forward for analysis in Stata.

5.3.3.1 ***CHCHD5***

5.3.3.1.1 Validation in discovery cohort

All 46 blood DNA samples analysed on the HM450 were run on the Pyrosequencer in duplicate. Within the Pyrosequencing data there were 32 concordant pairs of data, from which an average methylation value for each sample was taken.

Spearman's rank correlation was performed in Stata to test the correlation between the two platforms. A weak correlation of 0.256 was observed between the HM450 data and Pyrosequencing data, however this was not statistically significant (p=0.173) (Figure 87).

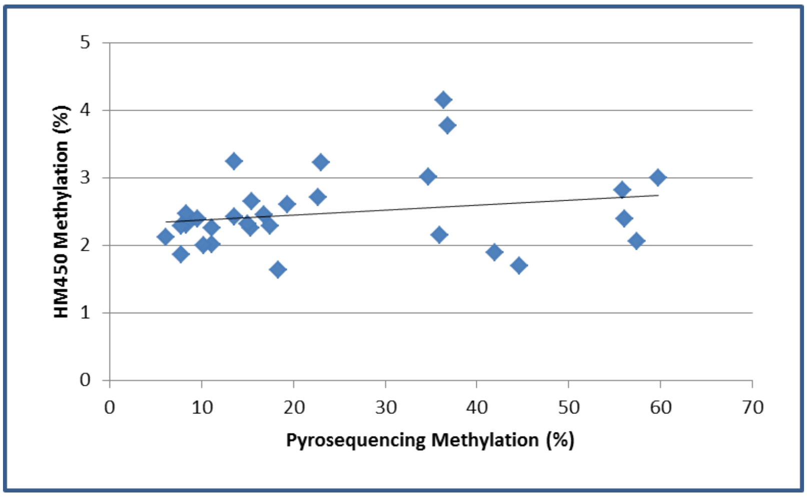


Figure 87: Correlation between methylation measured on the HM450 and Pyrosequencer at the *CHCHD5* locus.

Analysis of the HM450 data identified one CpG site within *CHCHD5* which was differentially methylated with naming score. At this CpG site, PD patients with a lower score on the naming test had an increased methylation level.

Figure 88 shows the mean methylation levels for naming score measured at the *CHCHD5* locus. The naming variable was a test scored out of 3, in which all participants scored either 2 or 3 out of the maximum 3 points. In both the HM450 and Pyrosequencing data, those participants achieving the lower score of 2 displayed increased methylation levels at the *CHCHD5* locus. The Pyrosequencer measured much higher levels of methylation than the HM450 and data were also much more variable as shown by the standard deviation bars, however the same pattern can be seen in relation to the naming score, with those scoring lower on the naming test exhibiting increased mean methylation levels.

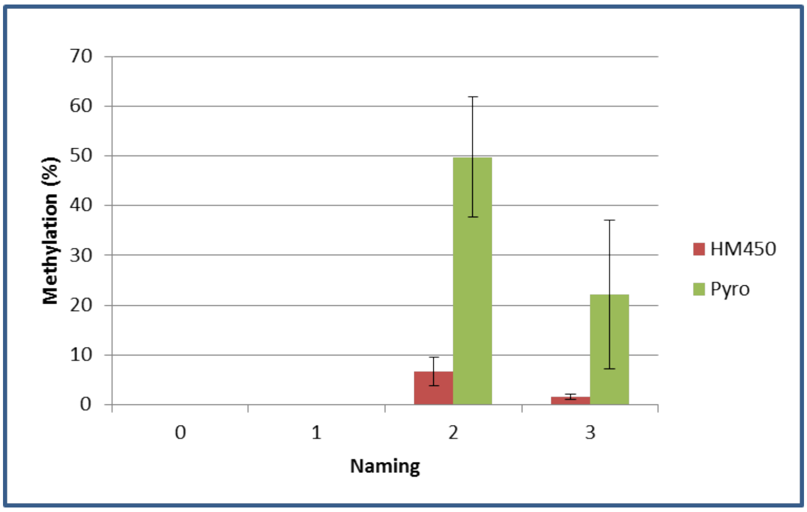


Figure 88: Mean methylation levels for naming score measured at the *CHCHD5* locus on the HM450 and Pyrosequencer.

The analysis model used to assess the relationship between naming and the HM450 data was applied to the Pyrosequencing data. This was to test whether the association between methylation and naming was still present using the Pyrosequencing data. In Stata, regression was performed (as in the analysis of HM450 data) to test for association between methylation at the index site and naming, using sex and age as covariates. The results of the HM450 were validated using Pyrosequencing (t=-4.53, p=1.16E-04). The methylation data were found to have a non-normal distribution using the Shapiro Wilk normality test. Therefore, Spearman’s rank correlation was used to assess whether the association observed between *CHCHD5* DNA methylation and naming was also observed in the Pyrosequencing data. The results of the HM450 were validated using Pyrosequencing at this locus using a non-parametric test (Rho = -0.520, p=0.003).

5.3.3.1.2 Validation in replication cohort

The Pyrosequencing assay was run across a larger independent sample subset (n=105) to test whether the association with naming score, at the index CpG site, was still observed. Of the samples run on the Pyrosequencer, 82 concordant pairs of data were identified. A mean methylation value for each sample was taken and used in the following analysis. Using both regression and Spearman's rank (due to the non-normal distribution of the methylation data) to analyse the Pyrosequencing data, a significant association was observed between methylation at the *CHCHD5* locus and naming score (regression: t=-3.07, p=0.003; Spearman's: Rho=-0.325, p=0.006). This increase in sample size led to a slight weakening of the significance level between DNA methylation and naming score; however there was still a significant association between naming and DNA methylation, indicating that this is likely to be a real association.

5.3.3.1.3 Comparison with healthy controls

In order to determine whether the observed differences in DNA methylation were unique to Parkinson's disease, DNA samples from 90 healthy controls were also included in the Pyrosequencing assay in duplicate. 79 pairs of concordant data were identified and a mean methylation value was calculated for each sample. In controls, there was no association between *CHCHD5* methylation and naming (Rho=0.012, p=0.918). Significant differences were observed in the DNA methylation levels at the *CHCHD5* locus between PD cases and controls (z=3.727, p=0.0002). The results of the methylation analysis for the *CHCHD5* locus in both controls and PD cases are displayed in Figure 89. This difference between PD and controls suggest that the differential methylation observed in both the HM450 and Pyrosequencing data, when analysing the DNA methylation data in relation to naming, is unique to PD cases. This suggests that there may be other variables influencing DNA methylation levels in PD cases, as naming score is not associated with DNA methylation in controls.

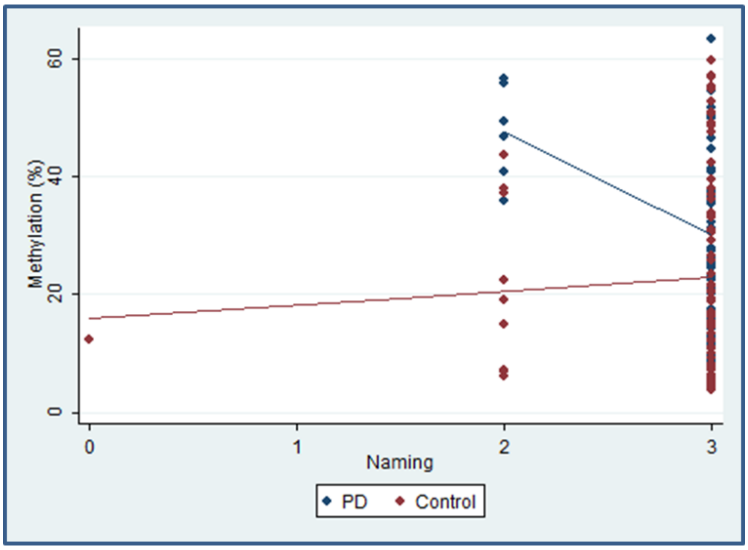


Figure 89: *CHCHD5* DNA methylation analysis with language score in both PD cases and controls. PD cases (Pyrosequencing cohort only) are shown in blue and controls are displayed in red.

DNA methylation at the *CHCHD5* locus was not found to be associated with any other cognitive or motor outcome variable.

5.3.4 Further Pyrosequencing analysis

Due to the high failure rate of the assays designed for the index CpG sites, an additional approach was adopted. Primers for seven Pyrosequencing assays failed to be optimised (Section 5.3.2). For five of these genes, predesigned Pyromark CpG Assays (Qiagen, UK) were available. None of these assays targeted the index CpG site, however it was still of interest to assess whether any other CpG sites within the gene exhibited the differential methylation seen between methylation and outcome in the HM450 data. Pyromark CpG Assays were obtained for the following genes: *C7orf50*, *C16orf45*, *MOGAT1*, *PGS1* and *RUNDC3A*. Due to very limited sample quantity a reduced number of samples were run in each assay compared to *CHCHD5*.

Like previously tested assays, a number failed when ICICLE DNA samples were run on the Pyrosequencer. Pyrosequencing data are not available for *C16orf45* and *C7orf50* due to all samples repeatedly failing. The remaining three assays were successfully run across a number of DNA samples.

5.3.5 RUNDC3A

Methylation at a CpG site within *RUNDC3A* was found to be associated with language score. The language test is scored out of 2, with the majority of participants scoring the maximum of 2 points. The HM450 identified those scoring 0 points on the language test had lower levels of methylation compared to those scoring >1 point (5.1% vs 10.2%, p=7.91E-06). The Pyrosequencing data cannot be used to validate these findings as the assay does not include the index CpG site; however the Pyrosequencing data were used to assess whether this differential methylation observed at a CpG site within *RUNDC3A* is also observed at other CpG sites within the gene. Four CpG sites were included in the Pyrosequencing array. 26 pairs of concordant data were available for *RUNDC3A* Pyrosequencing analysis. Spearman's rank correlation was used to test for association between language score and DNA methylation at each individual CpG site as well as the mean methylation across the region (i.e. mean across 4 CpG sites). Results of these tests are displayed in Table 50. At neither CpG site, nor across the region, was DNA methylation found to be associated with language score.

CpG site	**Rho**	**P value**
Site 1	0.247	0.224
Site 2	0.194	0.344
Site 3	0.172	0.401
Site 4	0.052	0.800
Region	0.170	0.408

Table 50: Association between *RUNDC3A* methylation and language score.

In addition to the PD samples, DNA from twelve healthy age and sex matched controls were also run across the *RUNDC3A* assay. Eight pairs of concordant data were identified and a mean methylation value calculated for each sample. DNA methylation at these specific loci within *RUNDC3A* did not differ between controls and PD cases. The results of the methylation analysis for the *RUNDC3A* region in both controls and PD cases are displayed in Figure 90.

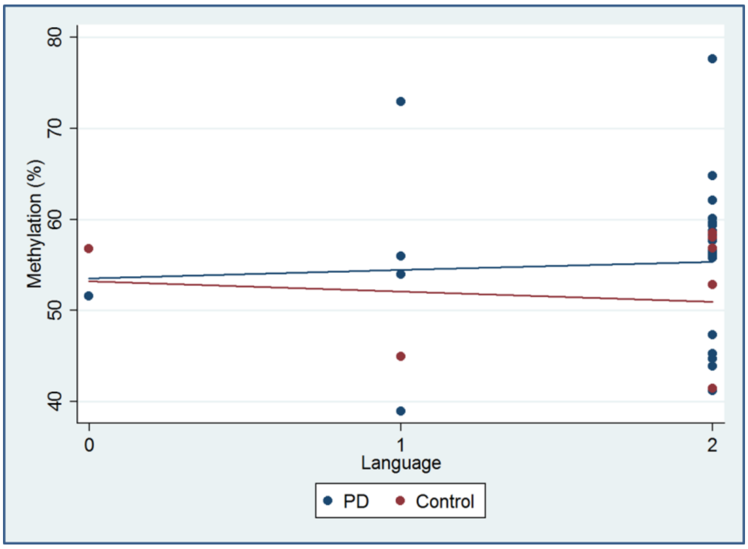


Figure 90: Regional *RUNDC3A* DNA methylation analysis with language score in both PD cases and controls. PD cases are shown in blue and controls are displayed in red.

5.3.6 *MOGAT1*

Methylation at a CpG site within *MOGAT1* was found to be associated with MDS-UPDRS III. The HM450 identified those with a higher score on the MDS-UPDRS III scale and therefore more severe motor impairment had lower levels of methylation compared to those whose motor disease was not as severe (effect size= 0.050, p=4.99E-06). The Pyrosequencing data cannot be used to validate these findings as the assay does not include the index CpG site; however the Pyrosequencing data were used to assess whether this differential methylation observed at a CpG site within *MOGAT1* is also observed at other CpG sites within the gene. Five CpG sites were included in the Pyrosequencing array, however due to limited number of data points available for the final two CpG sites included in the assay, only the first three CpG sites were included in the analysis. For CpG sites 1 and 2, there were 62 pairs of concordant data available for *MOGAT1* Pyrosequencing analysis, and in CpG site 3 there were 43 pairs of concordant data available for Pyrosequencing analysis. Spearman’s rank correlation was used to test for association between MDS-UPDRS III and DNA methylation at each individual CpG site, as well as the mean methylation across the region (i.e. mean across 3 CpG sites). Results of these tests are displayed in Table 51. At neither CpG site, nor across the region, was DNA methylation found to be associated with MDS-UPDRS III.

CpG site	Rho	P value
Site 1	0.167	0.253
Site 2	0.119	0.412
Site 3	0.190	0.267
Region	0.250	0.142

Table 51: Association between *MOGAT1* methylation and MDS-UPDRS III.

5.3.7 PGS1

Methylation at a CpG site within *PGS1* was found to be associated with tremor dominant phenotype. The HM450 identified those with a more severe tremor had increased levels of methylation compared to those whose motor disease was not as severe (effect size= -0.070, p=1.59E-06). The Pyrosequencing data however cannot be used to validate these findings as the assay does not include the index CpG site. As above, the Pyrosequencing data were used to assess whether this differential methylation observed at the CpG site (included on the HM450 BeadChip) within *PGS1* is also observed at other CpG sites within this gene. Four CpG sites were included in the Pyrosequencing array. 23 pairs of concordant data were available for *PGS1* Pyrosequencing analysis. Spearman's rank correlation was used to test for association between tremor dominant phenotype score and DNA methylation at each individual CpG site as well as the mean methylation across the region (i.e. mean across 4 CpG sites). Results of these tests are displayed in Table 52. At neither CpG site, nor across the region, was DNA methylation found to be associated with tremor dominant phenotype.

CpG site	Rho	P value
Site 1	-0.182	0.407
Site 2	-0.143	0.515
Site 3	-0.149	0.497
Site 4	0.097	0.659
Region	0.002	0.995

Table 52: Association between *PGS1* methylation and tremor dominant phenotype.

Due to MDS-UPDRS III and tremor dominant phenotype being motor related outcomes and therefore not present in healthy controls, there is no control data available for comparison for either *MOGAT1* or *PGS1*.

In summary, no HM450 findings were validated using the predesigned assays for three of the genes (*RUNDC3A*, *MOGAT1*, *PGS1*). This may be due to different CpG sites within the gene being investigated.

5.4 Discussion

This work aimed to identify potential biomarkers of PD-MCI using blood samples taken following idiopathic PD diagnosis. In the ICICLE cohort, differences between MCI and NCI samples were as expected, with the MCI cases scoring worse on all cognitive outcomes when compared to NCI samples. MCI cases were also more likely to have had fewer years in education and suffer from hypertension and depression. In addition to being more cognitively impaired, the MCI group were also more likely to have a worse motor function than the NCI group.

The HM450 array identified 46 CpG sites, in which the gene they were located in had previously been found to be differentially expressed in PD, to be associated with cognitive deficits in the ICICLE cohort. In addition, six CpG sites that had previously been found to be differentially expressed in PD were found to be associated with motor deficits in the ICICLE cohort. Following filtering of top hits, three assays were run across a larger independent ICICLE cohort to confirm whether the association observed was replicated when measured using Pyrosequencing. One of these assays (*CHCHD5*) successfully validated the HM450 findings, suggesting that the observations were confirmed using an alternative technology and that this may be a potential biomarker for PD-MCI.

These novel results were promising; however the function of *CHCHD5* is unclear. A study by Banci *et al.* (2012) showed that it is recognised as a substrate for the Mia40-dependent mitochondrial membrane import machinery, suggesting that CHCHD5 is taken up into the mitochondrial membrane space, where it may perform its yet to be determined role (Banci *et al.*, 2012). Another CHCHD family member, *CHCHD10* has previously been implicated in pathologically proven frontotemporal dementia-amyotrophic lateral sclerosis (FTD-ALS). A missense point mutation (c.176C>T; p.Ser59Leu) was discovered in this gene in a family with a history of FTD-ALS (Bannwarth *et al.*, 2014). Similarly to CHCHD5, the authors showed that CHCHD10 is located within the mitochondrial membrane space indicating that they may have related roles. A later study on a larger cohort of FTD-ALS patients discovered a second mutation (c.100C>T; p.Pro34Ser) (Chaussenot *et al.*, 2014). Although the overall prevalence of these mutations was low (n=115, 2.6%), these discoveries still indicate a possible role for CHCHD family proteins in mitochondrial function related to cognitive and motor decline.

These results suggest that the CpG site within *CHCHD5* could be a potential biomarker of PD-MCI. However, only one CpG site within *CHCHD5* was found to be associated with one cognitive outcome, the naming test. Due to only one CpG site being included in the Pyrosequencing assay, it is therefore unknown whether the differential methylation observed at this CpG site is also observed at neighbouring CpG sites, or if this is a feature of this CpG site alone. More extensive interrogation of the region is warranted and this could be achieved using a technology such as targeted bisulphite sequencing (Herman *et al.*, 1996). In addition, the accuracy of the naming test also needs to be considered. The naming test was scored out of 3, and all participants (cases and controls) had a score of either 2 or 3. This suggests that there is not much variation between participants so the differential methylation observed at the *CHCHD5* locus may not have the range, or power, to discriminate between relatively subtle differences in cognitive traits. The validity of this biomarker as a predictor of later cognition is therefore uncertain. DNA methylation levels were measured at the *CHCHD5* locus in PD cases and healthy controls. Interestingly, the differential methylation observed at this locus was only observed in the PD cases, and the healthy controls did not exhibit any differences in methylation that were attributable to naming score. This suggests that there may be another factor, in addition to, or independently of naming score, contributing to DNA methylation in PD cases. This may be an unmeasured confounder unrelated to cognition.

As discussed in other Chapters in this thesis, potential biomarkers may have been lost during the selection and filtering of HM450 top hits. Firstly, CpG sites were selected based on methylation p value and publicly available expression data. Only those hits showing an inverse methylation-expression relationship were selected for the follow up. This inverse relationship is only expected within promoter regions (Jjingo *et al.*, 2012; Jones, 2012) so this criterion could have potentially removed possible biomarkers from the selection. The CpG sites identified to be SNPs were removed from the selection of top hits to be taken forward for Pyrosequencing validation. The SNPs identified as being associated with outcome were not investigated in this thesis, but it is possible that they could also be useful biomarkers of disease. This would be a possible field for further research in this area.

Secondly, potential biomarkers may have been removed from the selection when designing primers using the Pyrosequencing assay design software. Some CpG sites were not amenable for Pyrosequencing due to densely populated CpG island regions and potential mispriming sequences. For this reason, these CpG sites were dropped from the selection, however an additional technique could be used to interrogate DNA methylation at these sites, such as the Sequenom® Epityper™ (Ehrich *et al.*, 2005) or bisulphite sequencing (Herman *et al.*, 1996).

Finally, a number of assays could not be optimised or failed validation (could not measure the full range of methylation values). For five of these assays, predesigned Pyromark assays were used, however these assays did not target the index CpG site and so were not able to be used to validate the HM450 findings, but were able to assess whether DNA methylation at other CpG sites within the same gene are differentially methylated with outcome. In none of the assays assessed was DNA methylation found to be associated with outcome. This approach was only adopted for those assays which could not be optimised, however this approach could also be adopted for those which failed validation or for assays which could not be designed using the Pyrosequencing Assay Design software.

Small sample sizes pose problems in most statistical tests. A test of correlation assumes a linear relationship and roughly even distribution of values across a scale. With few samples, these assumptions can easily be violated and the correlation estimates can therefore be imprecise. A source of bias with potentially greater influence is outliers. A single outlier can completely change the slope of a regression line and therefore the correlation coefficient. Firm conslusions therefore cannot be drawn from correlation value alone. More clarity on the true correlation of the methylation generated would be gained with a larger sample size and close scrutiny of the impact of outliers.

Due to the heterogeneous nature of blood, it is possible that any differences in DNA methylation observed between outcome groups is due to the cellular composition of the blood. Methylation beta values were adjusted for cellular composition to determine whether the differential methylation observed in relation to the outcome variables (both cognitive and motor) were actually due to differences in cellular composition. Adjusting for cellular composition revealed very little difference in significance, indicating that differences in cellular composition have no significant effect on the methylation differences observed between outcome groups in this Chapter.

In summary, limited evidence was generated to support the potential use of DNA methylation signatures as predictive biomarkers of PD-MCI. Conclusions were limited largely by technical issues encountered (such as the inability to design assays to validate discovery observations) and more extensive analysis of DNA methylation using other approaches is warranted. Due to the resources available, many potential biomarkers remained under-investigated and should form the basis of further studies.

Chapter 6. DNA methylation signatures as exposure indicators in post-stroke dementia and Parkinson’s disease

6.1 Introduction

As previously discussed in Chapter 1, PSD and PD are likely to be caused by a combination of genetic, epigenetic and environmental factors. Several environmental factors have been identified in the literature. These disease-specific exposures will be explored in more detail in Sections 6.1.1 and 6.1.2. However, evidence for the associations between exposure and outcome has sometimes been contradictory and inconclusive. One possible explanation for this is the inaccuracy inherent in the measurement of exposures, and biomarkers are increasingly being applied to index such exposures. For example, smoking can be difficult to accurately measure due to recall and reporting bias on self-reported questionnaires. Smoking is likely to be therefore under reported. Cotinine is a metabolite of nicotine, detectable in blood or urine, and is commonly used as a biomarker to estimate the smoking level. However, cotinine has a short half-life (16 hours) and is unable to inform about former smoking habits (Vartiainen *et al.*, 2002).

An emerging area in the field of epigenetics is to utilise DNA methylation as an indicator of exposure. DNA methylation has previously been identified as an indicator of smoking status and has been used to distinguish between smokers and never smokers (Elliott *et al.*, 2014) and never and former smokers (Shenker *et al.*, 2013).

Given the potential of environmental factors to permanently alter DNA methylation, these markers could give an indication of exposures that occurred many years before (Section 1.3.4) and therefore be a longer term archive of exposure compared to biomarkers such as cotinine. Using DNA methylation as a biomarker for exposure could help to eliminate the problem of recall and reporting bias, especially for long term historical exposures.

The use of biomarkers in this way does not necessarily require the biomarker to be on the causal pathway between exposure and outcome. For example CRP is a very useful clinical biomarker of atherosclerosis but it is not causally involved in the development of atherosclerosis, it is simply a marker of the inflammatory processes that occur in atherosclerosis (Patel *et al.*, 2001; Timpson *et al.*, 2005; Trion *et al.*, 2005). Evidence such as this underscores that biomarkers can have clinical utility without the necessity for causal involvement in a disease pathway. Conversely, if a causal relationship is observed, this may fuel intervention to prevent disease initiation or progression, so aiming to establish causal relationships is imperative. This applies equally to establishing the causal (or non-causal) role of DNA methylation linking exposures with diseases, so that the relevance of DNA methylation as a predictive biomarker or intervention target can be prioritised.

The aim of the work presented in this Chapter is to assess whether DNA methylation profiles can act as indicators/biomarkers for environmental exposures of PSD and PD and to determine in such cases whether these biomarkers are directly mediating the effects of exposure on disease or simply acting as a non-causal proxy. Figure 91 shows the associations tested in this thesis to investigate whether DNA methylation can act as a mechanism linking exposure with disease.

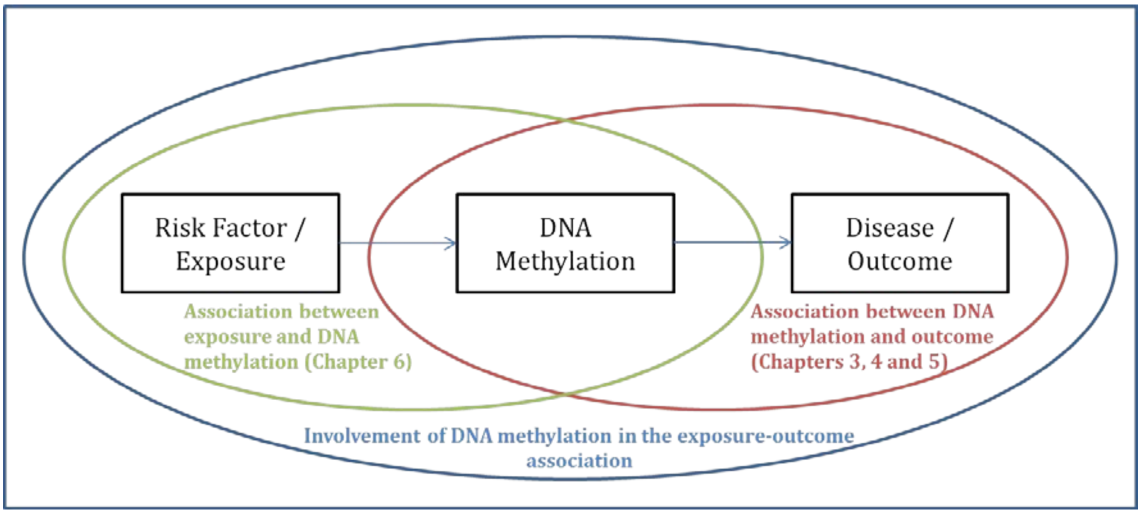

Risk Factor / Exposure
DNA Methylation
Disease / Outcome
Association between exposure and DNA methylation (Chapter 6)
Association between DNA methylation and outcome (Chapters 3, 4 and 5)
Involvement of DNA methylation in the exposure-outcome association

Figure 91: Associations tested in this thesis to assess the role of DNA methylation in the exposure-outcome association (Adapted from Michels, 2012).

6.1.1 Exposures affecting post-stroke dementia

PSD most commonly affects the elderly, those with a lower level of education and those with other co-morbidities (Leys *et al.*, 2005). Many of the risk factors for stroke are also thought to affect the risk of PSD. There are a number of medical conditions which have been associated not only with an increased risk of stroke, but also an increased risk of PSD. Cardiovascular risk factors are major contributors to the risk of PSD (Leys *et al.*, 2005). Hypertension and atrial fibrillation are considered, after age, the second largest risk factors for PSD. Possible mechanisms linking these disorders to cognitive decline include a reduced perfusion of the brain caused by a reduced and impaired regularity of cerebral blood flow, which may lead to subsequent brain damage and cognitive decline (Ott *et al.*, 1997). A variety of heart problems such as ischaemic heart disease, angina and cardiac failure may also affect the risk of stroke in the same way (Paciaroni and Bogousslavsky, 2013). Another related indicator of high risk is intermittent claudication caused by reduced blood flow to the legs causing muscle pain (Tilvis *et al.*, 2004).

The severity of the stroke indicated by the size and location of the infarct has also been linked to the risk of dementia. Multiple lesions, seen in patients with recurrent stroke are also strongly associated with a doubling in risk of PSD (Pendlebury and Rothwell, 2009). Imaging techniques can provide a lot of information regarding the severity of stroke, providing evidence for location, number and size of lesions, degree of cerebral atrophy, all of which may indicate the risk of PSD (Gorelick, 1997). In addition, symptoms suffered at the time of stroke may also be able to indicate those at an increased risk of developing dementia.

Factors relating to a poor lifestyle including low levels of physical activity and obesity, and smoking are also thought to impact upon PSD risk. Due to the increased risk of cardiovascular related disease, obesity and poor physical activity have been associated with increased risk of stroke and PSD (Leys *et al.*, 2005). Smoking is a known risk factor for both stroke and dementia. It has been suggested that, compared to never smokers, heavy smoking can double the risk of both AD and VaD (Ott *et al.*, 1998; Sahathevan *et al.*, 2012). Alcohol has also been associated with an increased risk of cognitive decline among the elderly (Thomas and Rockwood, 2001) although this association has not been studied with respect to post-stroke dementia.

6.1.2 Exposures affecting mild cognitive impairment in Parkinson’s disease

Like PSD, there are a number of exposures which have been associated with an increased risk of PD-MCI. A number of demographic risk factors have been associated with PD-MCI including increasing age, male sex, working in the agricultural sector (Section 1.2.4.2) and low levels of education (Palavra *et al.*, 2013).

Lifestyle factors have also been associated with PD-MCI risk, including smoking and alcohol consumption. Data on smoking as a risk factor for PD-MCI is inconsistent with some papers reporting a protective effect (Wang *et al.*, 2010; Schapira and Jenner, 2011) whilst others suggest it is a significant risk factor for cognitive impairment (Anstey *et al.*, 2007). This disparity could be due to the difficulty in gaining accurate information on smoking behaviour due to the tendency of participants to under report smoking habits. As described in Section 6.1.1, alcohol consumption has been associated with an increased risk of dementia in the elderly population so may also be a relevant risk factor for PD-MCI.

A decrease in body mass index (BMI) following Parkinson's disease diagnosis has been associated with an increased risk of cognitive decline although it is not clear whether this is a causal relationship (Kim *et al.*, 2012a) as it may be indicative of underlying disease processes that confound the associations.

A number of biochemical measures can also indicate risk of MCI in the general population and possibly can be extrapolated to PD. Reduced red cell folate (RCF) and vitamin B12 levels have been shown to lead to increased levels of homocysteine. In turn, elevated homocysteine levels have been associated with both MCI and AD. In addition to RCF and B12 levels, homocysteine levels can also be affected by Levodopa usage, the golden standard treatment for PD (Rodriguez-Oroz *et al.*, 2009).

A number of medical conditions may also increase the risk of developing PD-MCI. Diabetes, hypertension and hypercholesterolaemia have all been associated with increased risk of PD-MCI with a higher prevalence of these diseases observed in PD-MCI cases compared to cognitively normal PD cases. However, findings in this field have not proven consistent (Kandiah *et al.*, 2013). Finally, greater depression scores have also been reported to be higher in PD-MCI cases than PD cases with normal cognition (Yarnall *et al.*, 2013).

6.2 Experimental design

6.2.1 COGFAST exposures

COGFAST is a well characterised study with a wealth of exposure data providing an excellent opportunity to study the effects of exposures on post-stroke dementia and DNA methylation. A number of exposures were recorded at the time of stroke which relate to the index stroke (i.e. not previous or subsequent strokes that may have occurred). In addition, a number of exposure variables were recorded at recruitment into COGFAST including baseline CAMCOG scores. Table 53 describes all exposures collected in COGFAST that were used in the analysis reported in this Chapter.

Exposure Type	Exposure	Description	Time recorded
Stroke-related	OCSP class	The type of stroke experienced according to the Oxfordshire Community Stroke Project	Index stroke
Stroke-related	Side of body	The side of the body, if any, affected by the stroke	Index stroke
Stroke-related	Degree of weakness arm	The degree of weakness experienced in arm following stroke	Index Stroke
Stroke-related	Degree of weakness leg	The degree of weakness experienced in leg following stroke	Index Stroke
Stroke-related	Dysphasia	Whether dysphasia was a symptom following the stroke	Index Stroke
Medical history	Hypertension	Present or absent, defined by blood pressure >140/90 mmHg	Recruitment
Medical history	Atrial fibrillation	Present or absent, defined by WHO criteria	Recruitment
Medical history	IHD	Present or absent, defined by WHO criteria	Recruitment
Medical history	Angina	Present or absent, defined by WHO criteria	Recruitment
Medical history	Cardiac failure	Present or absent, defined by WHO criteria	Recruitment
Medical history	Intermittent claudication	Present or absent, defined by WHO criteria	Recruitment
Medical history	No of CVD risk factors	Number of Cardiovascular risk factors (Smoking, Diabetes, Hyperlipidaemia, Peripheral vascular disease, IHD, Atrial fibrillation, Hypertension)	Recruitment
Lifestyle	Smoking	Current, ex or never smoker	Recruitment
Lifestyle	Alcohol	Average units of alcohol per week	Recruitment
Baseline cognition	MMSE	Score at baseline (max=30)	Recruitment
Baseline cognition	Orientation	Score at baseline (max=10)	Recruitment
Baseline cognition	Language comprehension	Score at baseline (max=9)	Recruitment
Baseline cognition	Language expression	Score at baseline (max=21)	Recruitment
Baseline cognition	Memory remote	Score at baseline (max=6)	Recruitment
Baseline cognition	Memory recent	Score at baseline (max=4)	Recruitment
Baseline cognition	Memory learning	Score at baseline (max=17)	Recruitment
Baseline cognition	Memory total	Score at baseline (max=27)	Recruitment
Baseline	Attention	Score at baseline (max=7)	Recruitment

cognition			
Baseline cognition	Praxis	Score at baseline (max=12)	Recruitment
Baseline cognition	Calculation	Score at baseline (max=2)	Recruitment
Baseline cognition	Abstract thinking	Score at baseline (max=8)	Recruitment
Baseline cognition	Perception	Score at baseline (max=11)	Recruitment
Baseline cognition	Executive function	Score at baseline (max=28)	Recruitment
Baseline cognition	Total CAMCOG score	Score at baseline (max=107)	Recruitment

Table 53: List of exposures measured in COGFAST.

6.2.2 ICICLE exposures

ICICLE is a very well characterised study with detailed exposure data upon which the utility of DNA methylation as a marker of exposure and the relationship between exposures and PD-MCI can be investigated. A number of exposures were recorded at recruitment into the ICICLE study. They are described in Table 54.

Exposure Type	Exposure	Description	Time recorded
Demographic	Education	Years in education	Recruitment
Demographic	NART	National Adult Reading Test score	Recruitment
Lifestyle	Alcohol	Average units of alcohol per week	Recruitment
Lifestyle	Smoking	Current, ex or never smoker	Recruitment
Anthropometric	Height	Height (m)	Recruitment
Anthropometric	Weight	Weight (kg)	Recruitment
Anthropometric	BMI	Body Mass Index (kg/m^2)	Recruitment
Biochemical	RCF	Red cell folate measurement (µg/L)	Recruitment
Biochemical	B12	Vitamin B12 measurement (ng/L)	Recruitment
Biochemical	Homocysteine	Homocysteine measurement (µmol/L)	Recruitment
Medical history	IHD	Present or absent, defined by WHO criteria	Recruitment
Medical history	Diabetes	Present or absent, defined by WHO criteria	Recruitment
Medical history	Hypertension	Present or absent, defined by WHO criteria	Recruitment
Medical history	Hypercholesterolaemia	Present or absent, defined by WHO criteria	Recruitment
Medical history	GDS	Geriatric Depression Scale score	Recruitment
Medical history	Levodopa dosage	Levodopa daily dose (mg)	Recruitment

Table 54: List of exposures measured in ICICLE.

6.2.3 Relationship between exposures and outcomes

In Stata, the association between exposure and outcome was assessed using either Spearman’s rank, Kruskal Wallis, Wilcoxon rank sum, t tests or chi squared.

The outcome variables tested in COGFAST are described in detail in the Glossary (Table 2). In summary these were; diagnosis, Braak staging, MMSE, orientation, language comprehension, language expression, memory remote, memory recent, memory learning, memory total, attention, praxis, calculation, abstract thinking, perception, executive function and total CAMCOG score. For the analyses using COGFAST data, tests with p values $<1.05\text{x}10^{-4}$ were accepted as indicating an association between exposure and outcome after applying Bonferroni correction based on the number of tests conducted and $\alpha=0.05$.

The outcomes tested in ICICLE are described in detail in the Glossary (Table 4). In summary the motor outcomes were; Hoehn and Yahr, MDS-UPDRS II, MDS-UPDRS III, tremor dominant phenotype and PIGD phenotype. The cognitive outcomes tested in ICICLE were; MoCA, MMSE, total FAS, animals, pentagon copying, naming, language, language total, NMSQ memory, NMSQ concentration, total number of NMS, power of attention, digit vigilance accuracy, cognitive complaint, MCI, PRM, SRM, PAL and OTS. For the analyses using ICICLE data, tests with p values $<1.30\text{x}10^{-4}$ were accepted as indicating an association between exposure and outcome after applying Bonferroni correction based on the number of tests conducted and $\alpha=0.05$.

6.2.4 Associations between exposures and DNA methylation

The CpGassoc package (Barfield *et al.,* 2012), which performs multiple linear regression analyses with continuous predictor variables and ANOVA for categorical predictor variables, was implemented in R (version 2.15.0). For all analyses, age, sex and chip (i.e. the microarray chip upon which the samples were scanned) were included in the analysis model as covariates. Significant results were considered as those reaching genome-wide significance ($p<1.13\text{x}10^{-7}$). To assess the influence of cellular composition in the samples, HM450 data were subsequently adjusted for cellular composition using the method described by Houseman *et al.* (2012) and the CpGassoc analyses were repeated, as a sensitivity analyses, to assess whether adjustment for cellular composition impacted on the initial analysis.

6.2.5 Using methylation scores to predict smoking exposure status

The method used here has been described by Elliott *et al.* (2014). Methylation values published by Zeilinger *et al.* (2013) were used as a reference data set. Zeilinger *et al.* (2013) identified 187 CpG sites to be associated with smoking. Four of these CpG sites in the ICICLE cohort were dropped in the initial QC (Section 5.3.1). Methylation at the remaining 183 CpG sites was used to calculate weighted methylation scores for each DNA sample. Median methylation values of never smokers identified in Zeilinger *et al.* (2013) were used as the reference values and the associated effect sizes were used as weights. The difference between the ICICLE beta value and reference beta value was calculated for each CpG site. The sum of all CpG site scores was calculated to give a final weighted score in each person. The threshold score able to distinguish current from never smokers was calculated using random forests (Breiman, 2001) in R (Version 3.1.0)

6.2.6 Identifying CpG sites associated with both an exposure and outcome

To assess whether any CpG sites were associated with both an exposure and outcome the analyses using the CpGassoc package (Barfield *et al.,* 2012) were repeated. First, the ANOVA analysis was performed to look for association between outcome variables and methylation of those CpG sites significantly associated (i.e. reaching genome-wide significance) with an exposure variable. Then, the reverse was tested to look for association between exposure and methylation at only those CpG sites reaching genome-wide significance with an outcome. Any CpG sites which were significant with both an exposure and outcome were then tested for interaction.

6.2.7 Testing for interaction

First, multiple regression was used to test whether the addition of DNA methylation into the regression model altered the relationship between exposure and outcome. If DNA methylation was found to alter the relationship between exposure and outcome then DNA methylation could be considered to be involved in the relationship between them. If DNA methylation was thought to link the exposure with outcome, an interaction term was included in the model, to test whether the effect of methylation on outcome is affected by the presence or absence of the exposure. All analyses were performed in Stata.

6.3 Results

6.3.1 Relationship between exposure and outcome

Both cohorts (COGFAST and ICICLE) measured a range of exposures, detailed in 6.2.1 and 6.2.2. For each cohort, the relationship between exposure and outcome in the entire cohort was assessed using either Spearman’s rank correlation, Kruskal Wallis, Wilcoxon rank sum, t tests or a chi squared test, as appropriate.

6.3.1.1 ***Post-stroke dementia***

All measured exposures relating to the index stroke, lifestyle factors and medical history previously associated with PSD in the literature were compared to outcome variables, to test whether these exposure variables are associated with outcome in the COGFAST cohort. The associations were tested on all COGFAST participants that had a complete data set (including final diagnosis and Braak staging information). Participants who were still living and did not have information on final diagnosis and Braak staging were excluded from this analysis. All associations between exposure and outcome in PSD where the Bonferroni corrected p<0.05 are displayed in Table 55. Complete results tables can be found in Appendix C.

Exposure Type	Exposure	Outcome	Test statistic	P value
Baseline Cognition	Orientation	MMSE	0.732	6.40E-06¶
Baseline Cognition	Orientation	Language expression	0.661	0.0001¶
Baseline Cognition	Orientation	Memory learning	0.675	0.0001¶
Baseline Cognition	Orientation	Memory total	0.726	8.36E-06¶
Baseline Cognition	Orientation	Attention	0.705	1.99E-05¶
Baseline Cognition	Orientation	Total CAMCOG	0.672	0.0001¶
Baseline Cognition	Language comprehension	Language expression	0.657	0.0001¶
Baseline Cognition	Language expression	Language expression	0.689	3.64E-05¶
Baseline Cognition	Memory remote	Memory remote	0.650	0.0001¶
Baseline Cognition	Memory remote	Memory total	0.681	4.72E-05¶
Baseline Cognition	Perception	Perception	0.671	0.0001¶
Baseline Cognition	MMSE	MMSE	0.776	7.77E-07¶
Baseline Cognition	MMSE	Orientation	0.674	0.0001¶
Baseline Cognition	MMSE	Language comprehension	0.666	0.0001¶
Baseline Cognition	MMSE	Language expression	0.698	2.54E-05¶
Baseline Cognition	MMSE	Memory learning	0.771	1.01E-06¶
Baseline Cognition	MMSE	Memory total	0.735	5.62E-06¶
Baseline Cognition	MMSE	Attention	0.695	2.88E-05¶
Baseline Cognition	MMSE	Praxis	0.827	3.14E-08¶
Baseline Cognition	MMSE	Total CAMCOG	0.807	1.28E-07¶
Baseline Cognition	Total CAMCOG	Diagnosis	5.066	0.0001§
Baseline Cognition	Total CAMCOG	MMSE	0.766	1.31E-06¶
Baseline Cognition	Total CAMCOG	Orientation	0.703	2.10E-05¶
Baseline Cognition	Total CAMCOG	Language expression	0.765	1.35E-06¶
Baseline Cognition	Total CAMCOG	Memory learning	0.702	2.23E-05¶
Baseline Cognition	Total CAMCOG	Memory total	0.726	8.14E-06¶
Baseline Cognition	Total CAMCOG	Attention	0.675	0.0001¶
Baseline Cognition	Total CAMCOG	Praxis	0.686	3.98E-05¶
Baseline Cognition	Total CAMCOG	Executive function	0.659	0.0001¶
Baseline Cognition	Total CAMCOG	Total CAMCOG	0.795	2.63E-07¶

Table 55: Significant associations between exposures and outcomes in the COGFAST cohort. § = t test, ¶ = Spearman's rank. Number of tests performed = 476. Statistical threshold = $p<1.05x10^{-4}$.

Several baseline CAMCOG scores were found to be associated with the last CAMCOG score recorded prior to death. A positive correlation was observed between all significant associations between CAMCOG variables. Baseline orientation, language comprehension, language expression, memory remote, perception, MMSE and total CAMCOG scores were associated with several CAMCOG scores recorded at the final follow up prior to death. In all cases, a worse baseline score was associated with a worse score prior to death. A lower baseline total CAMCOG score was also found to be associated with a later diagnosis of dementia. Previous work using the COGFAST cohort had identified significant associations between baseline executive function score (CAMCOG test) and onset of dementia as well as memory total (combined score across all CAMCOG memory tests; recent, remote and learning) and onset of dementia (Allan *et al.*, 2011). None of these baseline CAMCOG scores were found to be associated with later cognitive outcome in the COGFAST samples used in this study. In addition, previously published work using the entire COGFAST cohort identified the number of CVD risk factors to be associated with onset of dementia (Allan *et al.*, 2011), however, this observation was not identified in the cohort sub-sample used in this thesis. Another study using the COGFAST cohort identified hypertension to be associated with cognition at baseline (three months post-stroke) (Rowan *et al.*, 2005), but hypertension was not found to be associated with cognitive outcome prior to death in this study. Although the majority of associations did not attain statistical significance many of the effect estimates were in the right direction. All exposure-outcome relationships were considered for analysis irrespective of whether the more widely reported association was observed in the sub-sample of COGFAST data utilised here, due to the power limitations of replicating such associations in this particular study.

6.3.1.2 ***Mild cognitive impairment in Parkinson's disease***

All measured exposures relating to the lifestyle, anthropometric and medical history previously associated with PD-MCI in the literature were compared to both motor and cognitive outcome variables, to test whether these exposure variables are associated with outcome in the ICICLE cohort. In addition to assessing cognitive decline, associations with motor outcomes, a strong hallmark of PD, were also explored here. All associations between exposures and motor outcomes which reach the stringent significance threshold (as described in Section 6.2.3) are displayed in Table 56. All associations between exposures and cognitive outcomes with p<1.30E-04 are displayed in Table 57. A comprehensive results table can be found in Appendix C.

Exposure Type	Exposure	Outcome	Test statistic	P value
Medical history	GDS	MDS-UPDRS II	0.632	2.13E-14¶
Medical history	GDS	MDS-UPDRS III	0.459	1.97E-07¶
Medical history	GDS	PIGD phenotype	0.553	1.06E-10¶

Table 56: Significant associations between exposures and motor outcomes in the ICICLE cohort. ¶ = Spearman's rank. Number of tests performed = 384. Statistical threshold = $p<1.30\times10^{-4}$.

Of all exposures tested, depression, as measured by the GDS was the only exposure found to be associated with motor outcome when the stringent significance threshold of 1.13×10^{-4} was applied. A higher score on the GDS (where > 5 is categorised as depressed) was associated with a higher score on all three associated motor tests, indicative of a worse motor function, however these results are not able to indicate whether depression is causally linked with motor outcome. Other associations were observed however they did not reach significance once Bonferroni correction was applied. Again, all exposure-motor outcome relationships were considered for analysis irrespective of whether the more widely reported association was observed in the ICICLE data, due to the power limitations of replicating such associations in this particular study.

Exposure Type	Exposure	Outcome	Test statistic	P value
Demographic	Education	MoCA	0.336	0.0001¶
Demographic	Education	MMSE	0.338	0.0001¶
Demographic	Education	Total FAS	0.426	4.25E-07¶
Demographic	Education	Animals	0.392	3.92E-06¶
Demographic	Education	Digit vigilance accuracy	0.386	5.59E-06¶
Demographic	Education	PRM	0.419	7.23E-07¶
Demographic	NART	MoCA	0.369	1.60E-05¶
Demographic	NART	MMSE	0.379	8.88E-06¶
Demographic	NART	Total FAS	0.459	3.87E-07¶
Demographic	NART	PRM	0.401	2.21E-06¶
Demographic	NART	PAL	-0.361	2.49E-05¶
Demographic	NART	Language total	0.327	0.0001¶
Demographic	NART	MCI	4.283	1.84E-05‡
Medical history	Levodopa daily dose	Total number NMS	0.482	1.41E-08¶

Table 57: Significant associations between exposures and cognitive outcomes in the ICICLE cohort. ‡ = Wilcoxon rank sum, ¶ = Spearman's rank. Number of tests performed = 384. Statistical threshold = $p<1.30x10^{-4}$.

Education and NART were found to be associated with a number of cognitive outcomes when the Bonferroni correction was applied. As expected, a higher number of years in education is associated with a better score on all cognitive tests (Table 57). Similarly, a higher NART score is associated with better cognition in all tests, as shown in Table 57. Due to the moderate correlation between education and NART (Rho=0.577, p=3.33E-15), many of these associations do, as expected, overlap. These results are similar to a previous study using the ICICLE cohort which also identified education to be associated with poorer cognitive outcome (Yarnall *et al.*, 2014). A higher Levodopa daily dose is associated with an increased number of non-motor symptoms (NMS). Again, these tests do not give any indication of the causal relationship between the two variables (exposure and outcome), however all exposure variables used in this analysis have been previously associated with an increased risk of PD and are considered possible risk factors for this disease in the literature. Previous studies using the ICICLE cohort found participants with a higher score on the GDS and therefore had a more severe depression were more likely to have PD-MCI (Yarnall *et al.*, 2014). Although some exposures (such as other medical conditions and smoking) were not found to be associated with cognitive outcome in the ICICLE cohort, there is enough evidence in the literature to suggest that all exposures tested could have an impact on disease risk. The effect estimates were in the expected direction in the majority of cases, although the p value may not have attained statistical significance. For these reasons, all exposures were included in subsequent analysis.

6.3.2 Relationship between exposure and DNA methylation

6.3.2.1 *Post-stroke dementia*

ANOVA was performed on all beta values measured against the exposure variables in Table 53. Baseline age, sex and chip were included in the analysis model as covariates when analysing blood DNA samples. Age at death, sex and chip were included as covariates in the analysis model using DLPFC and hippocampal DNA samples.

6.3.2.1.1 Baseline blood samples

To test for association between methylation and exposures, ANOVA was performed on all beta values calculated across 455,173 CpG sites measured in 29 baseline blood samples, against the exposures in Table 53. Manhattan plots are shown for exposure variables for which significant associations were observed with CpG sites in blood DNA. Summary and test statistics for each CpG site significantly associated (reaching genome-wide significance) with the exposure are displayed in the following tables. A number of variables (Side of body - Figure 92, Table 58; Degree of weakness arm, and Degree of weakness leg - Figure 93, Table 59), which can be used to indicate the severity of the stroke, were associated with methylation at several CpG sites. Due to the high degree of correlation identified between degree of weakness in the arm and leg (coefficient = 0.718, p<0.001), results for the degree of weakness in the arm are omitted.

Interestingly, DNA methylation at the same four CpG sites was associated with both degree of weakness in the arm and leg. This may be due to the fact that these variables are closely correlated. The CpG site associated with the side of the body affected by the stroke is unique to that variable and does not appear to be significantly associated with degree of weakness in either the arm or leg. None of these CpG sites associated with an exposure were found to reach genome-wide significance with any outcome in Chapter 3. In addition, none of these CpG sites have previously been associated with stroke or stroke-related variable, although literature in this domain is still very limited.

When the analyses were repeated after adjustment for cellular composition, using the method described by Houseman *et al.* (2012), no large deviations in associations were observed for any CpG site. This indicates that cellular composition is not causing the differences in methylation associated with the exposure measures displayed here.

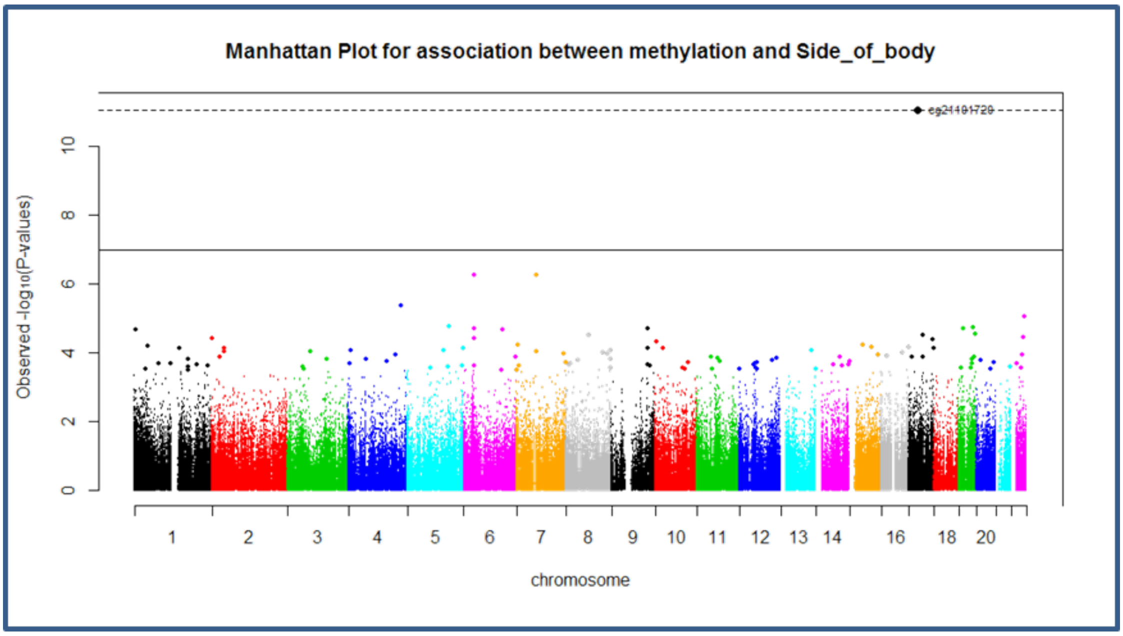


Figure 92: Manhattan plot for association between blood DNA methylation and the side of the body affected by stroke. Each circle represents an individual CpG site. The continuous line marks the $p<1.1x10^{-7}$ significance threshold. CpG sites with p values below the threshold are labelled with their Illumina ID.

CpG site	Gene ID	Relation to CpG island	Effect size	F statistic	P value
cg21101720	*ANKRD13B*	CpG Island	0.14	257.70	9.05E-12

Table 58: Summary and test statistics for the CpG site in blood associated with the side of the body affected by stroke.

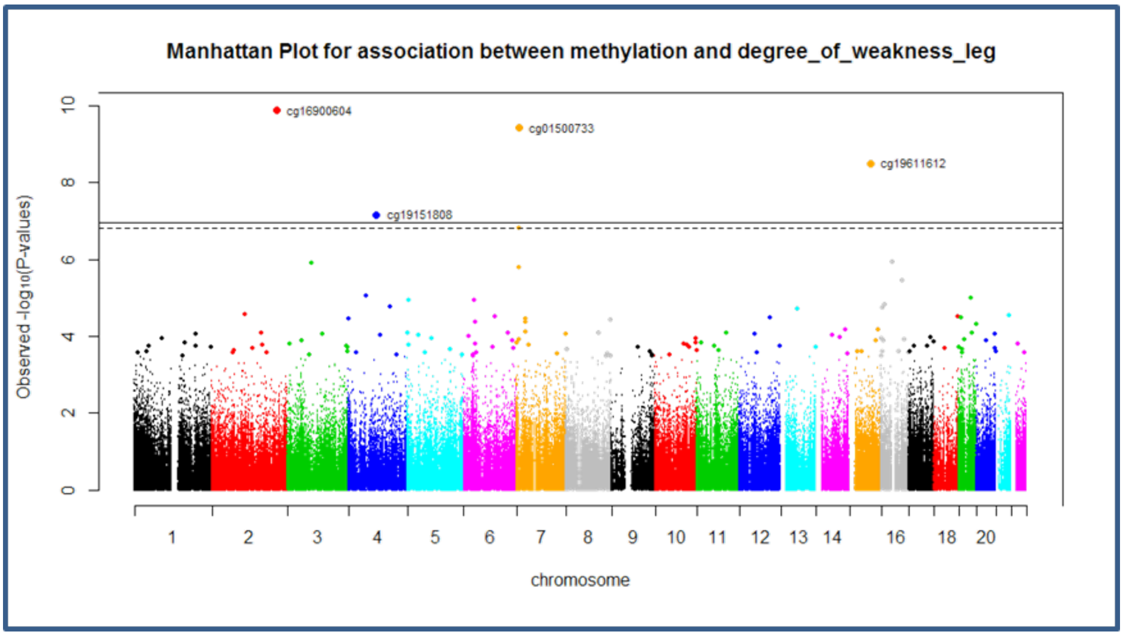


Figure 93: Manhattan plot for association between blood DNA methylation and degree of weakness in the leg. Each circle represents an individual CpG site. The continuous line marks the $p<1.1x10^{-7}$ significance threshold. CpG sites with p values below the threshold are labelled with their Illumina ID.

CpG site	Gene ID	Relation to CpG island	Effect size	F statistic	P value
cg16900604			0.26	260.48	1.30E-10
cg01500733	*C7orf28A*	North shore	0.13	128.19	3.71E-10
cg19611612			0.35	94.53	3.14E-09
cg19151808	*HERC3A*		0.57	60.19	6.82E-08

Table 59: Summary and test statistics for the 4 CpG sites in blood associated with degree of weakness in leg.

6.3.2.1.2 Dorsolateral prefrontal cortex samples

To test for association between methylation and exposures, ANOVA was performed on all beta values calculated across 431,822 CpG sites measured in 24 DLPFC samples, against the exposures in Table 53. Manhattan plots are shown for exposure variables for which significant associations were observed with CpG sites in the DLPFC. Summary and test statistics for each CpG site significantly associated (reaching genome-wide significance) with the exposure are displayed in the following tables. A number of variables (Side of body - Figure 94, Table 60; degree of weakness in arm (not presented) and degree of weakness in leg - Figure 95, Table 61), which indicate the severity of stroke, were significantly associated with methylation at a number of CpG sites. In addition, intermittent claudication was also associated with methylation at five CpG sites (Figure 96, Table 62). Due to the strong correlation between degree of weakness in the arm and leg, only data for the leg are presented.

None of the CpG sites highlighted here have previously been associated with stroke or stroke-related variables. CpG sites associated with intermittent claudication and the side of the body affected by stroke are not significantly associated with any other exposure. None of these CpG sites were found to be associated with an outcome variable in Chapter 4. In addition, none of the CpG sites associated with an exposure in the DLPFC were also identified in the blood.

When the analyses were repeated after adjustment for cellular composition, using the method described by Guintivano *et al.* (2013), no large deviations in associations were observed for any CpG site. This indicates that cellular composition is not causing the differences in methylation associated with the exposure measures displayed here.

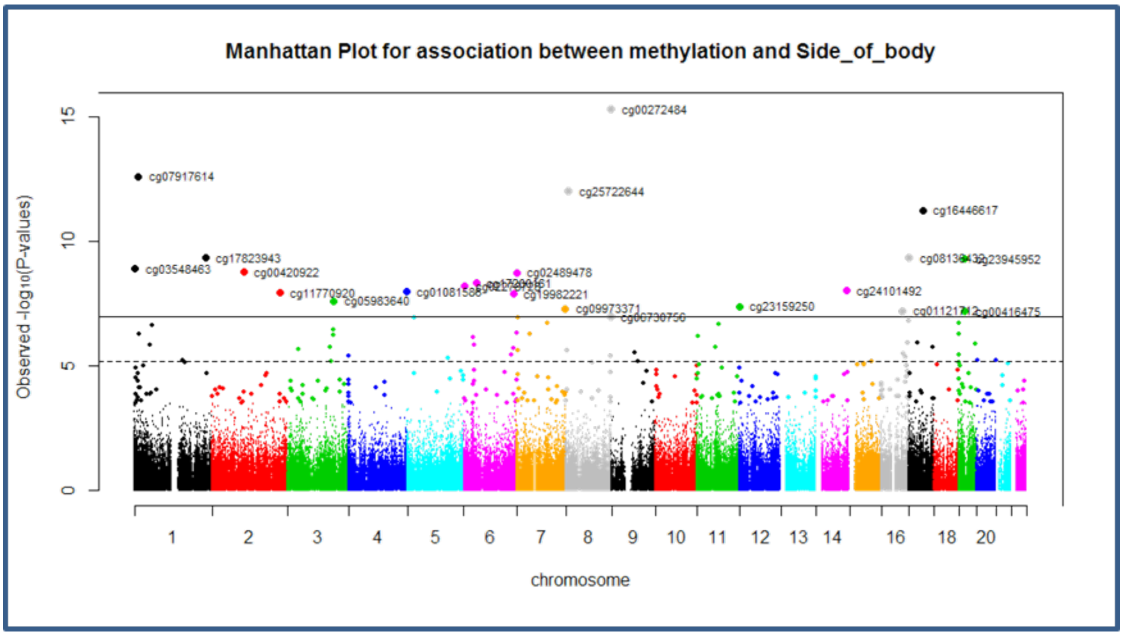


Figure 94: Manhattan plot for association between DLPFC DNA methylation and the side of the body affected by stroke. Each circle represents an individual CpG site. The continuous line marks the p<1.1x10^{-7} significance threshold. CpG sites with p values below the threshold are labelled with their Illumina ID.

CpG site	Gene ID	Relation to CpG island	Effect size	F statistic	P value
cg00272484	*TSNARE1*	South shore	-0.54	5251.05	7.79E-16
cg07917614	*CASZ1*		-0.78	1231.02	1.08E-12
cg25722644	*XKR5*	South shore	-0.58	1146.85	1.54E-12
cg16446617			-0.75	805.42	8.94E-12
cg08136432	*GALNS*	CpG island	-0.65	782.10	6.70E-10
cg17823943	*OBSCN*	North shelf	-0.67	319.79	8.65E-10
cg23945952	*CRTC1*	South shore	-0.13	283.54	1.56E-09
cg03548463	*CDK11B*	South shore	-0.57	267.82	2.07E-09
cg00420922	*TBC1D8*	CpG island	0.21	258.75	2.45E-09
cg02489478		North shore	-0.51	256.78	2.54E-09
cg17200161	*KCNK5*	CpG island	0.09	203.96	7.84E-09
cg02278728		North shelf	-0.34	196.90	9.32E-09
cg00416475	*GMIP*	CpG island	-0.21	180.89	1.41E-08
cg01081586		CpG island	0.12	173.71	1.71E-08
cg24101492		CpG island	0.08	173.65	1.72E-08
cg11770920	*DIRC3*		-0.33	167.68	2.04E-08
cg19982221			-0.35	167.60	2.04E-08
cg05983640	*HPS3*	CpG island	0.11	141.89	4.57E-08
cg01121712	*TPPP3*	CpG island	0.18	134.60	5.89E-08
cg04508405	*ARPC1B*	North shore	-0.13	122.93	9.12E-08
cg06730756	*DENND3*		-0.75	120.59	1.00E-07

Table 60: Summary and test statistics for the 21 CpG sites in DLPFC associated with the side of body affected by stroke.

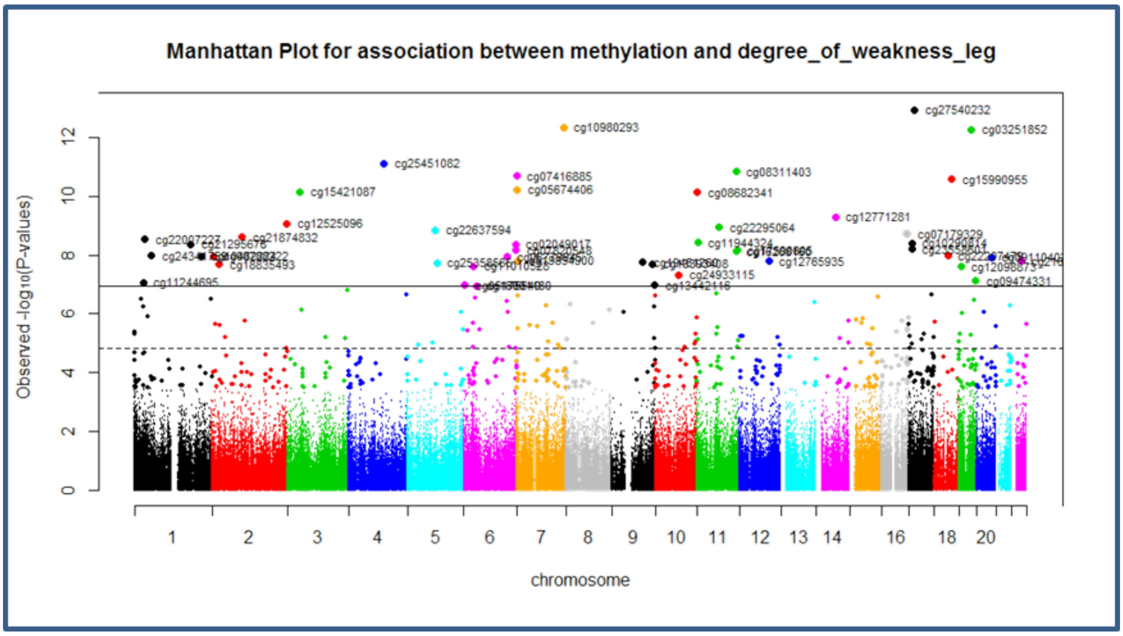


Figure 95: Manhattan plot for association between DLPFC DNA methylation and the degree of weakness in the leg. Each circle represents an individual CpG site. The continuous line marks the p<1.1x10^{-7} significance threshold. CpG sites with p values below the threshold are labelled with their Illumina ID.

CpG site	Gene ID	Relation to CpG island	Effect size	F statistic	P value
cg27540232			0.49	2177.16	6.32E-14
cg10980293	*ASB10*		0.63	1638.83	2.60E-13
cg03251852	*PRX*		0.87	1595.43	2.98E-13
cg25451082	*ARSJ*	South shore	-0.38	900.84	5.12E-12
cg15990955	*NEDD4L*		0.69	891.34	5.40E-12
cg07416885		CpG island	0.54	798.79	9.31E-12
cg08311403	*ROBO3*	CpG island	-0.12	768.09	1.13E-11
cg08682341	*INPP5A*	South shore	-0.51	605.35	3.69E-11
cg05674406	*INTS1*	North shore	0.68	529.30	7.18E-11
cg15421087	*CTNNB1*	CpG island	-0.44	445.81	1.68E-10
cg12771281	*TRMT5*	CpG island	-0.17	592.41	2.80E-10
cg12525096		South shore	0.19	352.17	5.38E-10
cg22295064	*CTTN*	CpG island	-0.08	326.82	7.77E-10
cg10290814	*TNK1*	CpG island	-0.19	306.20	1.07E-09
cg07179329	*CDH13*		0.81	435.93	1.10E-09
cg10482224	*PTPN14*	CpG island	0.27	296.73	1.25E-09
cg22637594		South shelf	-0.36	290.58	1.39E-09
cg22007227	*ZBTB8B*	CpG island	-0.09	272.90	1.89E-09
cg11944324	*TOLLIP*	North shore	0.28	266.29	2.13E-09
cg02049017	*RPS6KA2*	South shelf	0.17	263.24	2.25E-09
cg23558601	*TNK1*	CpG island	-0.20	262.06	2.30E-09
cg21874832	*FAM178B*	CpG island	0.27	252.77	2.75E-09
cg07820548		South shore	0.23	243.78	3.28E-09
cg17590805	*FLI1*	CpG island	0.54	242.90	3.34E-09
cg18835493	*KLHL29*	South shore	0.18	235.47	3.89E-09
cg22207479	*SLC14A2*		0.51	231.50	4.22E-09
cg16268160	*ROBO3*	CpG island	-0.13	221.48	5.24E-09
cg19461260	*TRIM14*	CpG island	0.34	211.23	6.61E-09
cg09110402	*PREX1*	CpG island	-0.10	205.14	7.63E-09
cg00788222		North shelf	0.34	204.82	7.68E-09

cg25358565	*FAM172A*	CpG island	-0.07	201.15	8.39E-09
cg21634064	*SMTN*	CpG island	-0.11	199.70	8.69E-09
cg24341759	*SSBP3*		0.42	198.95	8.86E-09
cg12765935	*METAP2*		0.65	195.79	9.58E-09
cg13393408	*GPR107*		0.22	182.41	1.35E-08
cg12098873	*TRIP10*	North shore	0.50	178.19	1.51E-08
cg11010528	*HCG4P6*	CpG island	-0.28	176.86	1.57E-08
cg19854900	*SDK1*		0.20	171.27	1.84E-08
cg24933115		CpG island	0.07	322.09	2.26E-08
cg21295678			0.12	143.34	4.35E-08
cg06779945	*TNFAIP3*		-0.01	142.58	4.46E-08
cg09474331	*TTYH1*	North shore	-0.09	141.76	4.59E-08
cg13442116	*KCNT1*	North shelf	0.39	140.04	4.87E-08
cg16711291		South shelf	0.55	129.13	7.20E-08
cg18001180	*KCNK17*	CpG island	-0.24	126.34	8.00E-08
cg05175540	*SERPINB9*	North shore	0.25	125.71	8.19E-08
cg11244695	*SLC9A1*		0.10	125.27	8.33E-08

Table 61: Summary and test statistics for the 47 CpG sites in DLPFC associated with degree of weakness in leg.

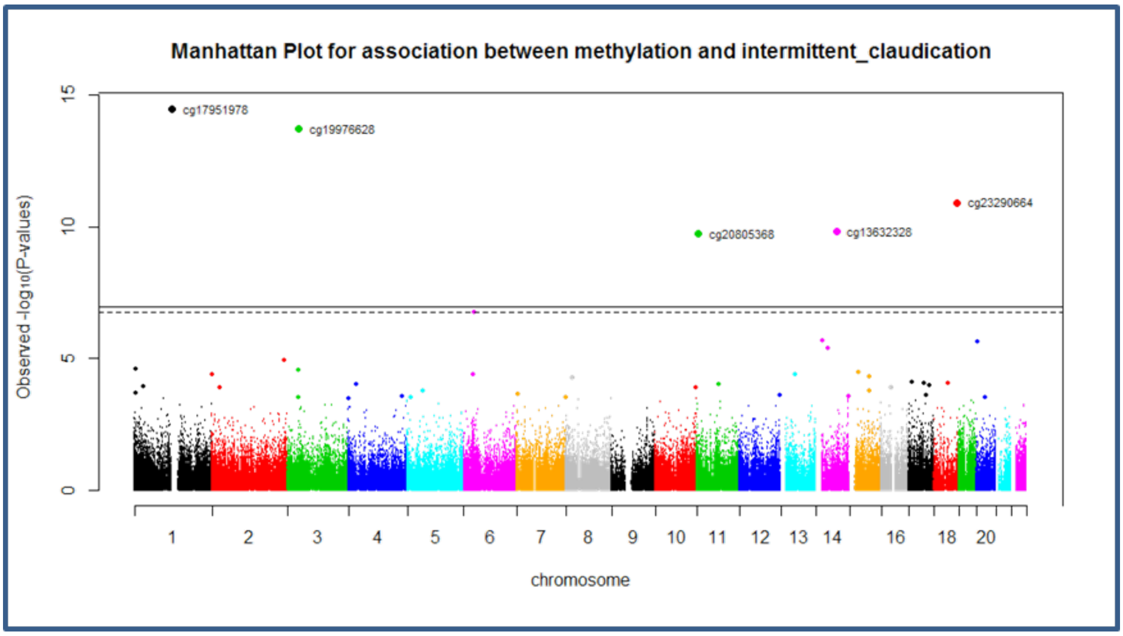


Figure 96: Manhattan plot for association between DLPFC DNA methylation and intermittent claudication. Each circle represents an individual CpG site. The continuous line marks the $p<1.1x10^{-7}$ significance threshold. CpG sites with p values below the threshold are labelled with their Illumina ID.

CpG site	Gene ID	Relation to CpG island	Effect size	F statistic	P value
cg17951978	*HSD3B2*		-0.05	3589.19	3.47E-15
cg19976628		North shelf	-0.05	2618.46	1.95E-14
cg23290664	*TSHZ1*	CpG island	0.71	807.92	1.20E-11
cg13632328		North shore	-0.02	685.41	1.52E-10
cg20805368			-0.05	486.43	1.87E-10

Table 62: Summary and test statistics for the 5 CpG sites in DLPFC associated with intermittent claudication.

6.3.2.1.3 Hippocampus samples

To test for association between methylation and exposures, ANOVA was performed on all beta values calculated across 434,422 CpG sites measured in 28 hippocampal samples, against the exposures in Table 53. Manhattan plots are shown for exposure variables for which significant associations were observed with CpG sites in the hippocampus. Summary and test statistics for each CpG site significantly associated (reaching genome-wide significance) with the exposure are displayed in the following tables. A number of variables (Side of body - Figure 97, Table 63; degree of weakness in arm (not presented) and degree of weakness in leg - Figure 98, Table 64) associated with the severity of stroke were associated with methylation at a number of CpG sites. Again, due to the high correlation between degree of weakness in the arm and leg, only data relating to the leg are presented.

Like in the blood and DLPFC, the CpG sites associated with the side of the body affected by the stroke were not associated with any other exposure in the hippocampus. None of the CpG sites associated with a stroke-related variable here have previously been reported to be associated with stroke or any stroke-related variable. Additionally, none of these CpG sites were also significantly associated with any outcome variable in Chapter 4. No CpG site significantly associated with an exposure in the hippocampus were also associated with an exposure in the blood, however there were a few overlapping CpG sites between the DLPFC and the hippocampus. Two CpG sites (cg12765935 and cg05674406) were associated with both the degree of weakness in the arm and leg in the DLPFC and the degree of weakness in the arm and leg in the hippocampus. One CpG (cg22207479) associated with both degree of weakness in arm and leg in the DLPFC was also associated with the degree of weakness in the arm in the hippocampus.

Adjustment for cellular composition using the method described by Guintivano *et al.* (2013) identified no large deviations in associations observed for any CpG site. This indicates that cellular composition is not causing the differences in methylation associated with the exposure measures displayed here.

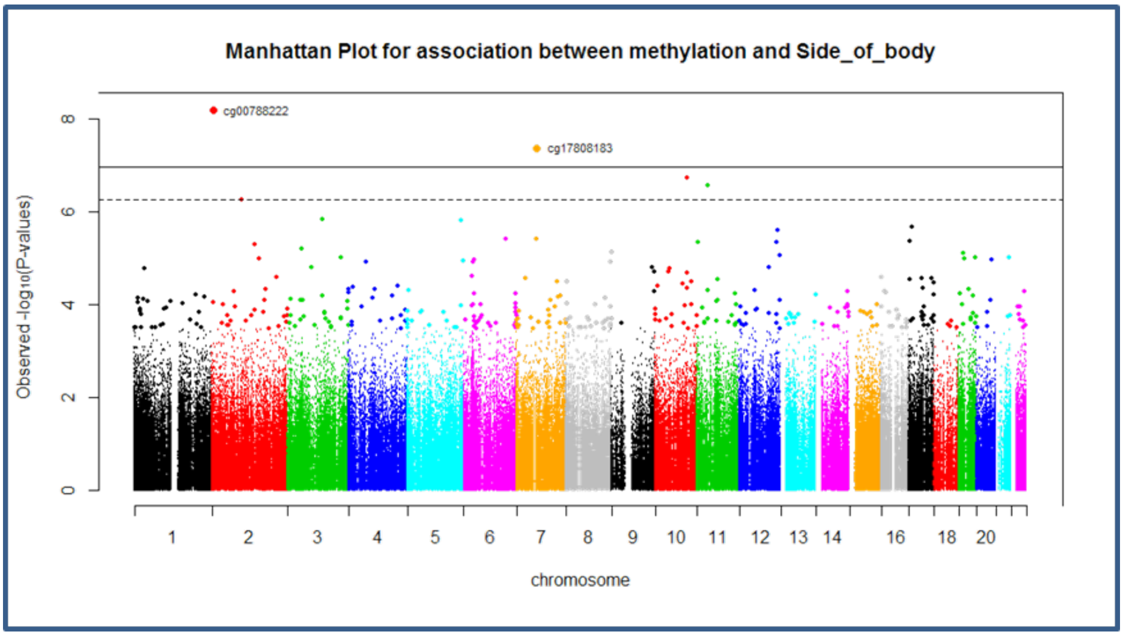


Figure 97: Manhattan plot for association between hippocampal DNA methylation and the side of body affected by stroke. Each circle represents an individual CpG site. The continuous line marks the $p<1.1x10^{-7}$ significance threshold. CpG sites with p values below the threshold are labelled with their Illumina ID.

CpG site	Gene ID	Relation to CpG island	Effect size	F statistic	P value
cg00788222		North shelf	-0.15	96.21	6.60E-09
cg17808183			-0.22	72.04	4.27E-08

Table 63: Summary and test statistics for the 2 CpG sites in the hippocampus associated with the side of body affected by stroke.

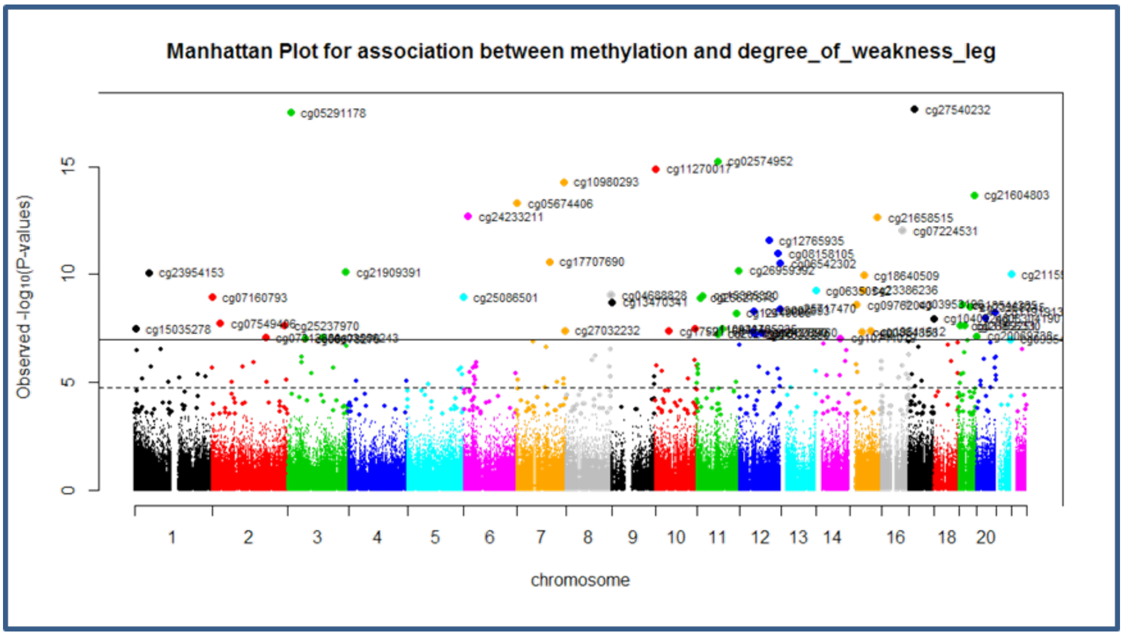


Figure 98: Manhattan plot for association between hippocampal DNA methylation and the degree of weakness in leg. Each circle represents an individual CpG site. The continuous line marks the p<1.1x10^{-7} significance threshold. CpG sites with p values below the threshold are labelled with their Illumina ID.

CpG site	Gene ID	Relation to CpG island	Effect size	F statistic	P value
cg27540232			0.60	2324.30	2.20E-18
cg05291178	*C3orf24*	North shore	0.65	2207.15	3.16E-18
cg02574952	*DPF2*	South shelf	0.93	1050.60	5.56E-16
cg11270017	*ADARB2*	North shelf	0.75	931.43	1.28E-15
cg10980293	*ASB10*		0.58	760.80	5.24E-15
cg21604803	*CPT1C*	CpG island	-0.33	623.05	2.09E-14
cg05674406	*INTS1*	North shore	0.68	550.95	4.89E-14
cg24233211			0.18	449.80	1.98E-13
cg21658515		South shelf	0.61	442.46	2.22E-13
cg07224531	*B3GNT9*	CpG island	0.28	359.41	9.29E-13
cg12765935	*METAP2*		0.47	307.98	2.68E-12
cg08158105	*BRI3BP*	North shore	0.45	539.40	1.05E-11
cg17707690	*ORC5L*		-0.12	221.38	2.54E-11
cg06542302		South shore	0.37	216.08	3.00E-11
cg26959392	*SPATA19*		0.38	192.45	6.56E-11
cg21909391	*LPP*		0.38	187.99	7.69E-11
cg23954153	*ARTN*	CpG island	-0.19	186.35	8.15E-11
cg21159390		CpG island	0.20	180.78	1.00E-10
cg18640509	*SORD*	CpG island	-0.23	178.42	1.09E-10
cg23386236	*CAPN3*		0.55	140.54	5.41E-10
cg06350542	*MCF2L*	North shelf	0.23	505.30	5.70E-10
cg04688828	*LYNX1*	CpG island	-0.15	129.58	9.29E-10
cg15365320	*DKFZp686O24166*	CpG island	-0.20	128.87	9.63E-10
cg07160793			0.50	125.64	1.14E-09
cg25086501		North shore	0.25	125.41	1.15E-09
cg25627675	*RPL27A*	North shelf	0.69	124.09	1.24E-09
cg13470341	*DMRT3*	CpG island	-0.09	115.79	1.96E-09
cg03953196	*COX4NB*		0.12	113.19	2.27E-09
cg09762043	*TUBGCP5*	CpG island	-0.26	111.07	2.57E-09
cg18544365	*DNM2*		0.54	110.48	2.67E-09

cg23365345	*ZNF792*	CpG island	-0.05	107.61	3.17E-09
cg25717470			0.21	104.58	3.82E-09
cg19008693	*VDR*		0.17	100.31	5.02E-09
cg11919138	*ATP5E*	CpG island	-0.11	98.48	5.67E-09
cg12449685	*ROBO4*		-0.39	134.03	6.19E-09
cg08304190	*MIR663*	CpG island	-0.39	89.62	1.05E-08
cg10409248		CpG island	-0.17	88.99	1.10E-08
cg07549406	*CGREF1*	CpG island	-0.09	81.61	1.92E-08
cg03162251	*MLLT1*	CpG island	0.74	106.15	2.34E-08
cg25237970	*DIS3L2*		0.17	78.95	2.38E-08
cg26856330	*NR2C2AP*	CpG island	-0.06	78.85	2.40E-08
cg21765235	*ODZ4*		-0.31	74.98	3.31E-08
cg11094040	*DOCK1*	CpG island	0.12	74.77	3.37E-08
cg15035278	*PRDM16*	North shelf	-0.10	74.31	3.50E-08
cg27032232	*DPP6*	CpG island	-0.23	72.51	4.10E-08
cg19611612			0.54	72.14	4.24E-08
cg17521156		CpG island	0.40	71.65	4.42E-08
cg00834858		North shelf	0.03	71.09	4.65E-08
cg01182690	*PPHLN1*	North shore	-0.22	70.99	4.69E-08
cg24373760	*LGR5*	CpG island	-0.05	69.96	5.15E-08
cg20210376	*LTBP3*	CpG island	0.21	68.84	5.71E-08
cg04033850		South shelf	0.13	67.95	6.20E-08
cg20069738	*ZBTB45*	CpG island	-0.05	66.25	7.28E-08
cg07312880			-0.46	64.37	8.73E-08
cg03556243	*ZBTB20*		-0.19	64.34	8.75E-08
cg10744079		CpG island	-0.91	63.98	9.07E-08
cg06678279	*ERC2*	CpG island	-0.04	63.67	9.35E-08
cg09354050	*UBASH3A*		0.09	62.84	1.02E-07

Table 64: Summary and test statistics for the 58 CpG sites in the hippocampus associated with degree of weakness in leg.

6.3.2.2 *Parkinson's disease*

ANOVA was performed on all beta values across 455,981 CpG sites against the exposure measures described in Table 54. Age, sex and chip were included in the analysis model as covariates.

Manhattan plots are shown for exposure variables, for which significant associations were observed with CpG sites in blood. Summary and test statistics for each CpG site significantly associated (reaching genome-wide significance) with the exposure are displayed in the following tables. One CpG site was associated with homocysteine levels (Figure 99, Table 66), 8 CpG sites were associated with IHD (Figure 100, Table 67), 24 CpG sites were associated with smoking status (Figure 101, Table 68) and 7 CpG sites were associated with weight (Figure 102, Table 69).

No CpG sites were identified as being significantly associated with more than one exposure. Interestingly five of the 24 CpG sites significantly associated with smoking were within the *GATA3* gene, however none of the CpG sites associated with smoking here have previously been reported to be differentially methylated between smokers and non-smokers. It may be that the sample size used here (with only two current smokers) was too small to detect the usual hits associated with smoking such as *AHRR* (Shenker *et al.*, 2013; Zeilinger *et al.*, 2013). A power calculation was performed to calculate the number of smokers required to detect methylation differences at the *AHRR* locus. To detect a methylation difference of 5% with alpha=$1.1\text{x}10^{-7}$, 26 smokers would be required.

One of the CpG sites associated with smoking is located within the *DLG2* gene which has previously been associated with Parkinson's disease. No other CpG site associated with an exposure in the ICICLE cohort has previously been associated with the exposure in the literature. In addition, no CpG sites identified here also reached genome wide significance when an association between methylation and an outcome variable was tested in Chapter 5.

When the CpGassoc analyses were repeated on data adjusted for cellular composition (Houseman *et al.*, 2012), very few differences were observed. No CpG site associated with weight, IHD or homocysteine were affected by cellular composition adjustment.

Two CpG sites associated with smoking status showed very slight alterations (Table 65). Adjusting for cellular composition very slightly increased the significance level and F statistic at both CpG sites. These results suggest that cellular composition is not a major influence on methylation at any CpG site associated with any of the exposures highlighted in the ICICLE cohort.

Variable	CpG site	Pre-adjustment		Post-adjustment	
		F statistic	P value	F statistic	P value
Smoking status	cg09769113	30.01	3.75E-08	30.15	3.57E-08
Smoking status	cg04318855	29.18	5.05E-08	31.47	2.25E-08

Table 65: CpG sites affected by cellular composition.

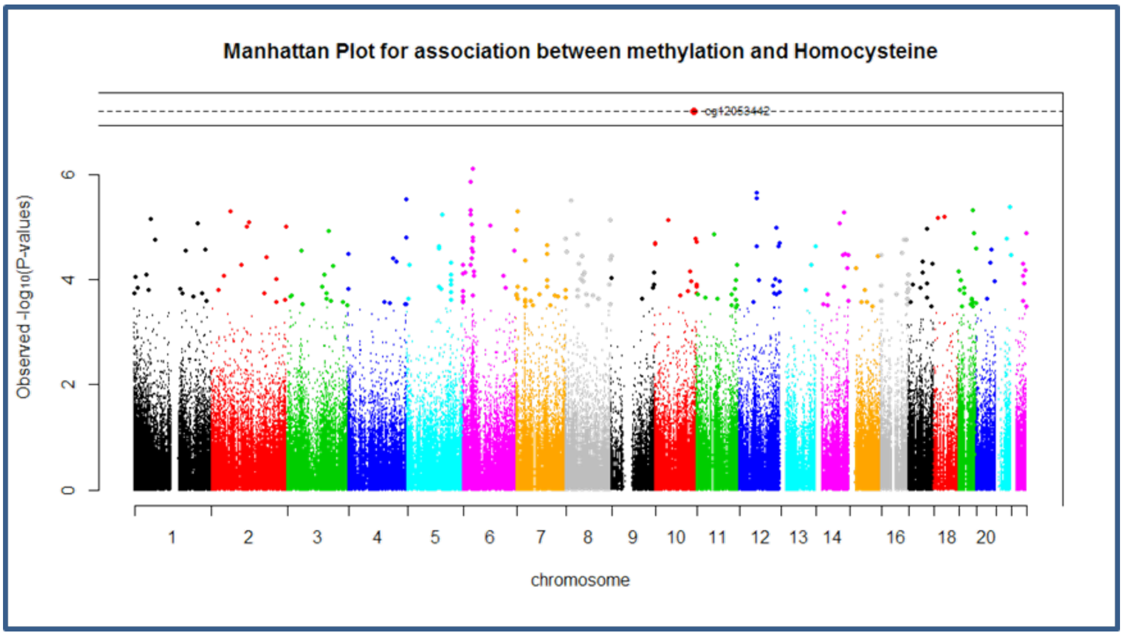


Figure 99: Manhattan plot for association between blood DNA methylation and homocysteine. Each circle represents an individual CpG site. The continuous line marks the p<1.1x10^{-7} significance threshold. CpG sites with p values below the threshold are labelled with their Illumina ID.

CpG site	Gene ID	Relation to CpG island	Effect size	T statistic	P value
cg12053442			0.44	-6.89	6.20E-08

Table 66: Summary and test statistics for the CpG site associated with homocysteine.

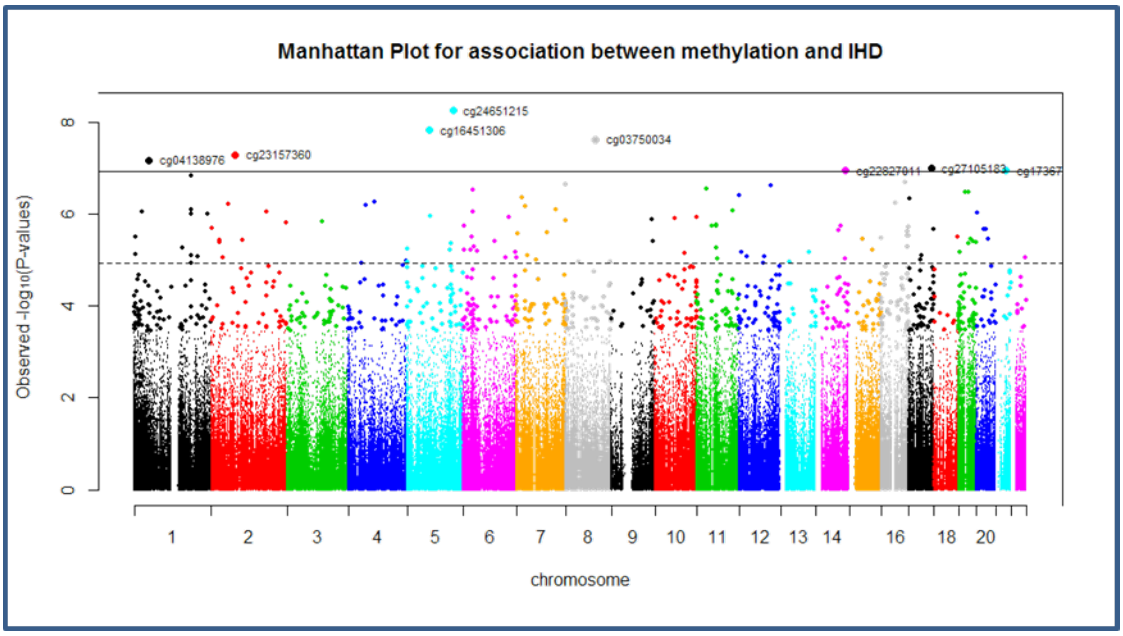


Figure 100: Manhattan plot for association between blood DNA methylation and IHD. Each circle represents an individual CpG site. The continuous line marks the $p<1.1x10^{-7}$ significance threshold. CpG sites with p values below the threshold are labelled with their Illumina ID.

CpG site	Gene ID	Relation to CpG island	Effect size	F statistic	P value
cg24651215		North shelf	0.15	58.57	5.59E-09
cg16451306	*MAP1B*		0.19	53.51	1.50E-08
cg03750034			0.19	52.05	2.39E-08
cg23157360			0.22	47.62	5.13E-08
cg04138976	*ATPAF1*	South shore	0.15	46.35	6.77E-08
cg27105183		South shore	-0.12	44.55	1.01E-07
cg17367472	*KRTAP11-1*		0.17	44.12	1.11E-07
cg22827011	*SERPINA6*		0.23	44.07	1.12E-07

Table 67: Summary and test statistics for the 8 CpG sites associated with IHD.

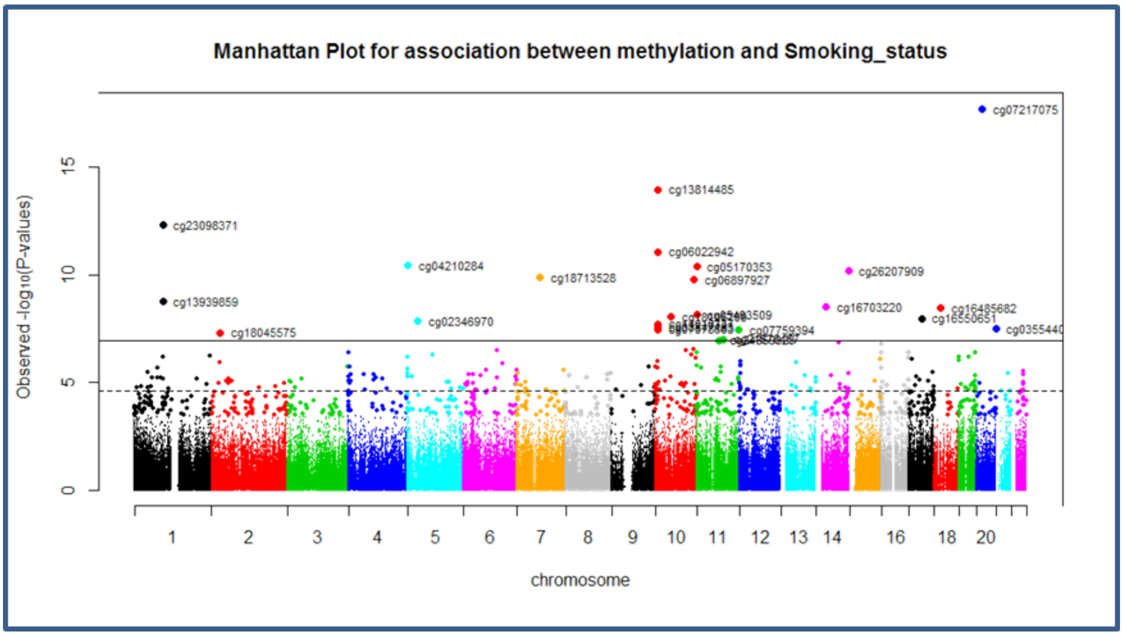


Figure 101: Manhattan plot for association between blood DNA methylation and smoking status. Each circle represents an individual CpG site. The continuous line marks the p<1.1x10^{-7} significance threshold. CpG sites with p values below the threshold are labelled with their Illumina ID.

CpG site	Gene ID	Relation to CpG island	Effect size	F statistic	P value
cg07217075	*BFSP1*	CpG island	0.05	178.25	2.05E-18
cg13814485	*GATA3*	CpG island	0.07	99.14	1.11E-14
cg23098371		CpG island	-0.28	75.46	4.88E-13
cg06022942	*GATA3*	CpG island	0.08	60.76	8.65E-12
cg04210284	*SLC6A3*	CpG island	0.21	54.19	3.75E-11
cg05170353		North shelf	-0.10	53.60	4.31E-11
cg26207909	*CKB*	North shore	-0.04	51.79	6.62E-11
cg18713528	*SRRM3*	CpG island	0.11	48.93	1.34E-10
cg06897927	*HMX3*	CpG island	0.12	47.98	1.71E-10
cg13939859	*BARHL2*	CpG island	0.13	39.52	1.74E-09
cg16703220			-0.30	37.51	3.18E-09
cg16485682	*GATA6*	CpG island	0.13	37.18	3.52E-09
cg05493509	*STK32C*	CpG island	-0.26	34.85	7.31E-09
cg18105709	*C10orf71*	CpG island	-0.03	34.25	8.88E-09
cg16550651	*KAT2A*	North shore	-0.17	33.64	1.09E-08
cg02346970	*PDZD2*	CpG island	0.04	32.82	1.42E-08
cg11018337	*GATA3*	CpG island	0.09	31.96	1.90E-08
cg08347183	*GATA3*	CpG island	0.06	30.87	2.77E-08
cg03554406	*SS18L1*	CpG island	-0.05	30.43	3.23E-08
cg07578663	*GATA3*	CpG island	0.03	30.25	3.45E-08
cg07759394	*GLB1L2*	CpG island	0.18	30.00	3.75E-08
cg18045575	*CGREF1*	CpG island	0.08	29.18	5.05E-08
cg21578457	*DLG2*		-0.16	27.32	1.00E-07
cg24359323	*FGF3*	CpG island	0.10	27.03	1.12E-07

Table 68: Summary and test statistics of the 24 CpG sites associated with smoking status.

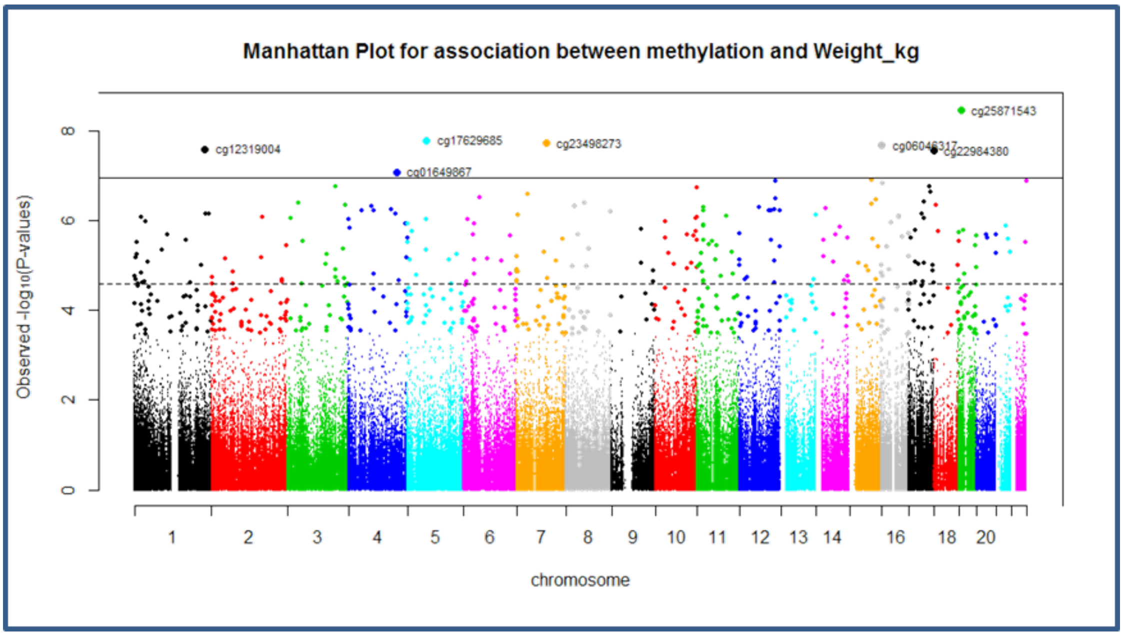


Figure 102: Manhattan plot for association between blood DNA methylation and weight. Each circle represents an individual CpG site. The continuous line marks the $p<1.1x10^{-7}$ significance threshold. CpG sites with p values below the threshold are labelled with their Illumina ID.

CpG site	Gene ID	Relation to CpG island	Effect size	T statistic	P value
cg25871543	*XAB2*	CpG island	0.28	-7.81	3.55E-09
cg17629685	*NDUFAF2*	CpG island	-0.08	7.28	1.65E-08
cg23498273	*SGCE*	North shore	-0.10	7.24	1.90E-08
cg06046317		North shore	0.42	-7.20	2.14E-08
cg12319004	*LEFTY1*	South shelf	0.55	-7.13	2.58E-08
cg22984380	*RPTOR*	South shelf	0.38	-7.11	2.77E-08
cg01649867			0.13	-6.73	8.46E-08

Table 69: Summary and test statistics for the 7 CpG sites associated with weight.

6.3.3 Is DNA methylation a good predictor of exposure status?

If DNA methylation is to be used as a robust predictor of exposure status, it must be measured in a readily accessible tissue such as the blood. Analysis in this Section was therefore limited to blood DNA methylation in both ICICLE and COGFAST.

In the ICICLE cohort, DNA methylation was found to be significantly associated with IHD, weight, homocysteine and smoking (Section 6.3.2.2). As previously mentioned, DNA methylation has previously been used to predict smoking status. The same approach as used by Elliott *et al.* (2014) was utilised here to demonstrate whether a methylation score could be used to distinguish between smokers and never smokers in ICICLE.

Methylation scores were calculated for each ICICLE participant with HM450 methylation data (n=46). Using random forests, the average threshold score segregating current smokers from never smokers in 500 trees was 13.66. This score distinguished current smokers from never smokers (Figure 103). The majority of former smokers had a methylation score below 13.66 and were indistinguishable from never smokers. This supports previously reported data that suggests that following smoking cessation, methylation levels return, over time, to similar levels found in never smokers (Zeilinger *et al.*, 2013; Elliott *et al.*, 2014).

Two ICICLE participants had unknown smoking status, however their prior smoking exposure could be predicted using the methylation scores generated. Both subjects had a score below 0 (-0.23 and -1.92) indicating that these were likely to be either never or former smokers. The smoking score of 13.66 is of similar magnitude to the score of 17.55 identified by Elliott *et al.* (2014), indicating that using the methylation data published by Zeilinger *et al.* (2013) as a reference, a methylation score can be used to successfully distinguish between current and non-smokers in this data set.

No difference was observed between smoking score and self-reported smoking with respect to the association with outcomes in this study population.

Homocysteine, IHD and weight were also found to be associated with DNA methylation and although DNA methylation could potentially refine the exposure status of each of these variables, none were taken forward.

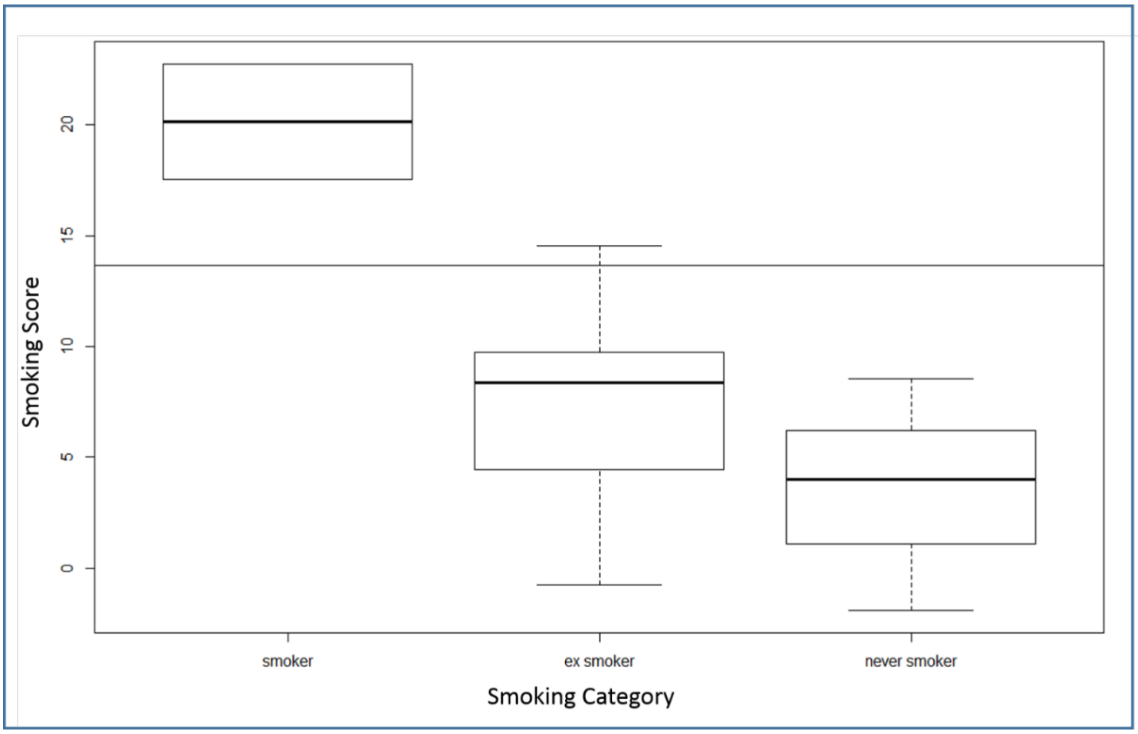


Figure 103: Smoking scores by reported smoking status (current, ex, never). Boxplots show the median and interquartile ranges. The line at 13.66 is the smoking score threshold separating smokers from non-smokers. Individuals above the line were considered to be smokers.

The analysis looking for associations between COGFAST blood DNA methylation and exposures (Section 6.3.2.1.1) yielded several useful candidates; Side of body affected by stroke, degree of weakness in arm and degree of weakness in leg. In this study, these variables were directly recorded from clinical notes and therefore would not have benefitted from the addition of DNA methylation data.

6.3.4 Is methylation associated with both an exposure and outcome?

It was of interest, in this Chapter, to investigate whether the CpG sites associated with an exposure were also associated with an outcome variable (as well as the reverse; whether CpG sites associated with an outcome were also associated with an exposure variable) and whether there was any interaction between methylation and the exposure to affect outcome.

6.3.4.1 *Parkinson's disease*

In Section 6.3.2.2, methylation at a number of CpG sites was found to be associated with a range of exposures (Homocysteine, 1 CpG site; IHD, 8 CpG sites; Smoking, 24 CpG sites; Weight, 7 CpG sites). For each exposure, the CpG sites reaching genome-wide significance were analysed using ANOVA to look for association between methylation and motor outcome (Hoehn and Yahr, MDS-UPDRS II, MDS-UPDRS III, tremor dominant phenotype and PIGD phenotype), and methylation and cognitive outcome (MoCA, MMSE, total FAS, animals, NMSQ memory, NMSQ concentration, power of attention, digit vigilance accuracy, PRM, PAL, SRM, OTS, pentagon copying, naming, language, language total, cognitive complaint, MCI and total number of NMS). No association was observed between an outcome variable and any CpG site associated with homocysteine, IHD and weight. Of the 24 CpG sites associated with smoking, nine of these were also associated with language (part of the MoCA test where participants have to verbally repeat sentences). Table 70 summarises these results.

CpG site	**Gene ID**	**Relation to CpG island**	**Smoking**			**Language**		
			Effect size	F stat	P value	Effect size	T stat	P value
cg18045575	*CGREF1*	Island	0.08	29.18	5.05E-08	0.03	-4.11	2.29E-04
cg11018337	*GATA3*	Island	0.09	31.96	1.90E-08	0.04	-4.05	2.69E-04
cg13814485	*GATA3*	Island	0.07	99.14	1.11E-14	0.02	-3.74	6.62E-04
cg06022942	*GATA3*	Island	0.08	60.76	8.65E-12	0.02	-3.62	9.11E-04
cg23098371		Island	-0.28	75.46	4.88E-13	-0.07	3.56	0.001
cg02346970	*PDZD2*	Island	0.04	32.82	1.42E-08	0.01	-3.53	0.001
cg26207909	*CKB*	North shore	-0.04	51.79	6.62E-11	-0.01	3.40	0.002
cg13939859	*BARHL2*	Island	0.13	39.52	1.74E-09	0.04	-3.33	0.002
cg24359323	*FGF3*	Island	0.10	27.03	1.12E-07	0.03	-3.23	0.003

Table 70: 9 CpG sites associated with both smoking and language.

Regression was used to highlight an association between smoking and language (F = 3.48, p = 0.041). At each CpG site, methylation was found to be associated with smoking (Table 71).

CpG site	**F statistic**	**P value**
cg18045575	11.02	0.002
cg11018337	-12.91	0.003
cg13814485	9.54	0.004
cg06022942	7.12	0.011
cg23098371	5.30	0.027
cg02346970	10.48	0.002
cg26207909	4.25	0.046
cg13939859	5.38	0.025
cg24359323	5.47	0.024

Table 71: Methylation is associated with smoking at 9 CpG sites.

An interaction term was used in the regression model to test for an interaction between methylation and smoking. There was no evidence that the presence or absence of smoking changes the effect of methylation on language at any CpG site. Figure 104 shows the results for cg18045575. Results for all other CpG sites were very similar and are therefore not presented.

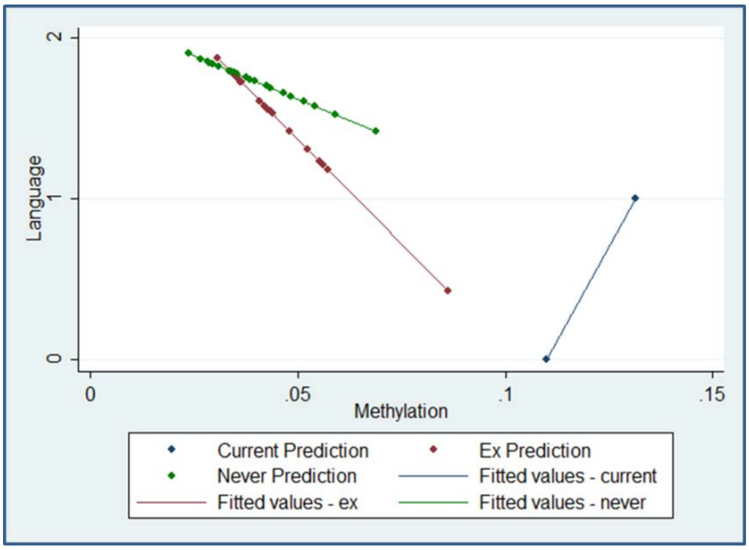


Figure 104: Graph of predicted values and linear fit lines from regression model for cg18045575. Blues points indicate current smokers, green indicates ex-smokers and red indicates never smokers.

CpG sites which were found to be associated with an outcome (Chapter 5) were analysed using ANOVA to look for association between methylation and an exposure variable. MDS-UPDRS III (a variable assessing the motor symptoms of PD) was found to be associated with methylation at one CpG site (T statistic = -7.06, p = 3.20E-08). Analysis using ANOVA revealed that this CpG site was also significantly associated with weekly alcohol consumption (T statistic = 2.49, p = 0.018), however no association was observed between MDS-UPDRS III and alcohol, so this CpG site was not investigated further.

In conclusion, there was no evidence that the presence or absence of any exposure changes the effect of methylation on any outcome variable in the ICICLE cohort.

6.3.4.2 *Post-stroke dementia*

The same methodology was applied to the post-stroke dementia data. Table 72 summarises the observed associations in each tissue. Two CpG sites were associated with both an exposure and outcome in the blood, whilst one CpG site was found to be associated with both an exposure and outcome variable using the DNA extracted from the hippocampus. No associations were observed between DLPFC DNA methylation in both an exposure and outcome.

Tissue	CpG site	Exposure	Exposure - Methylation Statistic	Outcome	Methylation - Outcome Statistic	Exposure - Outcome Statistic
B	cg07438999	Side of body	F=44.92, p=3.94E-08	Braak staging	F=17.00, p=1.40E-04	F=0.84, p=0.44
B	cg02490189	No of CVD risk factors	t=2.13, p=0.049	Total CAMCOG score	t=8.78, p=6.33E-08	F=-2.51, p=0.126
H	cg22666015	Degree of weakness in arm	F=4.45, p=0.032	Recent memory	t=-8.88, p=8.60E-08	F=6.19, p=0.007

Table 72: Statistics for associations between exposures and outcomes identified in COGFAST samples. Tissue – B=Blood, H=Hippocampus. For each association statistic, the test statistic and p value are presented.

In both CpG sites identified in the blood as being associated with both an exposure and outcome, there was no observed association between exposure and outcome and therefore no possible interaction between DNA methylation and exposure. As an association was observed between the degree of weakness in the arm (exposure) and recent memory (outcome) in the hippocampus, multiple regression was performed to assess the association between exposure and outcome, with methylation and an interaction term included in the model. However, there was still a significant association between exposure and outcome (F = 10.09, p=0.0002) indicating that the effect of methylation is not altered by the presence or absence of degree of weakness in the arm. However, due to the small sample size, the robustness of this observation was questionable.

In conclusion, there was no evidence that the presence or absence of any exposure changes the effect of methylation on any outcome variable in the COGFAST cohort.

6.4 Discussion

The main aim of this Chapter was to identify whether DNA methylation was involved in any pathways linking an exposure or risk factor of PSD or PD with disease (outcome variable). Due to the small sample size of both cohorts, few exposures were found to be associated with outcome. However, due to a wealth of literature suggesting that many of the exposures measured in these cohorts are in fact associated with either PSD or PD, they were all considered as possible pathways acting through methylation.

DNA methylation was found to be associated with several exposure variables including variables relating to the severity of stroke (side of body affected by stroke, degree of weakness in arm and leg and intermittent claudication) in all tissues analysed (blood, DLPFC and hippocampus) in the COGFAST cohort and a range of variables (weight, IHD, homocysteine and smoking) in the ICICLE cohort. Associations between DNA methylation at two CpG sites and exposure variables were of particular interest. One of the CpG sites found to be associated with smoking in the ICICLE cohort was located within the *DLG2* gene. *DLG2* encodes a protein that has been shown to be involved with the formation of membrane-associated protein scaffolds at postsynaptic synapses (Kim et al 1996). It is thought that a SNP within the *DLG2* gene may increase the risk of PD (Fung et al 2006). In addition, one of the CpG sites associated with weight in the ICICLE cohort was within the *RPTOR* gene, a component of the signalling pathway regulating cell growth in response to nutrient, hormone and insulin levels (Kim *et al.*, 2002). *RPTOR* has also previously been associated with obesity (Berndt *et al.*, 2013).

DNA methylation had previously been associated with a range of outcome variables affecting cognition (and motor control) in PD. This suggested that methylation may be involved in a causal pathway linking exposure with disease phenotype. However, in both the PSD and PD cohorts utilised in this study, DNA methylation was not found to interact with any exposure to affect outcome. It may be that DNA methylation is not involved in the pathway linking exposure and outcome, or at least in mediating the effect of the limited number of exposures considered in this study.

Another aim of this Chapter was to determine whether DNA methylation could be used as an exposure indicator. Current evidence suggests that DNA methylation could provide a more refined measure of smoking exposure (Shenker *et al.*, 2013; Elliott *et* *al.*, 2014) and this technique could be applied in a similar way to measure the exposure status of other variables. This approach is particularly relevant to exposures which are difficult to accurately measure or exposures which are susceptible to recall or reporting bias. The aim of this Chapter was to use DNA methylation data to predict exposure status of such variables.

In the ICICLE cohort, DNA methylation was found to be associated with smoking. As this had already been studied and had successfully shown DNA methylation differences between smokers and non-smokers, the approach used by Elliott *et al.* (2014) was applied to the ICICLE cohort to test whether this method could distinguish smokers from non-smokers in the ICICLE cohort. The data presented in this Chapter shows that DNA methylation can be used to predict smoking status, with current smokers exhibiting a different methylation profile to never and ex-smokers, supporting previous research that suggests over time, the methylation profiles of ex-smokers reverts back to that of never smokers (Zeilinger *et al.*, 2013; Elliott *et al.*, 2014). This approach was also used to predict the smoking status of two participants whose smoking status was unknown. This approach does have some limitations however. The methylation score was only able to detect between current and never smokers and did not account for the number of cigarettes smoked. The addition of more detailed information such as pack years (the number of packs of cigarettes smoked per day multiplied by the number of years the person has smoked) would enable a more accurate methylation score to be calculated. Furthermore, ex smokers were regarded as ex smokers regardless of the time since smoking cessation. The addition of this information would also provide a more accurate calculation of DNA methylation score. Another factor that has not been accounted for in this model is passive smoking. A non smoker who has been subject to heavy passive smoking may exhibit a methylation profile different to a non smoker who has not been subject to any passive smoking, which could confound results (Elliott *et al.*, 2014). Before being used as a marker of smoking status, further validation is required to assess the effect of passive smoking on DNA methylation, in addition to more detailed information regarding smoking habits and time since since cessation.

In addition, associations between DNA methylation and IHD, and DNA methylation and weight were also observed in the ICICLE cohort. However, neither IHD nor weight could be estimated with more precision using DNA methylation data, as phenotypic data was complete, but methylation data could in theory, be used to predict missing data, should this be required. DNA methylation was also observed to be associated with homocysteine levels. Homocysteine was a possible candidate, as the accurate measurement of homocysteine relies on the quick sample processing of plasma samples (collected in EDTA tubes) to be immediately (within a matter of hours) centrifuged (Salazar *et al.*, 1999) upon collection to prevent the deterioration of the analyte. It is possible that DNA methylation data may therefore provide a better estimate for homocysteine status. However, due to the unavailability of a reference data set and a small sample size, which would have made results produced from separating the ICICLE cohort into a training and test set unreliable, homocysteine was not considered a suitable candidate for implementation of this approach on this occasion.

Using COGFAST data, DNA methylation was found to be associated with variables relating to stroke severity. In COGFAST, all of these variables were well recorded and DNA methylation data were not required to refine exposure status. However, it is important to note that DNA methylation data could be useful in the measurement of these stroke-related variables in certain situations, for example, in cohorts where these detailed clinical data were not available. If the biomarker is sufficiently robust, the methylation information can be used to predict these clinical characteristics. Methylation data can therefore be used to both refine exposure measures, but also in some instances, to provide a proxy or surrogate measure where the information is not available/has not been collected.

The ability to predict smoking status using DNA methylation data is a very valuable resource with great clinical utility due to the smoking-associated reporting bias. The success of this approach using DNA methylation to predict smoking status suggests that this could also be useful for estimating other exposure levels which are prone to the same reporting bias or are difficult to accurately measure. Using a larger sample set, so that samples can be split into a training and a testing set, would enable a wider range of exposures to be assessed in this way.

Analyses reported here are likely to be severely limited in their power to discern associations due to the small sample size. A much larger study sample is required to resolve these outstanding questions and the work presented here should be considered exploratory.

Chapter 7. Methylation age as a risk factor for post-stroke dementia and Parkinson’s disease

7.1 Introduction

Epigenetic factors are thought to be heavily involved in the ageing process with growing evidence of differential methylation seen in elderly subjects (Fraga *et al.*, 2005; Talens *et al.*, 2012), changes in chromatin organization with age (Oberdoerffer and Sinclair, 2007) and evidence suggesting that DNA damage can accelerate the ageing process (Campisi and Vijg, 2009). However, it is unclear whether the effect of this accumulation of DNA damage on ageing is tissue-specific (Campisi and Vijg, 2009). Due to associations between epigenetics and ageing it has been suggested that DNA methylation could be used to predict the age of certain tissues and cell types. Due to the tissue-specific nature of DNA methylation and the finding that ageing may affect DNA methylation profiles in a tissue-specific manner, it is thought that the DNA methylation age of different tissues may vary (Horvath, 2013).

A method to predict age using DNA methylation has been developed (Horvath, 2013). This method, termed the “Epigenetic clock” predicts the DNA methylation age of tissues and cell types by using a weighted average methylation value of 353 CpG sites. Using a calibration function, the weighted average can then be transformed to DNA methylation age. The method is able to predict methylation age to a high accuracy, the median error rate is just 3.6 years (Horvath, 2013). This method can be used to predict DNA methylation age in heterogeneous tissue as well as single cell types. It is also thought that the DNA methylation age is not affected by differences in cellular composition making it a suitable resource for predicting DNA methylation of heterogeneous tissue, such as whole blood and brain tissue. It is unclear whether the DNA methylation of easily accessible tissue such as blood and saliva could be used as a surrogate for more inaccessible tissues such as brain tissue (Horvath, 2013). The epigenetic clock method has been used to predict the DNA methylation age of cancer samples. Each affected tissue showed significantly increased age acceleration with an average of 36.2 years, meaning that the DNA methylation age was predicted to be 36.2 years older than the chronological age (Horvath, 2013). Other age predictors have been described in the literature (Hannum *et al.*, 2013; Weidner *et al.*, 2014) however, none that are so far able to achieve such a high level of accuracy across all tissue types as the epigenetic clock method (Weidner *et al.*, 2014).

Due to the findings using cancer samples, it would be interesting to see if age acceleration was found in other age-related disorders such as post-stroke dementia (PSD) and Parkinson’s disease (PD). This Chapter aims to predict the DNA methylation age of COGFAST and ICICLE participants using the epigenetic clock method (Horvath, 2013). It is hypothesised that participants with an older DNA methylation age than their chronological age will be more likely to develop cognitive deficits.

7.2 Experimental design

Using the HM450 data described in Section 3.2.3 and Section 5.2.3, the DNA methylation age of blood samples in COGFAST and ICICLE was calculated using the epigenetic clock method described by Horvath (2013). The epigenetic clock analysis was also performed on the HM450 data described in Section 4.2.3, to calculate the DNA methylation age of brain samples in COGFAST. In all epigenetic clock analyses, there were three output measures, as described in Table 73.

Output Variable	Description
DNA Methylation Age	Methylation age of the tissue
Age Acceleration Difference (AAD)	Difference between DNA methylation age and chronological age (DNAm age – chronological age)
Age Acceleration Residual (AAR)	Residual from regressing DNA methylation age on chronological age, i.e. DNA methylation age corrected for chronological age

Table 73: Description of the output variables resulting from epigenetic clock analysis.

HM450 data were then analysed using regression or ANOVA to investigate associations between the age acceleration residual (AAR) and outcome variables relating to disease. For COGFAST, the following outcome variables were considered; diagnosis (D/CN), Braak staging, MMSE, orientation, language comprehension, language expression, memory remote, memory recent, memory learning, memory total, attention, praxis, calculation, abstract thinking, perception, executive function and total CAMCOG score (Glossary, Table 2) . Following Bonferroni correction, associations between AAR and COGFAST outcomes were accepted as significant when p<0.0029. For ICICLE, the following motor outcome variables were used; Hoehn and Yahr, MDS-UPDRS II, MDS-UPDRS III, tremor dominant phenotype and PIGD phenotype (Glossary, Table 4). The cognitive outcomes testing in ICICLE were; MoCA, MMSE, total FAS, animals, pentagon copying, naming, language, language total, NMSQ memory, NMSQ concentration, total number of NMS, power of attention, digit vigilance accuracy, cognitive complaint, MCI, PRM, SRM, PAL and OTS (Glossary, Table 4). Following Bonferroni correction, associations between AAR and ICICLE outcomes were regarded as indicative of association when p<0.0021.

Analyses were also repeated to look for associations between AAR and exposure variables to ascertain determinants of age acceleration. The exposure variables measured in COGFAST and tested in the analysis were; OCSP class, Side of the body affected by stroke, degree of weakness in arm/leg, dysphasia, hypertension, atrial fibrillation, IHD/angina, cardiac failure, intermittent claudication, no of CVD risk factors, smoking, alcohol, baseline MMSE, baseline orientation, baseline language comprehension, baseline language expression, baseline memory remote, baseline memory recent, baseline memory learning, baseline memory total, baseline attention, baseline praxis, baseline calculation, baseline abstract thinking, baseline perception, baseline executive thinking and baseline total CAMCOG score (Glossary, Table 1). Associations between COGFAST exposures and AAR were regarded as significant if $p<0.002$, following Bonferroni correction. The exposure variables tested in ICICLE were; education, NART, alcohol, smoking, height, weight, BMI, RCF, B12, homocysteine, IHD, diabetes, hypertension, hypercholesterolaemia, GDS and Levodopa daily dose (Glossary, Table 3). Associations between ICICLE exposures and AAR were regarded as significant if $p<0.003$, following Bonferroni correction.

7.3 Results

7.3.1 DNA methylation age of COGFAST participants

The epigenetic clock method (Horvath, 2013) was used to calculate the DNA methylation age of individuals who had provided the blood, DLPFC and hippocampus samples. Results of the epigenetic clock method are displayed for blood (Table 74), DLPFC (Table 75) and hippocampus (Table 76).

Sample ID	Chronological age (Years)	DNA methylation age (Years)	Age acceleration difference (Years)	Age acceleration residual
20030067	80	79.00	-1.00	1.87
20060127	83	70.21	-12.79	-8.62
20040031	75	71.24	-3.76	-3.08
20040044	78	73.72	-4.28	-2.29
20080030	80	69.42	-10.58	-7.72
20080048	78	87.06	9.06	11.05
20060102	86	68.49	-17.51	-12.03
20070044	80	71.69	-8.31	-5.45
20090066	85	69.58	-15.42	-10.37
20080043	77	64.46	-12.54	-10.99
20030057	85	84.06	-0.94	4.11
20060068	77	70.50	-6.50	-4.95
20070008	86	83.14	-2.86	2.63
20120048	78	91.93	13.93	15.92
20080090	82	79.83	-2.17	1.57
20060001	80	78.73	-1.27	1.59
20110315	75	75.25	0.25	0.93
20110694	83	85.50	2.50	6.67
20080044	78	78.40	0.40	2.39
20060092	90	85.02	-4.98	2.25
20100307	86	82.99	-3.01	2.47
20100709	83	83.81	0.81	4.99
20040036	76	75.51	-0.49	0.63
20090073	81	82.75	1.75	5.05
20040032	86	90.56	4.56	10.04
20070004	85	79.61	-5.39	-0.34
20120082	85	73.48	-11.52	-6.47
20070087	85	76.06	-8.94	-3.89
20070057	92	85.93	-6.07	2.03

Table 74: COGFAST blood DNA epigenetic clock data. For each sample, the chronological age of the blood at collection, DNA methylation predicted age, age acceleration difference (DNA methylation age minus chronological age) and the age acceleration residual is presented.

Table 74 shows that the majority of samples (21/29) have a lower DNA methylation age than chronological age. Only eight samples have a chronological age that is younger than the DNA methylation age. The largest age acceleration difference (AAD) is -17.51 years indicating that the DNA methylation age is 17.51 years younger than the chronological age. The AAR is however a more accurate way of looking at differences in age, since this measure is the DNA methylation age accounted for chronological age. AARs in the blood of COGFAST participants range from -12.03 to 15.92. A paired t test revealed significant differences between predicted and actual age of the blood samples (F=2.850, p=0.008).

Sample ID	Chronological age (Years)	DNA methylation age (Years)	Age acceleration difference (Years)	Age acceleration residual
20030067	80	59.46	-20.54	-2.47
20040031	76	54.84	-21.16	-6.69
20040044	78	46.68	-31.32	-15.06
20080030	84	66.48	-17.52	4.13
20080048	82	73.67	-8.33	11.53
20060102	89	68.44	-20.56	5.58
20070044	84	62.05	-21.95	-0.30
20090066	91	57.93	-33.07	-5.14
20080043	82	65.99	-16.01	3.85
20070008	93	73.45	-19.55	10.18
20120048	88	71.90	-16.10	9.14
20080090	88	62.55	-25.45	-0.21
20110315	86	74.14	-11.86	11.59
20110694	92	62.59	-29.41	-0.57
20060092	94	71.36	-22.64	7.99
20100307	94	66.40	-27.60	3.03
20100709	93	62.42	-30.58	-0.85
20040036	79	66.70	-12.30	4.87
20090073	88	61.87	-26.13	-0.88
20040032	88	64.08	-23.92	1.33
20070004	92	45.60	-46.40	-17.56
20120082	95	58.08	-36.92	-5.40
20070087	90	48.94	-41.06	-14.03
20070057	98	59.71	-38.29	-4.07

Table 75: COGFAST DLPFC DNA epigenetic clock data. For each sample, the chronological age of the blood at collection, DNA methylation predicted age, age acceleration difference (DNA methylation age minus chronological age) and the age acceleration residual is presented.

Table 75 shows that every DLPFC sample has a younger DNA methylation age than chronological age when looking at AAD. Differences range from -8.33 to -46.40 years. However, the AAR, which is the methylation age adjusted for chronological age, reveals that actually only 13/24 have a younger predicted age than actual age. The remaining eleven samples have an older predicted age than their actual age. A paired t test highlighted significant differences between predicted and actual age (F=12.71, p=6.96E-12).

Sample ID	Chronological age (Years)	DNA methylation age (Years)	Age acceleration difference (Years)	Age acceleration residual
20030067	80	58.26	-21.74	-5.15
20060127	86	62.25	-23.75	-6.51
20040031	76	68.27	-7.73	8.42
20040044	78	53.69	-24.31	-7.95
20080030	84	58.68	-25.32	-8.30
20060102	89	69.94	-19.06	-1.50
20070044	84	64.42	-19.58	-2.56
20090066	91	59.26	-31.74	-13.95
20080043	82	61.73	-20.27	-3.47
20030057	88	70.39	-17.61	-0.15
20070008	93	75.15	-17.85	0.15
20120048	88	91.11	3.11	20.57
20080090	88	63.64	-24.36	-6.91
20060001	84	61.53	-22.47	-5.45
20110315	86	71.30	-14.70	2.53
20110694	92	83.75	-8.25	9.64
20030080	93	82.64	-10.36	7.65
20080044	84	71.22	-12.78	4.24
20060092	94	84.27	-9.73	8.38
20100307	94	72.77	-21.23	-3.12
20100709	93	70.47	-22.53	-4.53
20040036	79	69.02	-9.98	6.50
20090073	88	73.62	-14.38	3.07
20040032	88	81.87	-6.13	11.33
20070004	92	64.87	-27.13	-9.24
20120082	95	71.11	-23.89	-5.67
20070087	90	75.40	-14.60	3.08
20070057	98	78.36	-19.64	-1.09

Table 76: COGFAST hippocampal DNA epigenetic clock data. For each sample, the chronological age of the blood at collection, DNA methylation predicted age, age acceleration difference (DNA methylation age minus chronological age) and the age acceleration residual is presented.

The AAD in Table 76 shows that only one of the 28 hippocampal samples has a DNA methylation age older than the chronological age. The AAR however, shows that when predicted age is adjusted for chronological age there is more variation between samples. AARs range from -9.24 to 20.56 indicating that some hippocampal samples do have a much older predicted age than their actual age. A paired t test showed that there were significant differences between predicted and actual age (F=12.05, p=2.26E-12).

7.3.2 DNA methylation age of ICICLE participants

The epigenetic clock method (Horvath, 2013) was used to predict the age of the blood samples. Results of the epigenetic clock method are displayed in Table 77.

Sample ID	Chronological age (Years)	DNA methylation age (Years)	Age acceleration difference (Years)	Age acceleration residual
IN061	72.9	68.37	-4.53	0.56
IN102	73.6	71.44	-2.16	3.08
IN040	77.6	75.46	-2.14	3.98
IN075	70.7	73.39	2.69	7.29
IN107	76.3	73.07	-3.23	2.60
IN042	78	74.13	-3.87	2.33
IN127	74.5	67.53	-6.97	-1.53
IN135	64.3	64.52	0.22	3.41
IN039	67.9	63.74	-4.16	-0.17
IN045	65.6	60.98	-4.62	-1.14
IN121	63.6	59.19	-4.41	-1.37
IN104	74.5	62.47	-12.03	-6.59
IN091	77.8	55.39	-22.41	-16.24
IN090	62.9	56.14	-6.76	-3.87
IN081	61.2	52.60	-8.60	-6.09
IN067	79.2	72.36	-6.84	-0.36
IN055	61.8	56.39	-5.41	-2.76
IN009	67.3	67.95	0.65	4.50
IN108	68.5	63.92	-4.58	-0.46
IN025	61.9	64.91	3.01	5.68
IN074	64.2	61.23	-2.97	0.21
IN003	76.3	77.42	1.12	6.95
IN013	78.5	68.68	-9.82	-2.11
IN089	65.4	64.17	-1.23	1.89
IN080	62.1	55.23	-6.87	-4.91
IN083	69.8	66.37	-3.43	1.23
IN049	66.3	59.05	-7.25	-3.82
IN072	49.6	59.63	10.03	7.62
IN028	63.7	53.36	-10.34	-7.82
IN112	67.1	66.46	-0.64	3.08
IN058	69	61.89	-7.11	-2.73
IN115	73.9	67.13	-6.77	-0.67
IN001	69.2	68.87	-0.33	4.12
IN140	75.8	77.82	2.02	8.79
IN056	61.9	52.50	-9.40	-7.51
IN051	67.4	63.98	-3.42	0.40
IN094	75.5	72.50	-3.00	3.65
IN086	70.2	61.71	-8.49	-3.69
IN067	79.2	59.21	-19.99	-12.03
IN035	61.9	60.33	-1.57	0.32
IN046	67.6	63.26	-4.34	-0.44
IN019	75.8	77.24	1.44	8.21
IN012	77.3	74.63	-2.67	4.61
IN124	66.2	64.61	-1.59	1.81

Table 77: ICICLE blood DNA epigenetic clock data. For each sample, the chronological age of the blood at collection, DNA methylation predicted age, age acceleration difference (DNA methylation age minus chronological age) and the age acceleration residual is presented.

The results for the ICICLE blood DNA (Table 77) are similar to those of the COGFAST blood DNA samples (Table 74) in that some samples have an older predicted age and some have a younger predicted age when compared to their actual chronological age.

The AARs (Table 77) show that just over half of all ICICLE blood DNA samples (23/44) have an older predicted age than actual age. A paired t test was used to show that predicted age does significantly differ from actual age (F=5.207, p=5.12E-06).

7.3.3 Does the methylation age of the brain differ from the blood?

The COGFAST cohort provides a unique opportunity to assess whether the brain tissue, following stroke, ages differently to the blood and whether a peripheral blood sample can be used as a surrogate to predict DNA methylation age of the brain.

Figure 105 shows the DNA methylation age against the actual age of each of the 81 samples (blood, DLPFC and hippocampal samples). This graph depicts the AAD and AAR for each sample.

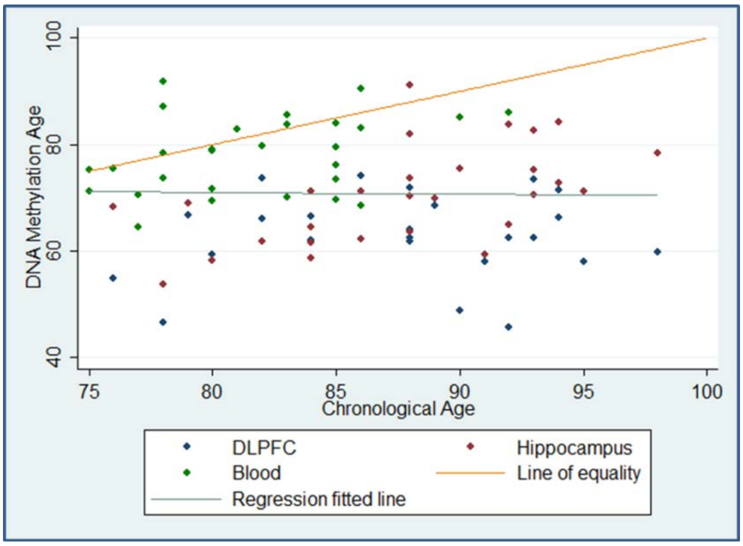


Figure 105: Scatter plot of DNA methylation age against chronological age for each of the three tissues analysed in COGFAST. Data points above the line of equality (orange) have a positive AAD indicating that their DNA methylation age is older than their actual age. The distance from the regression fitted line (grey) to each data point indicates the AAR. A positive AAR residual indicates a DNA methylation age above their actual age.

Figure 105 shows that more blood samples have a positive AAR (and therefore older DNA methylation age than actual age) than DLPFC and hippocampus samples. A multilevel linear regression model was used to compare AAR across tissue types while accounting for non-independence of multiple measurements within individuals. There was no evidence for a difference in AAR between peripheral blood and either DLPFC (95% confidence interval for difference -2.79, 3.29; p=0.873) or hippocampus (95% confidence interval for difference -2.81, 3.20; p=0.898). No difference was observed between the three tissue types indicating that; 1) the brain does not age differently to blood in the tissues and samples included in this analysis and 2) in this case, the blood can be used as a surrogate to predict DNA methylation age of brain tissue. This would be useful in future studies where blood samples were taken from living individuals.

In addition, correlation analyses were performed to assess whether inter-individual variation in DNA methylation age was correlated across tissues. A moderate correlation was observed between the two brain regions tested (Rho=0.392, p=0.064). Methylation age was observed to be more strongly correlated between the hippocampus and blood (Rho=0.744, p=4.69E-05) than the DLPFC and blood (Rho=0.195, p=0.373). This indicates that the blood may be a better surrogate for the hippocampus than the DLPFC.

7.3.4 Does an older methylation age increase the risk of developing cognitive decline?

Post-stroke dementia

Regression and ANOVA analyses were used to look for associations between blood-derived AAR and the outcome variables (Section 7.2) measured in COGFAST. Since this Chapter is concerned with the effect of DNA methylation age, the AAR was used over AAD in the analysis. AAR was found to differ between sexes (t=2.39, p=0.024) with females showing a negative mean AAR (-2.85), indicating a lower than expected DNA methylation age whereas males had a positive mean AAR (+2.66), indicating a DNA methylation age older than their actual age. For this reason, sex was included in the model as a covariate.

Figure 106 shows the correlation between predicted age and chronological age of the blood, coded for diagnosis (Dementia (D)/Cognitively Normal (CN)). A weak correlation (Rho=0.35, p=0.063) was observed between DNA methylation age and chronological age. When the correlation analysis was split by diagnosis, there was a strong correlation between predicted and chronological age in participants with dementia (Rho=0.727, p=0.001) and a weak negative correlation between predicted and chronological age in CN participants (Rho=-0.073, p=0.814). Figure 106 shows more CN samples have an older DNA methylation age than actual age (above the line of equality), however no significant difference in AAR was observed between stroke survivors with dementia and stroke survivors who remained cognitively normal prior to death (F=2.45, p=0.130) (Figure 107). No associations were observed between AAR and any outcome variable measured, suggesting that in the blood of stroke survivors, the methylation age does not increase the risk of developing PSD.

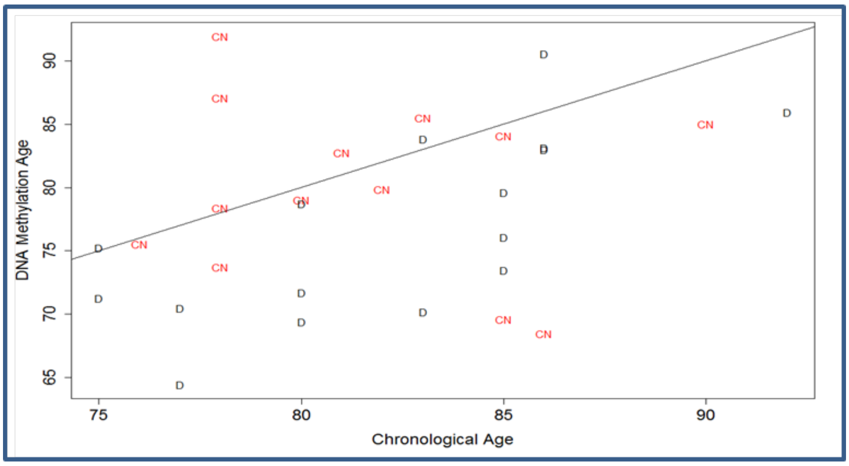


Figure 106: DNA methylation age against chronological age of the blood. Samples with dementia are coded D (black), cognitively normal samples are coded CN (red). Any samples above the line of equality have a DNA methylation age older than their actual age.

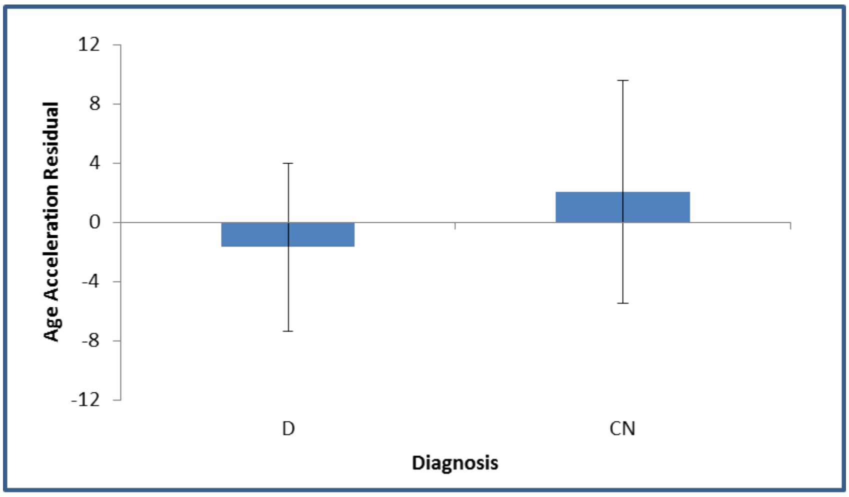


Figure 107: Age acceleration residual of the blood by diagnosis of post-stroke dementia.

To assess whether any exposures could explain the variances in AAR, ANOVA and regression analyses were performed to look for association between AAR and the exposure variables measured in COGFAST (Section 7.2). No associations reaching the Bonferroni corrected significance threshold were observed between AAR and any exposure. However, AAR was found to be associated with smoking status at the uncorrected significance threshold of p<0.05 (F=3.92, p=0.036). Figure 108 shows that ex-smokers have a positive AAR indicating a DNA methylation age above what would be predicted from their actual age, whereas current and never smokers have a negative AAR (DNA methylation age younger than actual age). This is not what would be expected based on methylation data that suggests the methylation profiles of ex-smokers reverts back to that of never smokers over time (Zeilinger *et al.*, 2013; Elliott *et al.*, 2014). Analysis by Elliott *et al.* (2014) and Zeilinger *et al.* (2013) used a subset of CpG sites sensitive to smoking, however the epigenetic clock method uses CpG sites which are very likely to behave differently. This would need to be tested in a much larger cohort before any conclusions can be drawn from this, as this finding may only be significant due to the small sample size used. A possible explanation for this seemingly strange finding could relate to reasons why participants stopped smoking. For example, the presence of any co-morbidities which affect the health of the participant may encourage this individual to quit smoking but may also result in changes to DNA methylation and the subsequent detection of accelerated ageing. No other significant associations were observed between AAR and exposure variables.

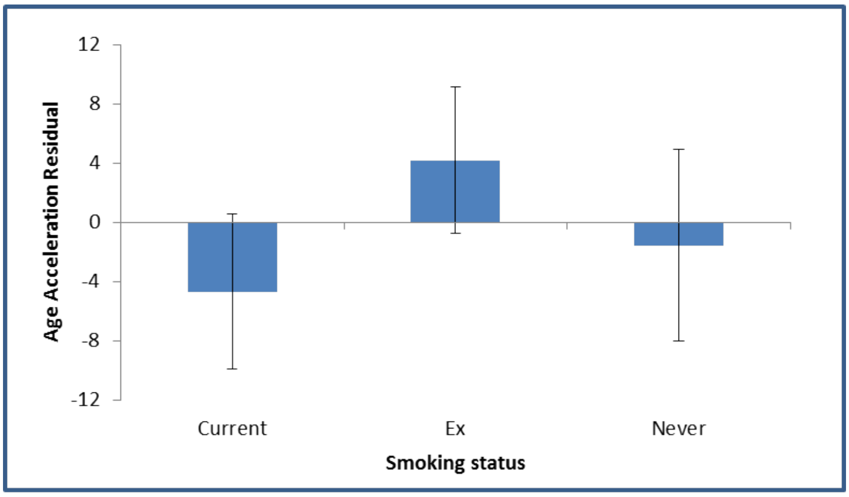


Figure 108: Age acceleration residual of the blood by smoking status. Data bars with error bars are presented for current, ex and never smokers. A bar above the line at y=0 indicates a positive AAR. A bar below the line at y=0 indicates a negative AAR.

The regression and ANOVA analyses were repeated in the DLPFC and hippocampus to look for associations between AAR of the brain and outcome variables (Section 7.2) and AAR and exposure variables (Section 7.2). Although sex wasn't found to be associated with the AAR in the DLPFC (t=0.134, p=0.895) or hippocampus (t=1.287, p=0.198), sex was still included in the model as it was associated with AAR in blood.

Figure 109 shows the correlation between DNA methylation age and actual age coded for diagnosis in the DLPFC (A) and hippocampus (B). No correlation was observed between DNA methylation and chronological age in the DLPFC (Rho = 0.076, p=0.72). When correlation analysis split by diagnosis was performed, a weak positive correlation was observed in the CN participants (Rho=0.248, p=0.463) and a weak negative correlation between the predicted age and chronological age in the participants with dementia (Rho=-0.099, p=0.747). A moderate correlation was observed between DNA methylation age and chronological age in the hippocampus (Rho= 0.550 p=0.002). When correlation was split by diagnosis, a moderate correlation was observed between predicted and chronological age in both participants with dementia (Rho = 0.657, p=0.006) and CN participants (Rho=0.541, p=0.069). No difference in AAR was observed between demented and cognitively normal participants in either tissue (DLPFC: F=0.56, p=0.464; hippocampus: F=0.70, p=0.409) (Figure 110).

AAR was not found to be associated with any outcome variable or any exposure variable in either the DLPFC or hippocampus. PM delay was not found to be associated with AAR in either tissue (DLPFC: t=0.36, p=0.720; hippocampus: t=-0.20, p=0.846).

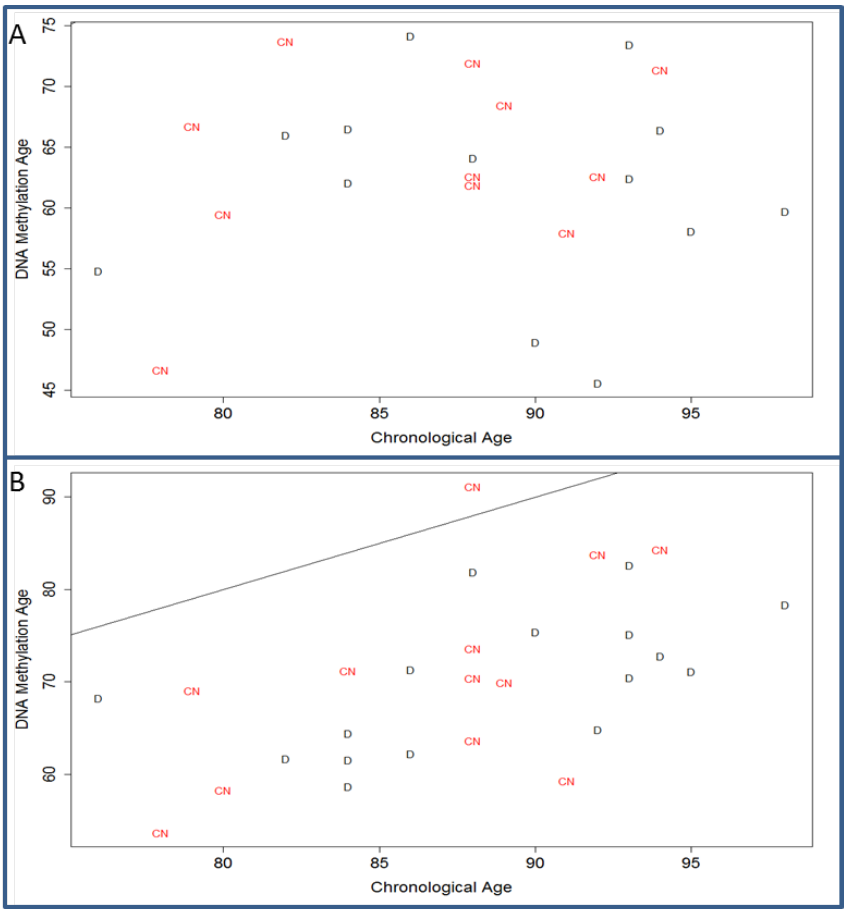


Figure 109: DNA methylation age against chronological age of the brain. A. DNA methylation age against chronological age in the DLPFC. B. DNA methylation age against chronological age in the hippocampus. Samples with dementia are coded D (black), cognitively normal samples are coded CN (red). Any samples above the line of equality have a DNA methylation age older than their actual age.

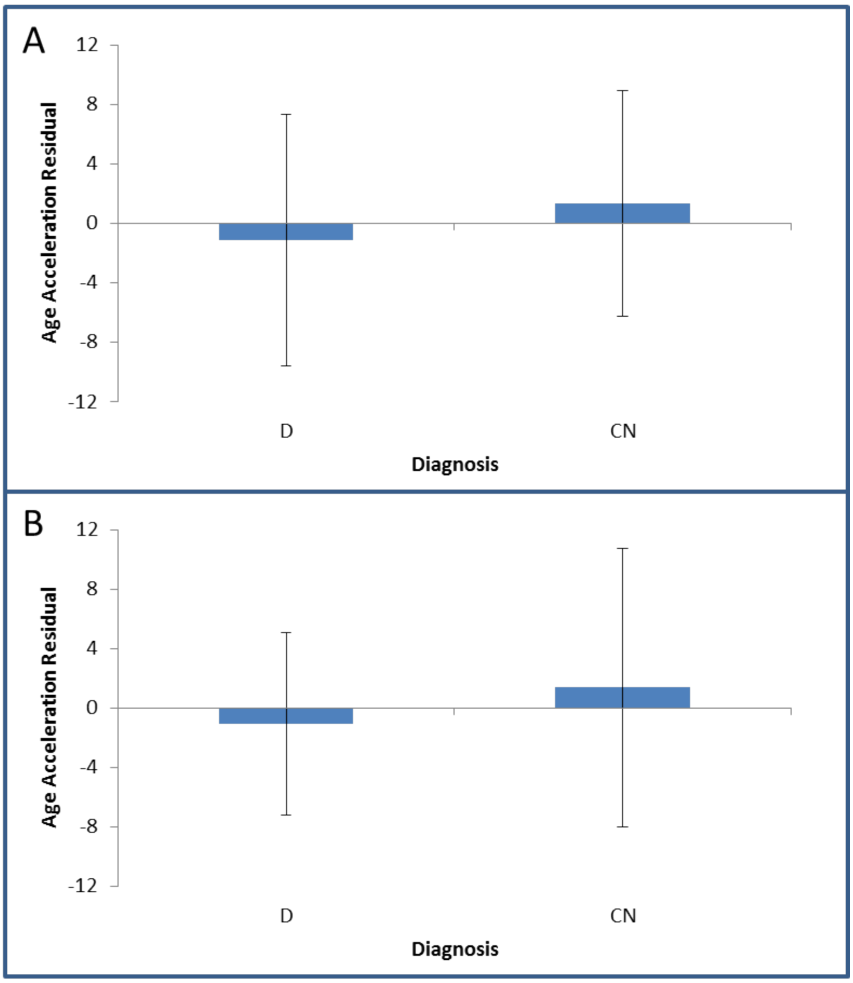


Figure 110: Age acceleration residual of the brain by diagnosis of post-stroke dementia. A. Age acceleration of the DLPFC. B. Age acceleration of the hippocampus.

Parkinson’s disease

The same approach was applied to test for associations between AAR and the outcome variables measured in ICICLE (Section 7.2). No sex differences in AAR were observed (t=0.47 p=0.642), however sex was still included in the model as a covariate.

Figure 111 shows the correlation between DNA methylation age and actual age of the blood, coded for MCI diagnosis (MCI/NCI). A strong correlation was observed between DNA methylation age and chronological age (Rho = 0.660 p=1.10E-06). When correlation analysis was repeated, split by MCI diagnosis, there was a moderate correlation between predicted and actual age in the MCI participants (Rho=0.452, p=0.045) and there was a strong correlation between predicted and chronological age in the NCI participants (Rho=0.886, p= 2.02E-07). No difference in AAR was observed between PD-MCI and PD-NCI participants (F=0.55, p=0.462) (Figure 112).

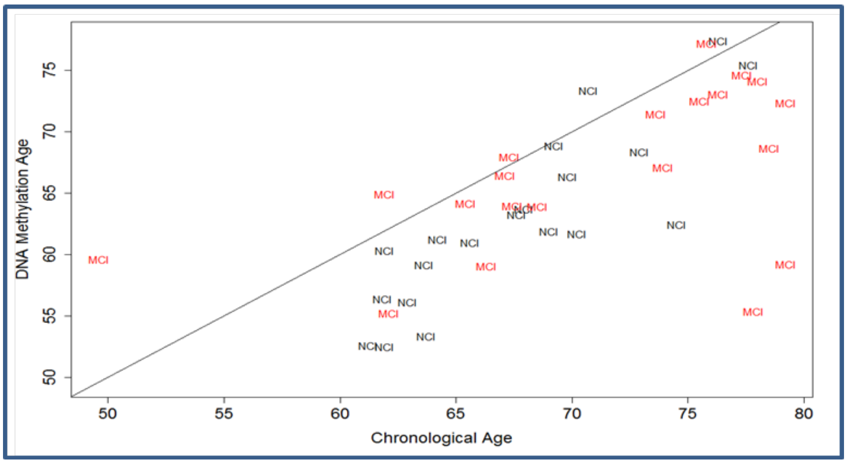


Figure 111: DNA methylation age against chronological age of the blood. Samples with MCI are coded MCI (red), cognitively normal samples are coded NCI (black). Any samples above the line of equality have a DNA methylation age older than their actual age.

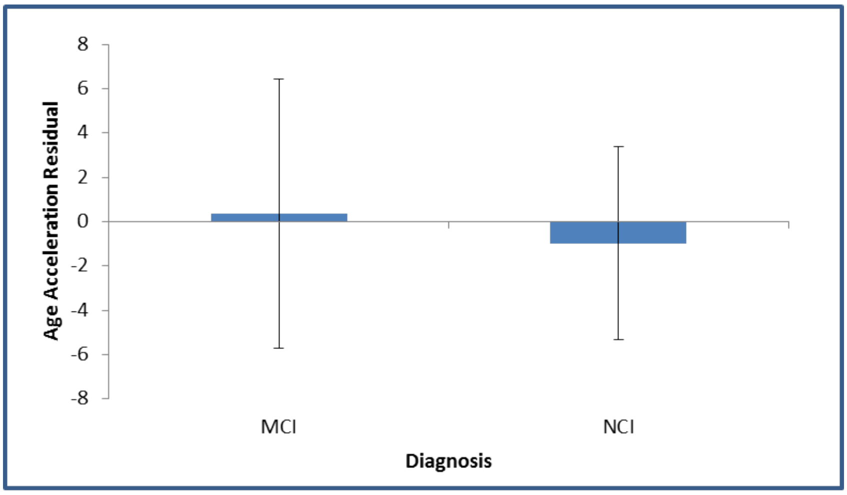


Figure 112: Age acceleration residual of the blood by diagnosis of PD-MCI.

AAR was not found to be associated with any cognitive outcome measured in ICICLE (Section 7.2), suggesting that the DNA methylation age of the blood does not affect the risk of developing cognitive deficits. AAR was found to be weakly correlated with MDS-UPDRS II (Rho= 0.310 p=0.041), a variable recording the self-evaluation of the activities of daily life, indicating that DNA methylation age of the blood may have some effect on basic motor ability (Figure 113). However, this did not reach the significance threshold after multiple test correction had been applied. No other association was observed between AAR and a motor outcome variable.

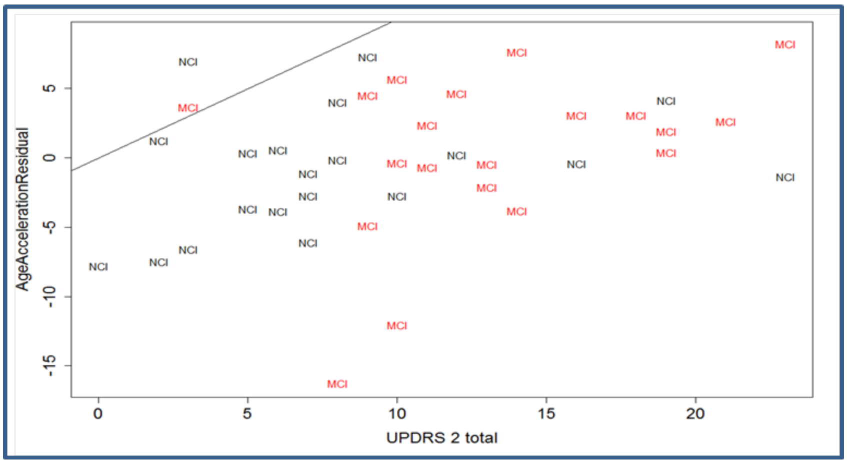


Figure 113: Correlation between AAR and MDS-UPDRS II coded for diagnosis of MCI.

Similarly to the COGFAST analysis, to test whether any exposure could explain the variation in AAR, ANOVA and regression analyses were performed to investigate associations between AAR and exposures variables measured in ICICLE (Section 7.2). No associations were observed that reached the Bonferroni corrected significance threshold (0.003), however one association did reach p<0.05. A difference in AAR was observed between participants with diabetes and participants without (F=3.44, p=0.042) (Figure 114). Participants with diabetes had a positive AAR and therefore had an older DNA methylation age compared with their actual age. No other significant association was observed between AAR and an exposure variable.

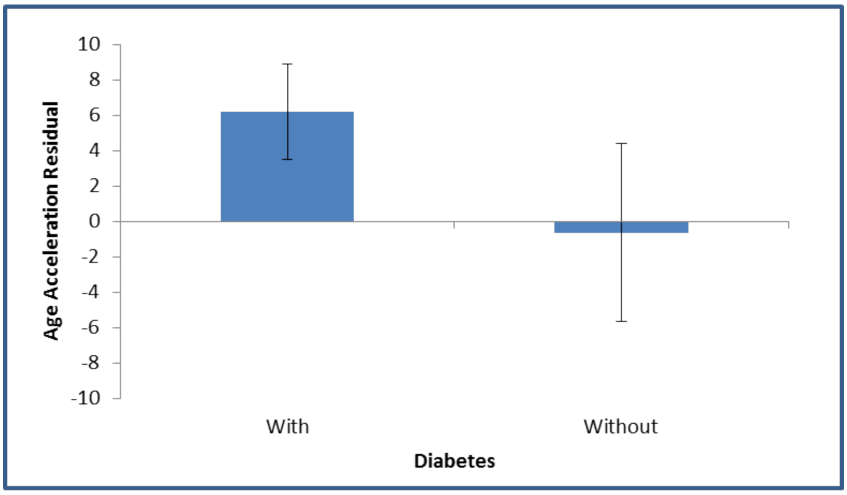


Figure 114: Age acceleration residual of the blood by diabetes.

AAR was associated with both an exposure (diabetes) and an outcome (MDS-UPDRS II) at the p<0.05 significance threshold in ICICLE. It is possible that diabetes may cause differences in AAR which then affects motor outcome, however regression analysis revealed that diabetes is not associated with MDS-UPDRS II (t=-1.03, p=0.310) making potential mediation by AAR unlikely. The relationship between AAR and MDS-UPDRS II is therefore unaffected by the presence or absence of diabetes.

7.4 Discussion

This Chapter aimed to identify the DNA methylation age of all samples used in the HM450 analyses of COGFAST and ICICLE and whether those participants with an older predicted age using methylation data than their actual age were more likely to develop cognitive decline.

The results in this Chapter outline that there were differences between DNA methylation age and chronological age in the blood of ICICLE samples and in each tissue (blood, DLPFC and hippocampus) from COGFAST participants.

In the PSD (COGFAST) samples, there were no strong associations observed between the age acceleration residual (AAR) and disease outcome in the blood or either brain region. Using the COGFAST samples there was an association, although it did not reach the stringent significance threshold following Bonferroni correction, between smoking status and AAR when regression was performed to look for association between exposures and AAR in the blood. Ex-smokers were seen to have an older DNA methylation age than chronological age whilst the current and never smokers had a younger DNA methylation age than actual age. These results are not seemingly biologically plausible due to the previously reported findings that the DNA methylation profiles of ex-smokers reverts back to that of a never smoker over time (Zeilinger *et al.*, 2013; Elliott *et al.*, 2014). For this reason, it would be expected that ex-smokers and never smokers had a more similar AAR than current and never smokers. It is counter-intuitive for ex-smokers to have a higher age than current and never smokers but the confidence intervals are very wide and therefore the age estimates are imprecise. However, to establish that these results are not just due to the small sample used in this analysis, this study would need to be replicated using a much larger cohort. A possible explanation for this finding is that the accelerated ageing observed in the ex-smokers is due to the presence of co-morbidities that caused them to stop smoking and also altered DNA methylation levels.

In ICICLE, no associations were observed between any outcome or exposure and AAR following Bonferroni correction. However, a weak correlation was observed between AAR and the outcome variable MDS-UPDRS II, a self-reported evaluation of the activities of daily life (Rho=0.310, p=0.041). In addition, PD samples with diabetes exhibited an older DNA methylation age compared with those participants without diabetes (F=3.44, p=0.042). However, no association was observed between diabetes and MDS-UPDRS II, indicating that they are unlikely to form part of a causal pathway.

The major limitation of this Chapter, as has already been alluded to, is the small sample size used, which has limited the statistical power to detect differences in AAR between samples. In addition a relatively small range of ages was used in these studies therefore limiting the ability to detect larger differences in accelerated ageing.

Since the method to predict the DNA methylation age of samples is a new technique, this is the first study (at the time of writing) that has attempted to look at the relationship between AAR and cognitive decline in a PSD and PD cohort. This technique has the ability to identify the risks of an older DNA methylation age, however, the sample sizes used in this current study were modest, with limited power to detect potential associations. It would be of interest to increase the sample size of the cohort used, to assess whether any associations, particularly those which reached the threshold of p<0.05, would be strengthened in a larger cohort. However, the lack of association between AAR and phenotype is not surprising. This has been the case in other studies which have used the epigenetic clock method or similar techniques to look at age-related DNA methylation and subsequent effects on phenotype (Bell *et al.*, 2012; Boks *et al.,* 2014). In a cohort of soldiers, trauma was associated with an accelerated ageing process. However, this accelerated ageing was not found to have an effect on phenotype, but rather a younger epigenetic age was associated with the development of post-traumatic stress disorder (Boks *et al.*, 2014). This supports the findings of this Chapter and indicates that although there are differences in age acceleration, these may not necessarily have a subsequent effect on phenotype. Another study highlighted many age-related DNA methylation differences in twins however very few of these loci were found to be associated with an age-related phenotype, further supporting these findings (Bell *et al.*, 2012).

The COGFAST cohort provided a great opportunity to assess whether there were any differences in the DNA methylation age of tissues. No differences were observed between the blood and either brain region. This supports the findings of Horvath (2013) who reported that there were very few differences in DNA methylation age between brain regions. It is interesting to note that in this case, blood does not show a different age acceleration to brain tissues, suggesting that blood could be used as a surrogate for assessing the predicted age of the brain in future studies. The use of an easily accessible tissue such as the blood would enable this method to be used in cohorts where brain tissue was not available. This would first require validation in a larger cohort with greater power to detect differences.

Chapter 8. Discussion

8.1 Summary of aims and objectives

This thesis addressed a total of 6 aims. Key observations and discussion relating to each of these aims are summarised below.

8.1.1 To identify blood-based biomarkers of post-stroke dementia

Due to the ageing population and a worsening of lifestyle factors such as a poor diet and sedentary lifestyle, the prevalence of stroke is increasing. A major consequence of stroke is that of post-stroke dementia (PSD). The mechanisms resulting in PSD are currently unknown and it is therefore unclear why some stroke survivors develop PSD whilst others remain with their cognition intact. Therefore, prognostic indicators that can help clinicians distinguish those stroke survivors most likely to develop PSD are required so that early preventative treatments can be implemented. DNA methylation has been identified as a valuable prognostic indicator in several cancers including medulloblastoma (Schwalbe *et al.*, 2013), myelodysplastic syndromes (Shen *et al.*, 2010) and ovarian cancer (Wei *et al.,* 2006).

The aim of Chapter 3 was to identify blood-based biomarkers which could be used to predict stroke survivors at risk of developing post-stroke dementia. To address this aim the COGFAST cohort was utilised. This cohort consists of stroke survivors over the age of 75 years who regained full cognition three months post-stroke. A blood sample taken at recruitment onto the study (three months post-stroke) was used in the epigenetic analysis. The methylation profile of the baseline blood samples was analysed using the HM450 BeadChip, to look for differentially methylated positions associated with cognitive outcome at the time of death. A selection of the HM450 top hits were analysed across a wider sample of the same cohort using Pyrosequencing.

These analyses identified two potential biomarkers; *APOB* and *NGF*, both of which have previously been implicated in cardiovascular (Gigante *et al.*, 2012) and dementia related diseases (Zhang *et al.*, 2013). Methylation at one CpG site within each of these genes was found to be significantly reduced in the individual with Braak stage VI pathology. Adjustment for cellular composition revealed that this difference in methylation was not due to differences in cellular composition between samples. This difference in DNA methylation was not seen in AD patients with Braak stage VI pathology who had not suffered a stroke, indicating this difference may be caused by a combination of stroke and Braak stage VI pathology. Since only one COGFAST participant was diagnosed with Braak stage VI pathology, this differential methylation could however just be a sample issue. Based on this study, it is unclear whether this is a ‘real’ (i.e. biologically meaningful) association and that all stroke patients with a Braak staging of VI would have a reduced methylation level at these CpG sites, or whether this methylation level is different in this individual to others. On further examination of the clinical background of this individual, it was found that this participant had had three recurrent strokes prior to the index stroke which saw the individual recruited onto the COGFAST study. This makes the individual unique and could explain why the methylation profile of this individual at these particular sites is different. It may be that this methylation difference is caused by recurrent strokes. This is explored further in Section 8.1.2. Alternatively, the observations, although validated using two complementary methods, could be a chance event peculiar to the individual concerned and be unrelated to dementia or stroke.

Due to the small sample size of COGFAST, with limited numbers of participants with the more severe Braak staging pathology, this study requires repeating in a larger cohort, to assess whether the differential methylation seen here is in fact associated with the combination of stroke and Braak stage VI pathology, or whether it is associated with recurrent strokes or some other unidentified variable. If analysis in a larger cohort identified a significant association between DNA methylation at these CpG sites within *APOB* and *NGF* and Braak staging, they could be used as biomarkers to predict later cognitive deficits. However, given the threshold effect observed, i.e. the association was not linear between methylation and Braak staging but only occurred at the most severe end of the Braak stage spectrum, these markers would only be able to predict those at risk of developing the most severe dementia. These biomarkers would not be suitable for predicting the development of dementia in the majority of stroke survivors.

In conclusion, two potential biomarkers were identified. Due to the small sample size, it is unclear whether the association between DNA methylation and Braak staging is a true association, or whether some other variable, such as recurrent strokes, is confounding the identified relationship. Replication in a larger cohort is required to determine the cause of the differential methylation.

8.1.2 To identify differentially methylated positions in the brain which are associated with impaired cognitive function in stroke patients

As previously mentioned, the mechanisms resulting in PSD are currently unknown. Whilst a biomarker used in the prediction of disease does not need to be causally implicated and can therefore be measured in any suitable and readily accessible tissue, e.g. blood/saliva, the search for a mechanistic link must be undertaken in the affected tissue, in this case the brain. The aim of Chapter 4 was to identify differentially methylated positions associated with cognitive decline in the brain, to try to elucidate the mechanisms involved in the pathogenesis of PSD.

To address this aim, brain samples (DLPFC and hippocampus) from the COGFAST cohort were utilised. The methylation profile of the brain samples were analysed using the HM450 BeadChip to look for differentially methylated positions associated with cognitive outcome and pathological markers at the time of death. A selection of the HM450 top hits were analysed across a wider cohort using Pyrosequencing.

These analyses identified two methylation markers of interest; cg18837178 and *NGF*. *NGF* was also identified as a potential blood-based biomarker, and as discussed previously, has been associated with AD in the literature. cg18837178 is a CpG site within a non-coding RNA (CT49). Both cg18837178 and the CpG site within *NGF* had a reduced methylation level in the only Braak stage VI individual studied, compared with individuals with Braak stage V or below. Again, these differences were not due to cellular composition and were not identified in non-stroke Braak stage VI AD cases. These differences were again found to be unique to the COGFAST individual with Braak stage VI pathology and a history of recurrent strokes. These data were therefore inconclusive with respect to the role of *NGF* and CT49 in mechanistic pathways leading to PSD. *NGF* is known to play a role in the development and maintenance of neurons, both during early development and throughout adulthood (Aloe *et al.*, 2012). This function relates to dementia and related cognitive decline, suggesting that NGF could be mechanistically involved in the pathogenesis of dementias. The biological function of CT49 is unknown.

To assess the possibility that DNA methylation may be involved in the mechanisms resulting in PSD, it was necessary to undertake some functional analyses. The quality of RNA was too poor to undertake gene expression analysis. For this reason, protein was extracted from brain tissue and Western blotting was performed to look for differences in NGF protein levels between samples from individuals of different Braak stages. As CT49 is a non-coding RNA, protein analysis was unable to be performed. Unfortunately, no brain tissue was available for the COGFAST participant with Braak stage VI pathology, so it remains unclear whether this differential methylation at *NGF* is translated at the protein level. There was no difference in NGF protein levels between non-stroke Braak stages, nor AD cases and controls. This does not support previous findings which have identified significant differences in NGF levels in the brains of AD patients compared to controls (Scott *et al.*, 1995; Fahnestock *et al.*, 1996; Mufson *et al.*, 2012b). Protein analysis in this case was unable (due to the omission of the individual with Braak stage VI pathology) to provide evidence for a mechanistic link between *NGF* methylation and the development of PSD.

The methylation differences identified within *NGF* and at cg18837178 were unique to the Braak stage VI individual with a history of recurrent strokes. The finding that *NGF* methylation levels are reduced in a PSD case with a history of recurrent strokes is of particular interest. Although brain protein levels were not measured in this individual, it can be tentatively postulated that increased NGF protein levels would be observed due to the anticipated relationship of an inverse relationship between methylation and expression. As a stroke is known to promote angiogenesis, and angiogenesis requires growth factors (Krupinski *et al.*, 1994; Font *et al.*, 2010), it could be postulated that following a recurrent stroke, there is an increase in angiogenesis and therefore an increase in growth factor levels, although this has yet to be proven. This suggests that following a recurrent stroke there may be an increased level of NGF in the brain. Due to the biological plausibility of this scenario, if suitable tissue was sought, it would be very interesting to assess whether the NGF protein levels in a recurrent stroke patient were increased compared to stroke survivors who had experienced only one stroke.

In conclusion, differential methylation at two CpG sites (*NGF* and cg18837178) was identified in post-stroke cases. At both CpG sites, Braak stage VI pathology was associated with a decreased methylation level. These sites could indicate involvement in the mechanistic pathway resulting in the development of PSD. However, as with the analysis of the blood samples, the involvement of these CpG sites is limited to only the most severe cases. Alternatively, methylation at these CpG sites may be associated with the recurrent strokes experienced prior to recruitment into the COGFAST study. To elucidate whether recurrent strokes or the combination of stroke plus Braak stage VI pathology are contributing to this differential methylation, replication in a larger cohort with an increased number of samples with the most severe form of brain pathology is required. The detection of aberrant methylation in this individual was observed and validated in DNA extracted from peripheral blood (described in Chapter 3) and in DNA from brain tissue, which provides strong evidence that these observations were real and not artefactual. The consistent differences across different tissues seen at the *NGF* locus could plausibly be explained by underlying genetic variation. Some corrections were made for known common variants that underlie HM450 probes but this may not have accounted for all possible genetic variation which is located further away from the CpG site but still exerts an effect on methylation levels. No genotype information was available for the individual concerned to persue this potential explanation. Despite removing SNPs (as suggested by Illumina) from the data analysis and using a genome browser to check for the presence of a SNP in the list of top hits, further investigation into recent literature revealed that this CpG site was in fact a SNP (Naeem *et al.*, 2014) which had not previously been detected during the data analysis phase of this project. This difference in DNA methylation observed at this CpG site is likely, therefore, to be due to genetic variation. This really underscores how fast the field of epigenetics is moving with this information being published in only the most recent literature and highlights the importance of regular literature searches.

A recent study by Lunnon *et al.* (2014) identified a number of differentially methylated loci associated with Braak staging. However, none of the loci identified in my study were also identified by Lunnon *et al.* (2014). This may be due to COGFAST being a much smaller cohort and therefore lacking the power to detect these more subtle methylation differences. Alternatively, the differences in the loci identified in each of these studies could be due to the different phenotypes used. The study by Lunnon *et al.* (2014) used AD samples, whereas COGFAST is comprised of stroke patients. This theory is strengthened by the fact that the AD (non-stroke) cases used as controls in my study did not show the same DNA methylation differences as the stroke patients, suggesting that a stroke may alter DNA methylation patterns at certain loci.

8.1.3 To identify blood-based biomarkers of mild cognitive impairment in Parkinson's disease

Parkinson's disease is predominantly known as a motor disorder (Saracchi *et al.*, 2014), however it is thought that up to 80% of PD patients could also experience cognitive deficits in the form of mild cognitive impairment (PD-MCI) or the more severe Parkinson's disease dementia (PDD) (Yarnall *et al.*, 2013). As is the case with PSD, the mechanisms which result in PD-MCI and PDD are unknown. It is therefore difficult to develop treatments when the mechanisms which result in MCI are not well understood. Cognitive deficits usually precede any motor impairment (Saracchi *et al.*, 2014) and so the ability to predict those likely to develop cognitive deficits would be highly beneficial in the administration of early preventative treatments. The aim of Chapter 5 was to identify a blood-based biomarker which may be used in the prediction of newly diagnosed PD patients at increased risk of developing PD-MCI.

To address this aim, the ICICLE cohort was utilised. This cohort consists of individuals with newly diagnosed idiopathic PD. A blood sample taken at recruitment was used in the epigenetic analysis undertaken in Chapter 5 of this thesis. The methylation profile of the baseline blood samples was analysed using the HM450 BeadChip, to look for differentially methylated positions associated with cognitive (and motor) outcome at the 18 month follow up. A selection of the HM450 top hits were analysed across a wider cohort using Pyrosequencing.

These analyses identified one potential biomarker, *CHCHD5*, a gene whose function is unclear, but is thought to perform a role in the mitochondrial membrane space. Methylation at one CpG site within the *CHCHD5* gene was found to be significantly reduced in the PD cases with a higher score on the naming test, a test that specifically evaluates the participant's ability to name pictured animals. Adjustment for cellular composition revealed that this difference in methylation was not due to differences in cellular composition between samples. These results suggest that the CpG site within *CHCHD5* could be a potential biomarker of PD-MCI. DNA methylation levels were also measured at this locus in healthy controls. No association between naming score and DNA methylation were observed in healthy controls, suggesting that this differential methylation is unique to PD cases. This suggests that differences in methylation associated with naming in the PD cases may be associated with another factor, in addition to, or independently of, naming score. This may be an unmeasured confounder unrelated to cognition. The accuracy of the naming test also needs to be considered. The naming test was scored out of 3, and all participants, both cases and controls, had a score of either 2 or 3. This suggests that there is not much variation between participants, so the differential methylation observed at the *CHCHD5* locus may not be effective in discerning subtle shifts in cognition. The validity of this biomarker as a predictor of later cognition is therefore uncertain and sensitivity and specificity analyses are warranted using more granular measures of cognition.

In addition to looking for biomarkers of PD-MCI, the methylation data were also used to look for biomarkers of motor impairment. No significant associations between DNA methylation and motor outcome were able to be validated.

In conclusion, only one potential biomarker for PD-MCI was identified. Methylation at *CHCHD5* was found to be lower in participants with a higher (better) score on the naming test. This association was only true in PD cases. There was no association between methylation of *CHCHD5* and naming score in healthy controls. This association observed between *CHCHD5* methylation and naming could be true, but it is also possible that there could be some other confounder, possibly relating to cognition which has an effect on the methylation level in PD cases. It would be interesting to perform functional analyses in a PD cohort to assess whether *CHCHD5* could be acting on a mechanistic pathway leading to PD-MCI. For this, the brains of PD cases would need to be used.

8.1.4 To assess whether DNA methylation can mediate the relationship between exposure and outcome

Both PSD and PD are thought to be caused by a combination of genetic, epigenetic and environmental factors. A number of environmental factors have been identified in the literature as outlined in Chapter 6. One of the aims of this Chapter was to assess whether any exposures or risk factors were associated with outcome in the cohorts (COGFAST and ICICLE) used throughout this thesis. It was also of interest to assess whether any of these associations were mediated by DNA methylation. Very few associations were identified between an exposure/risk factor and outcome variable in both cohorts, however many of the effect estimates were in the right direction, even though p values did not attain statistical significance. This may be due to the small sample sizes of both cohorts and therefore a lack of power. In the ICICLE cohort, methylation at nine CpG sites was associated with both smoking status (exposure) and language score (outcome), however there was no evidence that the presence or absence of smoking changes the effect of methylation on language at any CpG site, that is, there is no support for a mediating role of DNA methylation in this instance. In COGFAST, no CpG site was identified that was associated with both an exposure and outcome variable. In both the PSD and PD cohorts utilised in this thesis, methylation was not found to interact with any exposure to affect outcome. It may be that DNA methylation is not involved in the pathway linking any of the measured exposures with outcome. However, it may be a power issue and elucidating this relationship would benefit from repeating this analysis in a larger cohort.

8.1.5 To assess whether DNA methylation can be used as an indicator of exposure status

There is evidence to suggest that blood DNA methylation levels could be a useful indicator of exposure status (Shenker *et al.,* 2013; Elliott *et al.,* 2014). This is particularly useful for exposures which are frequently under-reported, difficult to accurately measure or simply unavailable. If DNA methylation was found to accurately predict exposure level, rather than relying on self-reported data that could be confounded by recall or reporting bias, DNA methylation could be used to predict the exposure level. DNA methylation has previously been identified as an indicator of smoking status and can be used to distinguish between smokers and never smokers (Elliott et al 2014) and never and former smokers (Shenker *et al.,* 2013). One of the aims of Chapter 6 was to assess whether DNA methylation could be used to predict exposure level in one of the cohorts used in this thesis.

Using the method outlined by Elliott *et al.* (2014), a methylation score was predicted for each ICICLE sample and a threshold score was used to distinguish between current and never smokers. The data presented in this Chapter show that DNA methylation can be used to successfully predict smoking status, with current smokers exhibiting a different methylation profile to never and ex-smokers. This finding supports previous research which suggests that over time, the methylation profiles of ex-smokers reverts back to that of never smokers (Zeilinger *et al.*, 2013; Elliott *et al.*, 2014).

In conclusion, DNA methylation was used to distinguish between current smokers and ex/never smokers in the ICICLE cohort. The ability to predict smoking status using DNA methylation data is a very valuable resource with high clinical utility due to smoking-associated reporting bias. The success of this approach in predicting smoking exposure suggests that it would also be useful in the prediction of other exposure levels that are prone to the same reporting bias' and inaccurate methods of measurement. Using a larger sample set to enable samples to be split into a training and testing set would allow more exposures, such as homocysteine, to be assessed in this way.

8.1.6 To assess whether age predicted from DNA methylation can identify individuals at risk of cognitive decline

Epigenetic factors are thought to affect the ageing process with growing evidence of differential methylation observed in elderly subjects (Fraga *et al.*, 2005; Talens *et al.*, 2012) and the suggestion that DNA damage can accelerate the ageing process (Campisi and Vijg, 2009). Due to these temporal associations and the tissue-specific nature of DNA methylation, it has been suggested that DNA methylation could be used to predict the age of specific tissues and cell types (Horvath, 2013). The aim of Chapter 7 was to predict the age of tissues using DNA methylation data and assess whether this predicted age could determine individuals at greatest risk of cognitive decline. It was also of interest to assess whether any exposure variables were associated with predicted age of tissues and discordance between predicted and actual age.

To address these aims, blood and brain samples (DLPFC and hippocampus) from the COGFAST cohort and blood samples from the ICICLE cohort were utilised. The methylation profile of each sample was analysed using the HM450 BeadChip and the epigenetic clock method (Horvath, 2013) was used to predict the age of each sample. Differences were observed between the chronological age and predicted age of the sample and an age acceleration residual (AAR) (DNA methylation age adjusted for chronological age) was calculated for each sample. Regression analyses were then performed to look for association between AAR and outcome variables, and AAR and exposure variables. In neither cohort, nor tissue, was there any association reaching the Bonferroni corrected significance threshold identified. A number of other associations significant at the uncorrected significance threshold of p<0.05 were identified however. In the ICICLE cohort, a weak association was observed between AAR and MDS-UPDRS III (Rho=-0.131, p=0.041), suggesting that a lower methylation age was associated with a more severe motor phenotype and between AAR and diabetes (F=3.44, p=0.042), suggesting that a higher methylation age was associated with an increased risk of diabetes. In the COGFAST cohort, the blood AAR was not associated with any outcome variable but was found to be associated with smoking status (F=3.92, p=0.036), indicating that individuals who smoke have a higher methylation age. No associations were observed between AAR and any exposure or outcome variable in the brain samples.

The COGFAST cohort provided an excellent opportunity to assess whether the blood could act as a surrogate to predict DNA methylation age of the brain. No differences were observed between the three tissue types (blood, DLPFC, hippocampus) indicating that; 1) the brain does not age differently to blood in the tissues and samples included in this analysis and 2) in this case, the blood can be used as a surrogate to predict DNA methylation age of brain tissue. This could be useful in future studies where blood samples were taken from living individuals. The observation that the blood and brain have similar methylation profiles is in agreement with previous reports (Davies *et al.*, 2012; Oh *et al.*, 2014), however here a methylation score is used representing 353 CpG sites.

In conclusion, the predicted age of samples did significantly differ from the chronological age of samples in both cohorts. However, no associations were observed between AAR and any exposure or outcome, indicating that in these cohorts, the AAR is not hugely influenced by exposures, nor is the AAR itself influencing disease outcome. The lack of association between AAR and exposure/outcome variables may indicate that the predicted age of samples based on DNA methylation levels (which could be interpreted as a form of biological ageing) is not involved in these disease pathways. Repetition of this study in a larger cohort could help to clarify this.

The main strength of this aim was that the COGFAST cohort enabled the relationship between AAR across the different tissues to be assessed. The results showed that the DNA methylation profile of the blood did not differ from either brain region. This suggests that the blood could therefore be used as a surrogate for brain tissue in studies analysing DNA methylation age. This would enable current cohorts, with living participants, to predict the DNA methylation age of the brain, without the need to wait until brain samples became available and would enable more cohorts to be involved in the analysis of DNA methylation age.

8.2 Strengths of the study

Biomarkers are being increasingly used to aid both the prediction and prognosis of a number of diseases including cardiovascular disease (Kim *et al.*, 2010) and cancer (Schwalbe *et al.*, 2013), but to date there are no published studies that have assessed blood-based biomarkers to aid the prediction of stroke survivors at risk of developing PSD. In addition, there have been no published studies with the aim of identifying biomarkers of PD-MCI. The two main aims of this thesis; the identification of biomarkers for PSD and PD-MCI are therefore addressing novel areas of epigenetic research. This thesis highlighted two potential biomarkers which may be useful in the prediction of PSD and one potential biomarker for the prediction of PD-MCI. These novel findings could, if robustly replicated, greatly benefit public health care services in the administration of preventative treatments and care.

A key strength of this study was the inclusion of brain tissue to assess possible mechanisms involved in the development of PSD. Post-mortem studies are still relatively uncommon, due to the difficulty in obtaining such samples, as are paired blood and post-mortem brain tissue from the same individuals. The latter allowed valuable comparisons to be made between different tissue types. Although there have been questions highlighted over the quality of post-mortem tissue and the different storage conditions that may be experienced (Ferrer *et al.*, 2008; Pidsley and Mill, 2011), this study identified no associations between DNA methylation at any CpG site included on the HM450 BeadChip and PM delay.

In addition to the novelty of this study, it was also strengthened by the addition of controls in both cohorts. The addition of blood and brain samples taken from non-stroke cases with dementia highlighted that DNA methylation differences were unique to COGFAST samples. There were no differences in DNA methylation across Braak stages in non-stroke cases with dementia in either brain tissue or blood detectable using the methods I applied. Similarly, the addition of healthy controls in the ICICLE cohort allowed for the discovery that DNA methylation was associated with naming score in PD cases alone.

This study also attempted to investigate, not only DNA methylation differences between samples, but also extended this to functional analyses. Although the quality of RNA was not suitable for gene expression analyses due to the long-term storage of the post-mortem tissue and resultant degradation of RNA, protein extracted from both brain regions was used for Western blotting. This was an important feature of the study required to assess possible mechanisms involved in the pathogenesis of PSD. Although no differences in NGF protein levels were observed across brain samples with a range of Braak staging, this may be due to the small sample number used and the lack of availability of protein from the COGFAST Braak stage VI sample.

A further strength of this study was the consideration of cellular composition which is recognised as a possible confounder in epigenetic studies. Due to the heterogeneous nature of both blood and brain tissue, it is possible that any differences in DNA methylation observed between outcome groups is due to the cellular composition of the blood and brain. Adjusting the methylation beta values for cellular composition enabled any effect of cellular composition on DNA methylation to be identified. In the blood samples, very little effect of cellular composition was identified. In the brain samples, cellular composition was seen to have a greater effect on DNA methylation than in the blood, however these effects were, for the most part, very small. Other confounders known to affect DNA methylation profiles (sex and age) were also accounted for in the analysis.

8.3 Limitations of the study

The key limitation of this study was the small sample sizes used, particularly that of the COGFAST cohort. Sample size is an issue affecting many EWAS due to the expense of performing such a study (Rakyan *et al.*, 2011; Michels *et al.*, 2013). In a recent review by Michels *et al.* (2013) which summarised 257 EWAS, 90 of these were identified to utilise 30 or less samples (i.e. were of the same or smaller size than the COGFAST samples used in the HM450 BeadChip), and 143 of these included 48 or less samples (i.e. were of the same or smaller size than the ICICLE samples used in the HM450 BeadChip) (Michels *et al.*, 2013).

As discussed in Chapters 3 and 4, the finding that Braak staging was associated with DNA methylation at certain loci was driven by DNA methylation levels in one individual. With the small sample size of COGFAST, it is impossible to tell whether this is a true biological association and a common finding in all stroke survivors with Braak stage VI pathology, or whether this single individual with Braak stage VI pathology phenotypically differs from the other COGFAST participants in some other way. One possible difference is the discovery that this individual had actually experienced recurrent strokes prior to recruitment. This could have affected the methylation level, as hypothesised in Chapter 4, or this could just be something inherent to that individual and DNA methylation differences may not be related to stroke plus Braak stage VI pathology or recurrent strokes. To elucidate which possible theory is correct, this study requires replication in a much larger cohort.

Due to there being only one COGFAST participant with Braak stage VI pathology, there was limited availability of this brain tissue. There was no tissue available from this individual for the protein analysis. The omission of this sample from the Western blotting prevented the relationship between *NGF* methylation and NGF protein levels to be deduced in this individual. It was therefore impossible to conclude whether *NGF* could have a mechanistic (or causal) role in the pathogenesis of PSD based upon the evidence at my disposal.

Another limitation of the study was that information relating to subsequent strokes experienced after recruitment into COGFAST was not available. The study design of COGFAST did not allow for recurrent strokes occurring after the index stroke to be recorded. Consequently, it is possible that a number of participants had multiple strokes once they were recruited onto the study. These unknown strokes may have influenced both DNA methylation levels and cognition recorded after baseline measures were taken. For example, it is possible that those participants who developed dementia following the index stroke had experienced additional strokes thereby making them more susceptible to cognitive decline. This information would be vital for assessing the effect of multiple strokes on cognition.

In Chapter 6, the small sample size of ICICLE also limited which exposure variables could be used to test the utility of DNA methylation as an exposure indicator. Homocysteine and smoking were identified as suitable candidate variables. Due to there not being an independent homocysteine reference data set and the size of the ICICLE cohort being too small to split the sample set into two, it was deemed neither ideal nor reliable to use DNA methylation data to predict homocysteine levels. If the sample size of ICICLE had been larger, it would have been possible to split the cohort into two and use one half as a training set and the other as a testing set. In this circumstance it would have been possible to use DNA methylation to try to better predict homocysteine levels. As an alternative, DNA methylation was used to predict smoking status which, although had already been done and therefore this Chapter did not add anything novel to current findings, it does support previous research and suggests that DNA methylation data can be used to predict smoking status in a PD cohort, and its predictive utility is unconfounded by disease status.

Another limitation of the study concerned the reliability of some of the outcome variables measured in ICICLE. In Chapter 5, DNA methylation at a CpG site within the *CHCHD5* gene was found to be associated with naming score in PD participants. The naming test was scored out of three. All participants scored either 2 or 3 points suggesting that either the test is not sensitive enough to detect large differences in naming ability, or that the participants’ ability did not differ greatly. This means that if the association between *CHCHD5* methylation and naming was successfully replicated in an independent cohort, the biomarker would only be clinically useful at predicting specific differences in cognition. Due to no other associations being identified between *CHCHD5* methylation and cognition, it is unlikely that the biomarker would be able to predict other cognitive deficits affecting PD patients.

Top HM450 hits were selected if they had an effect size greater than 5%, were not SNPs and if suitable primers for Pyrosequencing could be designed using the PSQ assay design software. There is some evidence that the Pyrosequencer cannot reliably measure differences of less than 5% (Mikeska *et al.*, 2011), hence an effect size cut off of 5% was implemented. This may have removed potential biomarkers of cognitive decline. Other hits which were not amenable to Pyrosequencing, such as those where suitable primers could not be designed, were also removed. These could however be potential biomarkers, but they were not considered in this analysis. These CpG sites were removed from the selection because they were not suitable for Pyrosequencing analysis. Another selection criterion was whether HM450 hits had previously been associated with differential expression in AD/PD. CpG sites were only taken forward if there was an inverse relationship between methylation and gene expression. This inverse relationship is expected in promoter regions, but it is less clear if this relationship is found in other regions of the gene. For this reason, potential biomarkers may have been removed from consideration because they did not show the typical inverse relationship. To widen the search for a potential biomarker, it may have been more suitable to increase the selection, to include all CpG sites which were within a gene previously shown to be differentially expressed in AD/PD. Furthermore, assays targeting potential biomarkers may have also been omitted from further consideration based on their failure to optimise or failure to validate on the Pyrosequencer. These failures do not make them unsuitable biomarkers. These assays may be better suited to analysis using an alternative platform, as suggested above. In Chapter 5 of this thesis, another approach was adopted to target assays which had failed to be optimised. Predesigned Pyromark assays (Qiagen, UK) were used for five of these assays, however these assays did not target the index CpG site and unsurprisingly, DNA methylation measured using these assays did not show any associations with outcome variables of interest.

Finally, cellular composition analysis was undertaken to assess whether observed differences in DNA methylation, were in fact, due to cellular composition. This analysis does strengthen the study, however the technique used does come with some limitations. Blood samples were adjusted for cellular composition using the algorithm described by Houseman *et al.* (2012) and the brain samples were adjusted for cellular composition using the CETS package (Guintivano *et al.*, 2013). Both of these methods use algorithms to estimate the proportions of cellular components in the samples. The cellular components of blood estimated using this technique are B cells, granulocytes, monocytes, NK cells, and T cells subsets and the proportions of neurons and glia are estimated in the brain samples. This is an estimation of cell counts and is therefore not entirely accurate and is susceptible to some error. The only way to accurately calculate the cellular proportions of each sample is to measure the proportion of each cellular component in each sample itself. In the blood, FACS could be performed to count the number of each cellular component and in the brain, laser capture microdissection (LCM) could be performed to separate neurons from glia. DNA methylation analysis would then be performed on each cell type to assess whether the DNA methylation profiles of samples differ due to cellular composition. Both techniques would be more time-consuming and would require considerably larger amounts of initial sample than using the estimation technique. However, the estimation techniques, although imperfect, are deemed suitable for use in many circumstances due to their high accuracy and time and sample limitations of other, more accurate methods.

8.4 Future work

There are a number of steps which could be taken to further the research performed in this thesis. Possible next steps are explored below.

As previously alluded to, this study requires replication in a larger cohort. The size of COGFAST, in particular, was too small to be able to determine whether the association between DNA methylation and Braak staging was true, or whether the differences in DNA methylation were associated with some other factor, such as recurrent stroke. Adding more samples, particularly those with the most severe Braak staging pathology, into the analysis would help to determine the reason for this association. COGFAST is still an ongoing living cohort with brain samples still to be collected. This, together with associated brain pathology data may increase the number of both blood and brain samples with an associated Braak stage VI pathology. Once all COGFAST brain samples are collected it would be of interest to repeat the analysis performed here to assess whether the increase in sample number provides a more robust detection of DNA methylation differences. Future work could also focus on exploring other reasons for this large difference in DNA methylation such as recurrent strokes. Including more recurrent stroke samples into the study would help to elucidate the cause of this large methylation difference. Another possible explanation, and possibly the most likely explanation for the large differences in DNA methylation observed in this project, is that of genetics. The study by Naeem *et al.* (2014) considers the *NGF* locus to be a SNP. Genotyping data were not available for this individual so the only means of obtaining conclusive evidence for this would be by way of sequencing.

A power calculation was performed to estimate the number of samples required to detect significant DNA methylation differences between groups. Based on the difference in blood *NGF* DNA methylation (the loci with the smallest difference in DNA methylation between groups) observed (14%) between Braak stage VI and the other Braak stages (I-V) with a power of 99.9% and alpha=0.05, I would require only one additional Braak stage VI sample to significantly detect these methylation differences, assuming that the difference in methylation observed in this individual is due to Braak stage pathology. This may be feasible within the COGFAST cohort.

In addition, there is scope for the ICICLE cohort to include brain samples for each participant. DNA methylation analysis of the blood, as was used in this thesis, enables us to look for biomarkers which may predict those participants most at risk of developing PD-MCI, but cannot be informative about mechanisms resulting in disease. The addition of brain samples into this cohort would enable the involvement of DNA methylation in the mechanisms resulting in PD-MCI to be assessed.

To be accepted as a clinically useful biomarker, any potential biomarkers would need to be analysed, and findings robustly replicated in a large independent cohort.

This study could also be expanded by using less rigorous selection criteria for top HM450 hits. Potential biomarkers may have been lost in the locus prioritisation process used in this thesis. SNPs were removed from the top HM450 hits due to Pyrosequencing not being the most ideal method of looking at SNPs. However, these SNPs may be useful biomarkers of either PSD or PD. It would be interesting to extend the work performed in this thesis and take forward the SNPs highlighted as being associated with disease for further analysis. All SNPs associated with outcome in this thesis were searched in the NHGRI GWAS Catalog (www.genome.gov/gwastudies), but none of these SNPs had previously been associated with PSD, PD or any related disease. This thesis may have identified novel SNPS so it would be of interest to follow up each of the SNPs identified here to assess whether any are potential risk factors for disease.

In addition to SNPs being removed, hits which had previously been found to be differentially expressed in disease were removed from the selection if they did not display an inverse relationship between DNA methylation and gene expression. As previously discussed, this relationship is only expected at promoter regions, so this selection criterion could have removed some potential biomarkers. If enough time and funds were available it would be interesting to assess each of these hits, regardless of the directionality to widen the pool of potential biomarkers.

Furthermore, HM450 hits were removed from the selection if suitable primers with a score >70 on the PSQ assay design software could not be designated. As this does not necessarily make them unsuitable biomarkers, these hits could be analysed using an alternative platform such as Sequenom EpiTYPER (Ehrich *et al.*, 2005) or bisulphite sequencing (Herman *et al.*, 1996). In addition, these techniques could also be applied to those assays which failed optimisation and validation.

In Chapter 4, Western blotting was used to assess the functional relevance of the differential methylation observed at the *NGF* locus. Due to the extracted RNA being too degraded for gene expression analysis, protein analysis was performed. The Western blotting technique used is able to assess levels of protein and this can then be linked with DNA methylation levels, however, this technique does not inform about causality and the direction of effect. Other approaches, such as a cell culture approach, could be taken to assess the directionality. Cells could be cultured in the presence of 5-azacytidine in order to cause genome-wide demethylation (Enright *et al.*, 2003). The expression of the target gene could be assessed at a transcript level using quantitative PCR (Yang *et al.*, 2002) and at a protein level using Western blotting (Palii *et al.*, 2008). This would determine whether the expression of the gene in question is affected by DNA methylation. The link between transcription and translation for the target gene could be assessed through the use of small interfering RNAs (siRNA) which would transiently knock down the level of gene transcript. Conversely, expression vectors could be used to drive the expression of the gene of interest (Alekseev *et al.*, 2009). The subsequent effects on protein level could then be assessed by Western blotting. In addition, since CT49 is a non-coding RNA, and only protein samples were available for the COGFAST cohort, functional analyses were not undertaken, however, it would be interesting to assess CT49 further to determine whether CT49 could provide a mechanistic link between DNA methylation and PSD. This could be assessed firstly by quantifying expression levels using quantitative PCR. If an association between DNA methylation and expression was established in these cases, a functional study could be performed. The biological function of CT49 is currently unknown, however it is thought that many long non-coding RNAs have a functional role (Mercer *et al.*, 2009; Dinger *et al.*, 2009), for example, in the regulation of gene-specific transcriptional regulation (Goodrich and Kugel, 2006). After driving the expression of CT49 using an expression construct, an expression array could be utilised to determine which genes, if any, are regulated by CT49. These targets could then be followed up to determine their role in PSD.

A common problem with EWAS is that the majority are cross-sectional studies (Michels *et al.*, 2013), with reverse causation being a possible problem. Therefore, a cross-sectional study, which is focussed on one time point cannot be used to infer causality (Relton and Davey Smith, 2010; Mill and Heijmans, 2013). To assess causality, multiple time points using the same tissue are required, or at least can be helpful in establishing temporal relationships (Ng *et al.*, 2012). In diseases affecting the brain this is impossible, since the brain is only accessible post-mortem. This study tried to overcome this by assessing the DNA methylation profiles of the blood prior to disease onset; however these methylation differences may reflect tissue-specificity rather than temporal differences and in this case cannot be used to strengthen causality. A method recently developed to overcome the problem with inferring causality in a cross-sectional study is that of two-step Mendelian randomisation (Relton and Davey Smith, 2012). Firstly, a SNP is used as a genetic proxy for the environmentally modifiable exposure which is investigated in regards to DNA methylation. This is used to determine whether the exposure causes DNA methylation. Secondly, an additional SNP is used as a proxy for DNA methylation levels to assess the relationship (and causality) between the methylation site of interest and disease phenotype or outcome (Relton and Davey Smith, 2012). In this study, the main aim was to identify potential biomarkers of PSD and PD. In this circumstance, disease causality was not of high priority as a biomarker does not need to be involved in disease aetiology in order to be a robust and sensitive biomarker for disease.

This thesis aimed to identify DMPs associated with disease. This approach identified single CpG sites which are associated with disease and was unable to inform about DNA methylation at neighbouring CpG sites. An alternative approach would be to look at differentially methylation regions (DMRs). DMRs are genomic regions in which correlated groups of CpG sites show differential methylation. The identification of DMRs, as opposed to DMPs, is considered to reduce the risk of false positives and technical artefacts. A DMR is thought to be more likely to have a biological role than a DMP, as neighbouring CpG sites also demonstrate differential methylation. HM450 data can be analysed in such a way that DMRs are identified rather than DMPs using methods such as bump hunting (Jaffe *et al.*, 2012).

Another way of reducing the risk of false positives is to target the search to only those within genes previously found to be associated with disease. Both hits identified as being associated with Braak staging in Chapter 3 had previously been associated with PSD related disorders. Had this strategy been applied, these hits would have still been identified; however this approach would obviously have limited the ability to identify novel biomarkers and novel genes associated with disease mechanisms. A candidate gene approach could be used in order to reduce the burden of multiple testing and to keep the number of hits to validate low.

8.5 Conclusion

In conclusion, this study identified possible biomarkers of both PSD and PD, however both studies were limited by sample size. Care must therefore be taken in the interpretation of the findings and further replication is of crucial importance.

References

Aarsland, D. and Kurz, M.W. (2010) 'The epidemiology of dementia associated with Parkinson's disease', *Brain Pathol*, 20(3), pp. 633-9.

Abrahams, S., Pickering, A., Polkey, C.E. and Morris, R.G. (1997) 'Spatial memory deficits in patients with unilateral damage to the right hippocampal formation', *Neuropsychologia*, 35(1), pp. 11-24.

Adair, S.J. and Hogan, K.T. (2009) 'Treatment of ovarian cancer cell lines with 5-aza-2'-deoxycytidine upregulates the expression of cancer-testis antigens and class I major histocompatibility complex-encoded molecules', *Cancer Immunol Immunother*, 58(4), pp. 589-601.

Ahmed, S.S., Santosh, W., Kumar, S. and Christlet, H.T. (2009) 'Metabolic profiling of Parkinson's disease: evidence of biomarker from gene expression analysis and rapid neural network detection', *J Biomed Sci*, 16, p. 63.

Alashwal, H., Dosunmu, R. and Zawia, N.H. (2012) 'Integration of genome-wide expression and methylation data: relevance to aging and Alzheimer's disease', *Neurotoxicology*, 33(6), pp. 1450-3.

Alekseev, O.M., Richardson, R.T., Alekseev, O. and O'Rand, M.G. (2009) 'Analysis of gene expression profiles in HeLa cells in response to overexpression or siRNA-mediated depletion of NASP', *Reprod Biol Endocrinol*, 7, p. 45.

Allan, L.M., Rowan, E.N., Firbank, M.J., Thomas, A.J., Parry, S.W., Polvikoski, T.M., O'Brien, J.T. and Kalaria, R.N. (2011) 'Long term incidence of dementia, predictors of mortality and pathological diagnosis in older stroke survivors', *Brain*, 134(Pt 12), pp. 3716-27.

Aloe, L., Rocco, M.L., Bianchi, P. and Manni, L. (2012) 'Nerve growth factor: from the early discoveries to the potential clinical use', *J Transl Med*, 10, p. 239.

Amarenco, P., Bogousslavsky, J., Callahan, A., 3rd, Goldstein, L.B., Hennerici, M., Rudolph, A.E., Sillesen, H., Simunovic, L., Szarek, M., Welch, K.M. and Zivin, J.A. (2006) 'High-dose atorvastatin after stroke or transient ischemic attack', *N Engl J Med*, 355(6), pp. 549-59.

Anderson, C.M., Ralph, J.L., Wright, M.L., Linggi, B. and Ohm, J.E. (2013) 'DNA methylation as a biomarker for preeclampsia', *Biol Res Nurs*.

Anderton, B.H. (2002) 'Ageing of the brain', *Mech Ageing Dev*, 123(7), pp. 811-7.

Anstey, K.J., von Sanden, C., Salim, A. and O'Kearney, R. (2007) 'Smoking as a risk factor for dementia and cognitive decline: a meta-analysis of prospective studies', *Am J Epidemiol*, 166(4), pp. 367-78.

Baccarelli, A., Wright, R., Bollati, V., Litonjua, A., Zanobetti, A., Tarantini, L., Sparrow, D., Vokonas, P. and Schwartz, J. (2010) 'Ischemic heart disease and stroke in relation to blood DNA methylation', *Epidemiology*, 21(6), pp. 819-28.

Bakulski, K.M., Dolinoy, D.C., Sartor, M.A., Paulson, H.L., Konen, J.R., Lieberman, A.P., Albin, R.L., Hu, H. and Rozek, L.S. (2012a) 'Genome-wide DNA methylation differences between late-onset Alzheimer's disease and cognitively normal controls in human frontal cortex', *J Alzheimers Dis*, 29(3), pp. 571-88.

Bakulski, K.M. and Fallin, M.D. (2014) 'Epigenetic epidemiology: promises for public health research', *Environ Mol Mutagen*, 55(3), pp. 171-83.

Bakulski, K.M., Rozek, L.S., Dolinoy, D.C., Paulson, H.L. and Hu, H. (2012b) 'Alzheimer's disease and environmental exposure to lead: the epidemiologic evidence and potential role of epigenetics', *Curr Alzheimer Res*, 9(5), pp. 563-73.

Ballard, C., Rowan, E., Stephens, S., Kalaria, R. and Kenny, R.A. (2003) 'Prospective follow-up study between 3 and 15 months after stroke: improvements and decline in cognitive function among dementia-free stroke survivors >75 years of age', *Stroke*, 34(10), pp. 2440-4.

Baltan, S., Besancon, E.F., Mbow, B., Ye, Z., Hamner, M.A. and Ransom, B.R. (2008) 'White matter vulnerability to ischemic injury increases with age because of enhanced excitotoxicity', *J Neurosci*, 28(6), pp. 1479-89.

Bamford, J., Sandercock, P., Dennis, M., Burn, J. and Warlow, C. (1991) 'Classification and natural history of clinically identifiable subtypes of cerebral infarction', *Lancet*, 337(8756), pp. 1521-6.

Banci, L., Bertini, I., Ciofi-Baffoni, S., Jaiswal, D., Neri, S., Peruzzini, R. and Winkelmann, J. (2012) 'Structural characterization of CHCHD5 and CHCHD7: two atypical human twin CX9C proteins', *J Struct Biol*, 180(1), pp. 190-200.

Bannwarth, S., Ait-El-Mkadem, S., Chaussenot, A., Genin, E.C., Lacas-Gervais, S., Fragaki, K., Berg-Alonso, L., Kageyama, Y., Serre, V., Moore, D.G., Verschueren, A., Rouzier, C., Le Ber, I., Auge, G., Cochaud, C., Lespinasse, F., N'Guyen, K., de Septenville, A., Brice, A., Yu-Wai-Man, P., Sesaki, H., Pouget, J. and Paquis-Flucklinger, V. (2014) 'A mitochondrial origin for frontotemporal dementia and amyotrophic lateral sclerosis through CHCHD10 involvement', *Brain*, 137(Pt 8), pp. 2329-45.

Barbey, A.K., Koenigs, M. and Grafman, J. (2013) 'Dorsolateral prefrontal contributions to human working memory', *Cortex*, 49(5), pp. 1195-205.

Barfield, R.T., Kilaru, V., Smith, A.K. and Conneely, K.N. (2012) 'CpGassoc: an R function for analysis of DNA methylation microarray data', *Bioinformatics*, 28(9), pp. 1280-1.

Barker, D.J. (2006) 'Adult consequences of fetal growth restriction', *Clin Obstet Gynecol*, 49(2), pp. 270-83.

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C.L., Serova, N., Davis, S., Soboleva, A. (2013) 'NCBI GEO:

archive for functional genomics data sets - update', *Nucleic Acids Res.* 41:D991-5. Available at: http://www.ncbi.nlm.nih.gov/geo/. Accessed: 02/09/14.

Bartolomei, M.S. (2009) 'Genomic imprinting: employing and avoiding epigenetic processes', *Genes Dev*, 23(18), pp. 2124-33.

Beck, S. and Rakyan, V.K. (2008) 'The methylome: approaches for global DNA methylation profiling', *Trends Genet*, 24(5), pp. 231-7.

Bell, J.T. and Spector, T.D. (2011) 'A twin approach to unraveling epigenetics', *Trends Genet*, 27(3), pp. 116-25.

Bell, J.T., Tsai, P.C., Yang, T.P., Pidsley, R., Nisbet, J., Glass, D., Mangino, M., Zhai, G., Zhang, F., Valdes, A., Shin, S.Y., Dempster, E.L., Murray, R.M., Grundberg, E., Hedman, A.K., Nica, A., Small, K.S., Dermitzakis, E.T., McCarthy, M.I., Mill, J., Spector, T.D. and Deloukas, P. (2012) 'Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population', *PLoS Genet*, 8(4), p. e1002629.

Benn, M. (2009) 'Apolipoprotein B levels, APOB alleles, and risk of ischemic cardiovascular disease in the general population, a review', *Atherosclerosis*, 206(1), pp. 17-30.

Benn, M., Nordestgaard, B.G., Jensen, G.B. and Tybjaerg-Hansen, A. (2007) 'Improving prediction of ischemic cardiovascular disease in the general population using apolipoprotein B: the Copenhagen City Heart Study', *Arterioscler Thromb Vasc Biol*, 27(3), pp. 661-70.

Berndt, S.I., Gustafsson, S., Magi, R., Ganna, A., Wheeler, E., Feitosa, M.F., Justice, A.E., Monda, K.L., Croteau-Chonka, D.C., Day, F.R., Esko, T., Fall, T., Ferreira, T., Gentilini, D., Jackson, A.U., Luan, J., Randall, J.C., Vedantam, S., Willer, C.J., Winkler, T.W., Wood, A.R., Workalemahu, T., Hu, Y.J., Lee, S.H., Liang, L., *et al.* (2013) 'Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture', *Nat Genet*, 45(5), pp. 501-12.

Bevan, S. and Markus, H.S. (2011) 'Genetics of common polygenic ischaemic stroke: current understanding and future challenges', *Stroke Res Treat*, 2011, p. 179061.

Bevan, S., Traylor, M., Adib-Samii, P., Malik, R., Paul, N.L., Jackson, C., Farrall, M., Rothwell, P.M., Sudlow, C., Dichgans, M. and Markus, H.S. (2012) 'Genetic heritability of ischemic stroke and the contribution of previously reported candidate gene and genomewide associations', *Stroke*, 43(12), pp. 3161-7.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J.M., Delano, D., Zhang, L., Schroth, G.P., Gunderson, K.L., Fan, J.B. and Shen, R. (2011) 'High density DNA methylation array with single CpG site resolution', *Genomics*, 98(4), pp. 288-95.

Biomarkers Definitions Working Group. (2001) 'Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework', *Clin. Pharmacol. Ther.*, 69, 89-95.

Bogdanov, M., Matson, W.R., Wang, L., Matson, T., Saunders-Pullman, R., Bressman, S.S. and Flint Beal, M. (2008) 'Metabolomic profiling to develop blood biomarkers for Parkinson's disease', *Brain*, 131(Pt 2), pp. 389-96.

Boks, M.P., Mierlo, H.C., Rutten, B.P., Radstake, T.R., De Witte, L., Geuze, E., Horvath, S., Schalkwyk, L.C., Vinkers, C.H., Broen, J.C. and Vermetten, E. (2014) 'Longitudinal changes of telomere length and epigenetic age related to traumatic stress and post-traumatic stress disorder', *Psychoneuroendocrinology*.

Bollati, V., Galimberti, D., Pergoli, L., Dalla Valle, E., Barretta, F., Cortini, F., Scarpini, E., Bertazzi, P.A. and Baccarelli, A. (2011) 'DNA methylation in repetitive elements and Alzheimer disease', *Brain Behav Immun*, 25(6), pp. 1078-83.

Booth, M.J., Branco, M.R., Ficz, G., Oxley, D., Krueger, F., Reik, W. and Balasubramanian, S. (2012) 'Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution', *Science*, 336(6083), pp. 934-7.

Boyes, J. and Bird, A. (1991) 'DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein', *Cell*, 64(6), pp. 1123-34.

Braak, H. and Braak, E. (1995) 'Staging of Alzheimer's disease-related neurofibrillary changes', *Neurobiol Aging*, 16(3), pp. 271-8; discussion 278-84.

Breiman, L. (2001) 'Random Forests', *Machine Learning*, (45), pp. 5-32.

Breitling, L.P., Yang, R., Korn, B., Burwinkel, B. and Brenner, H. (2011) 'Tobacco-smoking-related differential DNA methylation: 27K discovery and replication', *Am J Hum Genet*, 88(4), pp. 450-7.

Brennan, K., Garcia-Closas, M., Orr, N., Fletcher, O., Jones, M., Ashworth, A., Swerdlow, A., Thorne, H., Riboli, E., Vineis, P., Dorronsoro, M., Clavel-Chapelon, F., Panico, S., Onland-Moret, N.C., Trichopoulos, D., Kaaks, R., Khaw, K.T., Brown, R. and Flanagan, J.M. (2012) 'Intragenic ATM methylation in peripheral blood DNA as a biomarker of breast cancer risk', *Cancer Res*, 72(9), pp. 2304-13.

Brock, M.V., Gou, M., Akiyama, Y., Muller, A., Wu, T.T., Montgomery, E., Deasel, M., Germonpre, P., Rubinson, L., Heitmiller, R.F., Yang, S.C., Forastiere, A.A., Baylin, S.B. and Herman, J.G. (2003) 'Prognostic importance of promoter hypermethylation of multiple genes in esophageal adenocarcinoma', *Clin Cancer Res*, 9(8), pp. 2912-9.

Bruno, M.A., Leon, W.C., Fragoso, G., Mushynski, W.E., Almazan, G. and Cuello, A.C. (2009) 'Amyloid beta-induced nerve growth factor dysmetabolism in Alzheimer disease', *J Neuropathol Exp Neurol*, 68(8), pp. 857-69.

Brustovetsky, T., Li, V. and Brustovetsky, N. (2009) 'Stimulation of glutamate receptors in cultured hippocampal neurons causes Ca2+-dependent mitochondrial contraction', *Cell Calcium*, 46(1), pp. 18-29.

Burn, J., Dennis, M., Bamford, J., Sandercock, P., Wade, D. and Warlow, C. (1997) 'Epileptic seizures after a first stroke: the Oxfordshire Community Stroke Project', *BMJ*, 315(7122), pp. 1582-7.

Byun, H.M., Siegmund, K.D., Pan, F., Weisenberger, D.J., Kanel, G., Laird, P.W. and Yang, A.S. (2009) 'Epigenetic profiling of somatic tissues from human autopsy specimens identifies tissue- and individual-specific DNA methylation patterns', *Hum Mol Genet*, 18(24), pp. 4808-17.

Campisi, J. and Vijg, J. (2009) 'Does damage to DNA and other macromolecules play a role in aging? If so, how?', *J Gerontol A Biol Sci Med Sci*, 64(2), pp. 175-8.

Caramelli, P., Nitrini, R., Maranhao, R., Lourenco, A.C., Damasceno, M.C., Vinagre, C. and Caramelli, B. (1999) 'Increased apolipoprotein B serum concentration in Alzheimer's disease', *Acta Neurol Scand*, 100(1), pp. 61-3.

Cardarelli, R., Kertesz, A. and Knebl, J.A. (2010) 'Frontotemporal dementia: a review for primary care physicians', *Am Fam Physician*, 82(11), pp. 1372-7.

Casas, J.P., Bautista, L.E., Smeeth, L., Sharma, P. and Hingorani, A.D. (2005) 'Homocysteine and stroke: evidence on a causal link from mendelian randomisation', *Lancet*, 365(9455), pp. 224-32.

Castillo-Diaz, S.A., Garay-Sevilla, M.E., Hernandez-Gonzalez, M.A., Solis-Martinez, M.O. and Zaina, S. (2010) 'Extensive demethylation of normally hypermethylated CpG islands occurs in human atherosclerotic arteries', *Int J Mol Med*, 26(5), pp. 691-700.

Castro, R., Rivera, I., Struys, E.A., Jansen, E.E., Ravasco, P., Camilo, M.E., Blom, H.J., Jakobs, C. and Tavares de Almeida, I. (2003) 'Increased homocysteine and S-adenosylhomocysteine concentrations and DNA hypomethylation in vascular disease', *Clin Chem*, 49(8), pp. 1292-6.

Chaussenot, A., Le Ber, I., Ait-El-Mkadem, S., Camuzat, A., de Septenville, A., Bannwarth, S., Genin, E.C., Serre, V., Auge, G., Brice, A., Pouget, J. and Paquis-Flucklinger, V. (2014) 'Screening of CHCHD10 in a French cohort confirms the involvement of this gene in frontotemporal dementia with amyotrophic lateral sclerosis patients', *Neurobiol Aging*.

Chen, R.L., Balami, J.S., Esiri, M.M., Chen, L.K. and Buchan, A.M. (2010) 'Ischemic stroke in the elderly: an overview of evidence', *Nat Rev Neurol*, 6(5), pp. 256-65.

Chobanian, A.V. (2009) 'Shattuck Lecture. The hypertension paradox--more uncontrolled disease despite improved therapy', *N Engl J Med*, 361(9), pp. 878-87.

Chouliaras, L., Mastroeni, D., Delvaux, E., Grover, A., Kenis, G., Hof, P.R., Steinbusch, H.W., Coleman, P.D., Rutten, B.P. and van den Hove, D.L. (2013) 'Consistent decrease in global DNA methylation and hydroxymethylation in the hippocampus of Alzheimer's disease patients', *Neurobiol Aging*, 34(9), pp. 2091-9.

Chouliaras, L., Rutten, B.P., Kenis, G., Peerbooms, O., Visser, P.J., Verhey, F., van Os, J., Steinbusch, H.W. and van den Hove, D.L. (2010) 'Epigenetic regulation in the pathophysiology of Alzheimer's disease', *Prog Neurobiol*, 90(4), pp. 498-510.

Claus, R., Lucas, D.M., Stilgenbauer, S., Ruppert, A.S., Yu, L., Zucknick, M., Mertens, D., Buhler, A., Oakes, C.C., Larson, R.A., Kay, N.E., Jelinek, D.F., Kipps, T.J., Rassenti, L.Z., Gribben, J.G., Dohner, H., Heerema, N.A., Marcucci, G., Plass, C. and Byrd, J.C. (2012) 'Quantitative DNA methylation analysis identifies a single CpG dinucleotide important for ZAP-70 expression and predictive of prognosis in chronic lymphocytic leukemia', *J Clin Oncol*, 30(20), pp. 2483-91.

Crider, K.S., Yang, T.P., Berry, R.J. and Bailey, L.B. (2012) 'Folate and DNA methylation: a review of molecular mechanisms and the evidence for folate's role', *Adv Nutr*, 3(1), pp. 21-38.

Cumming, T. and Brodtmann, A. (2010) 'Dementia and stroke: the present and future epidemic', *Int J Stroke*, 5(6), pp. 453-4.

Cumming, T.B. and Brodtmann, A. (2011) 'Can stroke cause neurodegenerative dementia?', *Int J Stroke*, 6(5), pp. 416-24.

Cummings, J.L., Henchcliffe, C., Schaier, S., Simuni, T., Waxman, A. and Kemp, P. (2011) 'The role of dopaminergic imaging in patients with symptoms of dopaminergic system neurodegeneration', *Brain*, 134(Pt 11), pp. 3146-66.

Dallmeier, D. and Koenig, W. (2014) 'Strategies for vascular disease prevention: The role of lipids and related markers including apolipoproteins, low-density lipoproteins (LDL)-particle size, high sensitivity C-reactive protein (hs-CRP), lipoprotein-associated phospholipase A (Lp-PLA) and lipoprotein(a) (Lp(a))', *Best Pract Res Clin Endocrinol Metab*, 28(3), pp. 281-294.

Davies, M.N., Volta, M., Pidsley, R., Lunnon, K., Dixit, A., Lovestone, S., Coarfa, C., Harris, R.A., Milosavljevic, A., Troakes, C., Al-Sarraj, S., Dobson, R., Schalkwyk, L.C. and Mill, J. (2012) 'Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood', *Genome Biol*, 13(6), p. R43.

De Jager, P.L., Srivastava, G., Lunnon, K., Burgess, J., Schalkwyk, L.C., Yu, L., Eaton, M.L., Keenan, B.T., Ernst, J., McCabe, C., Tang, A., Raj, T., Replogle, J., Brodeur, W., Gabriel, S., Chai, H.S., Younkin, C., Younkin, S.G., Zou, F., Szyf, M., Epstein, C.B., Schneider, J.A., Bernstein, B.E., Meissner, A., Ertekin-Taner, N., Chibnik, L.B., Kellis, M., Mill, J. and Bennett, D.A. (2014) 'Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci', *Nat Neurosci*, 17(9), pp. 1156-63.

de Koning, I., van Kooten, F., Dippel, D.W., van Harskamp, F., Grobbee, D.E., Kluft, C. and Koudstaal, P.J. (1998) 'The CAMCOG: a useful screening instrument for dementia in stroke patients', *Stroke*, 29(10), pp. 2080-6.

Dedeurwaerder, S., Defrance, M., Bizet, M., Calonne, E., Bontempi, G. and Fuks, F. (2013) 'A comprehensive overview of Infinium HumanMethylation450 data processing', *Brief Bioinform*.

Delcuve, G.P., Rastegar, M. and Davie, J.R. (2009) 'Epigenetic control', *J Cell Physiol*, 219(2), pp. 243-50.

Desplats, P., Spencer, B., Coffee, E., Patel, P., Michael, S., Patrick, C., Adame, A., Rockenstein, E. and Masliah, E. (2011) 'Alpha-synuclein sequesters Dnmt1 from the nucleus: a novel mechanism for epigenetic alterations in Lewy body diseases', *J Biol Chem*, 286(11), pp. 9031-7.

Dietrich, D., Hasinger, O., Liebenberg, V., Field, J.K., Kristiansen, G. and Soltermann, A. (2012) 'DNA methylation of the homeobox genes PITX2 and SHOX2 predicts outcome in non-small-cell lung cancer patients', *Diagn Mol Pathol*, 21(2), pp. 93-104.

Di Francesco, A., Arosio, B., Falconi, A., Micioni Di Bonaventura, M.V., Karimi, M., Mari, D., Casati, M., Maccarrone, M. and D'Addario, C. (2014) 'Global changes in DNA methylation in Alzheimer's disease peripheral blood mononuclear cells', *Brain Behav Immun*.

Dinger, M.E., Amaral, P.P., Mercer, T.R. and Mattick, J.S. (2009) 'Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications', *Brief Funct Genomic Proteomic*, 8(6), pp. 407-23.

Donnan, G.A., Fisher, M., Macleod, M. and Davis, S.M. (2008) 'Stroke', *Lancet*, 371(9624), pp. 1612-23.

Dupont, J.M., Tost, J., Jammes, H. and Gut, I.G. (2004) 'De novo quantitative bisulfite sequencing using the pyrosequencing technology', *Anal Biochem*, 333(1), pp. 119-27.

Duret, L. and Galtier, N. (2000) 'The covariation between TpA deficiency, CpG deficiency, and G+C content of human isochores is due to a mathematical artifact', *Mol Biol Evol*, 17(11), pp. 1620-5.

Ehrich, M., Nelson, M.R., Stanssens, P., Zabeau, M., Liloglou, T., Xinarianos, G., Cantor, C.R., Field, J.K. and van den Boom, D. (2005) 'Quantitative high-throughput analysis of DNA methylation patterns by base-specific cleavage and mass spectrometry', *Proc Natl Acad Sci U S A*, 102(44), pp. 15785-90.

Elliott, H.R., Tillin, T., McArdle, W.L., Ho, K., Duggirala, A., Frayling, T.M., Davey Smith, G., Hughes, A.D., Chaturvedi, N. and Relton, C.L. (2014) 'Differences in smoking associated DNA methylation patterns in South Asians and Europeans', *Clin Epigenetics*, 6(1), p. 4.

Endres, M., Meisel, A., Biniszkiewicz, D., Namura, S., Prass, K., Ruscher, K., Lipski, A., Jaenisch, R., Moskowitz, M.A. and Dirnagl, U. (2000) 'DNA methyltransferase contributes to delayed ischemic brain injury', *J Neurosci*, 20(9), pp. 3175-81.

Enright, B.P., Kubota, C., Yang, X. and Tian, X.C. (2003) 'Epigenetic characteristics and development of embryos cloned from donor cells treated by trichostatin A or 5-aza-2'-deoxycytidine', *Biol Reprod*, 69(3), pp. 896-901.

Fahnestock, M., Michalski, B., Xu, B. and Coughlin, M.D. (2001) 'The precursor pro-nerve growth factor is the predominant form of nerve growth factor in brain and is increased in Alzheimer's disease', *Mol Cell Neurosci*, 18(2), pp. 210-20.

Fahnestock, M., Scott, S.A., Jette, N., Weingartner, J.A. and Crutcher, K.A. (1996) 'Nerve growth factor mRNA and protein levels measured in the same tissue from normal and Alzheimer's disease parietal cortex', *Brain Res Mol Brain Res*, 42(1), pp. 175-8.

Feigin, V.L., Lawes, C.M., Bennett, D.A., Barker-Collo, S.L. and Parag, V. (2009) 'Worldwide stroke incidence and early case fatality reported in 56 population-based studies: a systematic review', *Lancet Neurol*, 8(4), pp. 355-69.

Ferrer, I., Martinez, A., Boluda, S., Parchi, P. and Barrachina, M. (2008) 'Brain banks: benefits, limitations and cautions concerning the use of post-mortem brain tissue for molecular studies', *Cell Tissue Bank*, 9(3), pp. 181-94.

Fineberg, N.A., Haddad, P.M., Carpenter, L., Gannon, B., Sharpe, R., Young, A.H., Joyce, E., Rowe, J., Wellsted, D., Nutt, D.J. and Sahakian, B.J. (2013) 'The size, burden and cost of disorders of the brain in the UK', *J Psychopharmacol*, 27(9), pp. 761-70.

Fjell, A.M., Walhovd, K.B., Fennema-Notestine, C., McEvoy, L.K., Hagler, D.J., Holland, D., Brewer, J.B. and Dale, A.M. (2009) 'One-year brain atrophy evident in healthy aging', *J Neurosci*, 29(48), pp. 15223-31.

Foley, D.L., Craig, J.M., Morley, R., Olsson, C.A., Dwyer, T., Smith, K. and Saffery, R. (2009) 'Prospects for epigenetic epidemiology', *Am J Epidemiol*, 169(4), pp. 389-400.

Folstein, M.F., Folstein, S.E. and McHugh, P.R. (1975) '"Mini-mental state". A practical method for grading the cognitive state of patients for the clinician', *J Psychiatr Res*, 12(3), pp. 189-98.

Font, M.A., Arboix, A. and Krupinski, J. (2010) 'Angiogenesis, neurogenesis and neuroplasticity in ischemic stroke', *Curr Cardiol Rev*, 6(3), pp. 238-44.

Ford, E.S., Li, C. and Sniderman, A. (2013) 'Temporal changes in concentrations of lipids and apolipoprotein B among adults with diagnosed and undiagnosed diabetes, prediabetes, and normoglycemia: findings from the National Health and Nutrition Examination Survey 1988-1991 to 2005-2008', *Cardiovasc Diabetol*, 12, p. 26.

Fraga, M.F., Ballestar, E., Paz, M.F., Ropero, S., Setien, F., Ballestar, M.L., Heine-Suner, D., Cigudosa, J.C., Urioste, M., Benitez, J., Boix-Chornet, M., Sanchez-Aguilera, A., Ling, C., Carlsson, E., Poulsen, P., Vaag, A., Stephan, Z., Spector, T.D., Wu, Y.Z., Plass, C. and Esteller, M. (2005) 'Epigenetic differences arise during the lifetime of monozygotic twins', *Proc Natl Acad Sci U S A*, 102(30), pp. 10604-9.

Furie, K.L., Kasner, S.E., Adams, R.J., Albers, G.W., Bush, R.L., Fagan, S.C., Halperin, J.L., Johnston, S.C., Katzan, I., Kernan, W.N., Mitchell, P.H., Ovbiagele, B., Palesch, Y.Y., Sacco, R.L., Schwamm, L.H., Wassertheil-Smoller, S., Turan, T.N. and Wentworth, D. (2011) 'Guidelines for the prevention of stroke in patients with stroke or transient ischemic attack: a guideline for healthcare professionals from the american heart association/american stroke association', *Stroke*, 42(1), pp. 227-76.

Galloway, S., Takechi, R., Pallebage-Gamarallage, M.M., Dhaliwal, S.S. and Mamo, J.C. (2009) 'Amyloid-beta colocalizes with apolipoprotein B in absorptive cells of the small intestine', *Lipids Health Dis*, 8, p. 46.

Garcia-Berrocoso, T., Penalba, A., Boada, C., Giralt, D., Cuadrado, E., Colome, N., Dayon, L., Canals, F., Sanchez, J.C., Rosell, A. and Montaner, J. (2013) 'From brain to blood: New biomarkers for ischemic stroke prognosis', *J Proteomics*, 94, pp. 138-48.

Gelfman, S., Cohen, N., Yearim, A. and Ast, G. (2013) 'DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon-intron structure', *Genome Res*, 23(5), pp. 789-99.

Gemmell, E., Bosomworth, H., Allan, L., Hall, R., Khundakar, A., Oakley, A.E., Deramecourt, V., Polvikoski, T.M., O'Brien, J.T. and Kalaria, R.N. (2012) 'Hippocampal neuronal atrophy and cognitive function in delayed poststroke and aging-related dementias', *Stroke*, 43(3), pp. 808-14.

Gigante, B., Leander, K., Vikstrom, M., Frumento, P., Carlsson, A.C., Bottai, M. and de Faire, U. (2012) 'Elevated ApoB serum levels strongly predict early cardiovascular events', *Heart*, 98(16), pp. 1242-5.

Gilbert, S.F. (2009) 'Ageing and cancer as diseases of epigenesis', *J Biosci*, 34(4), pp. 601-4.

Gilks, W.P., Abou-Sleiman, P.M., Gandhi, S., Jain, S., Singleton, A., Lees, A.J., Shaw, K., Bhatia, K.P., Bonifati, V., Quinn, N.P., Lynch, J., Healy, D.G., Holton, J.L., Revesz, T. and Wood, N.W. (2005) 'A common LRRK2 mutation in idiopathic Parkinson's disease', *Lancet*, 365(9457), pp. 415-6.

Gillies, G.E., Pienaar, I.S., Vohra, S. and Qamhawi, Z. (2014) 'Sex differences in Parkinson's disease', *Front Neuroendocrinol*.

Goetz, C.G., Tilley, B.C., Shaftman, S.R., Stebbins, G.T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M.B., Dodel, R., Dubois, B., Holloway, R., Jankovic, J., Kulisevsky, J., Lang, A.E., Lees, A., Leurgans, S., LeWitt, P.A., Nyenhuis, D., Olanow, C.W., Rascol, O., Schrag, A., Teresi, J.A., van Hilten, J.J. and LaPelle, N. (2008) 'Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results', *Mov Disord*, 23(15), pp. 2129-70.

Goodrich, J.A. and Kugel, J.F. (2006) 'Non-coding-RNA regulators of RNA polymerase II transcription', *Nat Rev Mol Cell Biol,* 7(8), pp. 612-6.

Gorelick, P.B. (1997) 'Status of risk factors for dementia associated with stroke', *Stroke*, 28(2), pp. 459-63.

Graff, J. and Mansuy, I.M. (2009) 'Epigenetic dysregulation in cognitive disorders', *Eur J Neurosci*, 30(1), pp. 1-8.

Grant, P.A. (2001) 'A tale of histone modifications', *Genome Biol*, 2(4), p. REVIEWS0003.

Gravina, S. and Vijg, J. (2010) 'Epigenetic factors in aging and longevity', *Pflugers Arch*, 459(2), pp. 247-58.

Greenberg, D.A. and Jin, K. (2006) 'Growth factors and stroke', *NeuroRx*, 3(4), pp. 458-65.

Gretarsdottir, S., Thorleifsson, G., Manolescu, A., Styrkarsdottir, U., Helgadottir, A., Gschwendtner, A., Kostulas, K., Kuhlenbaumer, G., Bevan, S., Jonsdottir, T., Bjarnason, H., Saemundsdottir, J., Palsson, S., Arnar, D.O., Holm, H., Thorgeirsson, G., Valdimarsson, E.M., Sveinbjornsdottir, S., Gieger, C., Berger, K., Wichmann, H.E., Hillert, J., Markus, H., Gulcher, J.R., , *et al.* (2008) 'Risk variants for atrial fibrillation on chromosome 4q25 associate with ischemic stroke', *Ann Neurol*, 64(4), pp. 402-9.

Grundberg, E., Meduri, E., Sandling, J.K., Hedman, A.K., Keildson, S., Buil, A., Busche, S., Yuan, W., Nisbet, J., Sekowska, M., Wilk, A., Barrett, A., Small, K.S., Ge, B., Caron, M., Shin, S.Y., Lathrop, M., Dermitzakis, E.T., McCarthy, M.I., Spector, T.D., Bell, J.T. and Deloukas, P. (2013) 'Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements', *Am J Hum Genet*, 93(5), pp. 876-90.

Gschwendtner, A., Bevan, S., Cole, J.W., Plourde, A., Matarin, M., Ross-Adams, H., Meitinger, T., Wichmann, E., Mitchell, B.D., Furie, K., Slowik, A., Rich, S.S., Syme, P.D., MacLeod, M.J., Meschia, J.F., Rosand, J., Kittner, S.J., Markus, H.S., Muller-Myhsok, B. and Dichgans, M. (2009) 'Sequence variants on chromosome 9p21.3 confer risk for atherosclerotic stroke', *Ann Neurol*, 65(5), pp. 531-9.

Gudbjartsson, D.F., Arnar, D.O., Helgadottir, A., Gretarsdottir, S., Holm, H., Sigurdsson, A., Jonasdottir, A., Baker, A., Thorleifsson, G., Kristjansson, K., Palsson, A., Blondal, T., Sulem, P., Backman, V.M., Hardarson, G.A., Palsdottir, E., Helgason, A., Sigurjonsdottir, R., Sverrisson, J.T., Kostulas, K., Ng, M.C., Baum, L., So, W.Y., Wong, K.S., Chan, J.C., Furie, K.L., Greenberg, S.M., *et al.* (2007) 'Variants conferring risk of atrial fibrillation on chromosome 4q25', *Nature*, 448(7151), pp. 353-7.

Guintivano, J., Aryee, M.J. and Kaminsky, Z.A. (2013) 'A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression', *Epigenetics*, 8(3), pp. 290-302.

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J.B., Gao, Y., Deconde, R., Chen, M., Rajapakse, I., Friend, S., Ideker, T. and Zhang, K. (2013) 'Genome-wide methylation profiles reveal quantitative views of human aging rates', *Mol Cell*, 49(2), pp. 359-67.

Heart Protection Study Collaborative Group. (2002) 'MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20 536 high-risk individuals: a randomised placebo-controlled trial', *Lancet*, 360:7-22.

Heijmans, B.T. and Mill, J. (2012) 'Commentary: The seven plagues of epigenetic epidemiology', *Int J Epidemiol*, 41(1), pp. 74-8.

Heijmans, B.T., Tobi, E.W., Stein, A.D., Putter, H., Blauw, G.J., Susser, E.S., Slagboom, P.E. and Lumey, L.H. (2008) 'Persistent epigenetic differences associated with prenatal exposure to famine in humans', *Proc Natl Acad Sci U S A*, 105(44), pp. 17046-9.

Hellweg, R., Gericke, C.A., Jendroska, K., Hartung, H.D. and Cervos-Navarro, J. (1998) 'NGF content in the cerebral cortex of non-demented patients with amyloid-plaques and in symptomatic Alzheimer's disease', *Int J Dev Neurosci*, 16(7-8), pp. 787-94.

Hely, M.A., Reid, W.G., Adena, M.A., Halliday, G.M. and Morris, J.G. (2008) 'The Sydney multicenter study of Parkinson's disease: the inevitability of dementia at 20 years', *Mov Disord*, 23(6), pp. 837-44.

Herman, J.G., Graff, J.R., Myohanen, S., Nelkin, B.D. and Baylin, S.B. (1996) 'Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands', *Proc Natl Acad Sci U S A*, 93(18), pp. 9821-6.

Hernandez, D.G., Nalls, M.A., Gibbs, J.R., Arepalli, S., van der Brug, M., Chong, S., Moore, M., Longo, D.L., Cookson, M.R., Traynor, B.J. and Singleton, A.B. (2011) 'Distinct DNA methylation changes highly correlated with chronological age in the human brain', *Hum Mol Genet*, 20(6), pp. 1164-72.

Heyn, H., Carmona, F.J., Gomez, A., Ferreira, H.J., Bell, J.T., Sayols, S., Ward, K., Stefansson, O.A., Moran, S., Sandoval, J., Eyfjord, J.E., Spector, T.D. and Esteller, M. (2013) 'DNA methylation profiling in breast cancer discordant identical twins identifies DOK7 as novel epigenetic biomarker', *Carcinogenesis*, 34(1), pp. 102-8.

Higgins, G.A. and Mufson, E.J. (1989) 'NGF receptor gene expression is decreased in the nucleus basalis in Alzheimer's disease', *Exp Neurol*, 106(3), pp. 222-36.

Hindorff, L.A., MacArthur , J., Morales, J., Junkins, H.A., Hall, P.A., Klemm, A.K., Manolio, T.A. A catalog of published genome-wide association studies. Available at: www.genome.gov/gwastudies. Accessed: 02/09/14.

Hitchins, M.P., Rapkins, R.W., Kwok, C.T., Srivastava, S., Wong, J.J., Khachigian, L.M., Polly, P., Goldblatt, J. and Ward, R.L. (2011) 'Dominantly inherited constitutional epigenetic silencing of MLH1 in a cancer-affected family is linked to a single nucleotide variant within the 5'UTR', *Cancer Cell*, 20(2), pp. 200-13.

Hock, C., Heese, K., Muller-Spahn, F., Huber, P., Riesen, W., Nitsch, R.M. and Otten, U. (2000) 'Increased CSF levels of nerve growth factor in patients with Alzheimer's disease', *Neurology*, 54(10), pp. 2009-11.

Hoehn, M.M. and Yahr, M.D. (1967) 'Parkinsonism: onset, progression and mortality', *Neurology*, 17(5), pp. 427-42.

Hokama, M., Oka, S., Leon, J., Ninomiya, T., Honda, H., Sasaki, K., Iwaki, T., Ohara, T., Sasaki, T., Laferla, F.M., Kiyohara, Y. and Nakabeppu, Y. (2013) 'Altered Expression of Diabetes-Related Genes in Alzheimer's Disease Brains: The Hisayama Study', *Cereb Cortex*.

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

Houseman, E.A., Accomando, W.P., Koestler, D.C., Christensen, B.C., Marsit, C.J., Nelson, H.H., Wiencke, J.K. and Kelsey, K.T. (2012) 'DNA methylation arrays as surrogate measures of cell mixture distribution', *BMC Bioinformatics*, 13, p. 86.

Huang, Y., Pastor, W.A., Shen, Y., Tahiliani, M., Liu, D.R. and Rao, A. (2010) 'The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing', *PLoS One*, 5(1), p. e8888.

Hughes, A.J., Daniel, S.E., Kilford, L. and Lees, A.J. (1992) 'Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases', *J Neurol Neurosurg Psychiatry*, 55(3), pp. 181-4.

Huh, I., Zeng, J., Park, T. and Yi, S.V. (2013) 'DNA methylation and transcriptional noise', *Epigenetics Chromatin*, 6(1), p. 9.
Hwang, J.Y., Aromolaran, K.A. and Zukin, R.S. (2013) 'Epigenetic mechanisms in stroke and epilepsy', *Neuropsychopharmacology*, 38(1), pp. 167-82.

Illumina. (2010) 'GenomeStudio® Methylation Module v1.8 User Guide'

Inzitari, D., Di Carlo, A., Pracucci, G., Lamassa, M., Vanni, P., Romanelli, M., Spolveri, S., Adriani, P., Meucci, I., Landini, G. and Ghetti, A. (1998) 'Incidence and determinants of poststroke dementia as defined by an informant interview method in a hospital-based stroke registry', *Stroke*, 29(10), pp. 2087-93.

Iraola-Guzman, S., Estivill, X. and Rabionet, R. (2011) 'DNA methylation in neurodegenerative disorders: a missing link between genome and environment?', *Clin Genet*, 80(1), pp. 1-14.

Irizarry, R.A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., Ji, H., Potash, J.B., Sabunciyan, S. and Feinberg, A.P. (2009) 'The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores', *Nat Genet*, 41(2), pp. 178-86.

Iwamoto, K., Bundo, M., Ueda, J., Oldham, M.C., Ukai, W., Hashimoto, E., Saito, T., Geschwind, D.H. and Kato, T. (2011) 'Neurons show distinctive DNA methylation profile and higher interindividual variations compared with non-neurons', *Genome Res*, 21(5), pp. 688-96.

Iwasaki, Y.K., Nishida, K., Kato, T. and Nattel, S. (2011) 'Atrial fibrillation pathophysiology: implications for management', *Circulation*, 124(20), pp. 2264-74.
Jaenisch, R. and Bird, A. (2003) 'Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals', *Nat Genet*, 33 Suppl, pp. 245-54.

Jaffe, A.E., Murakami, P., Lee, H., Leek, J.T., Fallin, M.D., Feinberg, A.P. and Irizarry, R.A. (2012) 'Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies', *Int J Epidemiol*, 41(1), pp. 200-9.

Janvin, C.C., Larsen, J.P., Aarsland, D. and Hugdahl, K. (2006) 'Subtypes of mild cognitive impairment in Parkinson's disease: progression to dementia', *Mov Disord*, 21(9), pp. 1343-9.

Jirtle, R.L. and Skinner, M.K. (2007) 'Environmental epigenomics and disease susceptibility', *Nat Rev Genet*, 8(4), pp. 253-62.

Jjingo, D., Conley, A.B., Yi, S.V., Lunyak, V.V. and Jordan, I.K. (2012) 'On the presence and role of human gene-body DNA methylation', *Oncotarget*, 3(4), pp. 462-74.

Jones, P.A. (2012) 'Functions of DNA methylation: islands, start sites, gene bodies and beyond', *Nat Rev Genet*, 13(7), pp. 484-92.

Jowaed, A., Schmitt, I., Kaut, O. and Wullner, U. (2010) 'Methylation regulates alpha-synuclein expression and is decreased in Parkinson's disease patients' brains', *J Neurosci*, 30(18), pp. 6355-9.

Kaelin, W.G., Jr. and McKnight, S.L. (2013) 'Influence of metabolism on epigenetics and disease', *Cell*, 153(1), pp. 56-69.

Kahn, A. and Fraga, M.F. (2009) 'Epigenetics and aging: status, challenges, and needs for the future', *J Gerontol A Biol Sci Med Sci*, 64(2), pp. 195-8.

Kalaria, R.N. and Ballard, C. (2001) 'Stroke and cognition', *Curr Atheroscler Rep*, 3(4), pp. 334-9.

Kalaria, R.N., Viitanen, M., Kalimo, H., Dichgans, M. and Tabira, T. (2004) 'The pathogenesis of CADASIL: an update', *J Neurol Sci*, 226(1-2), pp. 35-9.

Kandiah, N., Mak, E., Ng, A., Huang, S., Au, W.L., Sitoh, Y.Y. and Tan, L.C. (2013) 'Cerebral white matter hyperintensity in Parkinson's disease: a major risk factor for mild cognitive impairment', *Parkinsonism Relat Disord*, 19(7), pp. 680-3.

Kasinski, A.L. and Slack, F.J. (2011) 'Epigenetics and genetics. MicroRNAs en route to the clinic: progress in validating and targeting microRNAs for cancer therapy', *Nat Rev Cancer*, 11(12), pp. 849-64.

Khoo, T.K., Yarnall, A.J., Duncan, G.W., Coleman, S., O'Brien, J.T., Brooks, D.J., Barker, R.A. and Burn, D.J. (2013) 'The spectrum of nonmotor symptoms in early Parkinson disease', *Neurology*, 80(3), pp. 276-81.

Kim, D.H., Sarbassov, D.D., Ali, S.M., King, J.E., Latek, R.R., Erdjument-Bromage, H., Tempst, P. and Sabatini, D.M. (2002) 'mTOR interacts with raptor to form a nutrient-sensitive complex that signals to the cell growth machinery', *Cell*, 110(2), pp. 163-75.

Kim, H.J., Oh, E.S., Lee, J.H., Moon, J.S., Oh, J.E., Shin, J.W., Lee, K.J., Baek, I.C., Jeong, S.H., Song, H.J., Sohn, E.H. and Lee, A.Y. (2012a) 'Relationship between changes of body mass index (BMI) and cognitive decline in Parkinson's disease (PD)', *Arch Gerontol Geriatr*, 55(1), pp. 70-2.

Kim, J.K., Samaranayake, M. and Pradhan, S. (2009) 'Epigenetic mechanisms in mammals', *Cell Mol Life Sci*, 66(4), pp. 596-612.

Kim, J.W., Kim, S.T., Turner, A.R., Young, T., Smith, S., Liu, W., Lindberg, J., Egevad, L., Gronberg, H., Isaacs, W.B. and Xu, J. (2012b) 'Identification of new differentially methylated genes that have potential functional consequences in prostate cancer', *PLoS One*, 7(10), p. e48455.

Kim, M., Long, T.I., Arakawa, K., Wang, R., Yu, M.C. and Laird, P.W. (2010) 'DNA methylation as a biomarker for cardiovascular disease risk', *PLoS One*, 5(3), p. e9692.

Kim, S.J., Moon, G.J. and Bang, O.Y. (2013) 'Biomarkers for Stroke', *J Stroke*, 15(1), pp. 27-37.

King, J.A., Trinkler, I., Hartley, T., Vargha-Khadem, F. and Burgess, N. (2004) 'The hippocampal role in spatial memory and the familiarity--recollection distinction: a case study', *Neuropsychology*, 18(3), pp. 405-17.

Kirvell, S.L., Elliott, M.S., Kalaria, R.N., Hortobagyi, T., Ballard, C.G., Francis, P.T. (2010) 'Vesicular glutamate transporter and cognition in stroke: a case–control autopsy study'. *Neurology,* 75, pp. 1803–1809.

Koivistoinen, T., Hutri-Kahonen, N., Juonala, M., Koobi, T., Aatola, H., Lehtimaki, T., Viikari, J.S., Raitakari, O.T. and Kahonen, M. (2011) 'Apolipoprotein B is related to arterial pulse wave velocity in young adults: the Cardiovascular Risk in Young Finns Study', *Atherosclerosis*, 214(1), pp. 220-4.

Konukoglu, D., Andican, G., Firtina, S., Erkol, G. and Kurt, A. (2012) 'Serum brain-derived neurotrophic factor, nerve growth factor and neurotrophin-3 levels in dementia', *Acta Neurol Belg*, 112(3), pp. 255-60.

Kovalchuk, A., Lowings, M., Rodriguez-Juarez, R., Muhammad, A., Ilnytskyy, S., Kolb, B. and Kovalchuk, O. (2012) 'Epigenetic bystander-like effects of stroke in somatic organs', *Aging (Albany NY)*, 4(3), pp. 224-34.

Krupinski, J., Kaluza, J., Kumar, P., Kumar, S. and Wang, J.M. (1994) 'Role of angiogenesis in patients with cerebral ischemic stroke', *Stroke*, 25(9), pp. 1794-8.

Kuo, Y.M., Emmerling, M.R., Bisgaier, C.L., Essenburg, A.D., Lampert, H.C., Drumm, D. and Roher, A.E. (1998) 'Elevated low-density lipoprotein in Alzheimer's disease correlates with brain abeta 1-42 levels', *Biochem Biophys Res Commun*, 252(3), pp. 711-5.

Lam, V., Takechi, R., Pallebage-Gamarallage, M.M., Galloway, S. and Mamo, J.C. (2011) 'Colocalisation of plasma derived apo B lipoproteins with cerebral proteoglycans in a transgenic-amyloid model of Alzheimer's disease', *Neurosci Lett*, 492(3), pp. 160-4.

Lee, M., Hong, K.S., Chang, S.C. and Saver, J.L. (2010) 'Efficacy of homocysteine-lowering therapy with folic Acid in stroke prevention: a meta-analysis', *Stroke*, 41(6), pp. 1205-12.

Lee, S.E., Shen, H., Taglialatela, G., Chung, J.M. and Chung, K. (1998) 'Expression of nerve growth factor in the dorsal root ganglion after peripheral nerve injury', *Brain Res*, 796(1-2), pp. 99-106.

Leeds, L., Meara, R.J., Woods, R. and Hobson, J.P. (2001) 'A comparison of the new executive functioning domains of the CAMCOG-R with existing tests of executive function in elderly stroke survivors', *Age Ageing*, 30(3), pp. 251-4.

Leszek, J., Sochocka, M. and Gasiorowski, K. (2012) 'Vascular factors and epigenetic modifications in the pathogenesis of Alzheimer's disease', *J Neurol Sci*, 323(1-2), pp. 25-32.

Letunic, I., Doerks, T., Bork, P. (2011) 'SMART 7: recent updates to the protein domain annotation resource', *Nucleic Acids Res*, 40: D302-5. Available at: http://smart.embl-heidelberg.de/. Accessed: 02/09/14.

Leys, D., Henon, H., Mackowiak-Cordoliani, M.A. and Pasquier, F. (2005) 'Poststroke dementia', *Lancet Neurol*, 4(11), pp. 752-9.

Li, Y.Y., Chen, T., Wan, Y. and Xu, S.Q. (2012) 'Lead exposure in pheochromocytoma cells induces persistent changes in amyloid precursor protein gene methylation patterns', *Environ Toxicol*, 27(8), pp. 495-502.

Lisanti, S., Omar, W.A., Tomaszewski, B., De Prins, S., Jacobs, G., Koppen, G., Mathers, J.C. and Langie, S.A. (2013) 'Comparison of methods for quantification of global DNA methylation in human cells and tissues', *PLoS One*, 8(11), p. e79044.

Lister, R., Mukamel, E.A., Nery, J.R., Urich, M., Puddifoot, C.A., Johnson, N.D., Lucero, J., Huang, Y., Dwork, A.J., Schultz, M.D., Yu, M., Tonti-Filippini, J., Heyn, H., Hu, S., Wu, J.C., Rao, A., Esteller, M., He, C., Haghighi, F.G., Sejnowski, T.J., Behrens, M.M. and Ecker, J.R. (2013) 'Global epigenomic reconfiguration during mammalian brain development', *Science*, 341(6146), p. 1237905.

Litvan, I., Goldman, J.G., Troster, A.I., Schmand, B.A., Weintraub, D., Petersen, R.C., Mollenhauer, B., Adler, C.H., Marder, K., Williams-Gray, C.H., Aarsland, D., Kulisevsky, J., Rodriguez-Oroz, M.C., Burn, D.J., Barker, R.A. and Emre, M. (2012) 'Diagnostic criteria for mild cognitive impairment in Parkinson's disease: Movement Disorder Society Task Force guidelines', *Mov Disord*, 27(3), pp. 349-56.

Lorenzen, J.M., Martino, F. and Thum, T. (2012) 'Epigenetic modifications in cardiovascular disease', *Basic Res Cardiol*, 107(2), p. 245.

Lovell, M.A. and Markesbery, W.R. (2007) 'Oxidative DNA damage in mild cognitive impairment and late-stage Alzheimer's disease', *Nucleic Acids Res*, 35(22), pp. 7497-504.

Lu, H., Liu, X., Deng, Y. and Qing, H. (2013) 'DNA methylation, a hand behind neurodegenerative diseases', *Front Aging Neurosci*, 5, p. 85.

Lubin, F.D., Gupta, S., Parrish, R.R., Grissom, N.M. and Davis, R.L. (2011) 'Epigenetic mechanisms: critical contributors to long-term memory formation', *Neuroscientist*, 17(6), pp. 616-32.

Lunnon, K., Smith, R., Hannon, E., De Jager, P.L., Srivastava, G., Volta, M., Troakes, C., Al-Sarraj, S., Burrage, J., Macdonald, R., Condliffe, D., Harries, L.W., Katsel, P., Haroutunian, V., Kaminsky, Z., Joachim, C., Powell, J., Lovestone, S., Bennett, D.A., Schalkwyk, L.C. and Mill, J. (2014) 'Methylomic profiling implicates cortical deregulation of ANK1 in Alzheimer's disease', *Nat Neurosci*, 17(9), pp. 1164-70.

Maes, O.C., Schipper, H.M., Chertkow, H.M. and Wang, E. (2009) 'Methodology for discovery of Alzheimer's disease blood-based biomarkers', *J Gerontol A Biol Sci Med Sci*, 64(6), pp. 636-45.

Markus, H.S. (2011) 'Stroke genetics', *Hum Mol Genet*, 20(R2), pp. R124-31.

Marques, S.C., Oliveira, C.R., Pereira, C.M. and Outeiro, T.F. (2011) 'Epigenetics in neurodegeneration: a new layer of complexity', *Prog Neuropsychopharmacol Biol Psychiatry*, 35(2), pp. 348-55.

Marsit, C.J., Koestler, D.C., Christensen, B.C., Karagas, M.R., Houseman, E.A. and Kelsey, K.T. (2011) 'DNA methylation array analysis identifies profiles of blood-derived DNA methylation associated with bladder cancer', *J Clin Oncol*, 29(9), pp. 1133-9.

Mashayekhi, F. and Salehin, Z. (2006) 'Cerebrospinal fluid nerve growth factor levels in patients with Alzheimer's disease', *Ann Saudi Med*, 26(4), pp. 278-82.

Masliah, E., Dumaop, W., Galasko, D. and Desplats, P. (2013) 'Distinctive patterns of DNA methylation associated with Parkinson disease: Identification of concordant epigenetic changes in brain and peripheral blood leukocytes', *Epigenetics*, 8(10).

Mastroeni, D., Grover, A., Delvaux, E., Whiteside, C., Coleman, P.D. and Rogers, J. (2011) 'Epigenetic mechanisms in Alzheimer's disease', *Neurobiol Aging*, 32(7), pp. 1161-80.

Mastroeni, D., McKee, A., Grover, A., Rogers, J. and Coleman, P.D. (2009) 'Epigenetic differences in cortical neurons from a pair of monozygotic twins discordant for Alzheimer's disease', *PLoS One*, 4(8), p. e6617.

Mathers, J.C., Strathdee, G. and Relton, C.L. (2010) 'Induction of epigenetic alterations by dietary and other environmental factors', *Adv Genet*, 71, pp. 3-39.

McKay, J.A., Xie, L., Harris, S., Wong, Y.K., Ford, D. and Mathers, J.C. (2011) 'Blood as a surrogate marker for tissue-specific DNA methylation and changes due to folate depletion in post-partum female mice', *Mol Nutr Food Res*, 55(7), pp. 1026-35.

McKeith, I.G. (2002) 'Dementia with Lewy bodies', *Br J Psychiatry*, 180, pp. 144-7.

Mercer, T.R., Dinger, M.E. and Mattick, J.S. (2009) 'Long non-coding RNAs: insights into functions', *Nat Rev Genet*, 10(3), pp. 155-9.

Michels, K.B. (2012) 'Epigenetic Epidemiology', Springer.

Michels, K.B., Binder, A.M., Dedeurwaerder, S., Epstein, C.B., Greally, J.M., Gut, I., Houseman, E.A., Izzi, B., Kelsey, K.T., Meissner, A., Milosavljevic, A., Siegmund, K.D., Bock, C. and Irizarry, R.A. (2013) 'Recommendations for the design and analysis of epigenome-wide association studies', *Nat Methods*, 10(10), pp. 949-55.

Mikeska, T., Felsberg, J., Hewitt, C.A. and Dobrovic, A. (2011) 'Analysing DNA methylation using bisulphite pyrosequencing', *Methods Mol Biol*, 791, pp. 33-53.

Mill, J. and Heijmans, B.T. (2013) 'From promises to practical strategies in epigenetic epidemiology', *Nat Rev Genet*, 14(8), pp. 585-94.

Miller, C.A. and Sweatt, J.D. (2007) 'Covalent modification of DNA regulates memory formation', *Neuron*, 53(6), pp. 857-69.

Morgan, D.G., May, P.C. and Finch, C.E. (1987) 'Dopamine and serotonin systems in human and rodent brain: effects of age and neurodegenerative disease', *J Am Geriatr Soc*, 35(4), pp. 334-45.

Moroney, J.T., Bagiella, E., Tatemichi, T.K., Paik, M.C., Stern, Y. and Desmond, D.W. (1997) 'Dementia after stroke increases the risk of long-term stroke recurrence', *Neurology*, 48(5), pp. 1317-25.

Moskowitz, M.A., Lo, E.H. and Iadecola, C. (2010) 'The science of stroke: mechanisms in search of treatments', *Neuron*, 67(2), pp. 181-98.

MRC Centre for Causal Analyses in Translational Epidemiology. 'GEoCoDE'. Available at: http://www.bristol.ac.uk/caite/geocode/. Accessed: 02/09/14.

Mufson, E.J., Binder, L., Counts, S.E., DeKosky, S.T., de Toledo-Morrell, L., Ginsberg, S.D., Ikonomovic, M.D., Perez, S.E. and Scheff, S.W. (2012a) 'Mild cognitive impairment: pathology and mechanisms', *Acta Neuropathol*, 123(1), pp. 13-30.

Mufson, E.J., He, B., Nadeem, M., Perez, S.E., Counts, S.E., Leurgans, S., Fritz, J., Lah, J., Ginsberg, S.D., Wuu, J. and Scheff, S.W. (2012b) 'Hippocampal proNGF signaling pathways and beta-amyloid levels in mild cognitive impairment and Alzheimer disease', *J Neuropathol Exp Neurol*, 71(11), pp. 1018-29.

Mufson, E.J., Kroin, J.S., Sendera, T.J. and Sobreviela, T. (1999) 'Distribution and retrograde transport of trophic factors in the central nervous system: functional implications for the treatment of neurodegenerative diseases', *Prog Neurobiol*, 57(4), pp. 451-84.

Mukherjee, D. and Patil, C.G. (2011) 'Epidemiology and the global burden of stroke', *World Neurosurg*, 76(6 Suppl), pp. S85-90.

Naeem, H., Wong, N.C., Chatterton, Z., Hong, M.K., Pedersen, J.S., Corcoran, N.M., Hovens, C.M. and Macintyre, G. (2014) 'Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array', *BMC Genomics*, 15, p. 51.

Nasreddine, Z.S., Phillips, N.A., Bedirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J.L. and Chertkow, H. (2005) 'The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment', *J Am Geriatr Soc*, 53(4), pp. 695-9.

Nelson, E.D. and Monteggia, L.M. (2011) 'Epigenetics in the mature mammalian brain: effects on behavior and synaptic transmission', *Neurobiol Learn Mem*, 96(1), pp. 53-60.

Ng, J.W., Barrett, L.M., Wong, A., Kuh, D., Smith, G.D. and Relton, C.L. (2012) 'The role of longitudinal cohort studies in epigenetic epidemiology: challenges and opportunities', *Genome Biol*, 13(6), p. 246.

Nguyen, S., Meletis, K., Fu, D., Jhaveri, S. and Jaenisch, R. (2007) 'Ablation of de novo DNA methyltransferase Dnmt3a in the nervous system leads to neuromuscular defects and shortened lifespan', *Dev Dyn*, 236(6), pp. 1663-76.

Numata, S., Ye, T., Hyde, T.M., Guitart-Navarro, X., Tao, R., Wininger, M., Colantuoni, C., Weinberger, D.R., Kleinman, J.E. and Lipska, B.K. (2012) 'DNA methylation signatures in development and aging of the human prefrontal cortex', *Am J Hum Genet*, 90(2), pp. 260-72.

Obeid, R., Schadt, A., Dillmann, U., Kostopoulos, P., Fassbender, K. and Herrmann, W. (2009) 'Methylation status and neurodegenerative markers in Parkinson disease', *Clin Chem*, 55(10), pp. 1852-60.

Oberdoerffer, P. and Sinclair, D.A. (2007) 'The role of nuclear architecture in genomic instability and ageing', *Nat Rev Mol Cell Biol*, 8(9), pp. 692-702.

Oh, G., Wang, S.C., Pal, M., Chen, Z.F., Khare, T., Tochigi, M., Ng, C., Yang, Y.A., Kwan, A., Kaminsky, Z.A., Mill, J., Gunasinghe, C., Tackett, J.L., Gottesman, II, Willemsen, G., de Geus, E.J., Vink, J.M., Slagboom, P.E., Wray, N.R., Heath, A.C., Montgomery, G.W., Turecki, G., Martin, N.G., Boomsma, D.I., McGuffin, P., Kustra, R. and Petronis, A. (2014) 'DNA Modification Study of Major Depressive Disorder: Beyond Locus-by-Locus Comparisons', *Biol Psychiatry*.

Ott, A., Breteler, M.M., de Bruyne, M.C., van Harskamp, F., Grobbee, D.E. and Hofman, A. (1997) 'Atrial fibrillation and dementia in a population-based study. The Rotterdam Study', *Stroke*, 28(2), pp. 316-21.

Ott, A., Slooter, A.J., Hofman, A., van Harskamp, F., Witteman, J.C., Van Broeckhoven, C., van Duijn, C.M. and Breteler, M.M. (1998) 'Smoking and risk of dementia and Alzheimer's disease in a population-based cohort study: the Rotterdam Study', *Lancet*, 351(9119), pp. 1840-3.

Paciaroni, M. and Bogousslavsky, J. (2013) 'Connecting cardiovascular disease and dementia: further evidence', *J Am Heart Assoc*, 2(6), p. e000656.

Pagonabarraga, J. and Kulisevsky, J. (2012) 'Cognitive impairment and dementia in Parkinson's disease', *Neurobiol Dis*, 46(3), pp. 590-6.

Palavra, N.C., Naismith, S.L. and Lewis, S.J. (2013) 'Mild cognitive impairment in Parkinson's disease: a review of current concepts', *Neurol Res Int*, 2013, p. 576091.

Palii, S.S., Van Emburgh, B.O., Sankpal, U.T., Brown, K.D. and Robertson, K.D. (2008) 'DNA methylation inhibitor 5-Aza-2'-deoxycytidine induces reversible genome-wide DNA damage that is distinctly influenced by DNA methyltransferases 1 and 3B', *Mol Cell Biol*, 28(2), pp. 752-71.

Panning, B. and Jaenisch, R. (1996) 'DNA hypomethylation can activate Xist expression and silence X-linked genes', *Genes Dev*, 10(16), pp. 1991-2002.

Patel, V.B., Robbins, M.A. and Topol, E.J. (2001) 'C-reactive protein: a 'golden marker' for inflammation and coronary artery disease', *Cleve Clin J Med*, 68(6), pp. 521-524, 527-34.

Pearce, M.S., McConnell, J.C., Potter, C., Barrett, L.M., Parker, L., Mathers, J.C. and Relton, C.L. (2012) 'Global LINE-1 DNA methylation is associated with blood glycaemic and lipid profiles', *Int J Epidemiol*, 41(1), pp. 210-7.

Pendlebury, S.T. and Rothwell, P.M. (2009) 'Prevalence, incidence, and factors associated with pre-stroke and post-stroke dementia: a systematic review and meta-analysis', *Lancet Neurol*, 8(11), pp. 1006-18.

Perry, R., Oakley, A. (1993) 'Coronal map of Brodmann areas in human brain'. In: Roberts GLP, editor. ed. Neuropsychiatric Disorders. London: Wolfe; 1-10.

Peters, R. (2006) 'Ageing and the brain', *Postgrad Med J*, 82(964), pp. 84-8.

Petersen, R.C. (2004) 'Mild cognitive impairment as a diagnostic entity', *J Intern Med*, 256(3), pp. 183-94.

Pidsley, R. and Mill, J. (2011) 'Epigenetic studies of psychosis: current findings, methodological approaches, and implications for postmortem research', *Biol Psychiatry*, 69(2), pp. 146-56.

Pohjasvaara, T., Erkinjuntti, T., Ylikoski, R., Hietanen, M., Vataja, R. and Kaste, M. (1998) 'Clinical determinants of poststroke dementia', *Stroke*, 29(1), pp. 75-81.
Portela, A. and Esteller, M. (2010) 'Epigenetic modifications and human disease', *Nat Biotechnol*, 28(10), pp. 1057-68.

Post, W.S., Goldschmidt-Clermont, P.J., Wilhide, C.C., Heldman, A.W., Sussman, M.S., Ouyang, P., Milliken, E.E. and Issa, J.P. (1999) 'Methylation of the estrogen receptor gene is associated with aging and atherosclerosis in the cardiovascular system', *Cardiovasc Res*, 43(4), pp. 985-91.

Qureshi, I.A. and Mehler, M.F. (2010) 'Emerging role of epigenetics in stroke: part 1: DNA methylation and chromatin modifications', *Arch Neurol*, 67(11), pp. 1316-22.

Rakyan, V.K., Down, T.A., Balding, D.J. and Beck, S. (2011) 'Epigenome-wide association studies for common human diseases', *Nat Rev Genet*, 12(8), pp. 529-41.

Rakyan, V.K., Preis, J., Morgan, H.D. and Whitelaw, E. (2001) 'The marks, mechanisms and memory of epigenetic states in mammals', *Biochem J*, 356(Pt 1), pp. 1-10.

Rechache, N.S., Wang, Y., Stevenson, H.S., Killian, J.K., Edelman, D.C., Merino, M., Zhang, L., Nilubol, N., Stratakis, C.A., Meltzer, P.S. and Kebebew, E. (2012) 'DNA methylation profiling identifies global methylation differences and markers of adrenocortical tumors', *J Clin Endocrinol Metab*, 97(6), pp. E1004-13.

Redon, J., Olsen, M.H., Cooper, R.S., Zurriaga, O., Martinez-Beneito, M.A., Laurent, S., Cifkova, R., Coca, A. and Mancia, G. (2011) 'Stroke mortality and trends from 1990 to 2006 in 39 countries from Europe and Central Asia: implications for control of high blood pressure', *Eur Heart J*, 32(11), pp. 1424-31.

Reed, K., Poulin, M.L., Yan, L. and Parissenti, A.M. (2010) 'Comparison of bisulfite sequencing PCR with pyrosequencing for measuring differences in DNA methylation', *Anal Biochem*, 397(1), pp. 96-106.

Reik, W. (2007) 'Stability and flexibility of epigenetic gene regulation in mammalian development', *Nature*, 447(7143), pp. 425-32.

Reik, W., Dean, W. and Walter, J. (2001) 'Epigenetic reprogramming in mammalian development', *Science*, 293(5532), pp. 1089-93.

Relton, C.L. and Davey Smith, G. (2010) 'Epigenetic epidemiology of common complex disease: prospects for prediction, prevention, and treatment', *PLoS Med*, 7(10), p. e1000356.

Relton, C.L. and 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(1), pp. 161-76.

Richter, J., Appenzeller, S., Ammerpohl, O., Deuschl, G., Paschen, S., Bruggemann, N., Klein, C. and Kuhlenbaumer, G. (2012) 'No evidence for differential methylation of alpha-synuclein in leukocyte DNA of Parkinson's disease patients', *Mov Disord*, 27(4), pp. 590-1.

Riedel, O., Dodel, R., Deuschl, G., Klotsche, J., Forstl, H., Heuser, I., Oertel, W., Reichmann, H., Riederer, P., Trenkwalder, C. and Wittchen, H.U. (2012) 'Depression and care-dependency in Parkinson's disease: results from a nationwide study of 1449 outpatients', *Parkinsonism Relat Disord*, 18(5), pp. 598-601.

Rodriguez-Oroz, M.C., Lage, P.M., Sanchez-Mut, J., Lamet, I., Pagonabarraga, J., Toledo, J.B., Garcia-Garcia, D., Clavero, P., Samaranch, L., Irurzun, C., Matsubara, J.M., Irigoien, J., Bescos, E., Kulisevsky, J., Perez-Tur, J. and Obeso, J.A. (2009) 'Homocysteine and cognitive impairment in Parkinson's disease: a biochemical, neuroimaging, and genetic study', *Mov Disord*, 24(10), pp. 1437-44.

Romenets, S.R., Wolfson, C., Galatas, C., Pelletier, A., Altman, R., Wadup, L. and Postuma, R.B. (2012) 'Validation of the non-motor symptoms questionnaire (NMS-Quest)', *Parkinsonism Relat Disord*, 18(1), pp. 54-8.

Rost, N.S., Wolf, P.A., Kase, C.S., Kelly-Hayes, M., Silbershatz, H., Massaro, J.M., D'Agostino, R.B., Franzblau, C. and Wilson, P.W. (2001) 'Plasma concentration of C-reactive protein and risk of ischemic stroke and transient ischemic attack: the Framingham study', *Stroke*, 32(11), pp. 2575-9.

Rowan, E., Morris, C.M., Stephens, S., Ballard, C., Dickinson, H., Rao, H., Saxby, B.K., McLaren, A.T., Kalaria, R.N. and Kenny, R.A. (2005) 'Impact of hypertension and apolipoprotein E4 on poststroke cognition in subjects >75 years of age', *Stroke*, 36(9), pp. 1864-8.

Sacco, R.L., Benjamin, E.J., Broderick, J.P., Dyken, M., Easton, J.D., Feinberg, W.M., Goldstein, L.B., Gorelick, P.B., Howard, G., Kittner, S.J., Manolio, T.A., Whisnant, J.P. and Wolf, P.A. (1997) 'American Heart Association Prevention Conference. IV. Prevention and Rehabilitation of Stroke. Risk factors', *Stroke*, 28(7), pp. 1507-17.

Sachdev, P.S., Chen, X., Brodaty, H., Thompson, C., Altendorf, A. and Wen, W. (2009) 'The determinants and longitudinal course of post-stroke mild cognitive impairment', *J Int Neuropsychol Soc*, 15(6), pp. 915-23.

Sahakian, B.J., Morris, R.G., Evenden, J.L., Heald, A., Levy, R., Philpot, M. and Robbins, T.W. (1988) 'A comparative study of visuospatial memory and learning in Alzheimer-type dementia and Parkinson's disease', *Brain*, 111 (Pt 3), pp. 695-718.

Sahathevan, R., Brodtmann, A. and Donnan, G.A. (2012) 'Dementia, stroke, and vascular risk factors; a review', *Int J Stroke*, 7(1), pp. 61-73.

Salazar, J.F., Herbeth, B., Siest, G. and Leroy, P. (1999) 'Stability of blood homocysteine and other thiols: EDTA or acidic citrate?', *Clin Chem*, 45(11), pp. 2016-9.
Sandoval, J., Heyn, H., Moran, S., Serra-Musach, J., Pujana, M.A., Bibikova, M. and Esteller, M. (2011) 'Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome', *Epigenetics*, 6(6), pp. 692-702.

Saracchi, E., Fermi, S. and Brighina, L. (2014) 'Emerging candidate biomarkers for Parkinson's disease: a review', *Aging Dis*, 5(1), pp. 27-34.

Sato, F., Tsuchiya, S., Meltzer, S.J. and Shimizu, K. (2011) 'MicroRNAs and epigenetics', *FEBS J*, 278(10), pp. 1598-609.

Sato, H., Suzuki, S., Kobayashi, H., Ogino, S., Inomata, A. and Arakawa, M. (1991) 'Immunohistological localization of apolipoproteins in the glomeruli in renal disease: specifically apoB and apoE', *Clin Nephrol*, 36(3), pp. 127-33.

Scavone, C., Munhoz, C.D., Kawamoto, E.M., Glezer, I., de Sa Lima, L., Marcourakis, T. and Markus, R.P. (2005) 'Age-related changes in cyclic GMP and PKG-stimulated cerebellar Na,K-ATPase activity', *Neurobiol Aging*, 26(6), pp. 907-16.

Schapira, A.H. and Jenner, P. (2011) 'Etiology and pathogenesis of Parkinson's disease', *Mov Disord*, 26(6), pp. 1049-55.

Scherzer, C.R., Eklund, A.C., Morse, L.J., Liao, Z., Locascio, J.J., Fefer, D., Schwarzschild, M.A., Schlossmacher, M.G., Hauser, M.A., Vance, J.M., Sudarsky, L.R., Standaert, D.G., Growdon, J.H., Jensen, R.V. and Gullans, S.R. (2007) 'Molecular markers of early Parkinson's disease based on gene expression in blood', *Proc Natl Acad Sci U S A*, 104(3), pp. 955-60.

Schrag, A., Jahanshahi, M. and Quinn, N. (2000) 'What contributes to quality of life in patients with Parkinson's disease?', *J Neurol Neurosurg Psychiatry*, 69(3), pp. 308-12.

Schwalbe, E.C., Williamson, D., Lindsey, J.C., Hamilton, D., Ryan, S.L., Megahed, H., Garami, M., Hauser, P., Dembowska-Baginska, B., Perek, D., Northcott, P.A., Taylor, M.D., Taylor, R.E., Ellison, D.W., Bailey, S. and Clifford, S.C. (2013) 'DNA methylation profiling of medulloblastoma allows robust subclassification and improved outcome prediction using formalin-fixed biopsies', *Acta Neuropathol*, 125(3), pp. 359-71.

Scott, S.A., Mufson, E.J., Weingartner, J.A., Skau, K.A. and Crutcher, K.A. (1995) 'Nerve growth factor in Alzheimer's disease: increased levels throughout the brain coupled with declines in nucleus basalis', *J Neurosci*, 15(9), pp. 6213-21.

Shadrina, M.I., Filatova, E.V., Karabanov, A.V., Slominsky, P.A., Illarioshkin, S.N., Ivanova-Smolenskaya, I.A. and Limborska, S.A. (2010) 'Expression analysis of suppression of tumorigenicity 13 gene in patients with Parkinson's disease', *Neurosci Lett*, 473(3), pp. 257-9.

Shehadeh, L.A., Yu, K., Wang, L., Guevara, A., Singer, C., Vance, J. and Papapetropoulos, S. (2010) 'SRRM2, a potential blood biomarker revealing high alternative splicing in Parkinson's disease', *PLoS One*, 5(2), p. e9104.

Shen, L., Kantarjian, H., Guo, Y., Lin, E., Shan, J., Huang, X., Berry, D., Ahmed, S., Zhu, W., Pierce, S., Kondo, Y., Oki, Y., Jelinek, J., Saba, H., Estey, E. and Issa, J.P. (2010) 'DNA methylation predicts survival and response to therapy in patients with myelodysplastic syndromes', *J Clin Oncol*, 28(4), pp. 605-13.

Shenker, N. and Flanagan, J.M. (2012) 'Intragenic DNA methylation: implications of this epigenetic mechanism for cancer research', *Br J Cancer*, 106(2), pp. 248-53.

Shenker, N.S., Ueland, P.M., Polidoro, S., van Veldhoven, K., Ricceri, F., Brown, R., Flanagan, J.M. and Vineis, P. (2013) 'DNA methylation as a long-term biomarker of exposure to tobacco smoke', *Epidemiology*, 24(5), pp. 712-6.

Siegmund, K.D., Connor, C.M., Campan, M., Long, T.I., Weisenberger, D.J., Biniszkiewicz, D., Jaenisch, R., Laird, P.W. and Akbarian, S. (2007) 'DNA methylation in the human cerebral cortex is dynamically regulated throughout the life span and involves differentiated neurons', *PLoS One*, 2(9), p. e895.

Sniderman, A.D., Islam, S., Yusuf, S. and McQueen, M.J. (2012) 'Discordance analysis of apolipoprotein B and non-high density lipoprotein cholesterol as markers of cardiovascular risk in the INTERHEART study', *Atherosclerosis*, 225(2), pp. 444-9.

Stebbins, G.T., Nyenhuis, D.L., Wang, C., Cox, J.L., Freels, S., Bangen, K., deToledo-Morrell, L., Sripathirathan, K., Moseley, M., Turner, D.A., Gabrieli, J.D. and Gorelick, P.B. (2008) 'Gray matter atrophy in patients with ischemic stroke with cognitive impairment', *Stroke*, 39(3), pp. 785-93.

Stefansson, O.A. and Esteller, M. (2013) 'Epigenetic modifications in breast cancer and their role in personalized medicine', *Am J Pathol*, 183(4), pp. 1052-63.

Stephens, S., Kenny, R.A., Rowan, E., Allan, L., Kalaria, R.N., Bradbury, M. and Ballard, C.G. (2004) 'Neuropsychological characteristics of mild vascular cognitive impairment and dementia after stroke', *Int J Geriatr Psychiatry*, 19(11), pp. 1053-7.

Stern, L.L., Mason, J.B., Selhub, J. and Choi, S.W. (2000) 'Genomic DNA hypomethylation, a characteristic of most cancers, is present in peripheral leukocytes of individuals who are homozygous for the C677T polymorphism in the methylenetetrahydrofolate reductase gene', *Cancer Epidemiol Biomarkers Prev*, 9(8), pp. 849-53.

Stroke Health Center. (2013) 'Ischemic versus haemorrhagic stroke'. Available at: http://www.webmd.com/stroke/ischemic-versus-hemorrhagic-stroke. Accessed: 02/09/14.

Sudlow, C.L. and Warlow, C.P. (1996) 'Comparing stroke incidence worldwide: what makes studies comparable?', *Stroke*, 27(3), pp. 550-8.

Sule, Z., Mracsko, E., Bereczki, E., Santha, M., Csont, T., Ferdinandy, P., Bari, F. and Farkas, E. (2009) 'Capillary injury in the ischemic brain of hyperlipidemic, apolipoprotein B-100 transgenic mice', *Life Sci*, 84(25-26), pp. 935-9.

Sun, Z., Chai, H.S., Wu, Y., White, W.M., Donkena, K.V., Klein, C.J., Garovic, V.D., Therneau, T.M. and Kocher, J.P. (2011) 'Batch effect correction for genome-wide methylation data with Illumina Infinium platform', *BMC Med Genomics*, 4, p. 84.

Sung, H.Y., Choi, E.N., Ahn Jo, S., Oh, S. and Ahn, J.H. (2011) 'Amyloid protein-mediated differential DNA methylation status regulates gene expression in Alzheimer's disease model cell line', *Biochem Biophys Res Commun*, 414(4), pp. 700-5.

Szyf, M. (2009) 'The early life environment and the epigenome', *Biochim Biophys Acta*, 1790(9), pp. 878-85.

Talens, R.P., Boomsma, D.I., Tobi, E.W., Kremer, D., Jukema, J.W., Willemsen, G., Putter, H., Slagboom, P.E. and Heijmans, B.T. (2010) 'Variation, patterns, and temporal stability of DNA methylation: considerations for epigenetic epidemiology', *FASEB J*, 24(9), pp. 3135-44.

Talens, R.P., Christensen, K., Putter, H., Willemsen, G., Christiansen, L., Kremer, D., Suchiman, H.E., Slagboom, P.E., Boomsma, D.I. and Heijmans, B.T. (2012) 'Epigenetic variation during the adult lifespan: cross-sectional and longitudinal data on monozygotic twin pairs', *Aging Cell*, 11(4), pp. 694-703.

Tarazi, F.I., Sahli, Z.T., Wolny, M. and Mousa, S.A. (2014) 'Emerging therapies for Parkinson's disease: From bench to bedside', *Pharmacol Ther*.

Tatemichi, T.K., Desmond, D.W., Paik, M., Figueroa, M., Gropen, T.I., Stern, Y., Sano, M., Remien, R., Williams, J.B., Mohr, J.P. and et al. (1993) 'Clinical determinants of dementia related to stroke', *Ann Neurol*, 33(6), pp. 568-75.

Teschendorff, A.E., Menon, U., Gentry-Maharaj, A., Ramus, S.J., Gayther, S.A., Apostolidou, S., Jones, A., Lechner, M., Beck, S., Jacobs, I.J. and Widschwendter, M. (2009) 'An epigenetic signature in peripheral blood predicts active ovarian cancer', *PLoS One*, 4(12), p. e8274.

The Gene Ontology Consortium. (2000) 'Gene ontology: tool for the unification of biology', *Nat Genet.* 25(1):25-9. Available at: http://geneontology.org/. Accessed: 02/09/14.

Thomas, V.S. and Rockwood, K.J. (2001) 'Alcohol abuse, cognitive impairment, and mortality among older people', *J Am Geriatr Soc*, 49(4), pp. 415-20.

Tilvis, R.S., Kahonen-Vare, M.H., Jolkkonen, J., Valvanne, J., Pitkala, K.H. and Strandberg, T.E. (2004) 'Predictors of cognitive decline and mortality of aged people over a 10-year period', *J Gerontol A Biol Sci Med Sci*, 59(3), pp. 268-74.

Timpson, N.J., Lawlor, D.A., Harbord, R.M., Gaunt, T.R., Day, I.N., Palmer, L.J., Hattersley, A.T., Ebrahim, S., Lowe, G.D., Rumley, A. and Davey Smith, G. (2005) 'C-reactive protein and its role in metabolic syndrome: mendelian randomisation study', *Lancet*, 366(9501), pp. 1954-9.

Toole, J.F., Malinow, M.R., Chambless, L.E., Spence, J.D., Pettigrew, L.C., Howard, V.J., Sides, E.G., Wang, C.H. and Stampfer, M. (2004) 'Lowering homocysteine in patients with ischemic stroke to prevent recurrent stroke, myocardial infarction, and death: the Vitamin Intervention for Stroke Prevention (VISP) randomized controlled trial', *JAMA*, 291(5), pp. 565-75.

Touleimat, N. and Tost, J. (2012) 'Complete pipeline for Infinium((R)) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation', *Epigenomics*, 4(3), pp. 325-41.

Trion, A., de Maat, M.P., Jukema, J.W., van der Laarse, A., Maas, M.C., Offerman, E.H., Havekes, L.M., Szalai, A.J., Princen, H.M. and Emeis, J.J. (2005) 'No effect of C-reactive protein on early atherosclerosis development in apolipoprotein E*3-leiden/human C-reactive protein transgenic mice', *Arterioscler Thromb Vasc Biol*, 25(8), pp. 1635-40.

Tzourio, C., Anderson, C., Chapman, N., Woodward, M., Neal, B., MacMahon, S. and Chalmers, J. (2003) 'Effects of blood pressure lowering with perindopril and indapamide therapy on dementia and cognitive decline in patients with cerebrovascular disease', *Arch Intern Med*, 163(9), pp. 1069-75.

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

Vartiainen, E., Seppala, T., Lillsunde, P. and Puska, P. (2002) 'Validation of self reported smoking by serum cotinine measurement in a community-based study', *J Epidemiol Community Health*, 56(3), pp. 167-70.

Vaverkova, H., Karasek, D., Novotny, D., Jackuliakova, D., Lukes, J., Halenka, M. and Frohlich, J. (2009) 'Apolipoprotein B versus LDL-cholesterol: Association with other risk factors for atherosclerosis', *Clin Biochem*, 42(12), pp. 1246-51.

Wagner, J.R., Busche, S., Ge, B., Kwan, T., Pastinen, T. and Blanchette, M. (2014) 'The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts', *Genome Biol*, 15(2), p. R37.

Walldius, G., Aastveit, A.H. and Jungner, I. (2006) 'Stroke mortality and the apoB/apoA-I ratio: results of the AMORIS prospective study', *J Intern Med*, 259(3), pp. 259-66.

Wang, C.C., Lu, T.H., Liao, W.C., Yuan, S.C., Kuo, P.C., Chuang, H.L., Lee, M.C. and Yen, C.H. (2010) 'Cigarette smoking and cognitive impairment: a 10-year cohort study in Taiwan', *Arch Gerontol Geriatr*, 51(2), pp. 143-8.

Wang, S.C., Oelze, B. and Schumacher, A. (2008) 'Age-specific epigenetic drift in late-onset Alzheimer's disease', *PLoS One*, 3(7), p. e2698.

Warlow, C., Sudlow, C., Dennis, M., Wardlaw, J. and Sandercock, P. (2003) 'Stroke', *Lancet*, 362(9391), pp. 1211-24.

Wei, S.H., Balch, C., Paik, H.H., Kim, Y.S., Baldwin, R.L., Liyanarachchi, S., Li, L., Wang, Z., Wan, J.C., Davuluri, R.V., Karlan, B.Y., Gifford, G., Brown, R., Kim, S., Huang, T.H. and Nephew, K.P. (2006) 'Prognostic DNA methylation biomarkers in ovarian cancer', *Clin Cancer Res*, 12(9), pp. 2788-94.

Weidner, C.I., Lin, Q., Koch, C.M., Eisele, L., Beier, F., Ziegler, P., Bauerschlag, D.O., Jockel, K.H., Erbel, R., Muhleisen, T.W., Zenke, M., Brummendorf, T.H. and Wagner, W. (2014) 'Aging of blood can be tracked by DNA methylation changes at just three CpG sites', *Genome Biol*, 15(2), p. R24.

Weisenberger, D.J., Campan, M., Long, T.I., Kim, M., Woods, C., Fiala, E., Ehrlich, M. and Laird, P.W. (2005) 'Analysis of repetitive element DNA methylation by MethyLight', *Nucleic Acids Res*, 33(21), pp. 6823-36.

Wesnes, K.A., McKeith, I.G., Ferrara, R., Emre, M., Del Ser, T., Spano, P.F., Cicin-Sain, A., Anand, R. and Spiegel, R. (2002) 'Effects of rivastigmine on cognitive function in dementia with lewy bodies: a randomised placebo-controlled international study using the cognitive drug research computerised assessment system', *Dement Geriatr Cogn Disord*, 13(3), pp. 183-92.

Wickelgren, I. (2012) 'Trying to forget', *Scientific American Mind*, 33-9.

Wu, H. and Zhang, Y. (2011) 'Mechanisms and functions of Tet protein-mediated 5-methylcytosine oxidation', *Genes Dev*, 25(23), pp. 2436-52.

Xu, Y., Yan, J., Zhou, P., Li, J., Gao, H., Xia, Y. and Wang, Q. (2012) 'Neurotransmitter receptors and cognitive dysfunction in Alzheimer's disease and Parkinson's disease', *Prog Neurobiol*, 97(1), pp. 1-13.

Yang, A.S., Estecio, M.R., Doshi, K., Kondo, Y., Tajara, E.H. and Issa, J.P. (2004) 'A simple method for estimating global DNA methylation using bisulfite PCR of repetitive DNA elements', *Nucleic Acids Res*, 32(3), p. e38.

Yang, H.H., Hu, N., Wang, C., Ding, T., Dunn, B.K., Goldstein, A.M., Taylor, P.R. and Lee, M.P. (2010) 'Influence of genetic background and tissue types on global DNA methylation patterns', *PLoS One*, 5(2), p. e9355.

Yang, Q., Shan, L., Yoshimura, G., Nakamura, M., Nakamura, Y., Suzuma, T., Umemura, T., Mori, I., Sakurai, T. and Kakudo, K. (2002) '5-aza-2'-deoxycytidine induces retinoic acid receptor beta 2 demethylation, cell cycle arrest and growth inhibition in breast carcinoma cells', *Anticancer Res*, 22(5), pp. 2753-6.

Yarnall, A.J., Breen, D.P., Duncan, G.W., Khoo, T.K., Coleman, S.Y., Firbank, M.J., Nombela, C., Winder-Rhodes, S., Evans, J.R., Rowe, J.B., Mollenhauer, B., Kruse, N., Hudson, G., Chinnery, P.F., O'Brien, J.T., Robbins, T.W., Wesnes, K., Brooks, D.J., Barker, R.A. and Burn, D.J. (2014) 'Characterizing mild cognitive impairment in incident Parkinson disease: the ICICLE-PD study', *Neurology*, 82(4), pp. 308-16.

Yarnall, A.J., Rochester, L. and Burn, D.J. (2013) 'Mild cognitive impairment in Parkinson's disease', *Age Ageing*, 42(5), pp. 567-76.

Yesavage, J.A., Brink, T.L., Rose, T.L., Lum, O., Huang, V., Adey, M. and Leirer, V.O. (1982) 'Development and validation of a geriatric depression screening scale: a preliminary report', *J Psychiatr Res*, 17(1), pp. 37-49.

Yu, L., Chibnik, L.B., Srivastava, G.P., Pochet, N., Yang, J., Xu, J., Kozubek, J., Obholzer, N., Leurgans, S.E., Schneider, J.A., Meissner, A., De Jager, P.L. and Bennett, D.A. (2015) 'Association of Brain DNA Methylation in SORL1, ABCA7, HLA-DRB5, SLC24A4, and BIN1 With Pathological Diagnosis of Alzheimer Disease', *JAMA Neurol*, 72(1), pp. 15-24.

Zeilinger, S., Kuhnel, B., Klopp, N., Baurecht, H., Kleinschmidt, A., Gieger, C., Weidinger, S., Lattka, E., Adamski, J., Peters, A., Strauch, K., Waldenberger, M. and Illig, T. (2013) 'Tobacco smoking leads to extensive genome-wide changes in DNA methylation', *PLoS One*, 8(5), p. e63812.

Zhang, D., Cheng, L., Badner, J.A., Chen, C., Chen, Q., Luo, W., Craig, D.W., Redman, M., Gershon, E.S. and Liu, C. (2010) 'Genetic control of individual differences in gene-specific methylation in human brain', *Am J Hum Genet*, 86(3), pp. 411-9.

Zhang, Y., Yang, R., Burwinkel, B., Breitling, L.P. and Brenner, H. (2014a) 'F2RL3 methylation as a biomarker of current and lifetime smoking exposures', *Environ Health Perspect*, 122(2), pp. 131-7.

Zhang, Y.W., Chen, Y., Liu, Y., Zhao, Y., Liao, F.F. and Xu, H. (2013) 'APP regulates NGF receptor trafficking and NGF-mediated neuronal differentiation and survival', *PLoS One*, 8(11), p. e80571.

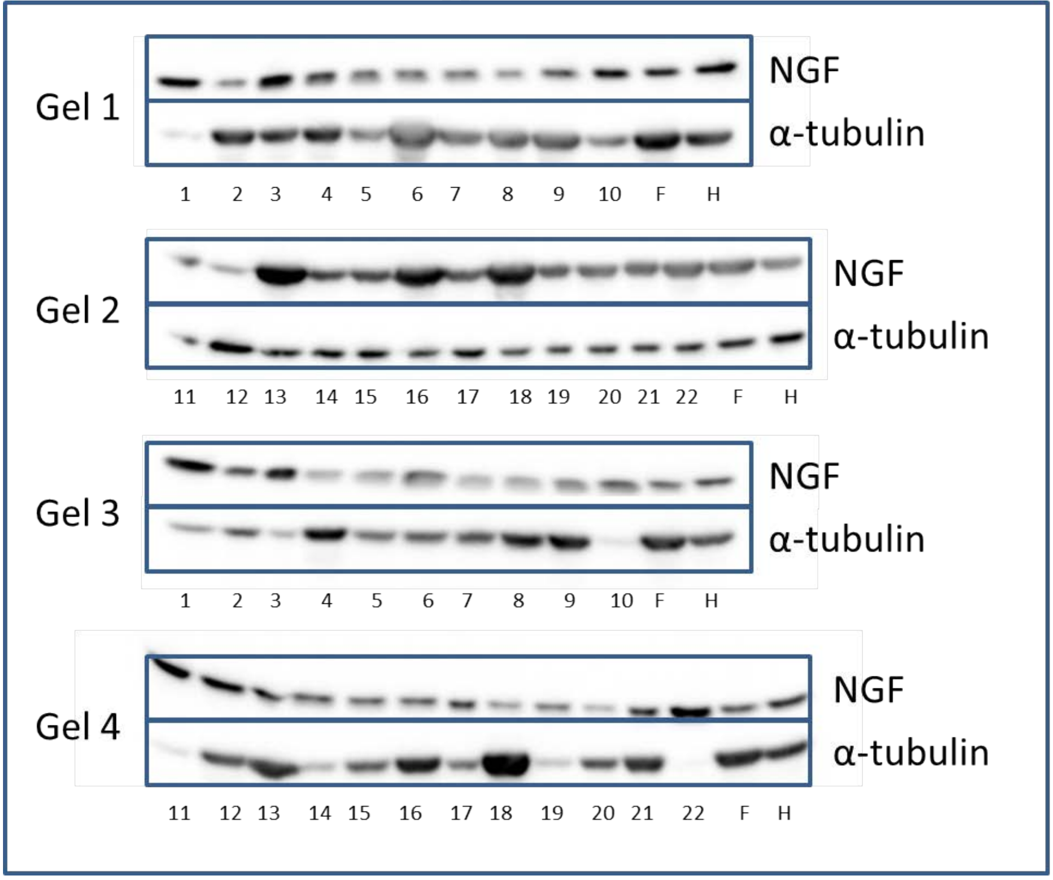
Zhang, Z.L., Wu, W.C., Liu, J.Q., Yao, Y.B., Pan, M.D., Yang, C.B., Wang, J.G., Huang, X.W. and Lin, J.Y. (2014b) 'Screening of differentially expressed genes related to ischemic stroke and functional analysis with DNA microarray', *Eur Rev Med Pharmacol Sci*, 18(8), pp. 1181-8.

Appendix A: Pyrosequencing dispensation sequences

Cohort	Assay	Dispensation Order
COGFAST	*CISD1*	CATCGATATATCAGATCATCGATAC
COGFAST	cg21463981	TGTCGTAG
COGFAST	*LIMK2*	CTACAGACAC
COGFAST	*HSPB3*	TATCGATAGATGATGAGTGATCGTTATTAGAGTTGATCGAT
COGFAST	*EIF4E3*	ATCGTTATAG
COGFAST	cg02490189	TATCGTTAGTCGTATGTCGTGATCGTAGTAGTCGTT
COGFAST	*APOB*	GTCGATAGTCGAT
COGFAST	*C17orf101*	GATGTCGTATGTCGTAGTTATGTCGAGGTATCAGTCGTGTTGATCGTG
COGFAST	cg18837178	TCGACATACGACTC
COGFAST	cg01063243	ATATCAGATACCAGATCAGATATATCAGATATATCAGATATATCAGATCAGACATATCAGACATCGACTC
COGFAST	*NGF*	ATCTGTCGTATTCGTGATTCGTG
COGFAST	cg12010173	CGACATCTCTCAACATATTACACACACATATAATCATCAGATCGATATACTACGACAC
ICICLE	*PROCA1*	ATCGATGT
ICICLE	*CHCHD5*	AGTCTGTCAGTCGTG
ICICLE	*TRIM15*	ATCTGTCGAGA
ICICLE	*C7orf50**	GTCGATATGGTCTAGTGGTCGTAGAGTGAGTATCG
ICICLE	*C16orf45**	TGTTCGAGTCGCTCAGATCGTGAGTGTTGTGAGTCG
ICICLE	*PGS1**	GTCGTGATCGTAGTCTATGAGATTGTCGTCG
ICICLE	*RUNDC3A**	GTCGTTCGTCATCGATGTGTAGAGTCG
ICICLE	*MOGAT1**	GTCGTATCAGTATGATGATCGTCGTCGTGTAGATCG

* indicates a pre-designed assay (Qiagen, UK).

Appendix B: Western blots



Gels 1 and 2=DLPFC samples. Gels 3 and 4=Hippocampal samples. F=pooled DLPFC standard. H=pooled hippocampal standard. NGF=27kDA, α-tubulin=55kDA.

Sample Number	Diagnosis	Sex	Age at death	Braak staging
1	SC	F	83	1
2	SD	F	80	1
3	SC	M	83	2
4	SD	M	83	2
5	SC	M	84	3
6	SD	M	84	3
7	SC	F	83	4
8	SD	F	84	4
9	SC	M	88	5
10	SD	M	84	5
11	DS	M	92	6
12	DS	M	90	6
13	DS	F	86	6
14	DS	F	89	6
15	DS	M	80	5
16	DS	F	83	5
17	CA	M	87	2
18	CA	F	89	2
19	CA	M	83	2
20	CA	F	87	3
21	CA	F	88	3
22	CA	M	91	2

SC=COGFAST cognitively normal, SD=COGFAST demented, DS=Alzheimer's non-stroke, CA=healthy control. F=female, M=male.

Appendix C: Associations between exposures and outcome

PSD

Associations for stroke-related and lifestyle variables with outcome. P values in red show a significant association. * = Kruskal Wallis, † = chi squared, ‡ = Wilcoxon rank sum, §= t test, ¶ = Spearman’s rank

	Stroke-related variables					Lifestyle factors	
Outcome Variable	OCSP class	Side of body	Degree of weakness arm	Degree of weakness leg	Dysphasia	Smoking	Alcohol
Diagnosis	1.903 0.610†	0.283 0.868†	0.081 0.947†	0.346 0.841†	0.251 0.616†	2.343 0.307†	-1.034 0.301‡
Braak staging	9.579 0.699†	11.99 0.348†	8.631 0.343†	5.406 0.727†	2.531 0.693†	8.490 0.207†	7.397 0.116*
MMSE	3.176 0.365*	0.133 0.936*	0.473 0.789*	2.955 0.228*	0.121 0.728‡	0.730 0.694*	0.093 0.262¶
Orientation	3.178 0.365*	3.139 0.208*	0.095 0.954*	1.465 0.481*	0.270 0.6033‡	0.847 0.655*	0.083 0.326¶
Language comprehension	2.288 0.515*	1.437 0.488*	0.645 0.724*	4.604 0.100*	2.302 0.129‡	0.824 0.662*	0.088 0.301¶
Language expression	0.898 0.826*	0.517 0.772*	0.765 0.682*	0.413 0.814*	0.250 0.803‡	0.361 0.835*	-0.035 0.682¶
Memory remote	0.820 0.845*	4.397 0.111*	1.004 0.605*	1.036 0.596*	1.298 0.194‡	4.716 0.095*	0.083 0.326¶
Memory recent	3.887 0.274*	0.593 0.743*	1.575 0.455*	1.428 0.490*	0.264 0.794‡	1.975 0.373*	0.064 0.451¶
Memory learning	2.703 0.440*	3.261 0.196*	0.769 0.681*	2.798 0.247*	-1.208 0.227‡	1.346 0.510*	0.015 0.863¶
Memory total	1.267 0.737*	0.578 0.749*	2.932 0.231*	2.954 0.228*	-0.579 0.563‡	1.199 0.549*	0.072 0.396¶
Attention	6.874 0.076*	0.757 0.685*	1.244 0.537*	2.879 0.237*	-0.241 0.809‡	3.794 0.150*	0.090 0.288¶
Praxis	5.915 0.116*	0.555 0.758*	3.843 0.146*	4.467 0.107*	-0.645 0.519‡	0.324 0.850*	-0.029 0.735¶
Calculation	5.360 0.147*	0.267 0.875*	0.788 0.674*	1.311 0.519*	0.938 0.348‡	5.255 0.072*	-0.001 0.992¶
Abstract thinking	2.684 0.443*	0.338 0.845*	1.033 0.597*	3.123 0.210*	1.187 0.235‡	0.445 0.801*	0.076 0.368¶
Perception	3.972 0.265*	1.597 0.450*	0.278 0.870*	1.258 0.533*	-1.798 0.072‡	3.945 0.139*	0.140 0.097¶
Executive function	4.506 0.212*	0.633 0.729*	1.268 0.530*	2.954 0.228*	-0.579 0.563§	1.199 0.549*	0.072 0.396¶
Total CAMCOG	4.104 0.250*	0.555 0.758*	0.304 0.859*	3.395 0.183*	-0.400 0.689‡	1.491 0.474*	0.112 0.184¶

Associations for medical history related variables with outcome. P values in red show a significant association. * = Kruskal Wallis, † = chi squared, ‡ = Wilcoxon rank sum, §= t test, ¶ = Spearman’s rank

Outcome Variable	**HT**	**AF**	**IHD/Angina**	**Cardiac failure**	**IC**	**No of CVD risk factors**
Diagnosis	0.366 $0.545^{\dagger}$	1.919 $0.239^{\dagger}$	0.838 $0.360^{\dagger}$	0.340 $0.560^{\dagger}$	1.065 $0.390^{\dagger}$	-0.736 $0.462^{\ddagger}$
Braak staging	1.251 $0.890^{\dagger}$	4.555 $0.330^{\dagger}$	2.569 $0.743^{\dagger}$	3.524 $0.469^{\dagger}$	8.077 $0.169^{\dagger}$	2.733 0.603*
MMSE	1.438 $0.150^{\ddagger}$	-1.133 $0.257^{\ddagger}$	-0.002 $0.998^{\ddagger}$	-1.145 $0.248^{\ddagger}$	-1.263 $0.207^{\ddagger}$	-0.030 $0.716^{\P}$
Orientation	0.894 $0.371^{\ddagger}$	-0.848 $0.396^{\ddagger}$	0.301 $0.764^{\ddagger}$	-0.965 $0.335^{\ddagger}$	-1.891 $0.059^{\ddagger}$	-0.014 $0.868^{\P}$
Language comprehension	0.174 $0.862^{\ddagger}$	0.919 $0.358^{\ddagger}$	-1.256 $0.209^{\ddagger}$	-1.596 $0.111^{\ddagger}$	-0.621 $0.535^{\ddagger}$	0.040 $0.637^{\P}$
Language expression	0.751 $0.452^{\ddagger}$	-0.365 $0.715^{\ddagger}$	0.761 $0.447^{\ddagger}$	0.859 $0.391^{\ddagger}$	-1.427 $0.154^{\ddagger}$	0.003 $0.971^{\P}$
Memory remote	1.711 $0.087^{\ddagger}$	-0.538 $0.590^{\ddagger}$	0.007 $0.995^{\ddagger}$	0.189 $0.850^{\ddagger}$	-2.763 $0.006^{\ddagger}$	0.011 $0.893^{\P}$
Memory recent	0.998 $0.318^{\ddagger}$	-0.690 $0.491^{\ddagger}$	-1.145 $0.252^{\ddagger}$	-1.322 $0.186^{\ddagger}$	-1.461 $0.144^{\ddagger}$	0.070 $0.406^{\P}$
Memory learning	0.283 $0.777^{\ddagger}$	-1.738 $0.082^{\ddagger}$	1.874 $0.061^{\ddagger}$	0.205 $0.838^{\ddagger}$	-0.207 $0.836^{\ddagger}$	-0.016 $0.854^{\P}$
Memory total	1.109 $0.267^{\ddagger}$	-1.661 $0.097^{\ddagger}$	1.364 $0.172^{\ddagger}$	0.037 $0.970^{\ddagger}$	-1.288 $0.198^{\ddagger}$	0.001 $0.995^{\P}$
Attention	1.037 $0.300^{\ddagger}$	-0.455 $0.649^{\ddagger}$	0.553 $0.580^{\ddagger}$	-0.249 $0.804^{\ddagger}$	-1.382 $0.167^{\ddagger}$	0.004 $0.966^{\P}$
Praxis	1.919 $0.055^{\ddagger}$	-1.131 $0.258^{\ddagger}$	0.425 $0.671^{\ddagger}$	0.451 $0.652^{\ddagger}$	0.754 $0.451^{\ddagger}$	-0.143 $0.090^{\P}$
Calculation	0.166 $0.868^{\ddagger}$	0.081 $0.935^{\ddagger}$	0.305 $0.760^{\ddagger}$	1.038 $0.299^{\ddagger}$	-0.188 $0.851^{\ddagger}$	0.003 $0.972^{\P}$
Abstract thinking	0.503 $0.615^{\ddagger}$	-0.674 $0.500^{\ddagger}$	1.021 $0.307^{\ddagger}$	0.156 $0.876^{\ddagger}$	-1.957 $0.051^{\ddagger}$	0.014 $0.090^{\P}$
Perception	0.560 $0.575^{\ddagger}$	-0.377 $0.706^{\ddagger}$	-0.572 $0.567^{\ddagger}$	0.194 $0.847^{\ddagger}$	-1.050 $0.294^{\ddagger}$	0.014 $0.866^{\P}$
Executive function	0.703 $0.483^{\S}$	-0.096 $0.924^{\ddagger}$	0.941 $0.348^{\S}$	-0.288 $0.773^{\ddagger}$	-2.515 $0.013^{\S}$	0.046 $0.589^{\P}$
Total CAMCOG	1.768 $0.077^{\ddagger}$	-1.065 $0.287^{\ddagger}$	0.590 $0.555^{\ddagger}$	0.194 $0.847^{\ddagger}$	-1.050 $0.294^{\ddagger}$	0.099 $0.241^{\P}$

Associations for baseline cognition variables with outcome part 1. P values in red indicate an association where $p<0.05$. * = Kruskal Wallis, † = chi squared, ‡ = Wilcoxon rank sum, §= t test, ¶ = Spearman's rank. B=baseline. Lang comp = language comprehension

Outcome Variable	**B MMSE**	**B lang comp**	**B language expression**	**B memory remote**	**B memory recent**	**B memory learning**	**B memory total**
Diagnosis	4.528 0.0003§	3.661 0.002§	3.555 0.001§	2.282 0.023‡	2.418 0.016‡	2.539 0.011‡	0.352 0.002§
Braak staging	5.643 0.343*	7.743 0.171*	8.319 0.140*	5.358 0.374*	11.90 0.036*	11.60 0.041*	8.626 0.125*
MMSE	0.776 7.77E-07¶	0.614 0.0004¶	0.605 0.001¶	0.471 0.010¶	0.473 0.010¶	0.518 0.004¶	0.606 0.001¶
Orientation	0.674 0.0001¶	0.593 0.001¶	0.520 0.004¶	0.360 0.055¶	0.556 0.002¶	0.451 0.014¶	0.534 0.003¶
Lang comp	0.666 0.0001¶	0.397 0.033¶	0.434 0.019¶	0.241 0.208¶	0.410 0.027¶	0.406 0.029¶	0.449 0.015¶
Language expression	0.698 2.54E-05¶	0.657 0.0001¶	0.689 3.64E-05¶	0.273 0.152¶	0.536 0.003¶	0.505 0.005¶	0.576 0.001¶
Memory remote	0.508 0.005¶	0.576 0.001¶	0.241 0.208¶	0.650 0.0001¶	0.402 0.031¶	0.367 0.050¶	0.461 0.012¶
Memory recent	0.479 0.009¶	0.457 0.013¶	0.120 0.537¶	0.385 0.039¶	0.449 0.015¶	0.387 0.038¶	0.437 0.018¶
Memory learning	0.771 1.01E-06¶	0.577 0.001¶	0.418 0.024¶	0.509 0.005¶	0.592 0.001¶	0.563 0.002¶	0.642 0.0002¶
Memory total	0.735 5.62E-06¶	0.574 0.001¶	0.524 0.004¶	0.681 4.72E-05¶	0.601 0.001¶	0.503 0.006¶	0.626 0.0003¶
Attention	0.695 2.88E-05¶	0.592 0.001¶	0.651 0.0001¶	0.340 0.071¶	0.280 0.141¶	0.440 0.017¶	0.510 0.005¶
Praxis	0.827 3.14E-08¶	0.402 0.031¶	0.578 0.001¶	0.323 0.087¶	0.339 0.072¶	0.290 0.127¶	0.376 0.045¶
Calculation	0.370 0.048¶	0.539 0.003¶	0.304 0.110¶	0.347 0.065¶	0.496 0.006¶	0.202 0.293¶	0.304 0.109¶
Abstract thinking	0.328 0.083¶	0.166 0.388¶	0.211 0.273¶	0.217 0.258¶	0.366 0.051¶	0.257 0.178¶	0.331 0.080¶
Perception	0.640 0.0002¶	0.476 0.009¶	0.376 0.044¶	0.324 0.087¶	0.623 0.0003¶	0.398 0.032¶	0.484 0.008¶
Executive function	0.586 0.001¶	0.407 0.028¶	0.529 0.003¶	0.268 0.159¶	0.442 0.016¶	0.331 0.079¶	0.430 0.020¶
Total CAMCOG	0.807 1.28E-07¶	0.601 0.001¶	0.580 0.001¶	0.446 0.015¶	0.565 0.001¶	0.532 0.003¶	0.624 0.0003¶

Associations for baseline cognition variables with outcome part 2. P values in red indicate an association where p<0.05. * = Kruskal Wallis, † = chi squared, ‡ = Wilcoxon rank sum, §= t test, ¶ = Spearman’s rank. B=baseline. Lang=language. Comp=comprehension. Orient=Orientation. Calc=calculation. Perc=perception

Outcome Variable	**B Orient**	**B Attention**	**B Praxis**	**B Calc**	**B Abstract Thinking**	**B Perc**	**B Executive Function**	**B Total CAMCOG score**
Diagnosis	3.024 0.003‡	3.320 0.001‡	2.490 0.013‡	1.904 0.057‡	0.817 0.421§	3.145 0.0004§	1.280 0.211§	5.066 0.0001§
Braak staging	7.213 0.205*	6.428 0.267*	7.827 0.166*	2.858 0.722*	1.493 0.914*	2.444 0.785*	3.383 0.641*	5.252 0.386*
MMSE	0.732 6.40E-06¶	0.456 0.013¶	0.475 0.009¶	0.475 0.009¶	0.211 0.271¶	0.502 0.006¶	0.425 0.022¶	0.766 1.31E-06¶
Orient	0.597 0.001¶	0.469 0.010¶	0.351 0.062¶	0.469 0.010¶	0.234 0.223¶	0.571 0.001¶	0.408 0.028¶	0.703 2.10E-05¶
Lang comp	0.560 0.002¶	0.478 0.009¶	0.638 0.0002¶	0.565 0.001¶	0.172 0.371¶	0.366 0.051¶	0.355 0.059¶	0.606 0.001¶
Lang expression	0.661 0.0001¶	0.360 0.055¶	0.406 0.029¶	0.539 0.003¶	0.274 0.150¶	0.554 0.002¶	0.446 0.015¶	0.765 1.35E-06¶
Memory remote	0.482 0.008¶	0.206 0.285¶	0.177 0.358¶	0.175 0.003¶	0.073 0.708¶	0.136 0.483¶	0.075 0.698¶	0.386 0.039¶
Memory recent	0.489 0.007¶	0.412 0.026¶	0.039 0.840¶	0.210 0.275¶	-0.066 0.734¶	0.213 0.267¶	0.077 0.691¶	0.387 0.038¶
Memory learning	0.675 0.0001¶	0.577 0.001¶	0.293 0.123¶	0.433 0.019¶	0.156 0.420¶	0.501 0.005¶	0.239 0.211¶	0.702 2.23E-05¶
Memory total	0.726 8.36E-06¶	0.457 0.013¶	0.394 0.034¶	0.409 0.028¶	0.196 0.309¶	0.500 0.006¶	0.298 0.116¶	0.726 8.14E-06¶
Attention	0.705 1.99E-05¶	0.455 0.013¶	0.448 0.015¶	0.500 0.006¶	0.057 0.770¶	0.348 0.064¶	0.282 0.140¶	0.675 0.0001¶
Praxis	0.570 0.001¶	0.585 0.001¶	0.643 0.0002¶	0.541 0.002¶	0.125 0.519¶	0.511 0.005¶	0.300 0.114¶	0.686 3.98E-05¶
Calc	0.352 0.061¶	0.317 0.094¶	0.203 0.292¶	0.488 0.007¶	0.195 0.310¶	0.249 0.194¶	0.363 0.053¶	0.410 0.027¶
Abstract thinking	0.169 0.381¶	0.047 0.811¶	0.234 0.222¶	0.179 0.354¶	0.338 0.073¶	0.336 0.074¶	0.350 0.062¶	0.395 0.034¶
Perc	0.563 0.002¶	0.488 0.007¶	0.397 0.033¶	0.518 0.004¶	0.313 0.098¶	0.671 0.0001¶	0.397 0.033¶	0.638 0.0002¶
Executive function	0.462 0.012¶	0.387 0.038¶	0.483 0.008¶	0.443 0.016¶	0.327 0.084¶	0.511 0.005¶	0.453 0.014¶	0.659 0.0001¶
Total CAMCOG	0.672 0.0001¶	0.514 0.004¶	0.456 0.013¶	0.484 0.008¶	0.235 0.220¶	0.588 0.001¶	0.404 0.030¶	0.080 2.63E-07¶

PD

Associations for lifestyle and anthropometric variables with motor outcomes. Test statistics and p values in red indicate an association where p<0.05. * = Kruskal Wallis, † = chi squared, ‡ = Wilcoxon rank sum, §= t test, ¶ = Spearman's rank

Outcome variable	Education	NART	Alcohol	Smoking	Height	Weight	BMI
Hoehn and Yahr	7.348 0.062*	1.890 0.596*	3.017 0.389*	2.921 0.842†	4.273 0.206*	5.083 0.166*	2.734 0.434*
MDS-UPDRS II	-0.146 0.0968¶	-0.102 0.250¶	-0.176 0.058¶	1.092 0.579*	0.088 0.321¶	0.133 0.130¶	0.068 0.440¶
MDS-UPDRS III	-0.255 0.003¶	-0.149 0.092¶	-0.087 0.352¶	2.451 0.264*	0.052 0.556¶	0.038 0.665¶	-0.002 0.985¶
Tremor dominant phenotype	-0.194 0.027¶	-0.072 0.419¶	-0.114 0.223¶	0.064 0.969*	0.037 0.677¶	0.031 0.725¶	0.046 0.603¶
PIGD phenotype	-0.271 0.002¶	-0.118 0.183¶	-0.126 0.174¶	2.779 0.249*	-0.034 0.702¶	0.001 0.993¶	0.014 0.877¶

Associations for medical history variables with motor outcomes. Test statistics and p values in red indicate an association where p<0.05. * = Kruskal Wallis, † = chi squared, ‡ = Wilcoxon rank sum, §= t test, ¶ = Spearman's rank Homocys=homocysteine.

Outcome variable	RCF	B12	Homocys	IHD	DM	HT	HC	GDS	LD
Hoehn and Yahr	15.00 0.002*	4.060 0.255*	9.928 0.019*	11.41 0.033†	5.852 0.085†	1.593 0.697†	2.388 0.449†	17.61 0.001*	2.919 0.404*
MDS-UPDRS II	0.037 0.693¶	-0.004 0.970¶	0.174 0.061¶	-1.421 0.158§	-0.388 0.708§	-0.647 0.519§	0.335 0.738	0.632 2.13E-14¶	0.182 0.050¶
MDS-UPDRS III	-0.071 0.450¶	-0.215 0.020¶	0.182 0.049¶	-1.778 0.078§	-1.512 0.131§	-2.778 0.006‡	-2.034 0.044§	0.459 1.97E-07¶	0.056 0.548¶
Tremor dominant phenotype	-0.006 0.948¶	0.041 0.663¶	0.033 0.728¶	-1.143 0.255§	-0.639 0.524§	-1.456 0.150§	-1.612 0.109§	0.086 0.357¶	-0.147 0.114¶
PIGD phenotype	0.016 0.867¶	-0.086 0.359¶	0.236 0.011¶	-3.011 0.003‡	-1.304 0.192‡	-1.926 0.054‡	-0.176 0.860‡	0.553 1.06E-10¶	0.096 0.301¶

Associations for lifestyle and anthropometric variables with cognitive outcomes. Test statistics and p values in red indicate an association where p<0.05. * = Kruskal Wallis, † = chi squared, ‡ = Wilcoxon rank sum, §= t test, ¶ = Spearman's rank. Conc=concentration. Dig vig accuracy = digit vigilance accuracy.

Outcome variable	Education	NART	Alcohol	Smoking	Height	Weight	BMI
MoCA total	0.336 0.0001¶	0.369 1.60E-05¶	0.177 0.050¶	5.708 0.058*	0.023 0.795¶	0.028 0.751¶	0.045 0.615¶
MMSE total	0.338 0.0001¶	0.379 8.88E-06¶	0.174 0.055¶	2.951 0.229*	0.038 0.667¶	-0.066 0.457¶	-0.056 0.527¶
Total FAS	0.426 4.25E-07¶	0.459 3.87E-08¶	0.059 0.519¶	0.020 0.990*	-0.048 0.586¶	-0.068 0.442¶	-0.039 0.664¶
Animals	0.392 3.92E-06¶	0.274 0.002¶	0.035 0.698¶	6.181 0.046*	0.009 0.916¶	0.084 0.342¶	0.126 0.154¶
NMSQ memory	1.309 0.191‡	0.536 0.592‡	0.899 0.369‡	4.392 0.111†	-0.396 0.692‡	0.352 0.725§	0.603 0.548§
NMSQ conc	-0.603 0.547‡	1.056 0.291‡	-0.641 0.522‡	1.097 0.503†	-1.732 0.083‡	-1.305 0.194§	-0.198 0.844§
Power of attention	-0.201 0.022¶	-0.320 0.0002¶	-0.120 0.199¶	2.799 0.247*	-0.026 0.773¶	0.018 0.837¶	0.024 0.790¶
Dig vig accuracy	0.386 5.59E-06¶	0.320 0.0002¶	0.044 0.641¶	2.316 0.314*	-0.120 0.173¶	-0.1124 0.203¶	-0.032 0.716¶
PRM	0.419 7.23E-07¶	0.401 2.21E-06¶	0.271 0.003¶	6.841 0.033*	0.085 0.339¶	0.042 0.631¶	0.003 0.970¶
PAL	-0.316 0.0002¶	-0.361 2.49E-05¶	-0.153 0.099¶	9.383 0.009*	-0.028 0.752¶	-0.088 0.317¶	-0.089 0.314¶
SRM	0.232 0.008¶	0.284 0.001¶	0.158 0.089¶	3.277 0.194*	0.050 0.570¶	-0.043 0.631¶	-0.024 0.896¶
OTS	0.284 0.001¶	0.236 0.007¶	0.238 0.010¶	2.777 0.250*	0.093 0.295¶	0.124 0.159¶	0.094 0.288¶
Pentagon copying	0.216 0.014¶	0.276 0.002¶	0.152 0.102¶	0.706 0.703*	0.120 0.175¶	0.017 0.857¶	-0.012 0.893¶
Naming	0.293 0.001¶	0.225 0.010¶	0.182 0.050¶	0.392 0.822*	0.103 0.242¶	0.096 0.279¶	0.038 0.671¶
Language	0.135 0.127¶	0.278 0.001¶	0.233 0.011¶	6.170 0.046*	-0.005 0.955¶	-0.026 0.769¶	-0.018 0.837¶
Language total	0.215 0.014¶	0.327 0.0001¶	0.260 0.005¶	4.791 0.091*	0.053 0.547¶	0.030 0.733¶	0.013 0.881¶
Cognitive complaint	1.615 0.106‡	1.335 0.182‡	0.098 0.922‡	1.645 0.47†	-1.027 0.305‡	-0.540 0.590§	0.188 0.851§
MCI	3.358 0.001‡	4.283 1.84E-05‡	2.169 0.030	3.469 0.209†	0.332 0.741§	0.828 0.409§	0.793 0.429§
TOTAL number NMS	-0.145 0.109¶	-0.076 0.401¶	-0.070 0.440¶	0.716 0.699*	0.186 0.039¶	0.046 0.610¶	-0.061 0.498¶

Associations for medical history variables with cognitive outcomes. Test statistics and p values in red indicate an association where p<0.05. * = Kruskal Wallis, † = chi squared, ‡ = Wilcoxon rank sum, §= t test, ¶ = Spearman’s rank. Conc=concentration. Homocys=homocysteine. LD=Levodopa dose. Dig vig accuracy = digit vigilance accuracy.

Outcome variable	**RCF**	**B12**	**Homocys**	**IHD**	**DM**	**HT**	**HC**	**GDS**	**LD**
MoCA total	-0.086 0.346¶	0.104 0.254¶	-0.134 0.139¶	2.351 0.019‡	2.086 0.037‡	2.206 0.027‡	1.330 0.183‡	-0.253 0.005¶	0.040 0.661¶
MMSE total	-0.065 0.475¶	-0.003 0.971¶	-0.093 0.304¶	1.376 0.169‡	1.194 0.233‡	3.782 0.0002‡	1.098 0.272‡	-0.265 0.003¶	0.010 0.911¶
Total FAS	-0.057 0.530¶	0.199 0.027¶	-0.161 0.075¶	2.483 0.013‡	1.803 0.071‡	2.298 0.022‡	0.085 0.932‡	-0.265 0.003¶	0.009 0.923¶
Animals	0.050 0.586¶	0.035 0.698¶	-0.153 0.092¶	2.041 0.049§	1.303 0.195§	2.007 0.047§	1.103 0.272§	-0.301 0.001¶	0.030 0.740¶
NMSQ memory	2.254 0.024‡	0.129 0.897‡	-1.707 0.088‡	0.874 0.350†	1.962 0.297†	0.241 0.624†	0.006 0.941†	-3.184 0.002‡	-1.158 0.002‡
NMSQ conc	0.283 0.777‡	-0.774 0.439‡	-0.134 0.894‡	0.054 0.816†	0.230 0.632†	0.678 0.410†	0.158 0.778†	-3.301 0.001‡	-3.301 0.001§
Power of attention	0.197 0.034¶	-0.155 0.096¶	-0.025 0.788¶	-1.308 0.191‡	-1.064 0.287‡	-1.097 0.273‡	-1.687 0.092‡	-0.076 0.419¶	0.228 0.014¶
Dig vig accuracy	-0.121 0.194¶	0.241 0.009¶	-0.116 0.215¶	0.033 0.973‡	0.796 0.426‡	0.967 0.334‡	1.241 0.215‡	-0.031 0.740¶	-0.267 0.004¶
PRM	-0.024 0.796¶	0.147 0.114¶	-0.209 0.024¶	1.247 0.212‡	3.048 0.002‡	3.680 0.0002‡	0.460 0.645‡	0.067 0.475¶	-0.143 0.125¶
PAL	0.128 0.170¶	-0.055 0.555¶	0.101 0.278¶	-1.303 0.193‡	-1.543 0.123‡	-2.893 0.004‡	-1.419 0.156‡	-0.013 0.890¶	0.196 0.034¶
SRM	-0.134 0.151¶	0.289 0.002¶	-0.170 0.068¶	0.125 0.901‡	1.704 0.091‡	0.527 0.598§	-0.456 0.648‡	-0.156 0.093¶	-0.159 0.087¶
OTS	0.028 0.761¶	0.020 0.830¶	-0.130 0.164¶	1.863 0.062‡	1.097 0.273‡	2.577 0.010‡	1.118 0.264‡	0.100 0.283¶	-0.220 0.017¶
Pentagon copying	-0.191 0.039¶	-0.027 0.771¶	-0.178 0.056¶	0.763 0.445‡	2.784 0.005‡	2.290 0.022‡	1.358 0.175‡	-0.001 0.992¶	-0.188 0.043¶
Naming	0.024 0.795¶	0.021 0.820¶	0.124 0.184¶	0.456 0.659‡	1.262 0.207‡	0.564 0.573‡	0.355 0.723‡	0.115 0.219¶	-0.080 0.392¶
Language	0.021 0.826¶	0.151 0.105¶	-0.160 0.085¶	1.458 0.145‡	2.558 0.011‡	-0.221 0.825‡	0.048 0.962‡	0.138 0.139¶	-0.160 0.085¶
Language total	0.035 0.705¶	0.101 0.277¶	-0.101 0.278¶	1.470 0.142‡	2.404 0.016‡	0.120 0.904‡	0.229 0.819‡	0.164 0.077¶	-0.136 0.144¶
Cognitive complaint	1.456 0.156‡	0.469 0.639‡	-0.791 0.429‡	0.137 0.711†	0.985 0.482†	0.031 0.860†	0.0003 0.986†	-1.650 0.099	-3.200 0.001
MCI	-1.368 0.171‡	1.214 0.225‡	-2.034 0.042‡	2.168 0.141†	1.577 0.283†	5.889 0.015†	0.515 0.218†	-0.140 0.889	-2.317 0.021
TOTAL number NMS	-0.019 0.833¶	0.018 0.842¶	0.026 0.777¶	-1.48 0.140‡	1.593 0.111‡	0.911 0.364§	0.238 0.812‡	0.225 0.012¶	0.482 1.41E-08¶

Appendix D: Publication

The role of longitudinal cohort studies in epigenetic epidemiology: challenges and opportunities
Jane WY Ng, Laura M Barrett, Andrew Wong, Diana Kuh, George Davey Smith, Caroline L Relton.
Genome Biology, 2012, 13(6), pp246-58.

Genome **Biology**

REVIEW

The role of longitudinal cohort studies in epigenetic epidemiology: challenges and opportunities

Jane WY Ng[1,2], Laura M Barrett[1], Andrew Wong[3], Diana Kuh[1], George Davey Smith[4] and Caroline L Relton[1]*

Abstract

Longitudinal cohort studies are ideal for investigating how epigenetic patterns change over time and relate to changing exposure patterns and the development of disease. We highlight the challenges and opportunities in this approach.

Keywords Epigenomics, DNA methylation, life course, longitudinal studies

Introduction

Interest in the role of epigenetic processes in common complex diseases continues to increase [1,2]. Epigenetics is a potentially major mechanism by which environmental factors can affect physiological function and disease risk. Research into epigenetics promises to reveal many of the causes that remain undiscovered after extensive investigation of common genetic variation [3].

Epidemiological approaches can be used to identify whether epigenetic processes are involved in mediating the association between risk factors (environmental, genetic, lifestyle, socioeconomic and so on) and common complex disease [4,5]. For example, longitudinal cohort studies have been a cornerstone of observational epidemiology for many years. Long-term follow-up of adult cohorts has identified important risk factors for cardiovascular disease, chronic bronchitis, and cancers, and follow-up of cohorts from birth or childhood has been equally successful at identifying the importance of early exposures (especially the childhood social environment) and developmental characteristics for adult health (for example, [6-10]). Longitudinal studies, particularly those that start in early life, can contribute to our understanding of how the epigenome changes over time, as a result of varying environmental exposures, and how disease phenotypes evolve. Longitudinal studies are costly to instigate and maintain, and cross-sectional studies (a less expensive alternative study design) have more often been used to assess the relationship between exposures and the epigenome and/or the epigenome and disease. However, cross-sectional studies cannot capture the dynamic nature of epigenetic mechanisms [11], making it difficult to identify the influences of the environment and/or disease state (or sub-clinical features of disease) on the epigenome and thus establish the direction of causality. As a result of this, study designs that make use of multiple time points are being increasingly recognized as the most suitable to analyze the epigenetics of common complex diseases. Because longitudinal studies track the same cohort at multiple time points throughout their lifetime, enabling the temporal relationship between exposure and disease to be established, they are ideally placed for exploitation in epigenetic investigations.

Advances in genomic technologies have opened up the possibility of large-scale population-based assessment of epigenetic patterns to help understand their influence on disease. How should such studies be conducted to maximize their impact and what can epigenetics researchers learn from previous approaches to population-based studies? Here we focus on how epidemiological approaches, including the design of cohort studies, can help investigate the role of epigenetic variation in common complex disease. Furthermore, the dynamic nature of epigenetic patterns means that they can be altered by disease-related factors (a process called ‘reverse causation’) as well as a host of confounding factors (such as age, sex, socioeconomic position, diet, or smoking). Many relevant approaches have been developed in the context of both genetic and life course epidemiology that could be fruitfully applied to epigenetics; examples are methods for dealing with biases, confounding, and reverse causation and also longitudinal statistical modeling techniques [12,13]. We first assess what epigenetic markers have been measured within existing life course studies before

*Correspondence: caroline.relton@ncl.ac.uk
[1] Institute of Genetic Medicine, Newcastle University, NE1 3BZ, UK
Full list of author information is available at the end of the article

discussing how the epidemiologist's toolkit can be applied to epigenomics.

Epigenetic studies within longitudinal cohorts

Since 2010, 34 life course studies have included measurements of DNA methylation, and just four of these have included analysis of epigenetic features at more than one time point (Table 1). In line with the vast majority of other epigenetic studies, the focus is on DNA methylation as this is the most straightforward form of epigenetic modification to measure, and the only currently feasible option in archived DNA samples. Prospective sample collection will permit the analysis of chromatin modifications and microRNA. Three of the studies analyzing more than one time point (Table 1) report findings relating specifically to age-related changes in childhood [14] or adulthood [15,16], and all three focus on gene-specific DNA methylation of a small panel of (different) loci and report differences that were modest in size (generally <5%). A further study considers changes in DNA methylation over a relatively short time period (28 to 180 days) in relation to air pollution exposure [17]. Although there was some indication of lower global DNA methylation in repetitive elements across the genome in this study [17] at 90 days of exposure, there was no evidence of a dose response, casting doubt on the biological importance of this association. In summary, very little has been done in this area.

Table 2 summarizes additional examples in which case-control studies of DNA methylation have been nested within existing large-scale longitudinal cohorts; this approach has been applied so far exclusively in the context of cancer. Analyses in this instance have been limited to gene panels (generally established tumor suppressor or oncogenes) and have been undertaken either (i) to assess the utility of epigenetic signatures as early biomarkers of cancer risk [18-20] or (ii) to consider the determinants of a perturbed methylation state (methylator phenotype), which has been implicated in numerous cancers [21-25]. With improved knowledge of methylation variable regions associated with diseases other than cancer (for example, cardiovascular disease, dementia, and rheumatoid arthritis), the same approach could be adopted in the context of longitudinal cohort studies.

The paucity of DNA methylation measurements undertaken in cohorts that have collected serial samples from the same individuals is clear, indicating that the potential richness of longitudinal data and sampling in these studies has yet to be fully exploited. Few studies have routinely collected serial samples from the same individuals at multiple points in the life course (for example, the Avon Longitudinal study of Parents and Children (ALSPAC) [26,27], and the Normative Aging Study [17,28-32]), but others are planning serial sampling in light of the interest in epigenetics (such as the Medical Research Council National Survey of Health and Development [33] and the Southall And Brent REvisited (SABRE) cohort [34]). Given the temporal variation in epigenetic patterns, serial sampling of any longitudinal cohort would be advised where possible.

Of the studies published so far, the variety of tissues analyzed is limited mainly to easily accessible peripheral blood, cord blood or buccal cells, the studies are modest in size compared with those used for genetic research, and the range of different methods that have been used to quantify DNA methylation have led to an overall lack of comparability between studies. It is clear from these observations that more can be done with respect to the collection and analysis of biological samples from longitudinal cohorts so that they are optimal for epigenetic studies.

Attributes of longitudinal cohort studies

Ideally, longitudinal epigenetic studies should include extensive, prospectively collected data and biological samples at multiple time points across the life course. Many existing longitudinal cohort studies are population-based, although some focus on a specific sub-group of the general population. For example, the SABRE cohort focuses on groups that are first or second generation migrants to the UK of non-European ethnicity to examine particular health issues, in this case the marked discordance in disease risk observed in migrant groups compared with Europeans living in the UK [34]. Longitudinal epigenetic studies can add value to existing resources, such as data from genome-wide association studies - for example, ALSPAC [26,27] and the Relationship between Insulin Sensitivity and Cardiovascular disease (RISC) cohort [35]. Exposures commonly captured in longitudinal studies include lifestyle factors, such as smoking, alcohol intake, diet, and physical activity patterns, and also socioeconomic measures across the life course. Common phenotypes on which longitudinal studies tend to focus include physical and anthropometric measures, cognitive, cardiovascular, metabolic, respiratory, and musculoskeletal function, and a range of blood-based intermediate biomarkers. Of particular value are birth cohorts with trans-generational and across-life samples from birth onwards, allowing an appraisal of epigenetic changes associated with *in utero* and early life exposures, a period when the epigenome is believed to be particularly plastic.

The epidemiological toolkit

Applying principles of life course epidemiology to epigenetic research

Research in life course epidemiology investigates developmental, aging, and risk factor trajectories and how

Table 1. Epigenetic studies in longitudinal cohorts: a summary of recent literature (2010 to 2012)

Cohort	DNA analysis time points	Tissue	Form of DNA methylation analysis	Loci	Exposure (if applicable)	Outcome (if applicable)	n	References
Avon Longitudinal Study of Parents and Children	1	Cord blood	Proxy genotype	*TACSTD2*	Postnatal growth	Childhood adiposity	6,990	[53]
Avon Longitudinal Study of Parents and Children	1	Cord blood	Illumina GoldenGate Cancer Panel	1,576 CpG sites		Childhood body composition	178	[56]
1958 British Birth Cohort Study	1	Peripheral blood	MeDIP-chip	Genome-wide methylation	Socioeconomic position		40	[92]
Columbia Children's Center for Environmental Health Northern Manhattan Mothers and Newborns Study	1	Cord blood	Methylamp	Global methylation	Polycyclic aromatic hydrocarbons, benzo[*a*] pyrene-DNA adducts		164	[93]
Detroit Neighborhood Health Study	1	Peripheral blood	Illumina HM27 BeadChip	*SLC6A4*	Traumatic events	Post-traumatic stress disorder	100	[94]
Detroit Neighborhood Health Study	1	Peripheral blood	Illumina HM27 BeadChip	Genome-wide methylation	Traumatic events	Post-traumatic stress disorder	100	[95]
Detroit Neighborhood Health Study	1	Peripheral blood	Illumina HM27 BeadChip	Genome-wide methylation		Depression	100	[96]
Detroit Neighborhood Health Study	1	Peripheral blood	Illumina HM27 BeadChip	Genome-wide methylation	Post-traumatic stress disorder		100	[97]
Dutch Famine Cohort	1	Peripheral blood	Pyrosequencing MethyLight	*Sat2*, LINE-1, *LUMA*	Pre-natal famine		947	[98]
Environmental Risk (E-Risk) Longitudinal Twin Study	2	Buccal cells	Sequenom MassArray	*DRD4*, *SERT*, *MAOA*	Change over time		182	[14]
Epigenetic Birth Cohort	1	Cord blood and placenta	Pyrosequencing	LINE-1	Gestational age and birth weight		319 mother-child dyads	[99]
Longitudinal Study of Adolescent Health	1	Buccal cells	Sequenom MassArray	*5HTT*		Depression	150	[100]
Longitudinal Study of Child Development	1	Buccal cells	Microarray	Genome-wide methylation	Childhood adversity		109	[101]
Lovelace Smokers' Cohort	1	Sputum	Methylation specific PCR	Lung cancer genes	Wood smoke exposure	Chronic obstructive pulmonary disease	1,827	[102]
Netherlands Twin Registry	2	Peripheral blood and buccal cells	Sequenom MassArray	*IL10*, *NR3C1*, *TNF*, *IGF2R*, *GRB10*, *LEP*, *CRH*, *ABCA1*, *IGF2*, *INSIGGF*, *KCNQ1OT1*, *MEG3*, *APOC1*, *GNASAS*, *GNAS A/B*	Cell counts, change over time		64	[15]
Newcastle Preterm Birth Growth Study	1	Peripheral blood	Pyrosequencing	*TACSTD2*	Postnatal growth	Childhood adiposity	121	[53]
New York Women's Birth Cohort	1	Peripheral blood	Pyrosequencing	*Sat2*, Alu, LINE-1	Prenatal tobacco smoke		90	[103]

Continued overleaf

Table 1. Continued

Cohort	DNA analysis time points	Tissue	Form of DNA methylation analysis	Loci	Exposure (if applicable)	Outcome (if applicable)	n	References
Normative Aging Study	1	Peripheral blood	Pyrosequencing	*CRAT*, *F3*, *GCR*, *ICAM*, *IFNγ*, *IL6*, *iNOS*, *OGG1*, *TLR2*		Lung function	756	[28]
Normative Aging Study	1	Peripheral blood	Pyrosequencing	Alu, LINE-1, *F3*, *TLR2*, *ICAM-1*	Air pollution	Fibrinogen, ICAM-1, VCAM-1, and CRP	704	[29]
Normative Aging Study	1-5	Peripheral blood	Pyrosequencing	*CRAT*, *F3*, *GCR*, *ICAM*, *IFNγ*, *IL6*, *iNOS*, *OGG1*, *TLR2*	Age		784	[16]
Normative Aging Study	1	Peripheral blood	Pyrosequencing	Alu, LINE-1		Cancer	722	[30]
Normative Aging Study	1-3	Peripheral blood	Pyrosequencing	Alu, LINE-1	Air pollution		706	[17]
Normative Aging Study	1	Peripheral blood	Pyrosequencing	LINE-1		Ischemic heart disease and stroke	712	[31]
Normative Aging Study	1	Peripheral blood	Pyrosequencing	LINE-1		Inflammatory markers VCAM-1, ICAM-1 and CRP	593	[32]
North Cumbria Community Genetics Project	1	Cord blood and maternal peripheral blood	Pyrosequencing	*IGF2*, *IGFBP3*, *ZNT5*, *LUMA*	Maternal characteristics, folate metabolism and genotype		430	[104]
Project Viva	1	Cord blood and maternal peripheral blood	Pyrosequencing	LINE-1	Methyl donor nutrients		516 infants and 830 mothers	[105]
Bangladesh Birth Cohort	1	Cord blood and maternal peripheral blood	Pyrosequencing	Alu,LINE-1, *p53*, *p16*	Arsenic exposure		120 mother-child pairs	[106]
Bangladesh Birth Cohort	1	Cord blood and maternal peripheral blood	Pyrosequencing	Alu, LINE-1	Arsenic exposure		114 mother-child pairs	[107]
Psychological, Social and Behavioral Determinants of Ill Health	1	Peripheral blood	Methylamp	Global methylation	Socioeconomic position		239	[108]
Rhode Island Child Health Study	1	Placenta	Pyrosequencing	*HSD11B2*	Maternal characteristics	Neurobehavioral outcomes	185	[109]
Singapore Chinese Health Study	1	Peripheral blood	MethyLight	Alu, *Sat2*	B vitamins	Cardiovascular disease	286	[110]
Southampton Women's Study	1	Umbilical cord	Sequenom MassArray	*eNOS*		Bone mineral content	66	[111]
Southampton Women's Study	1	Umbilical cord	Sequenom MassArray	*RXRA*, *eNOS*, *SOD1*, *PIK3CD*, *IL-8*		Childhood adiposity	66 + 239	[57]
Five studies	1	Peripheral blood	Pyrosequencing	Alu, LINE-1	Age, gender, smoking, alcohol, body mass index		1,465	[67]

Search strategy: a literature search retrieved >350 publications published between 1 January 2010 and 13 April 2012. Only established longitudinal cohort studies were included in the summary table (in as far as this could be ascertained from the information available). Of more than 350 publications retrieved for review only 34 included analyses of DNA methylation in established longitudinal cohort studies and of these only four included analyses at more than one time point. The remaining 30 studies measured DNA methylation at a single time point and considered this with respect to longitudinal data relating to exposures or outcomes.

Table 2. Nested case-control epigenetic studies in longitudinal cohorts: a summary of recent literature (2010 to 2012)

Cohort	DNA analysis time points	Tissue	Form of DNA methylation analysis	Loci	Exposure (if applicable)	Outcome (if applicable)	n	References
EPIC & Breakthrough Generations Study & KConFab	1	Peripheral blood	Pyrosequencing	*ATM*, LINE-1		Breast cancer	1,381	[112]
EPIC-Lung	1	Peripheral blood	Pyrosequencing	*CDKN2A RASSF1A*, *GSTP1*, *MTHFR*, *MGMT*	B vitamins, smoking	Lung cancer	93	[18]
EPIC-Norfolk	1	Tumor tissue	Pyrosequencing	*MLH1*		Colorectal cancer	185	[21]
EPIC-EURGAST	1	Tumor tissue	Pyrosequencing	*CHRNA3*, *DOK1*, *MGMT*, *RASSF1A*, *p14ARF*, *CDH1*, *MLH1*, *ALDH2*, *GNMT*, *MTHFR*		Gastric cancer	162	[22]
Iowa Women's Health Study	1	Tumor tissue	MethyLight	CpG island methylator phenotype	Smoking	Colorectal cancer	555	[23]
Netherlands Cohort Study	1	Tumor tissue	Methylation specific PCR	*CACNA1G*, *IGF2*, *NEUROG1*, *RUNX3*, *SOCS1*	Body size and physical activity	Colorectal cancer	734	[113]
New York University Women's Health Study	1	Peripheral blood	Methylation specific PCR	*RASSF1A*, *GSTP1*, *APC*, *RARβ2*		Breast cancer	200	[19]
Northern Sweden Health and Disease Study	1	Tumor tissue	MethyLight	CpG island methylator phenotype	B vitamins	Colorectal cancer	570	[24]
Nurse's Health Study	1	Tumor tissue	MethyLight	*CACNA1G*, *CDKN2A*, *CRABP1*, *IGF2*, *MLH1*, *NEUROG1*, *RUNX3*, *SOCS1CHFR*, *HIC1*, *IGFBP3*, *MGMT*, *MINT-1*, *MINT-31*, *p14*, *WRN*	B vitamins and alcohol	Colorectal cancer	761	[25]
Shanghai Women's Health Study	1	Peripheral blood	Pyrosequencing	Alu, LINE-1		Gastric cancer	576	[20]

Search strategy: a literature search retrieved >350 publications published between 1 January 2010 and 13 April 2012. Nested case-control studies within established longitudinal cohort studies are included. There are many further examples in which DNA methylation has been analyzed in a case-control study design, cross-sectional study or randomized controlled trial that includes some element of data collection over time (often retrospective), but these have not been included in the tables presented.

dynamic relationships unfold over time, and takes into account potential confounding, mediating, or interactive effects of lifetime biological, psychological, and social risk factors [36]. This conceptual framework is relevant for epigeneticists investigating long-term associations that may be biased, confounded or due to reverse causation. Life course epidemiologists have investigated various different methods for modeling risk factor trajectories (particularly growth trajectories) in relation to later health outcomes and have developed a novel structured approach [37] to distinguish critical, sensitive, and accumulation life course models [38]. They use a range of approaches for modeling repeat continuous and binary outcome measures, such generalized estimating equations or mixed models that consider correlated data such as repeat measures from the same individuals over time, and for modeling time to an event, such as survival and event history analysis. This toolkit is relevant to epigeneticists, whether studying lifetime environmental exposures that promote particular epigenetic signatures over time or how these signatures themselves may affect not just the level (intercept) of function (such as blood pressure) at a point in time but also its rate of change (slope) over time. Such statistical approaches have not been widely applied to epigenetic data, although examples can be found in Madrigano *et al.* [16,17], who illustrate the use of mixed models to analyze changes in methylation over time while accounting for the correlation among measurements within the same individual. Further discussion of this subject is provided below in the section on data analysis considerations.

Several research collaborations involving cohort studies, such as HALCyon (Healthy Aging across the Life Course) [39], FALCon (Function Across the Life Course) [40] and GEoCoDE (Genomic and Epigenomic Complex Disease Epidemiology) [41] have been formed. These have increased the sample size and power to investigate lifetime risk factors on longitudinal phenotypes and to test whether findings are replicated across cohorts in a systematic way, and they will be useful to epigenetics research. The collaborations have developed experience in data harmonization to derive comparable phenotypes across the cohorts, and in cross-cohort methods (for example, [42]). Those running epigenetic studies may want to make use of these collaborations for similar reasons, and a coordinated approach is likely to advance the science and be appealing to funders. Coordinating the cohorts has led to more effective ways of gaining knowledge of the various datasets and metadata as well as facilitating data sharing and encouraging good practice in data management.

From genetic to epigenetic epidemiology

Incorporating epigenetic measures into epidemiological studies is often done in the context of genetic epidemiology resources. However, studying epigenetic factors - which are, partly at least, phenotypic - is more similar to conventional epidemiology than it is to genetic epidemiology. Several aspects of germline genetic variation lead to special-case conditions that allow relaxation of usual epidemiological principles: reverse causation (disease influencing the variable being measured rather than *vice versa*) is clearly not an issue in genetic epidemiology, and confounding - which often vitiates conventional epidemiology - generally relates only to ancestry in genetic epidemiology [43], and this can be accounted for by using principal components from genome-wide data as control variables. Germline genetic variation can be assessed on samples taken at any stage of life, does not change over time, and can be assayed with high precision and low measurement error. Effect sizes for the influence of common genetic variants on common complex diseases tend to be small, which means that very large sample sizes are required. Given these circumstances, the genetic epidemiology study design of choice became large case-control studies, with the controls not being carefully selected to represent the source population - and sometimes (as in the case of the landmark Wellcome Trust Case Control Consortium (WTCCC) [44]) control groups shared for comparison with several disease groups. For example, in the WTCCC the common control groups consisted of blood donors (who are very unrepresentative in terms of factors that would be important confounders in conventional epidemiological studies, such as health-related behaviors and social class) and participants in the 1958 birth cohort - all of the same age, which in some cases barely overlapped with the age of the cases.

However, such study designs are not appropriate for epigenetic epidemiology, as confounding, bias, and reverse causation are all serious problems when studying phenotypic exposures. It is important that the successes of genetic epidemiology are not translated into failures for epigenetic epidemiology [1,5,45]. Prospective studies are the ideal type of study, including documented exposure (epigenetic) measures collected before the outcomes and temporal changes, detailed assessment of confounding factors, and consideration of measurement error. Currently, the effect sizes of associations in epigenetic studies are poorly delineated, but it is likely that, unlike the situation in the early days of molecular genetic epidemiology, the problem will not be one of relatively few robust associations, but rather many real observational associations will exist and the issue will be the separation of causal associations from those generated by confounding and bias. Various methods that have been developed to strengthen causality in conventional epidemiology - including collaborative analysis of multiple cohorts in which confounding structures differ [46], comparisons of plausible and implausible associations

[47,48], and the use of instrumental variables [47] - can be applied to epigenetic epidemiology studies.

An instrumental variables method that uses germline genetic variants as the instruments - Mendelian randomization - is increasingly used to strengthen causality with respect to environmentally modifiable exposures for which genetic variants can serve as proxy measures [49-51]. Mendelian randomization can be extended to the investigation of epigenetic profiles as the potentially modifiable exposure. This method - 'two step epigenetic Mendelian randomization' - is currently under development, and details can be found elsewhere [5,52].

A further complexity of epigenetic studies is the tissue-specific nature of epigenetic patterns. Given that they are integrally involved in the process of cell and tissue differentiation, it is no surprise that epigenetic patterns differ between tissue sources. Genetic comparisons within and between studies can be made using a variety of sources of DNA to generate genotype data; however, this is not the case in an epigenetic context. Population-based studies often have to rely on easily accessible DNA sources (such as blood, saliva, buccal cells; Table 1). These serve as a surrogate for the target tissue involved in the disease of interest, but there is inevitable heterogeneity in both specific cell type represented and sample processing, which may bias epigenetic measurement (see the section below on data analysis considerations). Despite these limitations, epigenetic epidemiological studies are emerging and include strategies such as Mendelian randomization approaches [53] or inter-tissue comparisons [15] to interrogate the functional relevance and casual nature of observations.

Inter-generational epigenetic studies

Family-based sampling of both siblings and multiple generations can have particular value in epigenetic studies. The fact that epigenetic states are often established in early (in particular antenatal) development makes birth cohorts with recruitment and sample collection from pregnant women and sample collection on offspring from birth onwards of particular value [26,27]. There is considerable interest in the role of epigenetic mechanism in the developmental origins of adult disease, to which longitudinal cohort studies are making a valuable contribution [4,53-59].

Data analysis considerations

Most research undertaking longitudinal analysis of molecular biomarker data assumes that there are predictable biological changes over time associated with a given exposure or disease process. However, in the context of epigenetic studies, change over time can be due to technical [60] or genetic factors [61], tissue type [62,63], changes with normal aging, and stochastic changes [64]. These sources of data 'noise' threaten the detection of the biological signal of interest. Thus, as is often the case, the first and most critical step to performing longitudinal DNA methylation analysis is careful study design and data collection with meticulous recording of technical factors and factors that vary between people. Given that data collection may occur months, years or even decades apart, the awareness and/or control of such sources of variability are paramount to making valid conclusions regarding within-individual changes over time as it may be impossible to account for these factors after the fact. Pre-processing of data is often necessary to generate comparable data from samples between and within individuals over time. International initiatives to address and reach consensus on such issues are in progress [65]. Equally important is that many of these methods seek to optimize the signal-to-noise ratio. These two considerations are critical to generating valid and reproducible results. Prudent use of pre-processing that matches the study design and data, and experimentation with several different methods are strongly encouraged. In addition, the threat of time-varying artifacts masquerading as biological signal is constantly present in longitudinal studies. This possibility should be formally tested as an automatic addition to the primary study hypothesis.

An example of a 'noise' source that is just beginning to be understood is the role of genetic factors in determining the degree of variability in DNA methylation over time. This is suggested by familial clustering of DNA methylation variability over time [61]. From the perspective of individual loci, there is also evidence of CpG site-dependent differential stability [15]. This indicates that loci should be carefully selected that demonstrate greater inter- than intra-individual variation over time. The mechanisms underlying this are unknown but could reasonably be related to overlying genetic architecture (for example, interaction with other epigenetic marks and possibly even the DNA itself) or the cellular milieu, as suggested by tissue-specific difference in stability in the same loci [63]. With the success of next-generation sequencing and its falling costs, we can look forward to a clearer view of the effect of genetic factors on DNA methylation and time-dependent variability.

As alluded to earlier, the vast majority of longitudinal cohort studies that are in a position to consider including epigenetic assessment have used biological specimens collected from peripheral blood. Reliance on leukocyte DNA extracted from peripheral blood introduces a potential source of measurement error [66]. Given the labile nature of leukocyte subtype populations over time, this variation may make an important contribution to intra-individual changes in DNA methylation. For instance, shifts in leukocyte populations can occur as a result of normal development and aging, inflammation from infectious, rheumatological, or oncological diseases, or normal response to medications (such as non-steroidal anti-inflammatory drugs). The most definitive solution is to isolate cell types (for example, through magnetic-activated or fluorescence-activated cell sorting), so as to perform comparisons within relatively homogenous leukocyte populations. However, this is possible only with freshly collected samples; one of the advantages of prospective longitudinal studies is the potential to collect appropriate samples relevant for epigenetic studies.

When analysis of relatively homogeneous cell types was unavailable, Zhu and colleagues [67] used total and differential leukocyte count (from a sample drawn concurrent with the methylation sample) to control for this variation in regression models. These researchers found that the proportion of leukocyte cell types correlated with levels of LINE-1 methylation. Importantly though, statistical adjustment for this did not alter the association between LINE-1 and Alu methylation levels and individual characteristics (age, gender, smoking habits, alcohol intake, and body mass index). Candidate gene studies of methylation have reached similar conclusions [15,16]. This could mean that leukocyte populations contribute a negligible amount of variance relative to the specified model factors. Alternatively, it may be that controlling for leukocyte population in this manner inadequately captures the effect of this noise. The possibility that using the direct measure of an unwanted variable in a regression equation may sub-optimally reduce noise was explored by Teschendorff and colleagues [60]. Using Illumina HumanMethylation27 BeadChip data, they proposed a variation of surrogate variable analysis in which confounders are modeled as statistically independent components. Using these components instead of the original measures in regression analysis, they found a stronger association between methylation of Polycomb-family gene loci and their phenotype of interest, age. From this, they concluded that the effect of confounders on the DNA methylation data was better represented by independent components than the original covariates.

Lastly, in cases where no information on cell counts is available, a potential solution may arise from the DNA methylation data itself. Such a possibility is presented by Houseman and colleagues through their software methylSpectrum [68]. The authors propose an algorithm to infer the contribution of different leukocyte sub-populations to whole blood DNA methylation patterns. This software is not designed to examine changes over time and requires a suitable reference sample from which to make inferences, which would reasonably require multiple age-appropriate references in a longitudinal study setting.

In summary, we need formal comparisons of these methods in heterogeneous and homogeneous samples from the same specimen. International efforts to create reference epigenomes from homogeneous cell samples will be highly beneficial [65]. However, variation due to cellular and tissue heterogeneity is just one example of the wide breadth of issues regarding noise that require detailed and systematic study.

Modeling epigenetic change over time

There are several issues that need to be considered when analyzing epigenetic change over time, such as the unit of DNA methylation change under examination (Box 1) and the analytic technique. The unit of analysis must consider several issues. For example, how is DNA methylation measured? What is the question under investigation? Is the research focused on testing site-specific changes in DNA methylation related to exposures and/or outcomes or is it seeking to explore a network of gene regulation? What type of *a priori* information is available? How does this information contribute to understanding of error or covariance of methylation measurements? Are individuals compared using categorical or continuous variables?

Guided by the selected unit of DNA methylation change, we now turn to examples of modeling intra-individual variation over time that is due to disease and/or environmental factors. The selection of an appropriate modeling technique has important implications for study power and calculations of statistical significance. We limit this discussion to longitudinal studies with three or more time points, as two time points can at most infer a difference rather than the nature of change. Much of this work is borrowed from other fields, particularly gene expression studies, and uses data-driven or knowledge-driven techniques, or combinations of both.

Several techniques use comparisons between two groups (such as controls versus cases) to determine differential time courses [69,70]. Some of these methods can be extended to comparisons between more than two groups (for example, [71]). An alternative to this individual-based approach is to find time course patterns that distinguish one group of individuals from another (for example, [72,73]). Methods that capitalize on other biological knowledge (such as genomes, transcriptomes, or nucleosomes) may allow us to better infer the nature of methylation in the context of how functional regulation of the genome relates to exposures or disease processes. This is especially powerful to detect signals that are expected to be subtle but consistent among jointly regulated loci [74]. An example is longitudinal gene set analysis [75] using annotations from databases such as Gene Ontology. The parallel analysis of different sources of high-throughput data has so far only been explored in cross-sectional methylation studies but could in theory be applied to longitudinal analysis. However, such longitudinal analysis will require advanced multi-dimensional

Box 1: Potential units of change to examine epigenetic mechanisms

- A single gene or gene region of interest
- Single gene loci that have different temporal patterns between biological groups
- A family of genes of known biological or clinical importance (such as those previously known to show exposure-related differential methylation)
- A group of functionally related genes (for example, as identified by Gene Ontology or Kyoto Encyclopedia of Genes and Genomes (KEGG) terms)
- A network of co-regulated genes (for example, using intersection with concurrent gene expression data or from previous literature)
- Genes related by their linear proximity on the DNA strand (such as regional grouping, as done to examine differential methylation between and within individuals [70])
- Genes related to the overlying chromatin architecture (such as knowledge of nucleosome position or histone modifications)
- Genes that show similar patterns of change (for example, gene curve [71])

techniques (Box 2). These techniques require pre-processed data that are relatively free of noise. Another approach may use data reduction techniques to extract meaningful features from data noise while simultaneously considering the time-varying nature of DNA methylation. For example, group-independent component analysis with temporal concatenation of microarray data would assume that there are common sites of epigenetic activity but that the course of change may be different for each individual. Most experience in this type of technique comes from the analysis of neuroimaging data, where the goal is to uncover areas of the brain that are activated similarly among individuals in an experimental group over time [76]. The translation of such ideas to molecular data, which often have far lower temporal resolution but higher 'spatial' resolution (gene loci as opposed to areas of the brain), would be a challenging but also potentially promising avenue.

The promise of epigenetic studies of longitudinal cohorts

Future longitudinal epigenetic studies will undoubtedly integrate greater levels of genomic, biologic and/or phenomic information. For example, our expanding knowledge of factors influencing chromatin architecture may soon allow the analysis of methylation marks within context of the broader chromatin state. Examples of such data are nucleosome mapping [77], histone modifications [78], and chromosome conformation capture [79]. The influence of the underlying and overlying chromatin architecture (interaction with protein, RNA, and DNA primary and secondary 'structure' [80]) on differential locus stability over time remains to be elucidated. Analysis of DNA methylation is clearly only scratching the surface of the epigenetic information that regulates gene expression, but longitudinal cohort studies provide a tractable opportunity to contribute to our knowledge base in this area and, as our understanding of the wider epigenome improves, additional epigenetic features may also be added to such studies.

Increasingly, studies are pushing to provide a broader mechanistic picture of cellular function and regulation by juxtaposing data from two or more kinds of high-throughput data [81,82]. So far, these data are often extracted from different materials or individuals (such as DNA methylation from whole blood and RNA from cell culture). This limits interpretation of functional relevance. However, advances in biotechnology that reduce the amount of specimen required and increase automation, in conjunction with falling costs, are likely to overcome this problem. Biobanked samples, such as plasma, DNA, and RNA from longitudinal cohorts, could make a valuable contribution to developments in this area. Furthermore, the development of nested recall studies for intensive phenotyping within established cohorts will greatly enhance research opportunities in this area.

As multi-dimensional datasets evolve and the ability to mine the information within them improves, it will be imperative that this information is made as accessible as possible to the wider scientific community. Although it is currently possible to access some information relating epigenetic data to common genetic variation and gene expression, providing an integrative approach, this is not available at multiple time points. Longitudinal studies can offer considerable added value in these settings and profiling using a comprehensive range of high-throughput methods can be overlaid on a wealth of exposure and phenotypic data, allowing researchers to explore specific hypotheses *in silico* and thus helping to prioritize resources for more detailed investigations.

In summary, longitudinal cohorts can offer a great deal in the context of epigenetic epidemiology, including identification of the major determinants of epigenetic variation in populations and a better understanding of the relationship between genetic and epigenetic variation. They provide an unprecedented opportunity to increase our understanding of the dynamic nature of epigenetic patterns and how changes occur in response to a wide range of environmental, lifestyle, and behavioral factors. Population-based studies will improve our knowledge of the extent and topography of inter-individual variation in epigenetic patterns and permit assessment of effect sizes of shifts in epigenetic patterns on health-related

Box 2: Longitudinal modeling strategies for high-dimensional data

Many techniques determine differential time courses based on comparison of two groups of variables (for example, [69,70,84-86]). When there are more than two groups, Yuan and colleagues [71] have demonstrated the utility of their method using hidden Markov models. Multi-group comparisons are also possible; Yuan and colleagues have demonstrated the utility of hidden Markov models to classify genes based upon their temporal expression patterns, which, rather than ignoring, takes advantage of the information contained in time course data. If no groups are present, an alternative is to group genes that show similar temporal patterns (for example, [72]). Another approach is to group genes using *a priori* knowledge of biological similarities and reduce the amount of multiple comparisons. Using Gene Ontology annotation to group 'functionally' related genes, Zhang *et al.* [75] developed a non-parametric longitudinal gene set analysis of gene expression data to detect time-exposure interaction effects. This method is suitable for unbalanced data with missing time points. It is also appropriate for heteroscedastic variance (where variance is uneven across a given data distribution) and non-normal data distributions.

Another consideration is the anticipated type of time course. If a cyclical pattern is expected - for instance, in the study of circadian rhythms or cell cycles - Li *et al.* [73] propose functional clustering using an autoregressive moving-average process. If the goal is to identify groups of co-expressed genes showing gradual changes over time that may be linked to disease progression, Qiu *et al.* [87] have developed a method to study gene expression in cancer tissue at various stages of malignant transformation, which may be applicable to epigenetic data.

Units that consider genes as groups or networks may require a transition from viewing DNA methylation data as a two-dimensional entity (such as disease group and time) to a three-dimensional one (such as disease group, gene locus and time), or even data 'blocks' with greater dimensions. The family of matrix and tensor decompositions (such as independent component analysis, canonical correlation analysis, non-negative tensor factorization, and canonical-decomposition/parallel factor analysis) used in areas such as psychometrics and chemometrics have been proposed as powerful representations of biological multi-dimensional data [88,89]. Translation of such methods to DNA methylation is sure to follow.

Although having multiple time points is advantageous for several reasons, a complication is that similar patterns of change in any group of people can start at different times (such as onset of puberty). This may obscure detection of meaningful but overlapping patterns. This can be unraveled using methods that account for lag between individuals, such as by using parallel factor analysis-related models [90] or spline-based models [91].

outcomes. A wealth of statistical approaches can be borrowed and adapted from related fields and be applied to longitudinal epigenetic analysis - an area of biostatistics that is likely to grow exponentially as high-throughput datasets become increasingly multi-dimensional. Insights into the temporal relationship between changes in epigenetic patterns and functional and health-related outcomes that can be gleaned from longitudinal studies will assist in defining causality. This, and other epidemiological methods to strengthen causal inference, will contribute to the identification of predictive epigenetic biomarkers and modifiable targets for intervention.

The ultimate goal of observational data generated in epidemiological investigations is to feed forward into clinical practice or public health. There is already evidence of translation of longitudinal biological data to clinical applications [83]. The incorporation of epigenetic biomarkers to enhance clinical tools for prediction and prognosis is beginning to emerge [5] (Table 2), and longitudinal cohorts will undoubtedly help in this domain.

Abbreviations

ALSPAC, Avon Longitudinal Study of Parents and Children; SABRE, Southall And Brent REvisited; WTCCC, Wellcome Trust Case Control Consortium.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

All authors contributed to the preparation of the manuscript.

Acknowledgements

JN receives funding support from the Clinical Investigator Program at University of British Columbia and GEoCoDE (EUFP7); LB is an MRC-funded PhD student; AW and DK are supported by the MRC; GDS is funded from multiple sources, including the MRC, Wellcome Trust and NIH; CR is supported by funding from multiple sources, including BBSRC, MRC, EUFP7, NIHR and the Wellcome Trust. This review was not supported from any single source.

Author details

[1]Institute of Genetic Medicine, Newcastle University, NE1 3BZ, UK. [2]Clinician Investigator Program, University of British Columbia, Vancouver, BC V6Z 1Y6, Canada. [3]MRC Unit for Lifelong Health & Aging, London, WC1B 5JU, UK. [4]MRC Centre for Causal Analyses in Translational Epidemiology (CAiTE), School of Social and Community Medicine, University of Bristol, Bristol BS8 2BN, UK.

Published: 29 June 2012

References

1. Relton CL, Davey Smith G: **Is epidemiology ready for epigenetics?** *Int J Epidemiol* 2012, **41:**5-9.
2. Davey Smith G: **Epigenetics for the masses: more than Audrey Hepburn and yellow mice?** *Int J Epidemiol* 2012, **41:**303-308.
3. Maher B: **Personal genomes: The case of the missing heritability.** *Nature* 2008, **456:**18-21.
4. Michels KB: **The promises and challenges of epigenetic epidemiology.** *Exp Gerontol* 2010, **45:**297-301.
5. Relton CL, Davey Smith G: **Epigenetic epidemiology of common complex disease: prospects for prediction, prevention, and treatment.** *PLoS Med* 2010, **7:**e1000356.
6. Pearson H: **Children of the 90s: coming of age.** *Nature* 2012, **484:**155-158.
7. Moayyeri A, Hammond CJ, Valdes AM, Spector TD: **Cohort Profile: TwinsUK and Healthy Ageing Twin Study.** *Int J Epidemiol* 2012. doi:10.1093/ije/dyr207.
8. Deary IJ, Gow AJ, Pattie A, Starr JM: **Cohort profile: The Lothian Birth Cohorts of 1921 and 1936**. *Int J Epidemiol* 2011. doi: 10.1093/ije/dyr197.
9. Wadsworth M, Kuh D, Richards M, Hardy R: **Cohort profile: the 1946 National Birth Cohort (MRC National Survey of Health and Development)**. *Int J Epidemiol* 2006, **35:**49-54.

10. Power C, Elliott J: **Cohort profile: 1958 British birth cohort (National Child Development Study).** *Int J Epidemiol* 2006, **35:**34-41.
11. Foley DL, Craig JM, Morley R, Olsson CA, Dwyer T, Smith K, Saffery R: **Prospects for epigenetic epidemiology.** *Am J Epidemiol* 2009, **169:**389-400.
12. De Stavola BL, Nitsch D, dos Santos Silva I, McCormack V, Hardy R, Mann V, Cole TJ, Morton S, Leon DA: **Statistical issues in life course epidemiology.** *Am J Epidemiol* 2006, **163:**84-96.
13. Kuh D, Ben-Shlomo Y, Lynch J, Hallqvist J, Power C: **Life course epidemiology.** *J Epidemiol Community Health* 2003, **57:**778-783.
14. Wong CC, Caspi A, Williams B, Craig IW, Houts R, Ambler A, Moffitt TE, Mill J: **A longitudinal study of epigenetic variation in twins.** *Epigenetics* 2010, **5:**516-526.
15. Talens RP, Boomsma DI, Tobi EW, Kremer D, Jukema JW, Willemsen G, Putter H, Slagboom PE, Heijmans BT: **Variation, patterns, and temporal stability of DNA methylation: considerations for epigenetic epidemiology.** *FASEB J* 2010, **24:**3135-3144.
16. Madrigano J, Baccarelli A, Mittleman MA, Sparrow D, Vokonas PS, Tarantini L, Schwartz J: **Aging and epigenetics: longitudinal changes in gene-specific DNA methylation.** *Epigenetics* 2012, **7:**63-70.
17. Madrigano J, Baccarelli A, Mittleman MA, Wright RO, Sparrow D, Vokonas PS, Tarantini L, Schwartz J: **Prolonged exposure to particulate pollution, genes associated with glutathione pathways, and DNA methylation in a cohort of older men.** *Environ Health Perspect* 2011, **119:**977-982.
18. Vineis P, Chuang SC, Vaissiere T, Cuenin C, Ricceri F, Johansson M, Ueland P, Brennan P, Herceg Z: **DNA methylation changes associated with cancer risk factors and blood levels of vitamin metabolites in a prospective study.** *Epigenetics* 2011, **6:**195-201.
19. Brooks JD, Cairns P, Shore RE, Klein CB, Wirgin I, Afanasyeva Y, Zeleniuch-Jacquotte A: **DNA methylation in pre-diagnostic serum samples of breast cancer cases: results of a nested case-control study.** *Cancer Epidemiol* 2010, **34:**717-723.
20. Gao Y, Baccarelli A, Shu XO, Ji BT, Yu K, Tarantini L, Yang G, Li HL, Hou L, Rothman N, Zheng W, Gao YT, Chow WH: **Blood leukocyte Alu and LINE-1 methylation and gastric cancer risk in the Shanghai Women's Health Study.** *Br J Cancer* 2012, **106:**585-591.
21. Gay LJ, Arends MJ, Mitrou PN, Bowman R, Ibrahim AE, Happerfield L, Luben R, McTaggart A, Ball RY, Rodwell SA: **MLH1 promoter methylation, diet, and lifestyle factors in mismatch repair deficient colorectal cancer patients from EPIC-Norfolk.** *Nutr Cancer* 2011, **63:**1000-1010.
22. Balassiano K, Lima S, Jenab M, Overvad K, Tjonneland A, Boutron-Ruault MC, Clavel-Chapelon F, Canzian F, Kaaks R, Boeing H, Meidtner K, Trichopoulou A, Laglou P, Vineis P, Panico S, Palli D, Grioni S, Tumino R, Lund E, Bueno-de-Mesquita HB, Numans ME, Peeters PH, Ramon Quirós J, Sánchez MJ, Navarro C, Ardanaz E, Dorronsoro M, Hallmans G, Stenling R, Ehrnström R, *et al.*: **Aberrant DNA methylation of cancer-associated genes in gastric cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC-EURGAST).** *Cancer Lett* 2011, **311:**85-95.
23. Limsui D, Vierkant RA, Tillmans LS, Wang AH, Weisenberger DJ, Laird PW, Lynch CF, Anderson KE, French AJ, Haile RW, Harnack LJ, Potter JD, Slager SL, Smyrk TC, Thibodeau SN, Cerhan JR, Limburg PJ: **Cigarette smoking and colorectal cancer risk by molecularly defined subtypes.** *J Natl Cancer Inst* 2010, **102:**1012-1022.
24. Van Guelpen B, Dahlin AM, Hultdin J, Eklof V, Johansson I, Henriksson ML, Cullman I, Hallmans G, Palmqvist R: **One-carbon metabolism and CpG island methylator phenotype status in incident colorectal cancer: a nested case-referent study.** *Cancer Causes Control* 2010, **21:**557-566.
25. Schernhammer ES, Giovannucci E, Baba Y, Fuchs CS, Ogino S: **B vitamins, methionine and alcohol intake and risk of colon cancer in relation to BRAF mutation and CpG island methylator phenotype (CIMP).** *PLoS ONE* 2011, **6:**e21102.
26. Boyd A, Golding J, Macleod J, Lawlor DA, Fraser A, Henderson J, Molloy L, Ness A, Ring S, Davey Smith G: **Cohort Profile: The 'Children of the 90s' – the index offspring of the Avon Longitudinal Study of Parents and Children.** *Int J Epidemiol* 2012. doi:10.1093/ije/dys064.
27. Fraser A, Macdonald-Wallis C, Tilling K, Boyd A, Golding J, Davey Smith G, Henderson J, Macleod J, Molloy L, Ness A, Ring S, Nelson SM, Lawlor DA: **Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort.** *Int J Epidemiol* 2012. doi:10.1093/ije/dys066.
28. Lepeule J, Baccarelli A, Motta V, Cantone L, Litonjua AA, Sparrow D, Vokonas PS, Schwartz J: **Gene promoter methylation is associated with lung function in the elderly: The Normative Aging Study.** *Epigenetics* 2012, **7:**261-269.
29. Bind MA, Baccarelli A, Zanobetti A, Tarantini L, Suh H, Vokonas P, Schwartz J: **Air pollution and markers of coagulation, inflammation, and endothelial function: associations and epigene-environment interactions in an elderly cohort.** *Epidemiology* 2012, **23:**332-340.
30. Zhu ZZ, Sparrow D, Hou L, Tarantini L, Bollati V, Litonjua AA, Zanobetti A, Vokonas P, Wright RO, Baccarelli A, Schwartz J: **Repetitive element hypomethylation in blood leukocyte DNA and cancer incidence, prevalence, and mortality in elderly individuals: the Normative Aging Study.** *Cancer Causes Control* 2011, **22:**437-447.
31. Baccarelli A, Wright R, Bollati V, Litonjua A, Zanobetti A, Tarantini L, Sparrow D, Vokonas P, Schwartz J: **Ischemic heart disease and stroke in relation to blood DNA methylation.** *Epidemiology* 2010, **21:**819-828.
32. Baccarelli A, Tarantini L, Wright RO, Bollati V, Litonjua AA, Zanobetti A, Sparrow D, Vokonas P, Schwartz J: **Repetitive element DNA methylation and circulating endothelial and inflammation markers in the VA Normative Aging Study.** *Epigenetics* 2010, **5:**222-228.
33. Kuh D, Pierce M, Adams J, Deanfield J, Ekelund U, Friberg P, Ghosh AK, Harwood N, Hughes A, Macfarlane PW, Mishra G, Pellerin D, Wong A, Stephen AM, Richards M, Hardy R; NSHD scientific and data collection team: **Cohort Profile: updating the cohort profile for the MRC National Survey of Health and Development: a new clinic-based data collection for ageing research.** *Int J Epidemiol* 2011, **40:**e1-e9.
34. Tillin T, Forouhi NG, McKeigue PM, Chaturvedi N: **Southall And Brent REvisited: Cohort profile of SABRE, a UK population-based comparison of cardiovascular disease and diabetes in people of European, Indian Asian and African Caribbean origins.** *Int J Epidemiol* 2012, **41:**33-42.
35. Hills SA, Balkau B, Coppack SW, Dekker JM, Mari A, Natali A, Walker M, Ferrannini E: **The EGIR-RISC STUDY (The European group for the study of insulin resistance: relationship between insulin sensitivity and cardiovascular disease risk): 1. Methodology and objectives.** *Diabetologia* 2004, **47:**566-570.
36. Kuh D, Ben-Shlomo Y: *A Life Course Approach to Chronic Disease Epidemiology: Tracing the Origins of Ill-health from Early to Adult Life.* 2nd edition. Oxford: Oxford University Press; 2004.
37. Mishra G, Nitsch D, Black S, De Stavola B, Kuh D, Hardy R: **A structured approach to modelling the effects of binary exposure variables over the life course.** *Int J Epidemiol* 2009, **38:**528-537.
38. Kuh D Ben-Shlomo Y, Lynch J, Hallqvist J, Power C: **Glossary for life course epidemiology.** *J Epidemiol Community Health* 2003, **57:**778-783
39. **Healthy Ageing across the Life Course** [http://www.halcyon.ac.uk]
40. **FALCon project** [http://www.nshd.mrc.ac.uk/collaborations/falcon.aspx]
41. **Bristol University: MRC Centre for Causal Analyses in Translational Epidemiology: GEoCoDE** [http://www.bristol.ac.uk/caite/geocode/]
42. Wills AK, Lawlor DA, Matthews FE, Sayer AA, Bakra E, Ben-Shlomo Y, Benzeval M, Brunner E, Cooper R, Kivimaki M, Kuh D, Muniz-Terrera G, Hardy R: **Life course trajectories of systolic blood pressure using longitudinal data from eight UK cohorts.** *PLoS Med* 2011, **8:**e1000440.
43. Davey Smith G, Lawlor DA, Harbord R, Timpson N, Day I, Ebrahim S: **Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology.** *PLoS Med* 2007, **4:**e352.
44. Wellcome Trust Case Control Consortium: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007, **447:**661-678.
45. Heijmans BT, Mill J: **Commentary: The seven plagues of epigenetic epidemiology.** *Int J Epidemiol* 2012, **41:**74-78.
46. Brion MJ, Zeegers M, Jaddoe V, Verhulst F, Tiemeier H, Lawlor DA, Davey Smith G: **Intrauterine effects of maternal prepregnancy overweight on child cognition and behavior in 2 cohorts.** *Pediatrics* 2011, **127:**e202-211.
47. Davey Smith G: **Assessing intrauterine influences on offspring health outcomes: can epidemiological studies yield robust findings?** *Basic Clin Pharmacol Toxicol* 2008, **102:**245-256.
48. Davey Smith G: **Negative control exposures in epidemiologic studies.** *Epidemiology* 2012, **23:**350-351; author reply 351-352.
49. Davey Smith G, Ebrahim S: **'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?** *Int J Epidemiol* 2003, **32:**1-22.
50. Davey Smith G: **Use of genetic markers and gene-diet interactions for interrogating population-level causal influences of diet on health.** *Genes Nutr* 2011, **6:**27-43.
51. Timpson NJ, Wade KH, Davey Smith G: **Mendelian randomization:**

application to cardiovascular disease. *Curr Hypertens Rep* 2012, **14**:29-37.

52. Relton CL, Davey Smith G: Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. *Int J Epidemiol* 2012, **41**:161-176.
53. Groom A, Potter C, Swan DC, Fatemifar G, Evans DM, Ring SM, Turcot V, Pearce MS, Embleton ND, Smith GD, Mathers JC, Relton CL: Postnatal growth and DNA methylation are associated with differential gene expression of the TACSTD2 gene and childhood fat mass. *Diabetes* 2012, **61**:391-400.
54. Waterland RA, Michels KB: Epigenetic epidemiology of the developmental origins hypothesis. *Annu Rev Nutr* 2007, **27**:363-388.
55. Waterland RA: Epigenetic epidemiology of obesity: application of epigenomic technology. *Nutr Rev* 2008, **66 Suppl 1**:S21-S23.
56. Relton CL, Groom A, St Pourcain B, Sayers AE, Swan DC, Embleton ND, Pearce MS, Ring SM, Northstone K, Tobias JH, Trakalo J, Ness AR, Shaheen SO, Davey Smith G: DNA methylation patterns in cord blood DNA and body size in childhood. *PLoS ONE* 2012, **7**:e31821.
57. Godfrey KM, Sheppard A, Gluckman PD, Lillycrop KA, Burdge GC, McLean C, Rodford J, Slater-Jefferies JL, Garratt E, Crozier SR, Emerald BS, Gale CR, Inskip HM, Cooper C, Hanson MA: Epigenetic gene promoter methylation at birth is associated with child's later adiposity. *Diabetes* 2011, **60**:1528-1534.
58. Gabory A, Attig L, Junien C: Developmental programming and epigenetics. *Am J Clin Nutr* 2011, **94(6 Suppl)**:1943S-1952S.
59. Heijmans BT, Tobi EW, Stein AD, Putter H, Blauw GJ, Susser ES, Slagboom PE, Lumey LH: Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc Natl Acad Sci U S A* 2008, **105**:17046-17049.
60. Teschendorff AE, Zhuang J, Widschwendter M: Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics* 2011, **27**:1496-1505.
61. Bjornsson HT, Sigurdsson MI, Fallin MD, Irizarry RA, Aspelund T, Cui H, Yu W, Rongione MA, Ekström TJ, Harris TB, Launer LJ, Eiriksdottir G, Leppert MF, Sapienza C, Gudnason V, Feinberg AP: Intra-individual change over time in DNA methylation with familial clustering. *JAMA* 2008, **299**:2877-2883.
62. Tawa R, Ueno S, Yamamoto K, Yamamoto Y, Sagisaka K, Katakura R, Kayama T, Yoshimoto T, Sakurai H, Ono T: Methylated cytosine level in human liver DNA does not decline in aging process. *Mech Ageing Devel* 1992, **62**:255-261.
63. Ono T, Tawa R, Shinya K, Hirose S, Okada S: Methylation of the c-myc gene changes during aging process of mice. *Biochem Biophys Res Comm* 1986, **139**:1299-1304.
64. Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, Heine-Suñer D, Cigudosa JC, Urioste M, Benitez J, Boix-Chornet M, Sanchez-Aguilera A, Ling C, Carlsson E, Poulsen P, Vaag A, Stephan Z, Spector TD, Wu YZ, Plass C, Esteller M: Epigenetic differences arise during the lifetime of monozygotic twins. *Proc Natl Acad Sci U S A* 2005, **102**:10604-10609.
65. Eckhardt F, Beck S, Gut IG, Berlin K: Future potential of the Human Epigenome Project. *Expert Rev Mol Diagn* 2004, **4**:609-618.
66. Martin GM: Epigenetic drift in aging identical twins. *Proc Natl Acad Sci U S A* 2005, **102**:10413-10414.
67. Zhu ZZ, Hou L, Bollati V, Tarantini L, Marinelli B, Cantone L, Yang AS, Vokonas P, Lissowska J, Fustinoni S, Pesatori AC, Bonzini M, Apostoli P, Costa G, Bertazzi PA, Chow WH, Schwartz J, Baccarelli A: Predictors of global methylation levels in blood DNA of healthy subjects: a combined analysis. *Int J Epidemiol* 2010. doi:10.1093/ije/dyq154.
68. Software by E. Andres Houseman: methylSpectrum [http://people.oregonstate.edu/~housemae/software/]
69. Ma P, Zhong W, Liu J: Identifying differentially expressed genes in time course microarray data. *Stat Biosci* 2009, **1**:144-159.
70. Storey JD, Xiao W, Leek JT, Tompkins RG, Davis RW: Significance analysis of time course microarray experiments. *Proc Natl Acad Sci U S A* 2005, **102**:12837-12842.
71. Yuan M, Kendziorski C: Hidden Markov models for microarray time course data in multiple biological conditions. *J Am Stat Assoc* 2006, **101**:1323-1332.
72. Yuan Y, Li CT, Wilson R: Partial mixture model for tight clustering of gene expression time-course. *BMC Bioinformatics* 2008, **9**:287.
73. Li N, McMurry T, Berg A, Wang Z, Berceli SA, Wu R: Functional clustering of periodic transcriptional profiles through ARMA(p,q). *PLoS ONE* 2010, **5**:e9894.
74. Choi H, Pavelka N: When one and one gives more than two: challenges and opportunities of integrative omics. *Front Genet* 2011, **2**:105.
75. Zhang K, Wang H, Bathke AC, Harrar SW, Piepho HP, Deng Y: Gene set analysis for longitudinal gene expression data. *BMC Bioinformatics* 2011, **12**:273.
76. Cole DM, Smith SM, Beckmann CF: Advances and pitfalls in the analysis and interpretation of resting-state fMRI data. *Front Syst Neurosci* 2010, **4**:8.
77. Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, Wei G, Zhao K: Dynamic regulation of nucleosome positioning in the human genome. *Cell* 2008, **132**:887-898.
78. Cedar H, Bergman Y: Linking DNA methylation and histone modification: patterns and paradigms. *Nat Rev Genet* 2009, **10**:295-304.
79. Tiwari VK, McGarvey KM, Licchesi JDF, Ohm JE, Herman JG, Schübeler D, Baylin SB: PcG Proteins, DNA methylation, and gene repression by chromatin looping. *PLoS Biol* 2008, **6**:e306.
80. Edwards JR, O'Donnell AH, Rollins RA, Peckham HE, Lee C, Milekic MH, Chanrion B, Fu Y, Su T, Hibshoosh H, Gingrich JA, Haghighi F, Nutter R, Bestor TH: Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. *Genome Res* 2010, **20**:972-980.
81. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, Gilad Y, Pritchard JK: DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol* 2011, **12**:R10.
82. Schadt EE, Bjorkegren JL: NEW: network-enabled wisdom in biology, medicine, and health care. *Sci Transl Med* 2012, **4**:115rv111.
83. Zhang Y, Tibshirani RJ, Davis RW: Predicting patient survival from longitudinal gene expression. *Stat Appl Genet Mol Biol* 2010, **9**:Article41.
84. Minas C, Waddell SJ, Montana G: Distance-based differential analysis of gene curves. *Bioinformatics* 2011, **27**:3135-3141.
85. Wang Y, Xu M, Wang Z, Tao M, Zhu J, Wang L, Li R, Berceli SA, Wu R: How to cluster gene expression dynamics in response to environmental signals. *Brief Bioinform* 2012, **13**:162-174.
86. Tai YC, Speed TP: On gene ranking using replicated microarray time course data. *Biometrics* 2009, **65**:40-51.
87. Qiu P, Gentles AJ, Plevritis SK: Discovering biological progression underlying microarray samples. *PLoS Comput Biol* 2011, **7**:e1001123.
88. Rubingh CM, Bijlsma S, Jellema RH, Overkamp KM, van der Werf MtJ, Smilde AK: Analyzing longitudinal microbial metabolomics data. *J Proteme Res* 2009, **8**:4319-4327.
89. Phan AH, Cichocki A: Tensor decompositions for feature extraction and classification of high dimensional data. *Nonlinear Theory Applications* 2010, **1**:37-68.
90. Puig AT, Wiesel A, Zaas AK, Woods CW, Ginsburg GS, Fleury G, Hero AO: Order-Preserving factor analysis - application to longitudinal gene expression. *IEEE Trans Signal Process* 2011, **59**:4447-4458.
91. Smith AA, Vollrath A, Bradfield CA, Craven M: Similarity queries for temporal toxicogenomic expression profiles. *PLoS Comput Biol* 2008, **4**:e1000116.
92. Borghol N, Suderman M, McArdle W, Racine A, Hallett M, Pembrey M, Hertzman C, Power C, Szyf M: Associations with early-life socio-economic position in adult DNA methylation. *Int J Epidemiol* 2012, **41**:62-74.
93. Herbstman JB, Tang D, Zhu D, Qu L, Sjödin A, Li Z, Camann D, Perera FP: Prenatal exposure to polycyclic aromatic hydrocarbons, benzo[a]pyrene-DNA adducts, and genomic DNA methylation in cord blood. *Environ Health Perspect* 2012, **120**:733-738.
94. Koenen KC, Uddin M, Chang SC, Aiello AE, Wildman DE, Goldmann E, Galea S: SLC6A4 methylation modifies the effect of the number of traumatic events on risk for posttraumatic stress disorder. *Depress Anxiety* 2011, **28**:639-647.
95. Uddin M, Galea S, Chang SC, Aiello AE, Wildman DE, de los Santos R, Koenen KC: Gene expression and methylation signatures of MAN2C1 are associated with PTSD. *Dis Markers* 2011, **30**:111-121.
96. Uddin M, Koenen KC, Aiello AE, Wildman DE, de los Santos R, Galea S: Epigenetic and inflammatory marker profiles associated with depression in a community-based epidemiologic sample. *Psychol Med* 2011, **41**:997-1007.
97. Uddin M, Aiello AE, Wildman DE, Koenen KC, Pawelec G, de Los Santos R, Goldmann E, Galea S: Epigenetic and immune function profiles associated with posttraumatic stress disorder. *Proc Natl Acad Sci U S A* 2010, **107**:9470-9475.
98. Lumey L, Terry MB, Delgado-Cruzata L, Liao Y, Wang Q, Susser E, McKeague I, Santella RM: Adult global DNA methylation in relation to pre-natal nutrition. *Int J Epidemiol* 2012, **41**:116-123.
99. Michels KB, Harris HR, Barault L: Birthweight, maternal weight trajectories and global DNA methylation of LINE-1 repetitive elements. *PloS ONE* 2011, **6**:e25254.

100. Olsson CA, Foley DL, Parkinson-Bates M, Byrnes G, McKenzie M, Patton GC, Morley R, Anney RJ, Craig JM, Saffery R: Prospects for epigenetic research within cohort studies of psychological disorder: a pilot investigation of a peripheral cell marker of epigenetic risk for depression. *Biol Psychol* 2010, **83:**159-165.
101. Essex MJ, Thomas Boyce W, Hertzman C, Lam LL, Armstrong JM, Neumann SM, Kobor MS: Epigenetic vestiges of early developmental adversity: childhood stress exposure and DNA methylation in adolescence. *Child Dev* 2011. doi: 10.1111/j.1467-8624.2011.01641.x.
102. Sood A, Petersen H, Blanchette CM, Meek P, Picchi MA, Belinsky SA, Tesfaigzi Y: Wood smoke exposure and gene promoter methylation are associated with increased risk for COPD in smokers. *Am J Respir Crit Care Med* 2010, **182:**1098-1104.
103. Flom JD, Ferris JS, Liao Y, Tehranifar P, Richards CB, Cho YH, Gonzalez K, Santella RM, Terry MB: Prenatal smoke exposure and genomic DNA methylation in a multiethnic birth cohort. *Cancer Epidemiol Biomarkers Prev* 2011, **20:**2518-2523.
104. McKay JA, Groom A, Potter C, Coneyworth LJ, Ford D, Mathers JC, Relton CL: Genetic and non-genetic influences during pregnancy on infant global and site specific DNA methylation: role for folate gene variants and vitamin B12. *PLoS ONE* 2012, **7:**e33290.
105. Boeke CE, Baccarelli A, Kleinman KP, Burris HH, Litonjua AA, Rifas-Shiman SL, Tarantini L, Gillman M: Gestational intake of methyl donors and global LINE-1 DNA methylation in maternal and cord blood: prospective results from a folate-replete population. *Epigenetics* 2012, **7:**253-260.
106. Kile ML, Baccarelli A, Tarantini L, Hoffman E, Wright RO, Christiani DC: Correlation of global and gene-specific DNA methylation in maternal-infant pairs. *PloS ONE* 2010, **5:**e13730.
107. Kile ML, Baccarelli A, Hoffman E, Tarantini L, Quamruzzaman Q, Rahman M, Mahiuddin G, Mostofa G, Hsueh YM, Wright RO, Christiani DC: Prenatal arsenic exposure and DNA methylation in maternal and umbilical cord blood leukocytes. *Environ Health Perspect* 2012. http://dx.doi.org/10.1289/ehp.1104173.
108. McGuinness D, McGlynn LM, Johnson PC, Macintyre A, Batty GD, Burns H, Cavanagh J, Deans KA, Ford I, McConnachie A, McGinty A, McLean JS, Millar K, Packard CJ, Sattar NA, Tannahill C, Velupillai YN, Shiels PG: Socio economic status is associated with epigenetic differences in the pSoBid cohort. *Int J Epidemiol* 2012, **41:**151-160.
109. Marsit CJ, Maccani MA, Padbury JF, Lester BM: Placental 11-Beta hydroxysteroid dehydrogenase methylation is associated with newborn growth and a measure of neurobehavioral outcome. *PLoS ONE* 2012, **7:**e33794.
110. Kim M, Long TI, Arakawa K, Wang R, Yu MC, Laird PW: DNA methylation as a biomarker for cardiovascular disease risk. *PLoS ONE* 2010, **5:**e9692.
111. Harvey NC, Lillycrop KA, Garratt E, Sheppard A, McLean C, Burdge G, Slater-Jefferies J, Rodford J, Crozier S, Inskip H, Emerald BS, Gale CR, Hanson M, Gluckman P, Godfrey K, Cooper C: Evaluation of methylation status of the eNOS promoter at birth in relation to childhood bone mineral content. *Calcif Tissue Int* 2012, **90:**120-127.
112. Brennan K, Garcia-Closas M, Orr N, Fletcher O, Jones M, Ashworth A, Swerdlow A, Thorne H; KConFab Investigators, Riboli E, Vineis P, Dorronsoro M, Clavel-Chapelon F, Panico S, Onland-Moret NC, Trichopoulos D, Kaaks R, Khaw KT, Brown R, Flanagan JM*:* Intragenic ATM methylation in peripheral blood DNA as a biomarker of breast cancer risk. *Cancer Res* 2012, **72:**2304-2313.
113. Hughes LA, Simons CC, van den Brandt PA, Goldbohm RA, de Goeij AF, de Bruine AP, van Engeland M, Weijenberg MP: Body size, physical activity and risk of colorectal cancer with or without the CpG island methylator phenotype (CIMP). *PLoS ONE* 2011, **6:**e18571.

doi:10.1186/gb-2012-13-6-246
Cite this article as: Ng JWY, *et al.*: The role of longitudinal cohort studies in epigenetic epidemiology: challenges and opportunities. *Genome Biology* 2012, **13:**246.