# Investigating the Genetics of Sporadic Early-Onset Alzheimer's Disease Using a Customised Genotyping Chip

Submitted to the

School of Life Sciences

of the

University of Nottingham

In fulfilment of the requirements for the degree of

Doctor of Philosophy

By

Imelda Barber, BSc, MSc

Supervisor

Professor Kevin Morgan

October 2016

# Contents

**Abstract** ..... VI
**Acknowledgments** ..... VII
**Declarations** ..... VIII

**1. Introduction** ..... 1
1.1. Dementia ..... 1
1.2. Alzheimer's disease ..... 3
1.2.1. Pathology ..... 3
1.2.2. Diagnosis ..... 7
1.2.3. Age at Onset ..... 8
1.3. Genetics Linked with Alzheimer's Disease ..... 10
1.3.1. Variants of high effect and rare frequency ..... 10
1.3.2. Variants of moderate effect and common frequency ..... 12
1.3.3. Variants of low effect and common frequency ..... 14
1.3.4. Variants of moderate effect and low frequency ..... 18
1.4. Metabolic Pathways Linked with Alzheimer's Disease ..... 20
1.4.1. Amyloid ..... 20
1.4.2. Tau ..... 23
1.4.3. Cholesterol ..... 25
1.4.4. Intracellular trafficking ..... 26
1.4.5. Oxidative stress and metal ions ..... 28
1.4.6. Cell cycle ..... 30
1.4.7. Neuroinflammation ..... 32
1.5. The NeuroX Genotyping Chip ..... 35
1.5.1. Illumina design ..... 35
1.5.2. Utility of the NeuroX ..... 38
1.5.3. Aims of the Project ..... 39

**2. DNA Preparation and *APOE* Genotyping** ..... 40
2.1. Methods ..... 40
2.1.1. DNA extraction ..... 40
2.1.2. Genotyping *APOE* ε ..... 42
2.2. Results ..... 48

**3. Screening sEOAD Samples for Causative Variants in *APP*** ..... 49
3.1. Methods ..... 49
3.1.1. Primer design ..... 51
3.1.2. Sanger sequencing sEOAD samples ..... 51
3.1.3. Effect of the 6bp deletion on cDNA from COS7 and BE(2)-C cell lines ..... 52
3.1.4. Effect of the 6bp deletion on cDNA from Patient brain cells ..... 54
3.1.5. Genotyping the 6bp deletion in LOAD and control samples ..... 54
3.1.6. Susceptibility to sEOAD and LOAD from the 6bp deletion ..... 54
3.2. Results and discussion ..... 55
3.2.1. Sanger sequencing sEOAD samples ..... 55
3.2.2. Effect of the 6bp deletion on cDNA from COS7 and BE(2)-C cell lines ..... 58
3.2.3. Effect of the 6bp deletion on cDNA from patient brain cells ..... 60
3.2.4. Genotyping the 6bp deletion in LOAD and control samples ..... 62
3.2.5. Susceptibility to sEOAD and LOAD from the 6bp deletion ..... 64

**4. NeuroX Data Quality Control in GenomeStudio** ..... 65
4.1. Methods ..... 65
4.1.1. Creating a project ..... 67
4.1.2. Uploading external information ..... 68
4.1.3. Excluding poor quality samples ..... 69
4.1.4. Automatic clustering of un-called SNPs ..... 69
4.1.5. Assessing non-autosomal SNPs ..... 69
4.1.6. Assessing autosomal SNPs ..... 71
4.1.7. Assessing specific SNPs ..... 72
4.1.8. Reintroducing poor quality samples ..... 72
4.1.9. Exporting calls to PLINK format ..... 72

**5. NeuroX Data Quality Control in PLINK** ..... 73
5.1. Methods ..... 73
5.1.1. Updating marker information ..... 74
5.1.2. Removing sample and markers failing data rate ..... 77
5.1.3. Gender ..... 78
5.1.4. Creating additional data ..... 79
5.1.5. Genetic ancestry ..... 80
5.1.6. Hardy Weinberg Equilibrium (HWE) ..... 82
5.1.7. Relatedness ..... 82
5.1.8. Heterozygosity ..... 83
5.1.9. Removing samples and markers ..... 83
5.1.10. Creating ancestry covariates ..... 84
5.1.11. Imputing *APOE* ε status ..... 85
5.2. Results ..... 86

**6. NeuroX Marker Annotation** ..... 87
6.1. Methods ..... 87
6.1.1. Reference and alternative allele ..... 87
6.1.2. Proximal genes ..... 88
6.1.3. Variant and predicted consequence ..... 89
6.2. Results and discussion ..... 90

**7. NeuroX Known Causative Variants in sEOAD** ..... 93
7.1. Methods ..... 93
7.1.1. List of causative variants ..... 93
7.1.2. Causative variants genotyped on the NeuroX ..... 94
7.1.3. Samples harbouring a causative variant ..... 96
7.1.4. Genotype validation ..... 96
7.2. Results and discussion ..... 97
7.2.1. Causative variants in *PARK2* ..... 100
7.2.2. Causative variant in *FUS* ..... 101
7.2.3. Causative variant in *GRN* ..... 102

**8. NeuroX Predicted Pathogenic Variants in sEOAD** ..... 103
8.1. Methods ..... 103
8.1.1. List of genes ..... 104
8.1.2. 16 familial genes ..... 104
8.1.3. 96 other genes ..... 105
8.2. Results and discussion ..... 106
8.2.1. 16 familial genes ..... 106
8.2.2. 96 other genes ..... 112

**9. Analyses of All NeuroX Markers in sEOAD** ..... 118
9.1. Methods ..... 118
9.1.1. All variants ..... 118
9.1.2. Variants from GWA studies ..... 119
9.1.3. All genes ..... 119
9.1.4. All pathways ..... 120
9.2. Results and discussion ..... 121
9.2.1. All variants ..... 121
9.2.2. Variants from GWA studies ..... 128
9.2.3. All genes ..... 133
9.2.4. All pathways ..... 137

**10. Overall Discussion** ..... 142
10.1. The NeuroX ..... 142
10.2. Genetics of sEOAD ..... 142

**11. Appendix - In-house Software** ..... 145
11.1. Description of use ..... 145
11.2. AnnotateWithGenomicRegion.pl ..... 146
11.3. ImissVsGender.r ..... 148
11.4. Relatedness.pl ..... 149
11.5. Allele.pl ..... 151
11.6. FormatVEP.pl ..... 153
11.7. MatchingVariants.pl ..... 156
11.8. SwitchPrimaryKey.pl ..... 159
11.9. FormatPrimaryKey.pl ..... 160
11.10. RemoveDuplicates.pl ..... 161
11.11. CombineTwoPrimaryKeys.pl ..... 162
11.12. GeneAnalysis.r ..... 163

**12. Appendix – External Software** ..... 166
12.1. Databases ..... 166
12.2. Quality control and analyses ..... 167
12.3. Programming ..... 167
12.4. PCR and sequencing ..... 168

**Publications** ..... 169
**Abbreviations** ..... 170
**References** ..... 171

# Abstract

Alzheimer's disease (AD) is the commonest form of dementia and is characterised with neuropathological hallmarks such as aggregated amyloid plaques and hyper-phosphorylated tau protein. One type of AD is autosomal dominant AD (ADAD) which is caused by highly penetrant variants in one of three genes (*APP*, *PSEN1* and *PSEN2*), other cases of AD are described as sporadic and can have a late onset of disease symptoms (> 65 years of age) or early onset (≤ 65 years or age). Late-onset Alzheimer's disease (LOAD) is estimated to be 70% heritable and is common. Conversely sporadic early-onset Alzheimer's disease (sEOAD) is estimated to 90% heritable but is relatively rare. The difference in prevalence between the two types of AD has resulted in genome wide association studies focusing on LOAD with sEOAD receiving little attention. Here we use an Illumina human exome genotyping chip customised with neurodegenerative markers (NeuroX) to genotype the coding region of sEOAD samples in a hope to elucidate the genetic aetiology of sEOAD.

Sanger sequencing exons 16 and 17 of *APP* was conducted in a sEOAD cohort (n=451) to screen for variants known to cause ADAD; 9% (n=4) of the cohort were heterozygous for known causative variants and where subsequently removed from the sEOAD NeuroX genotyping data before analyses. Screening also highlighted an intronic 6bp deletion downstream of exon 17 in *APP* with a non-significant increased minor allele frequency (MAF) in sEOAD, however it did not appear to influence splicing of exon 17. Screening the sEOAD cohort for other variants known to cause neurodegenerative disease was conducted using the NeuroX genotyping data (n=408) which identified two samples with variants in *PARK2*, these variants are thought to contribute susceptibility to Parkinson's disease (PD) suggesting these variants might elicit risk for multiple diseases. A further study with increased power would ascertain if the 6bp deletion and *PARK2* variants are associated with sEOAD.

Statistical analyses of the sEOAD NeuroX genotypes highlighted many variants, genes and pathways that could be contributing to susceptibility to disease; however no tests reached significance after adjusting for multiple testing. The genes most associated (*PDZK1*, *DCLK3*, *SLC33A1* and *BLOC1S2*) appear to be biologically relevant and would be ideal candidates for further study. Additionally, just under half of the variants that are significant associated with LOAD were genotyped on the NeuroX and two of these were significantly associated with sEOAD after correcting for multiple testing (rs3851179 and rs3764650). The genotypes of all the variants highlighted would need to be verified before their functionalities were investigated further.

# Acknowledgments

I'd like to thank Alzheimer's Research UK (ARUK) and the School of Life sciences for jointly funding my PhD. I'd like to thank my examiners Dr Patrick Tighe and Professor John Powell for spending time and effort examining my research. Thanks to all past and future colleagues, including Christopher Medway, Jenny Lord, James Turton, Keeley Brookes, Naomi Clement, Tulsi Patel, Helen Knight, Michelle Strickland and Hazel Wadeson. Special thanks to the following people:-

Professor Kevin Morgan has orchestrated, organised and campaigned for the lab's research; he is doing a fantastic job. Kevin has provided support, motivation, laughter and most importantly office space during the last four years, all of which have made my time enjoyable and my work rewarding.

Dr Sally Chappell has an amazing selfless capacity to take time out of her day and dedicate it to teaching others, she has on several occasion helped me. Sally is doing fantastic work on the Molecular Genetics and Diagnostics masters course which I was a student of before starting my PhD; she has juggled being a course organiser, lecturer and supervisor. The course was invaluable and gave me the knowledge and opportunity to undertake this PhD.

Dr Tamar Guetta-Baranes is the back bone to the lab. I fondly refer to Tamar as the lab goddess as she has kept the lab functioning smoothly without fault and she knows everything about performing molecular techniques. Tamar is nurturing and has taken all those that enter the lab under her wing, including PhD students such as myself and huge numbers of master/medical students.

Dr Anne Braae is a great person and good friend. Anne did the masters course and PhD at the same time as me and we helped each other where we could. Our parallel journey meant we often ranted to one another about life's ups and downs over a hot beverage.

Special love filled thanks to my partner David Buckingham who has shared every step of my experience, albeit usually via tired, ill expressed narratives. David has provided copious amounts of support, advice, love and laughter. David's mother Susan Buckingham, father Martin Buckingham and brother Michael Buckingham have always welcomed me and made me feel like part of their family. I'm especially grateful to Martin who was kind enough to cart my belongings around the country during my undergraduate studies.

Last but not least, special love filled thanks to my father Brett Barber, my mother Heather Barber and my siblings April Barber, Katrina Barber, Andrea Barber, Dean Barber and Liam Barber. They have kept me grounded and always invite me into their homes with open arms, lots of love and copious amounts of tea. They have never failed to put a smile on my face, nor supply me with entertaining drama filled anecdotes.

# Declarations

**Section 3.1.2** was performed in part by Mr Christopher Deacon and Ms Chidchanok Sakdapanichkul. **Section 3.1.3** was performed by Ms Jennyfer M. García-Cárdenas. **Section 3.1.5** was performed by Ms Gabriela Zapata Erazo.

# 1. Introduction

## 1.1. Dementia

Dementia is a syndrome that describes a vast array or disorders with a multitude of causes. Alzheimer's disease (AD) is neurodegenerative and is the most common cause of dementia. Other closely related neurodegenerative dementias include Frontotemporal Dementia (FTD) and Dementia with Lewy Bodies (DLB). **Table 1.1** lists the most common conditions causing dementia and demonstrates clearly the diversity of the syndrome. Definitions of dementia are varied, one description describes dementia as 'a syndrome of acquired persistent intellectual impairments characterised by deterioration in at least three of the following domains: memory, language, visuospatial skills, personality or behaviour and manipulation of acquired knowledge (including executive function)'.[1]

| Category | Examples |
|---|---|
| Infection | Syphilis, Lyme disease, chronic meningitis, PML, HIV, Whipple's disease |
| Neoplasms | Primary or metastatic tumours, paraneoplastic encephalitis |
| Traumatic brain disease | Chronic subdural hematoma, contusions, diffuse axonal injury, hydrocephalus |
| Autoimmune diseases | Multiple Sclerosis, primary CNS angiitis, lupus and other vasculidites, sarcoid |
| Metabolic disorders | Renal and hepatic failure, hyper/hypo-thyroidism/calcemia/natremia, Wilson's disease |
| Toxic disorders | Drugs: antidepressants, anxiolytics, sedatives, hypnotics, anticholinergics, neuroleptics<br>Metals: arsenic, thallium, lead, manganese<br>Industrial agents : $CC1_4$, $CS_2$, TCE, organophosphides<br>Radiation encephalopathy<br>Alcohol and other drugs of abuse |
| Nutrition/Deprivation | B12/Folate and other vitamin deficiencies, Wernicke-Korsakoff syndrome |
| Degenerative disorders | Alzheimer's disease, Frontotemporal and Parkinsonian dementias, Dementia with Lewy Bodies and Huntington's disease |
| Vascular diseases | Multiple infarct dementia, Binswanger's disease, Small vessel ischemic disease |
| Psychiatric disorders | Schizophrenia, Dementia syndrome of depression, Bipolar disorder, Malingering |

**Table 1.1** This table lists just some of the differential diagnosis of the dementia syndrome. Column one (Category) subdivides the differential diagnosis into types, column two gives examples of conditions causing dementia. Key: $CC1_4$, carbon tetrachloride; $CS_2$, carbon disulphide; CNS, central nervous system; PML, progressive multi-focal leukoencephalopathy; HIV, human immunodeficiency virus; TCE, trichloroethylene. Table was taken from Budsen and Kowall 2011 and modified.[1]

AD is a debilitating fatal disease, whereby patients undergo an insidious decline in memory, thinking and behaviour beyond what is considered a normal part of aging. The average duration of disease is eight years but patients can suffer the disease for up to several decades.[2] As with any complex disease, the clinical features between two patients are never the same and each case will present with a dominant clinical feature. AD conforms to four profiles, each profile is dependent on one dominant clinical feature and the most common profile is termed Dementia of Alzheimer's Type (DAT) where the dominant feature is progressive amnestic dysfunction. The dominant features of the other profiles

include progressive language, compartmental/executive and visual syndrome.[1] AD progresses through 3 stages: mild, moderate and severe. Mild cognitive impairment (MCI) appears to be a condition positioned between normal ageing and dementia, thus is arguably the initial stage of AD. **Table 1.2** demonstrates features present at each stage of AD.

| Feature | MCI | Mild | Moderate | Severe |
|---|---|---|---|---|
| Memory | | | | |
| Working | - | -/+ | ++ | +++ |
| Anterograde episodic | ++ | +++ | +++ | +++ |
| Remote | -/+ | -/+ | ++ | +++ |
| Semantic | -/+ | + | +++ | +++ |
| Attention and executive | -/+ | ++ | ++ | +++ |
| Language | - | -/+ | + | ++ |
| Visuospatial and perceptual | - | -/+ | ++ | ++ |
| Praxis | - | - | ++ | ++ |

**Table 1.2** Clinical features of Dementia of Alzheimer's Type (DAT) at each stage of disease; mild cognitive impairment (MCI), mild AD (Mild), moderate AD (Moderate) and severe AD (Severe). Key: -, absent; +, present; -/+, variable; MCI, mild cognitive impairment. Table taken from Hodges 2006.[3]

In 2013 dementia was the third largest cause of death in the UK; up 52% since 1990.[4] In the UK approximately 850,000 people were living with dementia in 2015.[5] Other high income countries are seeing similar figures and the World Health Organisation (WHO) reported AD and other dementias to be the fourth leading cause of death in high-income countries in 2012.[6] The overall economic impact of dementia in the UK was £26.3 billion in 2015,[5] despite the huge financial burden of dementia on the UK, the funding received for dementia research was a fraction of that given to other lower financial burden diseases such as cancer.[7]

The earliest documentation of Alzheimer's disease resulted from the work of Alois Alzheimer; who was a German psychiatrist recruited in 1903 to investigate the syndrome. His seminal work revolved around a female patient of 53 years of age, after her death he examined her brain and found tangled neurofibrils within nerve cells and miliar foci of a particular substance, we now know these features as neurofibrillary tangles (NFTs) and senile plaques (SPs). The disease was later named after Alois Alzheimer and these features are today the pathological hallmark of AD. Major research into AD began in 1974 Maryland, USA with the creation of the National Institute of Aging at the National Institute of Health (NIA-NIH).

## 1.2. Alzheimer's disease

### 1.2.1. Pathology

On a macroscopic level AD brain demonstrates brain atrophy and spongiosis. Brain atrophy in AD occurs in all main lobes in a bilaterally symmetrical fashion, specifically in the entorhinal cortex, hippocampus and amygdala where tau pathology[8] and neuronal loss predominate. On average, healthy elderly people lose less than one percent of their grey matter tissue per year, whereas those with AD lose approximately three percent.

AD brain is filled with bodies such as Granulo-vacuolara, Perisomatic granules and Hirano bodies along with accumulations of intracellular hyper phosphorylated TAU protein and extracellular amyloid protein. **Figure 1.1** illustrates the general structure and landscape of neurons and astrocytes.

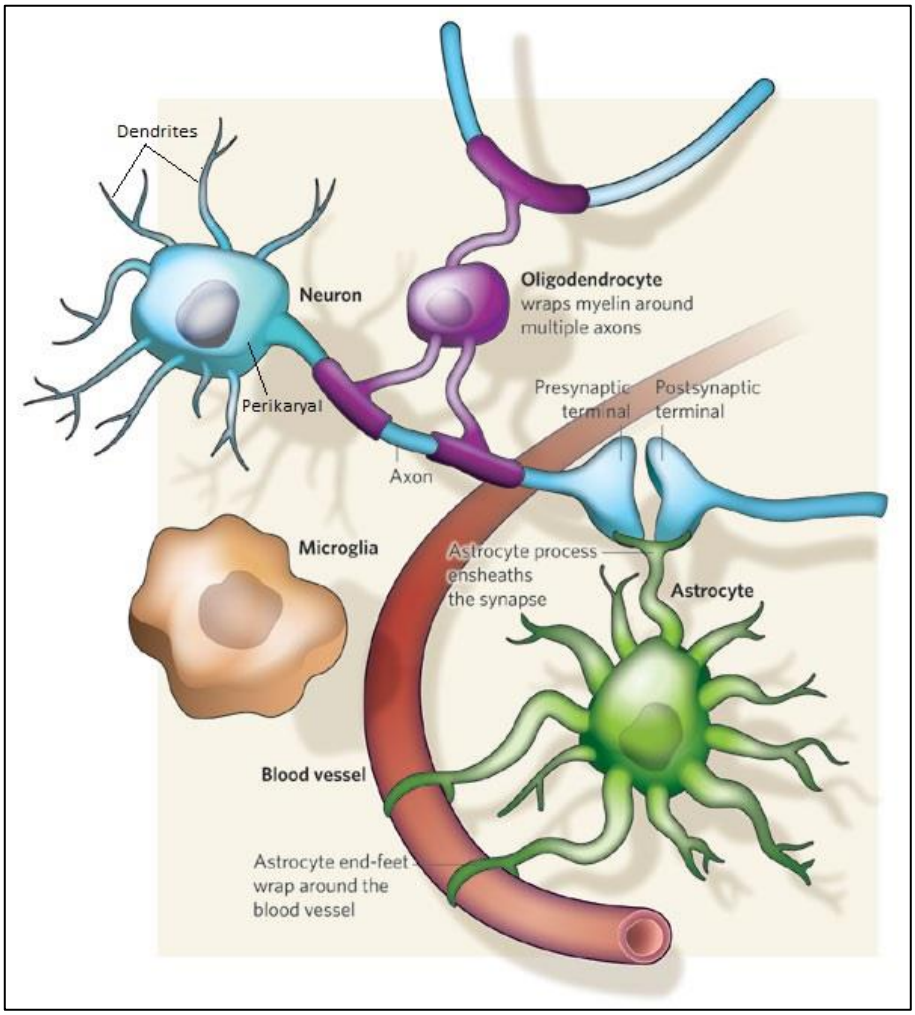



**Figure 1.1**. Figure depicts the anatomical structure of neurons (blue) and their connections to surrounding cells, such as astrocytes (green) and other neurons via synaptic terminals. Image taken from Allen and Barres 2009 and modified.[9]

Other microscopic occurrences in AD brains are the loss of certain features, such as neurons, synapses and neurotransmitters. The predominant microscopic features of AD are accumulations of amyloid protein and tau, both occur naturally in normal aging, however in disease they usually appear in greater quantities in specific areas of the brain and propagate through the brain in a particular pattern. Accumulations of amyloid protein form microscopic senile plaques (SP) which are mainly found in the grey matter of AD brain (**Figure 1.2A**), their shape can be described as stellate (star like), focal (focused) or diffuse (spread out).[10] The principle component of these deposits is Aβ which is detected with anti-Aβ immunohistochemistry (IHC), and amyloid (aggregated Aβ) which can be detected by thioflavin S staining or Congo red. Aβ deposits are located in the brain parenchymal (functional part) or the vessel walls of arteries, veins and capillaries where they are known collectively as cerebral amyloid angiopathy (CAA). Focal SPs have a core that is predominately Aβ and which are commonly surrounded by microglia cells. When focal SPs are surrounded by neuritic or astrocytic components (neuritic corona) they are also called neuritic plaques, those that are not surrounded by a neuritic corona are called cotton wool plaques (CWP). Hyper-phosphorylated tau deposits or inclusions are found within certain neurons and can be described as neurofibrillary tangles (NFT) if they occur in the perikaryal (cell body) (**Figure 1.2B**), or neuropil threads if they occur in the dendrites.[10] NFTs can also be seen extracellularly in severely affected regions as a result of cell death.

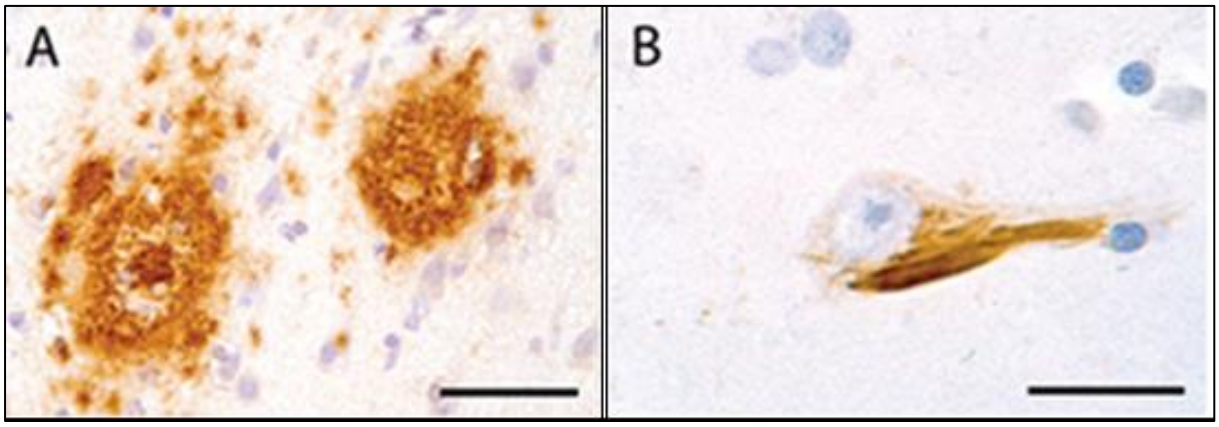



**Figure 1.2** An image of a senile plaque (SP) from the neocortex of a patient with Alzheimer's disease (AD) (A) and an image of a neurofibrillary tangle (NFT) from the neocortex of a patient with AD (B). Scale bars are 50µm (A) and 20µm (B). Image taken from Jucker *et al* 2013.[11]

It is consensus of opinion that amyloid pathology precedes tau pathology; supporting this view is the fact causative variants in *APP/PSEN1*/*PSEN2* leads to AD with amyloid and tau deposits, however causative variants in *MAPT* leads to FTD with tau deposits but not amyloid deposits.[12] Conversely however, amyloid and tau deposits appear to propagate from and to different areas of the brain (**Figure 1.3**).

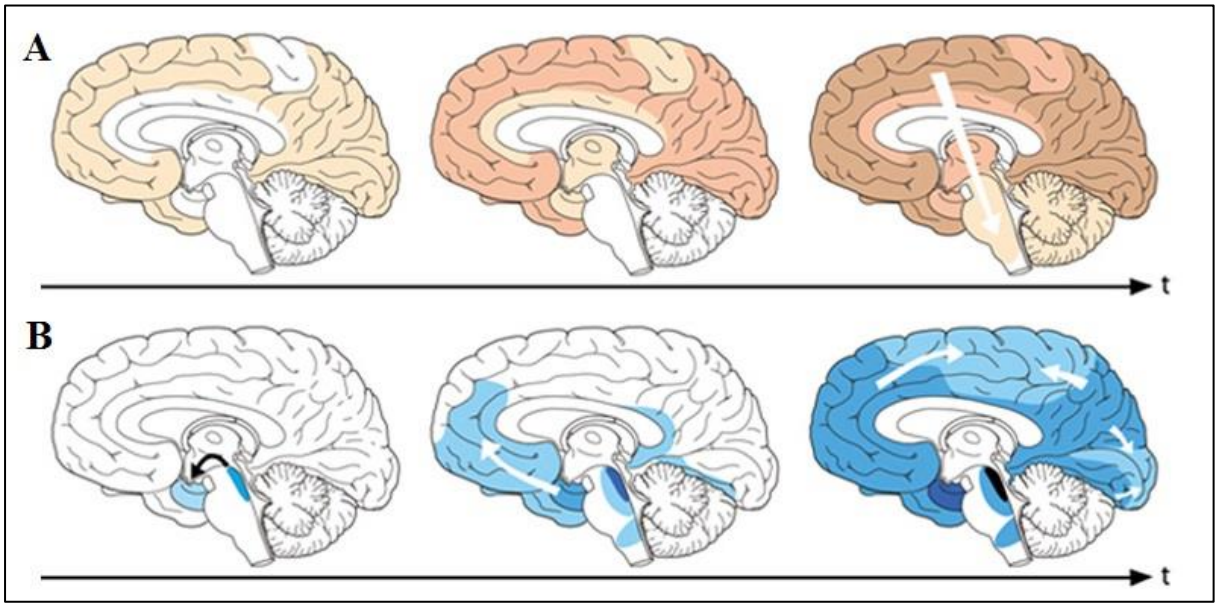



**Figure 1.3** Characteristic progression of amyloid beta deposits (A) and tau deposits (B) in Alzheimer's disease over time (t, black arrows), inferred from post-mortem analyses of brains. Three stages are shown in the medial brain and white arrows indicate the putative spread of the lesions. Image taken from Jucker *et al* 2013.[11]

Amyloid deposits do not show a strong correlation with severity of clinical features and are not believed to be a cause of disease, indeed cognitively normal individuals can have a high burden of amyloid deposits, and AD brains can be void of amyloid deposits altogether. Tau deposits[13] and Aβ oligomers[14,15] show the best correlation with severity of clinical features. It is believed that Aβ oligomers lead to tau pathology which in turn mediates neuronal injury and dysfunction, this observation is still under debate and much more confirmatory research is needed. **Figure 1.4** shows the time course of Aβ and tau biomarkers in relation to macroscopic and clinical features.

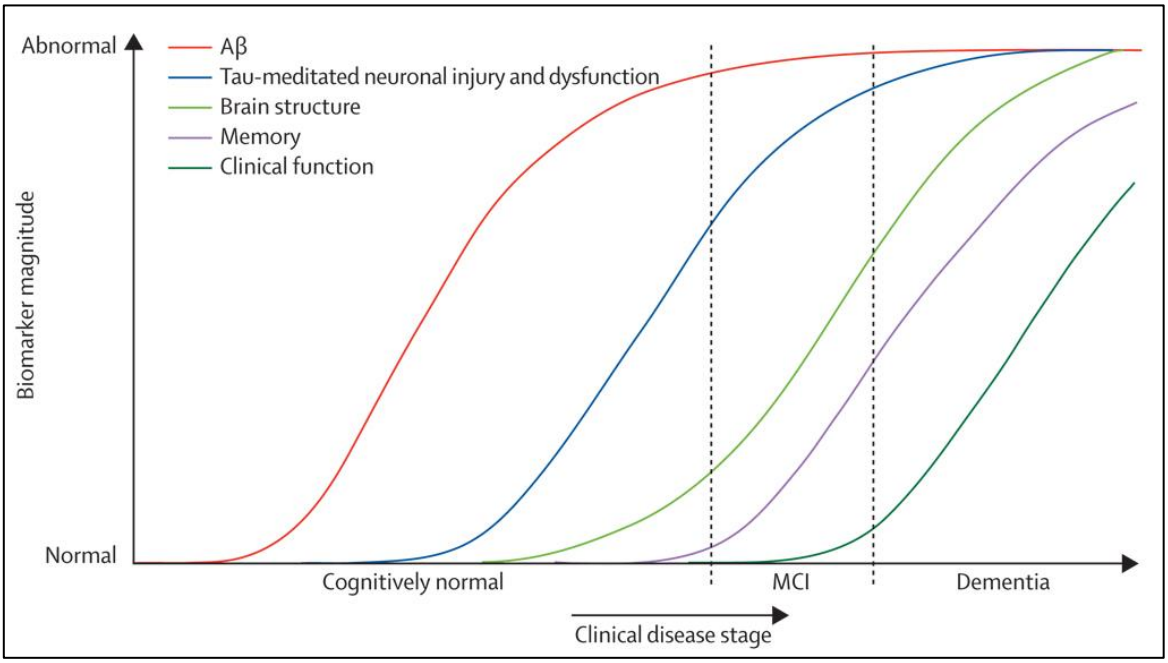



**Figure 1.4** Graph demonstrating the time course (X axis) of amyloid beta (Aβ, red line) and tau mediated effects (blue line) in relation to changes in brain structure (light green line), memory (Purple) and clinical functioning (dark green line) of Alzheimer's disease patients. Image taken from Jack *et al* 2011.[16]

### 1.2.2. Diagnosis

As with many neurodegenerative diseases, the clinical diagnosis is only confirmed upon port-mortem examination of the patient's brain. Although there is yet no cure or disease modifying drug for AD it is still important to diagnose patients with the correct type of dementia, this is especially true for dementia's that are treatable. In research a more accurate diagnosis means that research cohorts are less likely to suffer from noise or misleading results.

The recommended criteria for diagnosing AD is the NINCDS/ADRD (National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's disease and Related Disorders Association), alternative criteria include DSM-IV (Diagnostic and Statistical Manual of Mental Disorders), ICD-10 (International Statistical Classification of Disease and Related Health Problems) and CERAD (Consortium to Establish a Registry for Alzheimer's disease).

Pre-mortem examinations involve taking patient history, physical examination, blood tests, brain scans and questionnaires to access mental abilities. Neuroimaging helps distinguish AD from FTD and DLB, as scans in each have a typical anatomical, functional and molecular signature.[1] Computer tomography (CT) scans provide structural information; magnetic resonance imaging (MRI) provides detailed information about damage to blood vessel and brain atrophy. Single photon-emission computed tomography scan (SPECT) or a positron emission tomography (PET) scans are also used to image specific compounds in the brain, such as amyloid and tau. The imaging tracer PiB (Pittsburgh compound B) is a radioactive analog of thioflavin T which is used to image amyloid deposits in the brain. PiB has been around since 2002 and is the gold standard for imaging amyloid. It wasn't until recently that a tracer for imaging tau was developed; the tau tracer AV1451 has proven successful in recent studies.[17,18]

One of the current drives in AD research is finding a panel of biomarkers with the sensitivity and specificity to identify people who are likely/certain to develop AD, additionally this would ease patient testing, be more cost effective, provide quicker results and allow patients to be monitored through disease progression, which would greatly improve our understanding of the biological changes that occur through different stages of disease. Similarly, research is trying to identify a panel of genetic markers to be used in conjunction with biomarkers, these genetic markers would identify people at risk before physiological changes have begun.

### 1.2.3. Age at Onset

The terminology surrounding AD sub types is extensive. One simple classifier is age at symptom onset (AAO), using this classifier the vast majority of AD are described as late-onset Alzheimer's disease (LOAD) named because of their late onset of clinical symptoms (> 65 years of age); the remaining cases of AD are early-onset Alzheimer's disease (EOAD) which presents early (≤ 65 years of age).

Cases of AD that appear to run in families are often named familial Alzheimer's Disease (FAD), more often than not FAD are caused by highly penetrant (high effect) autosomal dominant variants; cases harbouring these are also known as autosomal dominant Alzheimer's disease (ADAD). Cases of LOAD can also run in families, these cases tend to be described as 'LOAD families'.

LOAD are often referred to as idiopathic and sporadic because the majority have no known single cause and the majority do not run in families. Here we refer to cases of idiopathic EOAD that do not run in families as sporadic early-onset Alzheimer's disease (sEOAD).

For ADAD the AAO is determined in part by the variant itself and thus the gene containing the causative variant (*APP*, *PSEN1* and *PSEN2*), **Figure 1.5** demonstrates the variability of AAO as determined by the gene. The median AAO for all high effect variants is between 45 and 55 years of age, with onset presenting as early as the early-20s or as late as the late-70s.[19] There is evidence that AAO is determined in part by a person's cognitive reserve. Testing cognition in patients with high and low cognitive reserve can create ceiling and floor effects respectively, thus underestimating or overestimating the stage of disease.[1] It is believed that 'pre-morbid intellect' buffers against the diminishing number of neurons, consequently however, once symptoms occur those with a high cognitive reserve endure shorter duration of disease.[20]

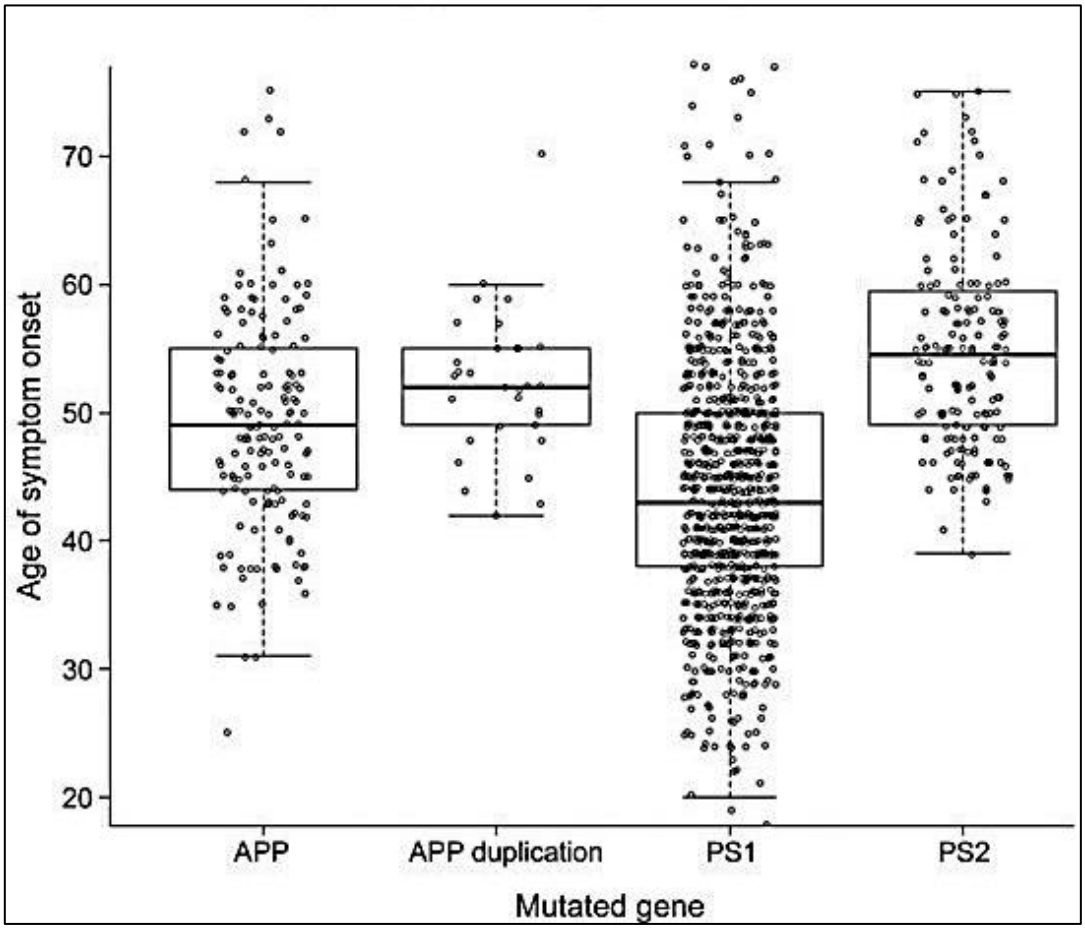



**Figure 1.5** Graph demonstrating the age at onset of Alzheimer's disease based on the gene harbouring a causative variant/duplication; *APP* (APP), *PSEN1* (PS1), *PSEN2* (PS2) or duplications of *APP* (APP duplication). Causative variants in *PSEN1* are the most common and result in a greater variability of onset age. Image taken from Ryman *et al* (2014).[19]

## 1.3. Genetics Linked with Alzheimer's Disease

Complex diseases such as AD are believed to be caused by a spectrum of variants. **Figure 1.6** demonstrates the spectrum of variants based on their effect size and frequency. It is believed that sEOAD and LOAD have a heritability of approximately 92-100% and 70% respectively.[21,22]

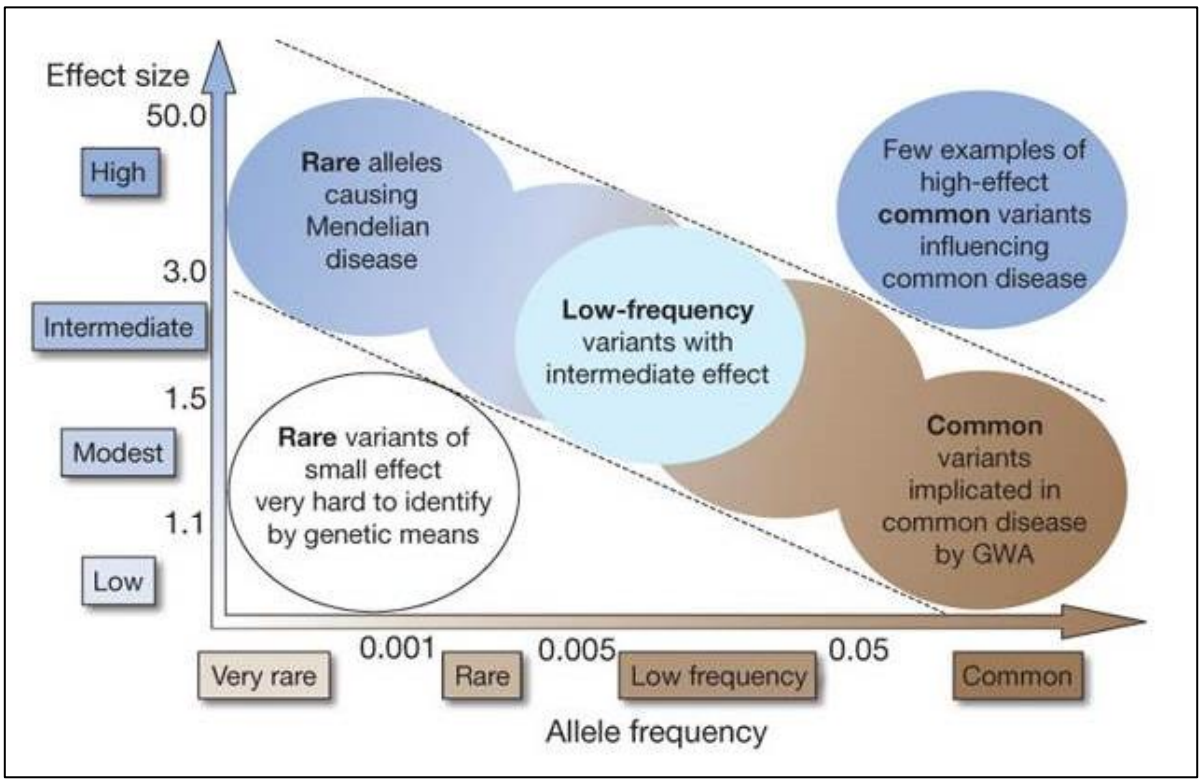



**Figure 1.6** Graph demonstrating the spectrum of all types of variants thought to be involved in the aetiology of complex diseases such as AD. Variants are plotted by effect size in terms of odds ratio on the y axis, and allele frequency on the x axis. Current sequencing practices are able to identify variants within the dotted band. Figure taken from Manolia *et al* (2009).[23]

### 1.3.1. Variants of high effect and rare frequency

It was discovered in the 1990's that single variants with high penetrance cause AD (causative variants).[24-26] Today there are several hundred of these variants documented, most of these variants are located in one of three genes, amyloid precursor protein (*APP*), presenilin 1 (*PSEN1*), presenilin 2 *(PSEN2*), recent sequencing studies have identified possible additional causative variants in progranulin (*PGRN*).[27-29] A causative variant affects either a single base or multiple bases in the form of a polymorphism, insertion or deletion. Variants of multiple bases can be on a large scale, for instance duplications of the entire *APP* gene or deletion(s)/insertion(s) of whole exons. When screening the three main genes, one study found that causative variations in *PSEN1* and *APP* account

for majority of the causative variants, approximately 55% and 13% respectively.[30] The prevalence of causative variants among AD cases is approximately 2%, thus making causative variants rare.

The AD&FTLD mutation database (AD&FTDMDB) and PD mutations database (PDMutDB) records causative variants found in pure AD, FTD, PD, ALS and mixed phenotypes.[31] **Table 1.3** lists the number of variants in the two databases by gene and their pathogenic nature as noted in the databases (accessed June-August 2013).

| Gene | N | Pathogenic Nature Unclear | Causative | Not Pathogenic |
|---|---|---|---|---|
| APP | 40 | 1 | 33 | 6 |
| C9orf72 | 13 | 1 | 1 | 11 |
| CHMP2B | 12 | 4 | 4 | 4 |
| FUS | 45 | 4 | 23 | 18 |
| GRN | 149 | 35 | 69 | 45 |
| LRRK2 | 128 | 68 | 6 | 54 |
| MAPT | 73 | 2 | 44 | 27 |
| PARK2 | 214 | 65 | 127 | 22 |
| PARK7 | 28 | 17 | 6 | 5 |
| PINK1 | 138 | 80 | 28 | 30 |
| PSEN1 | 197 | 8 | 185 | 4 |
| PSEN2 | 25 | 7 | 13 | 5 |
| SNCA | 27 | 1 | 25 | 1 |
| TARDBP | 45 | 2 | 34 | 9 |
| VCP | 20 | 2 | 18 | 0 |
| Total | 1154 | 297 | 616 | 241 |

**Table 1.3** The combined number of variants in the AD&FTDMDB and PDMutDB online databases (accessed June-August 2013). The table tallies the number of variants (N) by the gene they are located within (Gene). These counts are further subdivided in three categories based on the pathogenic nature of the variant: pathogenic nature unclear, causative or not pathogenic.

## 1.3.2. Variants of moderate effect and common frequency

Two single nucleotide polymorphisms (SNPs) in *APOE* were the first common variants to be identified that were significantly associated with AD; the two variants were rs429358 (Cys112Arg) and rs7412 (Arg158Cys). Having a major allele for both SNPs on the same haplotype is called the ε3 allele and is considered to have neutral effect in AD. Alternatively, having a minor allele for one SNP and a major allele for the second SNP on the same haplotype is called ε2 or ε4 allele. The population frequencies of ε2, ε3 and ε4 are estimated to be 8.4%, 77.9% and 13.7% respectively.[32] The ε4 allele frequency is increased to approximately 40% in AD and is considered risky. Interestingly, cases with CAA Type II (which involve capillaries) are four times more likely to have the ε4 allele.[33]

| | | | | **rs7412** | |
|---|---|---|---|---|---|
| | | | | **Arg158Cys** | |
| | | | | **Major** | **Minor** |
| | | | | **C** | **T** |
| **rs429358** | **Cys112Arg** | **Minor** | **C** | ε4 | ε1 |
| | | **Major** | **T** | ε3 | ε2 |

**Table 1.4** Box diagram demonstrating how the *APOE* ε1, ε2, ε3 and ε4 are calculated when alleles for two SNPs (rs7412 and rs429358) are on the same haplotype.

*APOE* ε was initially genotyped using restriction isotyping[34] in a design that also established the haplotype of the alleles, using the genotype for each SNP the *APOE* ε haplotype is calculated using **Table 1.4**. The ε1 haplotype has also been documented but appears to be extremely rare having only been identified in a handful of individuals.[35-37]

Techniques employed today which genotype the two SNPs, for example TaqMan, do not inherently establish haplotype; this is only problematic for individuals that are heterozygous for both SNPs as they can be one of two *APOE* ε statuses (ε3ε1 or ε4ε2), **Figure 1.7** demonstrates this situation, it is assumed that individuals with this genotype have the more common of the two statuses (ε4ε2).

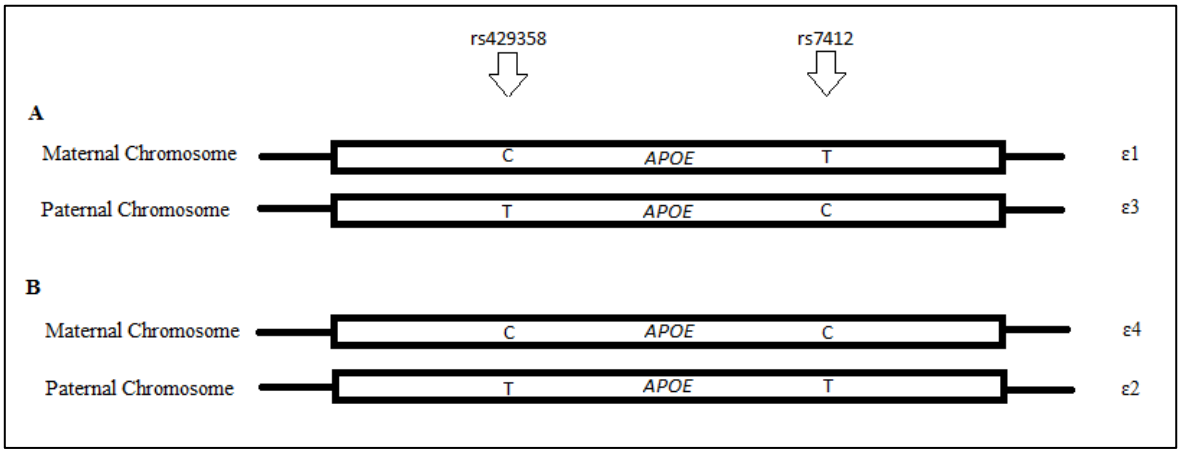



**Figure 1.7** Diagram demonstrating an individuals which are heterozygous (C/T) for the two *APOE* SNPs (rs429358 and rs7412) can have four different haplotypes and thus can have two *APOE* ε statuses; ε3ε1 or ε4ε2 (A or B respectively). The first combination (A) contains the rare ε1 haplotype so the alternative combination (B) is assumed more likely.

The *APOE* locus on chromosome 19 was identified through linkage analysis.[38] Corder and colleagues were the first to establish that the ε4 and ε2 allele were associated with sporadic LOAD and LOAD families with an increased risk[39] and decreased risk respectively.[40] The ε4 allele was also associated with sEOAD on several occasions.[41-46] A seminal meta-analysis that studied all combinations of genotypes was the first to solidify the effect of the alleles whilst taking into account age, gender and ethnicity.[32] **Table 1.5** shows the effect different *APOE* ε status had on the risk of sporadic AD from this study. The study also established that the ε4 allele has an effect on onset age.[32]

| | **ε2** | **ε3** | **ε4** |
|---|---|---|---|
| **ε2** | 0.6 (0.2-2.0) | 0.6 (0.5-0.8) | 2.6 (1.6-4.0) |
| **ε3** | | 1 | 3.2 (2.8-3.8) |
| **ε4** | | | 14.9 (10.8-20.6) |

**Table 1.5** The odds ratio of different *APOE* ε status on developing AD with upper and lower confidence intervals (95%) in brackets.

Since the seminal paper, the association of ε4 and ε2 with AD has been replicated in every AD genome wide association study (GWAS). It is now common practice to adjust genetic association studies for individual ε4 and ε2 dosage, along with age at onset and gender to limit the effect of confounding variables. Today the estimated population attributable fraction (PAF) for the *APOE* ε4 allele is 30.8%.[47]

### 1.3.3. Variants of low effect and common frequency

Before the age of next generation sequencing (NGS), genotyping chips were the high throughput technology of choice. These genotyping chips where designed to genotype all common single nucleotide polymorphisms (SNPs) in the human genome either directly or through proxy SNPs in high linkage disequilibrium with the SNP of interest (tag SNPs), this type of study was termed genome wide association study (GWAS) and identified what are called susceptibility loci. Many GWA studies have been conducted on AD cohorts and nine loci have consistently been associated with AD among these studies, given its high effect the largest signal has always been the *APOE* locus. A combined statistical analysis of these studies has been conducted by AlzForum (link provided in **Section 12**). More recently Lambert and colleagues performed a large GWA meta-analysis study that provided greater power than the initial GWA studies and enabled the identification of new loci, they also performed a replication study to bolster the results and overall the study identified 12 additional new loci.[48] **Table 1.6** lists all 29 variants linked to the 21 loci, overall 20 of the 21 loci have been replicated by the initial GWA studies or the replication stage of the meta-analyses,[48] the *DSG2* locus has not replicated and it thought to be a false positive.

| Locus N | Position (Variant) | Genetic Landscape | Gene Attributed to Signal | GWAS P | GWAS OR | PAF |
|---|---|---|---|---|---|---|
| 1 | 19:1046520 (rs3764650) | Intronic (ABCA7) | ABCA7 | 8.8E-22 | 1.23 (1.18-1.28) | |
| | 19:1063443 (rs4147929) | Intronic (ABCA7) | ABCA7 | 1.1E-15 | 1.15 (1.11-1.19) | 2.8 |
| 2 | 02:127894615 (rs744373) | Intronic (BIN1) | BIN1 | 1.6E-26 | 1.17 (1.13-1.20) | |
| | 02:127892810 (rs6733839) | Intergenic (BIN1 / CYP27C1, bias towards BIN1) | BIN1 | 6.9E-44 | 1.22 (1.18-1.25) | 8.2 |
| 3 | 06:47453378 (rs9349407) | Intronic (CD2AP) | CD2AP | 2.8E-09 | 1.18 (1.08-1.16) | |
| | 06:47487762 (rs10948363) | Intronic (CD2AP) | CD2AP | 5.2E-11 | 1.10 (1.07-1.13) | 2.6 |
| 4 | 19:51727962 (rs3865444) | Intergenic (SIGLEC7 / *CD33) | CD33 | 2.0E-10 | 0.89 (0.86-0.93) | |
| | 19:51727962 (rs3865444) | Intergenic (SIGLEC7 / *CD33, bias towards CD33) | CD33 | 3.0E-06 | 0.94 (0.91-0.96) | 1.8 |
| 5 | 08:27464519 (rs11136000) | Intronic (CLU) | CLU | 3.4E-23 | 0.88 (0.86-0.90) | |
| | 08:27467686 (rs9331896) | Intronic (CLU) | CLU | 2.8E-25 | 0.86 (0.84-0.89) | 5.1 |
| 6 | 01:207784968 (rs3818361) | Intronic (CR1) | CR1 | 4.7E-21 | 1.17 (1.14-1.21) | |
| | 01:207694049 (rs6656401) | Intronic (CR1) | CR1 | 5.7E-24 | 1.18 (1.14-1.22) | 3.5 |
| 7 | 07:143109139 (rs11767557) | lncRNA (EPHA1-AS1) | EPHA1 | N/A | 0.89 (0.83-0.96) | |
| | 07:143110762 (rs11771145) | lncRNA (EPHA1-AS1) | EPHA1 | 1.1E-13 | 0.90 (0.88-0.93) | 3.3 |
| 8 | 11:59939307 (rs610932) | 3’ UTR (MS4A6A) | MS4A6A | 1.8E-11 | 0.90 (0.88-0.93) | |
| | 11:59923508 (rs983392) | Intergenic (MS4A2 / *MS4A6A) | MS4A6A/MS4A4E | 6.1E-16 | 0.90 (0.87-0.92) | 3.8 |
| 9 | 11:85868640 (rs3851179) | Intergenic (*PICALM / EED) | PICALM | 2.9E-20 | 0.88 (0.86-0.90) | |
| | 11:85867875 (rs10792832) | Intergenic (*PICALM / EED, bias towards PICALM) | PICALM | 9.3E-26 | 0.87 (0.85-0.89) | 4.5 |
| 10 | 20:55018260 (rs7274581) | Intronic (CASS4) | CASS4 | 2.5E-08 | 0.88 (0.84-0.92) | 1.0 |
| 11 | 11:47557871 (rs10838725) | Intronic (CELF1) | CELF1 | 1.1E-08 | 1.08 (1.05-1.11) | 2.5 |
| 12 | 18:29088958 (rs8093731) | Intronic (DSG2) | DSG2 | 1.0E-04 | 0.73 (0.62-0.86) | |
| 13 | 14:53400629 (rs17125944) | Intronic (FERMT2) | FERMT2 | 7.9E-09 | 1.14 (1.09-1.19) | 1.2 |
| 14 | 07:37841534 (rs2718058) | Intronic (GPR141/EPDR1) | NME8 | 4.8E-09 | 0.93 (0.90-0.95) | 2.5 |
| 15 | 06:32578530 (rs9271192) | Intergenic (*HLA-DRB1 / HLA-DQA1) | HLA-DRB1 / HLA-DRB5 | 2.9E-12 | 1.11 (1.08-1.15) | 3.0 |
| 16 | 02:234068476 (rs35349669) | Intronic (INPP5D) | INPP5D | 3.2E-08 | 1.08 (1.05-1.11) | 3.8 |
| 17 | 05:88223420 (rs190982) | Intronic (MEF2C-AS1) | MEF2C | 3.2E-08 | 0.93 (0.90-0.95) | 2.8 |
| 18 | 08:27195121 (rs28834970) | Intronic (PTK2B) | PTK2B | 7.4E-14 | 1.10 (1.08-1.13) | 3.6 |
| 19 | 14:92926952 (rs10498633) | Intronic (SLC24A4) | SLC24A4 / RIN3 | 5.5E-9 | 0.91 (0.88-0.94) | 1.9 |
| 20 | 11:121435587 (rs11218343) | Intronic (SORL1) | SORL1 | 9.7E-15 | 0.77 (0.72-0.82) | 0.9 |
| 21 | 07:100004446 (rs1476679) | Intronic (ZCWPW1) | ZCWPW1 | 5.6E-10 | 0.91 (0.89-0.94) | 2.5 |

**Table 1.6** Variants highlighted from GWA studies, including studies before the meta-analysis highlighted in blue (from AlzGene, link provided in **Section 12**) and from the meta-analysis study highlighted in yellow.[48] The table lists the locus number (Locus N) along with the genomic position and variant reference ID of the GWAS variant (Position (Variant)), the location of the variant relative to the nearest coding gene(s) according to Ensembl/Havana merged annotation, asterix indicates the nearest gene if the variant is intergenic (Genetic Landscape), and the coding protein the association signal has been attributed to (Gene Attributed to Signal). Data from the GWA studies is given in the last two columns and includes the p-value (GWA P), the odds ratio with 95% confidence interval in brackets (GWA OR) and the population attributed factor in percentage (PAF).[47] Loci highlighted in green have been replicated and those highlighted in red have not. All information is based on build GRCh37. Key: N/A, not available; OR, odds ratio; P, p-value; GWAS, genome wide association study; PAF, population attributed factor.

All of the GWAS variants have been attributed to coding genes, there are several assumptions made by doing this, SNPs can have long range of effect and there is no denying that there could be physical properties of the human genome that have created the associations that we do not yet recognise. Despite this caution there appears to be validity to the attributed gene; additional SNPs that are in the same attributed gene and are in LD with the primary associated SNP show a higher than average significance, this effect can be clearly seen in the *CR1* locus **Figure 1.8A**, **Figure 1.8B** shows the absence of this effect in the *DSG2* locus; thus providing further evidence that the *DSG2* locus was a false positive.

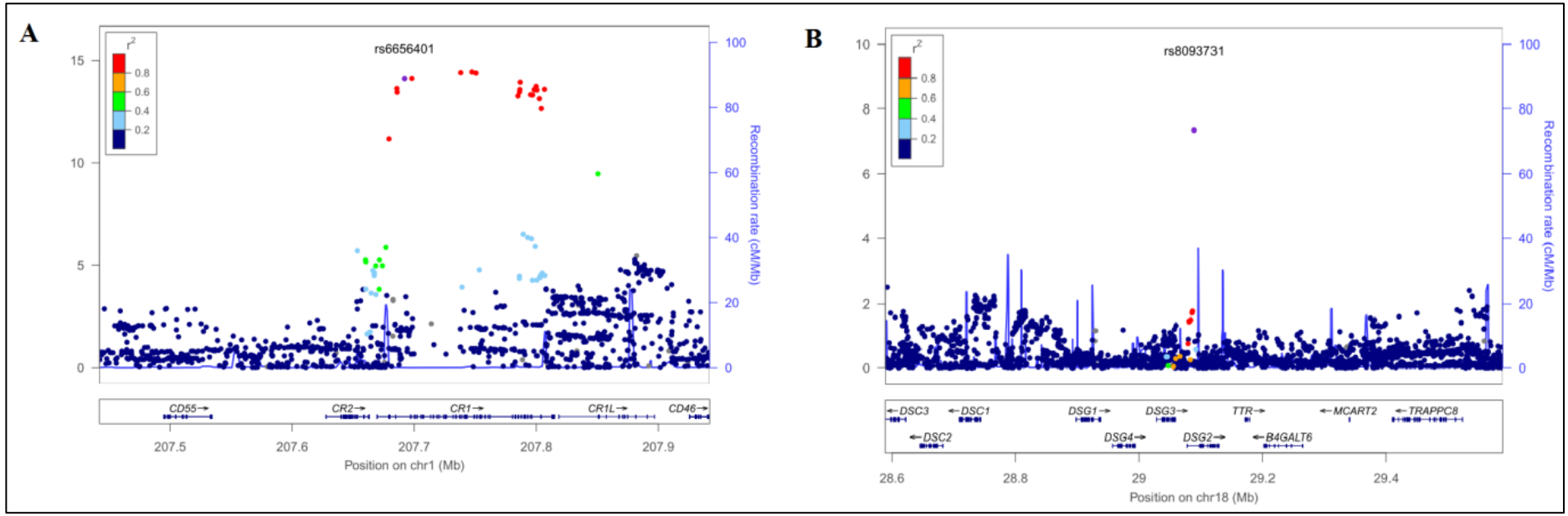



**Figure 1.8** This image includes supplementary figures 2A (A) and 7b (B) taken from the meta-analysis study[48] which plots the SNPs used in the study for the *CR1* and *DSG2* locus respectively. Figure A shows that the SNPs in linkage disequilibrium (dots in red) with the primary *CR1* SNP (dot in purple) show higher than average significance. Figure B shows the SNPs in linkage disequilibrium (dots in red) with the primary *DSG2* SNP (dot in purple) do not show higher than average significance.

### 1.3.4. Variants of moderate effect and low frequency

The number of samples required to identify variants of low frequency and intermediate effect has until recently hindered the ability to screen for these variants, as next generation sequencing technology was neither high throughput nor affordable. This has now changed and next generation sequencing is fast becoming higher throughput and more cost effective. Given the number of case samples required (tens of thousands) for 80% power to identify low frequency variants of intermediate effect with disease, research into such variants has only been conducted in sLOAD cases to date. Whole exome/genome sequencing has identified three such variants which show significant association with sLOAD. The three variants include a protective variant in the *APP* gene,[49] a risk variant in *TREM2* (Triggering Receptor Expressed On Myeloid Cells 2)[50,51] and a controversial risk variant in *PLD3* (Phospholipase D Family, Member 3).[52]

Thorlakur Jonsson and colleagues identified a protective variant (rs63750847, Ala673Thr) in *APP* which was significantly associated with sLOAD of Icelandic decent (p.value=6.92E-6 OR=0.13). This variant was seen to be significantly associated with cognitive performance in the general population (p.value 0.0021). The study also identified three individuals over the age of 65 years who harboured two alleles of this variant, they were found to be none demented suggesting this variant may confer protection in a dose dependent fashion.

Rita Guerreiro and colleagues identified a risk variant (rs75932628, Arg47His) in *TREM2* which was significantly associated with sLOAD of European and North American decent (p.value=9.0E-09, OR=5.05 (95% CI 2.77-9.16)). [50] This association was replicated by Thorlakur Jonsson and colleagues in a second study with sLOAD of Icelandic decent (p.value=3.42E-10, OR=2.92 (95% CI 2.09-4.09)) with a more significant p.value. [51] Having two alleles of other *TREM2* variants appear to be highly penetrant, for example individuals with two alleles of one of the three variants T66M, Y38C or Q33X present with early-onset frontotemporal dementia with leukodytrophy,[53] and individuals with two alleles of the variant Q33X can also present with early-onset dementia presenting with Nasu-Hakola.[54] All three variants combined (T66M, Y38C and Q33X) were found to be significantly associated with sLOAD (p.value 0.01).[50] TREM2 is a membrane protein that forms a complex which supresses the immune response and activates phagocytic pathway in macrophages and dendritic cells.[54] Compromised function of TREM2 could lead to increased death of neurons through increased systemic inflammatory response and/or affect the clearance of cell debris and possibly the removal of beta amyloid.[50]

Carlos Cruchaga and colleagues identified a risk variant (rs145999145, Val232Met) in *PLD3* which was significantly associated with sLOAD of European American descent (p.value=2.93E-5, OR=2.10 (95% CI 1.47–2.99)), it was also associated with LOAD families of European American descent (p.value=1.18E-6 OR=3.39 (95% CI 2.14–5.39)). This study also performed gene burden analysis

using 20 variants (including Val232Met) identified in a separate European cohort and found the gene to be significantly associated (p.value=1.44E-11 OR=2.75 (95% CI 2.05–3.68)), the gene was still significantly associated when variant Val232Met was removed (p.value=1.5E-8 OR=2.58 (95% CI 1.87–3.57)). Two other variants in *PLD3* could explain this association as they were both nominally associated with sLOAD; Met6Arg (p.value=0.02, OR=7.73 (95% CI 1.09–61)) and Ala442Ala (p.value=3.78E-7, OR=2.12 (95% CI 1.58–2.83)), Ala442Ala was found to reduce the levels of *PLD3* mRNA and transcripts containing exon 11. Functional work in this study demonstrated that lower levels of PLD3 are correlated with higher APP/Aβ levels and with more extensive AD pathology, suggesting PLD3 might be involved in APP processing.[52]

## 1.4. Metabolic Pathways Linked with Alzheimer's Disease

Research has identified many candidate genes, although looking at these genes in an isolated fashion has not thus far explained their link with disease; however, these genes appear to fall into particular pathways that make biological sense. Described briefly in this chapter are the main pathways implicated in the disease, together with the most notable supporting evidence.

### 1.4.1. Amyloid

One of the first pathways to be identified was the amyloid cascade, this was a direct result of identifying high levels of Aβ in senile plaques of AD brains.[55] The amyloid pathway is centralized around the amyloid precursor protein (APP) which is a ubiquitous expressed type 1 transmembrane protein coded by *APP*. APP is processed at several sites in the cell, including the cell surface, endosomal/lysosomal organelles and mitochondrial membrane.[56-58] APP undergoes proteolytic processing in a sequential manner to form three constituent peptides. Two processes are known to occur; the non-amyloidogenic process occurs primarily at the cell surface and the amyloidogenic process in the endocytic organelles,[59] both are depicted in **Figure 1.9**. The non-amyloidogenic sequence begins with cleavage of APP by the α-Secretase which releases the soluble APP alpha cleaved ectodomain (sAPPα) into the extracellular space, the residual 83 residue peptide (C83) is cleaved by the γ-secretase enzyme which releases the p3 peptide extracellularly and the amyloid intracellular domain (AAID) intracellularly. The amyloidogenic sequence is involved in the aetiology of AD and begins with cleavage of APP by the BACE1 enzyme which releases the soluble APP beta cleaved ectodomain (sAPPβ) into the extracellular space, the residual 99 residue peptide (C99) is cleaved by the γ-secretase enzyme which releases the amyloid beta peptide (Aβ) extracellularly and the amyloid intracellular domain (AICD) intracellularly (**Figure 1.9**). Aβ can have a length of 36 to 43 amino acids and has a property that allows it to self-aggregate, thus it can exist in three primary states; monomers, oligomers and fibrils (**Figure 1.10**).

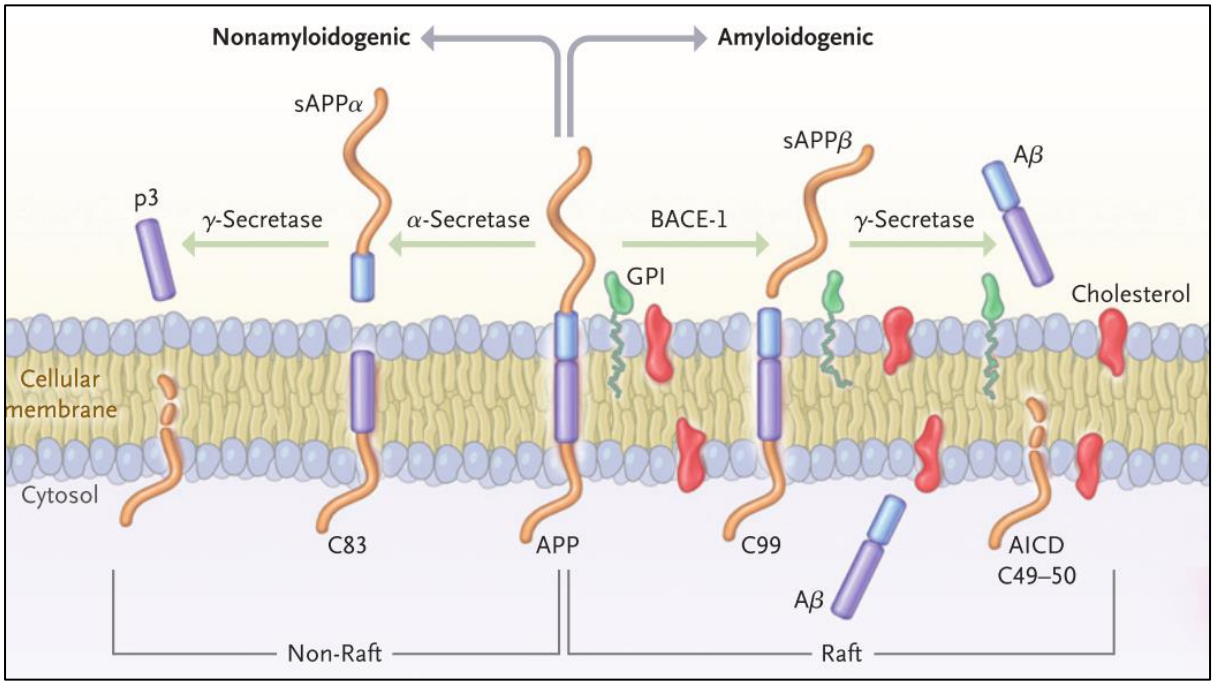



**Figure 1.9** The processing of amyloid precursor protein (APP) in the cell membrane. Two processes occur naturally in the cell; the non-amyloidogenic (left) and the amyloidogenic (right), which results in production of the amyloid beta (Aβ) peptide and is thought to be involved in the aetiology of Alzheimer's disease. The amyloidogenic process occurs on lipid rafts that are tightly packed membrane microenvironments enriched in sphingomyelin, cholesterol, and glycophosphatidylinositol (GPI)–anchored proteins. Image taken from Querfurth *et al* (2010) and modified.[60]

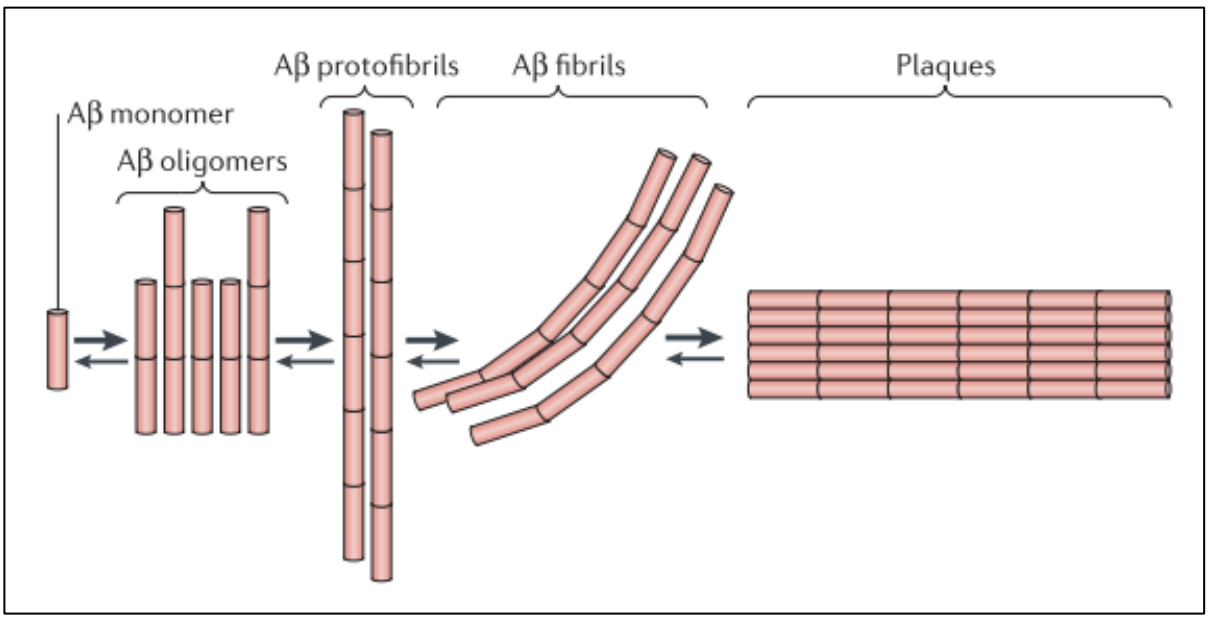



**Figure 1.10** The different states of amyloid beta (Aβ) protein, including monomer, protofibrils and fibrils, Aβ exists in a dynamic continuum of these states. Image taken from Heppner *et al* (2015).[61]

Finding causative variants in *APP*, *PSEN1* and *PSEN2* gave solidity to the involvement of the amyloid pathway in the aetiology of AD. APP is the precursor to Aβ and PSEN1/PSEN2 form the catalytic domain of the γ-secretase complex. Extensive evidence suggests that causative variants affect the cleavage site of γ-secretase which results in alteration of total Aβ levels and/or the ratio of $A\beta_{42}/A\beta_{40}$. Additionally, individuals with Downs Syndrome who have an extra copy of APP (APP is located on chromosome 21) tend to develop AD, even at a young age. Different states of Aβ have differing levels of toxicity; fibrils are the least toxic and $A\beta_{42}$ oligomers are the most neurotoxic.[62]

Aβ is cleared from the brain parenchyma through the interstitial fluid and the Virchow–Robin perivascular space. There is a 30% decreased clearance of $A\beta_{40}$ and $A\beta_{42}$ in AD cerebral spinal fluid (CSF) [63] compared to controls, which over 10 years is estimated to reach the levels of Aβ that are seen in AD brain. Additionally Aβ can be proteolytically cleaved; Aβ fibrils are largely resistant to this, however the soluble forms of Aβ are degraded by neprilysin (Nep) and insulin-degrading enzyme (IDE), it is also degraded by a lesser degree by endothelin converting enzyme 1 (ECE1), angiotensin converting enzyme (ACE), plasmin, matrix metalloprotease 9 (MMP9) and presequence peptidase (PreP). Nep is responsible for external degradation of peptides[64] and IDE is responsible for both external and internal degradation of peptides, including insulin.[65] Interestingly insulin acts to inhibit IDE from acting on Aβ due to its greater affinity for IDE[66] and insulin resistance increases the risk of developing AD.[67] Mouse models also suggest Aβ is degraded by protease Cystatin C (CST3) regulated by Cathepsin B (CTSB)[68] and variants in *CST3* have been significantly associated with AD.[69] Additionally, Aβ leads to the activation of caspase, which can cleave critical cellular proteins such as tau to initiate apoptosis.[70]

## 1.4.2. Tau

The second pathway to be suspected was that of tau, this was a direct result of identifying high levels of intracellular hyper-phosphorylated tau deposits in AD brain,[71,72] which are a histological hallmark of AD.

The cell cytoskeleton is made up of filaments containing tubulin (microtubule filaments), actin (actin filaments) or other subunits (intermediate filaments) (**Figure 1.11**). The cytoskeleton maintains cell shape, organisation, division and movement. The largest filaments are microtubules which are made up of α-tubulin and β-tubulin that dimerise to form profilaments, which then polymerise to form microtubules. Polymerization occurs fastest at the plus end, the minus end is anchored to microtubule organising centres (MOTs). Kinesin, dynein and myosin are proteins that move along microtubules (motor proteins) and transport cargo around the cell. Transport along microtubules can be long and short range, and in both directions; towards the soma (retrograde) is mainly motored by dynein protein and away from the soma (enterograde) is mainly motored by kinesin protein but which also has the ability to move retrograde. Tau is a microtubule (MT) associated protein coded by *MAPT* and like motor proteins is able to bind microtubules. The hydrophilic nature of tau results in a loose unfolded structure which has hindered attempts to identify exactly where tau binds to microtubules, saturation suggests that the ratio of tau to tubulin αβ-heterodimer is approximately 0.5.[73] Tau binds tubulin and directly promotes MT assembly through phosphorylation at tau's serine (S) and threonine (T) residues at Serine-Proline (SP) and Threonine-Proline (TP) motifs, kinases such as GSK-3β, CDK5 or MAPK phosphorylate tau at these motifs.[74,75] Tau also stabilizes microtubules by maintaining spacing between microtubules.[76,77] Indirectly tau generally inhibits motor proteins such as kinesin from binding to microtubules by providing competition for binding sites,[78-80] dynein tends to remain bound but reverses in direction when in the presence of tau due to its lower sensitivity.[78] Tau has several isoforms and the major isoforms are a result of two major differences; the number of a-terminal units (abbreviated as N) which are 29 amino acids in length and coded by exons 2 or 3, and secondly the number of repeat domains (abbreviated as R) at the c-terminal end which bind microtubules. The number of R domains in tau enhances binding to microtubules. Tau undergoes many types of modifications including proteolysis, glycosylation and phosphorylation; the latter obviously occurs in abundance in AD brain and as such it has been the most studied. Phosphorylation acts in very site specific manner; phosphorylation at sites within the R domains have a stronger inhibitory effect than phosphorylation sites outside of the R domains.[81] It is assumed that hyper-phosphorylation causes tau to self-associate and aggregate, however the evidence it not clear, some hyper-phosphorylated states appear to promote aggregation,[82] while certain sites seem to protect against aggregation.[83]

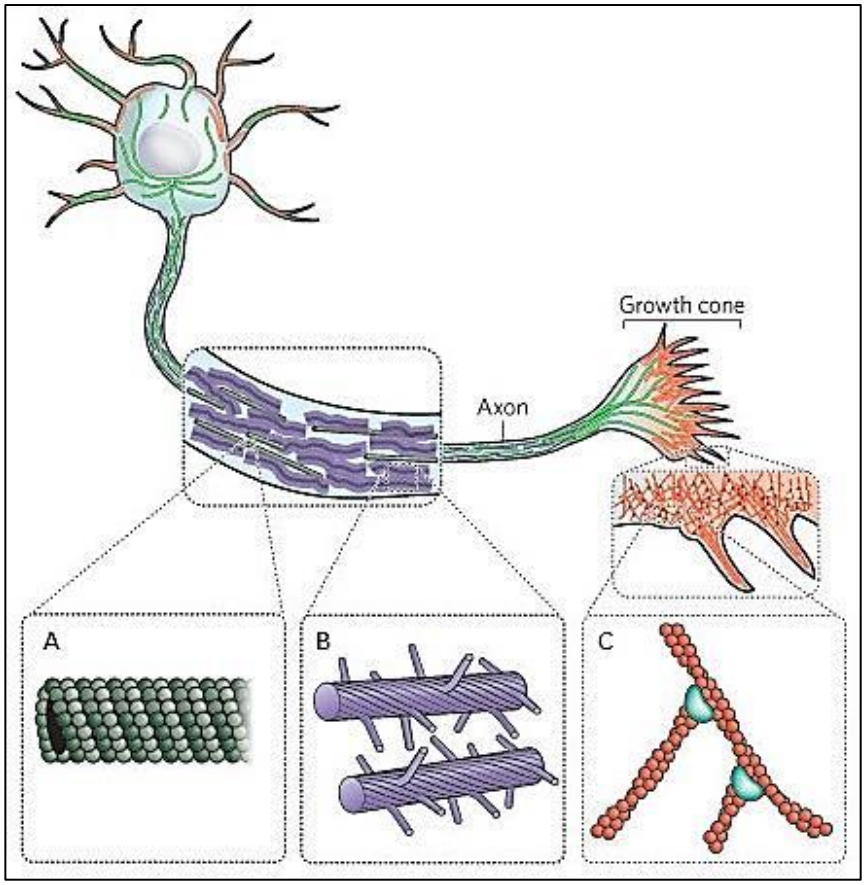



**Figure 1.11** Image shows the three types of polymer present in the cytoskeleton of neurons; microtubules (A), intermediate filaments (B) and actin filaments (C). Image taken from Fletcher *et al* (2010) and modified.[84]

The best evidence that the tau pathway is linked to AD are the autosomal dominant variants in *MAPT* that cause frontotemporal dementia and parkinsonism liked to chromosome-17 (FTD-17),[85] these variants inhibit the ability of tau to bind to microtubules resulting in a moderate decrease in stabilisation but dramatic increase in formation into paired helical filaments (PHFs). PHFs are the precursor to neurofibrillary tangles (NFTs),[86] which are the neuropathological feature that best correlates with cognitive decline in AD,[87] intermediate aggregates of tau are more cytotoxic[88] as is seen for Aβ.

Many other diseases are linked to axonal transport,[89] including some neurodegenerative diseases such as amyotrophic lateral sclerosis (ALS) and Spinal and Bulbar Muscular Atrophy (SBMA) and Huntington's disease. The tau pathway is linked to the amyloid pathway via APP deposition; vesicles containing APP are transported along microtubules via kinesin[79,90] to where Aβ is released.[91] Disruption of anterograde transport in AD leads to vesicles, organelles (e.g. mitochondria) and kinesin building up and causing axonal blockage well before amyloid deposition and onset of symptoms, it is hypothesised that longer exposure of APP to γ-secretase may lead to increased amyloidogenic turnover.

## 1.4.3. Cholesterol

Cholesterol is an important constituent of cell membranes and an important precursor of the hedgehog protein and steroid hormones. Cholesterol homeostasis is managed by an elaborate system in the human body which is supplied through dietary and hepatic means. Some of this cholesterol does reach the brain via high density lipoproteins (HDLs); however the presence of the blood-brain barrier (BBB) limits this distribution and as such cholesterol homeostasis in the brain is managed via *de novo* synthesis, intracellular redistribution and excretion to maintain strict levels. The brain contains 25% of the bodies' cholesterol; the majority of this cholesterol is on myelin sheaths, in neuron/astrocyte plasma membranes and endocytic recycling compartments.

Several cell types in the brain synthesise cholesterol to different levels, astrocytes synthesise the most, followed by microglia and neurons. Astrocytes are seen as the producer of cholesterol and neurons the receivers.[92] Astrocytes excrete cholesterol via the ATP-binding cassette transporter (ABCA1) and the primary cholesterol lipoprotein APOE mediates cellular uptake of the cholesterol via low-density lipoprotein receptors (LDLR), low density related protein 1 (LRP1) and heparin sulphate proteoglycan (HSPG).[93] Neurons and astrocytes require large amounts of cholesterol to make up their extensive membrane area; evidence suggests it is essential for axonal growth[94] and synaptic formation.[95,96]

Several lines of evidence support a link between cholesterol homeostasis in the brain and AD aetiology. Most notably several LOAD susceptibility loci are located within genes thought to be involved in cholesterol metabolism in the brain; these include *APOE*, *ABCA7* and *CLU (*also called *APOJ*). Evidently, the use of lipid lowering medications such as statins reduces the risk of AD.[97] Additionally, diseases affecting cholesterol homeostasis can lead to neurologic disorders; Smith–Lemli–Opitz syndrome is a disease that causes reduced levels of cholesterol and a build-up of cholesterol precursors which leads to mental retardation and reduced myelination.[98] Niemann-Pick type C is a disorder that causes depletion of the endosome and lysosome membrane proteins NPC1 or NPC2 leading to over accumulation of cholesterol and glycosphingolipids in late endosomal/lysosomal compartments, the disease presents with neurofibrillary tangles (NFTs) and results in dementia.[99]

There are also links between cholesterol and the amyloid/tau pathways. APOE binds Aβ (**Figure 1.12**) in an isoform specific manner.[100] Additionally, plasma membrane cholesterol is concentrated in sphingolipid islands (rafts) where amyloidogenic processing takes place.[101] Cholesterol directly regulates gamma secretase and reducing membrane cholesterol levels resulting in reduced Aβ generation. [102,103] APOE is found in senile plaques and NFTs of AD brain,[104,105] and lipid granules are a neuropathological feature of AD.[106] Cholesterol can also affect Aβ mediated tau proteolysis by the calcium dependent enzyme calpain[107] and also promote tau phosphorylation.[108]

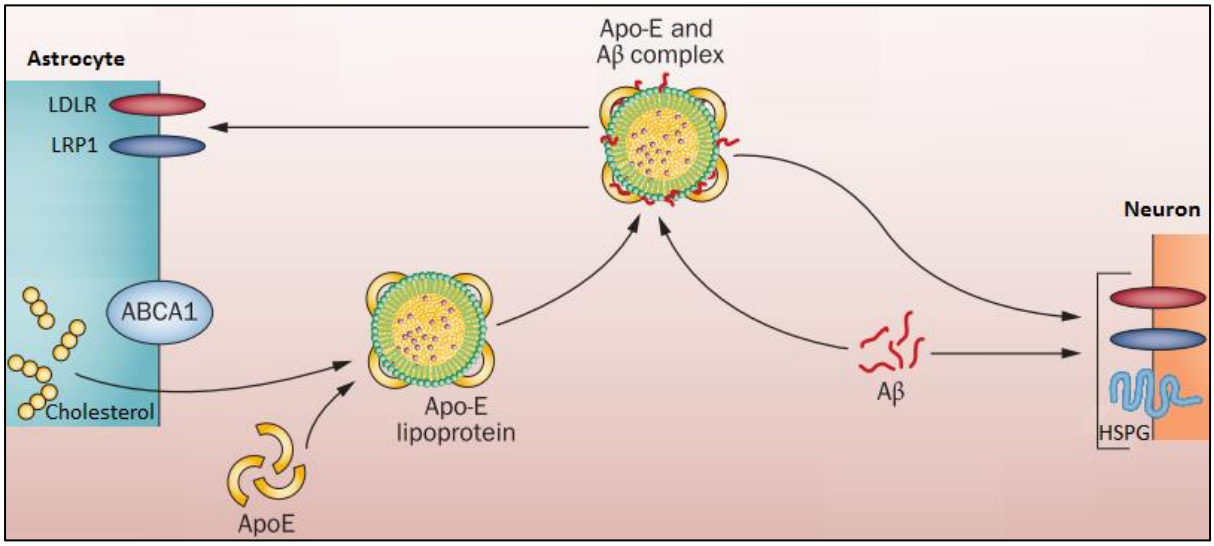



**Figure 1.12** Depiction of the physiological link between APOE and Aβ. Cholesterol is catalysed in astrocytes and excreted via ABCA1; it binds to APOE to form a lipoprotein before binding Aβ. Key: LRP1, low density related protein 1; LDLR, low-density lipoprotein receptor; Aβ, amyloid beta; ABCA1, ATP-binding cassette transporter; HSPG, heparin sulphate proteoglycan. Image taken from Liu *et al* (2013) and modified.[93]

## 1.4.4. Intracellular trafficking

Intracellular cargo can derive from intracellular processes or extracellular processes before being invaginated. Cargo are transported around the cell in vesicles either for utilization, recycling or excretion. Vesicles originate in the endoplasmic reticulum (ER) where the membrane lipids and proteins originate, the vesicles bud from the ER and then make their way through the several sacs of the Golgi apparatus (GA) via fission which starts at the cis-golgi and finishing at the trans-golgi, this movement it directed via microtubules and labels vesicles with address tags. The vesicles bud from the trans-golgi where they are directed to their appropriate destination. Vesicles can be endosomes, lysosomes which contain hydrolases that degrade macromolecules, phagosomes which contain microorganism particles or autophagosomes that contain cellular debris. Mature endosomes are termed late endosomes, which along with autophagosomes and phagosomes can fuse with lysosomes and/or excrete their contents from the cell in processes called exocytosis. Vesicles can also invaginate extracellular cargo from the plasma membrane in processes called endocytosis and phagocytosis. **Figure 1.13** illustrates these trafficking systems.

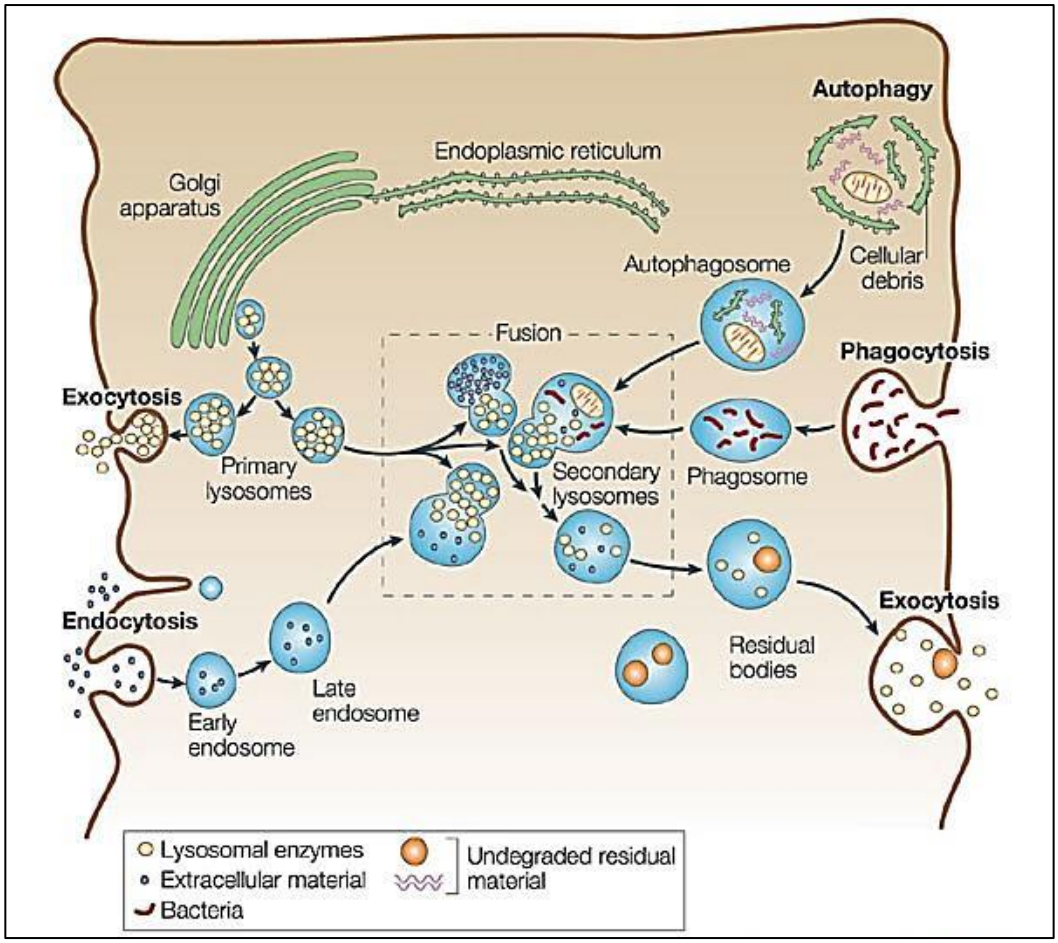



**Figure 1.13** Diagram illustrating the different systems utilised by the cell to traffic items around, in and out of the cell in vesicles. Image was taken from Desnick *et al* (2002).[109]

Endosomes, in particular those in neurons, travel long distances to reach dendrite and axon tips; up to one meter for some motor neurons. Endosomes transport cargo long distances using microtubules either towards the tip (enterograde) or towards the soma (retrograde). RAB (Ras-associated binding) proteins play an important role in vesicle trafficking, they bind proteins called effectors that facilitate different trafficking events such as budding, tethering and fusion. For example RAB5A can regulate early endosome trafficking along microtubules.[110]

Many neurologic disorders are associated with intracellular trafficking,[111] these include Schizophrenia and Amyotrophic Lateral Sclerosis (ALS). Schizophrenia is associated with variants in *DTNBP1* (dysbindin) which forms part of the BLOC1 complex,[112-115] the BLOC1 complex is needed to form recycling endosomes.[114] Amyloid Lateral Sclerosis (ALS) can be caused by highly penetrant variants in *SOD1* (Superoxide Dismutase 1),[116,117] SOD1 and ALS2 (alsin) control activation of RAB5A and recruitment of RAB5A to endosomes.[116,118]

The endosomal/lysosomal (EL) system is highly utilised in the amyloid pathway; APP, BACE1, C99 and Aβ are all moved around the cell and/or invaginated using this system,[119] additionally PSEN1 has

an ability to regulate lysosomal function.[120] Highlighting the importance of intracellular activities of Aβ are several lines of evidence in Down syndrome and transgenic mice that suggest $A\beta_{42}$ accumulates intracellularly before forming extracellular deposits,[121-131] and it accumulates within structurally damaged mitochondria of AD brains.[132] Furthermore, RAB5 and RAB7 are responsible for endocytosis of Aβ and are upregulated in AD brain.[133]

Several LOAD susceptibility loci are involved in vesicle trafficking, these include *EPHA1* (Erythropoietin-Producing Hepatoma Receptor A1), *BIN1* (Bridging Integrator 1), *CD2AP* (CD2-Associated Protein), *SORL1* (Sortilin-Related Receptor L) and *PICALM* (Phosphatidylinositol Binding Clathrin Assembly Protein).

## 1.4.5. Oxidative stress and metal ions

Reactive species (RS) tend to be reactive because they are free radicals, these are atoms or molecules capable of independent existence that have at least one unpaired valence electron ($^{\bullet}$). Free radicals are everywhere, in the environment they originate from natural sources such as cosmic radiation or manmade sources such as pollutants, the simplest free radicals are diatomic oxygen ($O_2^{\bullet}$) which has two unpaired electrons, and hydrogen (H) which has one unpaired electron. RS are also naturally generated in the human body, most of which are non-radicals. RS in the human body have important functions, including gene transcription, regulating guanylate cyclase activity, killing bacteria in macrophages and communication between neurons and glia.

Reactive species (RS) can be subdivided, the most studied biological RS are the reactive oxygen species (ROS), followed by nitrogen (RNS), bromine (RBS), chlorine (RCS) and sulphur (RSS). Reactive oxygen species (ROS) are derived from oxygen. Examples of ROS include superoxide ($O_2^{\bullet-}$), hydroxyl ($HO^{\bullet}$), nitric oxide ($NO^{\bullet}$), singlet diatomic oxygen ($^1O_2$), hydrogen peroxide ($H_2O_2$), peroxynitrite anion ($ONOO^-$) and hypochlorous acid (HOCl). Reactive nitrogen species (RNS) contain a nitrogen atom and are derived from nitric oxide and superoxide. Examples of RNS include nitrogen dioxide ($NO_2^{\bullet}$) and dinitrogen trioxide ($N_2O_3$). In biology ROS and RNS are the most studied because of their high utility in life. ROS are formed at several locations in the cell, including the endoplasmic reticulum (ER), cytosol, plasma membrane surface and mitochondria. Most ROS are produced by enzymes Complex I and Complex III in mitochondria. Metal ions are an important constituent of these oxidizing enzymes due to their ability to transfer single electrons. Despite their beneficial use in biological systems, RS can have unwanted side effects and a balance between RS generation and antioxidant defence moderates these unwanted effects.

Of all the organs in the human body, the brain is particularly susceptible to oxidative stress by ROS for several reasons; the brain has the highest consumption of oxygen for its size (20% of the basal oxygen we breath), this is needed to open and close ion channels which generate action potentials and neuro-secretions, for example the $Ca^{2+}$ trafficking across neuronal membranes. Additionally, the brain

is rich in lipids, redox-active metals, excitotoxic amino acids, autoxidizable neurotransmitters and low in antioxidant defences.[134,135]

Oxidative stress was one of the first pathways identified in AD. In 1986 research highlighted this pathway when it was found that the enzymes glucose-6-phosphate dehydrogenase (G6PD) and 6-phosphogluconate dehydrogenase (6PGD) had increased activity in AD brains potentially as a response to increased peroxide levels.[136] Following this, further research identified modifications to lipids,[137,138] nuclear DNA,[139] mitochondrial DNA[140] and proteins[141,142] as a result of ROS, all of these modification were found to be elevated in AD brain and occur before pathological changes.[143] These modifications are now seen as markers for ROS. **Figure 1.14** depicts the modifications made to lipids, DNA and proteins on a molecular level.

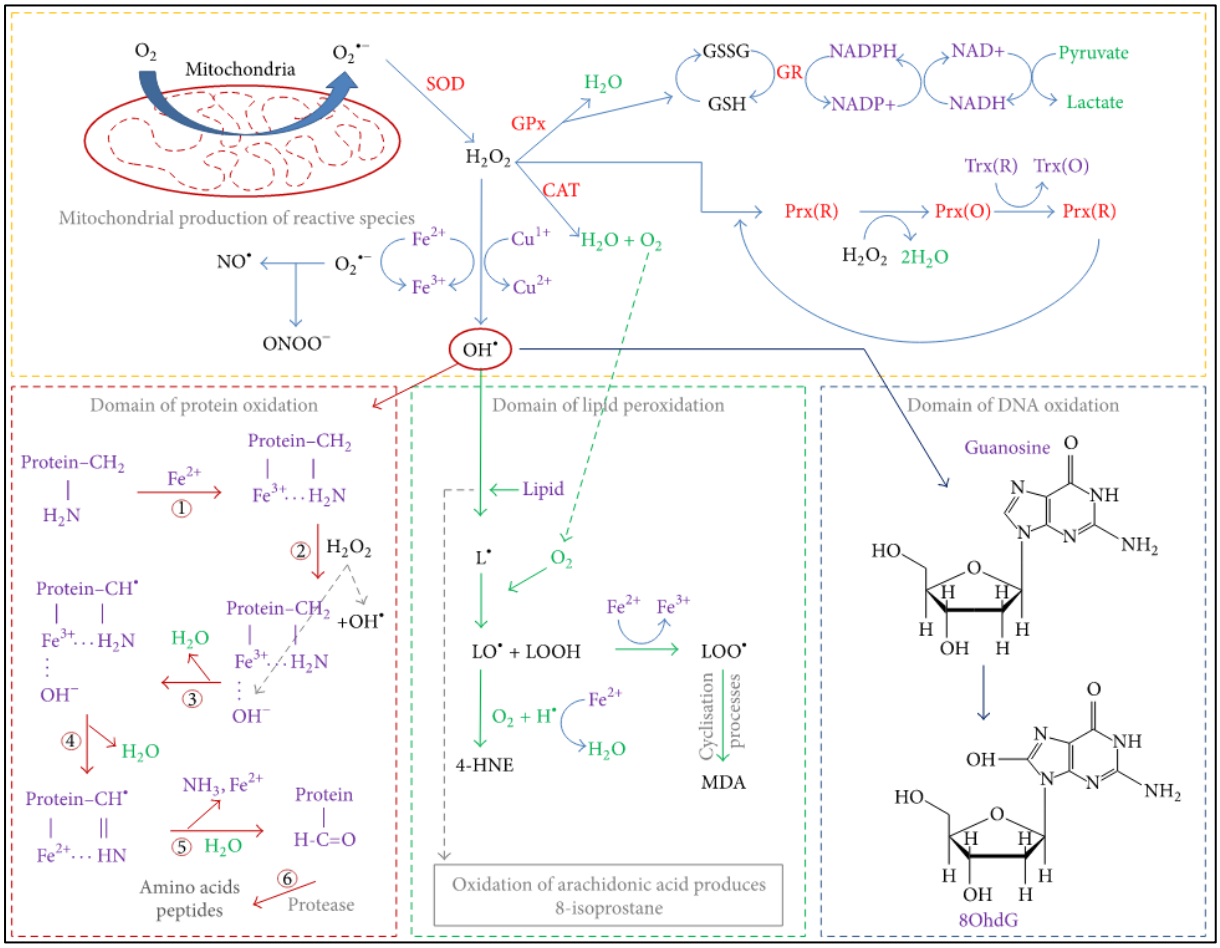



**Figure 1.14** Flow chart demonstrating the production of reactive species and the endogenous antioxidant system. Red: enzymes; green: other products; purple: cofactor/substrate; black: reactive species. Image taken from Fraunberger *et al* (2016).[144]

In terms of lipid peroxidation by-products, most research has focused on the by-product 4-hydroxynonenal (HNE), it can alter the phospholipid asymmetry of the membrane lipid bilayer, leading to apoptotic neuronal loss in both MCI and AD.[145] Additionally HNE and isoprostanes can lead to tau phosphorylation/ aggregation in cultured rat hippocampal neurons,[146] HNE can also reduce the uptake of glutamate and transport of GLUT3-mediated glucose in synaptosomes in cultured rat nerve cells.[147,148] LDL receptor-related protein 1 (LRP-1) is covalently modified by HNE.[149] The mitochondrial enzymes Complex IV (cytochrome oxidase/COX), α-ketoglutarate and pyruvate dehydrogenase have a reduced activity in AD brains.[150]

Aβ accentuates oxidative stress and metal ions appear to mediate this.[151] Zinc, iron and copper ions can bind Aβ to form complexes,[152] for example Aβ bound by zinc can enter lipid bilayers and form cation-selective channels that conduct $Ca^{2+}$ and other cations.[153,154] Aβ bound by copper generates hydrogen peroxide, a reaction that requires cholesterol, vitamin C, L-DOPA, DOPA and dopamine and is inhibited by zinc ions,[155] low levels of zinc ions have an antioxidant effect and high levels are oxidizing as they compete with copper. Aβ is found to accumulate within structurally damaged mitochondria in AD.[132] Metals ions also appear to play a part in the functioning of tau, such as copper,[156] zinc,[153,156,157] iron,[156,158-160] aluminium,[161] calcium[162] and magnesium.[162].

### 1.4.6. Cell cycle

The life cycle of a cell can be subdivided into two phases. The interphase (I) is a period of preparation (G1), DNA replication (S) and further preparation (G2). The interphase is followed by the mitosis phase (M) is when the cell finally divides; additionally the cell can enter a state of rest (G0). Controlling the shift between these time-points are cyclin dependent kinases (CDKs) which remain steadily expressed during the cell life cycle, but their activity are very precisely regulated by varying levels of proteins that bind them such as cyclins and CDK inhibitors (CKIs). **Figure 1.15** depicts the cell cycle and the proteins that regulate it. Also controlling CKD activity is the phosphorylation/de-phosphorylation of conserved threonine and tyrosine residues of CKDs, these modifications elicit a conformation change that allows better binding of cyclins. CKD2, CKD4 and CKD6 regulate interphase while CKD1 regulates mitosis. CKDs works to phosphorylate proteins (substrates), these substrates regulate transcription factors that regulate expression of proteins required at certain time-points, and some substrates also regulate CKDs such as Wee1 and Cdc25. Regulation of the cell cycle is bolstered by several check points that check DNA integrity, event completion and sufficient energy levels before ultimately committing to DNA replication or cell division, the two main check points are at the G1/S boundary and G2/M boundary. It is well established that the cell cycle is intrinsically linked to the cell's DNA damage response (DDR) system; differentiated neurons predominantly employ the non-homologous end joining (NHEJ) DNA repair mechanism.

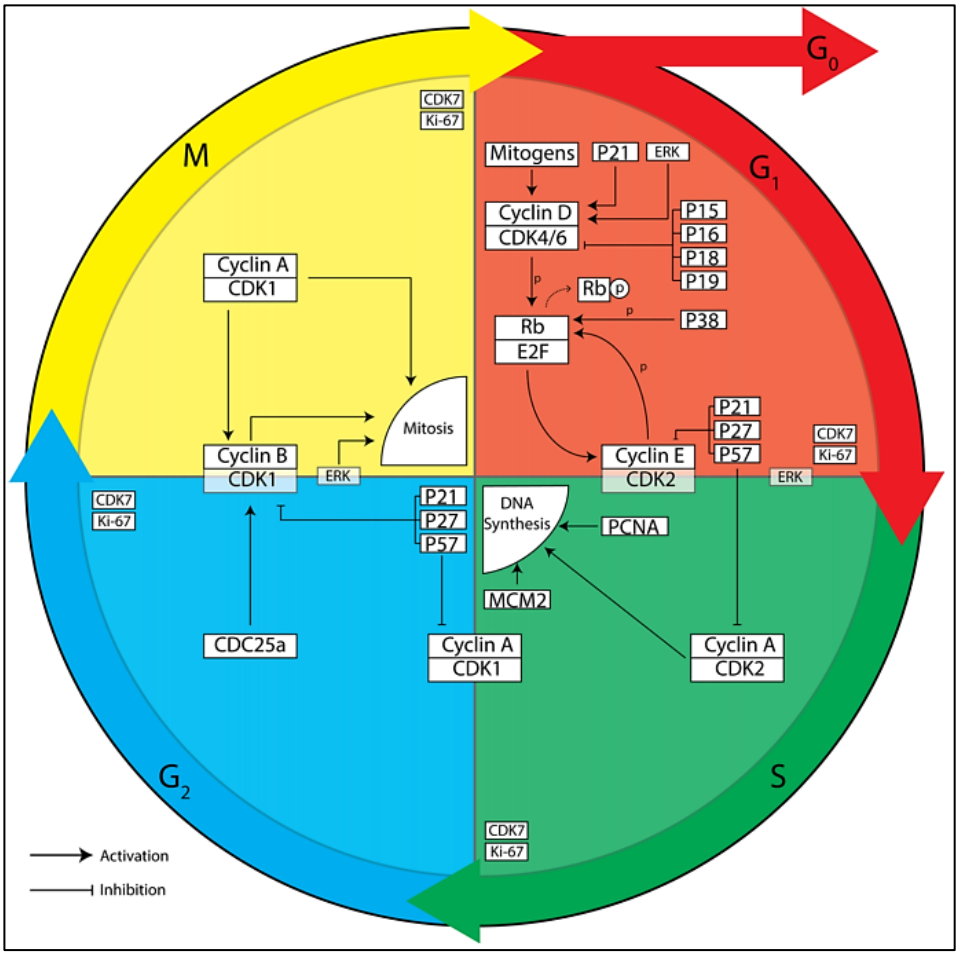



**Figure 1.15** Diagram illustrating the proteins involved in regulation of the cell cycle which involves a resting phase (G0) and four active phases (G1, S, G2 and M) **.** Image taken from Van Leeuwen *et al* (2015).[163]

Cancer is by far the most studied complex disease and is the result of the extended life of a cell, whereas neurodegeneration is the result of premature death of a cell, with this premise it is easy to imagine a link between the two diseases. Cancer is a highly genetic disease and caused by the accumulation of causative variants that directly or indirectly affect the CKDs that regulate the cell cycle; beyond a certain accumulation of causative variants the check points become ineffective which results in genomic instability (GI), chromosomal instability (CI) and unscheduled proliferation. Indeed CKIs are deranged in AD[164] and there is an inverse epidemiological association between AD and cancer.[165]

In adults, neurons are post-mitotic, as in they do not exhibit mitosis and cell division after foetal development is complete; however a small population of neural stem cells in the sub-ventricular and sub-granular zones of the hippocampus can undergo neurogenesis to produce new neurons. Despite most neurons not dividing, extensive evidence reports that cell cycle proteins are detected in neurons

of healthy and AD individuals, generally at lower levels when compared to healthy individuals,[163] they also co-localise with NFTs and senile plaques.[166,167] These cell cycle proteins stimulate cell cycle re-entry which progresses to the start of metaphase but does not initiate mitosis[163].

Whether cell cycle re-entry is a cause of neurodegeneration or an effect is unknown. Multiple lines of evidence suggests that neurons re-enter the cell cycle to elicit neuroplasticity of microtubules,[168,169] synapse formation[170-173] and initiate DNA repair.[174-176] In all cell types the DNA repair system is activated using the cell cycle and if the damage is irreparable then the cell proceeds to DNA replication which ultimately lead to apoptosis. Little evidence suggests re-entry could also cause neurodegeneration.[177-179]

The gene *PIN1* is associated with cancer[180] and inversely associated with AD,[181] supporting this association are lines of evidence that PIN1 can regulate gene expression,[182] the cell cycle,[183-185] tau phosphorylation,[186-188] Aβ production [189] and reduce the effect of oxidative stress.[190] Further supporting a link with oxidative stress, cell cycle proteins appear to be upregulated by HNE (4-hydroxynonenal) in cortical neurons,[191] and hydrogen peroxide can advance or halt the cell cycle depending on its concentration.[192]

### 1.4.7. Neuroinflammation

Humans have three levels of defence against microbial intruders, the physical barrier is the wall created by the skin and the mucous membrane of our intestinal tract, once biological invaders pass this barrier then the innate and adaptive immune systems provide agents that survey, tag and eliminate the invading microbes.

A system that all animals naturally have is the innate immune system. The agents of the innate system are cells such as macrophages, natural killer (NK) cells and the complement cells such as dendritic cells, **Figure 1.16** visually demonstrates the hierarchy of cells involved in the immune system. Macrophages and dendritic cells express toll-like receptors (TLRs) such as TLR2, TLR4 and TLR6 and co-receptors (CD14, CD36 and CD47), once a macrophage is activated it secretes cytokine/chemokine proteins.

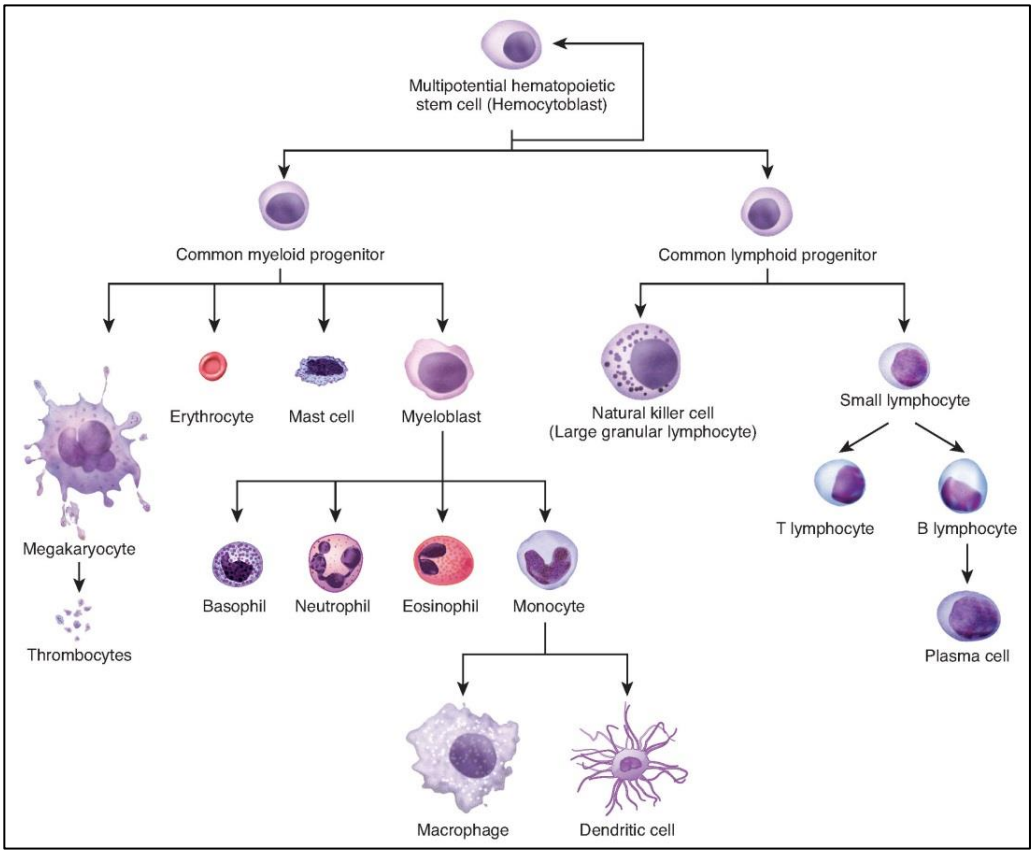



**Figure 1.16** Flow chart demonstrating the various fates of hemocytoblast stem cells, most of which are involved in the immune system (macrophages, dendritic cells, natural killer cells, T lymphocytes and B lymphocytes). Image taken from cnx.org.

The adaptive immune system is a system employed by many vertebrates. Adaptive immunity includes T lymphocyte cells (mature in the thymus) and plasma B lymphocyte cells (mature in the bone marrow). T cells have surface receptor proteins that allow surveillance of host and foreign peptides presented by MHC class I or II. B cells initially produce surface-bound antibody proteins but upon appropriate antigen recognition and co-stimulation, can multiple to produce both memory B cells, and moreover plasma cells which produce copious secreted antibody proteins specific for the antigen concerned. Antibodies bind antigens using the fragment antigen binding (Fab) region and bind to cells using the fragment constant (Fc) region. There are four types of Fc regions (IgA, IgD, IgE and IgM) and each can bind to different types of cells. The adaptive capability of the antibody comes from the Fab regions, there are approximately three billion plasma b cells in our blood, each cell is genetically coded with a different Fab region, there are approximately 100 million different Fab regions, and each is represented by approximately 30 individual plasma B cells in the blood. As a result of recombination events of nine different DNA modules and random insertions/deletions of DNA bases of these modules, this diversity allows the adaptive immune system to recognise almost any antigen.

There are cells and proteins that act as intermediates between the innate and adaptive systems to enhance their abilities, this system is called the complement system and is part of innate immunity. Complement includes proteins such as C1q, $C_3$, and $C_4$. The complement system is activated by the classical, lectin or alternative pathways which are subdivisions of the immune response.

Glial cells such as microglia (a type of macrophage) and astrocytes are the innate immune cells of the central nervous system (CNS). Microglia survey the brain for pathogens and support CNS homeostasis and plasticity. Microglia are activated by neurotransmitters, adenosine tri-phosphate (ATP), growth factors, cytokines, ions or surface molecules, which act as signals of cell death, apoptosis or injury. Once microglia are activated they release proinflammatory cytokines such as interleukin-1 (IL-1), interleukin-6 (IL-6) and tumour necrosis factor alpha (TNF-α). Microglia appear to be impaired by TNF, IL-1, IL-12 and IL-23, suggesting a negative feedback loop.

Neuroinflammation was first linked with AD by Alois Alzheimer but until recently inflammation was thought to occur in the later stages of AD as an effect rather than a cause. However recent evidence has highlighted its importance in the aetiology of AD, it is now consensus that inflammation contributes to and exacerbates the disease. All evidence implicates innate immunity in AD, as opposed to adaptive immunity which is known to promote other neuroinflammatory diseases such as multiple sclerosis (MS) and encephalitides.

Epidemiological studies show that non-steroidal anti-inflammatory drugs reduce risk of AD with prolonged treatment.[193-195] On a molecular level the brains and cerebral spinal fluid (CSF) of AD patients show upregulated levels of inflammatory markers (IL-1, IL-6, GM-CSF, TNF-α, IL-12, IL-23) which are seen to precede clinical symptoms.[196,197] Additionally, plaques and NFTs in AD brain contain complement proteins C1q, $C_3$ and $C_4$ and the proinflammatory cytokine IL-6,[104] interestingly individuals with high plaque and NFT burden and no AD show no neuroinflammation and no synaptic loss.[198]

Multiple lines of evidence suggest that microglia are the key players linking neuroinflammation and AD. Microglia are reactivated in AD brain,[199,200] and surround amyloid plaques in a morphological configuration that suggests they are responding,[201] microglia also secrete IDE and Nep which can degrade Aβ. Interestingly, microglia are activated by Aβ oligomers via the SR-A receptor and the KCa3.1 potassium channel[202] but conversely can also be impaired by Aβ.[203] Aβ can bind several receptors (CD14, CD36 and CD47, RAGE, TLRs, $\alpha_6\beta_1$ integrin, class A scavenger)[204-211] that are expressed by microglia. Several genetic susceptibility loci are also involved in innate immunity, including *CD33*, *MS4A6A*/*MS4A4E*, *CR1* and *TREM2*.[48]

## 1.5. The NeuroX Genotyping Chip

The NeuroChip (NeuroX) is a customised Illumina genotyping chip. The current version of the NeuroX v1 contains in total 267,607 markers, of which 91% (n=242,901) are Illumina content from the Infinium HumanExome BeadChip v1.1 and the remaining 9% (n=24,706) are additional custom content. According to the Illumina data sheet the standard content was designed to 'genotype exonic variation representing diverse populations such as European, African, Chinese and Hispanic, and a range of common disorders such as type 2 diabetes, cancer, metabolic and psychiatric', **Table 1.7** lists a breakdown of the standard content as stated in Illumina data sheet. The custom content was designed by a consortium led by University College London (UCL) that selected variation linked to neurodegenerative and/or neurologic disorders. Although the Illumina HumanExome chip has a capacity of 200k custom markers, most of the custom markers designed by the consortium did not pass quality control and thus only 24k were actually included in the final design. A limited number of NeuroX v1 chips were made for the consortium which were allocated to research groups, the design of NeuroX v1 has since been described in a review article.[212] At the time of writing this thesis, NeuroXs v2 is currently in the design phase and should be available by summer 2016.

| Categories | N |
|---|---|
| Total Markers | >240,000 |
| Number of unique RefSeq entries covered by at least one probe | >20,000 |
| Nonsynonymous SNPS (NCBI) | 219,621 |
| SNPs in splice site | 10,675 |
| Stop Variants | 5,637 |
| SNPs in promotor regions | 7,012 |
| SNPS in extended MHC region | 5,158 |
| GWAS tag markers* | 4,761 |
| HLA tags | 2,061 |
| Ancestry informative markers | 3,468 |
| Identity by decent markers | 3,369 |
| X / Y / Mitochondrial | 40 / 101 / 177 |
| Indels | 180 |

**Table 1.7** Breakdown of the markers genotyped on the Illumina HumanExome BeadChip v1.1 as stated by the Illumina data sheet for the chip. Markers are divided into categories (Category) and the number of markers in each is given (N) *Disease-associated tag markers identified from recent GWAS.

### 1.5.1. Illumina design

Illumina HumanExome chips use oligonucleotides (probes) that are 60 bases in length; each probe is bound within distinct wells on a plate via beads. Probes are designed to genotype variants in two different ways, either using Infinium I design or Infinium II design. Infinium I are designed to genotype ambiguous SNPs (A/T and G/C) by annealing up to and including the SNP, they contain two probe sequences, one for each allele, the two probes are extended by a single base; incorporated

A and T nucleotides are labelled with red coloured fluorophores and C and G green coloured fluorophores, genotype is determined based on if either or both of the allele specific beads fluoresce regardless of colour, if both beads fluoresce the sample is heterozygote for that SNP. Infinium II are designed to genotype unambiguous SNPs (A/G, A/C, T/C, and T/G) by annealing up to but not including the SNP, only one probe sequence is used which is extended by a single base like with Infinium I, genotype is determined by which of the two colour fluorophores light up from the bead, if both colours fluoresce the sample is heterozygote for that SNP. **Figure 1.17** illustrates both designs using an example SNP.

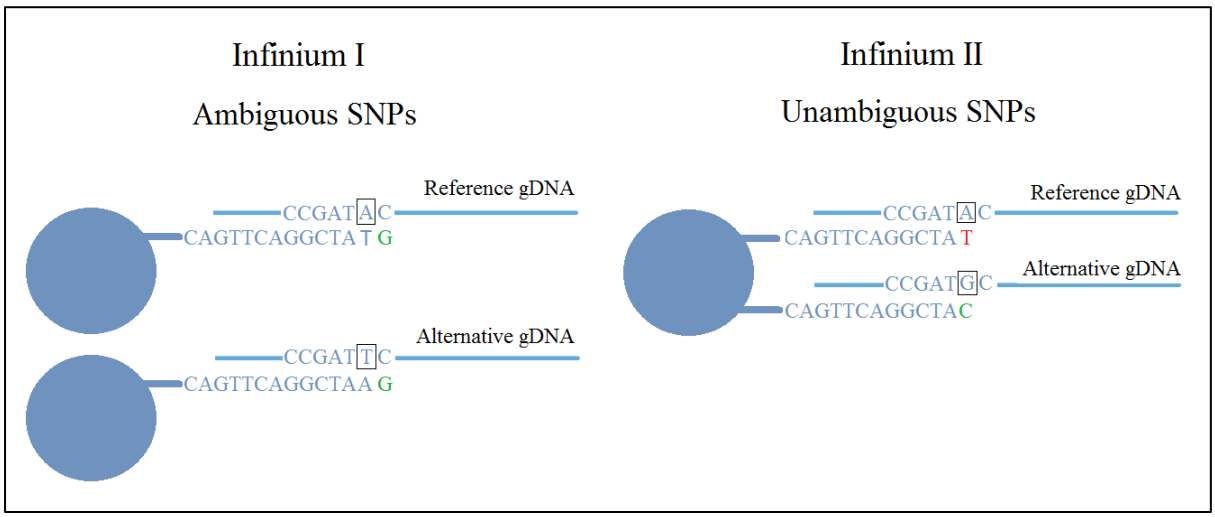



**Figure 1.17** Illustration of how Illumina Infinium I and Infinium II probe designs work using A>T and A>G SNPs as examples of an ambiguous and unambiguous SNP respectively. The solid filled blue circle represents the bead which has a unique oligonucleotide sequence attached (probe). Genomic DNA (light blue) hybridises to the probe and single base extension follows. The hollow black box marks the position of the SNP. The presence of fluorescence from either or both of the two Infinium I beads determines the genotype for ambiguous SNPs. The colour of the fluorescence from the one Infinium II bead determines the genotype for unambiguous SNPs.

Infinium II can suffer from inaccurate genotyping if the SNP has more than two alleles, for example, for a variant with an A reference allele and either a C or G alternative allele (A>C/G), the A allele will fluoresce red and both the C and G alleles will fluoresce green. Infinium I beads are the better design, however they take up more room on the genotyping chip and so as a result 99.6% of the markers on the NeuroX (266,608) are Infinium II.

For insertion and deletion (indels) Illumina uses the Infinium II design. The single base extension employed by Illumina genotyping chips means that indels are genotyped based on the first base of the indel and first base after the indel. Single base extension is not the best design for genotyping indels especially for those that are greater than one base. False positives will arise if a SNP overlaps the first base of the indel. False negatives will occur if the first base after the indel is the same as the first base

of the indel. There is also the real possibility that a positive finding could be the result of different indel at the same start position that just happens to have the same first base of/after the indel.

Illumina has designed its own system to discriminate strands of specific variants called the TOP/BOT. Traditionally, as research focused on candidate genes, forward/reverse strand and sense/antisense strand was used to define strand, both use the direction of transcription and thus the coding strand to define them, this system is still used today, however since the dawn of whole genome sequencing a second system was employed which has arbitrarily assigned a strand of each chromosome to plus or minus (+/-). Confusion has arisen when describing ambiguous SNPs (A/T and G/C) and as a result databases that hold SNP information such as dbSNP (link proved in **Section 12**) describe the strand of some SNPs incorrectly. Illumina decided to define their own system to describe SNP strand, which helps to definitively define the strand of ambiguous SNPs by using the sequence of surrounding bases. **Figure 1.18** demonstrates how the Illumina system works.

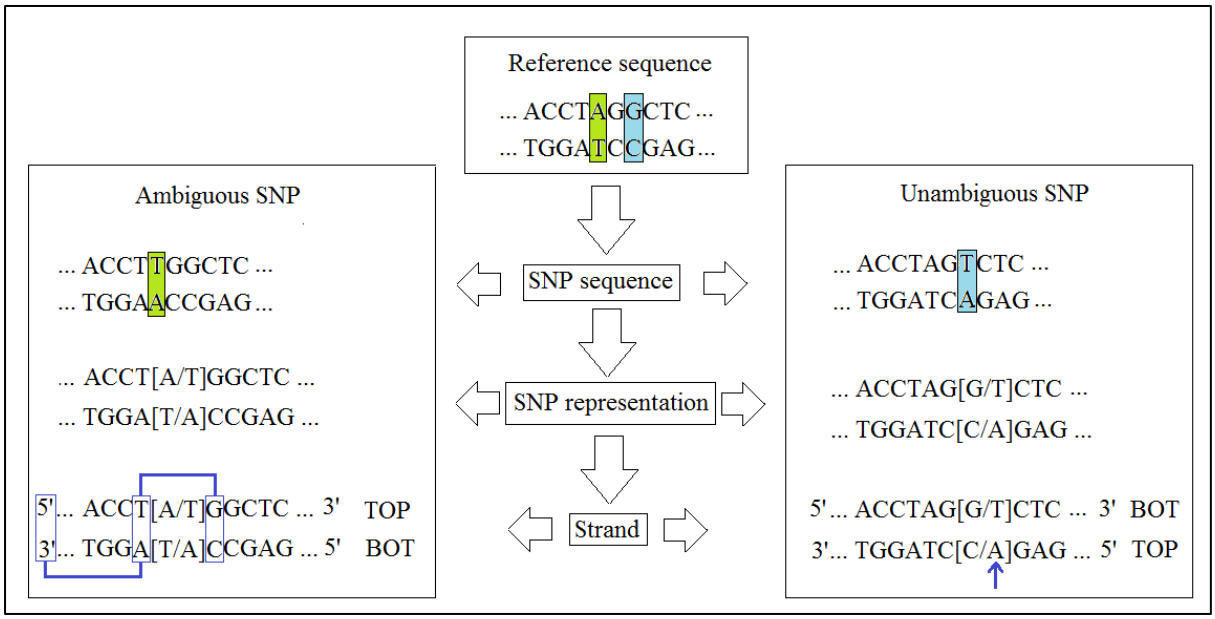



**Figure 1.18** Flow diagram showing how the Illumina TOP/BOT strand assignment is made using A>T and G>T SNPs as examples of an ambiguous and unambiguous SNP respectively. The reference sequence is given at the top, working down the flow chart the sequence for two SNPs is given; one ambiguous (highlighted in green) and one unambiguous SNP (highlighted in blue). For unambiguous SNPs (A/G, A/C, T/G and T/C), nucleotide A is on the TOP strand and nucleotide T is on the BOT strand. For ambiguous SNPs the nearest intra strand unambiguous pair (one base each side of the variant) is used instead, in the example given this is the first base either side as they give T/G A/C pairs, the strand with the unambiguous inter strand A/T pair on the 5' side is designated the TOP and the strand with the unambiguous inter strand A/T pair on the 3' side is the BOT strand.

Illumina has provided software to enable customers to quality control the intensity data generated for each marker on their chips, the software takes the raw red and green intensity data, normalises it and plots each sample against one another for every marker, the resulting 'SNP graph' enables the user to visualise the genotypes and access quality. The software also employs a clustering algorithm which

automatically clusters the data. The software calculates a score (GenTrain) using the shape, distance, intensity and number of clusters, this score is designed to mimic the visual evaluations made by a human and can be used to identify markers that need quality assessing. Samples are called (genotyped) for each marker using a Bayesian classification algorithm using the position of each sample relative to the clusters and the GenTrain score, the resulting call is given a score (GenCall).

### 1.5.2. Utility of the NeuroX

After GWAS identified common variants associated with disease, it was evident that there was still a considerable amount of heritability missing. It was thought this could be accounted for by low/rare frequency variants. However, a problem arose due to the number of samples required to ascertain power to associate these low/rare frequency variants with disease, tens of thousands of individuals would be required, and ideally you would NGS these individuals - the cost of NGS today means that these studies are rarely feasible. Rare variant genotyping chips were designed to address this issue, the NeuroX was one such example. The NeuroX was designed for researchers targeting neurodegenerative diseases that wanted to genotype rare variants.

There are many advantages to using the NeuroX, even though they will never match NGS they go some way to bridging the gap by providing an improvement to GWAS chips. Genotyping different disease types with the same panel of markers allows the identification of genetic crossover such as pleiotropic genes and pleomorphic risk, and genotyping different cohorts with the same panel of markers means merging data is much easier. The custom content has enabled the addition of variation which has been identified in recent research and including those that are not present in public databases; for example adding markers that genotype highly penetrant personal variants that are known to cause neurodegenerative disease, increasing the concentration of markers in susceptibility loci (fine-mapping), this former point also means that the NeuroX can be used as a screening tool in research and clinical practice. Practically, genotyping chips also offer cheaper data storage and quicker data analysis relative to NGS.

There are some limitations however. All genotyping chips lack the ability to genotype novel variation and are more limited when it comes to the type of variation they can successfully genotype; genotyping chips are designed for SNPs, there is little confidence in small indels and there is no design for large scale variants such as CNVs or rearrangements. The final design of the Illumina HumanExome chips has gradually evolved as a result of many years of exome sequencing studies culminating in 16 cohorts, exonic variants were chosen that were observed more than three times in at least two different cohorts, 75% of the individuals genotypes were of European/American ancestry, thus the design of the NeuroX is highly biased towards European/American cohorts.[213] The NeuroX does not genotype all variation; there is approximately 100 million variants documented in the current build of dbSNP (146), even considering the NeuroX's design to target coding variation, it still falls

considerably short of the 87 million variants located within genes, albeit these will include common variants. The decision to focus on coding variation has also limited its ability to genotype non-coding variation, which includes variants located within gene regions, such as in introns and in the 5' untranslated regions (UTR) and in the 3' UTR; variants in these regions regulate/modify transcription, translation and localisation. The exclusion of intergenic variants would be predicted not to impact too negatively, as currently we have limited ability to predict what these variants actually do.

Overall the NeuroX is beneficial for diseases thought to be highly genetic as a result of rare non-synonymous variants; diseases with Mendelian inheritance would definitely fall into this bracket, diseases with an early onset and high heritability might also benefit. For sEOAD the disease ranges from very early to medium onset (20's to ≤65 years of age) with a predicted heritability of approximately 92%, [21] so the genetics are likely to be of high to medium penetrance. The incidence of sEOAD means that cohorts are of limited numbers; there are only hundreds of sEOAD individuals available for study which is a relatively small number compared to the tens of thousands of LOAD individuals available. Using NGS to study sEOAD would be relatively cheap and achievable and so would be the ideal platform to study these cohorts, however using the NeuroX initially doesn't provide any disadvantage especially given the low cost of this approach.

Genotyping chips that focus on genotyping the human exome have existed for some time and Illumina and Affymetrix both have their own versions. Customised human exome chips are not a new idea either, for example the Immunochip focuses on immunological mediated diseases[214,215] and the Metabochip focuses on metabolic diseases.[216] Customised exome chips have not been utilised as extensively as the standard human exome chips ,'NeuroX' is mentioned in only nine articles to date (March 2016) according to Web of Science, of which six are studies that utilised NeuroX genotyping data.[217-222]

### 1.5.3. Aims of the Project

The project aims to identify loci and/or variants that contribute to the aetiology of sporadic early-onset Alzheimer's disease (sEOAD). DNA samples of sEOAD individuals have been genotyped using the NeuroX genotyping chip with the intention to genotype the exonic variation unbiasedly across the entire human genome. With this genotyping information, annotation tools have been used to filter the information and identify variants that are known or predicted to cause AD or other neurodegenerative/neurologic diseases. Additionally, statistical tests have been performed in a bid to identify loci and/or variants that are associated with sEOAD. This study design has not been performed with sEOAD before; hopefully it will help identify high effect genetic factors contributing to the disease and further our understanding of the disease process; furthering our understanding will create new drug targets and bring closer a drug that can modify or prevent the disease.

# 2. DNA Preparation and *APOE* Genotyping

## 2.1. Methods

### 2.1.1. DNA extraction

DNA was isolated from either brain tissue or blood and extracted 'in-house' using a standard Phenol: Chloroform procedure as explained below. The protocol for DNA extraction for either sample type was the same except the digestion step. All reagents were kept at room temperature unless otherwise stated.

The process of DNA extraction began with digestion of the tissue, to release the DNA from the cells and degrade the unwanted RNA and proteins. Digestion of brain tissue was achieved by adding an equal volume Lysis Buffer AL (QIAGEN), 30mAU/ml Proteinase K (QIAGEN) and incubating at 50°C/380RPM for 18 hours. Blood digestion was achieved by adding 100µg/ml RNase A (Thermo Scientific™), an equal volume Lysis Buffer AL (QIAGEN), 6mAU/ml Proteinase K (QIAGEN) and then incubating at 56°C for 10 minutes.



**Figure 2.1** DNA extraction of tissue using the Phenol: Chloroform procedure. Image shows the Falcon tube during the DNA separation stage where the tissue constituents separate into three layers after centrifugation.

After digestion, DNA separation allowed the isolation of the DNA from all other cell constituents. DNA separation was achieved by adding an equal volume of Phenol: Chloroform: Isoamyl Alcohol 25:24:1 (Sigma) to the sample and inverting a minimum of ten times, followed by centrifugation at 4°C/6000RPM for 15 minutes. After centrifugation the solution separated into three layers within the Falcon tube as shown in **Figure 2.1**. The top layer contains the DNA, middle layer the lipids and bottom layer the proteins, the top layer was removed and into a new tube. This step was repeated with a new tube until the middle layer became clear (if necessary up to an additional three times).

DNA precipitation was achieved by adding one tenth the volume of Sodium Acetate pH 5.2 and inverting a minimum of two times to mix, followed by adding one equal volume 100% ethanol (cooled on dry ice) and again gently inverting several times to mix. At this stage the DNA would usually precipitate into what looks like thin strands of cotton wool, however in some cases the DNA did not precipitate entirely and was hard to see with the naked eye. In both cases the tube was left to cool on dry ice for approximately 1 hour or in a -18°C freezer for 12

hours (overnight) as this can help the precipitation process. Extensive time in the cooling phase were avoided as this sometimes precipitated unwanted salt from the sample, this salt appeared similar in colour and texture to the DNA and could not be easily separated from the DNA.

After precipitation the DNA was further cleaned and desalted using ethanol/water solution. Cleaning the DNA was achieved by centrifuging the sample at 4°C/6000RPM for 15 minutes. If following centrifugation the DNA was present in a tight pellet bound to the base of the tube then the liquid phase was entirely removed, if the DNA was not visible or did not form a tight pellet then all but approximately 500µl of the liquid phase was removed. 1 ml 70% ethanol was added and the tube gently flicked to resuspend the pellet. At this stage, if the DNA that was not visible it would usually precipitate. This cleaning cycle was repeated a further two times.

After DNA cleaning, the DNA pellet was dried and re-suspended. Drying was achieved by leaving the tube to air dry for approximately 45 minutes or until dry. This process was encouraged by resting the tube at 135° angle to allow the liquid to run from the tubes. Re-suspension was achieved by adding approximately 50 µl Buffer AE (QIAGEN) to the tube, this volume differed between samples as each sample and tissue type yielded varying quantities of DNA. The tube was capped and placed on a heat block at 50°C for approximately 45 minutes or in the fridge (2°C) for at least 12 hours (overnight) to allow the DNA to re-suspend in the buffer solution. Finally the tubes were centrifuged briefly to settle any condensation and the entire solution transferred to DNase free tubes with screw lids.

The concentration of the DNA samples was assessed using a NanoDrop 3300 spectrometer (Thermo Scientific) following the manufacturer's instructions with all settings at default, the NanoDrop was also used to estimate purity using the 260/280 ratio. Quality and approximate molecular weight was assessed by electrophoresis on a 1% agarose gel at 70-80 volts for 40 minutes. Each well was loaded with 0.25-2µl DNA mixed with five times the volume of 6 X Loading Dye (Thermo Scientific™). The first well of each gel was loaded with 1-2µl Lambda HindIII DNA ladder (Thermo Scientific™). The resulting gels were photographed under UV light. **Figure 2.2** shows an example of one of the DNA quality control gels.

DNA samples were selected for genotyping based on phenotype of the individual and quality/quantity of their DNA. Case individuals were selected if they fulfilled the following criteria: European ethnicity, possible/probable/confirmed AD, void of clinical/physical features for non-AD diseases and had an age at onset (AAO) ≤ 65 years of age. Cases with lack of this data at the time of selection were also included in case this information became available in the future. Where AAO was not available, it was calculated using age at death and assuming eight years disease duration.[19] Control individuals were of European ethnicity, ≥ 65 years of age at sampling/death and were void of clinical/physical features of any disease. DNA was selected at a concentration ≥ 30ng/µl, a 260/280 ratio of 1.8-2.0 and a molecular weight approximately > 9,416bp.

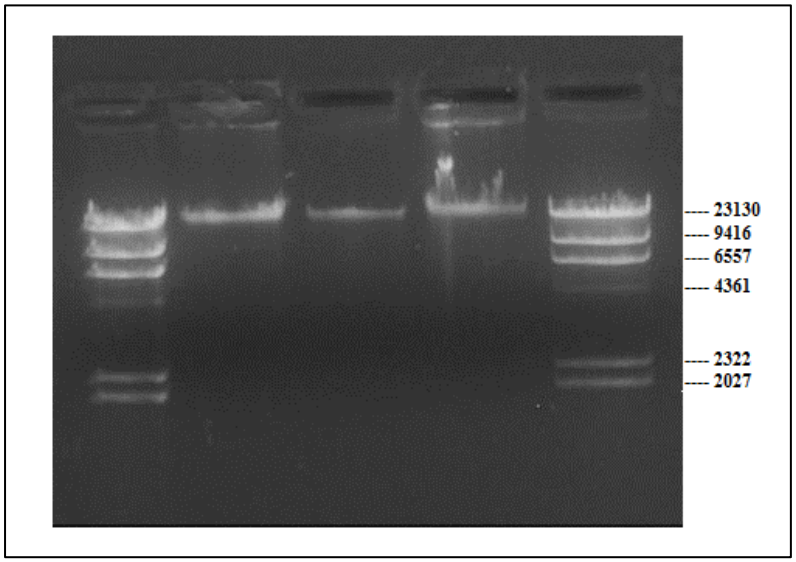



**Figure 2.2** Agarose gel containing genomic DNA made using the phenol chloroform protocol. Lane one and five contain the Lambda HindIII DNA ladder with the size of the reference bands marked (bp). Lanes one to three each contain DNA from three separate DNA extractions

DNA samples from the selected individuals were separated into three bins according to their concentration; > 70ng/µl, 50-69ng/µl or 30-49ng/µl. Each sample would contribute 720-1200ng of DNA for genotyping and thus the final volumes for the three bins were 12µl at 70ng/µl, 18ul at 50-69ng/µl or 24µl at 30-49ng/µl. DNA samples that did not fulfil these concentration requirements were excluded. The DNA samples were pipetted into standard 96 well plates and those at > 70 ng/µl were diluted to 70ng/µl by adding the relevant volume of Nuclease-free Water (QIAGEN) to the appropriate wells before adding the relevant volume of DNA and mixing. The plates were stored in a freezer at -80°C before being sent to the Laboratory of Neurogenetics (National Institute of Aging, USA).

## 2.1.2. Genotyping *APOE* ε

Selected samples with no *APOE* ε genotype documented were genotyped for the two SNPs required (rs7412 and rs429358); this was achieved using two TaqMan® assays, one for each SNP. Both SNPs are C/T; the probe that hybridised the C allele was attached to a VIC™ fluorophore and the probe that hybridised the T allele was attached to a FAM fluorophore. For each assay the reaction mixture (cocktail) was made in a total volume of 20µl with the following reagents: 1 X TaqMan® Universal PCR Master Mix (Applied Biosystems), 1 X Custom TaqMan® SNP Genotyping assay (Applied Biosystems), 3ng/µl DNA and molecular grade water to the required volume. The PCR was performed on MX3000P QPCR System (Agilent Technologies) and subjected to the following thermal profile: denaturation at 95°C for 10 minutes, followed by 55 cycles of 92°C for 15 seconds and 60°C for 60 seconds. Fluorescence recording and allele discrimination was conducted using the supplied software with ROX™, FAM™ and HEX™ filters selected; the HEX filter detects the same wavelength range emitted by the VIC™ fluorophore. The software plotted the samples based on the

intensities of the FAM™ and VIC™ fluorophores at the end of thermal cycling. Scatter plots were generated for each run and samples were assigned a genotype by manually clustering samples within one of four coloured boxes. Samples with high FAM and low VIC intensities were clustered in a red box to assign them as having just the T allele (Allele B), samples with low FAM and high VIC intensities were clustered into a blue box to assign them as having just the C allele (Allele A), samples with high FAM and high VIC were clustered into a green box to assign them as having both the T and the C alleles (Both) and samples with neither FAM or VIC intensities were clustered in gold box to assign them as having neither allele or being a no template control (NTC). Four positive controls samples with an *APOE* ε status already documented (M328, M302, AD236 and AD235) were genotyped first to access the assay design. FAM and VIC intensity signals were seen in all samples genotyped for rs429358, which was not expected as this would indicate all samples were heterozygous for this SNP, despite this, the scatter plot algorithm was able to differentiate true and false signals. When looking at the amplification plots the intensity lines that were straight and not the classic shape (logistic growth curve) were false signals. **Figure 2.3** shows the amplification plots generated when genotyping the four samples for both SNPs. **Figure 2.4** and **Figure 2.5** gives an example of the scatter plots generated when genotyping a large number of samples for the rs429358 and rs7412 SNPs respectively, both runs included the four positive control samples. The assigned genotypes for all samples genotyped using the TaqMan assays were downloaded from the software, using Excel the *APOE* ε status was calculated using **Table 2.1**.

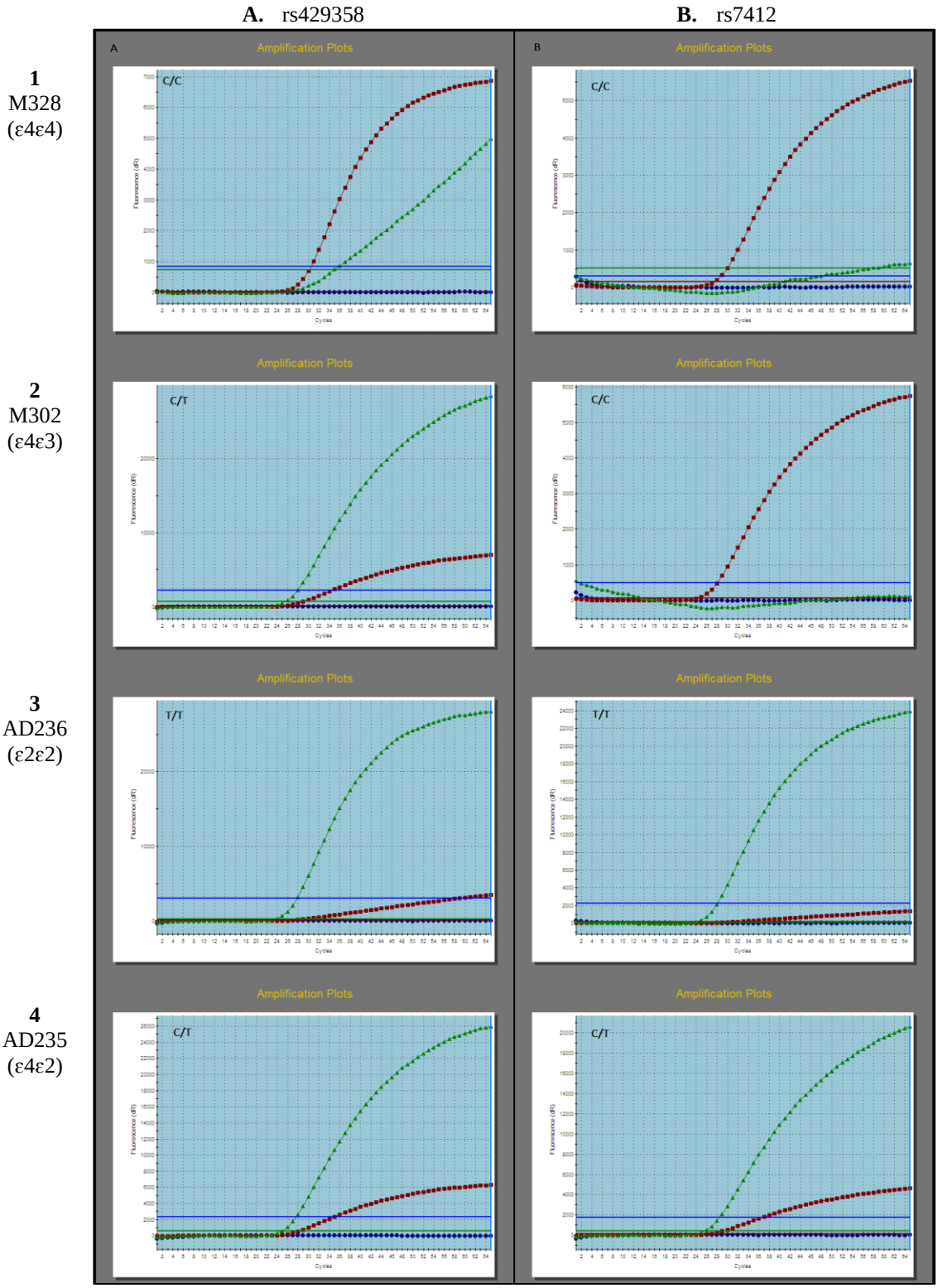



**Figure 2.3** The amplification plot generated by MX3000P QPCR System (Agilent Technologies) when genotyping four samples; M328 (row 1) M302 (row 2), AD236 (row 3) and AD235 (row 4) for SNPs rs429358 (column A) and rs7412 (column B) using the TaqMan genotyping assays. Intensity lines for the HEX fluorophore are in red (C allele) and the intensity lines for the FAM fluorophore are in green (T allele).
All four samples were positive controls as they had an *APOE* ε status already documented before genotyping, their status' are given on the left hand side below the corresponding sample. Given the *APOE* ε status for each sample the expected genotype for each SNP is written in the top left corner of the amplification plot. Intensity lines that are straight and not the classic shape (logistic growth curve) are false signals (1A, 1B, 3A and 3B)

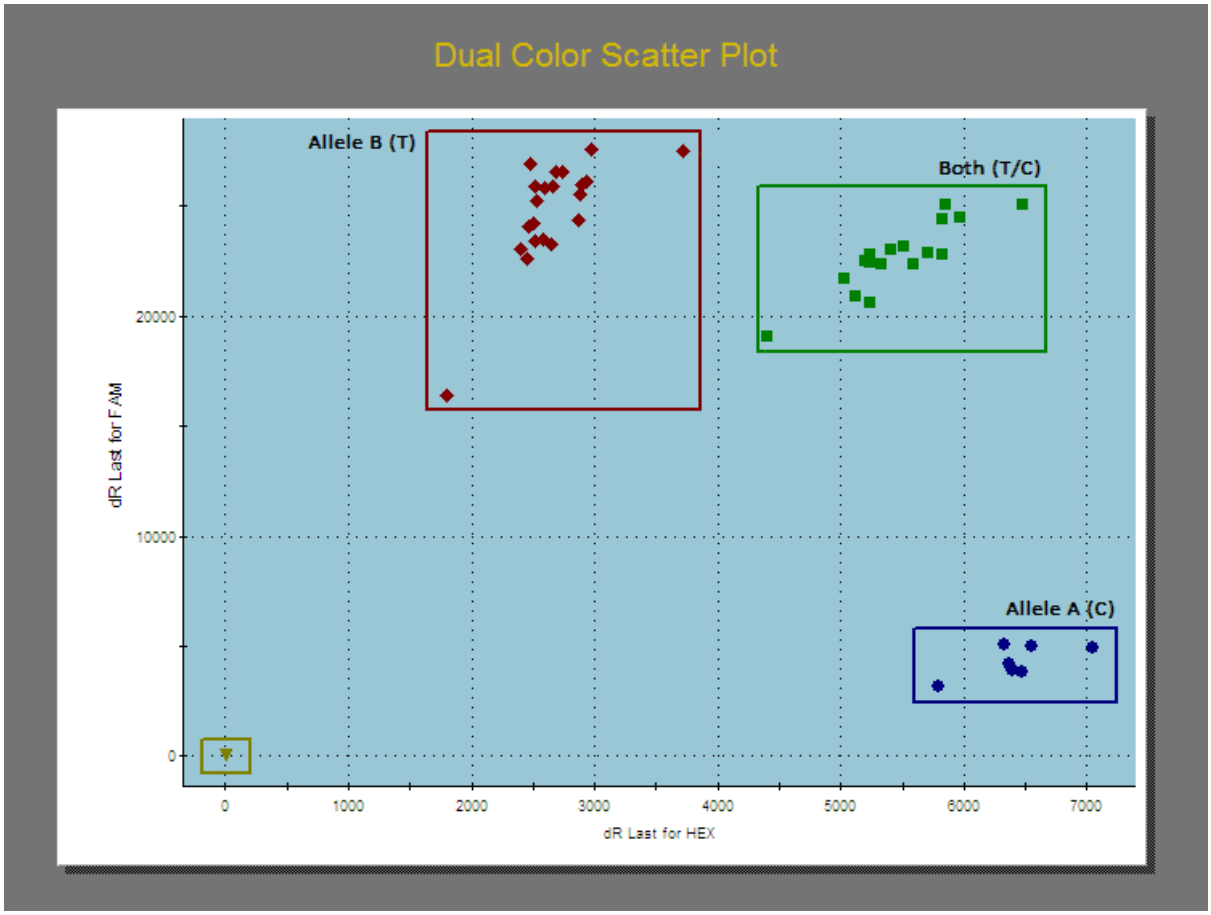



**Figure 2.4** An example scatter plot generated by MX3000P QPCR System (Agilent Technologies) when genotyping SNP rs429358 in 45 samples using the TaqMan assay the image has been modified to include black labels. Each sample is represented as a dot. Samples in the red box (Allele B) have just the T allele, those in the green box have both the T and C allele ('Both'), those in the blue box have just the C allele (Allele A) and the sample in the gold box is a no template control (NTC).

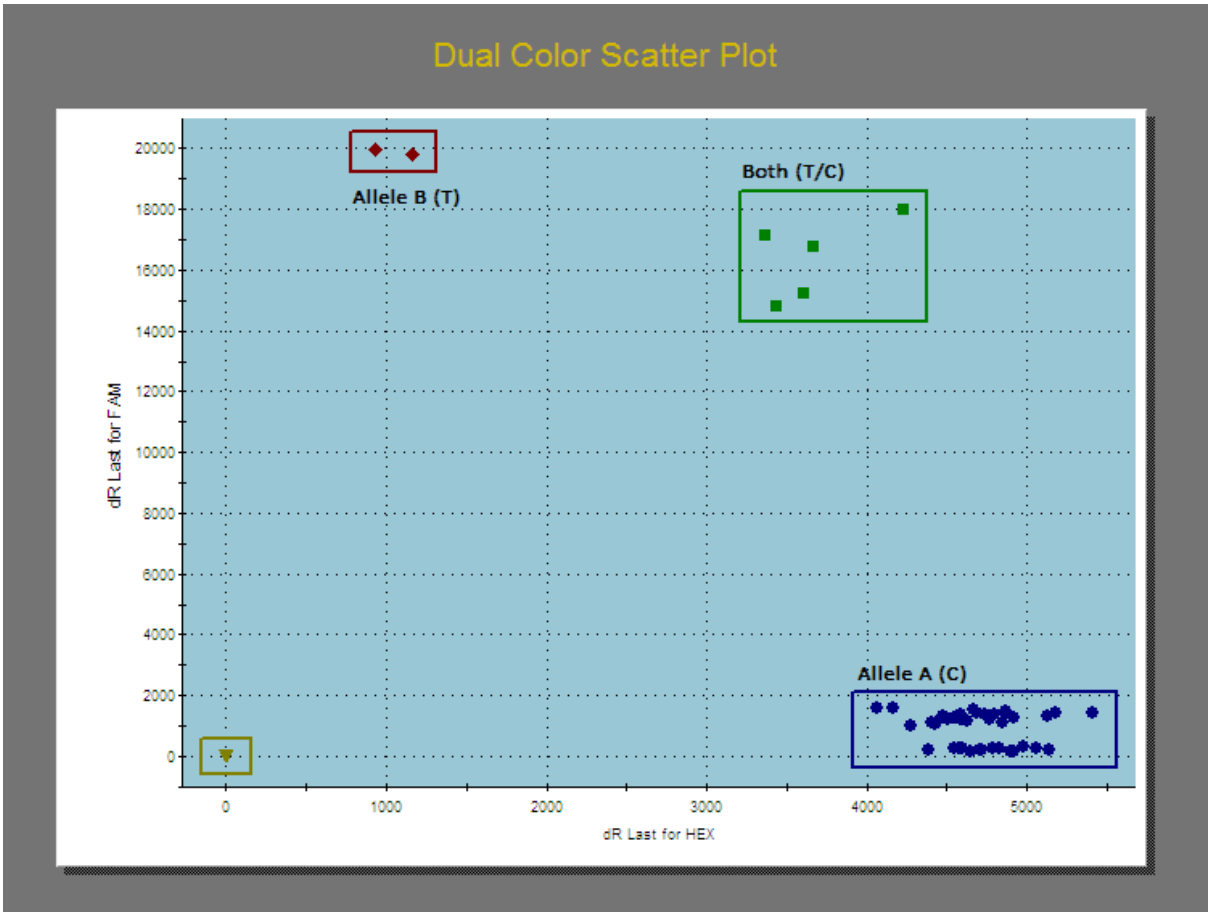



**Figure 2.5** An example scatter plot generated by MX3000P QPCR System (Agilent Technologies) when genotyping SNP rs7412 in 45 samples using the TaqMan assay, the image has been modified to include black labels. Each sample is represented as a dot. Samples in the red box (Allele B) have just the T allele, those in the green box have both the T and C allele ('Both'), those in the blue box have just the C allele (Allele A) and the sample in the gold box is a no template control (NTC).

| Assigned Genotype | | Actual Genotype | | Haplotype 1 | Haplotype 2 | *APOE* ε status |
|---|---|---|---|---|---|---|
| **rs429358** | **rs7412** | **rs429358** | **rs7412** | **rs429358 : rs7412** | **rs429358 : rs7412** | |
| Allele A | Allele A | CC | CC | C:C | C:C | 44 |
| Both | Allele A | CT | CC | C:C | T:C | 43 |
| | | TC | CC | T:C | C:C | 34 |
| Both | Both | CT | CT | C:C | T:T | 42 |
| | | TC | TC | T:T | C:C | 24 |
| | | CT | TC | C:T | T:C | 13 |
| | | TC | CT | T:C | C:T | 31 |
| Allele A | Both | CC | CT | C:C | C:T | 41 |
| | | CC | TC | C:T | C:C | 14 |
| Allele B | Allele A | TT | CC | T:C | T:C | 33 |
| Allele B | Both | TT | CT | T:C | T:T | 32 |
| | | TT | TC | T:T | T:C | 23 |
| Allele B | Allele B | TT | TT | T:T | T:T | 22 |
| Allele A | Allele B | CC | TT | C:T | C:T | 11 |

**Table 2.1** All possible genotype combinations for the two SNPs rs429358 and rs7412. Column one (Assigned Genotype) gives the genotype assigned to each sample/SNP as downloaded from the software, the software labels samples that have just the C allele as 'Allele A', just the T allele as 'Allele B' and having both the C and T allele as 'Both'. All possibilities of the actual genotype are written in column two (Actual Genotype). Column three and four (Haplotype 1 and Haplotype 2) are the two haplotypes, haplotype 1 is determined simply by taking the first base from column two; one from rs429358 and the second from rs7412, likewise haplotype 2 is determined by taking the second base. The final column (*APOE* ε status) is the *APOE* ε status which was determined by using Haplotype 1 and Haplotype 2 in **Table 1.4**. Rows highlighted in green are those used when directly translating the assigned genotype downloaded from the software to *APOE* ε status in Excel.

## 2.2. Results

A total of 335 samples had their DNA successfully extracted. A final set of 672 DNA samples were selected for genotyping on the NeuroX which included DNA from 480 cases (451 sEOAD) and 192 controls, the DNA for 263 of these cases were prepared as described in **Section 2.1.1**, the remaining DNA samples were either made by other members of our laboratory or were sent to us from other Alzheimer Research United Kingdom (ARUK) consortium members. The 672 DNA samples were sent to the Laboratory of Neurogenetics (National Institute of Aging USA) for genotyping on the NeuroX, data for a total of 640 samples was received; this included 450 cases and 190 controls. During quality control procedures it was evident that our UoN control data were of low quality and it was decided at this point to request data for additional European controls from University College London (UCL); they kindly permitted access to data for 560 UK controls (UCL controls). With this additional data our final data set had a total of 1200 samples; this included 450 cases and 750 controls. Data for all 1200 samples were included for quality control steps in **Section 4**.

47 sEOAD samples which did not have a documented *APOE* ε status were successfully genotyped for the two SNPs (rs7412 and rs429358) required. UCL control samples did not have an *APOE* ε status and DNA was not available to perform genotyping; it was decided to use the NeuroX data to genotype the UCL controls for the two SNPs required (**Section 5.1.2**).

# 3. Screening sEOAD Samples for Causative Variants in *APP*

## 3.1. Methods

451 sEOAD samples (**Table 3.1A**) were screened for causative mutations in exon 16 and 17 of the amyloid precursor protein gene (*APP*), this was achieved via Sanger sequencing in two sequencing reaction; one per exon. Additionally, there was a documented 6bp deletion (rs367709245; IVS17 83-88delAAGTAT) located downstream of exon 17 and the design of this PCR product also permitted genotyping of this variant. As rs367709245 appeared at slightly higher frequency in sEOAD and is located near an intron/exon boundary, its possible effect on splicing and generation of potential alternative isoforms of *APP* was investigated. Additionally, genotyping of rs367709245 was conducted in LOAD and controls using a custom designed KASP genotyping assay (**Table 3.1B Table 3.1C**)

**A sEOAD**

| Centre | N | Mean age at onset (±SD) | | Females (%) | | *APOE* ε4+ (%) | | *APOE* ε4ε4 (%) | | *APOE* ε4 MAF | D | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bristol | 24 | 53.5 | (5.3) | 11 | (45.8) | 12 | (50.0) | 3 | (12.5) | 0.31 | 24 | 0 |
| Manchester | 356 | 57.3 | (5.4) | 171 | (48.0) | 211 | (59.3) | 49 | (13.8) | 0.56 | 61 | 295 |
| Nottingham | 37 | 58.5 | (6.1) | 18 | (48.6) | 17 | (45.9) | 2 | (5.4) | 0.26 | 7 | 30 |
| Oxford | 34 | 55.5 | (4.2) | 19 | (55.9) | 20 | (58.8) | 4 | (11.8) | 0.35 | 24 | 10 |
| All | 451 | 57.0 | (5.5) | 219 | (48.6) | 260 | (57.6) | 58 | (12.9) | 0.51 | 116 | 335 |

**B LOAD**

| Centre | N | Mean age at onset (±SD) | | Females (%) | | *APOE* ε4+ (%) | | *APOE* ε4ε4 (%) | | *APOE* ε4 MAF | D |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Belfast | 295 | 78.1 | (6.3) | 187 | (63.4) | 161 | (54.6) | 24 | (8.1) | 0.31 | 295 |
| Bonn | 115 | 75.6 | (6.4) | 84 | (73.0) | 69 | (60.0) | 19 | (16.5) | 0.38 | 115 |
| Leeds | 103 | 79.4 | (6.1) | 55 | (53.4) | 67 | (65.0) | 0 | (0.0) | 0.40 | 103 |
| Nottingham | 71 | 78.1 | (6.0) | 48 | (67.6) | 40 | (56.3) | 3 | (4.2) | 0.30 | 71 |
| All | 584 | 77.8 | (6.4) | 374 | (64.0) | 337 | (57.7) | 61 | (10.4) | 0.34 | 584 |

**C CTRLS**

| Centre | N | Mean age at death (±SD) | | Females (%) | | *APOE* ε4+ (%) | | *APOE* ε4ε4 (%) | | *APOE* ε4 MAF |
|---|---|---|---|---|---|---|---|---|---|---|
| Belfast | 47 | 73.8 | (5.3) | 33 | (70.2) | 9 | (19.1) | 0 | (0.0) | 0.10 |
| Bonn | 134 | 74.9 | (7.6) | 74 | (55.2) | 41 | (30.6) | 3 | (2.2) | 0.16 |
| Leeds | 191 | 78.4 | (5.7) | 98 | (51.3) | 39 | (20.4) | 4 | (2.1) | 0.11 |
| Manchester | 3 | 83.7 | (6.1) | 2 | (66.7) | 0 | (00.0) | 0 | (0.0) | 0.00 |
| Nottingham | 111 | 77.4 | (7.5) | 41 | (36.9) | 22 | (19.8) | 0 | (0.0) | 0.10 |
| Oxford | 15 | 81.5 | (6.3) | 11 | (73.3) | 3 | (20.0) | 0 | (0.0) | 0.10 |
| Southampton | 27 | 77.1 | (5.5) | 12 | (44.4) | 6 | (22.2) | 0 | (0.0) | 0.11 |
| All | 528 | 76.9 | (6.8) | 271 | (51.3) | 120 | (22.7) | 7 | (1.3) | 0.12 |

**Table 3.1** Sample demographics of each cohort; sporadic early-onset Alzheimer's disease (sEOAD) (A), late-onset Alzheimer's disease (LOAD) (B) and control (CTRLS) (C). Each cohort contains samples from multiple centres and each centre is represented one per row. The number of samples from each centre (N) is given along with the mean age of onset with standard deviation (Mean age at onset (±SD)), the number and percentage of female samples (Females (%)), the number and percentage of samples harbouring at least one *APOE* ε4 allele (*APOE* ε4 + (%)), the number and percentage of samples with *APOE* ε4ε4 status (*APOE* ε4ε4 (%)), the minor allele frequency of the *APOE* ε4 allele (*APOE* ε4 MAF) and finally the number of samples classed as post mortem confirmed/definite Alzheimer's disease (D) or probable Alzheimer's disease (P). Key: N, number of samples; SD, standard deviation.

### 3.1.1. Primer design

Primers were designed using the following freely available software with default settings (links and versions provided in **Section 12**). Firstly, the DNA sequence of the base(s) of interest along with 50 bp upstream and downstream sequence was retrieved from ENSEMBL's genome browser, this sequence was in rich text format which highlighted all known variants. Using Primer3 and the sequence, primers where designed by marking highlighted variants as positions to avoid. *In Silico* PCR at UCSC's genome browser was used to check that the designed primers only bound to one unique site in the human genome. Finally SNPCheck was used to check the overall design of the primers. Where appropriate, NEBcutter was used to note the location and type of all restriction sites in the target sequence, thus enabling each of the primers used in **Section 3.1.3.1** to include a unique restriction site.

### 3.1.2. Sanger sequencing sEOAD samples

*APP* exon 16 was amplified using primers E16F 5' CAG-GTT-TCC-CTT-ACC-CTT-TC 3' and E16R 5' GCG-CTC-AGC-CTA-GCC-TAT-TT 3' (Eurogenomics), the product was 485bp in length. *APP* exon 17 was amplified using primers E17F 5' CAA-CCA-GTT-GGG-CAG-AGA-AT 3' and E17R 5' CAC-GGT-AAG-TTG-CAA-TGA-ATG 3' (Eurogenomics), the product was 482bp in length. Both amplicons were sequenced using the Sanger di-deoxy method in the forward direction using primer E17F or E16F, those found to harbour a causative variant were validated by sequencing in the reverse direction using reverse primer E16R or E17R.

gDNA was amplified in a final volume of 15µl using 2ng/µl gDNA, 1pM forward primer, 1pM reverse primer, 1X Buffer (BioLabs), 0.2mM dNTPs (Thermo Scientific), 0.1U/µl LongAmp Taq DNA polymerase (New England Biolabs) and molecular grade water to the required volume. The reaction was thermaly cycled with an initial denaturation step of 94°C for 2 minutes, followed by 35 cycles of 94°C for 30 seconds, 58°C (exon 16) or 61°C (exon 17) for 15 seconds, and 72°C for 45 seconds, with a final extension step at 72°C for 7 minutes. A reaction containing no DNA was included in all PCR batches as a negative control. The PCR products were cleaned using ExoSAP-IT (Affymetrix). Cleaned PCR products were Sanger sequenced in a final volume of 10µl using 4µl PCR product, 0.5pM primer, 1X BigDye Sequencing Buffer (Life Technologies), 0.25X BigDye Terminator v3.1 (Life Technologies) and molecular grade water to the required volume. The reaction was thermaly cycled for 25 cycles of 96°C for 30 seconds, 50°C for 15 seconds and 60°C for 4 minutes. The reactions were cleaned using Performa DTR Gel filtration Cartridges (Edge Biosystems) and eluent dried before sequencing was performed using the ABI 3130 automated sequencer. Sequence data was viewed using Sequence Scanner Software and compared using ClustalW (versions and links provided in **Section 12**).

### 3.1.3. Effect of the 6bp deletion on cDNA from COS7 and BE(2)-C cell lines

gDNA was amplified for two samples; one heterozygous for the deletion and the other wildtype. The PCR was achieved as described in **Section 3.1.2** using forward primer E17SF 5' CAA-ATA-GTC-GAC-CAA-CCA-GTT-GGG-CAG-AGA-AT 3' (Eurogenomics) which has an SalI restriction site at position 7 to 12, and reverse primer E17SR 5' GAG-CAG-TCT-AGA-CAC-GGT-AAG-TTG-CAA-TGA-ATG 3' (Eurogenomics) which has an Xbal restriction site at position 7 to 12. The product was 506bp in length which contained exon 17 (147bp) along with 167bp upstream and 192bp downstream intronic sequence.

Amplicon DNA was ligated with pCR 2.1-TOPO clone DNA (Invitrogen) and transformed into chemically competent One Shot TOP10 *E. coli* cells (Invitrogen). Cells were spread onto agar plates made using 4mg/ml Circlegrow (MP Biomedicals), 0.015g/ml of Agar (Invitrogen), 50µg/ml Ampicillin (Fisher) and 40µl/plate of 4% X-gal (Fisher) before they were incubated at 37°C for 16-18 hours. To confirm the presence/absence of the deletion in the *E. coli* cells the DNA was Sanger sequenced; white colonies were selected from the plates and DNA extracted using QIAprep Spin Miniprep Kit (QIAgen) and quantified using the NanoDrop 3300 Spectrometer (Thermo Scientific). The DNA was Sanger sequenced in a final volume of 10µl using 15-30ng DNA, 0.5pM forward primer M13F 5' GTA-AAA-CGA-CGG-CCA-G 3' (Invitrogen) or 0.5pM reverse primer M13R 5' GTC-ATA-GCT-GTT-TCC-TG 3' (Invitrogen), 1X BigDye Buffer (Applied Biosystems), 1X BigDye Terminator v3.1 (Life Technologies) and molecular grade water to the required volume. The reactions were thermal cycled and cleaned as in **Section 3.1.2**. The product was 660bp in length.

To obtain DNA inserts, a wild type and deletion clone were digested in a final volume of 20µl using 1µg clone, 1X digestion buffer (Thermo Scientific), 1U FastDigest SalI (Thermo Scientific), 1U FastDigest XbaI (Thermo Scientific) and molecular grade water to the required volume. Exon trap vector pET01 (MoBiTech) was digested in a final volume of 50µl using 1X digestion buffer (Thermo Scientific), 1U FastDigest SalI (Thermo Scientific), 1U FastDigest XbaI (Thermo Scientific), 2µl fast alkaline phosphatase (Fermentas) and molecular grade water to the required volume. Digestion was completed by incubating the reactions at 37°C for 30 minutes followed by 65°C for 20 minutes. Digestion products were separated using electrophoresis on a 1% agarose gel. Bands of DNA insert and linear vector (4017bp) were extracted from the gel using the QIAquick Gel Extraction Kit (QIAGEN).

DNA insert and pET01 vector were ligated using T4 DNA Ligase Kit (Invitrogen) and transformed into competent DH5-alpha *E. coli* cells (NEB). To confirm the presence/absence of the deletion in the *E. coli* cells, the DNA was Sanger sequenced as described above but using primer pET01S 5' GAT-CGA-TCC-GCT-TCC-TG 3' or pET01AS 5' GTC-ATA-GCT-GTT-TCC-TG 3' (Eurofin Genomics).

NucleoBond® Xtra Midi EF/Maxi EF kit (Macherey-Nagel) was used to extract pET01 vector (free of endotoxins) from the *E. coli* cells.

COS-7 and BE(2)-C cells were obtained from the European Collection of Cell Cultures (ECACC). COS-7 cells were cultured in Dulbecco's modified Eagle medium (Sigma) with 10% foetal bovine serum (Gibco), 2mM L-Glutamine, 100U/ml penicillin (Gibco), 100µg/ml streptomycin (Gibco) and 2.5µg/ml fungizone. BE(2)-C were cultured with 50% of Eagle's minimum essential medium and 50% Ham's F12 supplement with additional components added later; 1% of non-essential amino acids, 2mM L-Glutamine, 1mM Sodium Pyruvate, 10% foetal bovine serum, 100U/ml penicillin, 100µg/ml streptomycin and 100U/ml fungizone. The media was filtered through a 0.5µm filter, stored at 4°C, and warmed to 37°C before use. Cells were grown in 75cm$^2$ culture flasks in an incubator at 37°C with humidified atmosphere containing 5% carbon dioxide. Cells were split when they reached a confluence between 90-100%. Old medium was removed and cells were washed with PBS. Cells were detached from the bottom of the flask by trypsinisation using 3 ml of trypsin-EDTA solution and incubated at 37°C for 2-5 minutes. To stop trypsin action, 3ml of complete medium was added. An appropriate volume was transferred to a new flask and the volume made up to 15ml with complete medium. Cells were plated using 3.5x10$^5$ COS-7 cells or 6x10$^5$ BE(2)-C cells and 5ml of medium in 6cm culture wells followed by incubation at 37°C overnight.

Following overnight incubation, COS-7 and BE(2)-C cells were transfected with pET01 vector, repeated in triplicate for each cell line. A transfection mixture contained 1ug DNA, 9µl transfast (Promega) and 2ml of serum free medium was vortexed and incubated at room temperature for 15 minutes. Medium from the cells was removed and cells washed with 4ml PBS. Transfection mixture was added to the cells and incubated for 1 hour, before adding 4ml of complete medium and incubating for a further 24 hours. Cells were harvested by trypsinisation as describes in above and centrifuged for 5 minutes at 2350G.

Total RNA was extracted from the COS-7 and BE(2)-C cells using an RNeasy mini kit (Qiagen). RNA was treated with DNase (Ambion® TURBO DNA-free Kit) and cDNA was synthesised in duplicate using the AffinityScript Multi Temperature cDNA Synthesis Kit (Agilent) with either 0.3µg random primer (Agilent) or 0.5µg Oligo(dT) primers (Agilent), along with 2µg of RNA. Reactions containing no enzyme were included as a negative control.

cDNA was amplified in a final volume of 30µl using 500ng cDNA, 0.5pM sense primer pET01S (Eurofins Genomics), 0.5pM antisense primer pET01AS (Eurofins Genomics), 1X Buffer (Roche), 1.5mM $MgCl_2$ (Roche), 0.2mM dNTPs (Thermo Scientific), 5U/µl High Fidelity Taq DNA polymerase (Roche) and molecular grade water to the required volume. The product was 388bp in length if exon 17 was retained or 241bp if exon 17 was removed. The reaction mixture was thermally cycled with an initial extension at 94°C for 2 minutes followed by 35 cycles of 94°C for 30 seconds,

58°C for 30 seconds and 72°C for 1 minute, with a final extension step of 72°C for 7 minutes. 5µl of the PCR product was cleaned and Sanger sequenced as described in **Section 3.1.2** using forward primer pET01S for the cDNA from COS7 cells and reverse primer pET01AS for the cDNA from BE(2)-C cells.

### 3.1.4. Effect of the 6bp deletion on cDNA from Patient brain cells

Brain tissue from a case that harboured rs367709245 was used to investigate the resulting cDNA. Cerebellum and temporal cortex was obtained from patient O0483 from the Oxford OPTIMA Study; the tissue was received in dry ice and homogenised using liquid nitrogen and a mortar and pestle. Total RNA was extracted and cDNA synthesised as described in **Section 3.1.3**.

cDNA was amplified and sequenced as described in **Section 3.1.2** using 1ul cDNA, forward primer E16SF 5' TGG-ATG-CAG-AAT-TCC-GAC-ATG 3' (Eurofins Genomics) and reverse primer E18SR 5' TTC-TGC-TGC-ATC-TTG-GAC-AG 3' (Eurofins Genomics). The product was 253bp in length if exon 17 was retained or 106bp if exon 17 was removed. Reactions containing no cDNA or no enzyme (reverse transcriptase) were included as negative controls.

### 3.1.5. Genotyping the 6bp deletion in LOAD and control samples

Genotyping of rs367709245 in LOAD and control samples was performed in a final volume of 8ul using 1X MasterMix (KASP), 1X Assay (KASP), 1.25ng/µl DNA and molecular grade water to the required volume. Reactions were thermally cycled with an initial step of 94°C for 15 minutes, followed by 10 cycles of 94°C for 20 seconds, 61°C (decreasing 0.6°C every cycle) for 60 seconds, 32 cycles of 94°C for 20 seconds, 55°C for 60 seconds, and a finale step of 30°C for 60 seconds. A further 6 cycles of 94°C for 20 seconds and 55°C for 60 seconds was conducted to generate tighter clusters when appropriate. All samples positive for the deletion were confirmed by Sanger sequencing in the forward direction using the protocol described in **Section 3.1.2.**

### 3.1.6. Susceptibility to sEOAD and LOAD from the 6bp deletion

Power was calculated using Quanto in a 'gene-only' hypothesis and 'log-addictive' model. Fisher's exact test and logistic regression was performed to assess the level of association of rs367709245 with sEOAD and LOAD. Using Excel, genotyping results were translated into PLINK conventional files, PLINK[223] (link and version provided in **Section 12**) was used to perform fishers exact test and logistic regression test while adjusting for *APOE* ε2 and ε4 dosage together with gender.

```
$ plink --noweb --file plinksEOADControl --fisher --logistic --covar --covar-name
  GENDER,APOE2,APOE4 --out plinksEOADControlFisher
$ plink --noweb --file plinkLOADControl --fisher --logistic --covar --covar-name
  AGE,GENDER,APOE2,APOE4 --out plinkLOADControlFisher
```

## 3.2. Results and discussion

### 3.2.1. Sanger sequencing sEOAD samples

All 451 sEOAD samples were successfully Sanger sequenced with good coverage of all documented causative variations in exons 16 /17 and the 6bp deletion (rs367709245, IVS17 83-88delAAGTAT). None of the samples harboured a causative variant in exon 16, however sequencing identified four samples (9% of cohort) heterozygous for one of three known causative variants in exon 17, this is a higher frequency when compared to previous studies; Sassi and colleagues found causative variants in 2% (n=1) of 47 EOAD samples,[224] and Nicolas and colleagues found no causative variants in 264 EOAD samples.[225] Sequencing also identified five sEOAD samples heterozygous for rs367709245 and sequencing confirmed these genotypes. **Table 3.2** lists the nine sEOAD samples harbouring either a known causative variant or the 6bp deletion in *APP*, **Figure 3.1** and **Figure 3.2** show the Sanger sequence results for these samples.

| Sample | Centre | Gender | Diagnosis | AAO | *APOE ε* | Cohort | Variant |
|---|---|---|---|---|---|---|---|
| M562 | Manchester, UK | F | D | 62 | 33 | sEOAD | p.Ala713Thr |
| MRC045 | Nottingham, UK | F | Pr | 61 | 34 | sEOAD | p.Ala713Thr |
| O0691 | Oxford, UK | M | Pr | 50 | 44 | sEOAD | p.Val717Ile |
| M758 | Manchester, UK | F | Pr | 61 | 33 | sEOAD | p.Val717Gly |
| O0483 | Oxford, UK | M | D | 51 | 34 | sEOAD | rs367709245 |
| M111 | Manchester, UK | F | Pr | 54 | 33 | sEOAD | rs367709245 |
| M200 | Manchester, UK | F | Pr | 63 | 33 | sEOAD | rs367709245 |
| M400 | Manchester, UK | F | Pr | 50 | 34 | sEOAD | rs367709245 |
| M461 | Manchester, UK | F | Pr | 64 | 44 | sEOAD | rs367709245 |

**Table 3.2** Samples identified as harbouring either a known causative variant or the 6bp deletion in *APP* during Sanger sequencing. The sample ID is given (Sample) along with the origin of the sample (Centre), gender (Gender), diagnosis (Diagnosis), age at onset of disease (AAO), *APOE* ε status (*APOE* ε), cohort of the sample (Cohort) and finally the variant the sample harbours given as the protein change or rsID (Variant). Key: AAO, age at onset; UK, United Kingdom; F, female; M, Male; D, definite/confirmed Alzheimer's disease; Pr, probable Alzheimer's disease; N/A, not applicable; sEOAD, sporadic early onset Alzheimer's disease; rsID, reference single nucleotide polymorphism

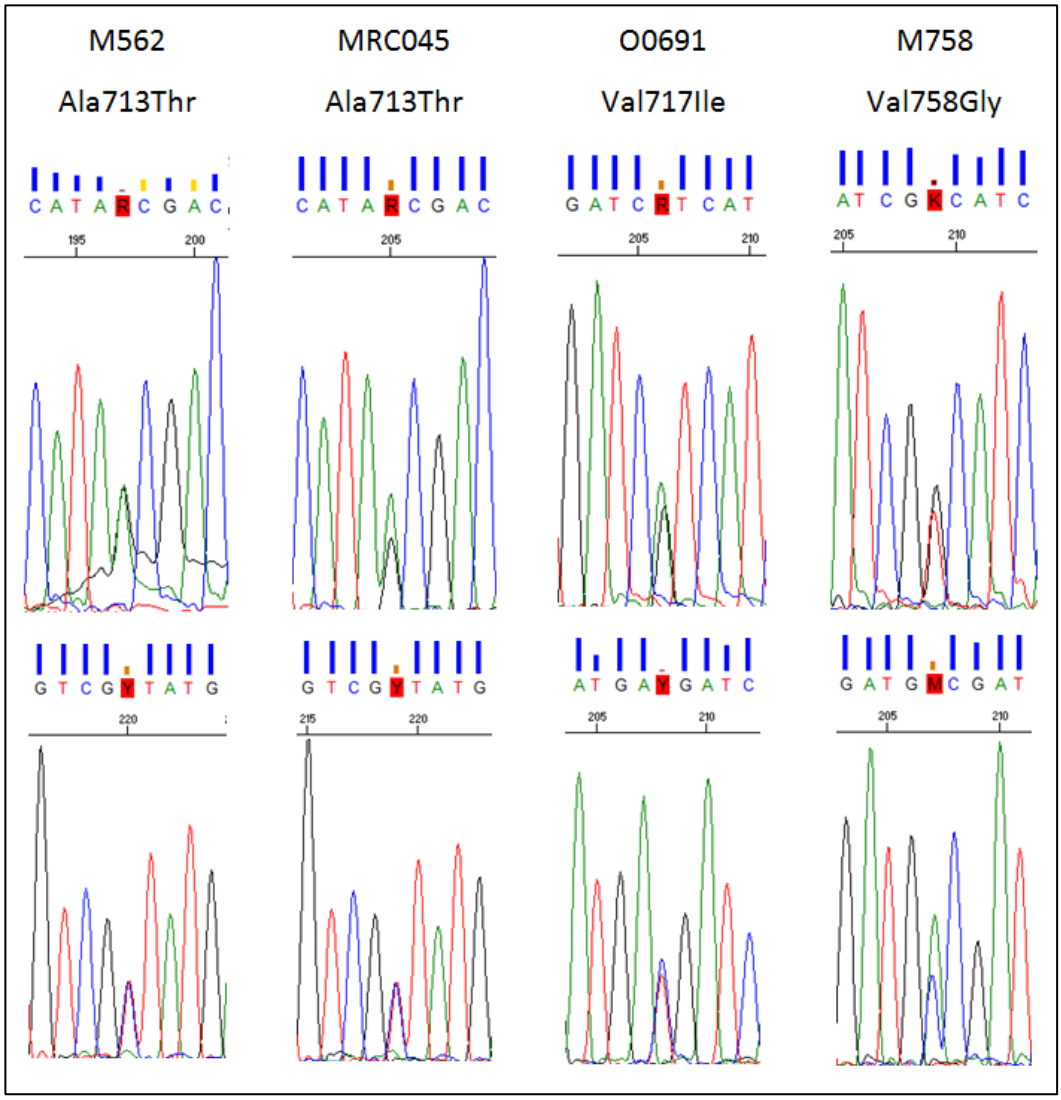



**Figure 3.1** Sequence chromatograms from four sEOAD samples harbouring a known causative variant in *APP*. Each sample is presented per column. The chromatograms were generated using Sequence Scanner Software (Applied Biosystems) and shows the heterozygous causative variant (highlighted in red) at the centre with 8 surrounding bases. The top image in each column shows the sequence in the forward direction using primer E17F and the bottom image sequence in the reverse direction using primer E17R. Abbreviation: APP, amyloid precursor protein.

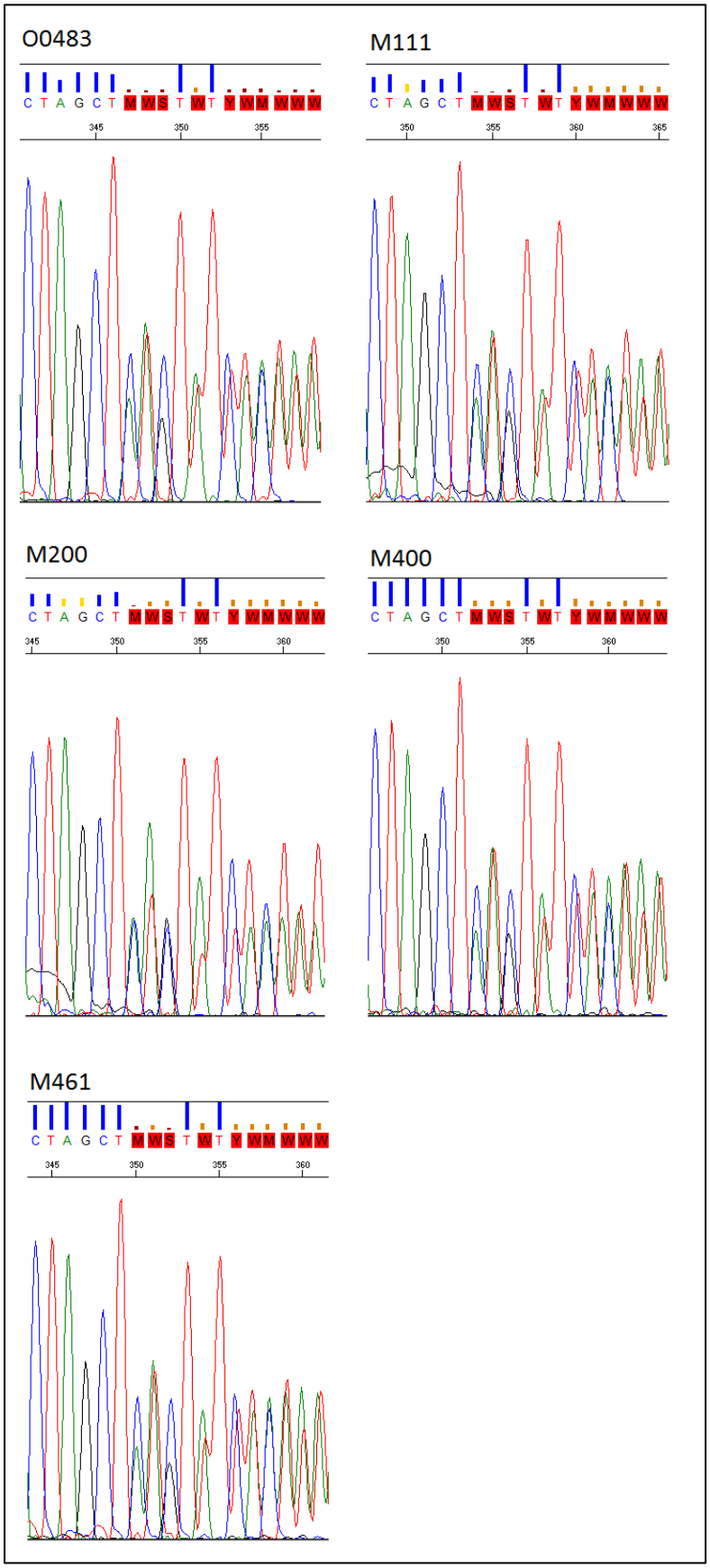



**Figure 3.2** Sequence chromatograms from five sEOAD samples harbouring the 6bp deletion (rs367709245, IVS17 83-88delAAGTAT).
The chromatograms were generated using Sequence Scanner Software (Applied Biosystems) and shows the heterozygous 6 bp deletion starting at base 7 of the chromatograms

### 3.2.2. Effect of the 6bp deletion on cDNA from COS7 and BE(2)-C cell lines

COS7 cells and BE(2)-C cells were transfected with minigene constructs with either the 6bp deletion (rs367709245, IVS17 83-88delAAGTAT) or wildtype insert. PCR of cDNA from these cells showed no alternative isoforms; a single amplicon was obtained of expected size in all instances (388 bp – **Figure 3.3**) and sequencing of all constructs confirmed exon 17 was retained. (**Figure 3.4**)

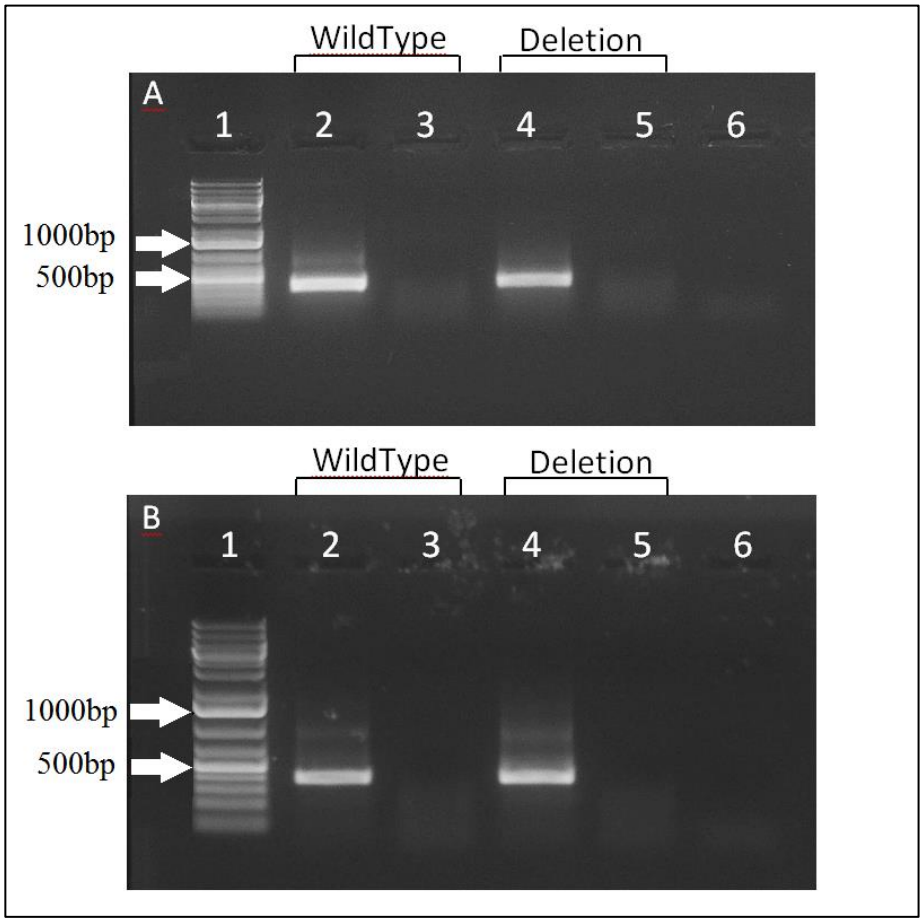



**Figure 3.3** RT-PCR products from COS7 cells (A) and BE(2)-C cells (B) transfected with pET01 vector containing wildtype or 6bp deletion (rs367709245, IVS17 83-88delAAGTAT) insert. Lane 1 contains the DNA ladder GeneRuler 1kb Plus (Life Technologies) which has 2 reference bands; 1000bp (top) and 500bp (bottom). Lane 2 contains the RT-PCR product from wildtype insert, lane 3 being its respective RT-PCR negative control (no enzyme). Lane 4 contains the RT-PCR product from the 6bp deletion insert with lane 5 being its respective RT-PCR negative control (no enzyme). Lane 6 contains the RT-PCR negative control (no RNA). Lanes 2 and 4 each contain a single band of 350-400bp, this is consistent with 388bp; the length of the amplicon when exon 17 is retained. Key: RT-PCR, reverse transcription polymerase chain reaction.

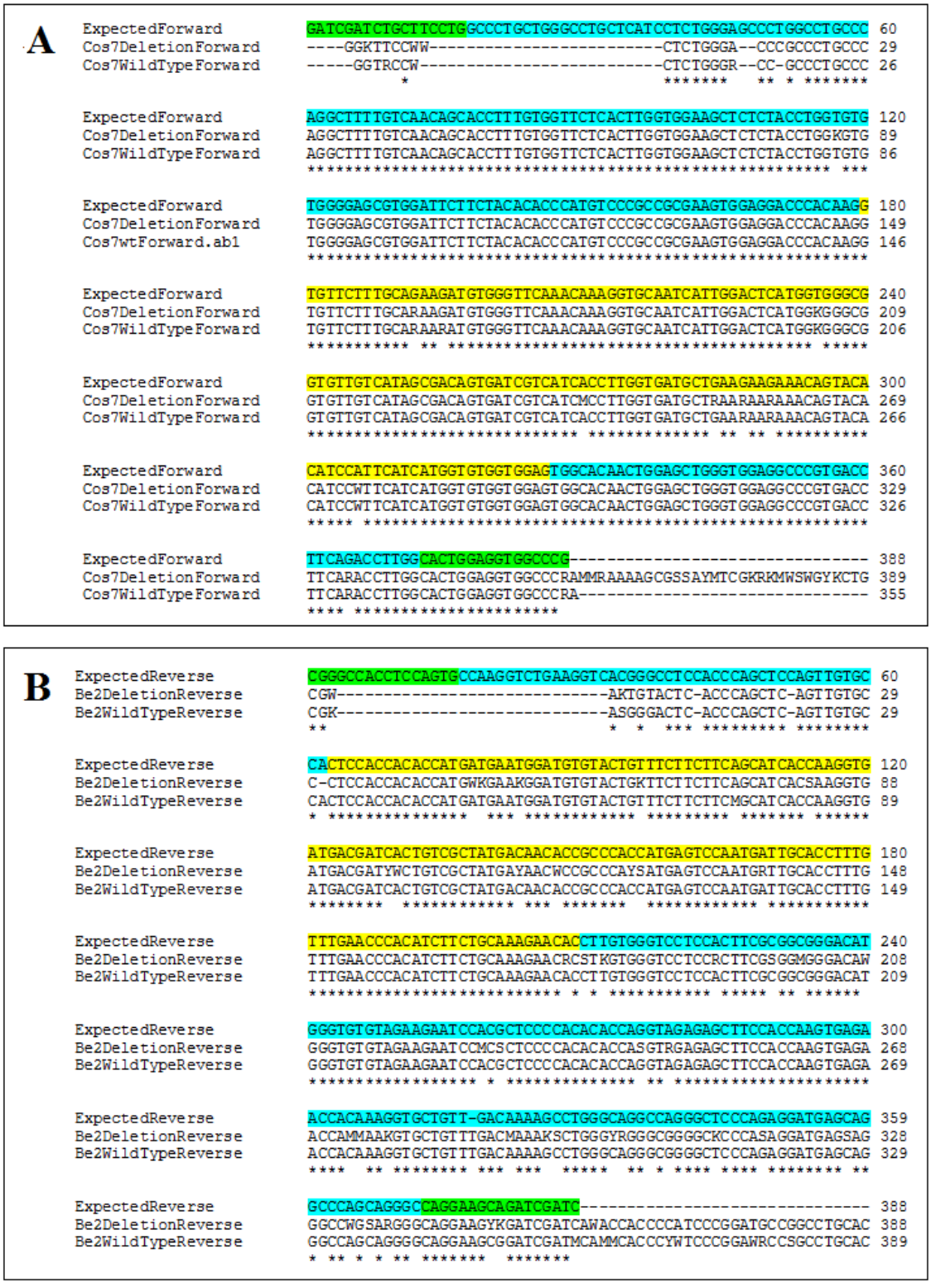



**Figure 3.4** Sequence results from Sanger sequencing the COS7 constructs (A) and BE(2)-C constructs (B). Each image shows the alignment results from ClustalW when comparing the expected sequence to the wildtype and the 6bp deletion (rs367709245, IVS17 83-88delAAGTAT) constructs. The expected sequence contains the exon of the pET01 vector (blue and green) with exon 17 inserted in the middle (yellow). Sequence highlighted in green is where the Sanger sequencing primers bound. All sequences match the expected sequences to a high degree, including the section containing exon 17 (yellow), confirming that exon 17 is retained in transcripts from both wildtype and 6bp deletion constructs.

### 3.2.3. Effect of the 6bp deletion on cDNA from patient brain cells

PCR of cDNA from cerebellum and temporal cortex from a patient harbouring the 6bp deletion (rs367709245, IVS17 83-88delAAGTAT) showed no alternative isoforms; a single amplicon was obtained of expected size in both instances (253bp – **Figure 3.5**) and sequencing of both cDNAs verified exon 17 was retained (**Figure 3.6**). In agreement with the results presented in **Section 3.2.2**, this approach indicates a lack of alternative splicing in the presence of rs367709245; given its location from the exon intron boundary (83bp), it could be too far away to influence splicing. However, it cannot be ruled out that it may have alternative functional effects on APP, for example, via gene regulation.

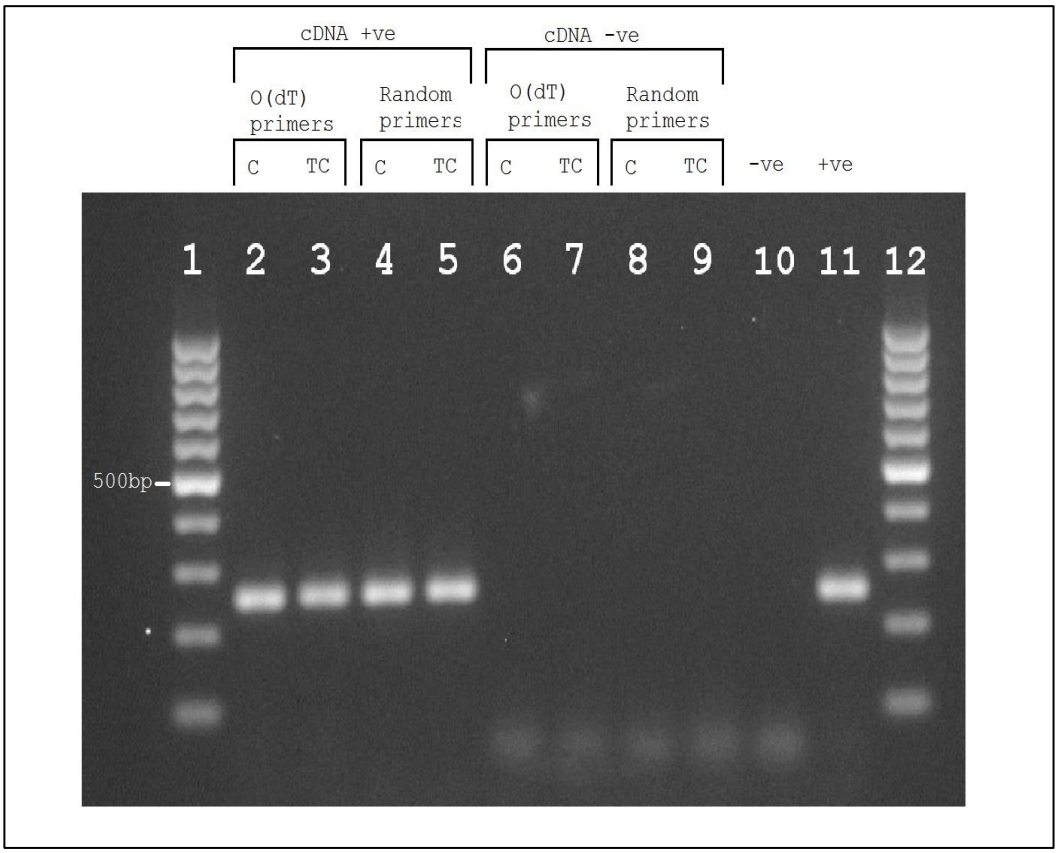



**Figure 3.5** RT-PCR products from sample O0483 who harbours rs367709245. Reverse transcriptase polymerase chain reaction (RT-PCR) products generated using Oligo(dT) or random primers and RNA from sample (O0483) cerebellum (C) and temporal cortex (TC). Lanes 1 and 12 contain GeneRuler 100bp Plus DNA ladder (ThemoFisher Scientific) containing 1 reference band marked at 500bp. Lanes 2 and 3 contain the RT-PCR products generated using Oligo(dT) primers. Lanes 4 and 5 contain the RT-PCR product generated using random primers. Lanes 6 to 9 are the negative controls for lanes 2 to 5 with no reverse transcriptase enzyme (cDNA -ve). Lane 10 contains an RNA negative control (-ve) and lane 11 contains a PCR positive control (+ve) which was cDNA generated from lymphoblastoid cell lines from the NHGRI sample repository.

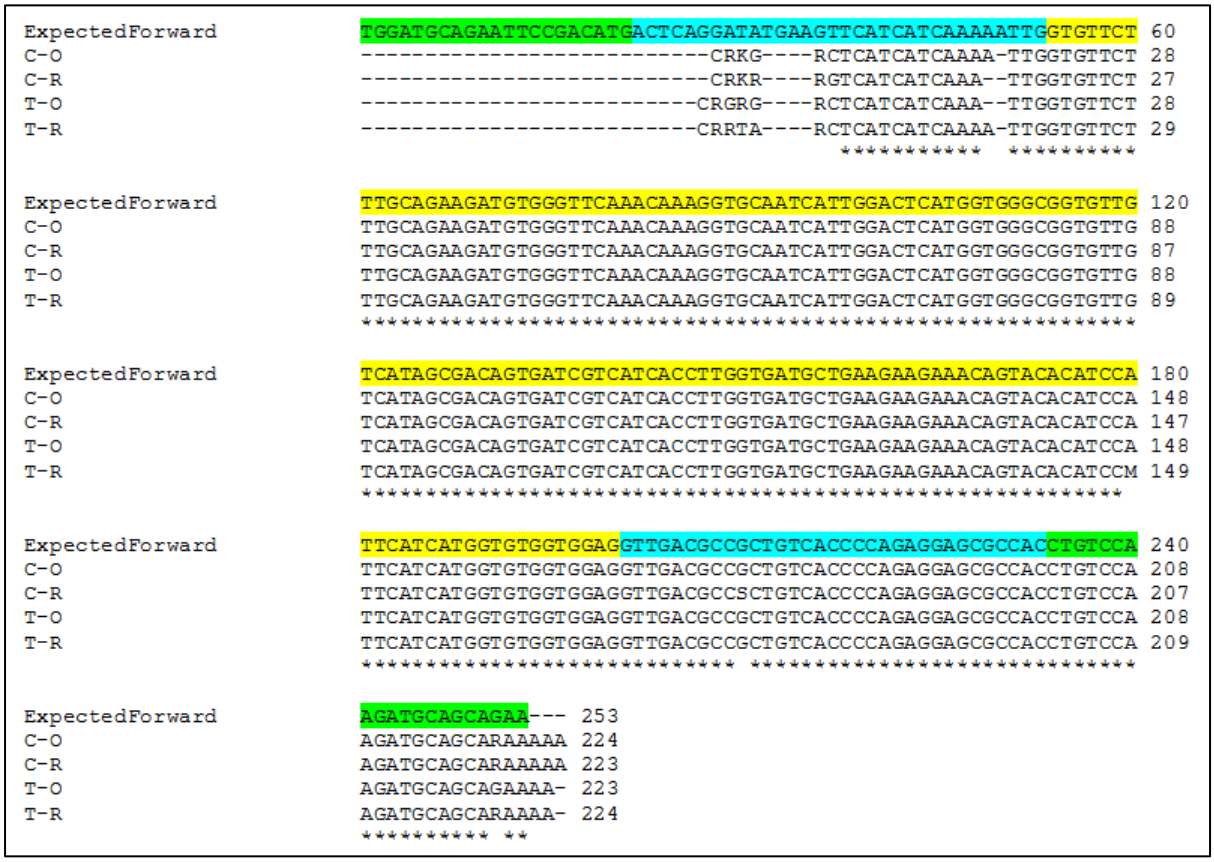



**Figure 3.6** Sequence results from Sanger sequencing the cDNA generated from cerebellum (C) and temporal cortex (T) using Oligo(dT) (O) or random primers (R) and RNA from a sample (O0483) which harboured the 6bp deletion (rs367709245). Each image shows the alignment generated from ClustalW when comparing the expected forward sequence to four cDNAs. The expected forward sequence contains part of exon 16 (green and blue) followed by exon 17 (yellow) and finally part of exon 18 (blue and green). Sequence highlighted in green is where the Sanger sequencing primers bound. All sequences match the expected sequences to a high degree, including the section containing exon 17 (yellow), confirming that exon 17 is retained in transcripts from all sources.

### 3.2.4. Genotyping the 6bp deletion in LOAD and control samples

584 LOAD and 528 control samples were successfully genotyped using TaqMan. Genotyping identified two LOAD and two control samples heterozygous for the 6bp deletion (rs367709245, IVS17 83-88delAAGTAT) and sequencing confirmed these genotypes. **Table 3.3** lists the four samples harbouring the 6bp deletion in *APP*; **Figure 3.7** shows the Sanger sequence results for these samples.

| Sample | Centre | Gender | Diagnosis | AAO/AAS | *APOE* ε | Cohort |
|---|---|---|---|---|---|---|
| LE328 | Leeds, UK | M | D | 70 | 34 | LOAD |
| LE431 | Leeds, UK | F | D | 76 | 34 | LOAD |
| BN0424 | Bonn, Germany | F | N/A | 84 | 33 | Control |
| BN1842 | Bonn, Germany | F | N/A | 68 | 33 | Control |

**Table 3.3** Controls identified as harbouring the 6bp deletion in *APP* (rs367709245) via high throughput KASP genotyping. The sample ID is given (Sample) along with the origin of the sample (Centre), gender (Gender), diagnosis (Diagnosis), age at onset of disease (AAO) or age at sampling (AAS), *APOE* ε status (*APOE* ε), cohort of the sample (Cohort). Key: AAO, age at onset; AAS, age at sampling; UK, United Kingdom; F, female; M, Male; D, definite/confirmed Alzheimer's disease; N/A, not applicable; LOAD, late onset Alzheimer's disease.

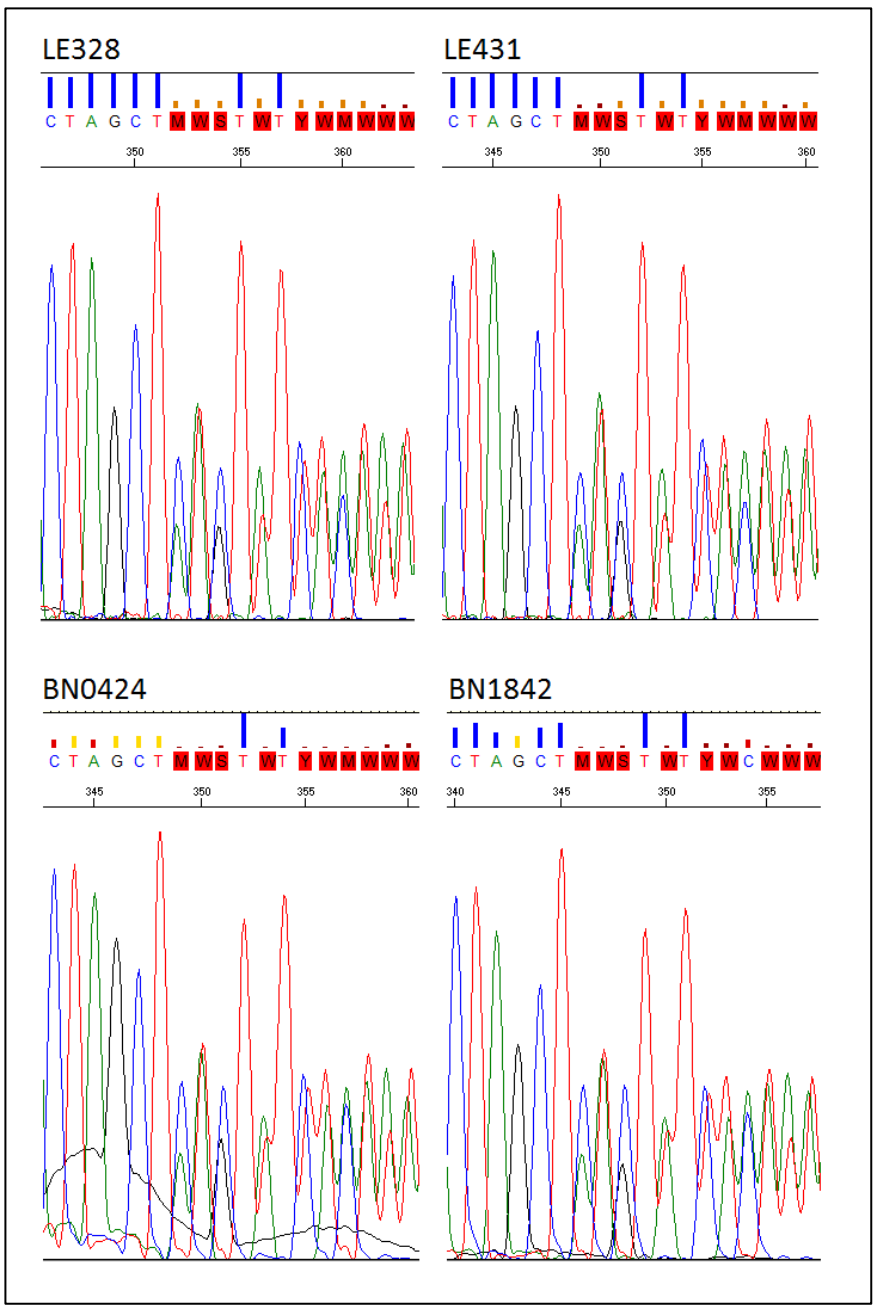



**Figure 3.7** Sequence chromatograms from four samples harbouring the 6bp deletion (rs367709245, IVS17 83-88delAAGTAT). The chromatograms were generated using Sequence Scanner Software (Applied Biosystems) and shows the heterozygous 6bp deletion starting at base 7 of the chromatograms

### 3.2.5. Susceptibility to sEOAD and LOAD from the 6bp deletion

Frequency calculations established the 6bp deletion (rs367709245, IVS17 83-88delAAGTAT) had a minor allele frequency (MAF) of 0.002 in the LOAD cohort (n=584) and control cohort (n=528), and a MAF of 0.006 in the sEOAD cohort (n=451). Assuming rs367709245 has a MAF of 0.002 and an odds ratio of two the power to detect association in this instance is 14% and 13% for LOAD and sEOAD cohorts respectively. Logistic regression failed to produce test statistics when testing rs367709245 in sEOAD and LOAD, most likely due to the small number of alleles involved. Fishers exact test established that rs367709245 was non-significant in sEOAD (p.value = 0.49) and non-significant in LOAD (p.value = 1); this non-significance in sEOAD is due to the lack of power. To verify the MAF in sEOAD and LOAD and to establish if rs367709245 is associated with either, a study using over 5,800 cases with an equal number of matched controls would be needed to achieve 80% power.

# 4. NeuroX Data Quality Control in GenomeStudio

## 4.1. Methods

Genotype calling for the raw NeuroX data was achieved using Illumina's GenomeStudio software (version and link given in **Section 12**). A cluster file generated by the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium was used from their best practices paper,[226] which was available to download freely from their webpage (link provided in **Section 12**). The CHARGE cluster file was created for an earlier un-customised version of the HumanExome BeadChip, version 0.1 as opposed to 0.2 which the NeuroX v1 uses. Despite this difference the overlap in markers appears to be considerable, and the CHARGE cluster file provides the only freely available cluster file which has been created from thorough inspection of the markers, it clusters 218,600 of the 267,607 markers on the NeuroX, and the remaining 49,007 markers fall into three categories:

- Markers that are part of the NeuroX custom content (n=24,706)
- Markers the CHARGE consortium discarded/zeroed (n=7,879)
- Markers whose names were different between the two versions of the chip (n=16,422)
  These markers have a '_ver' suffix on version 0.2 of the chip, for example the marker 'exm1917884' on version 0.1 is named 'exm1917884_ver3' on version 0.2

In GenomeStudio markers are referred to as SNPs despite some markers genotyping insertions and deletions on the chip, to avoid confusion in this chapter markers will be referred to as SNPs. Each SNP genotyped on the NeuroX had a SNP graph which contained clusters used to define the genotype of each sample (**Figure 4.1**).

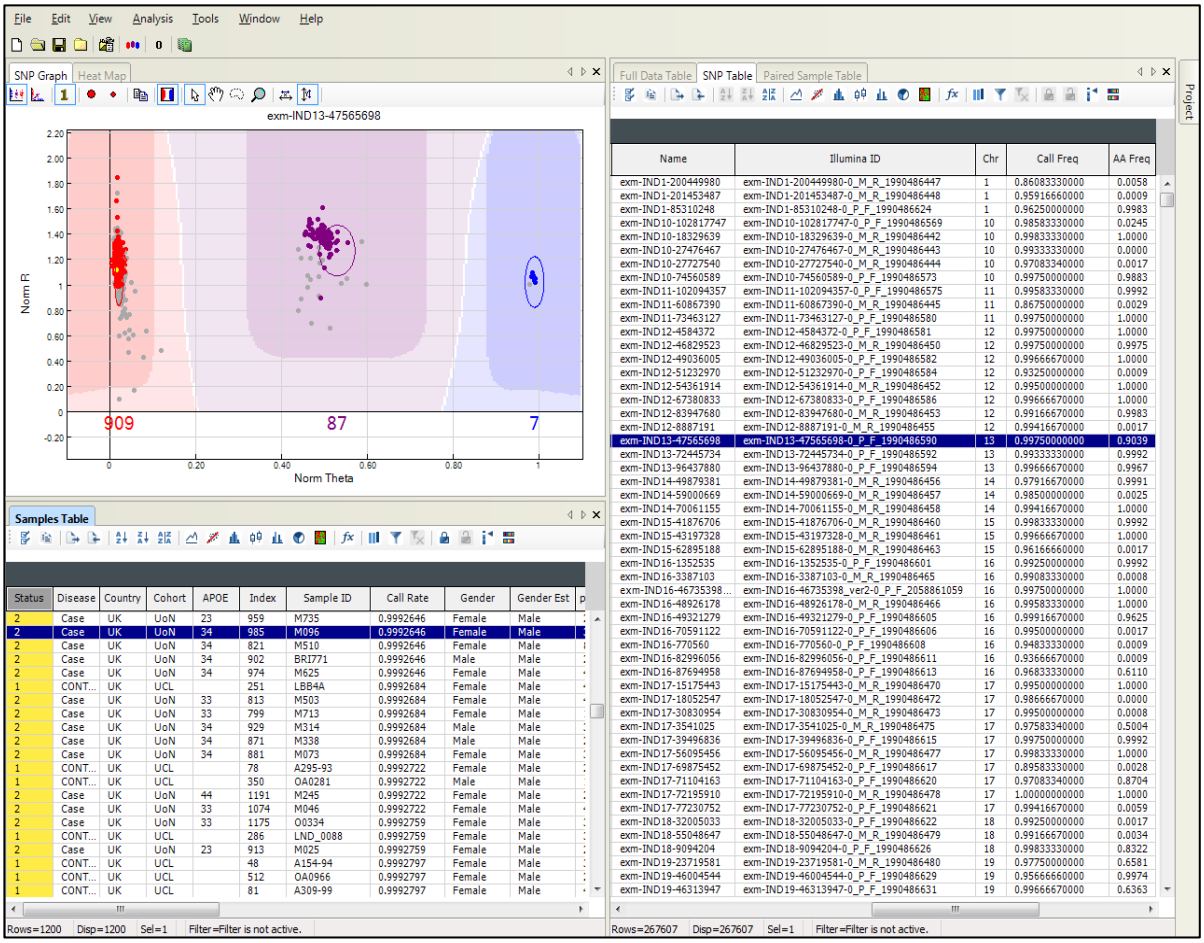



**Figure 4.1** Screenshot of Illumina's GenomeStudio software. Visible on the right is the SNP Table where each SNP is listed and one SNP is highlighted in blue. Bottom left is the Sample Table where each sample is listed and one sample is highlighted. Top left is the SNP Graph for the selected SNP, each sample appears as a coloured dot, the selected sample appears yellow and samples that have been excluded appear in grey. The relative position of each sample does not change on the graph, however the position, size and shape of the three hollow ovals on the graph can change and it is this that determines the cloud and the genotype of each sample.

Due to the huge quantity of SNPs on the NeuroX, they could not all be visually inspected to assess correctness, thus in this chapter different sets of criteria were used to isolate SNPs that potentially needed correcting. Each SNP isolated was visually inspected and either left as it was, corrected or zeroed. Two additional sources of information were consulted to help with this decision process:

1. CHARGE supplementary methods (CHARGE_ExomeChip_Best_Practices_V7.pdf) available from their webpage (link provided in **Section 12**)
2. Illumina technical note (technote_infinium_genotyping_data_analysis.pdf) available from their webpage (link provided in **Section 12**)

SNPs were zeroed if they fulfilled any of the following criteria:

- SNP failure

- GenomeStudio didn't allow correct clustering
  - Clusters were too close to one another
  - All clusters were shifted too far to one side of the graph
  - Clusters were placed slightly one on top of the other thereby encroaching on each cluster's vertical space
- SNP with a non-logical number of clusters
  - More than three clusters. These usually presented with four clusters (with two heterozygous clusters). Some SNP graphs were seen with as many as seven clusters
  - SNPs on a haploid chromosome with obvious heterozygous clusters
- SNPs with clusters that couldn't be defined with confidence
  - Low intensity
  - Too dispersed
  - Too horizontally elongated
  - Possible overlapping
  - One single cluster in the centre of the graph

### 4.1.1. Creating a project

Two data files were received for each sample, these took the form of intensity files (.idat), one file contained the intensities for red fluorescence and the other green fluorescence. The files were named according to the chip number (sentrix bar code) and the position on the chip (sentrix position) where the samples were loaded. The sample intensity data was located within subdirectories according to the sentrix bar code they were from. For example 9221305133_R01C01_grn.idat and 9221305133_R01C01_red.idat represents intensity data for one sample , the sentrix barcode for this sample was '9221305133', and this data was located in a subdirectory named '9221305133' along with other sample data with the same sentrix barcode.

Also required was a sample manifest that linked the sample ID to the sentrix bar code and position, this was provided along with the intensity data and additional information was added to aid with data quality control, such as gender, case/control and origin (UCL/UoN). Additionally, the NeuroX manifest was also used; this held all the information about the SNPs and was downloaded from the Illumina website.

A genotyping project was created using the 'Use sample sheet to load sample intensities' option in GenomeStudio. The following data were uploaded by pointing to the relevant location and SNP statistics were updated when prompted.

- NeuroX manifest (NeuroX_15036164_A.bpm)
- Sample intensity data for each sample (9221305133_R01C01_Grn.idat and 9221305133_R01C01_Red.idat for example)
- Sample manifest (Sample_Manifest_UoN_ControlsCases_and_UCL_Controls.csv)
- Cluster File (CHARGE_ExomeChip_v1.0_Cluster_File.egt)

Once the project was created and all data had been loaded several options were amended in 'Tools' → 'Options' → 'Project'; 'Exclude female Y-SNPs from SNP statistics' was checked, 'plot excluded samples' unchecked and 'use for all new projects' selected.

### 4.1.2. Uploading external information

An extra column was loaded into the SNP Table to annotate 121 SNPs that have recently been identified as located on the pseudo-autosomal region (described in **Section 5.1.1**). Table 4.1 lists the 121 SNPs.

| Marker | | | |
|---|---|---|---|
| exm1624485 | exm2263276 | exm1624887 | exm1625025_ver2 |
| exm1624661 | exm2263278_ver3 | exm1648651 | exm1625046 |
| exm1624677 | exm2263279 | exm1625041 | exm1625162 |
| exm1624783 | exm2263280 | exm1625556_ver2 | exm1625204 |
| exm1624797 | exm2263281 | exm1624641 | exm1625383_ver2 |
| exm1624862 | exm2268448 | exm1625029 | exm1625553 |
| exm1624867 | exm2273075 | exm1625819 | exm1625780 |
| exm1624886 | exm2273161 | exm1624434 | exm1648550 |
| exm1624907 | exm2273221 | exm1625064 | exm1667365 |
| exm1624944 | exm2273222 | exm1625554 | exm1624484 |
| exm1624960 | exm2273223 | exm1625555_ver3 | exm1624788 |
| exm1624975 | exm2273224 | exm1625807 | exm1624791 |
| exm1624995 | exm2273278 | exm1667356 | exm1624872 |
| exm1624999_ver3 | exm1624939_ver2 | exm1667360 | exm1624917 |
| exm1625001 | exm1624946 | exm2206194 | exm1624943_ver3 |
| exm1625200 | exm1625030_ver4 | exm1625045 | exm1624953 |
| exm1625277 | exm-rs5941436_ver2 | exm1625545 | exm1624965 |
| exm1625538 | exm-rs2573905 | exm1624766 | exm1624996_ver3 |
| exm1625569 | exm1625510 | exm1625068 | exm1625022 |
| exm1625571 | exm2264787 | exm1625216 | exm1625063 |
| exm1625573 | exm-rs525869_ver4 | exm1625222 | exm1625610 |
| exm1648538 | exm1624877 | exm1625253 | exm1625631 |
| exm1667350 | exm1667344 | exm1648569 | exm1648554 |
| exm1667504 | exm2273163 | exm2208735 | exm1667351 |
| exm1667508 | exm2273277 | exm1624792 | exm1667363 |
| exm1667537 | exm2264764 | exm1624794 | exm2209863 |
| exm1667541 | exm1624934 | exm1624804 | exm2248602 |
| exm2262791 | exm1667357 | exm1624869 | exm2248962 |
| exm2263170 | exm1625534 | exm1624931 | |
| exm2263174 | exm1625323 | exm1624954_ver3 | |
| exm2263176 | exm1624879 | exm1625005 | |

**Table 4.1** List of 121 markers updated to the pseudo-autosomal region (XY)

### 4.1.3. Excluding poor quality samples

Poor quality samples were temporarily removed during assessment of the SNPs to allow SNP graphs to be more accurate. On the Sample Table the call rate was sorted in ascending order. All samples with call frequency less than 0.99 (Call Freq<0.99) were excluded.

### 4.1.4. Automatic clustering of un-called SNPs

All SNPs that were not clustered by the CHARGE cluster file (n=49,007) were clustered using GenomeStudio's automated clustering function. To achieve this, SNPs were filtered to display only those with a call frequency equal to zero (Call Freq=0), the resulting list of SNPs were clustered by selecting all SNPs on the SNP Table and selecting Cluster Selected SNPs**.** SNP statistics were updated when prompted.

### 4.1.5. Assessing non-autosomal SNPs

SNPs on non-autosomal chromosomes were assessed separately to those on autosomal chromosomes as their cluster graphs are defined differently, this included SNPs on the Y chromosome, X chromosome, mitochondrial chromosome (MT) and pseudo autosomal region (PAR). Before inspection the number of expected clusters for SNPs on the Y or MT chromosomes (n=358) was updated to two, this was achieved by filtering the SNP Table to display only SNPs on the Y and MT chromosomes (Chr='Y' OR Chr='MT'), selecting all SNPs on the SNP Table, selecting SNP Properties and setting Expected Number of Clusters to 2. All filters were then removed from the SNP Table.

SNPs on the Y chromosome (n=139) were inspected first. These SNPs should not have any heterozygotes called and technically no females, however given the possibility that a small portion of samples would have the wrong gender, SNP graphs with a small portion of females called were allowed. The SNP Table was filtered to display only SNPs on the Y chromosome that were not in the list of markers located on recently identified PARs as described in **Section 4.1.2** (Chr='Y' AND PAR='No'). The Samples Table was sorted by gender and all female samples highlighted in the SNP Graph. Each SNP in the list was inspected individually as described in **Section 4.1** before removing all filters on the SNP Table.

SNPs on the MT chromosome (n=219) were inspected next. These SNPs should not have any heterozygotes and all samples should be called. The SNP Table was filtered to display only SNPs on the mitochondrial chromosome (Chr='MT'). Each SNP in the resulting list was inspected as described in **section 5.1** before removing all filters on the SNP Table.

SNPs on the X chromosome were then inspected. These SNPs should have all samples called and technically have no male heterozygotes, however given the possibility a small portion of samples would have the wrong gender, SNP graphs with a small portion of male heterozygote calls were

allowed. These markers included 5,644 of the 5,765 SNPs on the X chromosome; 121 of these SNPs have recently been located on PAR (**Section 4.1.2**), the SNP table was filtered to display markers on the X chromosome and excluding these 121 SNPs (Chr='X' AND PAR='No'). Given the high number of X chromosome SNPs in this list, additional filters where applied to isolate a smaller portion to visually inspect. The first filter applied identified SNPs with a high male heterozygote frequency (AB Freq); this was achieved by first calculating SNP statistics for just male samples by excluding female samples in the Samples Table and updating SNP statistics when prompted. SNPs were sorted in the SNP Table in AB Freq descending order and those with high male heterozygote frequency (AB Freq≥0.07) were zeroed by selecting those makers and selecting Zero Selected SNPs. To recalculate the SNP statistics on all samples, the AB Freq filter was removed and female samples were reintroduced by selecting Include Selected Samples and updating SNP statistics when prompted. The second filter applied identified SNPs with a low call frequency, a filter was applied to the SNPs Table to identify SNPs that had not yet been zeroed (active SNPs) and had a call frequency less than 99% (Call Freq!=0 AND Call Freq<0.99). Each SNP in the resulting list was inspected as described in **Section 4.1** before removing all filters on the SNP Table.

SNPs on pseudo-autosomal region (PAR) were inspected, this included 2 SNPs on PAR and an additional 121 SNPs which have recently been located on PAR (**Section 4.1.2**). These markers should produce a normal SNP Graph with one homozygous major cluster, one heterozygous cluster and one homozygous minor cluster with all samples called. The SNP Table was first filtered to display all SNPs on PAR (Chr=XY OR PAR='Yes'). Each SNP in the resulting list was inspected as described in **Section 4.1** before removing all filters on the SNP Table.

Finally all SNPs on non-autosomal regions were inspected as a whole. The SNP Table was first filtered into four separate lists, each initiated by displaying all SNPs on non-autosomal regions (Chr='Y' OR Chr='X' OR Chr='XY' OR Chr='MT') and followed by applying one of the following filters in addition:-

1. Clusters with low intensity

(AA R Mean < 0.2 AND AA Freq !=0) OR
(AB R Mean < 0.2 AND AB Freq !=0) OR
(BB R Mean < 0.2 AND BB Freq !=0)

2. Clusters with unexpected positions

(AA T Mean > 0.3 AND AA Freq !=0) OR
(AB T Mean < 0.3 AND AB Freq !=0) OR
(AB T Mean > 0.7 AND AB Freq !=0) OR
(BB T Mean < 0.7 AND BB Freq !=0)

3. Clusters too close together

Cluster separation < 0.4

4. Clusters too wide

(AA T Dev > 0.03) OR
(AB T Dev ≥ 0.07) OR
(BB T Dev > 0.03)

Each SNP in the resulting four lists were inspected as described in **Section 4.1** before removing all filters on the SNP Table.

### 4.1.6. Assessing autosomal SNPs

SNPs on the autosomal chromosomes (n=261,484) should produce a normal SNP Graph with one homozygous major cluster, one heterozygous cluster and one homozygous minor cluster with all samples called. The SNP Table was filtered to display markers on the autosomal chromosomes (Chr!='X' AND Chr!='Y' AND Chr!='XY' AND Chr!='MT'). SNPs that had less than perfect call rate were automatically clustered by applying an addition filter to display only those SNPs (Call Freq<1.00 AND Call Freq!=0), the resulting list of SNPs were then clustered. SNP statistics were updated when prompted before removing all filters on the SNP Table.

Given the high number of SNPs on the autosomal chromosomes, additional filters where applied to isolate portions of these SNPs to visually inspect. The SNP Table was first filtered into six separate lists, each begun by displaying all SNPs on autosomal chromosomes (Chr!='X' AND Chr!='Y' AND Chr!='XY' AND Chr!='MT') and followed by applying one of the following filters:-

1. SNPs with excess heterozygote calls

   AB Freq > 0.6 AND CallFreq != 0

2. SNPs with more heterozygote calls than expected

   Het excess > 0.1 AND CallFreq != 0

3. SNPs with less heterozygote calls than expected

   Het excess < -0.3 AND CallFreq != 0

4. Clusters with low intensity

   ((AA R Mean < 0.2) AND (AA Freq !=0) OR
   (AB R Mean < 0.2) AND (AB Freq !=0) OR
   (BB R Mean < 0.2) AND (BB Freq !=0)) AND
   (CallFreq != 0)

5. Clusters with unexpected positions

   (((AA T Mean >0.3 and AA Freq !=0) OR
   (AB T Mean < 0.3 and AB Freq !=0) OR
   (AB T Mean > 0.7 and AB Freq !=0) OR
   (BB T Mean < 0.7 and BB Freq !=0)) AND
   (CallFreq != 0)

6. Clusters too wide or too close together

   ((AA T Dev > 0.03) OR
   (AB T Dev ≥ 0.07) OR
   (BB T Dev > 0.03) OR
   (Cluster separation < 0.4)) AND
   (CallFreq != 0)

Each SNP in the resulting six lists was inspected as described in **Section 4.1** before removing all filters on the SNP Table.

### 4.1.7. Assessing specific SNPs

The SNP table was filtered to display causative variants as described in **Section 7.1**; these variants are known to cause familial forms of neurodegenerative diseases. Each SNP in the resulting list was inspected as described in **Section 4.1** before removing all filters on the SNP Table.

### 4.1.8. Reintroducing poor quality samples

Poor quality samples which were removed in **Section 4.1.3** were reintroduced by selecting all samples in the Sample Table, selecting Include Selected Samples and updating SNP statistics when prompted.

### 4.1.9. Exporting calls to PLINK format

The relevant PLINK plugin was added to GenomeStudio before exporting the genotype calls to PLINK format with all samples/SNPs included and alleles reported from the forward strand. Of the total 267,607 SNPs, 2,462 were zeroed and 265,145 remained active.

# 5. NeuroX Data Quality Control in PLINK

## 5.1. Methods

After exporting genotyping calls from GenomeStudio (**Section 4**) to PLINK format, PLINK was used to filter out further poor quality markers and poor quality samples (version and link given in **Section 12**). Entries to the command prompt/terminal are written within grey boxes after the relevant paragraph and each command in the box is prefixed with a dollar symbol ($). Data within text files are written within white boxes.

A new PLINK dataset was created whenever a change was made and each new dataset was named with a suffix of 1 through 9. **Figure 5.1** shows the three main file formats and flags to convert data between the three PLINK formats.

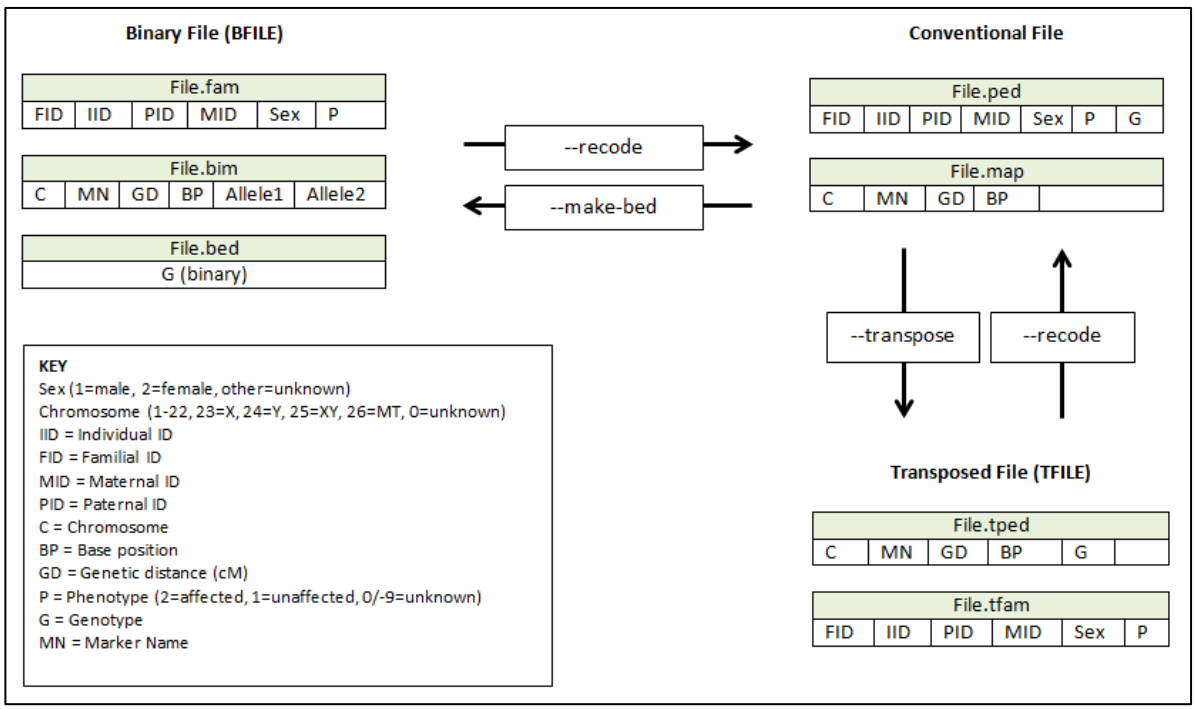



**Figure 5.1** Flow diagram of the PLINK file formats and the command line flags to convert between them. The three formats include binary, conventional and transposed. Each format has its data divided into two or three files; each file is represented with a box containing a green header. The data in all accessible files (not bed files) are white space delimited and can be opened in any software that can read text, such as Microsoft Excel. The data expected in each column of readable data files is given below the green header. The command flag used to convert between each format is stated in white boxes and preceded with '--'.

Quality control for the genotyping data was begun by converting the data from the conventional format (plink0.ped and plink0.map) to the binary format (plink1.fam, plink1.bim and plink1.bed). As PLINK doesn't have the capability to update phenotype, the sample phenotype and gender was manually added to plink1.fam using Excel.

```
$ plink --noweb --file plink0 --make-bed --out plink1
```

### 5.1.1. Updating marker information

Details about the design of each marker on the NeuroX was available from a comma separated values (CSV) file downloaded from the Illumina website. In the Illumina CSV file the sequence of each marker probe(s) was given along with the chromosome (Chr) and start base number of the variant it is designed to genotype (MapInfo), **Figure 5.2** shows the first 31 lines of this file. All but 29 markers had Chr and MapInfo in the Illumina CSV file, so this data was automatically exported from GenomeStudio into the .bim file. It was noted that the 29 markers lacking MapInfo were markers which genotyped variants listed in the three online databases as described in **Section 7.1**, the MapInfo for these 29 markers was manually calculated using both the information given in these databases and the UCSC genome browser hg18 (link given in **Section 12**) and LiftOver to convert to hg19 (link given in **Section 12**). The PLINK dataset was updated to reflect these changes; **Table 5.1** lists these markers and their final base position. It is noteworthy to mention that the reference genome builds 'hg18' and 'hg19' used by UCSC are actually the same builds as 'NCBI Build 36.1' and 'GRCh37' (the gold standard).

```
$ plink --noweb --bfile plink1 --update-map UpdateMarkerBP.txt --make-bed --out plink2
```

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Illumina | Inc. | | | | | | | | | | | | | | | | | |
| 2 | [Heading] | | | | | | | | | | | | | | | | | | |
| 3 | Descriptor File Name | NeuroX_15036164_A.csv | | | | | | | | | | | | | | | | | |
| 4 | Assay Format | Infinium HD Ultra | | | | | | | | | | | | | | | | | |
| 5 | Date Manufactured | 09/17/2012 | | | | | | | | | | | | | | | | | |
| 6 | Loci Count | 267607 | | | | | | | | | | | | | | | | | |
| 7 | [Assay] | | | | | | | | | | | | | | | | | | |
| 8 | IlmnID | Name | IlmnStrand | SNP | AddressA_ID | AlleleA_ProbeSeq | AddressB_ID | AlleleB_ProbeSeq | GenomeBuild | Chr | MapInfo | Ploidy | Species | Source | SourceVersion | SourceStrand | SourceSeq | TopGenomicSeq | BeadSetID |
| 9 | NeuroX_PARK7_Pro158del-1 | NeuroX_PARK7_Pro15 | PLUS | [D/I] | 17605557 | CTCTGAGAATCG | - | - | 37 | 1 | 0 | diploid | Homo s | genbank | 135 | PLUS | GCAGGTC | GCAGGTCATTAC | 763 |
| 10 | NeuroX_PINK1_23bp_del_ex7 | NeuroX_PINK1_23bp_de | PLUS | [I/D] | 78631227 | AAAAGGTTAGAT | - | - | 37 | 1 | 0 | diploid | Homo s | genbank | 135 | PLUS | TAGCCCA | TAGCCCATGGAT | 763 |
| 11 | NeuroX_PINK1_534_535insQ | NeuroX_PINK1_534_53 | PLUS | [D/I] | 90702357 | GCCCTGAAGAAT | - | - | 37 | 1 | 0 | diploid | Homo s | genbank | 135 | PLUS | GGTGAAC | GGTGAACATATT | 763 |
| 12 | NeuroX_PINK1_Asp525fs-135 | NeuroX_PINK1_Asp525 | PLUS | [I/D] | 58722836 | ATCTAAGCCTCT | - | - | 37 | 1 | 0 | diploid | Homo s | genbank | 135 | PLUS | CCGCAAA | CCGCAAATGTGC | 763 |
| 13 | NeuroX_PINK1_Cys549fs-135 | NeuroX_PINK1_Cys549 | MINUS | [D/I] | 98657160 | CGTTTCACACTC | - | - | 37 | 1 | 0 | diploid | Homo s | genbank | 135 | PLUS | GTTGGCT | GTTGGCTGGCTC | 763 |
| 14 | NeuroX_PINK1_Lys520fs-135 | NeuroX_PINK1_Lys520 | PLUS | [D/I] | 92691140 | GCCGCAAATGTG | - | - | 37 | 1 | 0 | diploid | Homo s | genbank | 135 | PLUS | CCATCTG | CCATCTGCCCGA | 763 |
| 15 | exm2268640-0_B_F_19848445 | exm2268640 | BOT | [T/C] | 9743988 | GGAGGAGCCCG | - | - | 37.2 | 1 | 762320 | diploid | Homo s | 1000gen | 0 | BOT | CCCCTGC | GGACCTGCCCGA | 688 |
| 16 | exm41-0_B_F_1921435147 | exm41 | BOT | [T/C] | 3645592 | CAGGGAAAAGTC | - | - | 37.1 | 1 | 861349 | diploid | Homo s | ExomeS | 0 | BOT | GCTCTGC | GAGCCCCGCATC | 662 |
| 17 | exm1916089-0_B_R_19276897 | exm1916089 | BOT | [T/C] | 93752267 | TCTTCAGGTTCC | - | - | 37.1 | 1 | 865545 | diploid | Homo s | ExomeS | 0 | TOP | cccacottcct | cccacottcctctcctcc | 662 |
| 18 | exm44-0_B_R_1921538602 | exm44 | BOT | [T/C] | 53658560 | TGTCACCATCGC | - | - | 37.1 | 1 | 865584 | diploid | Homo s | ExomeS | 0 | TOP | tgccccaccA | tgccccaccAGAACC | 662 |
| 19 | exm46-0_T_F_1921333919 | exm46 | TOP | [A/G] | 7696106 | CTGCCGCCCGGA | - | - | 37.1 | 1 | 865625 | diploid | Homo s | ExomeS | 0 | TOP | GTCGCCC | GTCGCCCTGCCT | 662 |
| 20 | exm47-0_T_F_1921489985 | exm47 | TOP | [A/G] | 90799171 | CCGCCCGGAAC | - | - | 37.1 | 1 | 865628 | diploid | Homo s | ExomeS | 0 | TOP | GCCCTGC | GCCCTGCCTGCC | 662 |
| 21 | exm51-0_B_R_1921417192 | exm51 | BOT | [T/C] | 57623126 | TCTTCATGATAC | - | - | 37.1 | 1 | 865662 | diploid | Homo s | ExomeS | 0 | TOP | GAACTCC | GAACTCCCAGCT | 662 |
| 22 | exm53-0_B_R_1921414362 | exm53 | BOT | [T/C] | 70604289 | TTCTCTTCATGA | - | - | 37.1 | 1 | 865665 | diploid | Homo s | ExomeS | 0 | TOP | CTCCCAG | CTCCCAGCTTCT | 662 |
| 23 | exm55-0_T_R_1921549846 | exm55 | TOP | [A/G] | 73680578 | CCTTCCTGTCCG | - | - | 37.1 | 1 | 865694 | diploid | Homo s | ExomeS | 0 | BOT | AGCGACG | GAGCCTTGCGCC | 662 |
| 24 | exm56-0_T_R_1921349466 | exm56 | TOP | [A/G] | 8606347 | CTTGCGCCTTCC | - | - | 37.1 | 1 | 865700 | diploid | Homo s | ExomeS | 0 | BOT | GGGAGTG | CCATCTGAGCCT | 703 |
| 25 | exm60-0_T_R_1921543826 | exm60 | TOP | [A/G] | 1664919 | CGGCTCCTACCT | - | - | 37.1 | 1 | 866429 | diploid | Homo s | ExomeS | 0 | BOT | GGGTCAG | ACGCACAGCACC | 662 |
| 26 | exm63-0_B_R_1919132154 | exm63 | BOT | [T/C] | 34684892 | CGCACAGCACG | - | - | 37.1 | 1 | 866438 | diploid | Homo s | ExomeS | 0 | TOP | GCCTCCA | GCCTCCAGAGCA | 662 |
| 27 | exm74-0_T_F_1921453331 | exm74 | TOP | [C/G] | 91691187 | CCTTCCCACCCT | 83810580 | CCTTCCCACCCTC | 37.1 | 1 | 871213 | diploid | Homo s | ExomeS | 0 | TOP | CAGGACG | CAGGACGGCAAC | 662 |
| 28 | exm76-0_B_F_1921429367 | exm76 | BOT | [T/C] | 25701245 | TATCCAGCGTCC | - | - | 37.1 | 1 | 871228 | diploid | Homo s | ExomeS | 0 | BOT | CCCACCC | TGTGCTCCTCAC | 662 |
| 29 | exm85-0_T_F_1921366461 | exm85 | TOP | [A/G] | 32791954 | GCCGCTGTGAAT | - | - | 37.1 | 1 | 871276 | diploid | Homo s | ExomeS | 0 | TOP | GAGCATC | GAGCATCAGAGC | 662 |
| 30 | exm88-0_B_R_1921389782 | exm88 | BOT | [T/C] | 39810141 | GAGGCGATTCGC | - | - | 37.1 | 1 | 874456 | diploid | Homo s | ExomeS | 0 | TOP | TGGGCAC | TGGGCACTCACT | 662 |
| 31 | exm94-0_B_F_1921415495 | exm94 | BOT | [T/C] | 88642314 | CCCCCGTATCAG | - | - | 37.1 | 1 | 874501 | diploid | Homo s | ExomeS | 0 | BOT | TGGGCCG | GGGACTGCCCCC | 662 |

**Figure 5.2** Screen shot of the first 31 lines of content in the comma separated values (CSV) file from Illumina, viewed using Excel. This file describes the markers on the NeuroX, including marker name (Name), alleles genotyped (SNP), probe sequence (AlleleA_ProbeSeq, AlleleB_ProbeSeq), chromosome (Chr) and start base position (MapInfo).

| Marker | Chr | MapInfo |
|---|---|---|
| NeuroX_PARK7_Pro158del | 1 | 8045015 |
| NeuroX_PINK1_23bp_del_ex7 | 1 | 20975486 |
| NeuroX_PINK1_534_535insQ | 1 | 20977040 |
| NeuroX_PINK1_Asp525fs | 1 | 20977011 |
| NeuroX_PINK1_Cys549fs | 1 | 20977085 |
| NeuroX_PINK1_Lys520fs | 1 | 20976995 |
| NeuroX_PARK2_Ala291fs | 6 | 161990449 |
| NeuroX_PARK2_Asn428fs | 6 | 161781121 |
| NeuroX_PARK2_Cys238fs | 6 | 162394356 |
| NeuroX_PARK2_Cys323fs | 6 | 161969996 |
| NeuroX_PARK2_Gln34fs_del_A | 6 | 162864412 |
| NeuroX_PARK2_Gln34fs_del_AG | 6 | 162864411 |
| NeuroX_PARK2_Gly179fs | 6 | 162475205 |
| NeuroX_PARK2_Pro133del | 6 | 162683570 |
| NeuroX_PARK2_Trp74fs | 6 | 162683748 |
| NeuroX_PARK2_Val324fs | 6 | 161969998 |
| NeuroX_LRRK2_Glu2490fs | 12 | 40761451 |
| NeuroX_LRRK2_IVS30-6C_T | 12 | 40704227 |
| NeuroX_LRRK2_IVS31+3A_G | 12 | 40704454 |
| NeuroX_LRRK2_IVS32+14G_A | 12 | 40707989 |
| NeuroX_LRRK2_IVS33+6T_A | 12 | 40709108 |
| NeuroX_LRRK2_IVS37-9A_G | 12 | 40716953 |
| NeuroX_LRRK2_IVS38+7C_T | 12 | 40717115 |
| NeuroX_LRRK2_IVS46-14T_A | 12 | 40753048 |
| NeuroX_LRRK2_IVS46-8delT | 12 | 40753054 |
| NeuroX_GRN_Ala237fs | 17 | 4031641 |
| NeuroX_GRN_Asn118fs | 17 | 42427596 |
| NeuroX_GRN_Asn119del | 17 | 42427601 |
| NeuroX_GRN_Thr52Hisfs | 17 | 42426809 |

**Table 5.1** 29 markers on the NeuroX which had their MapInfo (MapInfo) updated from zero to the correct base position (genome build GRCh37).

Additionally, as mentioned in **Section 4.1.1** there was 121 non-autosomal markers that had recently been located on the pseudo-autosomal region (PAR). The PLINK dataset for these 121 markers (**Table 4.1**) was updated to 'XY' to reflect these changes.

```
$ plink --noweb --bfile plink2 --update-map UpdateMarkerChrm.txt --update-chr --make-bed --out
  plink3
```

### 5.1.2. Removing sample and markers failing data rate

At this stage, both the samples and marker average data rate (call rate) was 0.96. Samples or markers with a low data rate are indicative of poor quality DNA and poor marker probe design respectively, thus filters were applied that removed both, as leaving them in would influence the accuracy of future statistical analyses.

Samples were removed with a data rate less than 95%. A low sample call rate is indicative of poor quality DNA and the standard approach is to remove samples before markers, as there is a greater chance for poor quality samples over poor quality markers, which in this case is true with the UoN controls. A total of 199 samples (3 cases and 196 controls) were removed which resulted in 1001 samples remaining with an average data rate of 0.990.

```
$ plink --noweb --bfile plink3 --mind 0.05 --make-bed --out plink4
```

Markers were removed with a data rate less than 90%. A low marker call rate is indicative of poor quality marker probe design; this cut-off was lower than that used for samples in order to accommodate the markers on chromosome Y which have an inherently lower data rate (approximately 91%). A total of 2,466 markers were removed; 2,462 of these were zeroed in GenomeStudio and the other four were active but had a data rate less than 90%. This resulted in 265,141 markers remaining with an average call rate of 0.999.

```
$ plink --noweb --bfile plink4 --geno 0.1 --make-bed --out plink5
```

### 5.1.3. Gender

The gender of the 1001 samples was assessed using homozygosity of common markers (MAF ≥ 0.1) on chromosome X (F value). A graph was produced using ImissVsGender.r (**Section 11.3)** which plotted each sample based on their F value and missing data rate (**Figure 5.3**). The graph was used to generate an F value threshold for females (< 0.4) and for males (> 0.9). The gender for each sample was then calculated in Excel based on these cut-offs, the calculated gender was compared to the documented gender; 87 of the 1001 samples had an alternative gender. It was decided to update the plink files to reflect the calculated gender, however in the analyses when gender was corrected for as a covariate (**section 9**), the analyses would be run twice, with each of the different gender sets, thus allowing comparison of results.

```
$ plink --noweb --bfile plink5 --maf 0.1 --check-sex --out plink5CheckSex
$ R CMD BATCH ImissVsGender.r
$ plink --noweb --bfile plink5 --update-sex UpdateGenderCalculated.txt --make-bed --out plink6
```

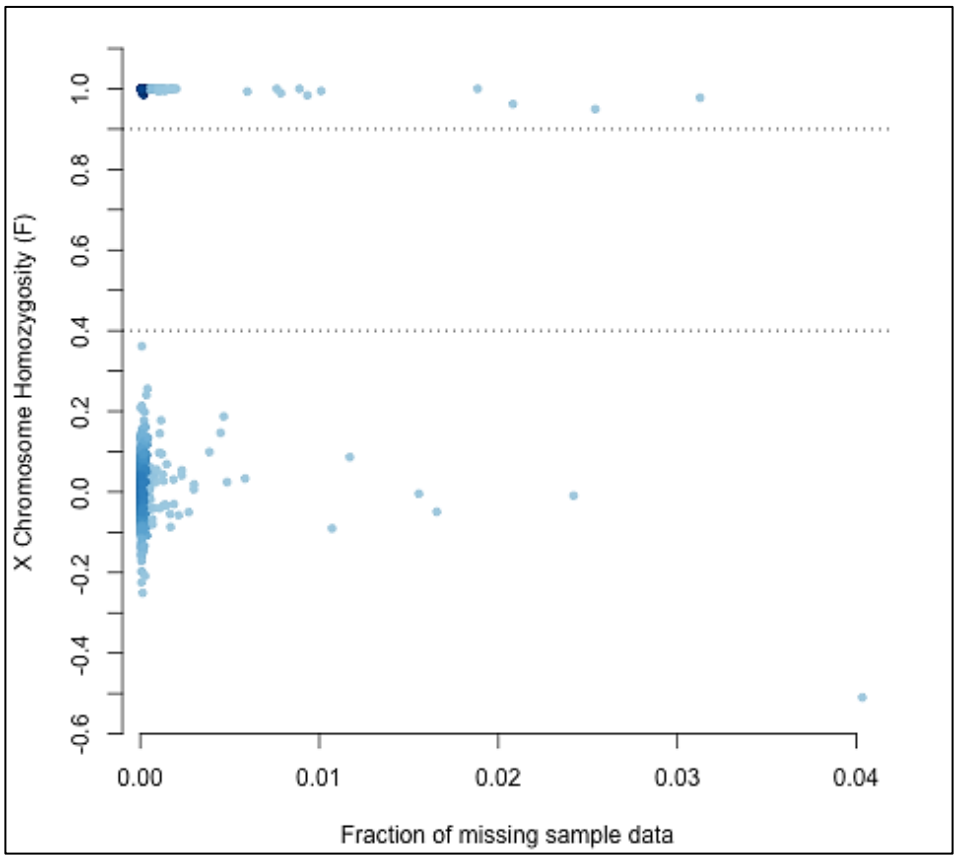



**Figure 5.3** Dot plot marking each sample against their missing data rate (X-axis) and homozygosity score (F) of common markers on chromosome X (Y-axis). Each dot represents one sample and the depth of blue indicates density of samples at that position. Female cutoff score (< 0.4) and male cutoff score (> 0.9) are shown as dotted horizontal lines.

### 5.1.4. Creating additional data

Using the plink6 dataset, two additional datasets were created to help assess other aspects of the data. The first additional dataset (HapMapAIMs2_plink6AIMs_merged2) contained the NeuroX ancestry informative markers (AIMs) merged with HapMap AIMs which was used for assessing genetic ancestry in **Section 5.1.5**. The difficulty in merging the dataset with HapMap was that merging depends solely on the names of the markers in each dataset; the marker names used in HapMap are not the same as in the NeuroX data; HapMap uses reference IDs and NeuroX uses the Illumina numbers which do not relate. To achieve merging, a dataset of NeuroX AIMs[227] and a dataset of HapMap AIMs were created and the names in the HapMap dataset updated to match the NeuroX. Once merged, markers were then removed that were not in the both dataset by simply applying a marker data rate cut-off of 99%. The resulting merged dataset contained 1,450 AIMs. The second additional dataset removed markers in linkage disequilibrium (LD) (plink6Pruned) and was used to assess relatedness of sample pairs in **Section 5.1.7** and a second assessment of sample quality in **Section 5.1.8**.

```
$ plink --noweb --bfile plink6 --extract ExtractAIMsNeuroX.txt --make-bed --out plink6AIMs
$ plink --noweb --bfile HapMap --extract ExtractAIMsHapMap.txt --make-bed --out HapMapAIMs
$ plink --noweb --bfile HapMapAIMs --update-map UpdateAIMNames.txt --update-name --make-bed
  --out HapMapAIMs2
$ plink --noweb --bfile HapMapAIMs2 --bmerge plink6AIMs.bed plink6AIMs.bim plink6AIMs.fam
  --make-bed --out HapMapAIMs2_plink6AIMs_merged
$ plink --noweb --bfile HapMapAIMs_plink6AIMs_merged --geno 0.01 --make-bed --out
  HapMapAIMs2_plink6AIMs_merged2
$ plink --noweb --bfile plink6 --maf 0.1 --indep-pairwise 50 5 0.2 --out plink6
$ plink --noweb --bfile plink6 --extract plink6.prune.in --make-bed --out plink6Pruned
```

### 5.1.5. Genetic ancestry

The AIM dataset created in **section 5.1.4** was converted to conventional PLINK format to allow compatibility with the EIGENSTRAT software (version and link given in **Section 12**), which contained the three programs SmartPCA and Ploteig.

Smartpca was used to generate the top two principle component eigenvectors for each sample using the parameters in the plink6SmartPCA.txt. This was followed with Ploteig which plotted the top two eigenvectors onto a two dimensional graph (**Figure 5.4**). Upon inspection of the graph it was concluded that most samples were of European ancestry. Approximately 12 samples deviated from the European cluster which would indicate mixed ancestry for those samples. One option would be to remove these samples from the dataset, however given the limited number of samples in our cohort it was decided instead to adjust for their ancestry differences by using the top eigenvectors for each sample as covariates in future analyses, this data was generated in **Sections 5.1.10**. For the purpose of Hardy Weinberg Equilibrium (HWE) testing however (**Section 5.1.6**), these 12 samples were added to the file RemoveSamplesFailingAncestry.txt for temporary removal.

plink6SmartPCA.txt

```
genotypename:  HapMapAIMs2_plink6AIMs_merged2.ped
snpname:       HapMapAIMs2_plink6AIMs_merged2.map
indivname:     HapMapAIMs2_plink6AIMs_merged2.ped
evecoutname:   HapMapAIMs2_plink6AIMs_merged2.evec
evaloutname:   HapMapAIMs2_plink6AIMs_merged2.eval
altnormstyle:  NO
numoutevec:    2
numoutlieriter: 0
familynames:   NO
grmoutname:    HapMapAIMs2_plink6AIMs_merged2.junk
```

```
$ plink --noweb --bfile HapMapAIMs2_plink6AIMs_merged2 --recode –out
  HapMapAIMs2_plink6AIMs_merged2
$ smartpca -p plink6SmartPCA.txt > plink6SmartPCA.log
$ perl ploteig -I HapMapAIMs2_plink6AIMs_merged2.evec -c 1:2:3:4:5:6 -p
  Control:Case:CEU:JPT:CHB:YRI -x -o HapMapAIMs2_plink6AIMs_merged2.plot.xtxt
```

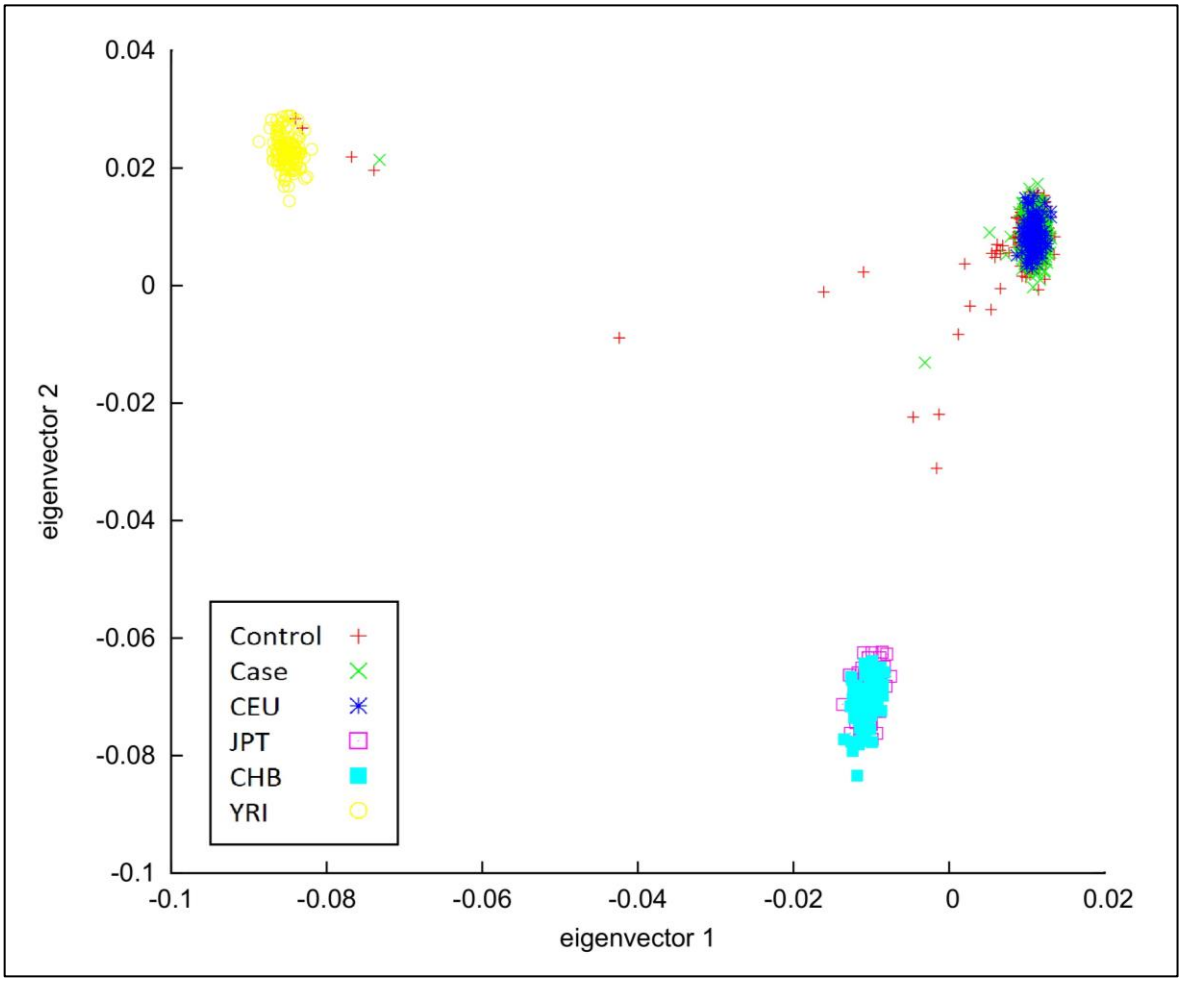



**Figure 5.4** Dot plot created by Ploteig and amended to incorporate the key in the bottom left corner. Each sample is plotted based on their two eigenvectors calculated by Smartpca. Each dot is colour/shape coded depending on the cohort they originate from. Samples from HapMap are from one of four geographical locations: CEU (Utah residents with ancestry from northern and western Europe, JPT (Japanese in Tokyo, Japan), CHB (Han Chinese in Beijing, China) and YRI (Yoruba people of Ibadan, Nigeria).

### 5.1.6. Hardy Weinberg Equilibrium (HWE)

Using PLINK, the observed and expected marker heterozygosity was calculated for all common markers (MAF > 0.05) along with the probability that the marker deviated from HWE by chance (p.value), samples failing ancestry were removed before calculating this. There were 39,161 common markers in the plink6 dataset and thus the probability that a marker would deviate from HWE by chance after correcting for multiple testing using the Bonferroni correction is 1 in 783,220 or 1.2E-6 (0.05 divided by 39,161 markers). Using Excel, markers were identified that significantly deviated from the HWE in controls by isolating those markers with a p-value less than 1.2E-6 in controls (n=69), the names of these markers were added to the text file (ExcludeMarkersFailingHWE.txt) for removal later.

```
$ plink --noweb --bfile plink6 --remove RemoveSamplesFailingAncestry --maf 0.05 --hardy –out
  plink6HWE
```

### 5.1.7. Relatedness

The percentage of markers sharing identity by state (IBS) for all possible pairs of samples in the plink6 dataset were calculated as an estimate for identity by decent (IBD), the IBD can be used to identify pairs of samples that have been involved in sample cross contamination or are related, a percentage cut-off of 18.75% was used as this is the average percentage of genetic similarity between second and third degree relatives. Using PLINK, IBS was calculated using the LD pruned plink6 dataset created in **Section 5.1.4** and only markers on autosomal chromosomes. The total number of percentages calculated was $1001^2$, as this was too many entries to access in Excel, Relatedness.pl (**Section 11.4**) was used to identify samples sharing ≥ 18.75% IBS. This program resulted in a list of 30 pairs of samples and 56 unique samples. A list of samples of the smallest number was devised by hand that accounted for all of the 30 pairs identified; this was a total of 30 unique samples (6 cases and 24 controls). The names of these 30 samples were added to the text file RemoveSamplesFailingIBS.txt for removal later.

```
$ plink --noweb --bfile plink6 --missing --out plink6Missing
$ plink --noweb --bfile plink6Pruned --exclude ExcludeNonAutosomalMarker.txt --genome --out
  plink6Genome
$ perl Relatedness.pl plink6Genome.genome plink6Missing.imiss FileOut1.txt
```

### 5.1.8. Heterozygosity

Samples with a high/low heterozygosity rate could indicate cross contamination and a low heterozygosity rate could also indicate inbreeding. Using PLINK, the observed sample homozygosity was calculated using the LD pruned plink6 dataset from **Section 5.1.4** and only markers on autosomal chromosomes. In Excel the heterozygosity rate for each sample was calculated. Heterozygosity rate was calculated using the formula (N-O)/N, where N is the number of none missing genotypes and O is the number of observed homozygotes. Overall standard deviation was calculated using Excel's in build function STDEV. A cut-off value of ± 3 times the standard deviation was used to isolate samples with a high/low heterozygosity rate, and there were 27 samples (4 cases and 23 controls) that failed this cut-off, their names were added to the file RemoveSamplesFailingHet.txt for removal later.

```
$ plink --noweb --bfile plink6Pruned --exclude ExcludeNonAutosomalMarkers.txt --het –out
  plink6Het
```

### 5.1.9. Removing samples and markers

Using PLINK, markers and samples that failed plink6 dataset checks were removed, this included markers failing Hardy Weinberg Equilibrium (**Section 5.1.6**) and samples failing relatedness (**Section 5.1.7**) and heterozygosity (**Section 5.1.8**). The samples listed in RemoveSamplesFailingIBS.txt and RemoveSamplesFailingHet.txt were combined into one file named RemoveSamplesFailingAll.txt for removal, this file contained a total of 55 unique samples (9 cases and 46 controls).

Additionally, samples were also removed that did not pass the criteria in **Section 2.1.1**. These samples included cases with one or more of the following: mixed phenotype, too old to be classified as sEOAD or harboured a causative mutation as identified in **Section 3**. These samples also included controls that did not have an age at sampling or were too young at age of sampling. This list contained a total of 131 samples (33 cases and 98 controls), 29 of these (4 cases and 25 controls) were not present in the dataset as they had already been removed prior, therefore a total of 102 samples were removed from the dataset. The final dataset (plink9) contained a total of 844 samples (408 cases and 436 controls) and 265,049 markers.

```
$ plink --noweb --bfile plink6 --exclude ExcludeMarkersFailingHWE.txt --make-bed --out plink7
$ plink --noweb --bfile plink7 --remove RemoveSamplesFailingAll.txt --make-bed --out plink8
$ plink --noweb --bfile plink8 --remove RemoveSamplesFailingPhenotype.txt --make-bed --out
  plink9
```

### 5.1.10. Creating ancestry covariates

Covariates were created for analyses (**Section 9**) that adjusted for genetic ancestry. To achieve this Twstats and SmartPCA was used from the EIGENSTRAT software (version and link given in **Section 12**). Smartpca was used as in **Section 5.1.5** using parameters in plink9aSmartPCA.txt (below), to caulculate the eigenvectors for the top ten principle components and the eigenvalues for all principle components. Twstats was used to calculate the significant of each principle component eigenvalue, the top 11 principle components had p-value <0.05, thus it was decided to correct for the top 10 eigenvectors in analyses.

Plink9aSmartPCA.txt

```
genotypename:  plink9AIMs.ped
snpname:       plink9AIMs.map
indivname:     plink9AIMs.ped
evecoutname:   plink9AIMs.evec
evaloutname:   plink9AIMs.eval
altnormstyle:  NO
numoutevec:    10
numoutlieriter: 0
familynames:   NO
grmoutname:    plink6AIMs.junk
```

```
$ plink --noweb --bfile plink9 --extract ExtractAIMsNeuroX.txt --recode --out plink9AIMs
$ smartpca -p plink9SmartPCA.txt > plink9SmartPCA.log
$ perl twstats -t twtable -i plink9AIMs.eval >> Out.txt
```

### 5.1.11. Imputing *APOE* ε status

Unlike to the UoN samples, the UCL control samples had no *APOE* ε status documented. Ideally these would have been genotyped using the TaqMan genotyping assay as in **Section 2.1.2**; however this option was not available. The NeuroX genotypes for rs429358 could not be used as the three markers genotyping rs429358 on the NeuroX had all failed and were zeroed during quality control (**Section 4.1.7**). To get around this issue, the whole of chromosome 21 was imputed for all UCL control samples using dataset plink7, data from the 1000 genomes project (phase 3) as a reference and the program IMPUTE2 (version and link given in **Section 12**). The program GTOOL (version and link given in **Section 12**) was used to convert chromosome 21 of plink7 into a format compatible with IMPUTE2, the native Linux command line tool SED was used to extract the data for rs429358 from the IMPUTE2 output. IMPUTE2 gave three probabilities for each sample; this was the probability of the sample being homozygous major for rs429358, heterozygous and homozygous minor. Finally the genotypes for one of the markers genotyping rs7412 (NeuroX_Seq2-rs7412) in plink7 was used and together with the highest probability score for rs429358 from IMPUTE2 the *APOE* ε status was calculated in Excel using **Table 2.1**.

```
$ plink --noweb --bfile plink7 --chr 19 --recode --out plink7Chr19
$ gtool -P --ped plink7Chr19.ped --map plink7Chr19.bim --og plink7Chr19.gens --os
  plink7Chr19.sample --binary_phenotype
$ impute2 -m genetic_map_chr19_combined_b37_phase3.txt -h 1000GP_Phase3_chr19.hap
  -l 1000GP_Phase3_chr19.legend -g plink7Chr19.gens -strand_g plink7Chr19.strand
  -int 42411941 48411941 -Ne 20000 -o impute2phase1.output
$ sed '44782q;d' impute2.output > impute2rs429358.output
$ sed '100532q;d' impute2.output2 > impute2rs429358.output2
$ plink --noweb --bfile plink7 --extract NeuroX_Seq2-rs7412.txt --recode
  --out plink7NeuroX_Seq2-rs7412
```

## 5.2. Results

The final data set (plink9) generated from quality control of the raw NeuroX data contained a total of 844 samples (408 cases and 436 controls) (**Table 5.3**) and 265,049 markers. All UoN controls were lost during quality control, leaving just UCL controls remaining. All UCL controls were successfully imputed to generate an *APOE* ε status, meaning there was no missing data.

**A sEOAD**

| Centre | N | Mean age at onset (±SD) | | Females (%) | | *APOE* ε4+ (%) | | *APOE* ε4ε4 (%) | | *APOE* ε4 MAF | Diagnosis |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Bristol, UK | 21 | 53.3 | (5.3) | 9 | (42.9) | 10 | (47.6) | 3 | (14.3) | 0.31 | 21:0 |
| Manchester, UK | 328 | 57.1 | (5.5) | 156 | (47.6) | 196 | (59.8) | 46 | (14.0) | 0.58 | 53:275 |
| Nottingham, UK | 26 | 58.2 | (6.3) | 12 | (46.2) | 11 | (42.3) | 1 | (3.8) | 0.23 | 5:21 |
| Oxford, UK | 33 | 55.6 | (4.2) | 19 | (57.6) | 19 | (57.6) | 3 | (9.1) | 0.33 | 24:9 |
| All | 408 | 56.8 | (5.5) | 196 | (48.0) | 236 | (57.8) | 53 | (13.0) | 0.53 | 103:305 |

**B CNTL**

| Centre | N | Mean age at onset (±SD) | | Females (%) | | *APOE* ε4+ (%) | | *APOE* ε4ε4 (%) | | *APOE* ε4 MAF |
|---|---|---|---|---|---|---|---|---|---|---|
| UK | 436 | 77.2 | (6.4) | 256 | (58.7) | 104 | (23.9) | 9 | (2.1) | 0.13 |

**Table 5.3** Sample demographics of sporadic early-onset Alzheimer's disease samples (A sEOAD) and control samples (B CNTL) remaining after quality control in PLINK. Each cohort contains samples from multiple centres; each centre is represented one per row. The number of samples from each centre (N) is given, along with the mean age of onset with standard deviation (Mean age at onset (±SD)), number and percentage of female samples (Females (%)), number and percentage of samples harbouring at least one *APOE* ε4 allele (*APOE* ε4 + (%)), number and percentage of samples with *APOE* ε4ε4 status (*APOE* ε4ε4 (%)), minor allele frequency of the *APOE* ε4 allele (*APOE* ε4 MAF) and finally the number of case samples classed as post mortem confirmed/definite Alzheimer's disease and probable Alzheimer's disease separated by a colon (Diagnosis). Key: N, number of samples; SD, standard deviation; UK, United Kingdom.

# 6. NeuroX Marker Annotation

Details about the design of each marker on the NeuroX was available from a comma separated values (CSV) file (**Figure 6.1**) downloaded from the Illumina website. The sequence of each marker probe(s) is given along with the chromosome (Chr), start base number of the variant it is designed to genotype (MapInfo), and the sequence used to design the probe(s) (SourceSeq). In the Illumina CSV file little information was given about the variants genotyped and thus multiple in-house and external programs were utilised to calculate more information.

## 6.1. Methods

Commands that were used to run programs in the command prompt/terminal are written within grey boxes and prefixed with a dollar symbol ($). Program input data files were created using Excel or in-house programs.

### 6.1.1. Reference and alternative allele

For each marker the Illumina CSV file states the two alleles which are genotyped, however these alleles are given in TOP/BOT strand format and not plus/minus, the base number where the variant ends is not given and deletions and insertions were both defined as 'Indels', this information was needed for ANNOVAR (**Section 6.1.2**) and VEP (**Section 6.1.3**). Alleles.pl (**Section 11.5**) calculated this data using two sources of information; the Illumina CSV file and reference sequence data (hg19) downloaded from UCSC (version and link given in **Section 12**). The output data from Alleles.pl stated insertions with the start base number = end base position - 1. For example an insertion of a C between bases 12600 and 12601 on the forward strand of chromosome 8 would be coded as 8:12600-12601. As a secondary outcome the program also identified if the start position was out by a number of bases either upstream or downstream of this point and thus was able to calculate a new start and end base position.

```
$ perl Alleles.pl FileIn.txt 14 FileOut1.txt FileOut2.txt
```

### 6.1.2. Proximal genes

The position of the markers in relation to genes was initially assessed using an in-house program AnnotateWithGenomicRegion.pl (**Section 11.2**) which used a list of 54,341 transcript reference sequences (hg19) downloaded from the UCSC (version and link given in **Section 12**). This annotation was superseded by using the external program ANNOVAR (version and link given in **Section 12**) with all flags submitted which annotated each marker with at least one gene.

ANNOVAR is designed to take specific variant information and annotate it with various forms of useful information, for example genes. The ANNOVAR package came with six perl scripts and required prerequisite data to be downloaded from UCSC's Golden Path (version and link given in **Section 12**). The main script 'table_annovar.pl' called on the five other scripts and the prerequisite data to annotate the markers. The input file for 'table_annovar.pl' was a space delimited text file with no header. The input file required the chromosome number and start base position. ANNOVAR bins all markers into a gene region, this includes markers that are located in intergenic regions which are binned into both the nearest upstream gene feature and nearest downstream gene feature. To gain variant specific annotation the reference allele, alternate allele and end base position for each marker were also required. All input data were taken from the results of Alleles.pl. To comply with ANNOVAR's input file format, the chromosomes were converted to the traditional coding (1-22, X, Y, M) with pseudo-autosomal (XY) recoded to X. The end base position for insertions as calculated from Alleles.pl was used as both the end and start base position, for example an insertion of a C between bases 12600 and 12601 on the forward strand of chromosome 8 would be coded as 8:12601-12601. Once markers had been annotated with genes, the number of genes covered on the NeuroX (gene coverage) was calculated in Excel by creating a list of unique genes using the remove duplicates function.

```
$ perl table_annovar.pl IN/AnnovarIn_AllAllele.txt humandb/ --buildver hg19
  --out ./OUT/AnnovarOut_AllAllele.txt --remove --protocol
  refGene,cytoBand,genomicSuperDups,esp6500siv2_all,1000g2014oct_all,1000g2014oct_afr,
  1000g2014oct_eas,1000g2014oct_eur,snp138,ljb26_all -operation g,r,r,f,f,f,f,f,f -nastring .
```

### 6.1.3. Variant and predicted consequence

Using Ensembl's variant effect predictor (VEP, version and link given in **Section 12**), each marker was annotated with a variant reference ID (rsID), a predicted consequence of the variant (most severe) and the MAF of the variant in the European cohort of the 1000 Genomes Project. The effect of each variant genotyped on the NeuroX was predicted using Ensembl's Variant Effect Predictor (VEP). VEP was installed locally and run over the command line with all flags submitted. The required input data were taken from the results of Alleles.pl and included chromosome, start base position, allele A and allele B. To comply with VEP's input file format the chromosomes were converted to the traditional coding (1-22, X, Y, M) with pseudo-autosomal (XY) recoded to X, and the start and end base position for insertions as calculated from Alleles.pl were switched around, for example an insertion of a C between bases 12600 and 12601 on the forward strand of chromosome 8 would be coded as 8:12601-12600. FormatVEP.pl (**Section 11.6**) was used to reformat the 'Extra' section of the VEP results.

```
$ perl /usr/local/bin/variant_effect_predictor.pl --i In.txt –o VEPOut.txt --cache --dir /var/lib/VEP/
  --maf_1kg --offline --assembly GRCh37 --everything --pick --humdiv --force_overwrite --fork 4
$ perl FormatVEP.pl VEPOut.txt FileOut.txt
```

## 6.2. Results and discussion

The in-house program Alleles.pl successfully annotated 99.3% (n=265,828) of the total 267,607 markers available for annotation. Annotation included variant type, strand, reference allele and alternative allele. As a secondary outcome to this, the program also identified 2,134 markers as having a position inconsistent with GRCh37 (27 Feb 2009) which is a major human genome build release, this is not surprising as a majority of the markers on the NeuroX had a position aligned to subsequent minor revision (GRCh37.1, GRCh37.2, GRCh37.3 and GRCh37.5). A corrected position was therefore calculated for each of these 2,134 markers to realign them to the major build, as this enabled the use of downstream software such as ANNOVAR and Ensembl's VEP.

ANNOVAR successfully annotated all 265,828 markers submitted to the program with a gene, and analysis of this annotation established that the markers covered 20,508 genes. Ensembl's Variant effect predictor (VEP) successfully annotated all 265,828 markers submitted to the program; this annotation included several key pieces of information relating to the variants genotyped and the most severe predicted consequence of these variants. Analysis of the VEP annotation established that 31% (n=81,803) had no documented minor allele frequency (MAF) in the European cohort of the 1000 Genomes Project, 33% (n=86,599) had a MAF of zero and 19% (n=49,964) were rare (MAF >0.00 and < = 0.01) (**Figure 6.1**). Assuming markers with a MAF of zero or none documented are also rare then the total number of rare variants genotyped could be as high as 83% (n=220,364). Additionally, VEP also generated some additional statistics about all the variants; they covered 20,317 genes, 5% (n=12,243) were novel, 95% (n=253,585) were already documented, and 80% (n=212,763) were missense (**Figure 6.2**) which were distributed with a bias towards the latter half (C terminas) of the coding region (**Figure 6.3**). Notably, the distribution of variants was not equal across chromosomes (**Figure 6.4**); however this appeared to be consistent with chromosome length. The gene coverage according to VEP is less than that predicted by ANNOVAR, however this is consistent with the fact intergenic variants were assigned a gene in ANNOVAR but not assigned a gene in VEP.

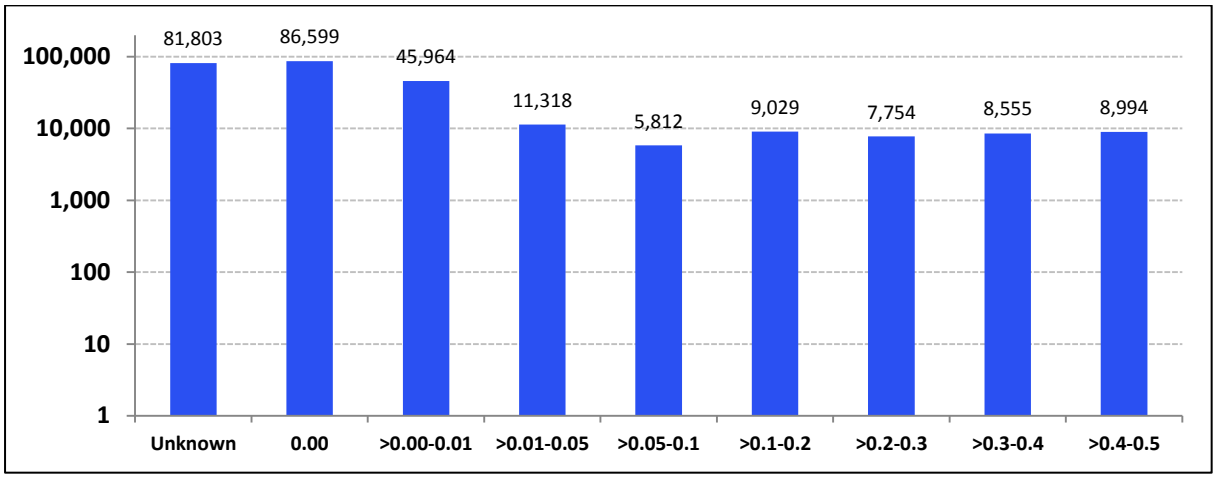



**Figure 6.1** Histogram constructed in Excel showing the number of variants on the NeuroX (Y-axis) classified by their minor allele frequency (MAF) as calculated in the European cohort of the 1000 genomes project and retrieved via VEP (X-axis). Variants with no documented MAF are classified as unknown.

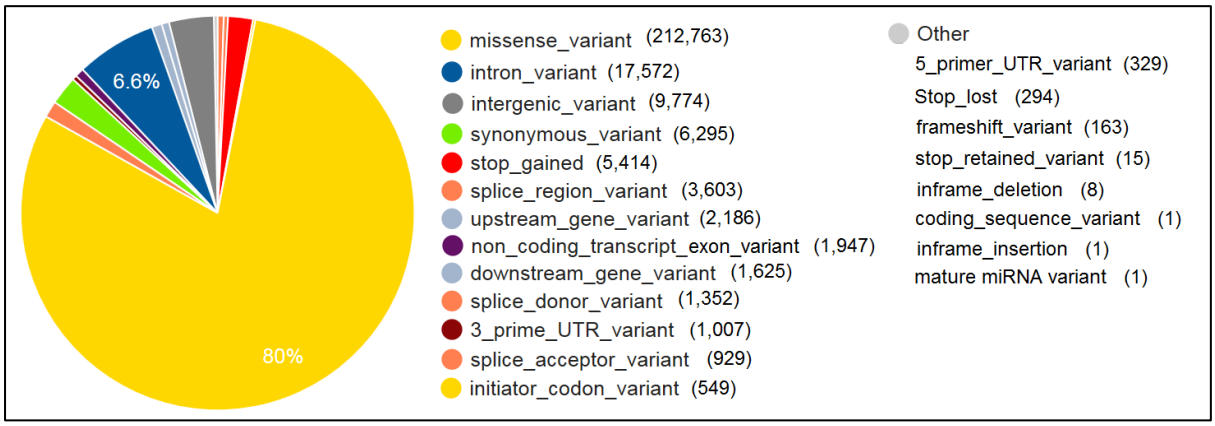



**Figure 6.2** Pie chart showing the type of variants genotyped by 99.3% of the markers on the NeuroX. The image was generated by VEP and modified to include the number of each type. The pie chart shows most of the variants genotyped are missense. The variant type is defined by their most severe predicted consequence.

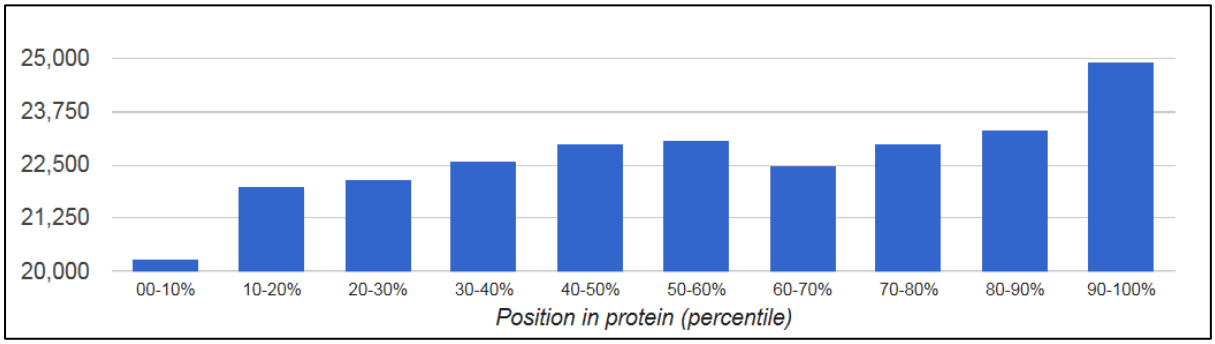



**Figure 6.3** Histogram generated by VEP showing the relative protein position (X-axis) of the intragenic variants genotyped by markers on the NeuroX (Y-axis). Image was taken from VEP.

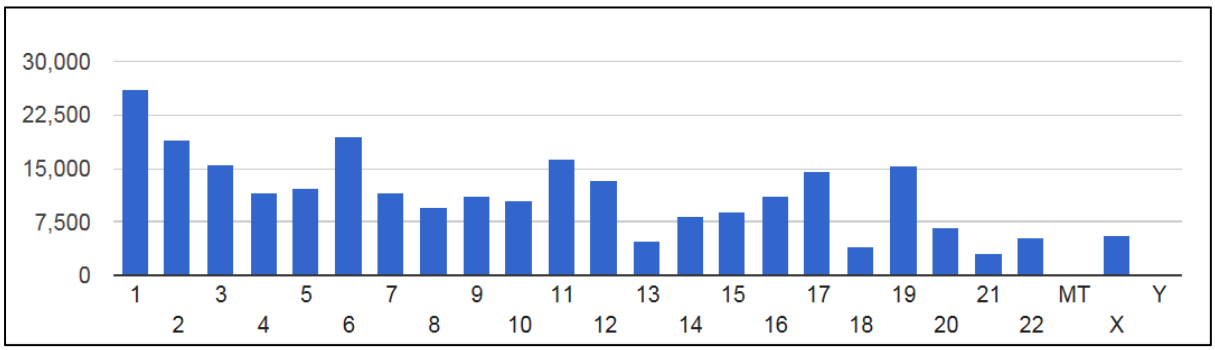



**Figure 6.4** Histogram showing the chromosomal location (X-axis) of the variants genotyped by 99.3% of the markers on the NeuroX (Y-axis). Image was taken from VEP.

# 7. NeuroX Known Causative Variants in sEOAD

## 7.1. Methods

As a way to identify causative variants in sEOAD, 16 genes linked to familial forms of neurodegeneration were screened using the final NeuroX dataset (plink9, **Section 5**). There are three online databases where researchers have documented variants found in patients with familial forms of Alzheimer's disease (AD), frontotemporal dementia (FTD), Parkinson's disease (PD) and Creutzfeldt-Jakob disease (CJD). These three databases are the Alzheimer Disease & Frontotemporal Dementia Mutation Database (AD&FTDMDB), Parkinson Disease Mutation Database (PDMutDB) and Human Prion Protein Database (HPPDB) (links provided in **Section 12**).

### 7.1.1. List of causative variants

The variant data contained in the AD&FTDMDB and the PDMutDB online databases includes the coordinates of the variant and the pathogenic nature (based on consensus of opinion). This information was coded in a browser extensible data file (BED) which allowed compatibility with genome browsers employed by ENSEMBL and UCSC. The BED file was downloaded from UCSC (version and link provided in **Section 12**) after selecting the link on each of the two database webpages, followed by 'My Data', 'Custom Tracks', view in: 'Table browser', group: 'All Tracks' and track: 'AD&FTDMDB' or 'PD Mutations'. The BED file contained the name of the variant which included allele information, chromosome, start base position (zero-based) and end position (one-based) based on the reference genome build NCBI36. The BED file also contained the variant pathogenicity coded as one of three red green blue (RGB) colour codes; 225.0.0 (red) pathogenic, 12.120.12 (green) not pathogenic and 255.153.102 (orange) pathogenic nature unclear. The variant base positions were converted to reference genome build hg19 using UCSC's LiftOver tool (version and link provided in **Section 12**) and then converting to 'one-based' by adding 1 to all start base positions.

Data from the HPPDB database were downloaded as a comma delimited value (CSV) file from the site using a link provided. The data did not contain the genomic coordinates so this was evaluated by hand using data from the CSV file and the UCSC Genome Browser hg19 (version and link provided in **Section 12**).

All data were combined into a single file which contained a total of 1196 database variants; **Table 7.1** lists the numbers and the type of these variants. 121 variants did not have their reference/alternative alleles documented because they were too large and so these were disregarded from further analysis, 1075 database variants remained.

| Type | N |
|---|---|
| Single nucleotide polymorphism | 965 |
| Multiple nucleotide polymorphism | 2 |
| Small insertion (≤ 19 nucleotides) | 27 |
| Large insertion (incl. multiplication)* | 54 |
| Small deletion (≤ 28 nucleotides) | 81 |
| Large deletion* | 64 |
| Repeat expansion* | 1 |
| Deletion insertion* | 2 |
| Total | 1196 |

**Table 7.1** The number (N) and type of variants (Type) listed in the three online databases; the Alzheimer Disease & Frontotemporal Dementia Mutation Database (AD&FTDMDB), Parkinson Disease Mutation Database (PDMutDB) and Human Prion Database (HPDB). The types of variants that did not have a reference/alternative allele documented are highlighted with an asterix.

### 7.1.2. Causative variants genotyped on the NeuroX

The in-house program MatchingVariants.pl (**Section 11.7**) was used to identify which of the 1075 database variants identified in **Section 7.1.1** were genotyped by markers on the NeuroX. The input data required the variant start position together with allele A and allele B for both the database and NeuroX variants. The input data for the NeuroX variants was taken from the results of Alleles.pl (**Section 6.1.1**).

```
$ perl MatchingVariants.pl FileIn1.txt FileIn2.txt N FileOut1.txt FileOut2.txt
```

MatchingVariants.pl established that 412 database variants were genotyped by markers on the NeuroX. **Table 7.2** lists these variants by gene, database and pathogenic nature and demonstrates that 407 of the 412 markers passed quality control and remained in the final (plink9) NeuroX dataset.

| | | | | | Pathogenic Nature Unclear | | | Causative | | | Not Pathogenic | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Gene** | **Disease** | **Database** | **Designed on NeuroX** | **Passed QC** | **Database** | **Designed on NeuroX** | **Passed QC** | **Database** | **Designed on NeuroX** | **Passed QC** | **Database** | **Designed on NeuroX** | **Passed QC** |
| APP | AD | 31 | 10 | 10 | 6 | 3 | 3 | 24 | 6 | 6 | 1 | 1 | 1 |
| C9orf72 | ALS/FTD | 12 | 0 | 0 | 11 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| CHMP2B | FTD | 12 | 7 | 7 | 4 | 0 | 0 | 4 | 3 | 3 | 4 | 4 | 4 |
| FUS | ALS/FTD | 44 | 21 | 20 | 18 | 3 | 3 | 22 | 15 | 15 | 4 | 3 | 2 |
| GRN | FTD | 146 | 33 | 33 | 45 | 10 | 10 | 66 | 13 | 13 | 35 | 10 | 10 |
| LRRK2 | PD | 128 | 71 | 71 | 54 | 15 | 15 | 6 | 3 | 3 | 68 | 53 | 53 |
| MAPT | FTD | 73 | 11 | 10 | 27 | 10 | 9 | 44 | 1 | 1 | 2 | 0 | 0 |
| PARK2 | PD | 144 | 70 | 70 | 22 | 4 | 4 | 59 | 27 | 27 | 63 | 39 | 39 |
| PARK7 | PD | 21 | 6 | 5 | 5 | 1 | 0 | 1 | 0 | 0 | 15 | 5 | 5 |
| PINK1 | PD | 130 | 52 | 52 | 30 | 12 | 12 | 20 | 7 | 7 | 80 | 33 | 33 |
| PSEN1 | AD | 197 | 76 | 74 | 4 | 3 | 3 | 185 | 69 | 67 | 8 | 4 | 4 |
| PSEN2 | AD | 25 | 9 | 9 | 5 | 1 | 1 | 13 | 5 | 5 | 7 | 3 | 3 |
| SNCA | PD | 5 | 1 | 1 | 1 | 1 | 1 | 3 | 0 | 0 | 1 | 0 | 0 |
| TARDBP | ALS/FTD | 45 | 26 | 26 | 9 | 3 | 3 | 34 | 21 | 21 | 2 | 2 | 2 |
| VCP | FTD | 20 | 14 | 14 | 0 | 0 | 0 | 18 | 13 | 13 | 2 | 1 | 1 |
| PRNP | Prion | 42 | 5 | 5 | 18 | 4 | 4 | 24 | 1 | 1 | 0 | 0 | 0 |
| Total | | 1075 | 412 | 407 | 259 | 70 | 68 | 523 | 184 | 182 | 293 | 158 | 157 |

**Table 7.2** Pathogenic nature and NeuroX coverage of variants from three online databases, including the Alzheimer Disease & Frontotemporal Dementia Mutation Database (AD&FTDMDB), Parkinson Disease Mutation Database (PDMutDB) and Human Prion Database (HPDB). Variants from the online databases were clustered into the gene they reside in (Gene) and the disease most associated with the gene (Disease). The number of variants in each gene is given (Database), followed by the number designed on the NeuroX (Designed on NeuroX) and finally the number that passed quality control in the final (plink9) NeuroX dataset (Passed QC). These variants are then subcategorised depending on the pathogenicity documented in the database; those that cause disease (Causative), those that have unclear pathogenic nature (Pathogenic Nature Unclear), and those that do not cause disease (Not Pathogenic). Key: AD; Alzheimer's disease, FTD; frontotemporal dementia, ALS; amyotrophic lateral sclerosis, PD; Parkinson's disease.

### 7.1.3. Samples harbouring a causative variant

Genotype counts (in 0/1/2 format) of the final (plink9) NeuroX dataset were created (--hardy), this genotype data was loaded into Excel along with the list of 407 markers and their pathogenicity (identified in **Section 7.1.2**). The combined data were filtered to retain the 407 markers that were also polymorphic and classed as pathogenic. Eight markers fulfilled this criterion. Sample genotype calls for these eight markers in 0/1/2 format were created using the final (plink9) NeuroX dataset (--recode12), using Excel the genotype calls was filtered to identify the samples harbouring a minor allele for any of these eight markers and all sEOAD samples were Sanger sequence in the forward and reverse direction. Control samples harbouring a minor allele for any of these eight markers were not Sanger sequenced as DNA was not available for the UCL controls

```
$ plink --noweb --bfile plink9 --hardy --out plink9Hardy
$ plink --noweb --bfile plink9 --extract ExtractCausative16FamilialGenesSNPs.txt --recode12
  --out plink9Causative16FamilialGenesSNPsRecode12
```

### 7.1.4. Genotype validation

Primers were designed and samples Sanger sequenced as described in **Section 3.1.1** using the following primers. *PARK2* NeuroX marker exm593516 was amplified in sample M177 using forward primer 5' GCT-CGT-GTG-GCA-GAA-CAA-TA 3' and reverse 5' ACA-CCC-CAC-CTC-TGA-CAA-G 3' (Eurogenomics), the product was 202 base pairs (bp) in length. *PARK2* NeuroX marker NeuroX_Gln34fs_Del_A and NeuroX_Gln34fs_Del_AG were amplified in sample M099 using forward primer 5' TCA-GGC-ATG-AAT-GTC-AGA-TTG 3' and reverse primer 5' CCT-TCC-AAT-TTC-CTT-GGT-CA 3' (Eurogenomics), the product was 272 bp in length. *FUS* NeuroX marker NeuroX_16:3119641 was amplified in sample M215 using forward primer 5' TGG-CAA-TCA-AGA-CCA-GAG-TG 3' and reverse 5' GAT-TCA-TGC-AAT-CCT-CCA-CA 3' (Eurogenomics), the product was 243 base pairs (bp) in length.

## 7.2. Results and discussion

The three online databases which document variants linked to familial forms of AD, FTD, PD or Prion disease (AD&FTDMDB, PDMutDB and HPPDB) contained 1196 variants which covered 16 genes. The program MatchingVariants.pl established that 38% (n=412) were designed to be genotyped by markers on the NeuroX. Analysis of the final (plink9) NeuroX dataset (**Section 5**) established that 407 markers passed QC and 182 of these were also classed as causative. sEOAD (n=408) and controls (n= 436) were screened using the final NeuroX dataset (plink9) for samples harbouring at least one minor allele for any of the 182 causative variants. The NeuroX identified three causative variants in three control samples and four causative variants in three sEOAD samples (**Table 7.3**). Sanger sequencing of the three sEOAD samples confirmed two of the four variants (**Figure 7.1**). These results suggest that these variants may contribute susceptibility to multiple types of neurodegeneration.

| Marker | Genomic Position | Variant | Base Change | MAF | Protein Change | Gene | Disease | Sample | Seq |
|---|---|---|---|---|---|---|---|---|---|
| exm593516 | 6:162394349 | rs137853054 | aCg/aTg | 2.8e-04 | p.T240M | *PARK2* | PD (R) | M177 (Case) | C |
| NeuroX_PARK2_Gln34fs_del_AG | 6:162864411-2 | rs55777503 | cga.cAG.ggg/cga.cgg.ggt | 2.9e-04 | p.Q34fs | *PARK2* | PD (R) | M099 (Case) | C |
| NeuroX_PARK2_Gln34fs_del_A | 6:162864412 | rs748142049 | cga.cAg.ggg/cga.cgg.ggg | 8.2e-06 | p.Q34fs | *PARK2* | PD (R) | M099 (Case) | U |
| NeuroX_16:31196410 | 16:31196410 | NA | gGc/gTc | U/K | p.G225V | *FUS* | ALS (R) | M215 (Case) | U |
| exm593474 | 6:161771240 | rs191486604 | gGc/gAc | 9.8e-05 | p.G430D | *PARK2* | PD (R) | DCR00239 (Control) | N/A |
| exm593521 | 6:162475167 | rs9456735 | Atg/Ctg | 0.005 | p.M192L | *PARK2* | PD (R) | OA0318 (Control) | N/A |
| NeuroX_PARK2_Met192Val | 6:162475167 | rs9456735 | Atg/Gtg | U/K | p.M192V | *PARK2* | PD (R) | OA0318 (Control) | N/A |
| NeuroX_rs63750768 | 17:42427630-3 | rs63750768 | gaT.AGT/ga | U/K | p.DS130fs | *GRN* | FTD (D) | 23 Cases & 11 Controls | N/A |

**Table 7.3** Causative variants identified in sEOAD and control samples. The name of the NeuroX marker (Marker) is followed by information about the causative variant genotyped by the marker, including the genomic position of the variant in genome build GRCh37 and format chromosome:base position (Genomic Position), the variant reference ID (Variant). The variant is given at nucleotide level (Base Change) on the sense strand, the minor allele frequency of the variant and base change in the general population according to ExAC (MAF, version and link provided in **Section 12**) and amino acid level (Protein Change). Also listed is the gene the variant resides in (Gene) and the disease most associated with the gene, the medelian pattern of inheritance is given in brackets (Disease). The sample(s) harbouring the variant (Sample) is followed by the results of Sanger sequencing (Seq). Key: U/K, unknown; NA, not available; U, Sanger sequencing did not confirm the genotype of this sample; C, Sanger sequencing confirmed the genotype of this sample; fs, frameshift; FTD, frontotemporal dementia; PD, Parkinson's disease; ALS, amyotrophic lateral sclerosis; R, recessive; D, dominant; Het, heterozygous.

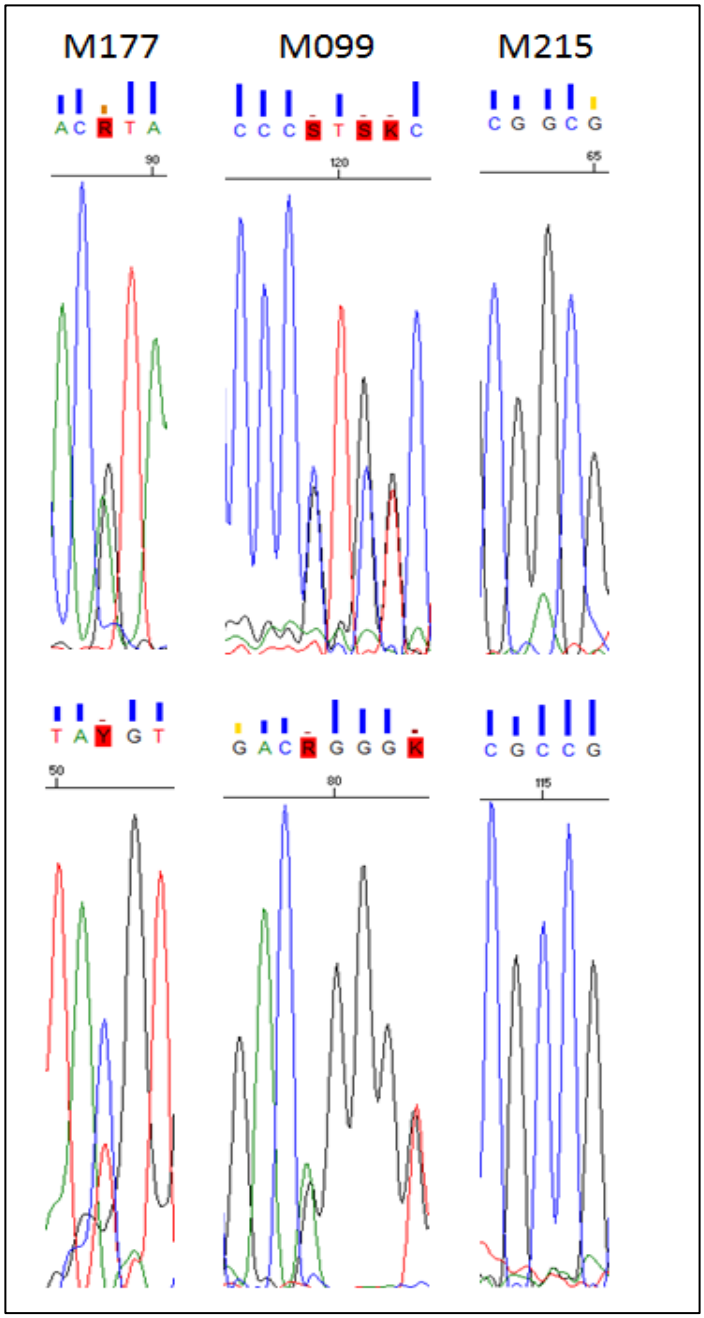



**Figure 7.1** Sequence chromatograms of all three sEOAD samples called by the NeuroX as harbouring one of three causative variants. Each sample was sequenced in the forward (top row) and reverse (bottom row) orientations. Note that the forward orientation does not always correspond to the sense strand. The images were taken from Sequence Scanner (Applied Biosystems) with the variant at the centre and surrounded by two/three bases either side. The sequence chromatograms confirm two of the three calls.

None of the eight causative variants identified were located in *PSEN1* or *PSEN2*, this is unexpected given that 9% of the sEOAD cohort where found to harbour a causative variant in *APP* (**Section 9.2.1**), but this can be explained by the fact the NeuroX only genotypes 37% of the causative variants in *PSEN1*/*PSEN2*. It is likely a small portion of the sEOAD cohort harbour a causative variant in *PSEN1* but have remained unidentified using the NeuroX.

### 7.2.1. Causative variants in *PARK2*

sEOAD sample M177 was confirmed heterozygous for *PARK2* p.T240M. Sample M099 was confirmed heterozygous for *PARK2* p.Q34fs (delAG) and the presence of *PARK2* p.Q34fs (delA) at the same start position was disproved; not surprisingly, on inspection of the Illumina CSV file it was evident that the probes for these two markers were Infinium II and their sequences were identical. As the first base of both variants were the same and first base after both variants were the same, the probes would not have been able to differentiate between the two variants and both were called.

Control sample OA0318 appeared to be heterozygous for two causative variants at the same position; p.M192L (A/C) and p.M192V (A/G) in *PARK2*. On inspection of the Illumina CSV file it was evident that the probes for these two markers were Infinium II and their sequences were identical. As the position has two ambiguous alternative alleles (either a G or C) the assay would not have been able to differentiate between the two variants and both would be called, thus this sample will most definitely harbour one of the two variants but without Sanger sequencing neither variant can be confirmed/disproved. Control sample DCR00239 appeared to be heterozygous for the causative variant p.G430D, however without Sanger sequencing this also cannot be confirmed. Sequencing could not be conducted as no DNA was available for these control samples.

Documented mutations in *PARK2* include point mutations and exon rearrangements, exon deletions and exon duplications, often as homozygote or compound heterozygotes in early-onset PD (EOPD). p.T240M has been seen in EOPD as a compound heterozygote with various exon deletions,[228-230] with exon duplication,[231] as a homozygote[232] and a heterozygote.[233] Heterozygote p.T240M has also been seen twice in late-onset PD (LOPD)[234,235] and five healthy family members.[230] A compound heterozygote with an exon deletion has also been seen in one healthy individual.[230]

In previous studies, p.Q34fs (delAG) has been seen in EOPD as a compound heterozygote with exon deletions,[236-242] with IVS1+1 G>A,[243] with p.C268* T>A,[244] as a homozygote[241,244] and a heterozygote.[245,246] A compound heterozygote with an exon deletion has also been seen in LOPD.[238]

Previous findings would suggest that p.T240M and p.Q34fs (delAG) elicit risk for PD, in particular EOPD when present as a homozygote or compound heterozygote. Finding a heterozygote p.T240M and p.Q34fs (delAG) in our sEOAD cohort would suggest that these mutations could elicit some risk to AD; however each of them was only found in one individual (0.25% of our sEOAD cohort) as a

heterozygote and we don't know if these individuals were compound heterozygotes. Both of these individuals (M117 and M099) had an *APOE* ε4ε4 status and there was nothing unusual about their presentation or progress, which suggests that a misdiagnosis is unlikely. However, M099 had a mother who was said to have had motor neuron disease (MND), so there was likely to have been physical signs in her, possibly consistent with a *PARK2* mutation. A large association study would be needed to establish if these two variants in isolation contribute to the risk of AD, additionally sequencing *PARK2* in these individuals would establish if they were compound heterozygote.

AD and PD are both neurodegenerative disease which have pathological similarities, however the neurodegeneration occurs in different areas of the brain and thus result in different symptoms. There could be a mechanistic overlap between the two diseases; one way to study this is using polygenic risk score to calculate the genetic overlap; as described in a seminal paper involving schizophrenia and bipolar disorder.[247] Notably, a recent study using polygenic risk analysis of PD and late-onset Alzheimer's disease (LOAD) found no significant overlap in the genetics of these two diseases.[248]

### 7.2.2. Causative variant in *FUS*

sEOAD sample M215 was not confirmed for *FUS* p.Q225V (G/T). The cluster plot was inspected to establish call quality and found to be good (**Figure 7.2**). M215 was sequenced from the original stock DNA with the same result as the initial sequencing (unconfirmed), thereby obviating a sample 'mix-up'. The probe for the assay is of Infinium II design and is designed to work for unambiguous SNPs (A/G, A/C, T/C, and T/G). Assuming the NeuroX is correct the sequence result could be the result of allele dropout.[249] It is interesting to note that the sequence chromatogram does show a small spike of an A base (**Figure 7.1**), it is possible that the DNA sample used to genotype case M215 was mosaic for a G/A SNP at the same position and the assay was sensitive enough to call it, however the rest of the sequence chromatogram is quite noisy, suggesting this peak could be the result of low level contamination or background noise. Alternatively, the NeuroX could be incorrect and the sample does not have a variant at this location.

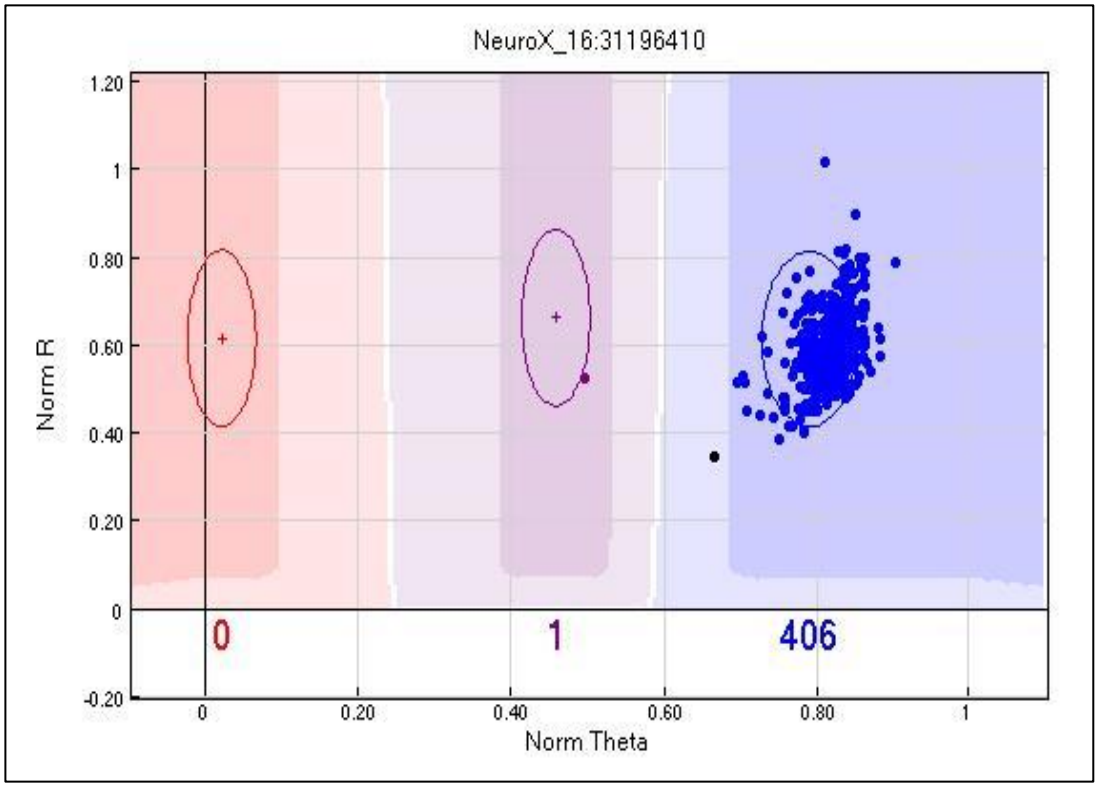



**Figure 7.2** SNP cluster graph of the marker genotyping p.G225V (NeuroX_16:31196410) taken from GenomeStudio and showing the 408 sEOAD samples. Each coloured circle represents one sample, those coloured red are homozygous mutant (n=0), purple are heterozygous (n=1, M215), blue are homozygous wildtype (n=406) and finally those coloured black are not called (n=0).

### 7.2.3. Causative variant in *GRN*

Also documented were 44 samples (23 sEOAD and 11 controls) that appeared to harbour the causative variant p.DS130fs (rs63750768, gaT.AGT/ga) in *GRN*. It had already been documented that there is a common synonymous SNP that overlaps the same start position; p.D130D (rs25646, gaT.agt/gaC.agt).[220] Both variants have a T base at the start of variants; however the deletion has a G base after the deletion which is unable to be differentiated from the C alternative allele of the SNP. As these samples were not Sanger sequenced at this position it cannot be disproved that these samples harbour the causative deletion.

# 8. NeuroX Predicted Pathogenic Variants in sEOAD

## 8.1. Methods

As a way to identify possible additional causative variants in sEOAD, 112 genes that are linked to neurologic diseases (**Table 8.1**) were screened using the final (plink9) NeuroX dataset (**Section 5**), and the predicted pathogenicity generated by VEP in **Section 6.1.3**. The 112 genes are linked to Alzheimer's disease (AD), frontotemporal dementia (FTD), prion disease, multiple sclerosis, PD, amyotrophic lateral sclerosis (ALS), myasthenia gravis, progressive supranuclear palsy (PSP), Charcot Marie Tooth and/or multiple system atrophy (MSA). This section is divided into two and includes the 16 genes linked to familial forms of neurodegeneration (as described in **Section 7**) and 96 other genes. As these genes include the 407 known causative variants already studied in **Section 7**, these variants have been excluded from this analysis.

| Genes | | | | | |
|---|---|---|---|---|---|
| APP | CLCN6 | FGF20 | LAMP3 | PINK1 | SQSTM1 |
| ABCA7 | CLN3 | FIG4 | LRRK2 | PLA2G6 | STBD1 |
| ACE | CLN5 | FUS | MAPT | PLD3 | STK39 |
| ACMSD | CLN6 | GAK | MATR3 | PPT1 | STX1B |
| ALS2 | CLN8 | GBA | MEF2C | PRNP | SYT11 |
| ANG | CLU | GIGYF2 | MFSD8 | PSAP | TAF15 |
| APOE | CR1 | GLA | MS4A4A | PSEN1 | TARDBP |
| ATP13A2 | CSF1R | GLB1 | MS4A6A | PSEN2 | TMEM106B |
| ATXN2 | CST3 | GPNMB | NEU1 | PTK2B | TPP1 |
| BACE1 | CTSD | GRN | NME8 | RIN3 | TREM2 |
| BCHE | DAO | GUSB | NOTCH3 | SETX | TREML2 |
| BIN1 | DCTN1 | HEXA | NPC1 | SGSH | TYROBP |
| BST1 | DNAJC5 | HEXB | NPC2 | SIGMAR1 | UBQLN2 |
| C9orf72 | DSG2 | HIP1R | OPTN | SLC24A4 | VAPB |
| CASS4 | ECE2 | HLA-DRB1 | PANK2 | SNCA | VCP |
| CD2AP | EPHA1 | HLA-DRB5 | PARK2 | SOD1 | VPS35 |
| CD33 | EWSR1 | HNRNPA1 | PARK7 | SORL1 | ZCWPW1 |
| CELF1 | FBXO7 | HNRNPA2B1 | PFN1 | SPG11 | |
| CHMP2B | FERMT2 | INPP5D | PICALM | SPTLC1 | |

**Table 8.1** List of the 112 genes linked neurologic/neurodegenerative diseases. These genes were screened for predicted pathogenic variants and can be broken down into the 16 genes linked to familial forms of neurodegeneration (highlighted in grey) and 96 other genes.

### 8.1.1. List of genes

Genotype counts for all markers in the final (plink9) NeuroX dataset were created (--hardy), this genotype data was loaded into Excel and combined with VEP annotation (**Section 6.1.3**), the list of 112 genes and the list of 407 markers (identified in **Section 7.1.2**).

```
$ plink --noweb --bfile plink9 --hardy --out plink9Hardy
```

### 8.1.2. 16 familial genes

The combined data were filtered to display only markers that fulfilled all the following criteria; were located in one of the 16 familial genes, were not one of the 407 markers, were polymorphic in sEOAD samples, had a minor allele frequency (MAF) < 0.01 in the European cohort of the 1000 Genomes Project and were predicted to be probably pathogenic; which is defined here as a 'deleterious effect' by SIFT (≤0.05) or 'probably damaging' by PolyPhen (≥0.909). Three markers fulfilled this criterion. Sample genotype calls for these three markers in 0/1/2 format were created using the final (plink9) NeuroX dataset and PLINK (--recode12), using Excel the genotype calls was filtered to identify the samples harbouring a minor allele for these three markers.

```
$ plink --noweb --bfile plink9 --extract ExtractPredicted16FamilialGenesSNPs.txt --recode12
  --out Plink9Predicted16FamilialGenesSNPs
```

All sEOAD samples identified were Sanger sequenced in the forward and reverse direction. Primers were designed and samples sequenced as described in **Section 3.1.1-3.1.2** using the following primers: Marker exm1330962 was amplified in sample M820 using forward primer 5' AGA-GCT-GAG-GCA-CCT-TGG-TA 3' and reverse 5' ATG-GCA-GGC-AAT-TCC-AGT-T 3' (Eurogenomics), the product was 235 base pairs (bp) in length. Marker exm1331018 and was amplified in samples M357, M143, M697 and M245 using primers forward primer 5' TCC-ACA-CGT-TCC-TCT-GCT-AA 3' and reverse primer 5' AGC-AGC-CTG-GTT-CTT-TTC-AA 3' (Eurogenomics), the product was 230 bp in length. Marker exm1331027 was amplified in samples M168, M691, M699, M172 and M382 using forward primer 5' GGC-TGG-TGT-TGA-CTC-TTG-GT 3' and reverse 5' TCT-TAC-CAG-AGC-TGG-GTG-GT 3' (Eurogenomics), the product was 206 base pairs (bp) in length.

### 8.1.3. 96 other genes

The combined data were filtered to display only markers that fulfilled all the following criteria; were located in one of the 96 other genes, were polymorphic, had a minor allele frequency < 0.01 in the European cohort of the 1000 Genomes Project and were predicted to be probably pathogenic; which is defined here as a 'deleterious effect' by SIFT (≤0.05) or 'probably damaging' by PolyPhen (≥0.909). 108 markers fulfilled this criterion. Sample genotype calls for these 108 markers in 0/1/2 format were created using the final (plink9) NeuroX dataset and PLINK (--recode12), using Excel the genotype calls was filtered to identify the samples harbouring a minor allele for these 108 markers.

```
$ plink --noweb --bfile plink9 --hardy --out plink9
$ plink --noweb --bfile plink9 --extract ExtractPredicted96OtherGenesSNPs.txt --recode12
  --out Plink9Predicted96OtherGenesSNPs
```

## 8.2. Results and discussion

sEOAD (n= 408) and controls (n= 436) were screened using the final NeuroX dataset (plink9, **Section 5**) for samples harbouring a minor allele for variants predicted to be pathogenic by SIFT or PolyPhen. A total of 112 genes that are linked to neurologic/neurodegenerative diseases were screened; this list includes 16 genes which are more closely linked to familial forms of neurodegeneration (familial genes). A total of 223 sEOAD samples were identified as harbouring at least one minor allele for 111 predicted pathogenic variants; three in the 16 familial genes and 108 in the 96 other genes.

The SIFT and Polyphen algorithms use machine learning to predict the pathogenicity of non-synonymous variants, both algorithms calculate a score of between 0 and 1 per variant. For SIFT the lower the score the more confident the variant is pathogenic, for Polyphen the higher the score the more confident the variant is pathogenic.

### 8.2.1. 16 familial genes

The NeuroX identified ten sEOAD samples harbouring one of three predicted pathogenic missense variants located in one of the 16 familial genes (**Table 8.1**). All ten sEOAD samples were Sanger sequenced in both directions and confirmed seven of the genotypes (**Figure 8.1**). The cluster plots for all three markers were visually examined and they were found to be good (**Figure 8.2**).

| Marker | Position | Variant | Base Change | MAF | Protein Change | Gene | Disease | Sample | Seq |
|---|---|---|---|---|---|---|---|---|---|
| exm1330962 | 17:44055753 | rs144397565 | gGc/gTc | 2.5e-5 | p.G107V | *MAPT* | FTD (D) | M820 | C |
| exm1331018 | 17:44067341 | rs143956882 | tCc/tTc | 0.001 | p.S427F | *MAPT* | FTD (D) | M357 | C |
| | | | | | | | | M143 | C |
| | | | | | | | | M697 | U |
| | | | | | | | | M245 | C |
| exm1331027 | 17:44068850 | rs143624519 | Gcc/Acc | 0.001 | p.A469T | *MAPT* | FTD (D) | M168 | C |
| | | | | | | | | M172 | U |
| | | | | | | | | M382 | U |
| | | | | | | | | M691 | C |
| | | | | | | | | M699 | C |

**Table 8.1** Three predicted pathogenic variants identified in sEOAD samples. The name of the NeuroX marker (Marker) is followed by information about the causative variant genotyped by the marker, including the genomic position of the variant in genome build GRCh37 and format chromosome:base position (Genomic Position), the variant reference ID (Variant). The variant is given at nucleotide level (Base Change) on the sense strand, the minor allele frequency of the variant and base change in the general population according to ExAC (MAF, version and link provided in **Section 12**) and amino acid level (Protein Change). Also listed is the gene the variant resides in (Gene) and the disease most associated with the gene, the medelian pattern of inheritance is given in brackets (Disease). The sample(s) harbouring the variant (Sample) is followed by the results of Sanger sequencing (Seq). Key: rsID, reference single nucleotide ploymorphism identification; FTD, frontotemporal dementia; D, dominant; C, Sanger sequencing confirmed the genotype of this sample; U, Sanger sequencing did not confirm the genotype of this sample.

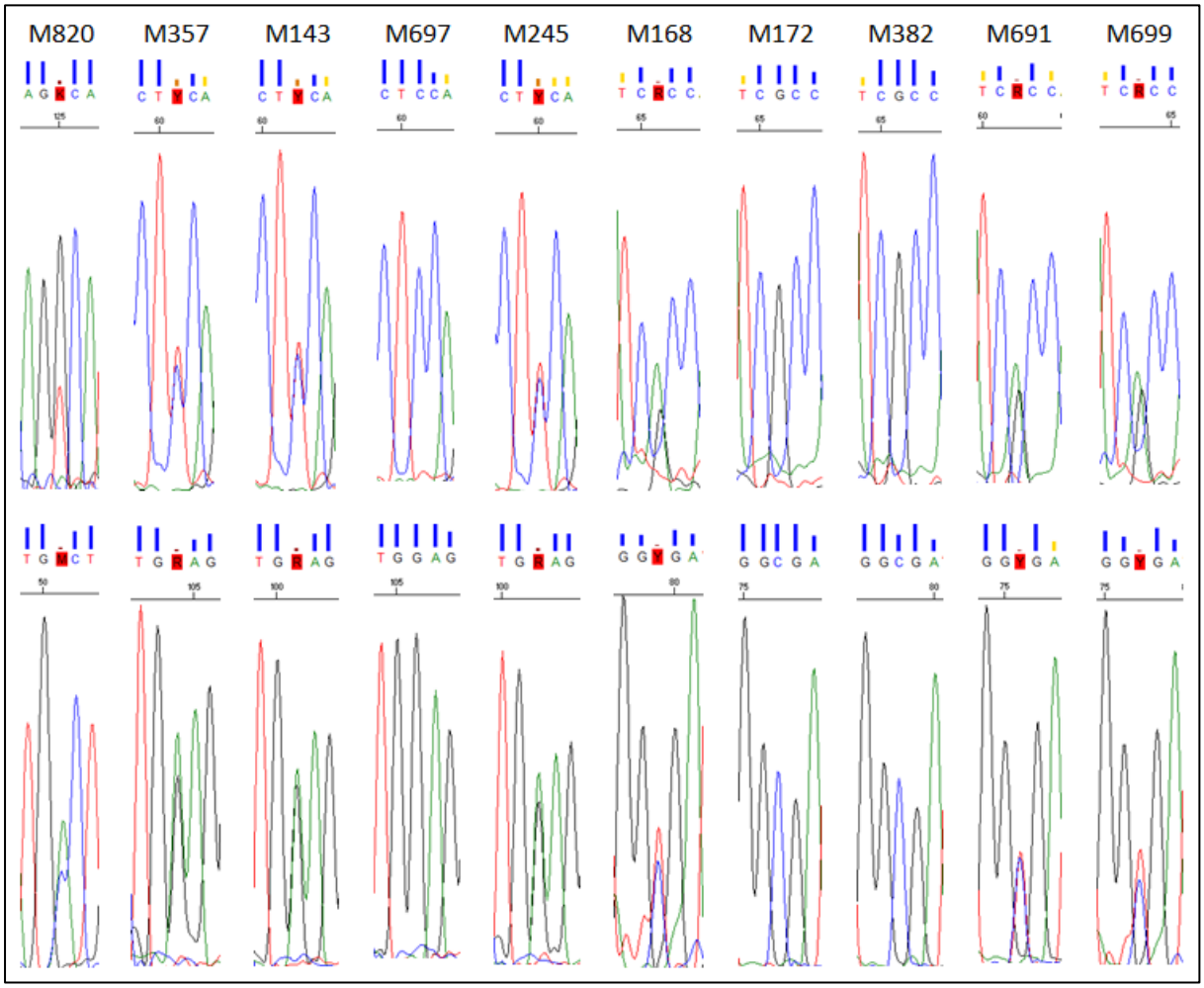



**Figure 8.1** Sequence chromatograms of all ten samples thought to harbour one of three predicted pathogenic variants in *MAPT*. Each sample was sequenced in forward (top row) and reverse (bottom row) orientations. Note that the forward orientation does not always correspond to the sense strand. The images were taken from Sequence Scanner (Applied Biosystems) with the heterozygous variant at the centre and surrounded by two or three bases either side. The sequencing results confirm seven of the samples are heterozygous.

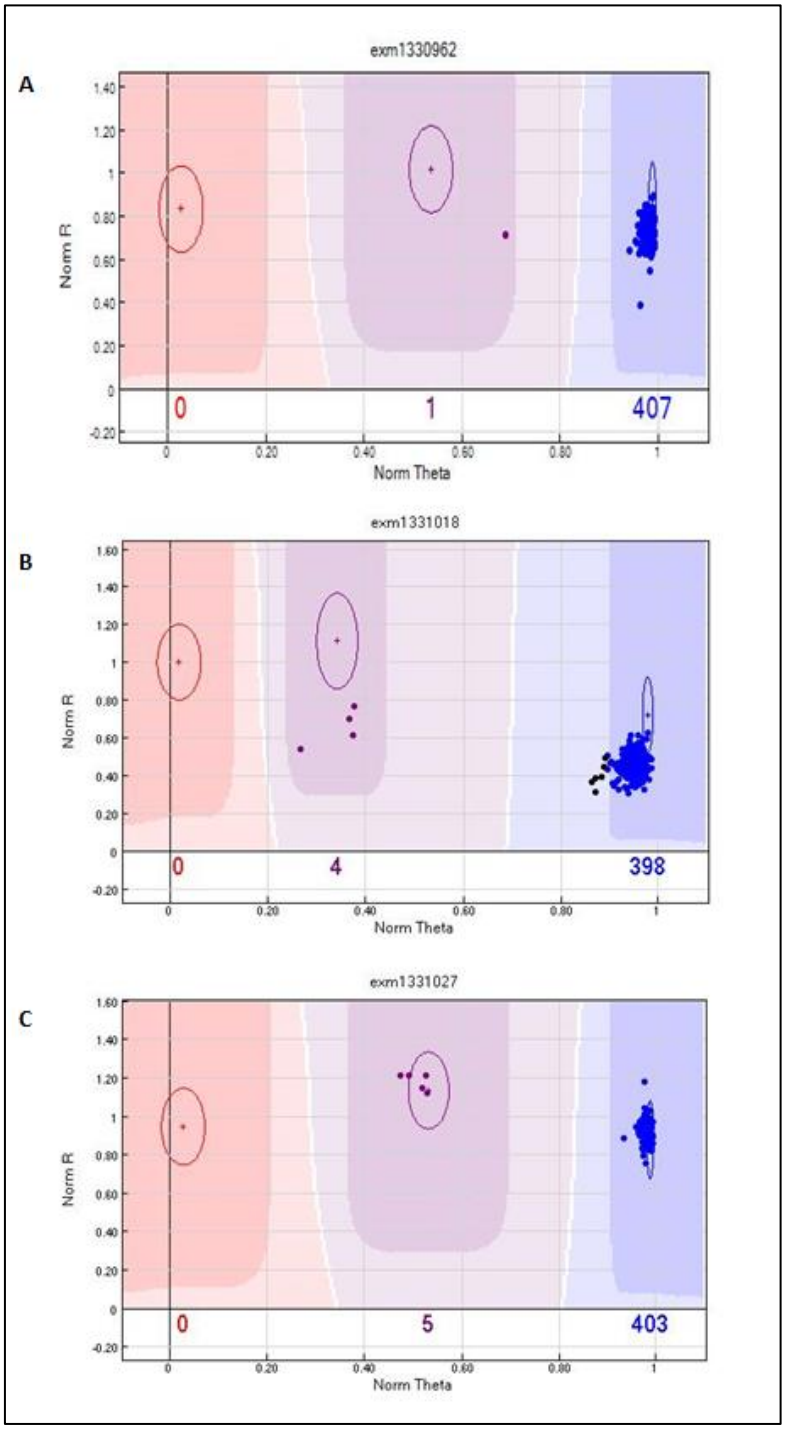



**Figure 8.2** SNP cluster graphs for markers genotyping G107V (A, exm1330962), S427F (B, exm1331018) and A469T (C, exm1331027) taken from GenomeStudio and showing the 408 sEOAD samples. Each coloured circle represents one sample, those coloured red are homozygous minor, purple are heterozygous, blue are homozygous major and finally those coloured black are not called. In the case of graph A, no samples are homozygote minor, one sample (M820) is heterozygous and the remainder 407 samples are homozygous major. In graph B no samples are homozygous minor, four samples (M357, M143, M697 and M245) are heterozygous, six samples are uncalled and the remainder 398 samples are homozygous major. In graph C no samples are homozygous minor , five samples (M168, M172, M382, M691 and M699) are heterozygous and the remainder 403 samples are homozygous major.

All three variant are located in the familial FTD gene *MAPT* which codes for the protein colloquially called tau, variants in tau are conventionally named in reference to a small tau isoform (ENST00000574436) which is 441 amino acid residues in length, has 11 exons and is named htau40 (**Figure 8.3**).

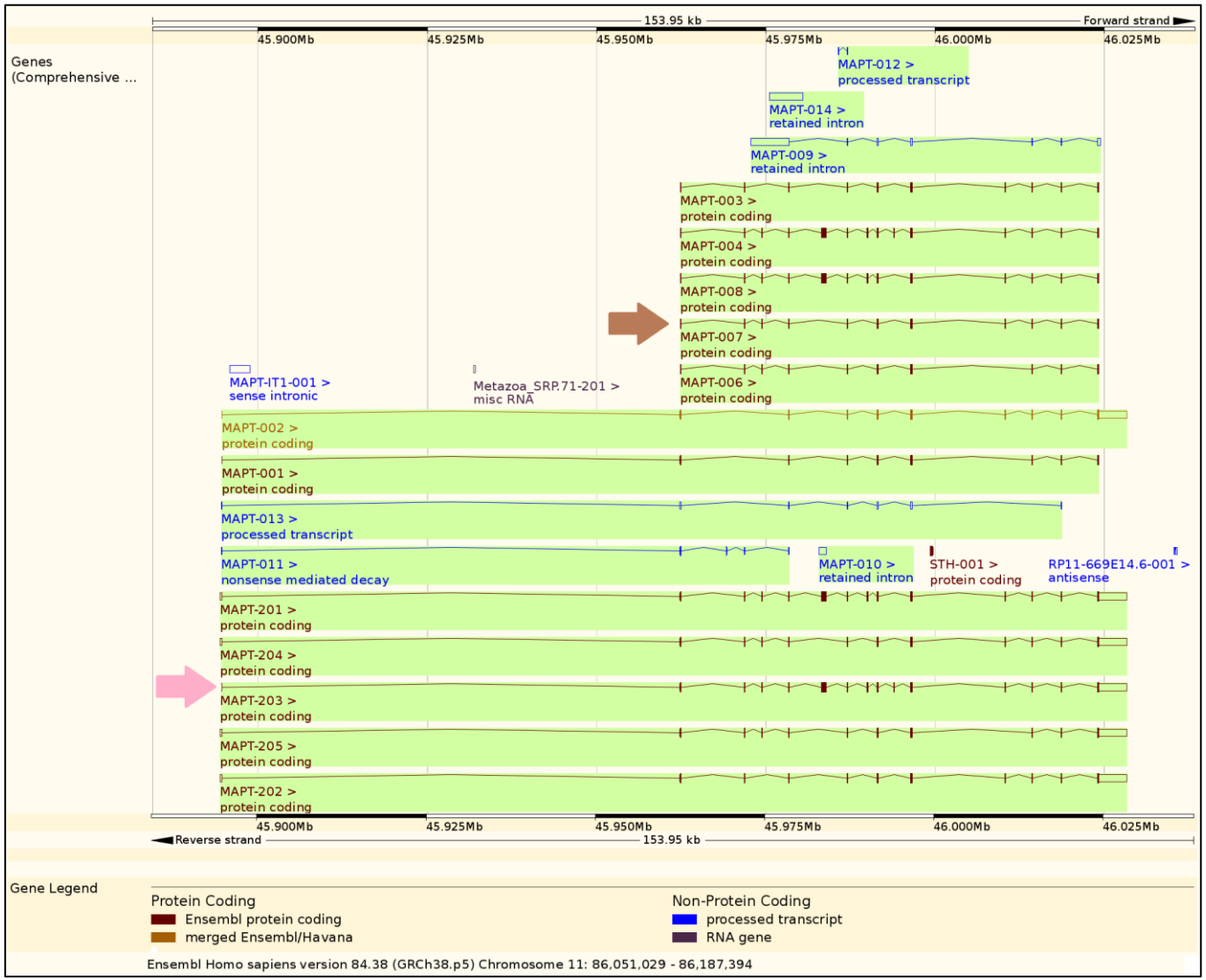



**Figure 8.3** Image illustrating the various transcripts of *MAPT*. Chromosome 17 is depicted horizontally at the top and bottom of the image in matching 25 kb black and white blocks. Between the chromosome depictions are transcript represented using blocks (exons) connected by lines (introns), below each transcript is the name and an arrow signifying direction. Each feature is colour coded and the key is given at the bottom of the figure labelled Gene Legend.
The transcript htau40 is indicated with the brown arrow, it has 11 exons named: 1,2,3,4,5,7,9,10,11,12,13 and 10.
The transcript tau-g is indicated with the pink arrow, it has 15 exons named: -1,1,2,3,4,4a,5,6,7,8,9,10,11,12,13 and 14.
Exons with the same name across both transcripts are coded by the same DNA. Image was taken from Ensembl GRCh38.p5.

sEOAD sample M820 appeared to be heterozygous for p.G170D (rs144397565, G/A), Sanger sequencing confirmed this sample was heterozygous for an alternative allele at the same position p.G170A (rs144397565, G/T). The assay for this marker is Infinium II design, which was unable to differentiate between the two alternative alleles (A and T). When p.G170A was re-run through VEP it was predicted be more pathogenic than p.G170D with a SIFT score of 0 and a Polyphen score of 1. Variant p.G170A is named in reference to the longest tau transcript (ENST00000344290) which is 776 residues in length, has 15 exons and is called tau-g (**Figure 8.3**). Variant p.G170A is located in exon 4 of *MAPT* where no causative variants have been documented.

sEOAD sample M357, M143 and M245 were confirmed to be heterozygous for the variant p.S427F (rs143956882), however M697 was found not to harbour the variant. Variant p.S427F had a SIFT score of 0.02 and a Polyphen score of 0.99. Variant p.S427F is named in reference to the tau-g transcript, it is located in exon 4a which is spliced out in htau40 and thus not present in human brain.[250,251] Not surprisingly no causative variants have been documented in exon4a.This variant has been seen in PD, AD and healthy individuals.[252-254] It is unlikely this variant has a functional effect in the brain if the transcripts containing this variant are not present.

sEOAD samples M168, M691 and M699 were confirmed heterozygous for the variant p.A469T (rs143624519); however M172 and M382 were found not to harbour the variant. Variant p.A469T had a SIFT score of 0.05 and a Polyphen score of 0.30. Variant p.A469T is named in reference to the tau-g transcript; in htau40 it is called p.A152T. This variant significantly increases the risk for both FTD (p.value=0.0005, OR=3.0 (CI: 1.6–5.6)) and AD (p.value=0.004, OR=2.3 (CI: 1.3–4.2)) when compared to controls.[255] *In vitro* site-directed mutagenesis of human tau cDNA showed the variant resulted in less efficient binding to microtubules and a pronounced increase in the formation of tau oligomers.[255] Isogenic human induced pluripotent stem cells generated from fibroblasts saw the variant result in axonal degeneration and cell death.[256] Whether heterozygous p.A469T in humans would cause the same effect is unknown. The variant is located in exon 7, the downstream residue (p.T153) is part of a Threonine-Proline motif that is phosphorylated during the cell cycle[257] and the upstream residue (p.I151) is seen to interact with microtubules using nuclear magnetic resonance[258]. Variant p.A469T could affect the upstream or downstream residue which could explain the experimental observations for this variant.

The majority of causative variants in MAPT are SNPs with an autosomal dominant inheritance and cause FTD. These three variants could elicit susceptibility for sEOAD, further functional work would be needed to establish the variants' true effects and a large association study would be needed to establish if these variants are associated with sEOAD.

### 8.2.2. 96 other genes

The NeuroX identified 213 sEOAD samples harbouring at least one minor allele for one of 108 predicted pathogenic variants located in the 96 other genes. These 108 variants span a total of 56 of the 96 genes. The cluster plots of all 108 markers were visually checked and all were found to be good quality. As the samples harbouring a predicted pathogenic variant in the 96 other genes were not Sanger sequenced their genotypes cannot be confirmed; noticibly, some variants appear to have a minor allele frequency greater than the applied filter of 0.01 (in the European cohort of the 1000 genomes), it is possible that these markers are also genotyping a common alternative allele at the same position (as described in **Section 1.5.1**), alternatively the calls could be correct and they have an increased minor allele frequency in sEOAD.

27 of the 108 markers were located in 14 genes linked to GWAS loci (**Table 8.2**). A similar study[220] which genotyped FAD cases with no known causative variants (n=37) identified two samples from the same family harbouring the same two predicted pathogenic variants; one in ABCA7 (p.L400V) and the second in ZCWPW1 (p.H351R), both are GWAS genes; however a control individual also carried this variant. A different family saw two individuals harbouring the same two predicted pathogenic variants; one in SORL1 (p.H1813Q) and the second in DSG2 (p.V56M), both are also GWAS genes however DSG2 is not thought to be a true susceptibility loci (as described in **Section 1.3.3**). Notably, none of these four variants are the same as those identified in **Table 8.2.** Of the variants listed in **Table 8.2**, 50 sEOAD samples were identified that harboured more than one variant listed in **Table 8.2.**

| Marker | Location | rsID | Gene | BC | AAC | PP | SIFT | Polyphen | Con | N |
|---|---|---|---|---|---|---|---|---|---|---|
| exm1398807 | 19:1043793 | rs146086314 | ABCA7 | Cgg/Tgg | R/W | 334 | D(0) | PR(0.976) | M | 1 |
| exm1398864 | 19:1047169 | rs144852598 | ABCA7 | cTc/cCc | L/P | 620 | D(0) | PR(1) | M | 2 |
| exm1398911 | 19:1049360 | rs199517653 | ABCA7 | Ggg/Agg | G/R | 826 | T(0.15) | PR(0.992) | M | 1 |
| exm1398927 | 19:1051006 | rs143718918 | ABCA7 | cGg/cAg | R/Q | 880 | D(0) | PR(1) | M | 3 |
| NeuroX_dbSNP_rs145632609 | 19:1056941 | rs145632609 | ABCA7 | tGt/tAt | C/Y | 1541 | D(0) | PR(1) | M | 2 |
| exm1399104 | 19:1057343 | rs117187003 | ABCA7 | Gtg/Atg | V/M | 1599 | D(0.02) | PR(1) | M | 5 |
| exm1399110 | 19:1057960 | rs141237099 | ABCA7 | Gtg/Atg | V/M | 1643 | D(0.02) | PR(0.965) | M | 1 |
| exm1399140 | 19:1059056 | rs114782266 | ABCA7 | cGc/cAc | R/H | 1812 | D(0.04) | B(0.049) | M | 13 |
| exm224963 | 2:127821511 | rs143820618 | BIN1 | aaC/aaA | N/K | 232 | D(0) | PR(0.998) | M, SR | 1 |
| exm1551807 | 20:55033660 | rs150016338 | CASS4 | Ctc/Ttc | L/F | 740 | T(0.49) | PR(0.993) | M | 0 (1) |
| exm144887 | 1:207760772 | - | CR1 | Acg/Gcg | T/A | 1858 | T(0.2) | PR(0.98) | M | 16 (1) |
| exm1381184 | 18:29104714 | - | DSG2 | Ata/Gta | I/V | 293 | D(0) | PO(0.932) | M | 62 (5) |
| exm1381259 | 18:29126108 | rs142841727 | DSG2 | gTa/gGa | V/G | 920 | T(0.51) | PR(0.967) | M | 10 |
| exm666292 | 7:143088584 | rs139482378 | EPHA1 | cGc/cAc | R/H | 966 | D(0) | PR(1) | M | 2 |
| exm666416 | 7:143095849 | rs140236236 | EPHA1 | cCg/cTg | P/L | 394 | D(0) | PR(0.995) | M | 4 |
| exm666428 | 7:143095979 | rs56006153 | EPHA1 | Cgt/Tgt | R/C | 351 | D(0.01) | PR(0.978) | M | 1 |
| exm535799 | 6:32497962 | - | HLA-DRB5 | Aag/Gag | K/E | 14 | D(0.02) | B(0) | M | 111 (6) |
| exm914223 | 11:59940535 | rs139777263 | MS4A6A | cTg/cAg | L/Q | 234 | D(0) | PR(1) | M | 1 |
| exm614838 | 7:37924776 | rs202221051 | NME8 | aGa/aCa | R/T | 390 | D(0.01) | PR(0.996) | M | 1 |
| exm691577 | 8:27277505 | rs140538134 | PTK2B | Gat/Aat | D/N | 100 | D(0.03) | PR(0.982) | M | 2 |
| exm1122447 | 14:93119307 | rs143827583 | RIN3 | aCg/aTg | T/M | 638 | D(0.02) | PR(1) | M | 1 |
| exm1122469 | 14:93142861 | rs147042536 | RIN3 | Tat/Cat | Y/H | 793 | D(0) | PR(1) | M | 6 |
| exm1122198 | 14:92792313 | rs150573991 | SLC24A4 | Aca/Gca | T/A | 78 | D(0.02) | B(0.451) | M | 1 |
| exm964272 | 11:121384931 | rs150609294 | SORL1 | aAt/aCt | N/T | 371 | D(0) | PO(0.507) | M | 1 |
| exm964491 | 11:121485586 | rs117725215 | SORL1 | aAt/aGt | N/S | 1809 | T(0.11) | PR(0.98) | M | 1 |
| exm964528 | 11:121495816 | rs140327834 | SORL1 | gAt/gTt | D/V | 2065 | D(0) | PR(1) | M | 1 |
| exm640645 | 7:100016781 | rs141450215 | ZCWPW1 | gAg/gGg | E/G | 105 | D(0) | PR(1) | M | 4 |

**Table 8.2** 27 markers located in 14 genes attributed to AD GWAS loci. These markers genotype variants that are predicted to be pathogenic by SIFT or Polyphen. The name of the marker (Marker) is followed by information about the variant, including the genomic position of the variant in genome build GRCh37 (Location). The marker information is preceeded with information from VEP, including the variant reference ID (rsID) and the gene it resides in (Gene), base change of the variant (BC), amino acid change (AAC), protein position affected (PP), SIFT score (SIFT), Polyphen score (Polyphen) and the predicted most severe consequence(s) (Con). Finally the number of heterozygote samples and homozygous mutant samples in sEOAD samples is given in the final column (N) with the number of homozygous mutant samples given in brackets where applicable. Key: D; deleterious, T; tolerated, PR; probably damaging, PO; possible damaging, B; benign, M; missense, SR; splice region.

The remaining 81 markers were located in an additional 42 genes (**Table 8.3**). One variant in *TREM2* (p.D87N) is of particular interest, this variant is called in one case (AD208) as a heterozygote. This variant has been significantly associated with AD (p.value=0.02) in a study by Guerreiro and collegues where a minor allele was seen in 6 of the 1091 cases and in none of the 1105 controls.[50]

| Marker | Location | rsID | Gene | B | AA | PP | SIFT | Polyphen | Con | N |
|---|---|---|---|---|---|---|---|---|---|---|
| exm1344029 | 17:61557724 | rs141543325 | ACE | Cgc/Tgc | R/C | 228 | D(0) | PR(1) | M | 1 |
| exm1344069 | 17:61559850 | rs150466411 | ACE | aCg/aTg | T/M | 381 | D(0) | PR(1) | M | 1 |
| exm1344071 | 17:61559883 | rs138418851 | ACE | aTg/aCg | M/T | 392 | D(0) | PR(0.999) | M | 1 |
| exm1344136 | 17:61562654 | rs147429960 | ACE | tCc/tGc | S/C | 660 | D(0.01) | PR(0.984) | M | 3 |
| exm1344195 | 17:61568577 | rs3730043 | ACE | aCg/aTg | T/M | 916 | D(0) | PR(1) | M | 10 |
| exm1344206 | 17:61570816 | rs141750591 | ACE | Gtg/Atg | V/M | 978 | D(0) | PR(1) | M | 1 |
| exm257473 | 2:202575717 | rs61757691 | ALS2 | atA/atG | I/M | 1373 | D(0.04) | B(0.18) | M | 3 |
| exm257513 | 2:202591249 | rs200706696 | ALS2 | gGa/gAa | G/E | 1069 | D(0) | PO(0.902) | M | 1 |
| exm257579 | 2:202619239 | rs201161419 | ALS2 | Gat/Aat | D/N | 543 | D(0) | PR(1) | M | 1 |
| exm257608 | 2:202625992 | rs200733209 | ALS2 | aTg/aCg | M/T | 242 | D(0) | PR(0.972) | M | 2 |
| exm22932 | 1:17316771 | rs200924194 | ATP13A2 | Cag/Gag | Q/E | 755 | D(0) | PO(0.542) | M | 1 |
| exm362831 | 3:165547569 | rs28933390 | BCHE | gGt/gTt | G/V | 418 | D(0.01) | B(0.351) | M | 2 |
| exm390825 | 4:15720540 | rs34163939 | BST1 | Ggg/Agg | G/R | 239 | D(0.03) | PR(1) | M | 1 |
| exm15449 | 1:11894056 | rs148630210 | CLCN6 | Ttt/Ctt | F/L | 499 | D(0.05) | B(0.001) | M | 1 |
| exm15466 | 1:11894592 | rs150830522 | CLCN6 | Gtg/Atg | V/M | 580 | D(0) | PO(0.946) | M | 1 |
| exm1229460 | 16:28488944 | rs146610181 | CLN3 | Cac/Aac | H/N | 404 | D(0.02) | PR(0.999) | M | 2 |
| exm1229500 | 16:28498845 | rs144770450 | CLN3 | aGt/aAt | S/N | 131 | D(0) | PR(0.958) | M | 1 |
| exm1072875 | 13:77574606 | rs138611001 | CLN5 | aaC/aaA | N/K | 242 | T(0.05) | PR(1) | M | 17 |
| exm494786 | 5:149456964 | rs146406037 | CSF1R | aAc/aTc | N/I | 255 | D(0.03) | PO(0.88) | M | 1 |
| exm494835 | 5:149460542 | rs56048668 | CSF1R | gTg/gGg | V/G | 32 | D(0) | PR(1) | M | 6 |
| exm1034555 | 12:109293187 | rs143550642 | DAO | cGg/cAg | R/Q | 283 | D(0) | PR(1) | M | 1 |
| exm203983 | 2:74592252 | rs72659383 | DCTN1 | cGg/cAg | R/Q | 1049 | T(0.13) | PR(1) | M | 2 |
| exm204005 | 2:74593944 | rs150928856 | DCTN1 | cCt/cGt | P/R | 811 | D(0.02) | PO(0.848) | M | 1 |
| NeuroX_dbSNP_rs139312819 | 20:62560745 | rs139312819 | DNAJC5 | gCg/gTg | A/V | 63 | D(0) | PR(1) | M | 1 |
| exm368506 | 3:183975459 | rs146013597 | ECE2 | aCg/aTg | T/M | 132 | D_LC(0) | PR(0.989) | M | 2 |
| exm368611 | 3:184005719 | rs35875049 | ECE2 | cGg/cAg | R/Q | 571 | T(0.46) | PR(0.973) | M | 1 |
| NeuroX_dbSNP_rs149515968 | 3:184008878 | rs149515968 | ECE2 | Tcc/Acc | S/T | 747 | T(0.39) | PR(0.988) | M | 5 |
| exm2007036 | 22:32894494 | rs34316445 | FBXO7 | Gat/Cat | D/H | 516 | D(0) | PO(0.894) | M | 2 |
| NeuroX_dbSNP_rs121908287 | 6:110036336 | rs121908287 | FIG4 | aTt/aCt | I/T | 41 | D(0) | PR(1) | M | 1 |
| exm379355 | 4:845686 | rs55801437 | GAK | caG/caC | Q/H | 1120 | D(0.03) | PR(0.993) | M | 1 |
| exm379592 | 4:898459 | rs146710139 | GAK | cGg/cAg | R/Q | 164 | D(0.02) | B(0.13) | M | 2 |
| exm106217 | 1:155205634 | - | GBA | aAc/aGc | N/S | 409 | D(0.02) | PO(0.891) | M, SR | 4 |
| exm1649673 | X:100653420 | - | GLA | Gat/Tat | D/Y | 313 | D(0) | PR(0.999) | M | 4 |
| exm609425 | 7:23286529 | rs140122424 | GPNMB | cCa/cTa | P/L | 18 | T(0.25) | PR(1) | M | 5 |
| exm609485 | 7:23300345 | rs35363287 | GPNMB | cCg/cTg | P/L | 324 | T(0.09) | PR(0.983) | M | 2 |
| exm609529 | 7:23313763 | rs145920361 | GPNMB | Cgt/Tgt | R/C | 547 | D(0.04) | PR(0.986) | M | 1 |
| NeuroX_dbSNP_rs148199798 | 15:72641570 | rs148199798 | HEXA | tCt/tGt | S/C | 279 | D(0.01) | PR(0.991) | M | 1 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| exm1174638 | 15:72642925 | rs121907970 | HEXA | Cgg/Tgg | R/W | 247 | D(0) | PR(1) | M | 1 |
| exm461072 | 5:74014629 | rs28942073 | HEXB | cCg/cTg | P/L | 417 | D(0.01) | B(0.297) | M | 1 |
| exm1046873 | 12:123335398 | rs151322438 | HIP1R | gCg/gTg | A/V | 152 | D(0.01) | PO(0.954) | M | 1 |
| exm1046883 | 12:123338616 | rs144603149 | HIP1R | Gcc/Acc | A/T | 202 | T(0.1) | PR(0.987) | M | 1 |
| exm1046913 | 12:123340138 | rs149504879 | HIP1R | aAt/aGt | N/S | 345 | D(0.03) | B(0.151) | M | 1 |
| exm1046961 | 12:123341637 | rs140743610 | HIP1R | Cgg/Tgg | R/W | 564 | D(0.01) | PR(0.994) | M | 2 |
| exm1046974 | 12:123342688 | rs79610396 | HIP1R | Gag/Aag | E/K | 619 | D(0.04) | PO(0.791) | M | 1 |
| exm914293 | 11:60073664 | rs62000397 | MS4A4A | aCc/aTc | T/I | 213 | D(0.01) | B(0.19) | M | 1 |
| exm1436897 | 19:15289850 | rs55882518 | NOTCH3 | cAt/cTt | H/L | 1235 | D(0.02) | B(0) | M | 7 |
| exm1436908 | 19:15290007 | rs10408676 | NOTCH3 | Gtg/Atg | V/M | 1183 | D(0.02) | PO(0.747) | M | 3 |
| exm1436916 | 19:15290236 | - | NOTCH3 | caC/caA | H/Q | 1133 | D(0) | PO(0.66) | M | 18 |
| exm1437060 | 19:15302941 | rs147373451 | NOTCH3 | cAt/cGt | H/R | 170 | D(0.01) | PO(0.776) | M | 4 |
| exm1378433 | 18:21118528 | rs80358257 | NPC1 | Cct/Gct | P/A | 1007 | D(0) | PR(1) | M | 1 |
| exm1378449 | 18:21119839 | rs34302553 | NPC1 | Ggc/Agc | G/S | 911 | T(0.14) | PR(0.999) | M | 2 |
| exm1114740 | 14:74959940 | rs147602717 | NPC2 | cTc/cCc | L/P | 13 | D(0.01) | PR(0.984) | M | 1 |
| NeuroX_10:13174107 | 10:13174107 | rs377219791 | OPTN | gCg/gTg | A/V | 481 | - | PR(0.998) | M | 1 |
| exm1523061 | 20:3869884 | rs148036492 | PANK2 | gAt/gTt | D/V | 46 | D_LC(0.03) | B(0) | M | 2 |
| NeuroX_rs41279408 | 20:3903937 | rs41279408 | PANK2 | cCg/cTg | P/L | 570 | D_LC(0) | PR(0.981) | M | 1 |
| exm1282264 | 17:4849268 | rs140547520 | PFN1 | gAa/gGa | E/G | 117 | D(0) | PR(1) | M | 2 |
| exm2010252 | 22:38522424 | rs76718524 | PLA2G6 | Cgg/Tgg | R/W | 461 | D(0.04) | PR(0.999) | M | 1 |
| exm1608187 | 22:38541545 | rs142530390 | PLA2G6 | Cac/Gac | H/D | 109 | D(0.01) | B(0.037) | M | 0 (1) |
| NeuroX_dbSNP_rs148519599 | 10:73588831 | rs148519599 | PSAP | Cgt/Tgt | R/C | 127 | D(0.05) | PR(0.991) | M | 1 |
| exm790906 | 9:135140063 | rs202121071 | SETX | Cat/Tat | H/Y | 2533 | D(0.01) | B(0.034) | M | 1 |
| NeuroX_9:135201718 | 9:135201718 | - | SETX | tTt/tCt | F/S | 1756 | D(0) | PR(1) | M | 1 |
| NeuroX_dbSNP_rs112089123 | 9:135202325 | rs112089123 | SETX | Tgt/Ggt | C/G | 1554 | D(0.01) | B(0.108) | M | 3 |
| exm791119 | 9:135204506 | rs150532677 | SETX | Aaa/Gaa | K/E | 827 | D_LC(0.05) | B(0.024) | M | 1 |
| exm791198 | 9:135211691 | rs138538492 | SETX | tAt/tGt | Y/C | 237 | D(0) | PR(1) | M | 1 |
| exm791202 | 9:135211747 | rs117861188 | SETX | aaG/aaC | K/N | 218 | D(0.01) | PR(0.983) | M | 2 |
| exm791207 | 9:135218103 | rs145438764 | SETX | Ttg/Gtg | L/V | 158 | D(0.01) | PR(0.999) | M | 4 |
| exm747347 | 9:34635679 | rs11559048 | SIGMAR1 | Cgg/Tgg | R/W | 208 | D(0) | PR(1) | M | 1 |
| NeuroX_21:33038771 | 21:33038771 | - | SOD1 | aGt/aTt | S/I | 60 | D(0) | PR(0.999) | M | 1 |
| exm1157672 | 15:44918698 | rs144012151 | SPG11 | aTt/aCt | I/T | 692 | D(0) | PO(0.89) | M | 1 |
| exm1157728 | 15:44949354 | rs80338868 | SPG11 | Gtt/Att | V/I | 270 | T(0.13) | PR(0.975) | M | 8 |
| exm510196 | 5:179252184 | rs11548633 | SQSTM1 | Aag/Gag | K/E | 238 | D(0.02) | PO(0.94) | M | 1 |
| exm510247 | 5:179263445 | rs104893941 | SQSTM1 | cCg/cTg | P/L | 392 | D(0) | PR(1) | M | 1 |
| exm1313053 | 17:34149718 | - | TAF15 | tAt/tGt | Y/C | 122 | D_LC(0) | PO(0.943) | M | 2 |
| exm1313094 | 17:34171358 | - | TAF15 | cGt/cAt | R/H | 388 | D_LC(0) | PR(0.999) | M | 2 |
| exm886203 | 11:6636698 | rs146798796 | TPP1 | aAt/aTt | N/I | 414 | D(0) | PR(0.99) | M | 4 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| exm886228 | 11:6637982 | rs200138397 | TPP1 | Cgg/Tgg | R/W | 266 | D(0.02) | B(0.003) | M | 1 |
| exm545524 | 6:41129133 | rs142232675 | TREM2 | Gat/Aat | D/N | 87 | T(0.43) | PR(1) | M | 1 |
| exm545551 | 6:41162204 | rs115991880 | TREML2 | agC/agA | S/R | 248 | D(0.03) | B(0.034) | M | 5 |
| exm545584 | 6:41166063 | rs147506354 | TREML2 | Gtt/Ttt | V/F | 54 | D(0) | PR(1) | M | 1 |
| exm545598 | 6:41166155 | - | TREML2 | gAc/gGc | D/G | 23 | D(0.03) | B(0.044) | M | 88 (9) |
| exm1552869 | 20:57016076 | rs143144050 | VAPB | atG/atA | M/I | 170 | D(0.04) | B(0.154) | M | 2 |

**Table 8.3** Remaining 81 markers located in 42 additional genes. These markers genotype variants that are predicted to be pathogenic by SIFT or Polyphen. The name of the marker (Name) is followed by information about the variant, including the genomic position of the variant in genome build GRCh37 (Location). The marker information is preceeded with information from VEP, including the variant reference ID (rsID) and the gene it resides in (Gene), base change of the variant (BC), amino acid change (AAC), protein position effected (PP), SIFT score (SIFT), Polyphen score (Polyphen) and the predicted most severe consequence(s) (Con). Finally the number of heterozygote samples and homozygous mutant samples in sEOAD is given in the final column (N) with the number of homozygous mutant samples given in brackets where applicable. Key: D; deleterious, DL; deleterious low confidence, T; tolerated, PR; probably damaging, PO; possible damaging, B; benign, M; missense, SR; splice region.

# 9. Analyses of All NeuroX Markers in sEOAD

## 9.1. Methods

To ascertain which variants or sets of variants were most associated with sEOAD, several analyses were performed, including logistic regression to test all variants individually, SKAT-O to test variants in all genes and PLINK set-test to test variants in all pathways.

### 9.1.1. All variants

PLINK (version and link provided in **Section 12**) was used to perform single variant analysis. Logistic regression was performed on all 265,045 markers in the final (plink9) NeuroX dataset (**Section 5**). The logistic regression adjusted for sample gender, *APOE* ε4 dosage, *APOE* ε2 dosage and ancestry (top 10 principle components). Two runs were conducted, one adjusting for the gender calculated in **Section 5.1.3** and a second adjusting for the documented gender. A $-\log_{10}$ quantile-quantile (QQ) plot of the resultig p-values was created using R (version and link provided in **Section 12**). A significance cut-off of 1.89E-7 was calculated using Bonferroni correction. Power was calculated using Quanto (version and link provided in **Section 12**) in a 'gene-only' hypothesis and 'log-additive' model.

```
$ plink --noweb --bfile plink9 --logistic --covar Plink9Covariates.txt  --covar-name
  DocumentedGender,e2,e4,PCA1,PCA2,PCA3,PCA4,PCA5,PCA6,PCA7,PCA8,PCA9,
  PCA10 --hide-covar --ci 0.95 --out Plink9DatabaseGender
$ plink --noweb --bfile plink9 --logistic --covar Plink9Covariates.txt --covar-name
  CalculatedGender,e2,e4,PCA1,PCA2,PCA3,PCA4,PCA5,PCA6,PCA7,PCA8,PCA9,
  PCA10 --hide-covar --ci 0.95 --out Plink9CalculatedGender
```

For the 11 markers with the most significant p.values, VarElect was used to identify if there was a direct or indirect link between the genes the markers resided in and the search terms "Alzheimer" OR "Dementia" OR "Neurodegeneration". Gene ID was taken from the results of VEP (**Section 6.1.3**) where possible, for intergenic markers the nearest upstream and downstream protein coding genes in Ensembl Genome Browser GRCh38 (link provided in **Section 12**) were used. Linkage disequilibrium (LD) of the 11 markers was established using SNP Annotation and Proxy Search (SNAP, version and link provided in **Section 12**) using default parameters 1000 Genomes Pilot 1 data, $r^2$ threshold of 0.8, CEU population and a distance limit of 500bp. Features overlapping the position of the 11 markers was established using Ensembl Genome Browser GRCh38.p5 (link provided in **Section 12**), features included genes, transcripts and multicell regulatory regions.

### 9.1.2. Variants from GWA studies

To date 29 variants have been linked to AD via GWA studies, these variants can be broken down into two lists, those that were found as a result of the recent meta-analysis[48] and those that were identified before this time which are catalogued in the AlzGene online database (link provided in **Section 12**).

MatchingVariants.pl (**Section 11.7**) was used to identify which of the 29 GWAS variants were designed on the NeuroX. The input data for MatchingVariants.pl required the variant start base position together with alleles A and B for both the GWA variants and NeuroX variants; for the GWA variants this data was taken from the meta-analysis paper[48] and AlzGene, for the NeuroX variants this data was taken from the results of Alleles.pl (as described in **Section 6.1.1**).

```
$ perl MatchingVariants.pl FileIn1.txt FileIn2.txt N FileOut1.txt FileOut2.txt
```

### 9.1.3. All genes

Using GeneAnalysis.r (**Section 11.12**), SKAT-O (link provided in **Section 12**) was used to perform gene-based burden analysis. SKAT-O required a gene;marker set file which was created using the ANNOVAR annotation (as described in **Section 6.1.2**). The set file for SKAT-O was created using three programs in series SwitchPrimaryKey.pl (**Section 11.8**) to create a gene;marker list, FormatPrimaryKey.pl (**Section 11.9**) and finally RemoveDuplicates.pl (**Section 11.10**). The resulting set file contained gene;marker with each of the 265,828 markers binned into 20,439 genes where one marker can occur in more than one gene. To accommodate the data format required by SKAT-O, Excel was used to re-code the marker names to numbers (1 to 267,607) and change how the data was delimited; from semi colon to tab.

All 265,045 markers in the final (plink9) NeuroX dataset (**Section 5**) that were binned within one of the ANNOVAR 20,439 genes were used in SKAT-O analysis. SKAT-O was performed whilst adjusting for gender, *APOE* ε4 dosage, *APOE* ε2 dosage and ancestry (top 10 principle components). Two runs were conducted, one adjusting for the gender calculated in **Section 5.1.3** and a second adjusting for the documented gender. A significance cut-off of 2.45E-6 was calculated using Bonferroni correction.

For the ten genes with the most significant p.values, VarElect was used to identify if there was a direct or indirect link between the genes and the search terms "Alzheimer" OR "Dementia" OR "Neurodegeneration". The level of linkage disequilibrium (LD) of the markers in the ten genes was established using SNAP as in **Section 9.1.1**.

```
$ plink --noweb --bfile plink9 --update-map UpdateMarkerNamesToNumber.txt
  --update-name --make-bed --out Plink9SKATO
$ perl SwitchPrimaryKey.pl MarkerGene.txt GeneMarker.txt
$ perl FormatPrimaryKey.pl GeneMarker.txt GeneMarkerSkato.txt 1
$ perl RemoveDuplicates.pl GeneMarkerSkato.txt GeneMarkerSkatoUnique.txt 0
$ R CMD BATCH GeneAnalysis.r
$ R CMD BATCH GeneAnalysis.r
```

### 9.1.4. All pathways

Using PLINK (version and link provided in **Section 12**), set-test was used to perform pathway analysis. Set-test required a pathway;marker set file which was created in two steps. Firstly a pathway;gene set was created; 54,341 reference gene sequences (refSeq) were downloaded from UCSC Table Browser (version and link provided in **Section 12**), each reference gene sequence is a transcript and was annotated with a gene name. RemoveDuplicates.pl (**Section 11.10**) was used to create a list of unique gene names (26,455) which was uploaded to Reactome Analysis Tool (version and link provided in **Section 12**). Reactome created a pathway;gene file where one gene can occur in one or more of the 1,889 pathways, this annotation was combined with the marker;gene annotation from ANNOVAR (as described in **Section 6.1.2**) using CombineTwoPrimaryKeys.pl (**Section 11.11**) to create the pathway;marker file. To accommodate the file format required by PLINK set-test the pathway;marker data was reformatted using FormatPrimaryKey.pl (**Section 11.9**). The resulting data contained markers binned into one or more of the 1,889 pathways.

All 265,045 markers in the final (plink9) NeuroX dataset (**Section 5**) that where binned within one of the 1,889 pathways were used in set-test analysis. Set-test was performed with 10,000 permutations while adjusting for documented gender, *APOE* ε4 dosage, *APOE* ε2 dosage and ancestry (top 10 principle components). Two runs were conducted, one adjusting for the gender calculated in **Section 5.1.3** and a second adjusting for the documented gender. A significance cut-off of 2.65E-5 was calculated using Bonferroni correction.

```
$ perl RemoveDuplicates.pl Genes.txt GenesUnique.txt 0
$ perl CombineTwoPrimaryKeys.pl PathwayGene.txt MarkerGene.txt PathwayMarker.txt
$ perl FormatPrimaryKey.pl PathwayMarker.txt PathwayMarkerPlink.txt 2
$ plink --noweb --bfile plink9 --set-test --set PathwayMarkerPlink.txt --mperm 10000
  --covar Plink9Covariates.txt --covar-name
  CalculatedSex,e2,e4,PCA1,PCA2,PCA3,PCA4,PCA5,PCA6,PCA7,PCA8,PCA9,PCA10
  --logistic --out Plink9PathwayAnalysisCalculatedSex
$ plink --noweb --bfile plink9 --set-test --set PathwayMarkerPlink.txt --mperm 10000
  --covar Plink9Covariates.txt --covar-name
  DatabaseSex,e2,e4,PCA1,PCA2,PCA3,PCA4,PCA5,PCA6,PCA7,PCA8,PCA9,PCA10
  --logistic --out Plink9PathwayAnalysisDatabaseSex
```

## 9.2. Results and discussion

Association analyses was performed using the final (plink9) NeuroX dataset (**Section 5**), this dataset included genotype data for 265,049 markers, 408 sEOAD samples and 436 control samples. Analyses included logistic regression of all variants using PLINK, gene analysis using SKAT-O and pathway set-test using PLINK. Additionally, variants associated with LOAD were genotyped on the NeuroX and these were highlighted in the logistic regression results. No genes or pathways were significant after adjusting for multiple testing; however two LOAD associated variants were significantly associated with sEOAD.

### 9.2.1. All variants

Adjusting for documented gender in logistic regression gave the same top 11 markers as adjusting for calculated gender; the only difference was the order of significance. **Figure 9.1** shows that $-\log_{10}$ quantile-quantile plot of the observed p-values. The plot shows a slight underrepresentation of all p-values signifying that the data may have been overcorrected; this is suprising considering Twstats demonstrated that the top 11 principle components showed significant eigenvalues (p-value < 0.05).

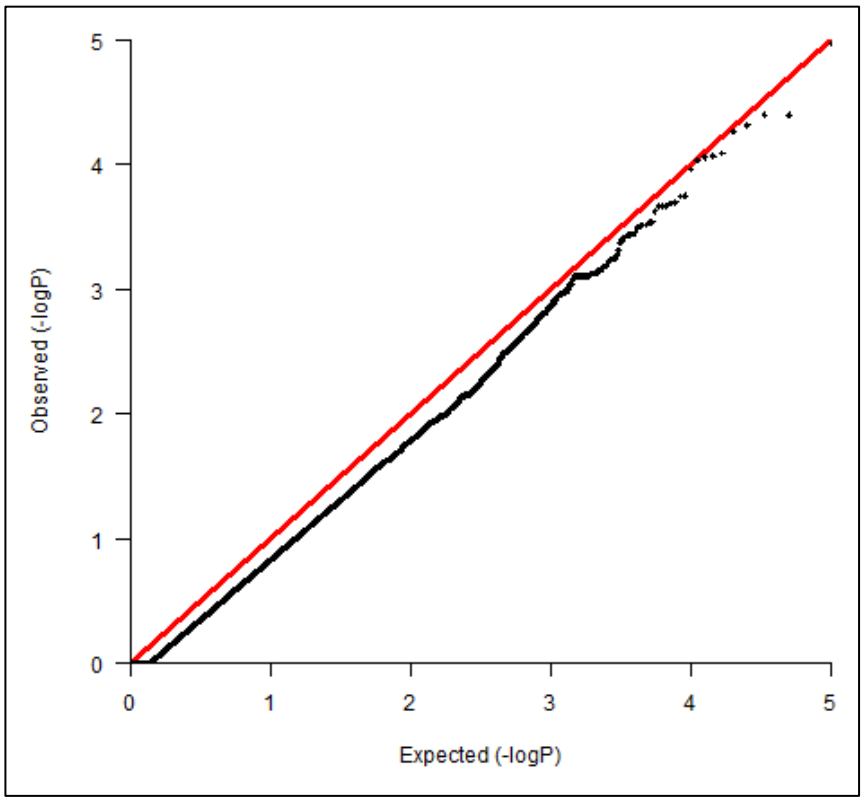



**Figure 9.1** Graph showing the $-\log_{10}$ quantile-quantile plot of the p-values resulting for single variant analysis of 265,049 markers on the NeuroX. The red line is the expected p-values by chance and the black dots are the observed p-values.

**Table 9.1** lists the top 11 markers when adjusting for calculated gender and shows that none reached significant association with sEOAD after correcting for multiple testing of all variants (p.value < 1.89E-7). **Table 9.2** lists the details of the top markers.

| Marker | p-value | OR | Power (%) |
|---|---|---|---|
| exm2264899 | 1.1E-05 | 1.7 (1.3–2.1) | 100 |
| exm849171 | 4.1E-05 | 2.8 (1.7–4.5) | 100 |
| exm849207 | 4.1E-05 | 2.8 (1.7–4.5) | 100 |
| exm2269839 | 4.9E-05 | 1.6 (1.3–2.0) | 100 |
| exm-rs12093699 | 5.5E-05 | 0.6 (0.5–0.8) | 99 |
| exm-rs10242744 | 8.2E-05 | 1.9 (1.4–2.6) | 100 |
| exm928279 | 8.3E-05 | 1.6 (1.3–2.0) | 100 |
| exm-rs3851179 | 8.8E-05 | 0.6 (0.5–0.8) | 99 |
| exm367051 | 9.4E-05 | 0.3 (0.1–0.5) | 73 |
| exm849115 | 1.1E-04 | 2.9 (1.7–5.0) | 100 |
| exm849188 | 1.8E-04 | 2.8 (1.6–4.7) | 100 |

**Table 9.1** The top 11 most significant markers from logistic regression analysis with sEOAD. Markers highlighted in blue are in linkage disequilibrium. The name of the marker (Name) is given followed by details of the test, including the p-value (p-value), odds ratio with 95% confidence interval in brackets (OR) and power to detect association at the estimated OR and MAF, and alpha level of 0.05. Key: p.value, probability value; OR, odds ratio

| Marker | MAF | VEP Variant | VEP Varaint Type | VEP SIFT:Polyphen | VEP Gene | ANNOVAR Gene |
|---|---|---|---|---|---|---|
| exm2264899 | 0.38 (0.30) | rs1298954 | Intronic | - | PDZK1 (I) | PDZK1 (I)<br>NBPF10<br>NBPF20 |
| exm849171 | 0.08 (0.03) | rs33965491 | Missense | T(0.59):B(0) | BLOC1S2 (I) | BLOC1S2 (I) |
| exm849207 | 0.08 (0.03) | rs6584356 | Missense | D(0.03):B(0.01) | PKD2L1 (D) | PKD2L1 (D) |
| exm2269839 | 0.50 (0.41) | rs7678085 | 3' UTR | - | INTU (I) | INTU (I)<br>SLC25A31 (I) |
| exm-rs12093699 | 0.25 (0.32) | rs12093699 | Intergenic | - | - | APCS (D)<br>CRP (D) |
| exm-rs10242744 | 0.16 (0.11) | rs10242744 | Intronic | - | THSD7A (I) | THSD7A (I) |
| exm928279 | 0.36 (0.28) | rs13817 | Missense | T(0.41):PRD(0.97) | MUS81 (I) | MUS81 (I) |
| exm-rs3851179 | 0.30 (0.41) | rs3851179 | Downstream | - | RNU6-560P (U) | PICALM (D)<br>EED (U) |
| exm367051 | 0.02 (0.05) | rs12632177 | Missense | T(1.00):B(0.03) | MCF2L2 (I) | MCF2L2 (I) |
| exm849115 | 0.07 (0.03) | rs7073610 | Missense | T(0.32):B(0.11) | CWF19L1 (D) | CWF19L1 (D) |
| exm849188 | 0.07 (0.03) | rs12782963 | Missense | TLC(0.05):POD(0.91) | PKD2L1 (D) | PKD2L1 (D) |

**Table 9.2** The top 11 most significant markers from logistic regression analysis with sEOAD. The name of the marker (Marker) is given followed by the minor allele frequency in sEOAD and controls in brackets (MAF). The final columns contain details about the markers as ascertained by the VEP and ANNOVAR annotation, including the VEP variant reference ID (VEP Variant), type of variant from VEP (VEP Variant Type), SIFT and Polyphen classification with the score in brackets from VEP (VEP SIFT:Polyphen), the VEP gene with the relation from VarElect in brackets (VEP Gene) and the ANNOVAR gene with the relation from VarElect in brackets (ANNOVAR Gene). Genes highlighted in red were the result of incorrect annotation, due to an error in data from UCSC's Golden Path. Key: I, indirect; D, direct; U, unrelated; rsID, reference single nucleotide identification; MAF, minor allele frequency; P, probability value; OR, odds ratio; VEP, variant effect predictor; UTR, untranslated region; T, tolerated; TLC, tolerated low confidence; D, deleterious; B, benign; PRD, probably damaging; POD, possible damaging.

The marker with the most significant score (exm2264899) genotypes a variant located within three *PDZK1* coding transcripts (ENST00000417171, ENST00000451928 and ENST00000443667) between exons one and two but nowhere near the exon intron boundary, and does not overlap any known regulatory regions. PDZK1 (PDZ Domain Containing 1) is a scaffolding protein with a PDZ-domain which also goes by the name of NHERF3 ($Na^+/H^+$ Exchanger Regulatory Factor 3), the protein is involved in cholesterol homeostasis. PDZK1 regulates SCARB1 (Scavenger Receptor Class B, Member 1) which is a receptor mediating transfer of cholesterol to and from high density lipoproteins (HDL).[259] PDZK1 also regulates perivascular macrophages and modifies amyloid pathology in an AD mouse model.[260] The variant highlighted in this gene is a SNP that has been mentioned in two studies as part of a panel to predict serum uric acid levels.[261,262] Large variations that overlap this coordinate have been seen in several studies using cohorts of the general population, developmental delay[263] or cancer.[264,265] VarElect highlights proteins of interest that are linked to PDZK1 via interaction, for instance one protein that VerElect highlights is ABCA1 which as mentioned in **Section 1.4.3** is involved in cholesterol transport. Other proteins highlighted include CREB1, GFAP and GSR; CREB1 is a leucine zipper transcription factor that is impaired in AD brain,[266] GFAP is a glial fibrillary acidic protein whose levels significantly correlate with protein deposits in AD brain[267] and GSR is an antioxidant whose activity is significant decreased in AD brain.[268]

Four markers overlapping three genes are in LD (exm849171, exm849207, exm849115 and exm849188); SNAP confirmed the former two are in complete LD with each other ($r^2 = 1$) and the latter two variants are in complete LD with each other ($r^2 = 1$), each of the first two variants are in incomplete LD with each of the last two variants ($r^2 = 0.87$). The remaining nine markers were not genotyped in the 1000 Genomes Pilot 1 dataset. The four markers genotype variants located in three genes *CWF19L1* (n=1), *BLOC1S2* (n=1) and *PKD2L1* (n=2). Notably, one variant (rs33965491) overlaps the start of several *BLOC1S2* transcripts, this region is also predicted to be a regulatory feature; given its proximity to the start of *BLOC1S2* it most likely regulates transcription of *BLOC1S2* (**Figure 9.1**).

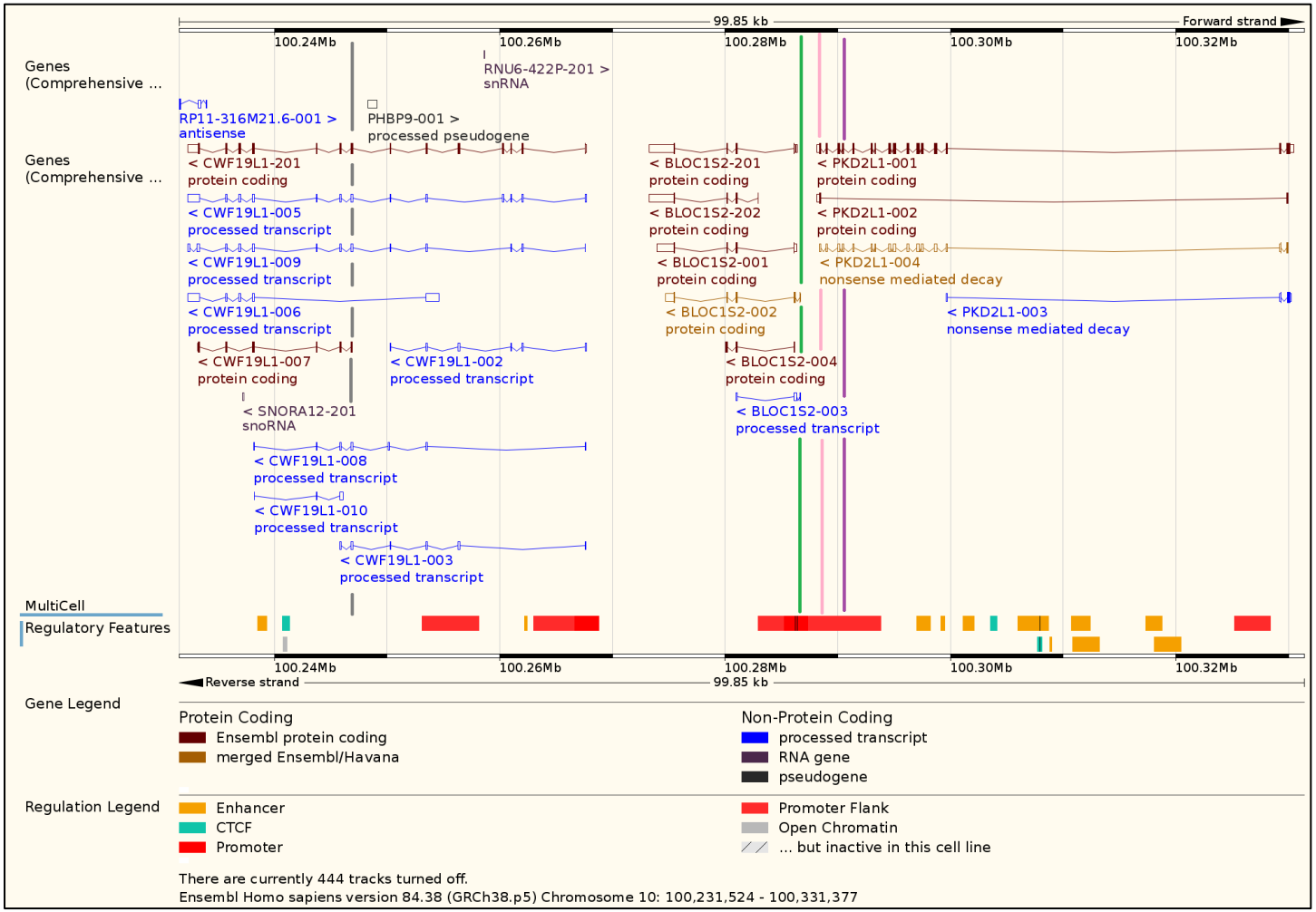



**Figure 9.1** Image illustrating the features overlapping four markers associated with sEOAD using logistic regression. Chromosome 10 is depicted horizontally at the top and bottom of the image in matching 10 kb black and white blocks, the location of the marker is marked using a pink vertical line. Between the chromosome depictions are chromosome features. Transcript features are represented using blocks (exons) connected by lines (introns) with the name of each feature written underneath and an arrow signifying direction. Regulatory features are depicted using coloured blocks below the transcripts. Each feature is colour coded with the key given at the bottom of the figure under Gene Legend and Regulatory Legend. Figure was taken from Ensembl GRCh38.p5. The vertical coloured lines mark the location of the variants genotyped by the four markers; rs7073610 (grey), rs33965491 (dark green), rs12782963 (pink) and rs6584356 (purple)

*BLOC1S2* (Biogenesis of Lysosomal Organelles Complex-1, Subunit 2) is involved in intracellular trafficking. BLOC1S2 is a subunit of the BLOC1 complex which is involved in forming recycling endosomes.[114] Notably, Schizophrenia is associated with variants in *DTNBP1* (dysbindin) which forms part of the BLOC1 complex.[112-115] VerElect highlights two proteins of interest that interact with BLOC1S2; CALB1 and CALB2 are part of a family of proteins called calmodulin that bind calcium (calbindin) and loss of calbindin immunoreactivity is seen in AD.[269]

*CWF19L1* (CWF19-Like 1, Cell Cycle Control (S. Pombe)) is involved in regulating cell cycle. Interestingly, a homozygous variant in intron 9 of *CWF19L1* causes exon skipping and loss of protein function,[270] the variant was seen in a family suffering with Spinocerebellar ataxia autosomal recessive 17 (SCAR17) which presents with dementia, notably this marker in this gene does not genotype the same variant.

*PKD2L1* (Polycystic Kidney Disease 2-Like 1) forms a calcium channel subunit in cilia that regulates sonic hedgehog (SHH) signalling.[271]. The locus has been previously associated with AD[272] and the gene functions in some neurons to trigger action potentials in response to extracellular acidity.[273]

One of the top markers is intergenic, located between *APCS* (Amyloid P Component, Serum) and *CRP* (C-Reactive Protein, Pentraxin-Related), both proteins are classic short pentraxins produced in the liver in response to the interleukin-6 (IL-6)[274] which is a cytokine involved in the innate immune response. CRP is considered a biomarker for chronic inflammation[275] and a high level of high sensitivity CRP (hsCRP) is a biomarker for Behçet's disease (BD) which can present with dementia.[276] APCS binds amyloid fibrils in a calcium dependent manner,[277] additionally, serum levels of APCS are associated with AD[278] and correlate with cognitive decline.[279]

Little is known about *MCF2L2* (MCF.2 Cell Line Derived Transforming Sequence-Like 2), however it comes from a family of proteins called Rho family-specific guanine nucleotide exchange factor. One member of the family in particular mediates cell signalling and cytoskeletal remodelling,[280] like PDZK1 it also contains a PDZ domain.[280] VarElect highlights several proteins that interact with MCF2L2; these include LRRK2, NOTCH1, NOTCH3 and SIRT1. Variants in LRRK2 and NOTCH3 cause diseases that present with dementia; PD[281] and CADASIL (Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy).[282] NOTCH1 is a substrate of γ-secretase[283] and levels are increased in AD brain.[284] SIRT1 is from a family of proteins called sirtuins that act in response to dietary stress,[285] levels of SIRT1 are significantly decreased in the parietal cortex of AD brain.[286]

Two of the top markers are located in non-coding regions of two genes; both genes are involved in the cell cytoskeleton. One marker is located in the 3' untranslated region of *INTU* (Inturned Planar Cell Polarity Protein) but does not overlap any known regulatory region. INTU is essential for the normal orientation of elongating ciliary microtubules. Furthermore, several protein databases predict that it

may have an effect on hedgehog signally due to sequence similarity. The second marker is located in an intron of *THSD7A* (Thrombospondin, Type I, Domain Containing 7A) overlapping a 'Promoter Flanking Region' (ENSR00002071984). Not much is known about THSD7A, it comes from a family of proteins called thrombospondins. THSD7A is thought to be involved in angiogenesis (blood vessel formation) as it is preferentially expressed in placenta endothelial cells and functions in cytoskeleton organisation.[287] Notably its paralog *SPON2* (Spondin 2, Extracellular Matrix Protein) promotes adhesion and outgrowth of hippocampal embryonic neurons.

MUS81 (MUS81 Structure-Specific Endonuclease Subunit) forms an endonuclease complex with EME1 and EME2 to repair DNA interstrand crosslinks (ICLs). Genetic defects in this pathway give rise to Fanconi Anemia. VarElect highlights CDK5 as interacting with MUS81. As mentioned in **Section 1.4.2** CDK5 phosphorylates tau, it functions in post-mitotic neurons via p35.[288]

One marker (exm-rs3851179) is downstream of the small nuclear RNA *RNU6-560P* which is located between *PICALM* and *EED* coding genes; this marker genotypes the same variant identified in GWAS and the susceptibility signal has been attributed to the *PICALM* gene, as such this marker is described in **Section 9.2.2**.

## 9.2.2. Variants from GWA studies

MatchingVariants.pl established that 12 of the 29 variants highlighted in GWA studies were designed on the NeuroX. All 29 GWAS variants and the logistic regression test results for association with sEOAD (**Section 9.2.1**) are listed in **Table 9.2**. Two of the 12 markers (exm-rs3764650 and exm-rs3851179) reached significance when correcting for multiple testing (p.value < 4.16E-3), the variants genotyped by these markers were attributed to the *ABCA7* and *PICALM* genes (as described in **Section 1.3.3**). The odds ratios of these two markers are in the same direction as previous studies, supporting the idea that the *ABCA7* locus provides risk for developing AD and the *PICALM* locus provides protective. *ABCA7* and *PICALM* have both received a substantial amount of research interest since being attributed to the GWAS susceptibility signals. ABCA7 (ATP-Binding Cassette, Sub-Family A (ABC1), Member 7) shares high sequence homology to ABCA1 which as described in **Section 1.4.3** is involved in trafficking and maintaining cholesterol homeostasis, however there is conflicting evidence supporting a similar function for ABCA7. In AD evidence suggests that ABCA7 is trafficked to the cell membrane along with low-density lipoprotein receptor-related protein (LRP)[289] which modulates APP processing and trafficking[290,291]. Additionally ABCA7 could function to facilitate microglia-mediated Aβ clearance given it is predominantly expressed in microglia. PICALM (Phosphatidylinositol Binding Clathrin Assembly Protein) interacts with AP2 (adaptor protein complex 2) and clathrin at cell membranes and mediates vesicle assembly at the neuromuscular junction, its primary function is to modulate trafficking. In AD PICALM is thought to promote transportation of APP-CTF to lysosomes for degradation and ultimately reduce the production of Aβ.[292,293]

| Locus N | Locus | Position (rsID) | GWAS P | GWAS OR | Marker | MAF | P | OR |
|---|---|---|---|---|---|---|---|---|
| 1 | ABCA7 | 19:1046520 (rs3764650) | 8.8E-22 | 1.23 (1.18-1.28) | exm-rs3764650 | 0.13 (0.08) | 1.7E-03 | 1.76 (1.23-2.50) |
| | ABCA7 | 19:1063443 (rs4147929) | 1.1E-15 | 1.15 (1.11-1.19) | - | - | - | - |
| 2 | BIN1 | 02:127894615 (rs744373) | 1.6E-26 | 1.17 (1.13-1.20) | exm-rs744373 | 0.30 (0.28) | 4.6E-01 | 1.09 (0.87-1.36) |
| | BIN1 | 02:127892810 (rs6733839) | 6.9E-44 | 1.22 (1.18-1.25) | - | - | - | - |
| 3 | CD2AP | 06:47453378 (rs9349407) | 2.8E-09 | 1.18 (1.08-1.16) | - | - | - | - |
| | CD2AP | 06:47487762 (rs10948363) | 5.2E-11 | 1.10 (1.07-1.13) | - | - | - | - |
| 4 | CD33 | 19:51727962 (rs3865444) | 2.0E-10 | 0.89 (0.86-0.93) | exm-rs3865444 | 0.32 (0.35) | 5.2E-01 | 0.93 (0.74-1.16) |
| | CD33 | 19:51727962 (rs3865444) | 2.0E-10 | 0.89 (0.86-0.93) | exm-rs3865444 | 0.32 (0.35) | 5.2E-01 | 0.93 (0.74-1.16) |
| 5 | CLU | 08:27464519 (rs11136000) | 3.4E-23 | 0.88 (0.86-0.90) | exm-rs11136000 | 0.35 (0.39) | 1.4E-01 | 0.85 (0.68-1.05) |
| | CLU | 08:27467686 (rs9331896) | 2.8E-25 | 0.86 (0.84-0.89) | NeuroX_rs9331896 | 0.36 (0.40) | 1.1E-01 | 0.84 (0.68-1.04) |
| 6 | CR1 | 01:207784968 (rs3818361) | 4.7E-21 | 1.17 (1.14-1.21) | exm-rs3818361 | 0.20 (0.15) | 1.6E-02 | 1.41 (1.07-1.86) |
| | CR1 | 01:207694049 (rs6656401) | 5.7E-24 | 1.18 (1.14-1.22) | exm-rs6656401 | 0.20 (0.15) | 9.8E-03 | 1.44 (1.09-1.91) |
| 7 | EPHA1 | 07:143109139 (rs11767557) | N/A | 0.89 (0.83-0.96) | exm-rs11767557 | 0.19 (0.21) | 7.9E-02 | 0.79 (0.60-1.02) |
| | EPHA1 | 07:143110762 (rs11771145) | 1.1E-13 | 0.90 (0.88-0.93) | exm2266485_ver2 | 0.34 (0.32) | 3.3E-01 | 1.12 (0.90-1.41) |
| 8 | MS4A6A | 11:59939307 (rs610932) | 1.8E-11 | 0.90 (0.88-0.93) | exm-rs610932 | 0.41 (0.43) | 2.2E-01 | 0.87 (0.70-1.09) |
| | MS4A6A | 11:59923508 (rs983392) | 6.1E-16 | 0.90 (0.87-0.92) | - | - | - | - |
| 9 | PICALM | 11:85868640 (rs3851179) | 2.9E-20 | 0.88 (0.86-0.90) | exm-rs3851179 | 0.30 (0.41) | 8.8E-05 | 0.63 (0.51-0.80) |
| | PICALM | 11:85867875 (rs10792832) | 9.3E-26 | 0.87 (0.85-0.89) | - | - | - | - |
| 10 | CASS4 | 20:55018260 (rs7274581) | 2.5E-08 | 0.88 (0.84-0.92) | - | - | - | - |
| 11 | CELF1 | 11:47557871 (rs10838725) | 1.1E-08 | 1.08 (1.05-1.11) | - | - | - | - |
| 12 | DSG2 | 18:29088958 (rs8093731) | 1.0E-04 | 0.73 (0.62-0.86) | - | - | - | - |
| 13 | FERMT2 | 14:53400629 (rs17125944) | 7.9E-09 | 1.14 (1.09-1.19) | - | - | - | - |
| 14 | HLA-DRB1/5 | 06:32578530 (rs9271192) | 2.9E-12 | 1.11 (1.08-1.15) | NeuroX_rs9271192 | 0.31 (0.28) | 7.4E-02 | 1.239 (0.98-1.57) |
| 15 | INPP5D | 02:234068476 (rs35349669) | 3.2E-08 | 1.08 (1.05-1.11) | - | - | - | - |
| 16 | MEF2C | 05:88223420 (rs190982) | 3.2E-08 | 0.93 (0.90-0.95) | - | - | - | - |
| 17 | NME8 | 07:37841534 (rs2718058) | 4.8E-09 | 0.93 (0.90-0.95) | - | - | - | - |
| 18 | PTK2B | 08:27195121 (rs28834970) | 7.4E-14 | 1.10 (1.08-1.13) | - | - | - | - |
| 19 | SLC24A4/RIN3 | 14:92926952 (rs10498633) | 5.5E-9 | 0.91 (0.88-0.94) | - | - | - | - |
| 20 | SORL1 | 11:121435587 (rs11218343) | 9.7E-15 | 0.77 (0.72-0.82) | - | - | - | - |
| 21 | ZCWPW1 | 07:100004446 (rs1476679) | 5.6E-10 | 0.91 (0.89-0.94) | - | - | - | - |

**Table 9.2** Logistic regression results of the GWA variants, including studies both before (blue) and after the meta-analysis[48] (yellow) (from AlzGene). The two loci highlighted green reached significance after correcting for multiple testing. The locus number (Locus N) and gene attributed to the locus (Locus) are given along with the genomic position of the GWAS variant in genome build GRCh37 and variant reference ID given in brackets (Position (rsID)), along with the data from the studies including the p.value (GWAS P) and the odds ratio with 95% confidence interval in brackets (GWAS OR). If it was designed on the NeuroX additional information is given, including the name of the marker (Marker), MAF in cases with controls on brackets (MAF), the test p.value (P) and finally the odds ratio with 95% confidence interval in brackets (OR). Key: rsID, reference single nucleotide polymorphism identification; N/A, not available; MAF, minor allele frequency; OR, odds ratio; P, p.value.

Noticeably, the marker exm-rs3851179 in the *PICALM* locus was also identified in the top 11 most significant markers associated with sEOAD using logistic regression (**Table 9.1**); VEP annotated this variant as residing downstream of the small nuclear RNA (snRNA) called *RNU6-560P* as appose to being upstream of *PICALM*, this is because the RNU6-560P feature is closer to the marker (approximately 4 kb downstream of RNU6-560P compared to approximately 100 kb upstream of *PICALM*) (**Figure 9.3**). Looking at the meta-analysis results for the *PICALM* region (**Figure 9.4**), approximately half of the significant SNPs are located in *PICALM*, however the most significant SNPs appear to be closer to *RNU6-560P*, additionally both features are close together and are on the same side of a recombination peak, thus they are likely to be in a high level of linkage disequilibrium. Given the known functionality of PICALM it would suggest that *PICALM* is the source of the susceptibility signal, but given its closer proximity to *RNU6-560P* and *EED* their involvement cannot be ruled out, furthermore the variant overlaps a regulatory feature that could regulate any of these three features, one paper suggests the primary meta-analysis SNP for this locus (rs10792832) to be an eQTL for *EED* and not *PICALM*.[294] EED comes from a family of proteins called polycombs which maintain epigenetic modifications and results in transcription repression. EED forms a complex with PRC2, EZH2 and SUZ12 to maintain epigenetic modifications,[295] however it may also function alone to activate integrin proteins and stimulate cell signalling; a process exploited by the HIV.[296] Interestingly the PRC2-EZH2-SUZ12-EED complex has central roles in development and cancer,[297] additionally the complex also represses *MYT1* (Myelin Transcription Factor 1).[298] RNU6-560P is a small non-coding RNA; from what is known about non-coding nuclear and nucleolar RNAs, they can be involved in quite diverse roles; including ligation, cleavage and modification of RNA species, methylation of DNA, telomere DNA synthesis and modulating protein function.[299]

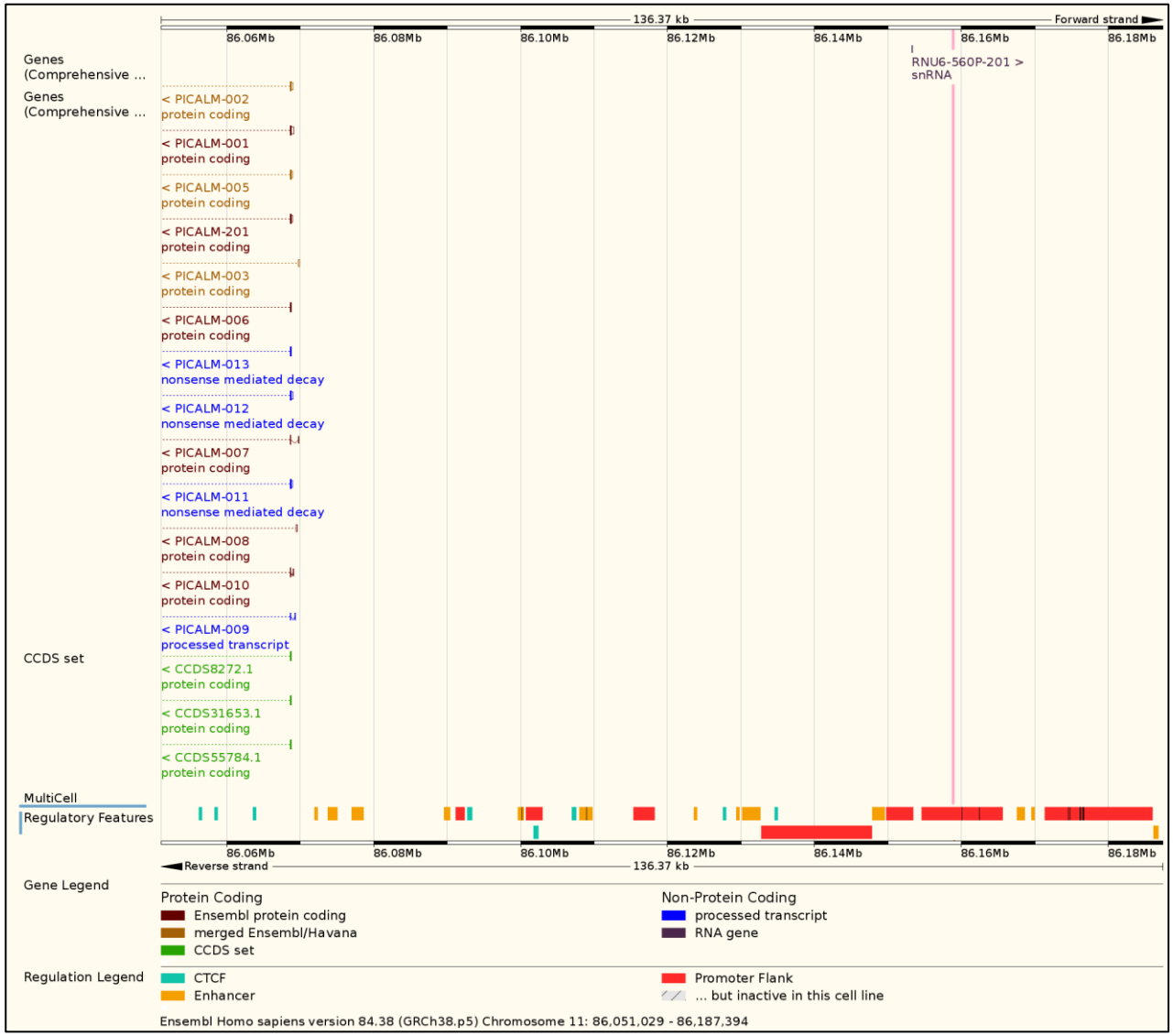



**Figure 9.3** Image illustrating the features overlapping the marker (exm-rs3851179) that genotypes a previously associated susceptibility variant (rs3851179) attributed to *PICALM*. Chromosome 11 is depicted horizontally at the top and bottom of the image in matching 10 kb black and white blocks, the location of the marker is indicated using a pink vertical line. Between the chromosome depictions are chromosome features. Transcript features are represented using blocks (exons) connected by dotted lines (introns), below each transcript is the name of the transcript and an arrow signifying direction. Regulatory features are represented using coloured blocks below the transcripts. Each feature is colour coded with the key given at the bottom of the figure under Gene Legend and Regulatory Legend. Figure was taken from Ensembl GRCh38.p5.

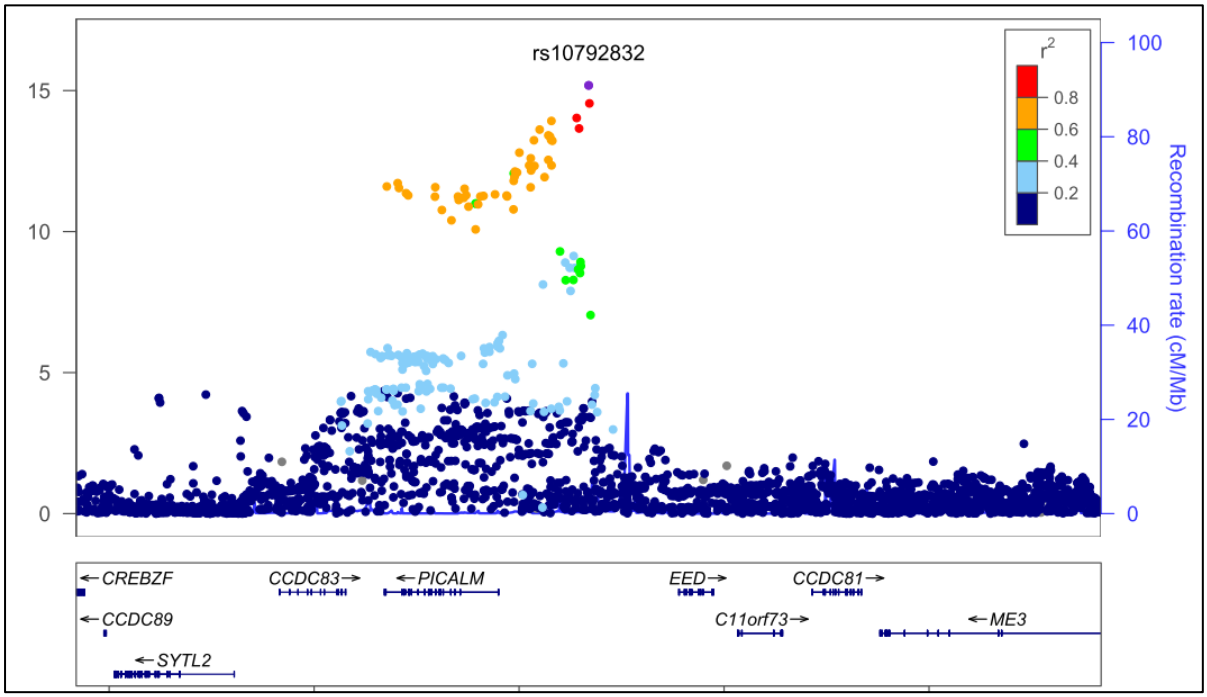



**Figure 9.4** This image is the **s**upplementary figure 4B taken from the meta-analysis study[48] showing the SNPs used in the study of the rs10792832 locus. This figure shows that the SNPs in linkage disequilibrium (dots in red/orange) with the primary SNP (dot in purple) have a higher than average significance.

### 9.2.3. All genes

Adjusting for documented gender in SKAT-O burden analysis gave the same top six genes as adjusting for calculated gender, the only difference was the order of significance; **Table 9.3** lists the top six genes when adjusting for calculated gender and shows that none reached significant association with sEOAD after correcting for multiple testing of all genes (p-value < 2.45E-6). **Table 9.4** lists the details of the top genes.

| ANNOVAR Gene | Description | p.value | N |
|---|---|---|---|
| DCLK3 | Double Cortin Like Kinase 3 | 1.5E-05 | 17 (09) |
| BLOC1S2 | Biogenesis of Lysosomal Organelles Complex-1, Subunit 2 | 2.6E-05 | 01 (01) |
| MINA | MYC Induced Nuclear Antigen | 3.6E-05 | 16 (04) |
| CMPK1 | Cystidine Monophosphate (UMP-CMP) Kinase 1, Cytosolic | 6.1E-05 | 05 (03) |
| IRAK3 | Interleukin-1 Receptor-Associated Kinase 3 | 7.1E-05 | 32 (09) |
| FOXD2 | Forkhead Box D2 | 8.1E-05 | 04 (04) |

**Table 9.3** List of six genes with the highest p.value when testing for association to sEOAD using the gene burden test SKAT-O. The gene ID generated by ANNOVAR is given in the first column (ANNOVAR Gene) followed by the definition of each gene in the second column (Description). Details of the test include p.value (p.value) along with the number of markers in each gene set followed by the number used in the test in brackets (N). Key: p.value, probability value; N, number.

| ANNOVAR Gene | Marker | MAF | VEP Variant | VEP Variant Type | VEP SIFT:Polyphen | VEP Gene |
|---|---|---|---|---|---|---|
| DCLK3 (I) | exm298345 | 0.0025 (0) | rs200497284 | Missense | D(0):POD(0.903) | DCLK3 (I) |
| | exm298346 | 0.03 (0.01) | rs35704209 | Missense | D(0.03):B(0.003) | |
| | exm298376 | 0.01 (0.01) | rs202070330 | Missense | T(0.29):B(0.146) | |
| | exm298404 | 0.0012 (0) | rs201773927 | Missense | T(0.11):PRD(0.994) | |
| | exm298409 | 0.0012 (0) | rs199872230 | Missense | T(0.16):POD(0.528) | |
| | exm298417 | 0.02 (0.01) | rs56070233 | Missense | T(0.1):POD(0.661) | |
| | NeuroX_rs9834970 | 0.50 (0.50) | rs9834970 | Intergenic | - | - |
| | NeuroX_rs1553656 | 0.46 (0.47) | rs1553656 | Intergenic | - | - |
| | NeuroX_rs4624519 | 0.35 (0.37) | rs4624519 | Intergenic | - | - |
| BLOC1S2 (I) | exm849171 | 0.08 (0.04) | rs33965491 | Missense | T(0.59):B(0) | BLOC1S2 (I) |
| MINA (I) | exm332285 | 0.52 (0.48) | rs2172257 | Missense | T(0.5):B(0) | MINA (I) |
| | exm332318 | 0.02 (0.02) | rs55907650 | Missense, Splice | D(0.03):PRD(0.995) | |
| | exm332325 | 0.02 (0.01) | rs56183666 | Missense | D(0):PRD(1) | |
| | exm332346 | 0.03 (0.01) | rs41265444 | Missense | D(0):PRD(1) | |
| CMPK1 (D) | exm2265006_ver2 | 0.01 (0.01) | rs7529388 | Intronic | - - | CMPK1 (D) |
| | exm57721 | 0.06 (0.06) | rs35687416 | Missense | D(0):POD(0.922) | |
| | exm57726 | 0.014 (0) | rs72553947 | Missense | T(0.35):B(0.001) | |
| IRAK3 (I) | exm1019681 | 0.02 (0.02) | rs35239505 | Missense | T(0.38):B(0.202) | IRAK3 (I) |
| | exm1019689 | 0.0012 (0) | rs137909830 | StopGained | - | |
| | exm1019698 | 0.0049 (0) | rs138559915 | Missense | D(0.03):PRD(0.972) | |
| | exm1019700 | 0.03 (0.02) | rs56001649 | Synonymous, Splice | - | |
| | exm1019702 | 0.08 (0.07) | rs1152888 | Missense, Splice | T(0.5):B(0.001) | |
| | exm1019713 | 0.0025 (0) | rs139342884 | Missense | T(0.08):POD(0.832) | |
| | exm2260104 | 0.19 (0.18) | rs1152915 | Intronic | - | |
| | exm1019738 | 0.01 (0.01) | rs138984535 | Missense | T(0.09):POD(0.656) | |
| | exm1019751_ver3 | 0.0025 (0) | rs146885838 | Missense | T(0.39):B(0.322) | |
| FOXD2 (I) | exm57776 | 0.05 (0.05) | rs78469326 | Missense | TLC(0.6):B(0.034) | FOXD2 (I) |
| | NeuroX_rs527430 | 0.0135 (0) | rs527430 | Intergenic | - | - |
| | exm2264833 | 0.48 (0.46) | rs1605672 | Intergenic | - | - |
| | exm-rs517384 | 0.10 (0.12) | rs517384 | Upstream | - | RP4-666O22.3 |

**Table 9.4** List of the top six genes from SKAT-O gene burden when testing for association with sEOAD. The genes highlighted in blue are close to one another in the human genome and might have some degree of linkage disequilibrium but there is not sufficient data available to permit this to be determined. The name of the gene as annotated by ANNOVAR is given with the relation of the gene from VarElect in brackets (ANNOVAR Gene), this is followed by the name of the markers used in the test (Markers) and the minor allele frequency in sEOAD and controls in brackets (MAF). The final columns contain details about the markers as ascertained by the VEP annotation, including the variant reference ID (VEP Variant), type of variant (VEP Variant Type), SIFT and Polyphen classification with the score in brackets (VEP SIFT:Polyphen) and the gene annotated by VEP and its relationship status provided by VarElect in brackets (VEP Gene). Key: I, indirect; D, direct; U, unrelated; rsID, reference single nucleotide identification; MAF, minor allele frequency; P, probability value; OR, odds ratio; VEP, variant effect predictor; UTR, untranslated region; T, tolerated; TLC, tolerated low confidence; D, deleterious; B, benign; PRD, probably damaging; POD, possible damaging.

Genes *CMPK1* and *FOXD2* are located next to each other in the human genome; SNAP was unable to establish if there was any level of linkage disequilibrium for the variants in these two genes as they were not genotyped in the 1000 Genomes Pilot 1 data set. The gene with the most significant score was *DCLK3* (1.5E-05).

DCLK3 (Double Cortin Like Kinase 3) is from a family of at least 11 paralogues in humans[300] which have one or two double cortin domain (DCX-domain), they can also contain a 'serine/threonine/proline' domain (SP-domain) which enables protein-protein interaction and/or a kinase domain that phosphorylates serine/threonine residues (Kinase-domain). The DCX domain has an ability to bind microtubules,[301] however the DCX-domain has a ubiquitin like fold[302] which is different to that if other microtubule associated proteins (MAPs), not surprisingly the domain binds at sites different to other MAPs, which are at the fenestrations localized between protofilaments.[303] Members of the family include DCX, DCLK1, DCLK2 and DCLK3. It is interesting to note that mutations in the DCX-domain of *DCX* result in Lissencephaly or subcortical band heterotopia (SBH) which are neuronal migration disorders. Furthermore, VarElect highlights CDK5 as interacting with DCLK3, as mentioned in **Section 1.4.2** CDK5 phosphorylates tau, it functions in post-mitotic neurons via p35.[288] The markers used in the DCLK3 test mostly genotype missense variants, however several of the markers are intergenic, all of the missense markers are present at higher frequency in sEOAD compared to controls.

BLOC1S2 (Biogenesis of Lysosomal Organelles Complex-1, Subunit 2) was also highlighted in logistic regression analysis of all markers; the marker exm849171 was the second most significant marker identified. *BLOC1S2* is described in more detail in **Section 9.2.1**.

MINA (MYC Induced Nuclear Antigen) is involved in cell cycle control. MINA acts as a histone lysine demethylase and a ribosomal histidine hydroxylase to regulate expression of the oncogene *MYC*, furthermore, phospho-MYC is increased in dystrophic neurites and neurons with neurofibrillary tangles in AD.[304] VarElect presents evidence that MINA interacts with UBC (ubiquitin) which is present in NFTs and senile plaques of AD brain.[305]

IRAK3 (Interleukin-1 Receptor-Associated Kinase 3) is part of the IL1R associated kinase family which is involved in the innate immune system. IRAK3 which is also known by IRAKM is a negative regulator of toll-like receptors (TLR). IRAK3 interacts with the myddome (IRAK1, IRAK2, IRAK4 and MyD88),[306] and exerts an inhibitory effect on translational control of cytokines and chemokines and also results in production of molecules that inhibit inflammatory pathways.[307] As described in **Section 1.4.7** neuroinflammation is a pathway closely linked with AD, furthermore interacting proteins IRAK1 and IRAK2 are differentially expressed in AD astrocytes.[308]

CMPK1 (Cytidine Monophosphate (UMP-CMP) Kinase 1, Cytosolic) and FOXD2 (Forkhead Box D2) are located close to one another in the human genome. CMPK1 is an enzyme that catalyses the transfer

of phosphate group to ribonucleotides to form diphosphate nucleotides. It was directly linked to the VarElect search terms due to proteome analysis of transgenic mice brain that found decreased levels of the protein as a result of aging and amyloid deposition.[309] FOXD2 (Forkhead Box D2) belongs to a family of transcription factor proteins that contain a forkhead domain. Additionally, VarElect highlights CREB1, CDK5, MARK1, MARK4 and PVALB as interacting with FOXD2. As mentioned previously CREB1 is a leucine zipper transcription factor that is impaired in AD brain,[266] and CDK5 phosphorylates tau.[288] Evidence suggests MARK4 and possibly MARK1 also phosphorylate tau.[310] PVALB is expressed in a subset of interneurons which appear to be particularly vulnerable in AD brains.[311] All the variants used in the test for CMPK1 are intragenic and one in particular appears to be driving the association of CMPK1, whereas a portion of those used in the test for FOXD2 are intergenic (not on the side of CMPK1) and it is the intergenic variants that appear to be driving the association of FOXD2. From this data it is not possible to know if one or both of these coding genes are a source or susceptibility, however given that CMPK1 is more associated, contains only intragenic markers and has a direct link to the VarElect search terms, CMPK1 seems a more likely candidate.

### 9.2.4. All pathways

Adjusting for documented gender in PLINK set-test analysis gave a slightly different list of top 6 pathways compared to when adjusting for calculated gender; **Table 9.4** lists the 6 pathways with the most significant p.value when adjusting for calculated gender. None of the 6 pathways reached significant association with sEOAD when adjusting for multiple testing (p.value < 2.65E-5), **Table 9.5** lists the details of the top pathways.

| Reactome Pathway | Description | p.value | N |
|---|---|---|---|
| 5619061 | Defective SLC33A1 causes spastic Paraplegia 42 (SPG42) | 0.5E-03 | 9 (1) |
| 3214841 | PKMTs methylate histone lysines | 1.5E-03 | 735 (2) |
| 432722 | Golgi Associated Vesicle Biogenesis | 2.1E-03 | 818 (2) |
| 5683371 | Defective ABCB6 causes isolated colobomatous microphthalmia 7 (MCOPCB7) | 3.0E-03 | 35 (1) |
| 1306955 | GRB7 events in ERBB2 signalling | 4.5E-03 | 134 (1) |
| 6783310 | Fanconi Anemia Pathway | 5.1E-03 | 737 (3) |

**Table 9.4** The top six most significant pathways from PLINK set-test analysis with sEOAD. The Reactome pathway ID number is given in the first column (Reactome Pathway) followed by the pathway description (Description), the test p.value (p.value) and the number of markers in each pathway followed by the number used in the test in brackets (N).

| Reactome Pathway | Marker | MAF | VEP rsID | VEP Type | VEP SIFT:Polyphen | VEP Gene | ANNOVAR Gene |
|---|---|---|---|---|---|---|---|
| 5619061 | exm360571 | 0.07 (0.03) | rs3804769 | Missense | T(1):B(0) | SLC33A1 (D) | SLC33A1 (D) |
| 3214841 | exm-rs486416_ver3 | 0.38 (0.34) | rs486416 | Intronic | - | EHMT2 (I) | EHMT2 (I) |
| | exm-rs3851179 | 0.30 (0.41) | rs3851179 | Downstream | - | RNU6-560P (U) | EED (I)<br>PICALM (D) |
| 432722 | exm592399 | 0.01 (0.003) | rs76130099 | Missense | T(0.07):B(0) | IGF2R (D) | IGF2R (D) |
| | exm-rs3851179 | 0.30 (0.41) | rs3851179 | Downstream | - | RNU6-560P (U) | PICALM (D)<br>EED (I) |
| 5683371 | exm267638 | 0.004 (0.01) | rs147445258 | Missense | T(0.09):PRD(0.979) | ABCB6 (D) | ABCB6 (D) |
| 1306955 | NeuroX_rs2439302 | 0.43 (0.50) | rs2439302 | Intronic | - | NRG1 (D) | NRG1 (D) |
| 6783310 | exm928279 | 0.36 (0.28) | rs13817 | Missense | T(0.41):PRD(0.966) | MUS81 (I) | MUS81 (I) |
| | exm356169 | 0.02 (0.01) | rs2229033 | Missense | TLC(0.2):B(0.177) | ATR (D) | ATR (D) |
| | exm-rs4932217 | 0.44 (0.41) | rs4932217 | Upstream | - | LINC00925 (NR) | MIR9-3HG (U)<br>POLG (D) |

**Table 9.4** List of the top six pathways from PLINK set-test when testing for association with sEOAD. The name of ID number of the pathway (Reactome Pathway) is given followed by the name of the markers used in the test (Marker) and the minor allele frequency in sEOAD and controls in brackets (MAF). The final columns contain details about the markers as ascertained by the VEP and ANNOVAR annotation, including the VEP variant reference ID (VEP rsID), type of variant from VEP (VEP Type), SIFT and Polyphen classification with the score in brackets from VEP (VEP SIFT:Polyphen), the VEP gene with the relation from VarElect in brackets (VEP Gene) and the ANNOVAR gene with the relation from VarElect in brackets (ANNOVAR Gene). Genes highlighted in red are not part of the pathway. Key: I, indirect; D, direct; U, unrelated; rsID, reference single nucleotide identification; MAF, minor allele frequency; P, probability value; OR, odds ratio; VEP, variant effect predictor; UTR, untranslated region; T, tolerated; TLC, tolerated low confidence; D, deleterious; B, benign; PRD, probably damaging; POD, possible damaging.

The pathway 'Defective SLC33A1 causes spastic Paraplegia 42 (SPG42)' only contains one gene; *SLC33A1* (Solute Carrier Family 33 (Acetyl-CoA Transporter), Member 1) which is also known as AT-1. When SLC33A1 is defective individuals present with Spastic Paraplegia (SPG42)[312] and congenital cataracts, hearing loss, and neurodegeneration (CCHLND).[313] In one study SPG42 patients were found to have upregulated BMP (bone morphogenetic proteins) cell signalling, function work in animal models suggest SLC33A1 plays a role in negatively regulating BMP cell signalling.[314] At a molecular level SLC33A1 encodes a transmembrane protein that transports acetyl coenzyme A (acetyl-CoA) from the cytosol to the lumen of the Golgi apparatus (GA) and the endoplasmic reticulum (ER), subsequently the acetyl-CoA is transferred to gangliosides and glycoproteins via acetyl transferase proteins NAT8B and NAT8 (also known as ATase1 and ATase2). ATase1 and ATase2 regulate the activity of BACE1 by transferring acetyl-CoA to BACE1 at seven lysine residues and affecting its ability to advance to the Golgi apparatus.[315-317] Inhibition of ATase1 and ATase2 in AD mouse models reduced the levels of Aβ and phosphorylated tau, improving synaptic plasticity and increasing lifespan.[318]

The pathway 'PKMTs methylate histone lysines' contains multiple genes. Two genes have been highlighted using PLINK set-test; *EED* (Embryonic Ectoderm Development) and *EHMT2* (Euchromatic Histone-Lysine N-Methyltransferase 2). The marker upstream of EED (exm-rs3851179) is the same marker identified in logistic regression of all variants (**Section 9.2.1**) and genotypes one of the variants highlighted in previous GWAS (**Section 9.2.2**). EED and EHMT2 are involved in modifying histones to repress expression; EHMT2 mono-methylates and di-methylates Lys-9 of histone H3 (H3K9me and H3K9me2, respectively) in euchromatin, EED is described in detail in **Section 9.2.2**. Epigenetic modifications are sub divided into those that modify histones and those that modify DNA, both function to regulate gene expression along with non-coding RNAs. Epigenetic modifications and gene expression have long been studied; however the role of histones has only recently been appreciated. There is no concrete evidence supporting that epigenetic regulation is involved in the aetiology of AD, however histone acetylation in the brain is important in memory function.[319]

The pathway 'Golgi Associated Vesicle Biogenesis' contains multiple genes. Two genes have been highlighted using PLINK set-test; *PICALM* (Phosphatidylinositol Binding Clathrin Assembly Protein) and *IGF2R* (Insulin-Like Growth Factor 2 Receptor). The marker upstream of *PICALM* (exm-rs3851179) is the same marker identified in logistic regression of all variants (**Section 9.2.1**) and genotypes one of the variants highlighted in previous GWAS (**Section 9.2.2**), this gene has been described in **Section 9.2.2**. IGF2R is a transmembrane protein receptor located in the cell membrane and in the lumen of the trans-Golgi network. On the cell membrane the receptor binds IGF2 (insulin growth factor 2) which accumulates and promotes growth, whereas in the lumen of the trans-Golgi network it binds to proteins (such as enzymes) containing M6P (mannose 6-phosphate). Activated

IGF2R accumulates and is internalised in clathrin coated vesicles. In AD IGF2 is decreased in the hippocampus of AD brain. IGF2R/IGF2 has shown to play a critical role in memory formation in the rat hippocampus,[320] similarly IGF2R/IGF2 signalling is important in the formation and maturation of synapses in mice hippocampus.[321] Notably, the mouse *IGF2R* orthologue is imprinted via histone modifications,[322] this could be linked to the previously mentioned pathway (PKMTs methylate histone lysines) which involves epigenetic modifications. In adult human *IGF2R* imprinting is polymorphic and rare,[323] however a human homolog of the imprinted mouse *Air* non-coding RNA which controls mouse *IGF2R* imprinting expression has been identified.[324]

The pathway 'Defective ABCB6 causes isolated colobomatous microphthalmia 7 (MCOPCB7)' contains just one gene; *ABCB6* (ATP-Binding Cassette, Sub-Family B (MDR/TAP), Member 6). When ABCB6 is defective individuals present with colobomatous microphthalmia-7 (MCOPCB7) a developmental defect of the eye coloboma[325]. *ABCB6* is also a candidate for neonatal metabolic syndrome, a disorder of mitochondrial function. ABCB6 is part of the same family of proteins as ABCA7; this gene was identified in GWAS and is described in **Section 9.2.2**. The ABC family of proteins are ubiquitous ATP-dependent membrane bound proteins that are found in plants, fungi, yeast and animals. Dimeric ABCB6 is located in the outer membrane of mitochondria where it is required for uptake of porphyrin and haem synthesis;[326] it could possibly function to prevent iron accumulation and thus oxidative stress in mitochondria.

The pathway 'GRB7 events in ERBB2 signalling' contains multiple genes. One gene has been highlighted by PLINK set-test; *NRG1* (Neuregulin 1). NRG1 is a well characterised ligand for the ErbB transmembrane receptors. It functions in cell signalling, growth and development. SNPs spanning the 3' or the 5' end of NGR1 (up to and including intron 2) are associated with schizophrenia;[327-329] and/or AD with psychosis (ADP).[330] Interestingly the marker highlighted here (NeuroX_rs2439302) genotypes a variant (rs2439302) located in intron 1 of most of the *NRG1* transcripts. It is thought that SNPs in the 5′ region of *NRG1* regulate its expression.[331] Most NRG1 isoforms contain a transmembrane domain which is processed by γ-secretase and NRG1 is a major substrate of BACE1.[332] Furthermore, NGR1 is upregulated in neuritic plaques.[333]

The 'Fanconi Anemia Pathway' involves multiple genes, when this pathway is defective it results in genetic instability and individuals present with aplastic anaemia, skeletal defects and leukaemia. The genes in this pathway involve DNA repair[334] which as described in **Section 1.4.6** is tightly coupled to the cell cycle. Three genes have been highlighted by PLINKs set-test; *MUS81* (MUS81 Structure-Specific Endonuclease Subunit), *POLG* (Polymerase (DNA Directed), Gamma) and *ATR* (ATR Serine/Threonine Kinase). The marker located in *MUS81* (exm928279) is the same marker identified in logistic regression of all variants (**Section 9.2.1**); this gene has been described in **Section 9.2.1**. The marker located upstream of *POLG* (exm-rs4932217) is also upstream of the non-coding RNA *MIR9-*

*3HG* (also known as LINC00925) which is not part of this pathway. POLG is the catalytic subunit of the mitochondrial DNA polymerase, interestingly, individuals with defects in this gene result in several disorders that affect the eye, muscles, mitochondria and brain, these disorders include progressive external ophthalmoplegia with mitochondrial DNA deletions 1 (PEOA1),[335] sensory ataxic neuropathy dysarthria and ophthalmoparesis (SANDO),[336] and mitochondrial DNA depletion syndrome 4A and 4B (MTDPS4A and MTDPS4B respectively).[337,338] ATR is a kinase that phosphorylates proteins including CHK1, RAD17, RAD9 and BRCA1. Individuals with causative variants in ATR present with Seckel syndrome-1 (SCKL1)[339]; a syndrome primarily involving deformity of the head and mental retardation.

# 10. Overall Discussion

## 10.1. The NeuroX

The NeuroX has the potential to genotype causative variants in *APP*/*PSEN1*/*PSEN2* and thus be used as a screening tool to identify ADAD individuals, unfortunately the NeuroX v1 genotypes only 35% of the causative variants documented in online resources. Ideally a screening tool for ADAD will have the capacity to genotype all causative variants in *APP*/*PSEN1*/*PSEN2* and have the capability of identifying duplications of *APP*, hopefully the next version of the chip will go further to achieving this goal.

The design of the probes on the NeuroX uses both the Illumina Infinium I and II designs. There have been several instances during this work where markers with Infinium II probe design have actually genotyped an alternative allele not described in the NeuroX manifesto, simply because the design has the capability to genotype two alternative alleles. This issue highlights the need to confirm NeuroX genotypes with Sanger sequencing if they are the result of Infinium II probes.

## 10.2. Genetics of sEOAD

Several sEOAD samples were identified as harbouring a known causative variant in *APP*, thus these samples were in fact ADAD and were subsequently removed from the final NeuroX dataset before commencing analyses. This highlights the need to screen sEOAD cohorts for all known causative variants. Additionally, several sEOAD samples were identified that harboured causative variants in *PARK2* which are thought to contribute to Parkinson's disease (PD), this suggests that these variants may elicit pleomorphic risk and supports the idea that AD and PD may have an overlapping genetic aetiology.

Many sEOAD samples were identified that harboured a predicted pathogenic variant within one of 112 neurologic related genes; some of these variants were located in LOAD GWAS genes. Harbouring multiple predicted pathogenic variants may elicit susceptibility to sEOAD and a set test (for example SKAT-O) with all predicted pathogenic variants placed in one set could be used to establish this.

Analysis of the NeuroX data generated from a sEOAD cohort identified many candidate variants, genes and pathways; these data could provide the basis for future research, however the genotypes of the variants would need to be confirmed with Sanger sequencing or high throughput genotyping such as TaqMan or KASP. These variants highlighted could contribute to the susceptibility of sEOAD, and together they may even have sufficient penetrance to cause early onset. A phenotype-genotype correlation analysis using the NeuroX genotype data would establish how well these variants or all variants on the NeuroX correlate with disease or even age at onset.

As with any association test there is no guarantee that an association seen is real as it could be the result of random chance, this is especially true if like in this instance no marker set reached significance after correcting for multiple testing. Replicating associations in an independent cohort makes an association less likely to be the result of random chance. As power was not able to be calculated in the gene and pathway analyses, it could be that the top associations would have reached significance had the power been sufficient. Follow up work would be to ascertain if power to detect significant association in the SKAT-O and PLINK set-test analyses can be calculated, but given the complex nature of the algorithms it may not be feasible.

Presented in this thesis are just the top results. Notably most of the genes highlighted are related to metabolic pathways already linked to the aetiology of AD and described in **Section 1**; these include the immune response, cytoskeleton, intracellular trafficking, cholesterol homeostasis, oxidative stress (iron homeostasis), cell cycle and DNA repair. Additional pathways identified which are not described in **Section 1** include gene expression, cell signalling and protein modification. **Figure 10.1** depicts the top results in the form of a cloud diagram to illustrate how each gene is linked to major themes.

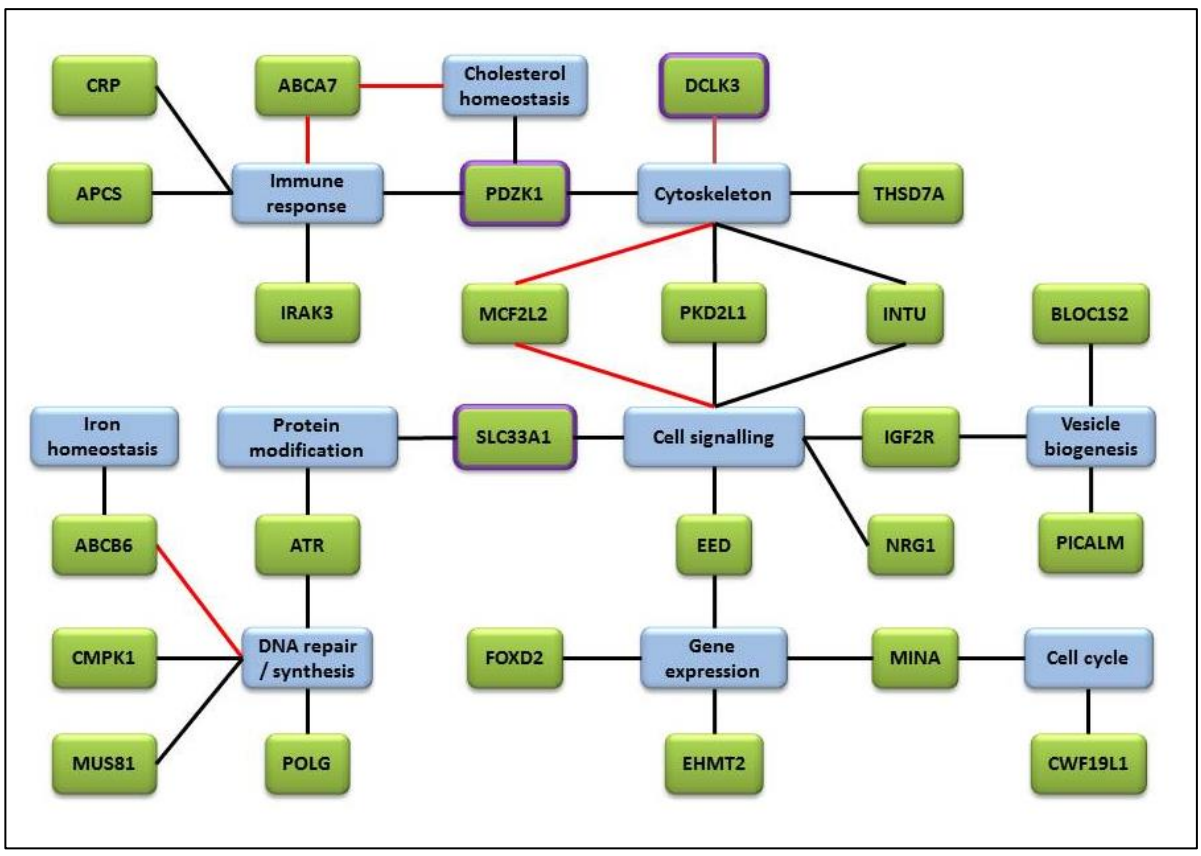



**Figure 10.1** Illustration of the top genes/proteins (green) identified through variant, gene and pathway analyses (**Section 9.2.1, Section 9.2.3** and **Section 9.2.4**) and their links (black and red lines) to general metabolic pathways (blue). Lines in red represent links with weaker supporting evidence. Each of the three genes/proteins encircled in purple are a top result from one of the analyses.

When assessing the results overall, two markers/loci stood out as interesting as they were highlighted in the results of several forms of analysis. The first marker (exm-rs3851179) genotypes the intergenic variant rs3851179. This marker was highlighted in variant, GWAS and pathway analyses, and is significantly associated with LOAD. The second marker (exm849171) genotypes the variant rs33965491. This variant was highlighted in both variant and gene analyses, despite the fact the *BLOC1S2* gene was not included in the pathway analysis due to its absence in the Reactome database, the relevant pathway (Golgi Associated Vesicle Biogenesis) was still highlighted in the results.

Variant rs33965491 appears to have two effects; it likely causes an amino acid change in one *BLOC1S2* transcript and it could also affect transcription of *BLOC1S2*. The variant rs3851179 overlaps a regulatory region and is located closest to a small non-coding RNA *RNU6-560P*, the nearest protein coding genes either side are *PICALM* and *EED*. This variant is thought to affect *PICALM* however evidence to support this is lacking; furthermore there is no evidence to discredit the idea that it might affect *RNU6-560P* and *EED*, indeed one paper shows this variant is an eQTL for EED.[294] It would be interesting to look at the data generated by the meta-analysis for this region in more detail. For instance, is there a significant SNP located within *RNU6-560P* or *EED* that is not in high linkage disequilibrium with SNPs located within *PICALM*? If so, this would suggest that *RNU6-560P*/EED is a contributing source of the susceptibility signal.

Future work for these two variants would be to verify the genotypes/MAF in sEOAD and control cohorts; this would be done with high throughput genotyping such as TaqMan or KASP genotyping. Further to this, the effect of these two variants on expression of the surrounding genes could be investigated using appropriate cell models.

The top gene from pathway analyses, *SLC33A1*, is linked to the functioning of BACE1 as is *NRG1* which was also highlighted in pathway analyses (**Section 9.2.4**). Both genes have one variant highlighted. BACE1 is the first enzyme involved in amyloidogenic processing of APP. Single variants in *APP*/*PSEN1*/*PSEN2* affect the functioning of the second amyloidogenic enzyme γ-secretase and alter the levels of APP processing; individually these variants have high penetrance and are sufficient to cause AD. As described in **Section 9.2.4** SLC33A1 is involved in transporting acetyl CoA to the lumen of the endoplasmic reticulum (ER) and Golgi apparatus (GA), where it is used to modify proteins such as BACE1. As described in **Section 9.2.4** NRG1 is involved in cell growth and is major substrate of BACE1. Given the function of these two proteins, the variant in SLC33A1 could affect the activity of BACE1 and result in altered levels of APP processing, the common variant in NRG1 could affect processing of NRG1 by BACE1 and result in altered cell growth.

Large-scale genetic analyses inevitably yield a new list of candidate genes and variants which require further study. Further experiments are needed to explore their function and establish their potential role in neurodegeneration and AD.

# 11. Appendix - In-house Software

Eleven programs have been written in-house. The programs perform processes that could not be computed by hand, to automate processes and allow replication.

Chromosomes 1 to 22 are coded accordingly; chromosome X, Y, XY (pseudo autosomal region) and M (mitochondrial) are coded 23, 24, 25 and 26 respectively.

## 11.1. Description of use

### Command line arguments

All commands are given in grey boxes and prefixed with a dollar symbol ($). Required arguments are given in square brackets and optional arguments in curly brackets. Programs should be run from a directory that contains two subdirectories; ‘IN’ and ‘OUT’. The IN directory should contain the input file(s) and the OUT directory is where the output file(s) are written. All Perl programs can be run on Linux, Windows or iOS providing Perl is installed.

### Input data

When input text files where created using Excel (version and link provided in **Section 12**) all data was written in the first column and the data was exported to text files by saving files to ‘tab-delimited txt’ on Windows or ‘windows formatted text’ on iOS. Some entries in input files are not computationally required but are still expected; these entries can be replaced with dummy data. Programs that use files created by PLINK are described in detail on the PLINK website (version and link provided in **Section 12**). All data are in text format and semi colon delimited unless otherwise stated, data should not contain semi colons unless used as a delimiter. The number of entries in input files will affect the run time linearly unless otherwise stated.

### Output data

All output files can be opened from within Excel; this allows the file to be displayed according to delimitation type and columns to be formatted appropriately; this is important for columns containing genes as the default Excel format will automatically convert some gene names into dates which cannot be undone, formatting columns containing gene names will stop this from occurring

### Pseudocode

For all Perl programs Pseudocode is given to explain simply how the main process functions.

## 11.2. AnnotateWithGenomicRegion.pl

This Perl program is designed to annotate points/positions on the genome with regions. For example, the point at chromosome 19 base pair 45412079 is SNP rs7412, this point would be annotated with five transcript regions; NM_001302688, NM_001302689, NM_001302690, NM_001302691 and NM_000041.

### Command line arguments

Four arguments can be entered; the name of the first input file (FileIn1), the name of the second input file (FileIn2), the name of the first output file (FileOut1) and the name of the second output file (FileOut2). FileIn1 lists all the points and FileOut2 lists all the regions.

```
$ perl AnnotateWithGenomicRegion.pl [FileIn1.txt] [FileIn2.txt] (FileOut1.txt) (FileOut2.txt)
```

### Input data

FileIn1 expects at least three entries per line and FileIn2 expects at least four. Entries for both files are described in **Table 11.1**. The number of lines in FileIn2 affects the computation time exponentially. RegionStart should always be smaller than RegionEnd.

| File | N | Name | Description | E | R |
|---|---|---|---|---|---|
| FileIn1 | 1 | PointName | Marker Name | Y | Y |
| | 2 | PointChromosome | Marker chromosome number (1-24, 26) | Y | Y |
| | 3 | PointStart | Base number at the point | Y | Y |
| | 4 | PointOptionalData | Additional data required by the user | N | N |
| FileIn2 | 1 | RegionName | Region Name | Y | Y |
| | 2 | RegionChromosome | Region chromosome number (1-24, 26) | Y | Y |
| | 3 | RegionStart | Base number at the start of the region | Y | Y |
| | 4 | RegionEnd | Base number at the end of the region | Y | Y |
| | 5 | RegionOptionalData | Additional data required by the user | N | N |

**Table 11.1** Description of the entries in each line of the two input files (File), including the entry number (N) and whether the entry is expected (E) and/or required (R) by the program. Each entry is named (name) and described (Description).

## Output data

FileOut1 and FileOut2 contain at least seven entries per line each. Entries for both files are described in **Table 11.2.** FileOut1 lists all the points (in the order given) annotated with the regions and FileOut2 lists all the regions (in the order given) annotated with the points. A point can be annotated with more than one region and a region can be annotated with more than one point.

| File | N | Name | Description |
|---|---|---|---|
| FileOut1 | 1 | PointName | Point Name |
| | 2 | PointChromosome | Point chromosome number (1-24, 26) |
| | 3 | PointStart | Base number at the point |
| | 4 | PointOptionalData | Additional point data stated by the user in FileIn1 |
| | 5 | RegionName | Region Name |
| | 6 | RegionChromosome | Region chromosome number (1-24, 26) |
| | 7 | RegionStart | Base number at the start of the region |
| | 8 | RegionEnd | Base number at the end of the region |
| | 9 | RegionOptionalData | Additional region data stated by the user in FileIn2 |
| FileOut2 | 1 | RegionName | Region Name |
| | 2 | RegionChromosome | Region chromosome (1-24, 26) |
| | 3 | RegionStart | Base number at the start of the region |
| | 4 | RegionEnd | Base number at the end of the region |
| | 5 | RegionOptionalData | Additional region data stated by the user in FileIn2 |
| | 6 | PointName | Point Name |
| | 7 | PointChromosome | Point chromosome (1-24, 26) |
| | 8 | PointStart | Base number at the point |
| | 9 | PointOptionalData | Additional point data stated by the user in FileIn1 |

**Table 11.2** Description of the entries in each line of the two output files (File), including the number of entries (N) generated. Each entry is named (name) and described (Description).

## Pseudocode

```
For each point
        Print point data to OutFile1.txt
        For each region
                If region chromosome equals point chromosome
                        If RegionStart is < PointStart and RegionStop is > than PointStart
                                Print region data to OutFile1.txt
                                Store point data by region data
        Move to new line in OutFile.txt
Print point data by region data to OutFile2.txt
```

## 11.3. ImissVsGender.r

This R script is used to visualise calculated gender for each sample. Each sample is plotted against their missing data rate and their X chromosome homozygosity score (F). Two clusters should be obvious; one cluster towards the top of the graph (male samples) and one at the bottom (female samples). The graph can be used to decide on F cut-off values for males and females.

### Command line arguments

No arguments are given when using R CMD BATCH statement; instead the names of two input and one output file (FileIn1, FileIn2 and FileOut1) are coded within the R script. FileIn1 is missing data rate file for samples (.imiss) created by PLINK and FileIn2 is a sexcheck file (.sexcheck) also created by PLINK.

```
$ R CMD BATCH ImissVsGender.r
```

### Input data

Both FileIn1 and FileIn2 expect six entries per line are semi colon delimited. Entries for both files are described in **Table 11.3**.

| File | N | Name | Description | E | R |
|---|---|---|---|---|---|
| FileIn1 | 1 | FID | Family ID | Y | N |
| | 2 | IID | Individual ID | Y | Y |
| | 3 | MISS_PHENO | Missing phenotype? (Y/N) | Y | N |
| | 4 | N_MISS | Number of missing SNPs | Y | N |
| | 5 | N_GENO | Number of non-obligatory missing genotypes | Y | N |
| | 6 | F_MISS | Proportion of missing SNPs (0-1) | Y | Y |
| FileIn2 | 1 | FID | Family ID | Y | N |
| | 2 | IID | Individual ID | Y | Y |
| | 3 | PEDSEX | Sex as determined in pedigree file (1=male, 2=female) | Y | N |
| | 4 | SNPSEX | Sex as determined by X chromosome | Y | N |
| | 5 | STATUS | Does PEDSEX match SNPSEX (PROBLEM/OK) | Y | N |
| | 6 | F | The actual X chromosome inbreeding (homozygosity) estimate | Y | Y |

**Table 11.3** Description of the entries in each line of the two input files (File), including the entry number (N) and whether the entry is expected (E) and/or required (R) by the program. Each entry is named (name) and described (Description).

### Output data

FileOut1 is a bitmap image file containing a graph that plots each sample on the same graph; each sample is plotted by their proportion of missing SNPs (F_MISS) versus their X chromosome homozygosity score (F).

# 11.4. Relatedness.pl

This program identifies pairs of samples that have identity by descent (IBD) greater than 0.1875.

## Command line arguments

Three arguments can be entered; the name of the first input file (FileIn1), name of the second input file (FileIn2) and the name of the output file (FileOut1). FileIn1 is missing data rate file for samples (.imiss) created by PLINK and FileIn2 is an IBD/IBS estimation file (.genome) created by PLINK.

```
$ perl Relatedness.pl [FileIn1] [FileIn2] (FileOut1)
```

## Input data

FileIn1 expects six entries per line and FileIn2 expects 14. Entries for both files are described in **Table 11.4**.

| File | N | Name | Description | E | R |
|---|---|---|---|---|---|
| FileIn1 | 1 | FID | Family ID | Y | N |
| | 2 | IID | Individual ID | Y | Y |
| | 3 | MISS_PHENO | Missing phenotype? (Y/N) | Y | N |
| | 4 | N_MISS | Number of missing SNPs | Y | N |
| | 5 | N_GENO | Number of non-obligatory missing genotypes | Y | N |
| | 6 | F_MISS | Proportion of missing SNPs (0-1) | Y | Y |
| FileIn2 | 1 | FID1 | Family ID for first individual | Y | N |
| | 2 | IID1 | Individual ID for first individual | Y | Y |
| | 3 | FID2 | Family ID for second individual | Y | N |
| | 4 | IID2 | Individual ID for second individual | Y | Y |
| | 5 | RT | Relationship type given PED file | Y | N |
| | 6 | EZ | Expected IBD sharing given PED file | Y | N |
| | 7 | Z0 | P(IBD=0) | Y | N |
| | 8 | Z1 | P(IBD=1) | Y | N |
| | 9 | Z2 | P(IBD=2) | Y | N |
| | 10 | PI_HAT | P(IBD=2)+0.5*P(IBD=1) ( proportion IBD ) | Y | Y |
| | 11 | PHE | Pairwise phenotypic code (1,0,-1 = AA, AU and UU pairs) | Y | N |
| | 12 | DST | IBS distance (IBS2 + 0.5*IBS1) / ( N SNP pairs ) | Y | N |
| | 13 | PPC | IBS binomial test | Y | N |
| | 14 | RATIO | Of HETHET : IBS 0 SNPs (expected value is 2) | Y | N |

**Table 11.4** Description of the entries in each line of the two input files (File), including the entry number (N) and whether the entry is expected (E) and/or required (R) by the program. Each entry is named (name) and described (Description).

## Output data

FileOut1 contains seven entries per line. Entries are described in **Table 11.5**.

| N | Name | Description |
|---|---|---|
| 1 | IID1 | Individual ID for first individual |
| 2 | FID1 | Family ID for first individual |
| 3 | F_MISS1 | Proportion of missing SNPs for first individual (0-1) |
| 4 | PI_HAT | P(IBD=2)+0.5*P(IBD=1) ( proportion IBD ) |
| 5 | IID1 | Individual ID for second individual |
| 6 | FID2 | Family ID for second individual |
| 7 | F_MISS2 | Proportion of missing SNPs for second individual (0-1) |

**Table 11.5** Description of the entries in each line of the output file (File), including the number of entries (N) generated. Each entry is named (name) and described (Description).

## Pseudocode

```
For each individual in FileIn1.txt
        Store individual missing data rate by individual ID
For each pair of individuals in FileIn2.txt
        If identity by decent is > 0.1875
                If the pair have not been seen before
                        Print to OutFile1.txt
```

## 11.5. Allele.pl

This program uses the information supplied in an Illumina CSV file to establish the alleles genotyped in plus/minus format; this format is required for programs such as variant effect predictor (VEP) and ANNOVAR. The Illumina CSV file contains information about each marker's probe design including the alleles genotyped; however these are coded in TOP/BOT format. This program utilises the source sequence in the Illumina CSV file to calculate this.

### Command line arguments

The command takes three arguments; the name of the input file (FileIn1), the size of the reference sequence to compare to MarkerSourceSequence (Flanking) and the name of the output file (FileOut1). FileIn1 lists all the markers and Flanking states the number of bases upstream/downstream of the marker start base (default is 14).

```
$ perl Alleles.pl [FileIn] (Flanking) (FileOut1)
```

### Input data

FileIn1 expects four entries per line (**Table 11.6**). Lines in FileIn1 should be ordered by MarkerChromosome. Markers on the pseudo autosomal region (25) should be recoded to 23 or 24. Also required are 25 fasta formatted files, named chr1.fa to chr24.fa and chr26.fa; each containing the chromosome sequence (Plus strand) for chromosomes 1-24 and 26.

| N | Name | Description | E | R |
|---|---|---|---|---|
| 1 | MarkerName | Marker Name | Y | Y |
| 2 | MarkerChromosome | Marker chromosome (1-24, 26) | Y | Y |
| 3 | MarkerStart | Base number of the start of the variant | Y | Y |
| 4 | MarkerSourceSequence | Marker source sequence from the Illumina CSV file which should be in the format *[ #/# ]*<br>The asterisk and hashtag are one or more IUPAC nucleotide letters (A, C, G, T, U, W, S, M, K, R, Y, B, D, H, V or N), either of the two hashtags can be a dash (-) to represent an indel | Y | Y |

**Table 11.6** Description of the entries in each line of the one input file (File), including the entry number (N) and whether the entry is expected (E) and/or required (R) by the program. Each entry is named (name) and described (Description).

## Output data

FileOut1 contains 14 entries; these are described in **Table 11.7**. If entries 6-11 could not be calculated then they are filled with a dash symbol. A second file called Errors.txt is created at run time which has included in it all the errors produced at run time. If one or more errors are written to this file, the user will be notified on screen at the end of the run.

| N | Name | Description |
|---|---|---|
| 1 | MarkerCount | Line number of the marker as stated in FileIn1 |
| 2 | MarkerName | Marker Name |
| 3 | MarkerChromosome | Marker chromosome (1-24, 26) |
| 4 | MarkerStart | Base number where the variant begins |
| 5 | MarkerSourceSequence | Marker source sequence from the Illumina CSV file which should be in the format *[ #/# ]*<br>The asterisk and hashtag are one or more IUPAC nucleotide letters (A, C, G, T, U, W, S, M, K, R, Y, B, D, H, V or N), either of the two hashtags can be a dash (-) to represent an indel |
| 6 | MarkerStrand | Strand of the MarkerSourceSequence, which can be Plus or Minus |
| 7 | MarkerStartNew | New base number where the variant begins |
| 8 | MarkerEnd | Base number where the variant ends |
| 9 | AlleleReference | Allele of the variant on the Plus strand, which is one or more IUPAC nucleotide letters (A, C, G, T, U, W, S, M, K, R, Y, B, D, H, V or N) or a dash (-) for an insertion. |
| 10 | AlleleAlternative | Allele of the variant on the Minus strand, which is one or more IUPAC nucleotide letters (A, C, G, T, U, W, S, M, K, R, Y, B, D, H, V or N) or a dash (-) for a deletion. |
| 11 | VariantType | Type of variant which can be ‘Insertion’, ‘Deletion’ or ‘SNP’ |
| 12 | ReferenceSequence | Chromosome Plus strand sequence |
| 13 | AllMatches | Number of matches when comparing ReferenceSequence to MarkerSourceSequence |
| 14 | AllMatchesCode | Code that displays the number of matches for each comparison |

**Table 11.7** Description of the entries in each line of the output file (File), including the number of entries (N) generated. Each entry is named (name) and described (Description).

## Pseudocode

```
For every marker in FileIn1.txt
        Grab all of reference sequence from the appropriate chromosome file
        Split MarkerSourceSequence into left sequence, first allele, second allele and right sequence
        Create all combinations of sequence and complement sequences of these
        Grab the reference sequence at MarkerBase with flanking upstream and downstream bases
        Compare reference sequence to all combinations of sequence
        If there is only one match
                Calculate MarkerStrand, MarkerStartNew, MarkerEnd, AlleleReference,
                AlleleAlternative, and VariantType
                Print all data to FileOut1.txt
        Go to new line in FileOut1.txt
```

# 11.6. FormatVEP.pl

This program takes the text data created by VEP (Link and version provided in **Section 12**). The data produced when running VEP locally, contains 'Extra' data at the end of each line, however where some variants have missing data a dummy variable isn't added, thus when this data is opened within Excel, data within a specific column are not of the same type.

## Command line arguments

The command takes two arguments; the names of the input file and the name of the output file (FileIn1 and FileOut1 respectively). FileIn1 lists the raw text data generated from running VEP locally and FileOut1 lists the 'Extra' column data reformatted.

```
$ perl FormatVEP.pl [FileIn1] (FileOut1)
```

## Input data

FileIn1 expects 14 entries per line (**Table 11.8**) which are tab delimited. Lines starting with a hashtag will be ignored.

| N | Name | Description | E | R |
|---|---|---|---|---|
| 1 | Uploaded | Name of variant | Y | Y |
| 2 | Location | Variant coordinate uploaded to VEP (chr:start or chr:start-end) | Y | N |
| 3 | Allele | Variant allele used to calculate the consequence | Y | N |
| 4 | Gene | Ensembl stable ID of affected gene | Y | N |
| 5 | Feature | Ensembl stable ID of feature | Y | N |
| 6 | Feature_type | Type of feature. Currently one of Transcript, RegulatoryFeature, MotifFeature | Y | N |
| 7 | Consequence | Consequence type of this variant | Y | N |
| 8 | cDNA_position | Relative position of base pair in cDNA sequence | Y | N |
| 9 | CDS_position | Relative position of base pair in coding sequence | Y | N |
| 10 | Protein_position | Relative position of amino acid in protein | Y | N |
| 11 | Amino_acids | Amino acid change of the variant. Only given if the variant affects the protein-coding sequence | Y | N |
| 12 | Codons | Alternative codons with the variant base in upper case | Y | N |
| 13 | Existing_variation | Known identifier of existing variant | Y | N |
| 14 | Extra | Extra information separated by semi colon (key=value) | Y | Y |

**Table 11.8** Description of the entries in each line of the two input files (File), including the entry number (N) and whether the entry is expected (E) and/or required (R) by the program. Each entry is named (name) and described (Description).

## Output data

FileOut1 contains 36 entries (**Table 11.9**). If no information for entries 2-36 exists then they are filled with a dash symbol.

| N | Name | Description |
|---|---|---|
| 1 | Uploaded | Name of variant |
| 2 | AA_MAF | Frequency of existing variant in NHLBI-ESP African American population |
| 3 | AFR_MAF | Frequency of existing variant in 1000 Genomes combined African population |
| 4 | AMR_MAF | Frequency of existing variant in 1000 Genomes combined American population |
| 5 | ASN_MAF | Frequency of existing variant in 1000 Genomes combined Asian population |
| 6 | BIOTYPE | Biotype of transcript or regulatory feature |
| 7 | CANONICAL | Indicates if transcript is canonical for this gene |
| 8 | CCDS | Indicates if transcript is a CCDS transcript |
| 9 | CLIN_SIG | Clinical significance of variant from dbSNP |
| 10 | DISTANCE | Shortest distance from variant to transcript |
| 11 | DOMAINS | The source and identifer of any overlapping protein domains |
| 12 | EA_MAF | Frequency of existing variant in NHLBI-ESP European American population |
| 13 | EAS_MAF | Frequency of existing variant in 1000 Genomes combined East Asian population |
| 14 | ENSP | Ensembl protein identifer |
| 15 | EUR_MAF | Frequency of existing variant in 1000 Genomes combined European population |
| 16 | EXON | Exon number(s) / total |
| 17 | GMAF | Minor allele and frequency of existing variant in 1000 Genomes combined population |
| 18 | HGNC_ID | Stable identifer of HGNC gene symbol |
| 19 | HIGH_INF_POS | A flag indicating if the variant falls in a high information position of the TFBP |
| 20 | IMPACT | Subjective impact classification |
| 21 | INTRON | Intron number(s) / total |
| 22 | MOTIF_NAME | The source and identifier of a transcription factor binding profile (TFBP) aligned at this position |
| 23 | MOTIF_POS | The relative position of the variation in the aligned TFBP |
| 24 | MOTIF_SCORE_CHANGE | The difference in motif score of the reference and variant sequences for the TFBP |
| 25 | PolyPhen | PolyPhen prediction and/or score |
| 26 | PUBMED | Pubmed ID(s) of publications that cite existing variant |
| 27 | SAS_MAF | Frequency of existing variant in 1000 Genomes combined South Asian population |
| 28 | SIFT | SIFT prediction and/or score |
| 29 | SOMATIC | Somatic status of existing variant |
| 30 | STRAND | Strand of the feature (1/-1) |
| 31 | SWISSPROT | UniProtKB/Swiss-Prot identifier |
| 32 | SYMBOL | Gene symbol (e.g. HGNC) |
| 33 | SYMBOL_SOURCE | Source of gene symbol |
| 34 | TREMBL | UniProtKB/TrEMBL identifier |
| 35 | TSL | Transcript support level |
| 36 | UNIPARC | UniParc identifier |

**Table 11.9** Description of the entries in each line of the output file (File), including the number of entries (N) generated. Each entry is named (name) and described (Description).

## Pseudocode

```
For each line that does not begin with a hash tag
        Print name of variant to FileOut1
        Reset all values of hash data structure to dash
        Separate data in Extra column by semi colon to retrieve entries
                Separate entries by equals sign to retrieve key and value
                Store value by key in hash data structure
        Print hash data to FileOut1
        Move to new line in FileOut1
```

# 11.7. MatchingVariants.pl

This program takes two lists of variant information and identifies if variants in one list are present in the other.

## Command line arguments

The command takes in five arguments; the name of the input files (FileIn1 and FileIn2), a 'Y' or a 'N' to state if the order of the alleles in both input file is important (Order) , and finally the output files (FileOut1 an FileOut2). FileIn1 lists the first set of variants, FileIn2 lists the second set of variants, FileOut1 lists the first set of variants annotated with the second set of variants and FileOut2 lists the second set of variants annotated with the first set of variants.

```
$ perl MatchingVariants.pl [FileIn1] [FileIn2] (Order) (FileOut1) (FileOut2)
```

## Input data

FileIn1 and FileIn2 each expects at least five entries per line (**Table 11.10**). Lines that are filled with a dash in entry 2/3 will be ignored.

| File | N | Entry Name | Description | E | R |
|---|---|---|---|---|---|
| FileIn1 | 1 | ListA _Name | Variant name | Y | Y |
| | 2 | ListA _Chromosome | Variant chromosome (1-24, 26) | Y | Y |
| | 3 | ListA _Start | Base number of the start of the variant | Y | Y |
| | 4 | ListA _Allele1 | Allele1 of the variant | Y | Y |
| | 5 | ListA _Allele2 | Allele2 of the variant | Y | Y |
| FileIn2 | 1 | ListB _Name | Variant name | Y | Y |
| | 2 | ListB _Chromosome | Variant chromosome (1-24, 26) | Y | Y |
| | 3 | ListB _Start | Base number of the start of the variant | Y | Y |
| | 4 | ListB _Allele1 | Allele1 of the variant | Y | Y |
| | 5 | ListB _Allele2 | Allele2 of the variant | Y | Y |
| | 6 | ListB _AdditionalData | Additional data required by the user | N | N |

**Table 11.10** Description of the entries in each line of the two input files (File), including the entry number (N) and whether the entry is expected (E) expected (N) and/or required (R) by the program. Each entry is named (name) and described (Description).

## Output data

FileOut1 and FileOut2 each contain 12 entries (**Table 11.11**).

| File | N | Entry Name | Description |
|---|---|---|---|
| FileIn1 | 1 | ListA _Name | Variant name |
| | 2 | ListA _Chromosome | Variant chromosome (1-24, 26) |
| | 3 | ListA _Start | Base number of the start of the variant |
| | 4 | ListA _Allele1 | Allele1 of the variant |
| | 5 | ListA _Allele2 | Allele2 of the variant |
| | 6 | ListA _AdditionalData | Additional data required by the user |
| | 7 | ListB _Name | Variant name |
| | 8 | ListB _Chromosome | Variant chromosome (1-24, 26) |
| | 9 | ListB _Start | Base number of the start of the variant |
| | 10 | ListB _Allele1 | Allele1 of the variant |
| | 11 | ListB _Allele2 | Allele2 of the variant |
| | 12 | ListB _AdditionalData | Additional data required by the user |
| FileIn2 | 1 | ListB _Name | Variant name |
| | 2 | ListB _Chromosome | Variant chromosome (1-24, 26) |
| | 3 | ListB _Start | Base number of the start of the variant |
| | 4 | ListB _Allele1 | Allele1 of the variant |
| | 5 | ListB _Allele2 | Allele2 of the variant |
| | 6 | ListB _AdditionalData | Additional data required by the user |
| | 7 | ListA _Name | Variant name |
| | 8 | ListA _Chromosome | Variant chromosome (1-24, 26) |
| | 9 | ListA _Start | Base number of the start of the variant |
| | 10 | ListA _Allele1 | Allele1 of the variant |
| | 11 | ListA _Allele2 | Allele2 of the variant |
| | 12 | ListA _AdditionalData | Additional data required by the user |

**Table 11.11** Description of the entries in each line of the two output files (File), including the number of entries (N) generated. Each entry is named (name) and described (Description).

## Pseudocode

```
For each variant in ListA
   Print ListA data to FileOut1
      Create complements of ListA_Allele1 and ListA_Allele2
         For each variant in ListB_Variant
            If the Chromosomes match
               If the Starts match
                  If ListA_Allele1 matches ListB_Allele1
                  and ListA_Allele2 matches ListB_Allele2
                     Print ListB data to FileOut1
                     Store ListA data in array
                  If ListA_Allele1complement matches ListB_Allele1
                  and ListA_Allele2 complement matches ListB_Allele2
                     Print ListB data to FileOut1
                     Store ListA data in array
                  If the order of alleles does not matter
                      If ListA_Allele1complement matches ListB_Allele2
                     and ListA_Allele2 complement matches ListB_Allele1
                        Print ListB data to FileOut1
                        Store ListA data in array
                     If ListA_Allele1complement matches ListB_Allele2
                     and ListA_Allele2 complement matches ListB_Allele1
                        Print ListB data to FileOut1
                        Store ListA data in array
For each element in array
    Print data to FileOut2
```

## 11.8. SwitchPrimaryKey.pl

This program takes in generic data that has primary keys. Each primary key and its associated terms are on a separate line, the program reformats the data so that each term becomes a new primary key. An example is given in **Table 11.12**.

| Input | Output |
|---|---|
| PrimaryKey1;Term1;Term2 | Term1;PrimaryKey1 |
| PrimaryKey2;Term2;Term3 | Term2;PrimaryKey1;PrimaryKey2 |
| | Term3;PrimaryKey2 |

**Table 11.12** Example data demonstrating how input data (Input) are reformatted to give the output data (Output)

### Command line arguments

The command takes two arguments; the names of the input file and name of the output file (FileIn1 and FileOut1 respectively).

```
$ perl SwitchPrimaryKey.pl [FileIn1] (FileOut1)
```

### Input data

FileIn1 expects at least one entry per line. The first entry of each line is assumed to be the primary key and all entries after are terms associated with that primary key.

### Pseudocode

```
For each line in FileIn1
    For each term
        Store term in hash if it doesn't already exist
For each hash element
    Print hash element to FileOut1
    For each line in FileIn1
        For each term
            If hash element data matches term
                Print primary key to FileOut1
        Print new line to FileOut1
```

## 11.9. FormatPrimaryKey.pl

This program takes in generic data that has primary keys. Each primary key and its associated terms are on a separate line, the program reformats the data as required by PLINKs set-test or SKAT-O. An example is given in **Table 11.13**.

| Input | SKAT-O | PLINK |
|---|---|---|
| PrimaryKey1;Term1;Term2 | PrimaryKey1;Term1 | PrimaryKey1 |
| PrimaryKey2;Term2;Term3 | PrimaryKey1;Term2 | Term1 |
| | PrimaryKey2;Term2 | Term2 |
| | PrimaryKey2;Term3 | END |
| | | |
| | | PrimaryKey2 |
| | | Term2 |
| | | Term3 |
| | | END |

**Table 11.13** Example data demonstrating how input data (Input) are reformatted to either of two types; SKAT-O (SKAT-O) or PLINK set-test (PLINK).

### Command line arguments

The command takes three arguments; the names of the input file and output file (FileIn1 and FileOut1 respectively) followed by a number 1 for SKAT-O format, or a number 2 for PLINK set-test format (Format).

```
$ perl FormatPrimaryKey.pl [FileIn1] [FileOut1] [Format]
```

### Input data

FileIn1 expects at least one entry per line. It is assumed the first entry of each line is the primary key and all entries after are terms associated with it

### Pseudocode

```
If format is SKAT-O
    For each line of FileIn1
        For each term
            Print PrimaryKey;Term\n to FileOut1
If format is PLINK set-test
    For each line in FileIn1
        Print PrimaryKey\n to FileOut1
        For each term
            Print Term\n to FileOut1
        Print END\n\n to FileOut1
```

## 11.10. RemoveDuplicates.pl

This program takes in generic data and removes duplicates. The program can focus on lines or entries at a particular position. The program removes lines that are duplicate lines or removes lines that contain a duplicate entry. An example is given in **Table 11.14**.

| Input | Line | Entry |
|---|---|---|
| Term1;Term2;Term3 | Term1;Term2;Term3 | Term2 |
| Term4;Term5;Term6 | Term4;Term5;Term6 | Term5 |
| Term7;Term5;Term8 | Term7;Term5;Term8 | |
| Term1;Term2;Term3 | | |

**Table 11.14** Example data demonstrating how duplicated input data (Input) are removed from either the entire line (Line) or a specific entry, for example number two (Entry) to create the output data.

### Command line arguments

The command takes three arguments; the names of the input file and output file (FileIn1 and FileOut1 respectively) followed by the number 0 to mean remove duplicate lines, or a number greater than zero to indicate remove duplicates at a particular entry position.

```
$ perl RemoveDuplicates.pl [FileIn1] [FileOut1] [Format]
```

### Input data

FileIn1 expects at least one entry per line and each line should have the same number of entries.

### Pseudocode

```
If remove duplicates from lines
    For each line in FileIn1
        Store line in array
    Create unique array using library ‘List::MoreUtils qw(uniq)’ function ‘uniq’
    For each element in unique array
        Print value to FileOut1
If remove duplicates from specific entry
    For each line in FileIn1
        If term at specific entry has not been seen before
            Store term in hash
    For each hash key sorted
        Print value of hash key to FileOut1
```

# 11.11. CombineTwoPrimaryKeys.pl

This program takes in two sets of generic data that have primary keys. Each primary key and its associated terms are on a separate line. The program creates a new set of data where the primary key is that used in the first input data file and the terms are the primary keys from the second input data file. The program uses the terms from both input files as a linker. An example is given in **Table 11.15**.

| First Input | Second Input | Output |
|---|---|---|
| PrimaryKey1;Term1;Term2 | PrimaryKey3;Term1;Term4 | PrimaryKey1;PrimaryKey3;PrimaryKey4 |
| PrimaryKey2;Term2;Term3 | PrimaryKey4;Term5;Term2 | PriamryKey2;PrimaryKey4 |

**Table 11.15** Example data demonstrating how the primary key from the first input file (First Input) and the primary key from the second input file (Second Input) are combined using their terms.

## Command line arguments

The command takes three arguments; the names of the first input file and second input file (FileIn1 and FileIn2 respectively) followed by the name of the output file (FileOut1).

```
$ perl CombineTwoPrimaryKeys.pl [FileIn1] [FileIn2] (FileOut1)
```

## Input data

FileIn1 expects at least one entry per line. The first entry of each line is assumed to be the primary key and all entries after are terms associated with that primary key.

## Pseudocode

```
For each line in FileIn1
    Empty array
    For each term
        For each line in FileIn2
            For each term
                If terms match
                    Store primary key from line of FileIn2 in array
    Create unique array using library ‘List::MoreUtils qw(uniq)’ function ‘uniq’
    Print primary key from FileIn1 to FileOut1
    For each value in unique array
        Print value to FileOut1
    Move to new line in FileOut1
```

## 11.12. GeneAnalysis.r

This R script is used to automate SKAT-O analysis using the R package 'SKAT' (Link and version provided in **Section 12**)

### Command line arguments

No arguments are given when using R CMD BATCH statement, instead the input file names are coded within the R script; the names of five input and one output file (FileIn1, FileIn2, FileIn3, FileIn4, FileIn5 and FileOut1) are coded within the R script. FileIn1 is a binary file containing genotype data (.bed) created by PLINK. FileIn2 is a text file containing phenotype data (.fam) created by PLINK. FileIn3 is a text file containing marker coordinate data (.bim) created by PLINK. FileIn4 is a text file containing covariate data. FileIn5 is a set file in SKAT-O format. FileOut1 is the text file containing a list of sets and their test p.values.

```
$ R CMD BATCH GeneAnalysis.r
```

## Input data

The script takes in five input files. Data in FileIn1 are binary encoded as created by PLINK. Entries in FileIn2 to FileIn5 are tab delimited and are described in **Table 11.16**. FileIn4 requires a header line.

| File | N | Name | Description | E | R |
|---|---|---|---|---|---|
| FileIn2 | 1 | FID | Family ID | Y | N |
| | 2 | IID | Individual ID | Y | Y |
| | 3 | PID | Paternal ID | Y | N |
| | 4 | MID | Maternal ID | Y | N |
| | 5 | G | Gender | Y | N |
| | 6 | P | Phenotype | Y | Y |
| FileIn3 | 1 | C | Chromosome | Y | N |
| | 2 | MN | Marker name | Y | Y |
| | 3 | GD | Genetic distance (cM) | Y | N |
| | 4 | BP | Base number of the start of the variant | Y | N |
| | 5 | Allele1 | Allele1 of the variant | Y | Y |
| | 6 | Allele2 | Allele2 of the variant | Y | Y |
| FileIn4 | 1 | FID | Family ID | Y | Y |
| | 2 | IID | Individual ID | Y | Y |
| | 3 | DatabaseGender | Database gender (1=male, 2=female) | Y | Y |
| | 4 | CalculatedGender | Calculated Gender (1=male, 2=female) | Y | Y |
| | 5 | Age | Age | Y | Y |
| | 6 | E2 | *APOE* ε2 dosage | Y | Y |
| | 7 | E4 | *APOE* ε4 dosage | Y | Y |
| | 8 | PCA1 | Principle component 1 | Y | Y |
| | 9 | PCA2 | Principle component 2 | Y | Y |
| | 10 | PCA3 | Principle component 3 | Y | Y |
| | 11 | PCA4 | Principle component 4 | Y | Y |
| | 12 | PCA5 | Principle component 5 | Y | Y |
| | 13 | PCA6 | Principle component 6 | Y | Y |
| | 14 | PCA7 | Principle component 7 | Y | Y |
| | 15 | PCA8 | Principle component 8 | Y | Y |
| | 16 | PCA9 | Principle component 9 | Y | Y |
| | 17 | PCA10 | Principle component 10 | Y | Y |
| FileIn5 | 1 | SetName | Set name | Y | Y |
| | 2 | MarkerName | Marker name | Y | Y |

**Table 11.16** Description of the entries in each line of the four text input files (File), including the entry number (N) and whether the entry is expected (E) and/or required (R) by the program. Each entry is named (name) and described (Description).

## Output data

FileOut1 contains five entries which are tab delimited and are described in **Table 11.17**.

| N | Entry Name | Description |
|---|---|---|
| 1 | Set | Variant name |
| 2 | P | Variant chromosome (1-24, 26) |
| 3 | NMarkerAll | Base number of the start of the variant |
| 4 | NMarkerTest | Allele1 of the variant |
| 5 | PFDR | Allele2 of the variant |

**Table 11.17** Description of the entries in each line of the output file (File), including the number of entries (N) generated. Each entry is named (name) and described (Description).

# 12. Appendix – External Software

Many pieces of commercial and free software were used during this project. Operating systems used including Windows 7, Windows 8, Ubuntu 12.04, Red Hat, iOS Yosemite and iOS Capitan. Commercial applications used include Microsoft Excel 2010 for Windows, Microsoft Excel 2011 for Mac and Illumina GenomeStudio 2011.1 with genotyping module 1.9.4. Listed below are the free applications and databases used.

## 12.1. Databases

Ensembl Genome Browser (GRCh37)
http://grch37.ensembl.org/Homo_sapiens/Info/Index
Ensembl Genome Browser
http://www.ensembl.org/Homo_sapiens/Info/Index
Ensembl VEP
http://www.ensembl.org/info/docs/tools/vep/index.html
UCSC Genome Browser
https://genome.ucsc.edu/cgi-bin/hgGateway
UCSC Golden Path (hg19)
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/
UCSC LiftOver
https://genome.ucsc.edu/cgi-bin/hgLiftOver
UCSC In-Silico PCR
https://genome.ucsc.edu/cgi-bin/hgPcr
UCSC Table Browser
https://genome.ucsc.edu/cgi-bin/hgTables
AlzGene
http://www.alzgene.org/TopResults.asp
AD&FTDMDB (accessed August 2013)
www.molgen.ua.ac.uk/admutations
PDMutDB (accessed August 2013)
www.molgen.vib-ua.be/PDMutDB
HPDB (accessed November 2014)
http://www.mad-cow.org/prion_point_mutations.html
Reactome v53
http://www.reactome.org/
ExAC (current release 0.3)
http://exac.broadinstitute.org/

## 12.2. Quality control and analyses

PLINK v1.07
http://pngu.mgh.harvard.edu/~purcell/plink/
EIGENSOFT v5.0 (includes Smartpca, Twstats and Ploteig)
https://genetics.med.harvard.edu/reich/Reich_Lab/Software.html
IMPUTE2
https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#home
GTOOL
http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html
SKAT-O v0.95 or v1.07
https://cran.r-project.org/web/packages/SKAT/index.html
ANNOVAR (used June 2015)
http://annovar.openbioinformatics.org/en/latest/
CHARGE consortium
http://web.chargeconsortium.com/main/exomechip
Illumina technical notes
http://www.illumina.com/Documents/products/technotes/
Quonto v1.2.4
http://biostats.usc.edu/Quanto.html

## 12.3. Programming

Strawberry Perl for Windows v5.22.0.1
http://strawberryperl.com/
Perl for Mac
Pre-installed on Mac
Perl for Linux
Pre-installed on Linux
R for Windows
https://cran.r-project.org/bin/windows/base/
R for Mac
https://cran.r-project.org/bin/macosx/
TextWrangler for Mac
http://www.barebones.com/products/textwrangler/
PadreIDE v0.94
http://padre.perlide.org/
RGui for Windows v3.1.0
https://www.r-project.org/
RStudio for Windows v0.98.109
https://www.rstudio.com/products/rstudio/download/
RStudio for Mac v0.99.467
https://www.rstudio.com/products/rstudio/download/

## 12.4. PCR and sequencing

NEBcutter v2.0
http://nc2.neb.com/NEBcutter2/
Primer3 v0.40
http://bioinfo.ut.ee/primer3-0.4.0/
SNPCheck3
https://secure.ngrl.org.uk/SNPCheck/snpcheck.html
SNAP v2.2
https://www.broadinstitute.org/mpg/snap/
Sequence Scanner Software v2
http://www.appliedbiosystems.com/absite/us/en/home/support/software-community/free-ab-software.html
ClustalW2
http://www.ebi.ac.uk/Tools/msa/clustalw2/

# Publications

**Barber IS**, García-Cárdenas JM, Sakdapanichkul C, Deacon C, Zapata Erazo G, Guerreiro R, Bras J, Hernandez D, Singleton A, Guetta-Baranes T, Braae A, Clement N, Patel T, Brookes K, Medway C, Chappell S, Mann DM; ARUK Consortium, Morgan K. **Screening exons 16 and 17 of the amyloid precursor protein gene in sporadic early-onset Alzheimer's disease**. *Neurobiol Aging*. 2015 Dec 29. pii: S0197-4580(15)00615-6. doi: 10.1016/j.neurobiolaging.2015.12.011

Guerreiro R, Escott-Price V, Darwent L, Parkkinen L, Ansorge O, Hernandez DG, Nalls MA, Clark L, Honig L, Marder K, van der Flier W, Holstege H, Louwersheimer E, Lemstra A, Scheltens P, Rogaeva E, St George-Hyslop P, Londos E, Zetterberg H, Ortega-Cubero S, Pastor P, Ferman TJ, Graff-Radford NR, Ross OA, **Barber I**, Braae A, Brown K, Morgan K, Maetzler W, Berg D, Troakes C, Al-Sarraj S, Lashley T, Compta Y, Revesz T, Lees A, Cairns NJ, Halliday GM, Mann D, Pickering-Brown S, Powell J, Lunnon K, Lupton MK; International Parkinson's Disease Genomics Consortium (IPDGC), Dickson D, Hardy J, Singleton A, Bras J. **Genome-wide analysis of genetic correlation in dementia with Lewy bodies, Parkinson's and Alzheimer's diseases**. *Neurobiol Aging*. 2016 Feb;38:214.e7-214.e10. doi: 10.1016/j.neurobiolaging.2015.10.028

Carrasquillo MM, **Barber I**, Lincoln SJ, Murray ME, Camsari GB, Khan QU, Nguyen T, Ma L, Bisceglio GD, Crook JE, Younkin SG, Dickson DW, Boeve BF, Graff-Radford NR, Morgan K, Ertekin-Taner N. **Evaluating pathogenic dementia variants in posterior cortical atrophy**. *Neurobiol Aging*. 2016 Jan;37:38-44. doi: 10.1016/j.neurobiolaging.2015.09.023

Bras J, Guerreiro R, Darwent L, Parkkinen L, Ansorge O, Escott-Price V, Hernandez DG, Nalls MA, Clark LN, Honig LS, Marder K, Van Der Flier WM, Lemstra A, Scheltens P, Rogaeva E, St George-Hyslop P, Londos E, Zetterberg H, Ortega-Cubero S, Pastor P, Ferman TJ, Graff-Radford NR, Ross OA, **Barber I**, Braae A, Brown K, Morgan K, Maetzler W, Berg D, Troakes C, Al-Sarraj S, Lashley T, Compta Y, Revesz T, Lees A, Cairns N, Halliday GM, Mann D, Pickering-Brown S, Dickson DW, Singleton A, Hardy J. **Genetic analysis implicates APOE, SNCA and suggests lysosomal dysfunction in the etiology of dementia with Lewy bodies**. *Hum Mol Genet*. 2014 Dec 1;23(23):6139-46. doi: 10.1093/hmg/ddu334

Cruchaga C, Karch CM, Jin SC, Benitez BA, Cai Y, Guerreiro R, Harari O, Norton J, Budde J, Bertelsen S, Jeng AT, Cooper B, Skorupa T, Carrell D, Levitch D, Hsu S, Choi J, Ryten M; UK Brain Expression Consortium, Hardy J, Ryten M, Trabzuni D, Weale ME, Ramasamy A, Smith C, Sassi C, Bras J, Gibbs JR, Hernandez DG, Lupton MK, Powell J, Forabosco P, Ridge PG, Corcoran CD, Tschanz JT, Norton MC, Munger RG, Schmutz C, Leary M, Demirci FY, Bamne MN, Wang X, Lopez OL, Ganguli M, Medway C, Turton J, Lord J, Braae A, **Barber I**, Brown K; Alzheimer's Research UK Consortium, Passmore P, Craig D, Johnston J, McGuinness B, Todd S, Heun R, Kölsch H, Kehoe PG, Hooper NM, Vardy ER, Mann DM, Pickering-Brown S, Brown K, Kalsheker N, Lowe J, Morgan K, David Smith A, Wilcock G, Warden D, Holmes C, Pastor P, Lorenzo-Betancor O, Brkanac Z, Scott E, Topol E, Morgan K, Rogaeva E, Singleton AB, Hardy J, Kamboh MI, St George-Hyslop P, Cairns N, Morris JC, Kauwe JS, Goate AM**. Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease**. *Nature*. 2014 Jan 23;505(7484):550-4. doi: 10.1038/nature12825

Benitez BA, Jin SC, Guerreiro R, Graham R, Lord J, Harold D, Sims R, Lambert JC, Gibbs JR, Bras J, Sassi C, Harari O, Bertelsen S, Lupton MK, Powell J, Bellenguez C, Brown K, Medway C, Haddick PC, van der Brug MP, Bhangale T, Ortmann W, Behrens T, Mayeux R, Pericak-Vance MA, Farrer LA, Schellenberg GD, Haines JL, Turton J, Braae A, **Barber I**, Fagan AM, Holtzman DM, Morris JC; 3C Study Group; EADI consortium; Alzheimer's Disease Genetic Consortium (ADGC); Alzheimer's Disease Neuroimaging Initiative (ADNI); GERAD Consortium, Williams J, Kauwe JS, Amouyel P, Morgan K, Singleton A, Hardy J, Goate AM, Cruchaga C. **Missense variant in TREML2 protects against Alzheimer's disease**. *Neurobiol Aging*. 2014 Jun;35(6):1510.e19-26. doi: 10.1016/j.neurobiolaging.2013.12.010

# Abbreviations

| | |
|---|---|
| AD | Alzheimer’s disease |
| sEOAD | Sporadic early-onset Alzheimer’s disease |
| LOAD | Late-onset Alzheimer’s disease |
| FTD | Frontotemporal Dementia |
| PD | Parkinson’s disease |
| DLB | Dementia with Lewy Bodies |
| NIA | National Institute of Aging |
| UCL | University College London |
| UN | University of Nottingham |
| APP | Amyloid precursor protein |
| SNP | Single nucleotide polymorphism |
| GWAS | Genome-wide association study |
| AAO | Age at onset |
| DNA | Deoxyribonucleic acid |
| RNA | Ribonucleic acid |
| PCA | Polymerase chain reaction |
| QC | Quality control |
| Min | Minute |
| Temp | Temperature |
| Sec | Second |
| RPM | Revolutions per minute |
| bp | Base pair |
| Indel | Insertion/deletion |
| MAF | Minor allele frequency |
| p.value | Probability value |
| PAR | Pseudo-autosomal region |
| PAF | Population attributable factor |
| SD | Standard deviation |

# References


1. Budson, A.E. & Kowall, N.W. *The Handbook of Alzheimer's Disease and Other Dementias Preface*, XV-XV (Wiley-Blackwell, Chichester, 2011).
2. Brookmeyer, R., Corrada, M.M., Curriero, F.C. & Kawas, C. Survival following a diagnosis of Alzheimer disease. *Archives of Neurology* **59**, 1764-1767 (2002).
3. Hodges, J.R. Alzheimer's centennial legacy: origins, landmarks and the current status of knowledge concerning cognitive aspects. *Brain* **129**, 2811-22 (2006).
4. Naghavi, M. *et al.* Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* **385**, 117-171 (2015).
5. Dementia UK: Second Edition. in *Dementia UK* (Alzheimer's Society, 2014).
6. WHO. Top Ten leading Causes of Death by Country Income Group. in *Top Ten Causes of Death* (World Health Organisation, 2012).
7. Luengo-Fernandez, R., Leal, J. & Gray, A. UK research spend in 2008 and 2012: comparing stroke, cancer, coronary heart disease and dementia. *BMJ Open* **5**, e006648 (2015).
8. Whitwell, J.L. *et al.* MRI correlates of neurofibrillary tangle pathology at autopsy - A voxel-based morphometry study. *Neurology* **71**, 743-749 (2008).
9. Allen, N.J. & Barres, B.A. Glia - more than just brain glue. *Nature* **457**, 675-677 (2009).
10. Duyckaerts, C., Delatour, B. & Potier, M.C. Classification and basic pathology of Alzheimer disease. *Acta Neuropathol* **118**, 5-36 (2009).
11. Jucker, M. & Walker, L.C. Self-propagation of pathogenic protein aggregates in neurodegenerative diseases. *Nature* **501**, 45-51 (2013).
12. Munoz, D.G. & Ferrer, I. Neuropathology of hereditary forms of frontotemporal dementia and parkinsonism. *Handbook of clinical neurology* **89**, 393-414 (2008).
13. Paulina V. Arriagada, J.H.G., E. Tessa Hedley-Whyte, Bradley T. Hyman,. Neurofibrillary tangles but not senile plaques parallel duration and severity of Alzheimer's disease. *Neurology* **42**, 631-9 (1992).
14. Naslund, J. *et al.* Correlation between elevated levels of amyloid beta-peptide in the brain and cognitive decline. *Jama-Journal of the American Medical Association* **283**, 1571-1577 (2000).
15. Lue, L.F. *et al.* Soluble amyloid beta peptide concentration as a predictor of synaptic change in Alzheimer's disease. *American Journal of Pathology* **155**, 853-862 (1999).
16. Jack, C.R., Jr. *et al.* Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurology* **9**, 119-128 (2010).
17. Schwarz, A.J. *et al.* Regional profiles of the candidate tau PET ligand18F-AV-1451 recapitulate key features of Braak histopathological stages. *Brain*, aww023 (2016).
18. Scholl, M. *et al.* PET Imaging of Tau Deposition in the Aging Human Brain. *Neuron* **89**, 971-82 (2016).
19. Ryman, D.C. *et al.* Symptom onset in autosomal dominant Alzheimer disease A systematic review and meta-analysis. *Neurology* **83**, 253-260 (2014).
20. Stern, Y. Cognitive reserve in ageing and Alzheimer's disease. *Lancet Neurology* **11**, 1006-1012 (2012).

21. Wingo, T.S., Lah, J.J., Levey, A.I. & Cutler, D.J. Autosomal recessive causes likely in early-onset Alzheimer disease. *Arch Neurol* **69**, 59-64 (2012).

22. Gatz, M. *et al.* Role of genes and environments for explaining Alzheimer disease. *Arch Gen Psychiatry* **63**, 168-74 (2006).

23. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-53 (2009).

24. Goate, A. *et al.* Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's-disease *Nature* **349**, 704-706 (1991).

25. Levylahad, E. *et al.* Candidate gene for the chromosome 1 familial Alzheimer's disease locus. *Science* **269**, 973-977 (1995).

26. Sherrington, R. *et al.* Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature* **375**, 754-760 (1995).

27. Finch, N. *et al.* Plasma progranulin levels predict progranulin mutation status in frontotemporal dementia patients and asymptomatic family members. *Brain* **132**, 583-591 (2009).

28. Jin, S.C. *et al.* Pooled-DNA sequencing identifies novel causative variants in PSEN1, GRN and MAPT in a clinical early-onset and familial Alzheimer's disease Ibero-American cohort. *Alzheimers Research & Therapy* **4**, 9 (2012).

29. Almeida, M.R. *et al.* Progranulin peripheral levels as a screening tool for the identification of subjects with progranulin mutations in a Portuguese cohort. *Neuro-degenerative diseases* **13**, 214-23 (2014).

30. Janssen, J.C. *et al.* Early-onset familial Alzheimer's disease: Mutation frequency in 31 families. *Neurobiology of Aging* **23**, S311-S311 (2002).

31. Cruts, M., Theuns, J. & Van Broeckhoven, C. Locus-specific mutation databases for neurodegenerative brain diseases. *Hum Mutat* **33**, 1340-4 (2012).

32. Farrer, L.A. *et al.* Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease - A meta-analysis. *Jama-Journal of the American Medical Association* **278**, 1349-1356 (1997).

33. Thal, D.R. *et al.* Two types of sporadic cerebral amyloid angiopathy. *Journal of Neuropathology and Experimental Neurology* **61**, 282-293 (2002).

34. Hixson, J.E. & Vernier, D.T. Restriction isotyping of human apolipoprotein E by gene amplification and cleavage with HhaI. *J Lipid Res* **31**, 545-8 (1990).

35. Persico, A.M. *et al.* Enhanced APOE2 transmission rates in families with autistic probands. *Psychiatric Genetics* **14**, 73-82 (2004).

36. Murrell, J.R. *et al.* The fourth Apolipoprotein E haplotype found in the Yoruba of Ibadan. *American Journal of Medical Genetics Part B-Neuropsychiatric Genetics* **141B**, 426-427 (2006).

37. Seripa, D. *et al.* The missing ApoE allele. *Annals of Human Genetics* **71**, 496-500 (2007).

38. Pericakvance, M.A. *et al.* Linkage studies in familial Alzheimer-disease - evidence for chromosome-19 linkage *American Journal of Human Genetics* **48**, 1034-1050 (1991).

39. Corder, E.H. *et al.* Gene dose of apolipoprotein E type-4 allele and the risk of Alzheimer's disease in late-onset families *Science* **261**, 921-923 (1993).

40. Corder, E.H. *et al.* Protective effect of apolipoprotein E type 2 allele for late-onset Alzheimer's disease *Nature Genetics* **7**, 180-184 (1994).

41. Poirier, J. *et al.* Apolipoprotein E polymorphism and Alzheimer's disease *Lancet* **342**, 697-699 (1993).

42. Chartierharlin, M.C. *et al.* Apolipoprotein E, ε4 allele as a major risk factor for sporadic early and late-onset forms of Alzheimer's disease: analysis of the 19q13.2 chromosomal region *Human Molecular Genetics* **3**, 569-574 (1994).

43. Dai, X.Y. *et al.* Association of apolipoprotein ε4 with sporadic Alzheimer's disease is more pronounced in early-onset type. *Neuroscience Letters* **175**, 74-76 (1994).

44. Kurz, A. *et al.* The apolipoprotein E epsilon-4 allele is a risk factor for Alzheimer's disease of early and late onset. *Nervenarzt* **65**, 775-779 (1994).

45. Okuizumi, K. *et al.* APOE-ε4 and early-onset Alzheimer's *Nature Genetics* **7**, 10-11 (1994).

46. Vanduijn, C.M. *et al.* Apolipoprotein E4 allele in a population-based study of early-onset Alzheimer's disease *Nature Genetics* **7**, 74-78 (1994).

47. Van Cauwenberghe, C., Van Broeckhoven, C. & Sleegers, K. The genetic landscape of Alzheimer disease: clinical implications and perspectives. *Genet Med* (2015).

48. Lambert, J.C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics* **45**, 1452-U206 (2013).

49. Jonsson, T. *et al.* A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. *Nature* **488**, 96-99 (2012).

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

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

52. Cruchaga, C. *et al.* Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. *Nature* **505**, 550-4 (2014).

53. Guerreiro, R.J. *et al.* Using Exome Sequencing to Reveal Mutations in TREM2 Presenting as a Frontotemporal Dementia-like Syndrome Without Bone Involvement. *Jama Neurology* **70**, 78-84 (2013).

54. Paloneva, J. *et al.* Mutations in two genes encoding different subunits of a receptor signaling complex result in an identical disease phenotype. *American Journal of Human Genetics* **71**, 656-662 (2002).

55. Kang, J. *et al.* The precursor of Alzheimer's-disease amyloid A4 protein resembles a cell-surface receptor *Nature* **325**, 733-736 (1987).

56. Xu, H.X., Greengard, P. & Gandy, S. Regulated formation of Golgi secretory vesicles containing Alzheimer beta-amyloid precursor protein *Journal of Biological Chemistry* **270**, 23243-23245 (1995).

57. Kinoshita, A. *et al.* Demonstration by FRET of BACE interaction with the amyloid precursor protein at the cell surface and in early endosomes. *Journal of Cell Science* **116**, 3339-3346 (2003).

58. Mizuguchi, M., Ikeda, K. & Kim, S.U. Differential distribution of cellular forms of beta-amyloid precursor protein in murine glial cell cultures *Brain Research* **584**, 219-225 (1992).

59. Haass, C., Kaether, C., Thinakaran, G. & Sisodia, S. Trafficking and proteolytic processing of APP. *Cold Spring Harb Perspect Med* **2**, a006270 (2012).

60. Querfurth, H.W. & LaFerla, F.M. Alzheimer's disease. *N Engl J Med* **362**, 329-44 (2010).

61. Heppner, F.L., Ransohoff, R.M. & Becher, B. Immune attack: the role of inflammation in Alzheimer disease. *Nature Reviews Neuroscience* **16**, 358-372 (2015).

62. Walsh, D.M. *et al.* Naturally secreted oligomers of amyloid beta protein potently inhibit hippocampal long-term potentiation in vivo. *Nature* **416**, 535-539 (2002).

63. Mawuenyega, K.G. *et al.* Decreased Clearance of CNS beta-Amyloid in Alzheimer's Disease. *Science* **330**, 1774-1774 (2010).

64. Huang, S.M. *et al.* Neprilysin-sensitive synapse-associated amyloid-beta peptide oligomers impair neuronal plasticity and cognitive function. *Journal of Biological Chemistry* **281**, 17941-17951 (2006).

65. Iwata, N. *et al.* Identification of the major A beta(1-42)-degrading catabolic pathway in brain parenchyma: Suppression leads to biochemical and pathological deposition. *Nature Medicine* **6**, 143-150 (2000).

66. Gasparini, L. *et al.* Stimulation of beta-amyloid precursor protein trafficking by insulin reduces intraneuronal beta-amyloid and requires mitogen-activated protein kinase signaling. *Journal of Neuroscience* **21**, 2561-2570 (2001).

67. Willette, A.A. *et al.* Association of Insulin Resistance With Cerebral Glucose Uptake in Late Middle-Aged Adults at Risk for Alzheimer Disease. *JAMA Neurol* **72**, 1013-20 (2015).

68. Sun, B. *et al.* Cystatin C-cathepsin B axis regulates amyloid beta levels and associated neuronal deficits in an animal model of Alzheimer's disease. *Neuron* **60**, 247-57 (2008).

69. Bertram, L., McQueen, M.B., Mullin, K., Blacker, D. & Tanzi, R.E. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat Genet* **39**, 17-23 (2007).

70. Rohn, T.T. The role of caspases in Alzheimer's disease; potential novel therapeutic opportunities. *Apoptosis* **15**, 1403-1409 (2010).

71. Grundkeiqbal, I. *et al.* Abnormal phosphorylation of the microtubule-associated protein T (tau) in Alzheimer cytoskeletal pathology *Proceedings of the National Academy of Sciences of the United States of America* **83**, 4913-4917 (1986).

72. Brion, J.P., Couck, A.M., Passareiro, E. & Flamentdurand, J. Neurofibrillary tangles of Alzheimer's-disease - an immunohistochemical study *Journal of Submicroscopic Cytology and Pathology* **17**, 89-96 (1985).

73. Gustke, N., Trinczek, B., Biernat, J., Mandelkow, E.M. & Mandelkow, E. Domains of tau-protein and interactions with microtubules *Biochemistry* **33**, 9511-9522 (1994).

74. Cleveland, D.W., Hwo, S.Y. & Kirschner, M.W. Physical and chemical properties of purified tau factor and role of tau in microtubule assembly. *Journal of Molecular Biology* **116**, 227-247 (1977).

75. Weingarten, M.D., Lockwood, A.H., Hwo, S.Y. & Kirschner, M.W. Protein factor essential for microtubule assembly *Proceedings of the National Academy of Sciences of the United States of America* **72**, 1858-1862 (1975).

76. Umeyama, T., Okabe, S., Kanai, Y. & Hirokawa, N. Dynamics of microtubules bundled by microtubule associated protein 2C (MAP2C) *Journal of Cell Biology* **120**, 451-465 (1993).

77. Chen, J., Kanai, Y., Cowan, N.J. & Hirokawa, N. Projection domains of map2 and tau determine spacings between microtubules in dendrites and axons *Nature* **360**, 674-676 (1992).

78. Dixit, R., Ross, J.L., Goldman, Y.E. & Holzbaur, E.L.F. Differential regulation of dynein and kinesin motor proteins by tau. *Science* **319**, 1086-1089 (2008).

79. Stamer, K., Vogel, R., Thies, E., Mandelkow, E. & Mandelkow, E.M. Tau blocks traffic of organelles, neurofilaments, and APP vesicles in neurons and enhances oxidative stress. *Journal of Cell Biology* **156**, 1051-1063 (2002).

80. Seitz, A. *et al.* Single-molecule investigation of the interference between kinesin, tau and MAP2c. *Embo Journal* **21**, 4896-4905 (2002).

81. Biernat, J., Gustke, N., Drewes, G., Mandelkow, E.M. & Mandelkow, E. Phosphorylation of Ser(262) strongly reduces binding of tau-protein to microtubules - distinction between PHF-like immunoreactivity and microtubule-binding *Neuron* **11**, 153-163 (1993).

82. Iqbal, K., Alonso, A.D. & Grundke-Iqbal, I. Cytosolic abnormally hyperphosphorylated tau but not paired helical filaments sequester normal MAPs and inhibit microtubule assembly. *Journal of Alzheimers Disease* **14**, 365-370 (2008).

83. Schneider, A., Biernat, J., von Bergen, M., Mandelkow, E. & Mandelkow, E.M. Phosphorylation that detaches tau protein from microtubules (Ser262, Ser214) also protects it against aggregation into Alzheimer paired helical filaments. *Biochemistry* **38**, 3549-3558 (1999).

84. Fletcher, D.A. & Mullins, D. Cell mechanics and the cytoskeleton. *Nature* **463**, 485-492 (2010).

85. Brandt, R., Hundelt, M. & Shahani, N. Tau alteration and neuronal degeneration in tauopathies: mechanisms and models. *Biochim Biophys Acta* **1739**, 331-54 (2005).

86. Barghorn, S. *et al.* Structure, microtubule interactions, and paired helical filament aggregation by tau mutants of frontotemporal dementias. *Biochemistry* **39**, 11714-11721 (2000).

87. Nelson, P.T. *et al.* Correlation of Alzheimer Disease Neuropathologic Changes With Cognitive Status: A Review of the Literature. *Journal of Neuropathology and Experimental Neurology* **71**, 362-381 (2012).

88. Khlistunova, I. *et al.* Inducible expression of tau repeat domain in cell models of tauopathy - Aggregation is toxic to cells but can be reversed by inhibitor drugs. *Journal of Biological Chemistry* **281**, 1205-1214 (2006).

89. Chevalier-Larsen, E. & Holzbaur, E.L. Axonal transport and neurodegenerative disease. *Biochim Biophys Acta* **1762**, 1094-108 (2006).

90. Koo, E.H. *et al.* Precursor of amyloid protein in Alzheimer-disease undergoes fast anterograde axonal-transport *Proceedings of the National Academy of Sciences of the United States of America* **87**, 1561-1565 (1990).

91. Lazarov, O., Lee, M., Peterson, D.A. & Sisodia, S.S. Evidence that synaptically released beta-amyloid accumulates as extracellular deposits in the hippocampus of transgenic mice. *Journal of Neuroscience* **22**, 9785-9793 (2002).

92. Pfrieger, F.W. & Ungerer, N. Cholesterol metabolism in neurons and astrocytes. *Prog Lipid Res* **50**, 357-71 (2011).

93. Liu, C.C., Kanekiyo, T., Xu, H. & Bu, G. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. *Nat Rev Neurol* **9**, 106-18 (2013).

94. Posse De Chaves, E.I., Rusinol, A.E., Vance, D.E., Campenot, R.B. & Vance, J.E. Role of lipoproteins in the delivery of lipids to axons during axonal regeneration. *Journal of Biological Chemistry* **272**, 30766-30773 (1997).

95. Fester, L. *et al.* Cholesterol-Promoted Synaptogenesis Requires the Conversion of Cholesterol to Estradiol in the Hippocampus. *Hippocampus* **19**, 692-705 (2009).

96. Goritz, C., Mauch, D.H. & Pfrieger, F.W. Multiple mechanisms mediate cholesterol-induced synaptogenesis in a CNS neuron. *Molecular and Cellular Neuroscience* **29**, 190-201 (2005).

97. Fassbender, K. *et al.* Simvastatin strongly reduces levels of Alzheimer's disease beta-amyloid peptides A beta 42 and A beta 40 in vitro and in vivo. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 5856-5861 (2001).

98. Yu, H. & Patel, S.B. Recent insights into the Smith-Lemli-Opitz syndrome. *Clinical Genetics* **68**, 383-391 (2005).

99. Madra, M. & Sturley, S.L. Niemann-Pick type C pathogenesis and treatment: from statins to sugars. *Clinical Lipidology* **5**, 387-395 (2010).

100. Ladu, M.J. *et al.* Isoform-specific binding of apolipoprotein E to beta-amyloid. *Journal of Biological Chemistry* **269**, 23403-23406 (1994).

101. Ehehalt, R., Keller, P., Haass, C., Thiele, C. & Simons, K. Amyloidogenic processing of the Alzheimer beta-amyloid precursor protein depends on lipid rafts. *Journal of Cell Biology* **160**, 113-123 (2003).

102. Simons, M. *et al.* Cholesterol depletion inhibits the generation of beta-amyloid in hippocampal neurons. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 6460-6464 (1998).

103. Wahrle, S. *et al.* Cholesterol-dependent gamma-secretase activity in buoyant cholesterol-rich membrane microdomains. *Neurobiology of Disease* **9**, 11-23 (2002).

104. Armstrong, R.A. The molecular biology of senile plaques and neurofibrillary tangles in Alzheimer's disease. *Folia Neuropathologica* **47**, 289-299 (2009).

105. Namba, Y., Tomonaga, M., Kawasaki, H., Otomo, E. & Ikeda, K. Apolipoprotein-e immunoreactivity in cerebral amyloid deposits and neurofibrillary tangles in alzheimers-disease and kuru plaque amyloid in creutzfeldt-jakob disease. *Brain Research* **541**, 163-166 (1991).

106. Foley, P. Lipids in Alzheimer's disease: A century-old story. *Biochim Biophys Acta* **1801**, 750-3 (2010).

107. Nicholson, A.M. & Ferreira, A. Increased Membrane Cholesterol Might Render Mature Hippocampal Neurons More Susceptible to beta-Amyloid-Induced Calpain Activation and Tau Toxicity. *Journal of Neuroscience* **29**, 4640-4651 (2009).

108. Fan, Q.W., Yu, W., Senda, T., Yanagisawa, K. & Michikawa, M. Cholesterol-dependent modulation of tau phosphorylation in cultured neurons. *Journal of Neurochemistry* **76**, 391-400 (2001).

109. Desnick, R.J. & Schuchman, E.H. Enzyme replacement and enhancement therapies: Lessons from lysosomal disorders. *Nature Reviews Genetics* **3**, 954-966 (2002).

110. Nielsen, E., Severin, F., Backer, J.M., Hyman, A.A. & Zerial, M. Rab5 regulates motility of early endosomes on microtubules. *Nature Cell Biology* **1**, 376-382 (1999).

111. Neefjes, J. & van der Kant, R. Stuck in traffic: an emerging theme in diseases of the nervous system. *Trends Neurosci* **37**, 66-76 (2014).

112. Straub, R.E. *et al.* Genetic variation in the 6p22.3 gene DTNBP1, the human ortholog of the mouse dysbindin gene, is associated with schizophrenia. *American Journal of Human Genetics* **71**, 337-348 (2002).

113. Schwab, S.G. *et al.* Support for association of schizophrenia with genetic variation in the 6p22.3 gene, dysbindin, in sib-pair families with linkage and in an additional sample of triad families. *American Journal of Human Genetics* **72**, 185-190 (2003).

114. Delevoye, C. *et al.* BLOC-1 Brings Together the Actin and Microtubule Cytoskeletons to Generate Recycling Endosomes. *Current Biology* **26**, 1-13 (2016).

115. Larimore, J. *et al.* The schizophrenia susceptibility factor dysbindin and its associated complex sort cargoes from cell bodies to the synapse. *Molecular Biology of the Cell* **22**, 4854-4867 (2011).

116. Li, Q., Spencer, N.Y., Pantazis, N.J. & Engelhardt, J.F. Alsin and SOD1(G93A) Proteins Regulate Endosomal Reactive Oxygen Species Production by Glial Cells and Proinflammatory Pathways Responsible for Neurotoxicity. *Journal of Biological Chemistry* **286**, 40151-40162 (2011).

117. Yang, Y. *et al.* The gene encoding alsin, a protein with three guanine-nucleotide exchange factor domains, is mutated in a form of recessive amyotrophic lateral sclerosis. *Nature Genetics* **29**, 160-165 (2001).

118. Otomo, A. *et al.* ALS2, a novel guanine nucleotide exchange factor for the small GTPase Rab5, is implicated in endosomal dynamics. *Human Molecular Genetics* **12**, 1671-1687 (2003).

119. LaFerla, F.M., Green, K.N. & Oddo, S. Intracellular amyloid-beta in Alzheimer's disease. *Nat Rev Neurosci* **8**, 499-509 (2007).

120. Lee, J.H. *et al.* Lysosomal Proteolysis and Autophagy Require Presenilin 1 and Are Disrupted by Alzheimer-Related PS1 Mutations. *Cell* **141**, 1146-U91 (2010).

121. Gyure, K.A., Durham, R., Stewart, W.F., Smialek, J.E. & Troncoso, J.C. Intraneuronal A beta-amyloid precedes development of amyloid plaques in Down syndrome. *Archives of Pathology & Laboratory Medicine* **125**, 489-492 (2001).

122. Mori, C. *et al.* Intraneuronal A beta 42 accumulation in Down syndrome brain. *Amyloid-Journal of Protein Folding Disorders* **9**, 88-102 (2002).

123. Chui, D.H. *et al.* Transgenic mice with Alzheimer presenilin 1 mutations show accelerated neurodegeneration without amyloid plaque formation. *Nature Medicine* **5**, 560-564 (1999).

124. Knobloch, M., Konietzko, U., Krebs, D.C. & Nitsch, R.M. Intracellular A beta and cognitive deficits precede beta-amyloid deposition in transgenic arcA beta mice. *Neurobiology of Aging* **28**, 1297-1306 (2007).

125. Kuo, Y.M. *et al.* The evolution of A beta peptide burden in the APP23 transgenic mice: Implications for A beta deposition in Alzheimer disease. *Molecular Medicine* **7**, 609-618 (2001).

126. Li, Q.X. *et al.* Intracellular accumulation of detergent-soluble amyloidogenic A beta fragment of Alzheimer's disease precursor protein in the hippocampus of aged transgenic mice. *Journal of Neurochemistry* **72**, 2479-2487 (1999).

127. Lord, A. *et al.* The Arctic Alzheimer mutation facilitates early intraneuronal A beta aggregation and senile plaque formation in transgenic mice. *Neurobiology of Aging* **27**, 67-77 (2006).

128. Oakley, H. *et al.* Intraneuronal beta-amyloid aggregates, neurodegeneration, and neuron loss in transgenic mice with five familial Alzheimer's disease mutations: Potential factors in amyloid plaque formation. *Journal of Neuroscience* **26**, 10129-10140 (2006).

129. Oddo, S. *et al.* Triple-transgenic model of Alzheimer's disease with plaques and tangles: Intracellular A beta and synaptic dysfunction. *Neuron* **39**, 409-421 (2003).

130. Wirths, O. *et al.* Intraneuronal A beta accumulation precedes plaque formation in beta-amyloid precursor protein and presenilin-1 double-transgenic mice. *Neuroscience Letters* **306**, 116-120 (2001).

131. Oddo, S., Caccamo, A., Smith, I.F., Green, K.N. & LaFerla, F.M. A dynamic relationship between intracellular and extracellular pools of A beta. *American Journal of Pathology* **168**, 184-194 (2006).

132. Hirai, K. *et al.* Mitochondrial abnormalities in Alzheimer's disease. *Journal of Neuroscience* **21**, 3017-3023 (2001).

133. Che, S. *et al.* Regional selectivity of rab5 and rab7 protein up regulation in mild cognitive impairment (MCI) and Alzheimer's disease (AD). *Society for Neuroscience Abstract Viewer and Itinerary Planner* **40**(2010).

134. Halliwell, B. Oxidative stress and neurodegeneration: where are we now? *Journal of Neurochemistry* **97**, 1634-1658 (2006).

135. Emerit, J., Edeas, A. & Bricaire, F. Neurodegenerative diseases and oxidative stress. *Biomedicine & Pharmacotherapy* **58**, 39-46 (2004).

136. Martins, R.N., Harper, C.G., Stokes, G.B. & Masters, C.L. Increased cerebral glucose-6-phosphate-dehydrogenase activity in Alzheimer's disease may reflect oxidative stress. *Journal of Neurochemistry* **46**, 1042-1045 (1986).

137. Sayre, L.M. *et al.* 4-hydroxynonenal-derived advanced lipid peroxidation end products are increased in Alzheimer's disease. *Journal of Neurochemistry* **68**, 2092-2097 (1997).

138. Prasad, M.R., Lovell, M.A., Yatin, M., Dhillon, H. & Markesbery, W.R. Regional membrane phospholipid alterations in Alzheimer's disease. *Neurochemical Research* **23**, 81-88 (1998).

139. Gabbita, S.P., Lovell, M.A. & Markesbery, W.R. Increased nuclear DNA oxidation in the brain in Alzheimer's disease. *Journal of Neurochemistry* **71**, 2034-2040 (1998).

140. Mecocci, P., Macgarvey, U. & Beal, M.F. Oxidative damage to mitochondrial DNA is increased in Alzheimer's disease. *Annals of Neurology* **36**, 747-751 (1994).

141. Smith, C.D. *et al.* Excess brain protein oxidation and enzyme dysfunction in normal aging and in Alzheimer disease. *Proceedings of the National Academy of Sciences of the United States of America* **88**, 10540-10543 (1991).

142. Smith, M.A. *et al.* Carbonyl-related posttranslational modification of neurofilament protein in the neurofibrillary pathology of Alzheimer's disease. *Journal of Neurochemistry* **64**, 2660-2666 (1995).

143. Nunomura, A. *et al.* Oxidative damage is the earliest event in Alzheimer disease. *Journal of Neuropathology and Experimental Neurology* **60**, 759-767 (2001).

144. Fraunberger, E.A., Scola, G., Laliberte, V.L.M., Duong, A. & Andreazza, A.C. Redox Modulations, Antioxidants, and Neuropsychiatric Disorders. *Oxidative Medicine and Cellular Longevity*, 14 (2016).

145. Lange, M.L.B. *et al.* Loss of phospholipid asymmetry and elevated brain apoptotic protein levels in subjects with amnestic mild cognitive impairment and Alzheimer disease. *Neurobiology of Disease* **29**, 456-464 (2008).

146. Mattson, M.P., Fu, W.M., Waeg, G. & Uchida, K. 4-hydroxynonenal, a product of lipid peroxidation, inhibits dephosphorylation of the microtubule-associated protein tau. *Neuroreport* **8**, 2275-2281 (1997).

147. Mark, R.J., Lovell, M.A., Markesbery, W.R., Uchida, K. & Mattson, M.P. A role for 4-hydroxynonenal, an aldehydic product of lipid peroxidation, in disruption of ion homeostasis and neuronal death induced by amyloid beta-peptide. *Journal of Neurochemistry* **68**, 255-264 (1997).

148. Keller, J.N. *et al.* 4-hydroxynonenal, an aldehydic product of membrane lipid peroxidation, impairs glutamate transport and mitochondrial function in synaptosomes. *Neuroscience* **80**, 685-696 (1997).

149. Owen, J.B. *et al.* Oxidative modification to LDL receptor-related protein 1 in hippocampus from subjects with Alzheimer disease Implications for A beta accumulation in AD brain. *Free Radical Biology and Medicine* **49**, 1798-1803 (2010).

150. Gibson, G.E., Sheu, K.F.R. & Blass, J.P. Abnormalities of mitochondrial enzymes in Alzheimer disease. *Journal of Neural Transmission* **105**, 855-870 (1998).

151. Smith, D.G., Cappai, R. & Barnham, K.J. The redox chemistry of the Alzheimer's disease amyloid beta peptide. *Biochimica Et Biophysica Acta-Biomembranes* **1768**, 1976-1990 (2007).

152. Nair, N.G., Perry, G., Smith, M.A. & Reddy, V.P. NMR Studies of Zinc, Copper, and Iron Binding to Histidine, the Principal Metal Ion Complexing Site of Amyloid-beta Peptide. *Journal of Alzheimers Disease* **20**, 57-66 (2010).

153. Arispe, N., Pollard, H.B. & Rojas, E. Zn2+ interaction with Alzheimer amyloid beta protein calcium channels. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 1710-1715 (1996).

154. Arispe, N., Pollard, H.B. & Rojas, E. Giant multilevel cation channels formed by Alzheimer disease amyloid β-protein [AβP-(1-40)] in bilayer membranes. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 10573-10577 (1993).

155. Opazo, C. *et al.* Metalloenzyme-like activity of Alzheimer's disease beta-amyloid - Cu-dependent catalytic conversion of dopamine, cholesterol, and biological reducing agents to neurotoxic H2O2. *Journal of Biological Chemistry* **277**, 40302-40308 (2002).

156. Lovell, M.A., Robertson, J.D., Teesdale, W.J., Campbell, J.L. & Markesbery, W.R. Copper, iron and zinc in Alzheimer's disease senile plaques. *Journal of the Neurological Sciences* **158**, 47-52 (1998).

157. Cuajungco, M.P. & Faget, K.Y. Zinc takes the center stage: its paradoxical role in Alzheimer's disease. *Brain Research Reviews* **41**, 44-56 (2003).

158. Focht, S.J. *et al.* Regional distribution of iron, transferrin, ferritin, and oxidatively-modified proteins in young and aged Fischer 344 rat brains. *Neuroscience* **79**, 255-261 (1997).

159. Smith, M.A., Harris, P.L.R., Sayre, L.M. & Perry, G. Iron accumulation in Alzheimer disease is a source of redox-generated free radicals. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 9866-9868 (1997).

160. Yamamoto, A. *et al.* Iron (III) induces aggregation of hyperphosphorylated tau and its reduction to iron (II) reverses the aggregation: implications in the formation of neurofibrillary tangles of Alzheimer's disease. *Journal of Neurochemistry* **82**, 1137-1147 (2002).

161. Scott, C.W., Fieles, A., Sygowski, L.A. & Caputo, C.B. Aggregation of tau-protein by aluminum *Brain Research* **628**, 77-84 (1993).

162. Yang, L.S. & Ksiezak-Reding, H. Ca2+ and Mg2+ selectively induce aggregates of PHF-tau but not normal human tau. *Journal of Neuroscience Research* **55**, 36-43 (1999).

163. van Leeuwen, L.A. & Hoozemans, J.J. Physiological and pathophysiological functions of cell cycle proteins in post-mitotic neurons: implications for Alzheimer's disease. *Acta Neuropathol* **129**, 511-25 (2015).

164. Ogawa, O. *et al.* Increased p27, an essential component of cell cycle control, in Alzheimer's disease. *Aging Cell* **2**, 105-110 (2003).

165. Driver, J.A., Zhou, X.Z. & Lu, K.P. Pin1 dysregulation helps to explain the inverse association between cancer and Alzheimer's disease. *Biochimica Et Biophysica Acta-General Subjects* **1850**, 2069-2076 (2015).

166. Zhu, X.W. *et al.* Activation of p38 kinase links tau phosphorylation, oxidative stress, and cell cycle-related events in Alzheimer disease. *Journal of Neuropathology and Experimental Neurology* **59**, 880-888 (2000).

167. Zhu, X.W. *et al.* Neuronal CDK7 in hippocampus is related to aging and Alzheimer disease. *Neurobiology of Aging* **21**, 807-813 (2000).

168. Schmetsdorf, S., Arnold, E., Holzer, M., Arendt, T. & Gaertner, U. A putative role for cell cycle-related proteins in microtubule-based neuroplasticity. *European Journal of Neuroscience* **29**, 1096-1107 (2009).

169. Korulu, S., Yildiz-Unal, A., Yuksel, M. & Karabay, A. Protein kinase C activation causes neurite retraction via cyclinD1 and p60-katanin increase in rat hippocampal neurons. *European Journal of Neuroscience* **37**, 1610-1619 (2013).

170. Odajima, J. *et al.* Cyclin E Constrains Cdk5 Activity to Regulate Synaptic Plasticity and Memory Formation. *Developmental Cell* **21**, 655-668 (2011).

171. Cheung, Z.H. & Ip, N.Y. Cdk5: a multifaceted kinase in neurodegenerative diseases. *Trends in Cell Biology* **22**, 169-175 (2012).

172. Hernandez-Ortega, K. & Arias, C. ERK activation and expression of neuronal cell cycle markers in the hippocampus after entorhinal cortex lesion. *Journal of Neuroscience Research* **90**, 2116-2126 (2012).

173. Hernandez-Ortega, K., Ferrera, P. & Arias, C. Sequential expression of cell-cycle regulators and Alzheimer's disease-related proteins in entorhinal cortex after hippocampal excitotoxic damage. *Journal of Neuroscience Research* **85**, 1744-1751 (2007).

174. Tomashevski, A., Webster, D.R., Grammas, P., Gorospe, M. & Kruman, I.I. Cyclin-C-dependent cell-cycle entry is required for activation of non-homologous end joining DNA repair in postmitotic neurons. *Cell Death and Differentiation* **17**, 1189-1198 (2010).

175. Schwartz, E.I. *et al.* Cell cycle activation in postmitotic neurons is essential for DNA repair. *Cell Cycle* **6**, 318-329 (2007).

176. Casafont, I., Palanca, A., Lafarga, V., Berciano, M.T. & Lafarga, M. Effect of ionizing radiation in sensory ganglion neurons: organization and dynamics of nuclear compartments of DNA damage/repair and their relationship with transcription and cell cycle. *Acta Neuropathologica* **122**, 481-493 (2011).

177. Rimkus, S.A. *et al.* Mutations in String/CDC25 inhibit cell cycle re-entry and neurodegeneration in a Drosophila model of Ataxia telangiectasia. *Genes & Development* **22**, 1205-1220 (2008).

178. Lee, H.G. *et al.* The Neuronal Expression of MYC Causes a Neurodegenerative Phenotype in a Novel Transgenic Mouse. *American Journal of Pathology* **174**, 891-897 (2009).

179. Feddersen, R.M., Ehlenfeldt, R., Yunis, W.S., Clark, H.B. & Orr, H.T. Disrupted cerebellar cortical development and progressive degeneration of purkinje cells in sv40-t antigen transgenic mice *Neuron* **9**, 955-966 (1992).

180. Bao, L. *et al.* Prevalent overexpression of prolyl isomerase Pin1 in human cancers. *American Journal of Pathology* **164**, 1727-1737 (2004).

181. Holzer, M. *et al.* Inverse association of Pin1 and tau accumulation in Alzheimer's disease hippocampus. *Acta Neuropathologica* **104**, 471-481 (2002).

182. Ryo, A., Nakamura, M., Wulf, G., Liou, Y.C. & Lu, K.P. Pin 1 regulates turnover and subcellular localization of beta-catenin by inhibiting its interaction with APC. *Nature Cell Biology* **3**, 793-801 (2001).

183. Wulf, G.M., Liou, Y.C., Ryo, A., Lee, S.W. & Lu, K.P. Role of Pin1 in the regulation of p53 stability and p21 transactivation, and cell cycle checkpoints in response to DNA damage. *Journal of Biological Chemistry* **277**, 47976-47979 (2002).

184. Zacchi, P. *et al.* The prolyl isomerase Pin1 reveals a mechanism to control p53 functions after genotoxic insults. *Nature* **419**, 853-857 (2002).

185. Zheng, H.W. *et al.* The prolyl isomerase Pin1 is a regulator of p53 in genotoxic response. *Nature* **419**, 849-853 (2002).

186. Zhou, X.Z. *et al.* Pin1-dependent prolyl isomerization regulates dephosphorylation of Cdc25C and tau proteins. *Molecular Cell* **6**, 873-883 (2000).

187. Lu, P.J., Wulf, G., Zhou, X.Z., Davies, P. & Lu, K.P. The prolyl isomerase Pin1 restores the function of Alzheimer-associated phosphorylated tau protein. *Nature* **399**, 784-788 (1999).

188. Hamdane, M. *et al.* Pin1 allows for differential Tau dephosphorylation in neuronal cells. *Molecular and Cellular Neuroscience* **32**, 155-160 (2006).

189. Pastorino, L. *et al.* The prolyl isomerase Pin1 regulates amyloid precursor protein processing and amyloid-beta production. *Nature* **440**, 528-534 (2006).

190. Brenkman, A.B. *et al.* The peptidyl-isomerase Pin1 regulates p27(kip1) expression through inhibition of Forkhead box O tumor suppressors. *Cancer Research* **68**, 7597-7605 (2008).

191. Ayala, A., Munoz, M.F. & Arguelles, S. Lipid peroxidation: production, metabolism, and signaling mechanisms of malondialdehyde and 4-hydroxy-2-nonenal. *Oxidative medicine and cellular longevity* **2014**, 360438 (2014).

192. Giorgio, M., Trinei, M., Migliaccio, E. & Pelicci, P.G. Hydrogen peroxide: a metabolic by-product or a common mediator of ageing signals? *Nature Reviews Molecular Cell Biology* **8**, 722A-728 (2007).

193. Monje, M.L., Toda, H. & Palmer, T.D. Inflammatory blockade restores adult hippocampal neurogenesis. *Science* **302**, 1760-1765 (2003).

194. Sastre, M. *et al.* Nonsteroidal anti-inflammatory drugs and peroxisome proliferator-activated receptor-gamma agonists modulate immunostimulated processing of amyloid precursor protein through regulation of beta-secretase. *Journal of Neuroscience* **23**, 9796-9804 (2003).

195. Bas, A. *et al.* Nonsteroidal antiinflammatory drugs and the risk of Alzheimer's disease. *New England Journal of Medicine* **345**, 1515-1521 (2001).

196. Brosseron, F., Krauthausen, M., Kummer, M. & Heneka, M.T. Body Fluid Cytokine Levels in Mild Cognitive Impairment and Alzheimer's Disease: a Comparative Overview. *Molecular Neurobiology* **50**, 534-544 (2014).

197. Tarkowski, E., Andreasen, N., Tarkowski, A. & Blennow, K. Intrathecal inflammation precedes development of Alzheimer's disease. *Journal of Neurology Neurosurgery and Psychiatry* **74**, 1200-1205 (2003).

198. Lue, L.F., Brachova, L., Civin, W.H. & Rogers, J. Inflammation, A beta deposition, and neurofibrillary tangle formation as correlates of Alzheimer's disease neurodegeneration. *Journal of Neuropathology and Experimental Neurology* **55**, 1083-1088 (1996).

199. Griffin, W.S.T. *et al.* Brain interleukin 1 and S-100 immunoreactivity are elevated in down syndrome and Alzheimer's disease. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 7611-7615 (1989).

200. McGeer, P.L., Itagaki, S., Boyes, B.E. & McGeer, E.G. Reactive microglia are positive for HLA-DR in the substantia nigra of Parkinson's and Alzheimer's disease brains *Neurology* **38**, 1285-1291 (1988).

201. Hohsfield, L.A. & Humpel, C. Migration of blood cells to β-amyloid plaques in Alzheimer's disease. *Experimental Gerontology* **65**, 8-15 (2015).

202. Maezawa, I., Zimin, P.I., Wulff, H. & Jin, L.W. Amyloid-beta protein oligomer at low nanomolar concentrations activates microglia and induces microglial neurotoxicity. *J Biol Chem* **286**, 3693-706 (2011).

203. Krabbe, G. *et al.* Functional Impairment of Microglia Coincides with Beta-Amyloid Deposition in Mice with Alzheimer-Like Pathology. *Plos One* **8**, 8 (2013).

204. Stewart, C.R. *et al.* CD36 ligands promote sterile inflammation through assembly of a Toll-like receptor 4 and 6 heterodimer. *Nat Immunol* **11**, 155-61 (2010).

205. Fassbender, K. *et al.* The LPS receptor (CD14) links innate immunity with Alzheimer's disease. *Faseb Journal* **18**, 203-205 (2004).

206. Koenigsknecht, J. & Landreth, G. Microglial phagocytosis of fibrillar beta-amyloid through a beta(1) integrin-dependent mechanism. *Journal of Neuroscience* **24**, 9838-9846 (2004).

207. Sheedy, F.J. *et al.* CD36 coordinates NLRP3 inflammasome activation by facilitating intracellular nucleation of soluble ligands into particulate ligands in sterile inflammation. *Nature Immunology* **14**, 812-+ (2013).

208. Paresce, D.M., Ghosh, R.N. & Maxfield, F.R. Microglial cells internalize aggregates of the Alzheimer's disease amyloid beta-protein via a scavenger receptor. *Neuron* **17**, 553-565 (1996).

209. Bamberger, M.E., Harris, M.E., McDonald, D.R., Husemann, J. & Landreth, G.E. A cell surface receptor complex for fibrillar beta-amyloid mediates microglial activation. *Journal of Neuroscience* **23**, 2665-2674 (2003).

210. ElKhoury, J. *et al.* Scavenger receptor-mediated adhesion of microglia to beta-amyloid fibrils. *Nature* **382**, 716-719 (1996).

211. Du Yan, S. *et al.* Amyloid-beta peptide-receptor for advanced glycation endproduct interaction elicits neuronal expression of macrophage-colony stimulating factor: a proinflammatory pathway in Alzheimer disease. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 5296-301 (1997).

212. Nalls, M.A. *et al.* NeuroX, a fast and efficient genotyping platform for investigation of neurodegenerative diseases. *Neurobiol Aging* **36**, 1605 e7-12 (2015).

213. Page, C.M. *et al.* Assessing the Power of Exome Chips. *Plos One* **10**(2015).

214. Cortes, A. & Brown, M.A. Promise and pitfalls of the Immunochip. *Arthritis Research & Therapy* **13**, 3 (2011).

215. Parkes, M., Cortes, A., van Heel, D.A. & Brown, M.A. Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat Rev Genet* **14**, 661-73 (2013).

216. Landi, S., Gemignani, F., Gioia-Patricola, L., Chabrier, A. & Canzian, F. Evaluation of a microarray for genotyping polymorphisms related to xenobiotic metabolism and DNA repair. *Biotechniques* **35**, 816-+ (2003).

217. Nalls, M.A. *et al.* Baseline genetic associations in the Parkinson's Progression Markers Initiative (PPMI). *Movement Disorders* **31**, 79-85 (2016).

218. Zhang, M. *et al.* Mutation analysis of CHCHD2 in Canadian patients with familial Parkinson's disease. *Neurobiology of Aging* **38**(2016).

219. Simon-Sanchez, J., Heutink, P., Gasser, T. & Ipdgc. Variation in PARK10 is not associated with risk and age at onset of Parkinson's disease in large clinical cohorts. *Neurobiology of Aging* **36**(2015).

220. Ghani, M. *et al.* Mutation analysis of patients with neurodegenerative disorders using NeuroX array. *Neurobiol Aging* **36**, 545 e9-14 (2015).

221. Hernández Bustos, O.E. & Hernández Bustos, L. Relación entre conductancias de sodio y potasio al estilo Hodgkin y Huxley y su efecto en la propagación de potenciales de acción a 40°C. *Revista Salud Uninorte* **27**, 223-235 (2011).

222. Carrasquillo, M.M. *et al.* Evaluating pathogenic dementia variants in posterior cortical atrophy. *Neurobiology of Aging* **37**, 38-44 (2016).

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

224. Sassi, C. *et al.* Exome sequencing identifies 2 novel presenilin 1 mutations (p.L166V and p.S230R) in British early-onset Alzheimer's disease. *Neurobiol Aging* **35**, 2422 e13-6 (2014).

225. Nicolas, G. *et al.* Screening of dementia genes by whole-exome sequencing in early-onset Alzheimer disease: input and lessons. *European Journal of Human Genetics* **24**, 710-716 (2016).

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

227. Guo, Y. *et al.* Illumina human exome genotyping array clustering and quality control. *Nat Protoc* **9**, 2643-62 (2014).

228. Periquet, M. *et al.* Parkin mutations are frequent in patients with isolated early-onset parkinsonism. *Brain* **126**, 1271-1278 (2003).

229. Amboni, M. *et al.* Cerebellar and Pyramidal Dysfunctions, Palpebral Ptosis and Weakness As Presenting Symptoms of PARK-2. *Movement Disorders* **24**, 303-305 (2009).

230. Deng, H. *et al.* Heterogeneous phenotype in a family with compound heterozygous parkin gene mutations. *Archives of Neurology* **63**, 273-277 (2006).

231. Sironi, F. *et al.* Parkin analysis in early onset Parkinson's disease. *Parkinsonism & Related Disorders* **14**, 326-333 (2008).

232. Madegowda, R.H., Kishore, A. & Anand, A. Mutational screening of the parkin gene among South Indians with early onset Parkinson's disease. *J Neurol Neurosurg Psychiatry* **76**, 1588-90 (2005).

233. Camargos, S.T. *et al.* Familial Parkinsonism and early onset Parkinson's disease in a Brazilian movement disorders clinic: phenotypic characterization and frequency of SNCA, PRKN, PINK1, and LRRK2 mutations. *Mov Disord* **24**, 662-6 (2009).

234. Schlitter, A.M. *et al.* Parkin gene variations in late-onset Parkinson's disease: comparison between Norwegian and German cohorts. *Acta Neurol Scand* **113**, 9-13 (2006).

235. Bras, J. *et al.* Analysis of Parkinson disease patients from Portugal for mutations in SNCA, PRKN, PINK1 and LRRK2. *BMC Neurol* **8**, 1 (2008).

236. Abbas, N. *et al.* A wide variety of mutations in the parkin gene are responsible for autosomal recessive parkinsonism in Europe. *Human Molecular Genetics* **8**, 567-574 (1999).

237. Hedrich, K. *et al.* Evaluation of 50 probands with early-onset Parkinson's disease for Parkin mutations. *Neurology* **58**, 1239-1246 (2002).

238. Lesage, S. *et al.* Rare heterozygous parkin variants in French early-onset Parkinson disease patients and controls. *J Med Genet* **45**, 43-6 (2008).

239. Guo, J.F. *et al.* Mutation analysis of Parkin, PINK1, DJ-1 and ATP13A2 genes in Chinese patients with autosomal recessive early-onset Parkinsonism. *Mov Disord* **23**, 2074-9 (2008).

240. Lohmann, E. *et al.* A multidisciplinary study of patients with early-onset PD with and without parkin mutations. *Neurology* **72**, 110-116 (2009).

241. Koziorowski, D. *et al.* Low frequency of the PARK2 gene mutations in Polish patients with the early-onset form of Parkinson disease. *Parkinsonism Relat Disord* **16**, 136-8 (2010).

242. Guo, J.F. *et al.* Mutation analysis of Parkin, PINK1 and DJ-1 genes in Chinese patients with sporadic early onset parkinsonism. *J Neurol* **257**, 1170-5 (2010).

243. Illarioshkin, S.N. *et al.* Mutation analysis of the parkin gene in Russian families with autosomal recessive juvenile parkinsonism. *Movement Disorders* **18**, 914-919 (2003).

244. Scherfler, C. *et al.* Striatal and cortical pre- and postsynaptic dopaminergic dysfunction in sporadic parkin-linked parkinsonism. *Brain* **127**, 1332-1342 (2004).

245. Brooks, J. *et al.* Parkin and PINK1 mutations in early-onset Parkinson's disease: comprehensive screening in publicly available cases and control. *J Med Genet* **46**, 375-81 (2009).

246. Bruggemann, N. *et al.* Frequency of heterozygous Parkin mutations in healthy subjects: need for careful prospective follow-up examination of mutation carriers. *Parkinsonism Relat Disord* **15**, 425-9 (2009).

247. International Schizophrenia, C. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-52 (2009).

248. Moskvina, V. *et al.* Analysis of Genome-Wide Association Studies of Alzheimer Disease and of Parkinson Disease to Determine If These 2 Diseases Share a Common Genetic Risk. *Jama Neurology* **70**, 1268-1276 (2013).

249. Blais, J. *et al.* Risk of Misdiagnosis Due to Allele Dropout and False-Positive PCR Artifacts in Molecular Diagnostics: Analysis of 30,769 Genotypes. *J Mol Diagn* **17**, 505-14 (2015).

250. Liu, F. & Gong, C.X. Tau exon 10 alternative splicing and tauopathies. *Molecular Neurodegeneration* **3**, 10 (2008).

251. Pittman, A.M., Fung, H.-C. & de Silva, R. Untangling the tau gene association with neurodegenerative disorders. *Human Molecular Genetics* **15**, R188-R195 (2006).

252. Schulte, E.C. *et al.* Rare variants in beta-Amyloid precursor protein (APP) and Parkinson's disease. *European Journal of Human Genetics* **23**, 1328-1333 (2015).

253. Schulte, E.C. *et al.* Variants in eukaryotic translation initiation factor 4G1 in sporadic Parkinson's disease. *Neurogenetics* **13**, 281-285 (2012).

254. Frigerio, C.S. *et al.* On the identification of low allele frequency mosaic mutations in the brains of Alzheimer's disease patients. *Alzheimers & Dementia* **11**, 1265-1276 (2015).

255. Coppola, G. *et al.* Evidence for a role of the rare p.A152T variant in MAPT in increasing the risk for FTD-spectrum and Alzheimer's diseases. *Hum Mol Genet* **21**, 3500-12 (2012).

256. Fong, H. *et al.* Genetic Correction of Tauopathy Phenotypes in Neurons Derived from Human Induced Pluripotent Stem Cells. *Stem Cell Reports* **1**, 226-234 (2013).

257. Illenberger, S. *et al.* The endogenous and cell cycle-dependent phosphorylation of tau protein in living cells: Implications for Alzheimer's disease. *Molecular Biology of the Cell* **9**, 1495-1512 (1998).

258. Mukrasch, M.D. *et al.* Structural Polymorphism of 441-Residue Tau at Single Residue Resolution. *Plos Biology* **7**, 399-414 (2009).

259. Kocher, O. & Krieger, M. Role of the adaptor protein PDZK1 in controlling the HDL receptor SR-BI. *Current Opinion in Lipidology* **20**, 236-241 (2009).

260. Thanopoulou, K., Fragkouli, A., Stylianopoulou, F. & Georgopoulos, S. Scavenger receptor class B type I (SR-BI) regulates perivascular macrophages and modifies amyloid pathology in

an Alzheimer mouse model. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 20816-20821 (2010).

261. Polasek, O. *et al.* Common Variants in SLC17A3 Gene Affect Intra-personal Variation in Serum Uric Acid Levels in Longitudinal Time Series. *Croatian Medical Journal* **51**, 32-39 (2010).

262. Gunjaca, G. *et al.* Predictive Value of 8 Genetic Loci for Serum Uric Acid Concentration. *Croatian Medical Journal* **51**, 23-31 (2010).

263. Cooper, G.M. *et al.* A copy number variation morbidity map of developmental delay. *Nature Genetics* **43**, 838-U44 (2011).

264. Walter, M.J. *et al.* Acquired copy number alterations in adult acute myeloid leukemia genomes. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 12950-12955 (2009).

265. Forbes, S.A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research* **43**, D805-D811 (2015).

266. Yamamoto-Sasaki, M., Ozawa, H., Saito, T., Rosler, M. & Riederer, P. Impaired phosphorylation of cyclic AMP response element binding protein in the hippocampus of dementia of the Alzheimer type. *Brain Research* **824**, 300-303 (1999).

267. Harpin, M.L. *et al.* Glial fibrillary acidic protein and beta-a4 protein deposits in temporal-lobe of aging brain and senile dementia of the Alzheimer type - relation with the cognitive state and with quantitative studies of senile plaques and neurofibrillary tangles *Journal of Neuroscience Research* **27**, 587-594 (1990).

268. Casado, A., Lopez-Fernandez, M.E., Casado, M.C. & de la Torre, R. Lipid peroxidation and antioxidant enzyme activities in vascular and Alzheimer dementias. *Neurochemical Research* **33**, 450-458 (2008).

269. Stefanits, H., Wesseling, C. & Kovacs, G.G. Loss of Calbindin immunoreactivity in the dentate gyrus distinguishes Alzheimer's disease from other neurodegenerative dementias. *Neuroscience Letters* **566**, 137-141 (2014).

270. Burns, R. *et al.* Homozygous splice mutation in CWF19L1 in a Turkish family with recessive ataxia syndrome. *Neurology* **83**, 2175-2182 (2014).

271. DeCaen, P.G., Delling, M., Vien, T.N. & Clapham, D.E. Direct recording and molecular identification of the calcium channel of primary cilia. *Nature* **504**, 315-+ (2013).

272. Grupe, A. *et al.* A scan of chromosome 10 identifies a novel locus showing strong association with late-onset Alzheimer disease. *American Journal of Human Genetics* **78**, 78-88 (2006).

273. Huang, A.L. *et al.* The cells and logic for mammalian sour taste detection. *Nature* **442**, 934-938 (2006).

274. Deban, L., Bottazzi, B., Garlanda, C., de la Torre, Y.M. & Mantovani, A. Pentraxins: Multifunctional proteins at the interface of innate immunity and inflammation. *Biofactors* **35**, 138-145 (2009).

275. Pearson, T.A. *et al.* Markers of inflammation and cardiovascular disease application to clinical and public health practice - A statement for healthcare professionals from the centers for disease control and prevention and the American Heart Association. *Circulation* **107**, 499-511 (2003).

276. Warren, J.E., Sandhu, A., Thompson, P.D. & Blumbergs, P.C. Dementia in a case of Behcet's disease. *Internal Medicine Journal* **36**, 611-616 (2006).

277. Tennent, G.A., Lovat, L.B. & Pepys, M.B. Serum amyloid P component prevents proteolysis of the amyloid fibrils of Alzheimer disease and systemic amyloidosis. *Proceedings of the National Academy of Sciences of the United States of America* **92**, 4299-4303 (1995).

278. Nishiyama, E., Iwamoto, N., Kimura, M. & Arai, H. Serum amyloid P component level in Alzheimer's disease. *Dementia* **7**, 256-259 (1996).

279. Kimura, M. *et al.* Assessment of cerebrospinal fluid levels of serum amyloid P component in patients with Alzheimer's disease. *Neuroscience Letters* **273**, 137-139 (1999).

280. Sato, K. *et al.* PLEKHG2/FLJ00018, a Rho family-specific guanine nucleotide exchange factor, is tyrosine phosphorylated via the EphB2/cSrc signaling pathway. *Cellular Signalling* **26**, 691-696 (2014).

281. Pchelina, S.N., Emelyanov, A.K. & Usenko, T.S. Molecular basis of Parkinsons's disease linked to LRRK2 mutations. *Molecular Biology* **48**, 1-10 (2014).

282. Pradotto, L., Azan, G., Doriguzzi, C., Valentini, C. & Mauro, A. Sporadic vascular dementia as clinical presentation of a new missense mutation within exon 7 of NOTCH3 gene. *Journal of the Neurological Sciences* **271**, 207-210 (2008).

283. Lleo, A. Activity of gamma-secretase on substrates other than APP. *Current Topics in Medicinal Chemistry* **8**, 9-16 (2008).

284. Berezovska, O., Xia, M.Q. & Hyman, B.T. Notch is expressed in adult brain, is coexpressed with presenilin-1, and is altered in Alzheimer disease. *Journal of Neuropathology and Experimental Neurology* **57**, 738-745 (1998).

285. Lalla, R. & Gonmez, G. The role of sirtuins in Alzheimer's disease. *Frontiers in Aging Neuroscience* **5**, 4 (2013).

286. Julien, C. *et al.* Sirtuin 1 Reduction Parallels the Accumulation of Tau in Alzheimer Disease. *Journal of Neuropathology and Experimental Neurology* **68**, 48-58 (2009).

287. Wang, C.-H. *et al.* Thrombospondin Type I Domain Containing 7A (THSD7A) Mediates Endothelial Cell Migration and Tube Formation. *Journal of Cellular Physiology* **222**, 685-694 (2010).

288. Kimura, T., Ishiguro, K. & Hisanaga, S.-i. Physiological and pathological phosphorylation of tau by Cdk5. *Frontiers in Molecular Neuroscience* **7**(2014).

289. Jehle, A.W. *et al.* ATP-binding cassette transporter A7 enhances phagocytosis of apoptotic cells and associated ERK signaling in macrophages. *Journal of Cell Biology* **174**, 547-556 (2006).

290. Zlokovic, B.V., Deane, R., Sagare, A.P., Bell, R.D. & Winkler, E.A. Low-density lipoprotein receptor-related protein-1: a serial clearance homeostatic mechanism controlling Alzheimer's amyloid beta-peptide elimination from the brain. *Journal of Neurochemistry* **115**, 1077-1089 (2010).

291. Castellano, J.M. *et al.* Low-density lipoprotein receptor overexpression enhances the rate of brain-to-blood A beta clearance in a mouse model of beta-amyloidosis. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 15502-15507 (2012).

292. Tian, Y., Bustos, V., Flajolet, M. & Greengard, P. A small-molecule enhancer of autophagy decreases levels of A beta and APP-CTF via Atg5-dependent autophagy pathway. *Faseb Journal* **25**, 1934-1942 (2011).

293. Tian, Y., Chang, J.C., Fan, E.Y., Flajolet, M. & Greengard, P. Adaptor complex AP2/PICALM, through interaction with LC3, targets Alzheimer's APP-CTF for terminal degradation via autophagy. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 17071-17076 (2013).

294. Karch, C.M., Ezerskiy, L.A., Bertelsen, S., Alzheimer's Disease Genetics, C. & Goate, A.M. Alzheimer's Disease Risk Polymorphisms Regulate Gene Expression in the ZCWPW1 and the CELF1 Loci. *PLoS One* **11**, e0148717 (2016).

295. Margueron, R. *et al.* Role of the polycomb protein EED in the propagation of repressive histone marks. *Nature* **461**, 762-U11 (2009).

296. Witte, V. *et al.* HIV-1 Nef mimics an integrin receptor signal that recruits the polycomb group protein Eed to the plasma membrane. *Molecular Cell* **13**, 179-190 (2004).

297. Danis, E. *et al.* Inactivation of Eed impedes MLL-AF9-mediated leukemogenesis through Cdkn2a-dependent and Cdkn2a-independent mechanisms in a murine model. *Experimental Hematology* **43**, 930-935 (2015).

298. Kirmizis, A. *et al.* Silencing of human polycomb target genes is associated with methylation of histone H3 Lys 27. *Genes & Development* **18**, 1592-1605 (2004).

299. Matera, A.G., Terns, R.M. & Terns, M.P. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat Rev Mol Cell Biol* **8**, 209-20 (2007).

300. Dijkmans, T.F., van Hooijdonk, L.W.A., Fitzsimons, C.P. & Vreugdenhil, E. The Doublecortin Gene Family and Disorders of Neuronal Structure. *Central Nervous System Agents in Medicinal Chemistry* **10**, 32-46 (2010).

301. Taylor, K.R., Holzer, A.K., Bazan, J.F., Walsh, C.A. & Gleeson, J.G. Patient mutations in doublecortin define a repeated tubulin-binding domain. *Journal of Biological Chemistry* **275**, 34442-34450 (2000).

302. Kim, M.H. *et al.* The DCX-domain tandems of doublecortin and doublecortin-like kinase. *Nature Structural Biology* **10**, 324-333 (2003).

303. Moores, C.A. *et al.* Mechanism of microtubule stabilization by doublecortin. *Molecular Cell* **14**, 833-839 (2004).

304. Ferrer, I., Blanco, R., Carmona, M. & Puig, B. Phosphorylated c-MYC expression in Alzheimer disease, Pick's disease, progressive supranuclear palsy and corticobasal degeneration. *Neuropathology and Applied Neurobiology* **27**, 343-351 (2001).

305. Perry, G., Friedman, R., Shaw, G. & Chau, V. Ubiquitin is detected in neurofibrillary tangles and senile plaque neurites of Alzheimer disease brains *Proceedings of the National Academy of Sciences of the United States of America* **84**, 3033-3036 (1987).

306. Gay, N.J., Symmons, M.F., Gangloff, M. & Bryant, C.E. Assembly and localization of Toll-like receptor signalling complexes. *Nature Reviews Immunology* **14**, 546-558 (2014).

307. Zhou, H. *et al.* IRAK-M mediates Toll-like receptor/IL-1R-induced NF kappa B activation and cytokine production. *Embo Journal* **32**, 583-596 (2013).

308. Cui, J.G., Li, Y.Y., Zhao, Y., Bhattacharjee, S. & Lukiw, W.J. Differential Regulation of Interleukin-1 Receptor-associated Kinase-1 (IRAK-1) and IRAK-2 by MicroRNA-146a and NF-kappa B in Stressed Human Astroglial Cells and in Alzheimer Disease. *Journal of Biological Chemistry* **285**, 38951-38960 (2010).

309. Carrette, O. *et al.* Changes of the cortex proteome and Apolipoprotein E in transgenic mouse models of Alzheimer's Disease. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences* **840**, 1-9 (2006).

310. Gu, G.J. *et al.* Role of Individual MARK Isoforms in Phosphorylation of Tau at Ser(262) in Alzheimer's Disease. *Neuromolecular Medicine* **15**, 458-469 (2013).

311. Brady, D.R. & Mufson, E.J. Parvalbumin-immunoreactive neurons in the hippocampal formation of Alzheimer's diseased brain. *Neuroscience* **80**, 1113-1125 (1997).

312. Lin, P. *et al.* A Missense Mutation in SLC33A1, which Encodes the Acetyl-CoA Transporter, Causes Autosomal-Dominant Spastic Paraplegia (SPG42). *American Journal of Human Genetics* **83**, 752-759 (2008).

313. Huppke, P. *et al.* Mutations in SLC33A1 Cause a Lethal Autosomal-Recessive Disorder with Congenital Cataracts, Hearing Loss, and Low Serum Copper and Ceruloplasmin. *American Journal of Human Genetics* **90**, 61-68 (2012).

314. Mao, F. *et al.* Identification and Functional Analysis of a SLC33A1: c.339T > G (p.Ser113Arg) Variant in the Original SPG42 Family. *Human Mutation* **36**, 240-249 (2015).

315. Hirabayashi, Y., Nomura, K.H. & Nomura, K. The acetyl-CoA transporter family SLC33. *Molecular Aspects of Medicine* **34**, 586-589 (2013).

316. Ding, Y. *et al.* Biochemical inhibition of the acetyltransferases ATase1 and ATase2 reduces beta-secretase (BACE1) levels and Abeta generation. *J Biol Chem* **287**, 8424-33 (2012).

317. Ko, M.H. & Puglielli, L. Two Endoplasmic Reticulum (ER)/ER Golgi Intermediate Compartment-based Lysine Acetyltransferases Post-translationally Regulate BACE1 Levels. *Journal of Biological Chemistry* **284**, 2482-2492 (2009).

318. Peng, Y. *et al.* Improved proteostasis in the secretory pathway rescues Alzheimer's disease in the mouse. *Brain* **139**, 937-952 (2016).

319. Volmar, C.-H. & Wahlestedt, C. Histone deacetylases (HDACs) and brain function. *Neuroepigenetics* **1**, 20-27 (2015).

320. Chen, D.Y. *et al.* A critical role for IGF-II in memory consolidation and enhancement. *Nature* **469**, 491-U63 (2011).

321. Schmeisser, M.J. *et al.* I kappa B Kinase/Nuclear Factor kappa B-Dependent Insulin-Like Growth Factor 2 (Igf2) Expression Regulates Synapse Formation and Spine Maturation via Igf2 Receptor Signaling. *Journal of Neuroscience* **32**, 5688-5703 (2012).

322. Yang, Y.W. *et al.* The histone code regulating expression of the imprinted mouse Igf2r gene. *Endocrinology* **144**, 5658-5670 (2003).

323. Frost, J.M. & Moore, G.E. The Importance of Imprinting in the Human Placenta. *Plos Genetics* **6**(2010).

324. Yotova, I.Y. *et al.* Identification of the human homolog of the imprinted mouse Air non-coding RNA. *Genomics* **92**, 464-473 (2008).

325. Wang, L.J. *et al.* ABCB6 Mutations Cause Ocular Coloboma. *American Journal of Human Genetics* **90**, 40-48 (2012).

326. Krishnamurthy, P.C. *et al.* Identification of a mammalian mitochondrial porphyrin transporter. *Nature* **443**, 586-589 (2006).

327. Stefansson, H. *et al.* Neuregulin 1 and susceptibility to schizophrenia. *American Journal of Medical Genetics* **114**, 747-747 (2002).

328. Li, D., Collier, D.A. & He, L. Meta-analysis shows strong positive association of the neuregulin 1 (NRG1) gene with schizophrenia. *Human Molecular Genetics* **15**, 1995-2002 (2006).

329. Munafo, M.R., Thiselton, D.L., Clark, T.G. & Flint, J. Association of the NRG1 gene and schizophrenia: a meta-analysis. *Molecular Psychiatry* **11**, 539-546 (2006).

330. Go, R.C.P. *et al.* Neuregulin-1 polymorphism in late onset Alzheimer's disease families with psychoses. *American Journal of Medical Genetics Part B-Neuropsychiatric Genetics* **139B**, 28-32 (2005).

331. Law, A.J. *et al.* Neuregulin 1 transcripts are differentially expressed in schizophrenia and regulated by 5 ' SNPs associated with the disease. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 6747-6752 (2006).

332. Fleck, D., Garratt, A.N., Haass, C. & Willem, M. BACE1 Dependent Neuregulin Processing: Review. *Current Alzheimer Research* **9**, 178-183 (2012).

333. Chaudhury, A.R. *et al.* Neuregulin-1 and ErbB4 immunoreactivity is associated with neuritic plaques in Alzheimer disease brain and in a transgenic model of Alzheimer disease. *Journal of Neuropathology and Experimental Neurology* **62**, 42-54 (2003).

334. Duxin, J.P. & Walter, J.C. What is the DNA repair defect underlying Fanconi anemia? *Current Opinion in Cell Biology* **37**, 49-60 (2015).

335. Lamantea, E. *et al.* Mutations of mitochondrial DNA polymerase gamma A are a frequent cause of autosomal dominant or recessive progressive external ophthalmoplegia. *Annals of Neurology* **52**, 211-219 (2002).

336. Schulte, C., Synofzik, M., Gasser, T. & Schols, L. Ataxia with ophthalmoplegia or sensory neuropathy is frequently caused by POLG mutations. *Neurology* **73**, 898-900 (2009).

337. Ferrari, G. *et al.* Infantile hepatocerebral syndromes associated with mutations in the mitochondrial DNA polymerase-gamma A. *Brain* **128**, 723-731 (2005).

338. Giordano, C. *et al.* Fatal congenital myopathy and gastrointestinal pseudo-obstruction due to POLG1 mutations. *Neurology* **72**, 1103-1105 (2009).

339. Mokrani-Benhelli, H. *et al.* Primary Microcephaly, Impaired DNA Replication, and Genomic Instability Caused by Compound Heterozygous ATR Mutations. *Human Mutation* **34**, 374-384 (2013).