Investigating the role of *CLU, PICALM* and *CR1* in Alzheimer's disease

Jenny Lord, BSc MSc

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy

July 2014

Abstract

In 2009, two large genome wide association studies (GWAS) found associations between common single nucleotide polymorphisms (SNPs) at three loci (*CLU, PICALM* and *CR1*) and Alzheimer's disease (AD) risk. The causal variants underlying these associations and how these impact on AD susceptibility remain unclear. Target enrichment and next generation sequencing (NGS) were used to completely resequence the three associated loci in 96 AD patients in an attempt to uncover potentially causative and rare variants that may explain the observed association signals. A pipeline was developed for the handling of pooled NGS data following a comparison of several different combinations of programs. 33 exonic SNPs were found within the three genes, along with over 1000 non-coding variants. To identify the variants most likely to be affecting AD risk, a two pronged approach was adopted. The variants were imputed in a large case-control cohort (2067 cases, 7376 controls) to test for association with AD, and the likely functional consequences of the variants were assessed using *in silico* resources. Several of the analysed variants showed suggestive or significant association with AD in the imputed data, and/or were predicted to have consequences on the function or regulation of the genes, suggesting avenues for future research in AD genetics. The whole method of pooled, targeted NGS and prioritisation using imputed data for association testing and *in silico* resources for functional analysis represents a new strategy for tracking down the illusive causation of GWAS signals.

Publications

Next generation sequencing of CLU, PICALM and CR1: pitfalls and potential solutions.
Lord J, Turton J, Medway C, Shi H, Brown K, Lowe J, Mann D, Pickering-Brown S, Kalsheker N, Passmore P, Morgan K; Alzheimer's Research UK(ARUK) Consortium.
International Journal of Molecular Epidemiology and Genetics. 2012;3(4):262-75. Epub 2012 Nov 15.

Rare coding variants in Phospholipase D3 (PLD3) confer risk for Alzheimer's disease
Carlos Cruchaga… **Jenny Lord**... et al.
Nature. Epub 2013 Dec 11.

Missense variant in TREML2 protects against Alzheimer's disease
Bruno A Benitez... **Jenny Lord**... Carlos Cruchaga et al.
Neurobiology of Aging. Epub 2013 Dec 21.

Textbook chapters:
Clusterin. **J Lord** and K Morgan.
Complement Component (3b/4b) Receptor 1 (CR1). **J Lord** and K Morgan.
PICALM. **J Lord** and K Morgan.
In: Morgan, K. and Carrasquillo, M. (eds), 2013. *Genetic Variants in Alzheimer's Disease*. Springer, published July 2013 (ISBN 978-1-4614-7308-4)

Acknowledgements

I would first like to thank Alzheimer's Research UK (ARUK) for the funding of this PhD studentship and the collection of samples included in the sequencing project. Both ARUK and the Big Lottery Fund provided financial support to the project. The next thanks go to the patients and families who kindly donated the samples which enabled this work to take place. Thanks also go to Agilent, Source Bioscience and The University of Nottingham (particularly the School of Life Sciences, and formerly the School of Molecular Medical Sciences). For access to the Exome project data which facilitated the validation of variants, thanks go to Rita Guerreiro and John Hardy at UCL.

Since this project has been largely bioinformatics based, thanks go to all the software developers who have written the programs utilised in this work. Thanks also go to the HapMap Project, the Welcome Trust Case Control Consortium, the ENCODE Project, the 1000 genomes project, UCSC and Ensembl, as well as other sources of publically available data, without which the study of genetics would be far less advanced. Thank you also to the groups that generated the ARUK/Mayo GWAS dataset which enabled association testing of variants which would not otherwise have been possible.

Thanks to the Springer New York publishing group for allowing the adaptation of chapters I contributed to the book Genetic Variants in Alzheimer's Disease within the Introduction section of this thesis.

A huge thank you goes to Kevin Morgan and Noor Kalsheker, whose supervision, support and advice has been invaluable throughout this project. A special thank you is needed for James Turton, friend and colleague, whose help and support over the past few years has been incredible. Thanks to Ng See May for conducting the TaqMan genotyping assays presented in Chapter 6, which formed part of her MSc dissertation project. For all the support, advice and fun times, I thank all the other people from the lab, past and present – Sally Chappell, Linda Morgan, Tamar Guetta-Baranes, Kris Brown, Christopher Medway, Hui Shi, James Bullock, Anne Braae, Imelda Barber and Helen Knight.

Thank you to all the friends and family who have helped me through the last few years (Nicky, Anna, Ess, Tal and the mumble folks to name just a few). A huge thank you goes to my parents – for all the love, support, advice, belief and much needed pep-talks. Finally, Ewan Stern. I cannot even explain how much you've done for me. Thanks for keeping me sane. Thanks for the faith, patience, silliness, love, friendship, proofreading, and endless supply of tea.

This work is dedicated to Taliesin Pearson, PhD: best friend, brilliant mathematician and constant source of inspiration. You will always be missed.

Contents

1. Introduction

1.1. Alzheimer's Disease

Alzheimer's disease (AD) is a devastating, incurable, neurodegenerative disorder, with numerous genetic and environmental risk factors, first described by Dr Alois Alzheimer in the early 1900s. Its prevalence has escalated since its discovery: in 2006, the worldwide prevalence of AD was estimated to be around 26.6 million, and it is thought this figure could rise fourfold in coming decades, to a predicted 107 million cases in 2050, as life expectancies across the world increase (Brookmeyer et al. 2007).

In the UK alone, it is estimated there are around 820,000 people living with dementia. AD is the most common form of dementia, the cost of which to the UK economy each year is estimated to be a staggering £23 billion (Alzheimer's Research UK 2013) – a figure which is set to increase along with the growing numbers of AD sufferers forecast over coming years.

1.2. AD in the clinic

Symptoms and diagnosis

Symptoms of early AD include memory loss, language difficulties, disorientation and behavioural or mood changes. As the disease progresses, these changes become more severe, often resulting in a complete inability to perform daily tasks or recognise loved ones, and constant care becomes a necessity. The symptomatic decline is accompanied by a shrinking of the brain and neuronal cell death. On average, AD patients live for around 8 years following diagnosis, with increasing cognitive impairment and weakening defences leaving them vulnerable to secondary infections, often pneumonia. A clinical diagnosis of possible or probable AD can be provided during the patient's lifetime, mainly on the basis of cognitive assessments, but there is considerable overlap with the symptoms of other forms of dementia, and a definitive diagnosis of AD can only be confirmed post mortem, upon the identification of the characteristic Aβ plaques and tau neurofibrillary tangles within the brain. The plaques and tangles are not specific features of AD, however, and can be seen in the brains of individuals without any cognitive impairment (Villemagne et al. 2008).

Mini Mental State Examination

There are a variety of tests used in the diagnosis of AD. The Mini Mental State Examination (MMSE) (Folstein et al. 1975) is often utilised as it is designed to be a quick way to establish an overview of a patient's mental state. It is a 30 question test designed to assess an individual's mental abilities, including memory, attention and language skills. The test can be affected by educational status and cultural background, but in general, a score of over 27/30 is

indicative of normal cognitive function, while lower scores can be indicative of mild cognitive impairment (MCI, which can, but does not always progress to AD) or AD itself. Scores between 10 and 26 are suggestive of MCI or mild-to-moderate AD, while scores below 10 are suggestive of severe AD. This can help guide treatment strategies. The test can be used over time to monitor decline or assess effectiveness of treatments (information from http://www.alzheimers.org.uk/site/scripts/documents_info.php?documentID=121).

Often, the results of such tests are taken in to consideration alongside other lines of evidence, such as personal and family history, physical examinations or brain scans (which can help identify alternative causes of AD-type symptoms, such as brain tumours, depression and infections). Patients with early stage dementia may require observation over a period of time to see if and how symptoms progress.

A number of frameworks for the diagnosis of AD have been developed over the years, such as the NINCDS-ADRDA (National Institute of Neurological and Communicative Disorders and Stroke and Alzheimer's Diseases and Related Disorders Association), seeking to standardise diagnostics of AD and other dementias, bringing together neuropsychological testing as well as biomarker profiling, MRI and PET scanning (McKhann et al. 1984; Dubois et al. 2007).

Treatments and prevention

Since the fundamental cause of AD remains unclear, it is difficult to develop effective treatments. Knowing what aspects of the pathology of the condition are causes and which are consequences might allow the development of therapies which could actually prevent or cure the disease. As it is, currently available treatments target the symptoms of the disease, so while they may be effective in reducing day to day manifestations of the disorder, they do not modify the course of the disease.

NICE approved treatments

Four drugs are currently approved by the National Institute for Health and Care Excellence (NICE) for the treatment of AD: acetylcholine esterase (AChE) inhibitors (donezepil, galantimine and rivastigmine) are recommended for the management of symptoms in mild to moderate AD, while memantine can be used in the treatment of severe AD, as well as for moderate AD in those patients that cannot be treated using AChE inhibitors (National Institute for Health and Care Excellence 2011). The aim of all these treatments is to manage the cognitive, behavioural and psychological symptoms of the condition in an attempt to maintain function, enabling independence for as great a time as possible. Aside from drug treatments, patients with AD may be aided in a variety of ways, depending on severity and circumstances, e.g. social support, community dementia care, home nursing, respite care and residential care homes.

The three AChE inhibitors, all similar in terms of treatment and cost effectiveness, share a common mechanism of action: increasing levels of acetylcholine (ACh) at the sites of neurotransmission in the brain. Such treatments arose following observations of reduced ACh release, reduced choline uptake and loss of cholinergic neurons in Alzheimer's brains, which lead to the development of the cholinergic theory of AD (Francis et al. 1999). If dysfunction in the AD brain does arise from deficiencies in ACh, inhibiting its degradation (e.g. using AChE inhibitors) should increase the available level of the neurotransmitter, and thus reduce the cognitive impairment seen. However, since the AChE inhibitor treatments only provide symptomatic relief without slowing the progression of the disease, it seems unlikely that they are targeting the fundamental cause of the condition, making it unlikely that ACh deficiency alone is the root cause of AD.

Memantine is a medium affinity, voltage-dependent, non-competitive N-methyl-D-aspartate (NMDA) receptor antagonist, which works by blocking the effects of the increased levels of glutamate seen in AD patients, which is thought to contribute to neuronal dysfunction.

Clearly the devastating effects and huge global burden of AD makes finding new treatments imperative, and a vast number of clinical and pre-clinical trials are currently underway for a plethora of drugs targeting various aspects of AD.

Aβ related treatments

Unsurprisingly, given its predominance as one of the major pathological hallmarks of AD, Aβ is the target of many drugs which have been developed in the fight against AD over the years. Various approaches to reducing Aβ levels in the brain have been attempted: reducing its production; preventing its aggregation; or promoting its clearance.

Since Aβ production occurs as a result of the sequential cleavage of APP by β- and then γ-secretases (discussed in greater detail later), they seem logical targets to inhibit to reduce Aβ production. The main issue with this is that each has multiple other target molecules, notably Notch for γ-secretase, which can lead to unacceptable side effects.

Two γ-secretase inhibitors that have reached clinical trials include LY450139 and Semagacestat, but both trials were halted. Treatment with LY450139 did appear to reduce plasma Aβ levels, but elicited unacceptable side effects and was discontinued at phase II (Fleisher et al. 2008). Semagacestat reached phase III trials, but was stopped when the treatment group showed worsening cognitive functions relative to the placebo group (Samson 2010). There remains potential for the use of γ-secretase modulators, which could shift Aβ production to smaller less toxic forms without interfering with the enzyme's other target molecules (Ozudogru and Lippa 2012).

Inhibitors of the monomeric β-secretase (BACE1) are expected to have less severe side effects. Knock-out animals lacking the catalytic components of γ-secretase (PSEN1 or PSEN2) fail to develop in to viable embryos, while gene knock-out of β-secretase is well tolerated (Luo et al. 2001). Despite this, and despite promising results of BACE1 inhibitors in animal models (Fukumoto et al. 2010; Chang et al. 2011), very few β-secretase targeted therapies have reached clinical trial. One promising result was obtained in the case of CTS-21166, a well tolerated BACE1 inhibitor which passed phase I clinical trials in humans, eliciting a dose dependent decrease in plasma Aβ levels, although little has been published on the molecule (Panza et al. 2009). Further research will be needed to establish whether this is a safe and effective Aβ reducing therapy in humans, and whether this has any positive effect on AD symptoms and progression.

The aim of Aβ aggregation inhibitors is to prevent or reverse the aggregation of Aβ in the brain, reducing the formation of the highly stable amyloid plaques. Whether this is a valid strategy, given increasing evidence that plaques may actually be the brain's defensive mechanism for dealing with Aβ remains to be determined. Two such therapies have reached clinical trials, following promising results in animal studies (McLaurin et al. 2006; Gervais et al. 2007). Tramiprosate was shown to significantly reduce amyloid burden in Tg2576 mice (APP Swedish mutation transgenic strain) (Gervais et al. 2007), and was well tolerated in human clinical trials. However, in phase III trials, cognitive assessments and MRI measures were unaffected in response to the treatment, the reasons for which remain unclear (Aisen et al. 2011). The other therapy, ELND005 (scyllo-inositol) was well tolerated in humans at lower doses, but as yet the therapy's effect on Aβ aggregation and on AD patients remains to be determined (Salloway et al. 2011; Schenk et al. 2012).

Another approach that has received significant attention is that of Aβ immunisation. There are two broad strategies for this; active and passive immunisation. Active immunisation involves the introduction of antigens (e.g. synthetic Aβ42, either in full or fragmented form), to which the body then elicits its own immune response. The alternative to this is passive immunisation, where the antibodies themselves are introduced to the patient, circumventing their own immune system's response. The mechanism by which these strategies work to reduce Aβ remains unclear, but may include prevention of aggregation or stimulation of phagocytosis by microglia (Schenk 2002; Schenk et al. 2012). These approaches are generally trialled on animal models expressing familial AD mutations which result in excessive Aβ production, and some promising findings have been reported, with reductions in neuropathological features, and even some improvements to cognitive performance evidenced after various Aβ immunisation strategies (Schenk et al. 1999; Janus et al. 2000; Morgan et al. 2000). The crossover from these animal models in to humans has been problematic, however. Notably, AN 1792 was approved for human trials following extensive testing on animal models, and

was found to be well tolerated in Phase I human trials. The Phase IIa trial however, featuring 300 patients receiving AN 1792, was halted after a number of patients (18 (6%)) developed meningoencephalitis (Orgogozo et al. 2003), drawing scepticism from the public and scientific community with regards to this approach. Despite this, there are a number of investigations of both active and passive immunisation ongoing, and time will tell whether an acceptable level of adverse effects can be maintained, and whether this can actually help patients suffering from AD (Schenk et al. 2012).

Under the traditional amyloid cascade hypothesis, Aβ is the actual causative factor for AD development, so the need to modulate its production would be crucial. However, over recent years, this theory has increasingly been called in to question, at least partly due to the failure of Aβ based therapies. Despite this, Aβ has been shown to have neurotoxic properties (Maltsev et al. 2011), so even if it is not the fundamental causative agent, reducing its presence early on in the disease process could help limit the damage seen in AD patients.

Who to treat?

One significant issue in attempting to combat AD is that of deciding who to treat. Currently therapies are given when patients present with symptoms of dementia and receive AD as a clinical diagnosis. However, it is thought that the damage which eventually manifests in this way has actually begun to occur much earlier, perhaps decades before any cognitive impairment is detected in the individual. Any treatment given at this time may simply come too late: neurons which have been lost cannot be replaced, and decimated neural connections cannot be repaired.

As such it is imperative that treatments are given as early in the disease process as possible. At the very least, early diagnosis is crucial, but ideally high risk, pre-symptomatic individuals should be identified (e.g. through biomarker screening, scanning to detect early (pre-symptomatic) changes in the brain, and identification of those with high risk genetic profiles). It is possible that in the plethora of drugs which have failed to elicit sufficient responses in clinical trials, there are useful therapies which have been disregarded because they were applied to patients for whom it was already too late to address the damage caused by AD. Prevention, in such a condition, may be the only viable form of cure.

It is interesting to note that aside from the prospect of an absolute cure, simply delaying the onset of the condition for a modest time period (1-2 years) could reduce the global case load and associated financial burden significantly (Brookmeyer et al. 2007).

Prevention

Longitudinal studies of individuals taking non-steroidal anti-inflammatory drugs (NSAIDs) for unrelated conditions, such as rheumatoid arthritis, reveal a lower incidence of AD when such medication is used before the typical age

of onset of AD (Breitner et al. 1994; Stewart et al. 1997). This has lead to trials treating AD patients with NSAIDs, but with disappointing results. It was demonstrated that when patients already diagnosed with AD are given such treatments, there is no improvement (or even a worsening) of symptoms (Aisen et al. 2003; Martin et al. 2008). Taken together, this suggests that before the symptoms of AD present, limiting inflammation can slow or prevent the development of AD, thus implying inflammation may play a key early role in AD pathogenesis. These drugs are relatively cheap, easily available and have demonstrably acceptable side effects, so may represent a promising strategy for reducing AD incidence in the future, if at risk individuals can be identified, and NSAIDs given before disease onset.

Similarly, it has been shown that the use of statins, taken to reduce cholesterol, are associated with a decreased risk of AD (Zamrini et al. 2004). This may be an effect of the drugs themselves; there is some evidence that statins may lead to enhanced degradation of extracellular Aβ by triggering the release of insulin-degrading enzyme from microglia (Tamboli et al. 2010). However, it is also possible that the effect is indirect, with the drugs simply lowering cholesterol, high levels of which are a known risk factor for AD (Kivipelto and Solomon 2006).

Exposure to environmental risk factors for AD can also affect susceptibility. Age is by far the most significant environmental risk factor for late onset AD, with risk increasing dramatically as aging occurs. Other factors include educational status, head injury, exercise, hypertension, and vitamin intake (Dosunmu et al. 2007; Dangour et al. 2010; Pogge 2010), although what effect modifications of these risk exposures could have on public health remains to be determined (Dangour et al. 2010).

1.3. AD pathological hallmarks

AD is a highly heterogeneous condition, with the physical manifestations of the disorder showing considerable overlap with other types of dementia (e.g. vascular and frontotemporal dementia). Indeed, the "pathological hallmarks" (Aβ plaques and tau tangles) generally taken to be diagnostic of AD can also be present in other types of dementia, as well as in non-demented brains.

One of the major changes seen in AD is at the gross level, with shrinkage of the brain apparent as a result of neuronal and synaptic losses particularly in the hippocampus, an area which is implicated in the formation of new memories, and is particularly affected in early AD (see Figure 1.1).

Figure 1.1 - Healthy vs. advanced AD brain

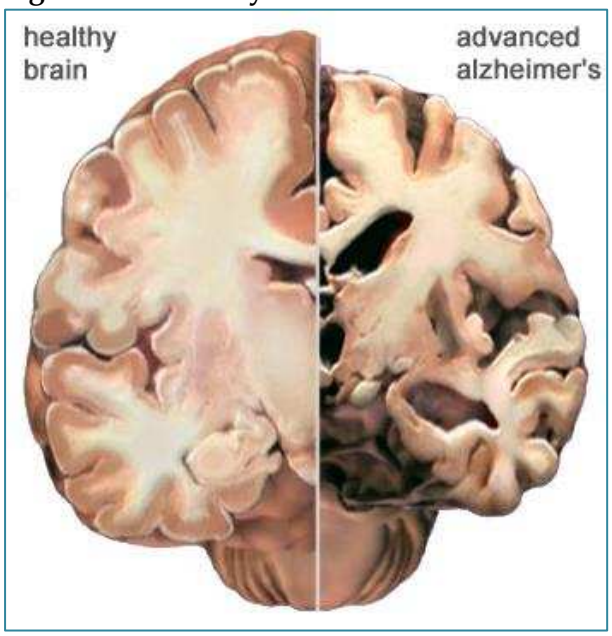


Image to show the difference in gross brain volume between healthy (left) and AD affected (right) brains, taken from http://www.alz.org/braintour/healthy_vs_alzheimers.asp. Large lesions apparent on the AD image are typical and indicative of neuronal cell death.

On a finer level, Aβ plaques and tau neurofibrillary tangles (NFTs) are two of the major characteristic features of AD. The structure and localisation of these is shown in figure 1.2.

Figure 1.2 - Amyloid plaques and tau neurofibrillary tangles

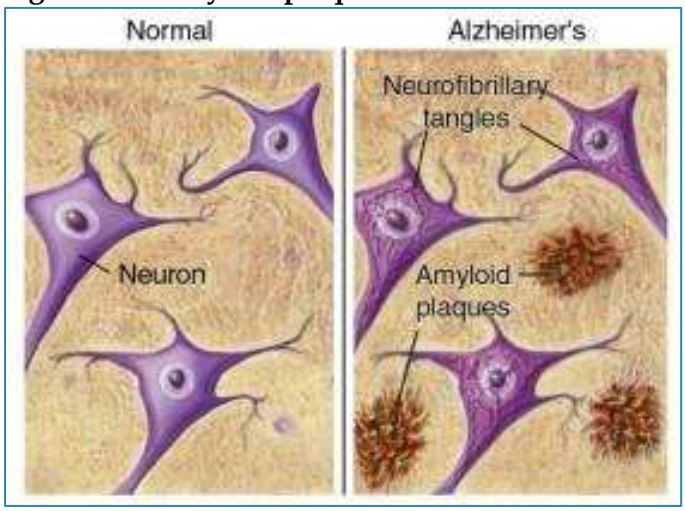


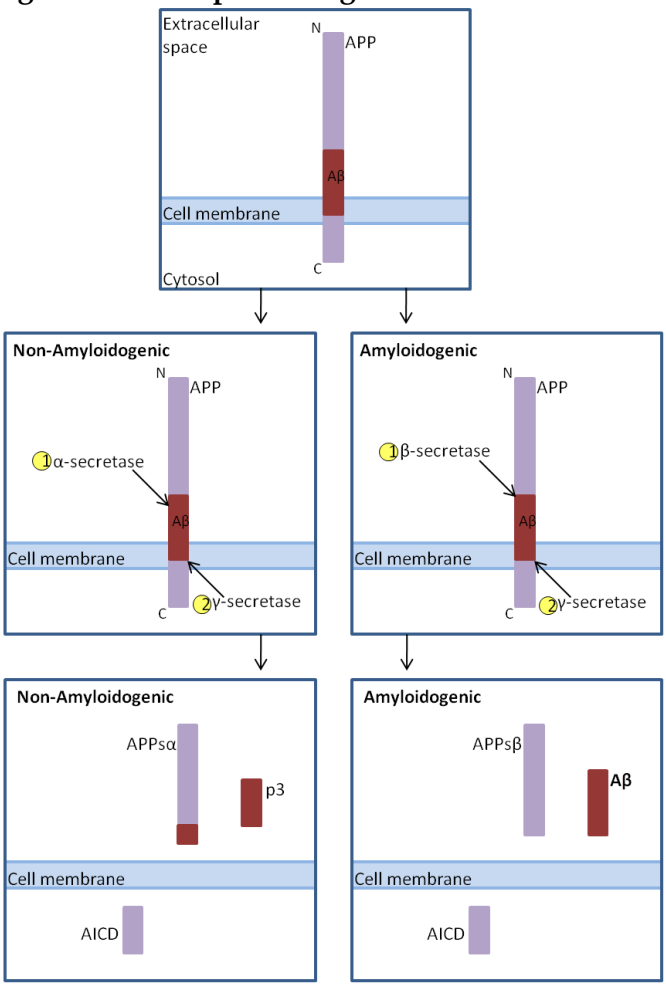
Image to show the differences in brain pathology in normal and AD affected brains, taken from http://www.brightfocus.org/alzheimers/about/understanding/plaques-and-tangles.html. The image on the left depicts a region of healthy brain from an unaffected individual, featuring normally functioning neurons and no extracellular accumulations of protein. On the right, an AD affected brain is shown. Here the neurons contain clumps of hyperphosphorylated tau, and there is extracellular accumulation of amyloid and other components which form Aβ or amyloid plaques.

Aβ plaques

Aβ or amyloid plaques are extracellular accumulations of protein, which are comprised of the hydrophobic Aβ peptide, along with a multitude of other proteins and cell constituents, including apolipoprotein E (ApoE), clusterin (or ApoJ) and several components of the complement cascade (Liao et al. 2004). Activated glial cells and immune complexes are often found in the vicinity of Aβ plaques, indicating immune involvement and activation (Eikelenboom et al. 2006).

The Aβ peptide is produced when the amyloid precursor protein (APP) is processed, which can follow either the amyloidogenic or non-amyloidogenic pathway, summarised in Figure 1.3, depending on which secretase enzyme is involved in the first of the sequential cleavage steps.

Figure 1.3 - APP processing



Schematic diagram of APP processing, adapted from Thinakaran 2008 (Thinakaran and Koo 2008). In the non-amyloidogenic pathway, α-secretase cleaves within the Aβ peptide sequence, meaning Aβ cannot be produced. Instead, APPsα, peptide P3 and AICD are generated. In the amyloidogenic pathway, cleavage by β- then γ-secretase leads to the release of Aβ, plus APPsβ and AICD.

Both APP processing pathways produce the amyloid precursor protein intracellular domain (AICD), which is thought to have important roles in transcriptional control, although the genes on which it exerts these effects are yet to be established (Chang and Suh 2010). It is the amyloidogenic pathway – with cleavage first by β- then γ-secretase - that generates the aggregation prone Aβ peptide. There are several species of Aβ peptide, as γ-secretase can cleave the protein at several different amino acid residues: Aβ-40 and Aβ-42 are the most common (~90% and <10% respectively), although shorter species have been reported (Selkoe and Wolfe 2007). Aβ-42 is particularly hydrophobic and prone to aggregation.

Neurofibrillary tangles

tau, the main component of neurofibrillary tangles, is a microtubule associated protein, which is involved in the stabilisation and regulation of microtubule bundles, crucial for cytoskeletal integrity and axonal transport (Roy et al. 2005). Tau's normal function is modulated by a fine balance between phosphorylation and dephosphorylation. When this balance is breached, tau hyperphosphorylation can occur, disrupting its normal functions, and giving the molecule a propensity to form paired helical filaments. These are the insoluble, aggregation prone building blocks of neurofibrillary tangles which form within the cytoplasm of neuronal cells (Goedert et al. 1995). Tau is also implicated in a number of other neurodegenerative disorders, such as frontal temporal dementia (Neumann et al. 2009). This can be caused by mutations within the gene encoding tau, microtubule associated protein tau (*MAPT*) (Goedert and Jakes 2005), which leads to neuronal cell loss and consequent cognitive decline. The formation of amyloid plaques is generally thought to precede the formation of tau tangles - both plaques and tangles are found to arise from mutations in APP, while mutations in tau generally only give rise to tangles (Lovestone 2000).

It is not clear how either of these signature lesions relate to AD pathology, and whether they are early, causative events in the development of AD, or late stage consequences of the disease process. Additional to these, inflammation of the brain, vascular involvement and cerebral amyloid angiopathy (CAA) are also often observed in AD brains compared to normal controls, but none of these are specific or necessary for a diagnosis of AD.

1.4. Types of AD

There are two general forms of AD, classified on the basis of the time of onset of symptoms - those showing symptoms before the age of 65 (generally between 50-65, but can be much earlier (Bird et al. 1996)) are classified as early onset, and account for about 5% of total AD cases. Late onset AD (LOAD), which constitutes around 95% of cases, shows the onset of symptoms after the age of 65. Although this distinction is often made, it is a rather arbitrary cut off

dividing a whole spectrum of ages of onset, likely due to differing contributions of genetic and environmental risk factors, in to two discrete categories.

Early onset AD

Early onset AD can be divided in to two groups - early onset familial (EOFAD) and early onset sporadic AD. The first, EOFAD is a monogenic disorder caused by rare mutations in one of three genes (*PSEN1*, *PSEN2*, or *APP*), inherited in a Mendelian fashion, which almost guarantee the onset of symptoms before the age of 65. All three of these genes are tightly linked to APP processing (*PSEN1* and *PSEN2* encode components of the γ-secretase enzyme), which mutations then disrupt, giving an accumulation of Aβ and formation of amyloid plaques at an earlier age than is seen in LOAD.

The first EOFAD causative mutation in APP was identified by Goate et al. in 1991 (Goate et al. 1991). Multiple EOFAD affected families had mutations which showed linkage to chromosome 21. The variant, V717I was termed the London mutation (Goate et al. 1991). Subsequent to this, a large number of other causative mutations have been identified, including 24 in *APP*, 185 in *PSEN1* and 14 in *PSEN2* (Tanzi 2012). The majority are fully penetrant autosomal dominant mutations, with one APP recessive mutation reported to date (Tanzi 2012). Most of the mutations in APP alter the relative ratios of $A\beta_{42}$:$A\beta_{40}$. The Swedish mutation, notable as it is often used as a model for AD in transgenic animals, encodes an amino acid change within the Aβ domain of APP (Lannfelt et al. 1994), which increases the overall production of all Aβ species, as well as enhancing the molecule's propensity to aggregate (Tanzi 2012). There is also known to be an APP mutation (A673T) that is protective against both AD and cognitive decline in elderly AD unaffected individuals (Jonsson et al. 2012).

Early onset sporadic AD is less easily defined. Individuals with this form of AD do not appear to follow the monogenic inheritance pattern typical of the familial form of AD but have an age at onset below 65 years. This may be due to complex interplay between different genetic and environmental risk factors, as for the late onset condition, or may be due to as yet undiscovered mutations, perhaps showing differing levels of penetrance, making it harder to track within families (Jin et al. 2012; Antonell et al. 2013).

Late onset AD

The vast majority of cases of AD are late onset (LOAD), with the onset of symptoms after the age of 65. LOAD is a complex disorder, caused by a combination of genetic and environmental risk factors, all of which alter risk for AD without alone being necessary or sufficient for the development of the disorder. Age is the most significant environmental risk factor for LOAD, and other factors which have been implicated in increasing or decreasing LOAD risk, affecting age of onset, or affecting disease progression include

educational status, head injury, hypertension, high cholesterol and vitamin intake (Dosunmu et al. 2007; Pogge 2010).

1.5. Genetics of LOAD

Since this thesis deals primarily with LOAD, LOAD shall simply be referred to as AD, and any mention of AD from this point forward refers to the late onset, complex disorder.

Although the heritability of AD is estimated to be around 70%-80% (Gatz et al. 2006), relatively little is understood about the genetics of the condition. None of the genes involved in EOFAD have been implicated in sporadic AD risk (Tanzi and Bertram 2005), and until recently, there was only one well established and replicated genetic risk factor for AD – the ε4 allele of Apolipoprotein E (*APOE*), which, due to its large effect size, was identified via linkage studies in the early 1990s (Pericak-Vance et al. 1991).

ApoE is a 229 amino acid glycoprotein with three major isoforms, created by combinations of non-synonymous single nucleotide polymorphisms (SNPs) at two variant sites within the gene, which generate the ε2, ε3 and ε4 alleles. The ε4 allele has been associated with an increased risk of AD, with those in possession of a single copy at a 2-3 fold increased risk, and ε4 homozygotes at around 12 times greater risk than those with no ε4 alleles; while the rare ε2 allele has a protective effect, and is associated with a decreased risk of AD (Farrer et al. 1997).

The main site of ApoE expression is in the liver, with the brain second (Kim et al. 2009). Most of the expression in the brain is accounted for by astrocytes (Grehan et al. 2001), rather than neuronal cells, although these can be induced to express low levels of ApoE under certain conditions (Xu et al. 2006). Its normal biological function is not yet fully understood, so it is unknown whether the increased risk of AD conferred by the ε4 allele stems from a loss of neuroprotective function, the gain of neurotoxic function, or a combination of the two. There is evidence that the different ApoE isoforms may have differing effects on neuroinflammation (Kim et al. 2009), Aβ deposition in the brain (Reiman et al. 2009) and Aβ clearance from the brain (Deane et al. 2008), any of which could be related to an alteration in AD risk. The SNPs in *APOE* that generate the different protein isoforms have striking effects on the structure of the molecule (Mahley et al. 2006), so it is unsurprising that its function is affected; it is just yet to be elucidated how, and how this impacts on AD risk.

1.6. Genome wide association studies

Despite two decades of research trying to find new genetic variants involved in AD risk, no other loci could be definitively confirmed and replicated as genetic risk factors for AD, although over 500 genes were investigated (Bertram et al. 2007), largely identified as being plausible biological candidates, based on what is known about the aetiology of the condition.

The problem with looking at biological candidate genes is that one is entirely limited by prior knowledge of the condition, and when that knowledge is limited or incomplete, such as is the case for AD (and many other complex disorders), genes which are involved, but do not necessarily fit with current understanding are bound to be overlooked. Additionally, any genes which are discovered via this approach are unlikely to greatly further the understanding of the aetiology of the condition since they will fit with mechanisms and pathways which are already thought to be involved.

In order to combat this, an unbiased method of searching for loci involved in disease risk was needed - a way of considering all genes in the human genome simultaneously, without any assumption as to which might be involved.

This much needed method was provided by the advent of genome wide association studies (GWAS), which became possible as a number of crucial components came together concurrently. Firstly, the HapMap project (HapMap 2003) was formed - an international collaboration that sought to catalogue all common human variation in a number of different populations, and document the patterns of linkage disequilibrium (LD) therein. Secondly, technology was developed that allowed hundreds of thousands of SNPs to be genotyped simultaneously, and relatively affordably. These two factors meant a panel of tag SNPs could be devised which would capture the majority of variance across the whole genome of an individual. Finally, sample sets that were large enough to give GWAS sufficient power became available through collaboration between different research groups.

Essentially, a GWAS is based on hundreds of thousands (or millions, with the most recent technology) of SNPs being genotyped in large numbers of cases and controls (generally several thousand). Variants which are significantly more common in one group than the other are said to be associated with the particular trait being considered – more common in controls, and the variant is associated with decreased risk, more common in cases, and the variant is associated with increased risk. Because of the vast numbers of simultaneous independent tests being conducted, the rate of false positives would be unacceptably high at the standard level of significance ($p < 0.05$), so a conservative Bonferroni correction is applied, making the level generally classed as reaching genome wide significance $p < 5x10^{-8}$, and anything falling between $5x10^{-5}$ to $5x10^{-8}$ is said to show suggestive significance.

Replication is crucial to ensure GWAS "hits" are not spurious false positives, both within the same population, and in different populations, although population specific differences in LD may mean that a SNP tagging a causal variant in one population may not do so in a different population, and rare causative variants may be population specific.

1.7. AD GWAS

Although there had previously been a number of GWAS in AD, they failed to detect any signals reaching genome wide significance, other than *APOE* (Coon et al. 2007; Grupe et al. 2007; Abraham et al. 2008; Beecham et al. 2009; Carrasquillo et al. 2009), and loci showing suggestive significance did not replicate well. This was likely due the relatively small numbers of case and control samples used, meaning studies were underpowered to discover variants of modest effect sizes. Furthermore, early genotyping chips did not have a very comprehensive coverage of the genome, since tag SNPs were at first selected based on distance, rather than LD, which could mean causal variants simply were not represented in these studies.

In September 2009, the first two truly large scale GWAS for AD, using the latest genotyping technologies, were simultaneously published in a single edition of Nature Genetics (Harold et al. 2009; Lambert et al. 2009). The two studies together identified three independent signals reaching genome wide significance, bringing the first major advancement in our knowledge of AD genetics in over 15 years. Each study used a two stage approach to garner compelling evidence for the loci identified, taking any SNPs showing evidence of association in the first stage, and replicating these in a second, independent sample set. The sample numbers and SNPs genotyped by each study are summarised in Table 1.1. Each paper found two genome wide significant hits: *CLU* and *PICALM* in the Harold et al. study (Harold et al. 2009), and *CLU* and *CR1* in the Lambert et al. paper (Lambert et al. 2009), with odds ratios and significance values summarised in Table 1.2. Since *CLU* was identified by both papers, replication for this locus was immediately available. In addition to this, *PICALM*, identified as significant by Harold et al. showed suggestive significance (OR = 0.88, $p = 2.8\text{x}10^{-3}$) in Stage 1 of the Lambert et al. paper, and *CR1*, identified as significant by the Lambert et al. paper showed suggestive significance (OR = 1.17, $p = 8.3\times10^{-6}$) in Stage 1 of the Harold et al. paper. Since then, further replication for all three loci has been published (Carrasquillo et al. 2010; Corneveaux et al. 2010; Jun et al. 2010; Seshadri et al. 2010), and a number of other genes have also been found to be associated with AD, including *BIN1* (Seshadri et al. 2010), *ABCA7*, the *MS4A* locus, *EPHA1*, *CD33* and *CD2AP* (Hollingworth et al. 2011; Naj et al. 2011). The basic design and summary results for *CLU, PICALM* and *CR1* in some of the replication GWAS are presented in Table 1.3.

Although each of these new findings has received replication in independent Caucasian sample groups, replication in other populations has been less successful (Jun et al. 2010; Lee et al. 2010; Li et al. 2011; Logue et al. 2011). Whether this stems from genuine aetiological differences across populations, or whether these studies have simply been underpowered, with insufficient sample sizes to detect associations of the expected magnitude remains to be elucidated.

Table 1.1 – Study design of the two major 2009 AD GWAS

Study	Stage	SNPs genotyped	Genotyping Platform	Cases	Controls	Total Samples
Harold et al. 2009	1	529,205	Illumina HumanHap550/300 BeadChips	3941	7848	11789
	2	2	Sequenom assays	2023	2340	4363
Lambert et al. 2009	1	537,029	Illumina Human610-Quad BeadChips	2032	5328	7360
	2	11	Taqman (Applied Biosystems) or Sequenom assays	3978	3297	7275

Information on the number of SNPs genotyped, genotyping platform used, and post QC sample numbers in the two major AD GWAS published in September 2009 (Harold et al. 2009; Lambert et al. 2009).

Table 1.2 – Main results of the two major 2009 AD GWAS

Gene	SNP	Paper	*p*-value	OR (combined)	95% CI
CLU	rs11136000	Harold et al.	8.5×10^{-10}	0.86	0.82-0.90
		Lambert et al.	7.5×10^{-9}	0.86	0.81-0.90
CR1	rs6656401	Lambert et al.	3.7×10^{-9}	1.21	1.14-1.29
PICALM	rs3851179	Harold et al.	1.3×10^{-9}	0.86	0.82-0.90

Results from the two major 2009 AD GWAS, including *p*-values and ORs of the SNPs reaching genome wide significance (Harold et al. 2009; Lambert et al. 2009).

Table 1.3 – *CLU, PICALM* and *CR1* replication GWAS

Study	Part/Design	Samples	Genotyped	Results
Cornevaux et al. 2010	Case-control	1019 Cases, 591 Controls (White - USA, UK, Netherlands)	34 SNPs	*CLU* – rs11136000 – OR 0.86, $p = 0.040$
				PICALM – rs541458 – OR 0.81, $p = 0.01$
				CR1 – rs6656401 – OR 1.28, $p = 0.008$
Carrasquillo et al. 2010	Case-control	1819 Cases, 2565 Controls (White – USA)	3 SNPs	*CLU* – rs11136000 – OR 0.82, $p = 8.6 \times 10^{-5}$
				PICALM – rs3851179 – OR 0.80, $p = 1.3 \times 10^{-5}$
				CR1 – rs3818361 – OR 1.15, $p = 0.014$
Jun et al. 2010	Meta-analysis	5935 Cases, 7034 Controls (9 white northern European cohorts)	17 SNPs	*CLU* – rs11136000 – OR 0.91, $p = 0.0007$
				PICALM – rs3851179 – OR 0.89, $p = 3.9 \times 10^{-5}$
				CR1 – rs3818361 – OR 1.14, $p = 6.1 \times 10^{-5}$
	Meta-analysis	1135 Cases, 1135 Controls (5 cohorts including African American, Israeli-Arab and Caribbean Hispanic)	17 SNPs	No significant associations
Schjeide et al. 2011	Combined case-control	2868 Cases, 1386 Controls (USA family samples, German unrelated cases and controls)	5 SNPs	*CLU* – rs11136000 – OR 0.88, $p = 0.04$
				PICALM – rs541458 – OR 0.82, $p = 0.01$
				CR1 – rs6656401 – OR 1.33, $p = 0.001$
Naj et al. 2011	Combined case-control	11840 Cases, 10931 Controls (14 cohorts of European ancestry)	9 Loci	*CLU* – rs1532278 – OR 0.89, $p = 8.3 \times 10^{-8}$
				PICALM – rs561655 – OR 0.87, $p = 7.0 \times 10^{-11}$
				CR1 – rs6701713 – OR 1.16, $p = 4.6 \times 10^{-10}$

Study design and results for some of the GWAS replicating the associations between SNPs in *CLU, PICALM* and *CR1* and AD (Carrasquillo et al. 2010; Corneveaux et al. 2010; Jun et al. 2010; Naj et al. 2011; Schjeide et al. 2011).

The new genes which have been identified as AD risk factors via GWAS are all involved in a small number of pathways (see figure 1.4), which include lipid and Aβ metabolism, the immune response, and cell membrane processes. This gives new insights in to the aetiology of the condition, which may in turn lead to new targets for treatments.

Figure 1.4 – Pathways of AD GWAS genes

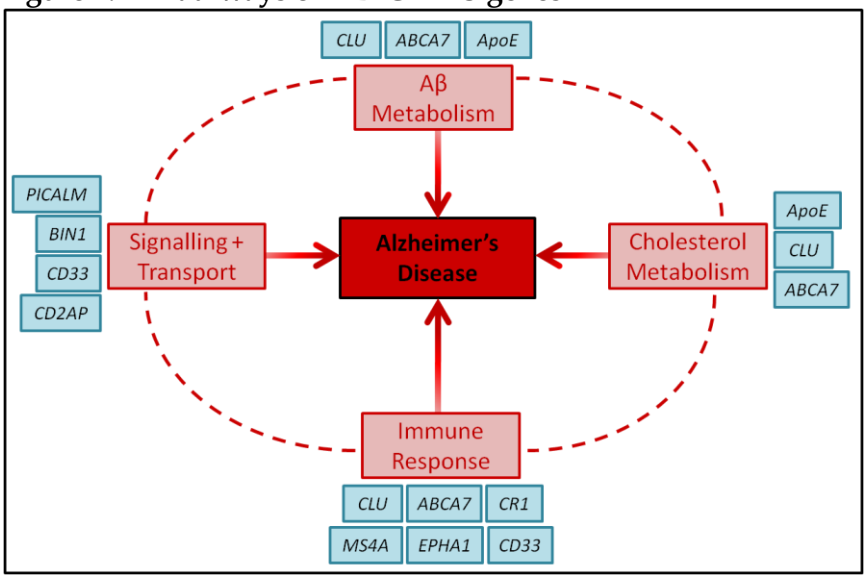


Diagram to show the way in which the new AD candidate genes identified by GWAS relate to potential pathways in AD pathogenesis, adapted from Morgan 2011 (Morgan 2011).

For the past two decades, the amyloid cascade hypothesis has been the dominant theory of AD pathogenesis. The amyloid cascade hypothesis (Hardy and Higgins 1992) centres on the belief that Aβ is the primary causative agent in the pathogenesis of AD, and all other features of the disease (neurofibrillary tangle formation, neuroinflammation, vascular damage, neuronal cell death etc.) are secondary consequences of this. The form of Aβ thought to be responsible for this has changed over the years since the hypothesis was first devised. Initially, it was thought that the Aβ plaques were responsible for the neurotoxic effects, however, since Aβ plaques can be seen in the brains of cognitively normal individuals (Villemagne et al. 2008); plaque presence and cognitive decline do not correlate well (Terry et al. 1991); and treatments which clear Aβ plaques from the brains of affected individuals give no long term symptomatic improvements (Holmes et al. 2008), this has become less popular than the theory that it may be pre-fibrillar, soluble oligomers of Aβ that are the source of the problem (Ferreira et al. 2007), perhaps via disruption of synaptic signalling. Under this version of the amyloid cascade hypothesis, it is possible that the Aβ plaques themselves are in fact the brain's attempt to sequester these harmful Aβ oligomers away from neurons, thus limiting damage (Maltsev et al. 2011).

There are numerous strands of support for the theory; individuals with trisomy 21 (with an extra copy of the APP gene), have increased Aβ

production and often develop AD at a young age (Olson and Shaw 1969), and the familial forms of AD involve disruptions to normal levels of Aβ production (Pimplikar 2009). Aβ has been shown to have neurotoxic effects both *in vivo* and *in vitro* (Maltsev et al. 2011).

Most of the evidence against the amyloid cascade hypothesis is related to the version of the hypothesis that postulates the plaques to be the harmful species, however, even the revised version of the theory cannot explain all aspects of the condition (Pimplikar 2009), and consequentially, a number of alternative hypotheses have been proposed over the years, each claiming a different primary cause of the condition. These hypothesised initial insults include tau hyperphosphorylation (Maccioni et al. 2010), decreased acetylcholine production (Francis et al. 1999), oxidative stress (Markesbery 1997), mitochondrial disregulation (Swerdlow and Khan 2004), neuroinflammation (Tan and Seshadri 2010; Zotova et al. 2010) and vascular damage (Marchesi 2011).

Each of these theories has its own relative strengths and weaknesses, but none can fully explain the entire catalogue of neurological features and symptoms in AD. It is hoped that the new AD associated genes will help to create a more unified theory of AD pathogenesis, which will eventually explain fully how the disease forms and what the causative agents are.

1.8. New genes in AD

The three most significant genes which were implicated in AD risk by the Lambert et al. and Harold et al. GWAS were *CLU, CR1* and *PICALM* (Harold et al. 2009; Lambert et al. 2009). As these three genes are the main focus of this thesis, they will be discussed in detail later (see *CLU, PICALM* and *CR1* sections 1.9, 1.10 and 1.11 respectively). Below is an overview of what is currently known about the other newly identified GWAS genes, and a brief review of the ways in which they may contribute to AD risk, given previous theories of the condition, and the pathways implicated by the new AD associated genes.

1.8.1. *BIN1*

The gene encoding bridging integrator 1 (*BIN1,* also known as amphiphysin II), was first implicated in AD risk by the Harold et al. 2009 GWAS, where a suggestive association ($p<10^{-6}$) was detected between two SNPs, rs744373 and rs7561528 ~30kb upstream of the gene and AD (Harold et al. 2009). In later studies with larger sample sizes, both SNPs have since reached genome wide significance (rs744373 $p = 1.59\times10^{-11}$, OR 1.13 (95% CI 1.06–1.21) (Seshadri et al. 2010); rs7561528 $p = 5.2\text{x}10^{-14}$, OR 1.17 (95% CI 1.12-1.22) (Naj et al. 2011); rs7561528 $p = 2.6\text{x}10^{-14}$, OR 1.17 (95% CI 1.12-1.21) (Hollingworth et al. 2011)).

The gene, situated on chromosome 2q14 is comprised of 20 exons, which are alternatively transcribed to give ~10 protein isoforms (Pruitt et al. 2007). The isoforms of the protein are differentially expressed between tissue types, with some displaying brain-specific expression, and show differences in subcellular localisation (Wechsler-Reya et al. 1997; DuHadaway et al. 2003). *BIN1* has tumour suppressor gene activity, and has been linked to both breast and prostate cancers (Kuznetsova et al. 2007). Like PICALM, BIN1 is implicated in the process of clathrin mediated endocytosis (CME), facilitating the recycling of neurotransmitters and synaptic vesicles (SVs) which is important for efficient signalling. CME is a complex process, with BIN1's role apparently in the recruitment of dynamin to the site of CME at the membrane, which is needed to "pinch off" developing SVs (Pant et al. 2009).

1.8.2. *ABCA7*

The gene encoding ABCA7 is situated on chromosome 19p13.3, spanning a region of around 25.5kb. It is a member of the ATP binding cassette (ABC) transporter gene super-family, with 49 documented ABC proteins, and 12 ABCA sub-family members. Members of these families are involved in the active transport of a range of substances (e.g. ions, sugars, peptides) across cellular and organelle membranes (Vasiliou et al. 2009). ABCA7, along with other ABCA family members, is expressed in the brain, and features two transmembrane domains, each comprised of six α helices. Although the function of ABCA7 is not completely understood, ABC transporters are thought to facilitate the uptake of glucose, amino acids and ions to the brain, as well as having lipid trafficking functions which may be relevant to AD pathology (Kim et al. 2008). Apolipoproteins A-I have been shown to be able to act as ligands for ABCA7, introducing the possibility that the transporter may be involved in Aβ clearance (Tanaka et al. 2011).

The gene was first implicated in AD risk by a GWAS conducted by Hollingworth et al. in 2011. In this study, intronic SNP rs3764650, greatly surpassed the threshold for genome wide significance after replication and meta-analyses including data from a companion study ($p = 5.0 \times 10^{-21}$, OR 1.23 (95% CI 1.17-1.28)) (Hollingworth et al. 2011; Naj et al. 2011).

1.8.3. *MS4A* locus

Several SNPs at the *MS4A* locus have been implicated in AD susceptibility by GWAS (Hollingworth et al. 2011; Naj et al. 2011). The *MS4A* gene region spans an area of around 800kb on chromosome 11q21, with 16 members of the *MS4A* family documented in this area to date, many encoding proteins with multiple isoforms (Liang and Tedder 2001). The SNPs implicating this region in AD risk fall within a block of high LD, making it difficult to pinpoint which of the many genes are involved, but implicating *MS4A2, MS4A6A, MS4A4E* and *MS4A4A. MS4A6A, MS4A4E* and *MS4A4A* are not well characterised, but homology in protein structure, and presence in a large gene cluster implies structural and functional similarities with MS4A1 and MS4A2 which have been experimentally characterised, and are thought to be involved in calcium

signalling via immunoglobulin receptor signalling complexes (Walshe et al. 2008). Many of the *MS4A* genes have been shown to be expressed in the brain, and in cell types associated with immunity and neuroinflammation (Liang and Tedder 2001), which may suggest mechanisms by which the alteration in AD risk is invoked.

This region was implicated in AD risk when two large GWAS published simultaneously in 2011 found several SNPs within the region reached genome wide significance. In a combined meta-analysis of data from both studies, three SNPs were found to have highly significant associations with AD (rs610932 - $p = 1.2x10^{-16}$, OR 0.91 (95% CI 0.88-0.93); rs670139 - $p = 1.1x10^{-10}$, OR 1.08 (95% CI 1.06-1.11); rs4938933 - $p = 8.2x10^{-12}$, OR 0.89 (95% CI 0.87-0.92)) (Hollingworth et al. 2011; Naj et al. 2011).

1.8.4. *EphA1*

The *EphA1*, or erythropoietin-producing human hepatocellular carcinoma gene spans a region of around 17.8kb on chromosome 7q34, featuring 18 exons which encode a 976 amino acid member of the receptor tyrosine kinase superfamily (Hirai et al. 1987; Maru et al. 1988). It was first implicated in AD risk by two GWAS published in 2011, where in the meta-analysis of data taken from each study, a SNP ~3kb upstream of the gene, rs11767557 reached genome wide significance ($p = 6x10^{-10}$, OR 0.87 (95% CI 0.78-0.96) (Hollingworth et al. 2011; Naj et al. 2011), a finding which has subsequently been replicated (Carrasquillo et al. 2011).

The product of the gene, which is expressed in multiple tissue types, is thought to be involved in cell-adhesion, cellular organisation and synaptic plasticity (Yamazaki et al. 2009; Hruska and Dalva 2012; Triplett and Feldheim 2012). EphA1 is important in synapse development during embryogenesis (Hruska and Dalva 2012), which could suggest that underlying differences in neuronal circuitry dictated by differences in EphA1 during development may render the brain more or less able to cope with the changes associated with AD (Chen et al. 2012). Alternatively, the gene's role in mature neurons, promoting maintenance of the synapse and synaptic plasticity could affect AD susceptibility (Chen et al. 2012).

1.8.5. *CD33*

The gene encoding *CD33* or siglec-3, (sialic acid binding immunoglobulin-like lectin-3) is situated on chromosome 19q13.3, spanning a region of around 14.9kb. CD33 is a transmembrane receptor for sialic acids, which is expressed in multipotent cells of the myeloid lineage, as well as in mature monocytes, macrophaghes, dendritic cells, basophiles and mast cells, with three isoforms validated to date (Andrews et al. 1983; Griffin et al. 1984; Valent and Bettelheim 1992; Yokoi et al. 2006). CD33 is thought to be involved in the body's immune responses, and may be a key player in dampening any innate immune responses triggered by the "self" (Crocker et al. 2007; Varki 2009). Siglec family members have also been shown to be implicated in the

endocytosis of various ligands (Biedermann et al. 2007; Tateno et al. 2007; Walter et al. 2008), although not via CME, which is emerging as a common factor between several AD risk genes (Tateno et al. 2007). Either of these processes could be linked to the aetiology of AD development.

CD33 was first implicated in AD when in a study looking at families with multiple affected individuals, Bertram et al. found an association between a variant (rs3826656) ~3kb upstream of the gene and AD risk ($p = 4x10^{-6}$), but which failed replication in an independent cohort (Bertram et al. 2008). In 2011, two companion GWAS papers were published, with meta-analysis of the combined data revealing a genome wide significant association between SNP rs3865444 and AD ($p = 1.6x10^{-9}$, OR 0.91 (95% CI 0.88-0.93)) (Hollingworth et al. 2011; Naj et al. 2011), a finding which has subsequently been replicated (Carrasquillo et al. 2011).

1.8.6. *CD2AP*

The gene *CD2AP* (CD2-associated protein) was implicated in AD when the SNP rs9349407, within intron one showed association with AD in the combined data of two large companion GWAS from 2011($p = 8.6x10^{-9}$, OR 1.11 (95% CI 1.07-1.15) (Hollingworth et al. 2011; Naj et al. 2011). Although this effect has failed to replicate in other studies of Caucasian populations (Carrasquillo et al. 2011), meta-analysis including the initial data showed a strengthened association ($p = 6.5x10^{-11}$).

CD2AP is a 149.5kb gene with 18 exons, situated on chromosome 6p12. Although the gene is known to be expressed in the brain (Su et al. 2004), the function of CD2AP is not completely understood. It has roles in the immune system (stabilising interactions between T cells and antigen presenting cells (Dustin et al. 1998)), cellular structural organisation (anchoring various cellular components to the actin cytoskeleton), cell adhesion, endocytosis and apoptosis (Monzo et al. 2005). Links to innate and adaptive immunity, vesicle trafficking, endocytosis and synaptic plasticity may provide insights in to how the gene is related to AD pathology. Alternatively, the gene is implicated in renal disease, which commonly leads to hypertension and can cause neurovascular damage, so the gene's link with AD may be an indirect one, altering susceptibility to known AD risk factors (Shih et al. 1999; Grunkemeyer et al. 2005).

The following sections (1.9. *CLU*, 1.10. *PICALM* and 1.11. *CR1*) are largely based on chapters written for the book Genetic Variants in Alzheimer's disease, published in 2013 by Springer New York (Morgan 2013).

1.9. *CLU*

An Introduction to *CLU*

CLU is one of the most robustly evidenced genetic risk factors for AD after Apoε4, given its high level of significance in multiple large GWAS. Although these were not the first time CLU had been implicated in AD, the GWAS evidence brought a new fervour to the investigation of how CLU could be mechanistically involved in the pathology of AD, and a drive to discover the specific genetic variations which convey the observed alteration in disease risk.

CLU is often referred to as an "enigmatic" molecule, as it plays a role in a wide variety of physiological functions, including lipid metabolism, complement inhibition, sperm maturation, DNA repair and cell cycle control. Cholesterol and Aβ metabolism, neuroinflammation and apoptosis are all strong candidate pathways linking CLU's function to AD, but despite extensive study, it remains unclear which of CLU's numerous biological roles is responsible for its relationship with AD risk.

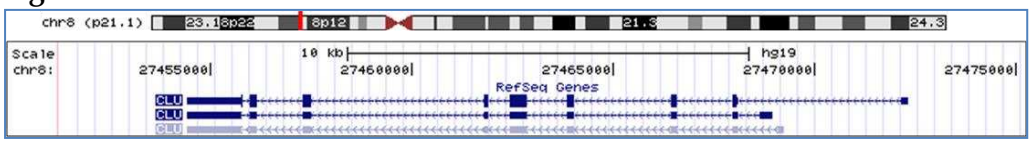
Clusterin – genetics and regulation

CLU is also known as apolipoprotein J (ApoJ); complement lysis inhibitor (CLI); sulfated glycoprotein 2 (SGP-2); testosterone-repressed prostate message 2 (TRPM-2); and secreted protein 40,40 (SP-40,40). This spectrum of nomenclature has arisen as a consequence of CLU's diverse physiological functions and wide-spread expression, which lead to independent "discovery" in a variety of species and contexts. CLU was originally identified in 1983 in the fluid of the rete testis of rams (Blaschuk et al. 1983), with ApoJ (de Silva et al. 1990), CLI (Jenne and Tschopp 1989) and SP-40,40 (Kirszbaum et al. 1992) subsequently identified in human serum, eventually coming to be considered a single protein species - CLU.

The *CLU* gene (NCBI - NG_027845.1, Ensembl - ENSG00000120885), comprising 9 exons, is situated on chromosome 8p21-p12, spanning a region of around 18kb (see Figure 1.5). There is a certain discrepancy between reported isoforms of *CLU* in different online databases and in the literature. NCBI lists three different transcripts for *CLU*, only one of which is said to be coding: 2877bp Isoform 1 (NM_001831.3, encoding NP_001822.3). Ensembl lists 21 transcripts, three of which are classed as coding (ENST00000316403 at 3080bp, ENST00000523500 at 2381bp and the 2098bp ENST00000405140), all of which are said to give rise to a single protein isoform of 449aa - CCDS47832. Reporting at the 5th International CLU workshop, Trougakos et al. stated there were two alternatively spliced *CLU* transcripts, the main gene transcript (termed isoform two) corresponding to the major, secreted, form of the CLU protein, and a second transcript, alternatively spliced to exclude exon two (Trougakos et al. 2009). Ling et al., in a recent paper investigating CLU isoforms and AD risk, reported the expression of two transcripts in the brain

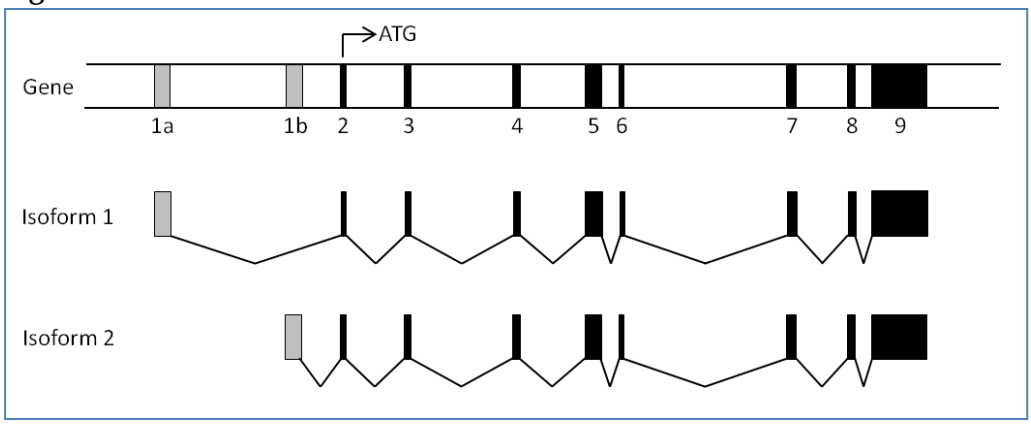
(termed *CLU1* and *CLU2* (with *CLU2* analogous to Trougakos's isoform two)). These were identical in exons 2-9 but had different, untranslated, first exons and different proximal promoters, and the expression of *CLU2* was consistently higher than *CLU1* (Ling et al. 2012). The structure of these isoforms is shown in Figure 1.6. This view of *CLU* transcripts differing in exon one but sharing identical translated exons is supported elsewhere in the literature, where a putative third transcript has been suggested which shares exons 2-9 with *CLU1* and *CLU2* but has another, different exon one (Andersen et al. 2007; Rizzi and Bettuzzi 2010). Leskov et al. have also reported a transcript lacking exon two (Leskov et al. 2003). Because the main isoforms only vary in the first, untranslated exon, differences between them are thought to be regulatory rather than coding. Indeed, there is evidence that different isoforms are differentially regulated by various stimuli (Cochrane et al. 2007; Schepeler et al. 2007), which could well be explained by the presence of alternative promoter regions for different transcripts.

Figure 1.5 - Genetic location of *CLU*



Location of the *CLU* gene on chromosome 8p21-p12 and the transcripts of the gene according to RefSeq. Image taken from UCSC Genome Browser (Kent et al. 2002) (http://genome.ucsc.edu/).

Figure 1.6 - *CLU* Isoforms



The structure of the two most commonly reported and experimentally confirmed isoforms of CLU, adapted from Rizzi 2010 (Rizzi and Bettuzzi 2010). Black boxes represent exons which are consistent between the two transcripts, while the gray boxes indicate the differing first exons. The translation start codon resides in the second exon, meaning both transcripts are thought to give rise to identical proteins.

Expression of CLU occurs in almost all mammalian tissues, with different levels of expression characteristic of specific tissue types. Expression within the brain is relatively high, along with the liver, testes and ovaries (de Silva et al. 1990). Within the brain, expression appears to be highest in astrocytes, which secrete CLU (Pasinetti et al. 1994; Saura et al. 2003). CLU is expressed at

low levels by neurons (Charnay et al. 2008), and shows regional variation in expression in different areas of the brain (Pasinetti et al. 1994). Morgan et al. demonstrated expression of CLU changes over the course of normal aging, with expression increasing in the corpus callosum and caudate-putamen (both white matter rich regions) and decreasing in interior and peripheral regions of the hilus (Morgan et al. 1999). Expression of CLU in mammals begins prenatally (around the 14th day of gestation in mice (Charnay et al. 2008)) and persists throughout adult life.

Lymar et al. demonstrated in rats that in cell types which normally express CLU at low levels, the proximal 266bp or 311bp of the *CLU* promoter are sufficient to give maximal expression of reporter genes (Lymar et al. 2000). However, in cell types which normally express high *CLU* levels, in this case Sertoli cells, a region from -426 to -311 was also needed for maximal reporter gene expression (Lymar et al. 2000). The *CLU* promoter features a number of binding sites for various stress related transcription factors, indicative that the expression of *CLU* can be modulated in response to various stressors, and it can also be affected by immune related molecules such as cytokines IL1β and IL2 (Zwain et al. 1994). TGFβ has been shown to be able to up-regulate *CLU* expression (Jin and Howe 1997), and a number of TGFβ inhibitory elements exist within the promoter and first intron of the gene (Michel et al. 1995). A multitude of other molecules have been shown to be able to regulate *CLU* expression, including heat shock factors (Michel et al. 1997), c-myc (Thomas-Tikhonenko et al. 2004), n-myc (Chayka et al. 2009), NFκB (Li et al. 2002), members of the AP1 complex (Jin and Howe 1999), insulin-like growth factor-1 and its receptor (Criswell et al. 2005) and H-ras (Kyprianou et al. 1991; Lund et al. 2006).

CLU is also subject to epigenetic regulation. Nuutinen et al. showed that inhibiting histone deacetylase could induce *CLU* expression, while gene methylation and deacetylation silenced *CLU* expression in neuronal cell lines studied (Nuutinen et al. 2005). Lund et al. demonstrated that in H-ras transformed cells, which have decreased *CLU* expression, methylation levels at the clusterin gene were 20-40% higher than in non-transformed cells (Lund et al. 2006). The group identified a region -560 to -314 (relative to transcription start site) where methylation of CpG dinucleotides was significantly higher in the transformed cells, particularly between -385 and -376. No classical CpG island was found to be present within the *CLU* promoter, but there was one present 14.5kb upstream of the gene (Lund et al. 2006) which was shown to be hyper-methylated in H-ras transformed cell lines. This same region was also shown to have methylation levels two fold higher in the colon and small intestine (medium and low *CLU* expression respectively) compared to the testis (high *CLU* expression) (Lund et al. 2006). Hypo-methylation of the *CLU* promoter region had previously been demonstrated in cells with high overall levels of *CLU* expression (Rosemblit and Chen 1994).

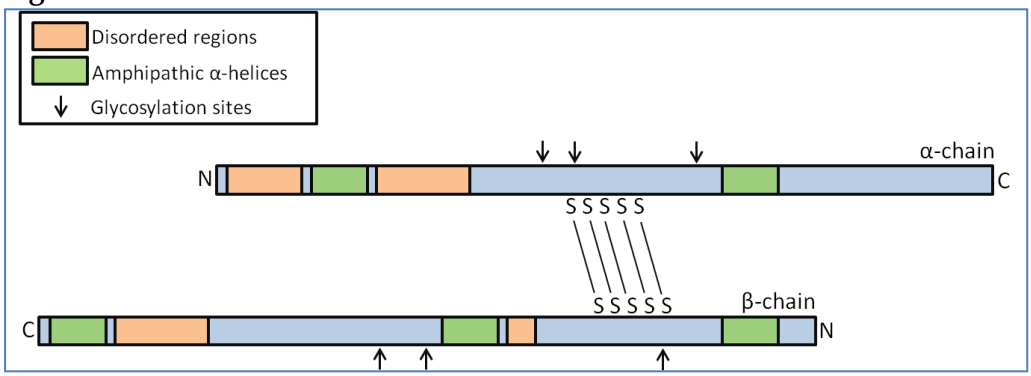
Given *CLU*'s complex expression pattern and wide variety of postulated physiological roles, it is unsurprising such a range of molecules and mechanisms contribute to its regulation.

Clusterin – the protein

There are two forms of CLU – the major form is secreted (sCLU), but there is also a nuclear isoform (nCLU). sCLU is thought to have largely pro-survival functions, while nCLU is pro-apoptotic, expressed in response to the presence of certain stressors.

sCLU is a heavily glycosylated heterodimer, both subunits of which are encoded by the *CLU* gene. Post-translational processing of the full 449 amino acid pre-protein (49kDa) results in a protein with the mature structure shown in Figure 1.7. A 22 amino acid hydrophobic signal sequence at the N-terminus of the full translated protein directs the molecule to the endoplasmic reticulum, where the signal peptide is cleaved and removed. A second cleavage site between Arg205 and Ser206 separates the α and β subunits, which are subsequently bound together by five disulphide bridges (Kirszbaum et al. 1992). On transit from the endoplasmic reticulum to the Golgi apparatus, from which sCLU will be secreted, the protein undergoes heavy glycosylation, giving the molecule its final molecular weight of 70-80kDa.

Figure 1.7 - sCLU Protein Structure



Structure of the mature sCLU protein with domains and glycosylation sites shown, adapted from Rizzi 2010 (Rizzi and Bettuzzi 2010). Other groups have speculated slightly different domains may be present, e.g. Jones 2002 postulated the presence of two coiled coil domains within the protein sequence (Jones and Jomary 2002).

sCLU has been shown to be able to interact with a wide variety of molecules, including lipids, Aβ, immunoglobulins and complement proteins, which has lead to it being implicated in a wide range of physiological processes. However, it has been speculated that rather than reflecting active involvement in these processes and genuine biological functions of CLU, interacting with such a range of molecules may rather reflect its capability to act as a chaperone (Humphreys et al. 1999).

Clusterin as a chaperone

Chaperone molecules are important in managing the aggregation of proteins. Stressed proteins can become partially denatured and unfold, exposing hydrophobic regions which would normally be masked internally. The exposure of these hydrophobic regions confers a propensity to aggregate and precipitate, which if unchecked can have cytotoxic effects.

It has been suggested that the structural and functional characteristics of CLU are indicative that the molecule's main function is as a chaperone (Nuutinen et al. 2009). Since most of CLU is secreted, this chaperone activity would occur largely in the extracellular space, although it may be capable of acting as a nuclear chaperone in times of cellular stress, when the nCLU isoform of the protein (discussed later) is generated.

The amphipathic α-helices seen in CLU's protein structure are typical of chaperone proteins (e.g. small heat shock proteins) (Law and Griswold 1994; Lakins et al. 2002), while the large disordered regions, or molten globule domains, represent flexible protein-protein interacting regions which allow CLU to interact with a low specificity to a range of target molecules (Bailey et al. 2001).

Unlike some chaperones, CLU cannot itself facilitate the refolding of stressed proteins, but can stabilise them, preventing aggregation (Poon et al. 2000). It may also be capable of enabling the clearance of such molecules from the extracellular space via receptor mediated endocytosis (Nuutinen et al. 2009). The stable regions of CLU can specifically interact with targets such as megalin/low density lipoprotein-related protein 2 (LRP2, a cell surface receptor which facilitates the endocytosis of various ligands) (Zlokovic 1996).

In addition to AD, CLU has been implicated in other disorders which feature protein aggregation as a prominent characteristic, including Creutzfeldt-Jakob disease, where it has been detected in prion clusters (Freixes et al. 2004), and in familial amyloidosis, where it has been linked to the prevention of lysozyme aggregation (Kumita et al. 2007).

As well as acting as a chaperone, CLU is speculated to be involved in a host of other processes that may link the protein to AD, including Aβ metabolism, lipid trafficking and metabolism, neuroinflammation and apoptosis. Additionally, CLU's ability to facilitate the transport of various ligands across the blood brain barrier (BBB) may be relevant in AD pathology.

nCLU

The other form of clusterin, nCLU, is less well characterised. Reddy et al. first demonstrated the presence of a 43kDa protein, which they speculated was a non-secretary form of CLU (Reddy et al. 1996). The group identified a second, in frame, ATG codon within the third exon, 99 bases downstream of the ATG from which translation of sCLU begins in exon two. It is thought this arises

from a transcript where the first and third exons are joined by alternative splicing (Leskov et al. 2003), resulting in an mRNA lacking exon two, and thus the normal start codon. Ling et al. failed to find such a transcript in their recent investigation of CLU isoforms in the brain, however, they speculated that since its expression is associated with cell death, it would be transient and therefore not necessarily detectable in the context of high CLU1 and CLU2 expression (Ling et al. 2012).

The nCLU protein lacks the first 33aa of the pre-sCLU protein, the region which contains the hydrophobic signal sequence, meaning the protein is not targeted to the endoplasmic reticulum, and thus is not secreted. Three potential nuclear localisation signals exist in the nCLU protein sequence, but mutational analysis has indicated these are not necessary in establishing the cellular location of nCLU (Scaltriti et al. 2004). There is evidence that nCLU may exist in the cytoplasm as an inactive precursor molecule, with induction and translocation to the nucleus occurring in response to certain stressors (e.g. ionising radiation (Yang et al. 2000), or TGF-β (Reddy et al. 1996) exposure). nCLU contains two coiled coil domains. The N-terminal coiled coil domain appears to be able to bind to the C-terminal one, suggesting the protein may be capable of oligomerisation. nCLU has been shown to be able to bind to Ku-70 (Yang et al. 2000), an interaction which appears dependent on three crucial leucine residues within the C-terminal coiled coil domain (Leskov et al. 2003). This interaction seems to be essential for nCLU's pro-apoptotic functions. Ku-70 is a crucial component of the double-strand DNA repair complex Ku70/Ku80. Binding of nCLU to Ku-70 could disrupt the repair complex, leading to a failure to repair damaged DNA, and ultimately apoptosis of the cell (Leskov et al. 2003).

As CLU is expressed in virtually all tissues and shows a high degree of conservation across mammalian species (see Figure 1.7), it may be assumed that its role is one of fundamental importance biologically. However, despite this, and despite CLU's suggested involvement in such a wide variety of physiological processes, experiments with CLU knockout mice have shown the absence of CLU is well tolerated, with mice developing and living normally (McLaughlin et al. 2000; Charnay et al. 2008). The lack of overt phenotype in CLU knockout mice may reflect the ability of other molecules to compensate for its absence. Other apolipoproteins may be able to fulfil some of CLU's roles, similar to how CLU has been speculated to compensate for ApoE deficiency in knockout mice (Anderson et al. 1998). McLaughlin et al. initiated myosin-induced autoimmune myocarditis in wild type and CLU knockout mice, and found a similar initiation of humoural and cell mediated inflammatory responses between the two. However, the severity of the inflammatory response was significantly increased, and significantly more cardiac tissue injury and long term impairment of cardiac function was observed in the CLU deficient animals (McLaughlin et al. 2000). These results suggest CLU may play a protective role against post-inflammatory destruction of tissue in autoimmune myocarditis (McLaughlin et al. 2000).

Imhof et al. induced ischemic cerebral injury in mice, which gave rise to long lasting expression of CLU in the astrocytes of wild type animals. CLU knockout mice displayed significantly slower tissue remodelling during recovery from such injury than did wild type mice (Imhof et al. 2006). Each of these findings of impaired recovery in CLU knockout animals implies a protective role of CLU, aiding in the recovery of tissues from various assaults. Impairment of such mechanisms could well contribute to the tissue damage and neuronal cell death observed in AD.

Figure 1.7 - Conservation in the CLU region

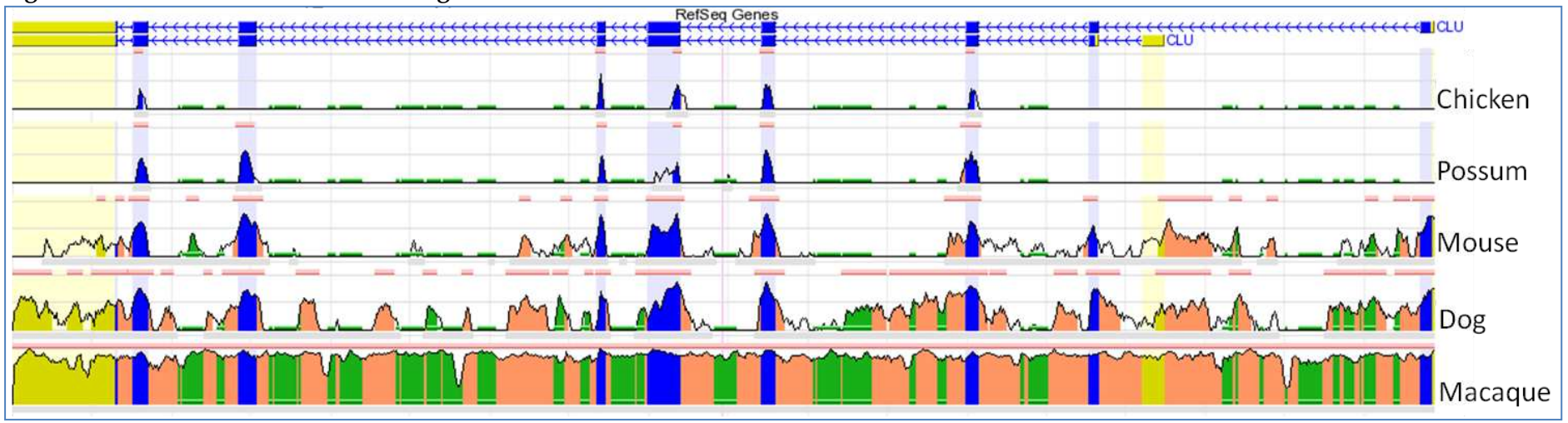


Image to show genetic conservation across selected vertebrate species in CLU and the surrounding chromosomal region, taken from ECR browser (Ovcharenko et al. 2004) (http://ecrbrowser.dcode.org/). Blue – exons, pink – introns, yellow – untranslated regions, red – intergenic regions, green – repetitive DNA elements. The height of the graph is proportional to the level of conservation between human and that particular species (shown on right).

As well as being associated with AD, there are a number of other disorders CLU has been implicated in, often showing altered expression levels in the disease condition versus controls. Such conditions include atherosclerosis (Ishikawa et al. 1998), systemic lupus erythematosus (Newkirk et al. 1999), type 2 diabetes (Kujiraoka et al. 2006), heart disease/myocardial infarction (Vakeva et al. 1993; Poulakou et al. 2008), polycystic kidney disease (Harding et al. 1991) and rheumatoid arthritis (Devauchelle et al. 2004), many of which feature inflammatory or autoimmune aspects. Gao et al. speculated that AD and Parkinson's disease (PD) may share common genetic risk factors, given that many PD patients suffer from dementia, and Parkinsonian movements are often seen in AD patients (Gao et al. 2011). The group found that the same SNP within *CLU* that was associated with AD (rs11136000) was also associated with PD, an effect which seemed independent of CLU's effect on dementia risk, suggesting the two disorders do indeed share some common aetiological factors (Gao et al. 2011).

With roles in both promoting cell survival and inducing apoptosis, it is unsurprising CLU has drawn extensive attention in relation to cancer development, progression and susceptibility treatment. Changes in CLU expression levels are seen in many cancers, including those of the prostate (Miyake et al. 2000), breast (Redondo et al. 2000), colon (Pucci et al. 2004) and bladder (Miyake et al. 2001). In general, CLU expression often appears to be decreased in naive cancer cells, with increased expression in cancers which have developed resistance to conventional treatments (Miyake et al. 2000; Cappelletti et al. 2008). It has been speculated that up-regulation of CLU may be part of the mechanism by which breast cancer cells become resistant to anti-oestrogen therapies, and thus down-regulation of CLU, in combination with conventional cancer treatments, may help combat the ability of tumours to evade the cytotoxicity of anti-cancer therapies. OGX-011 is an antisense 21 base oligonucleotide targeted against the exon two region of CLU mRNA, which contains the translation start site (Chi et al. 2005). A number of studies have been conducted using OGX-011 in combination with usual cancer therapies, but while some have yielded promising results (So et al. 2005; Laskin et al. 2012), others have seen little effect beyond the expected response to the conventional treatments alone (Chia et al. 2009).

Clusterin and AD

In addition to the evidence from GWAS that CLU is involved in AD risk, a host of other research connects the gene/protein to the disorder. CLU was first implicated in AD back in 1990, when May et al. demonstrated expression of the gene in the hippocampus was significantly increased in AD when compared to healthy controls (May et al. 1990). CLU protein levels have also been shown to be higher in the frontal cortex and hippocampus of AD patients (Lidstrom et al. 1998). Early studies failed to find a link between AD and CLU levels in cerebrospinal fluid (CSF) (Harr et al. 1996), however, using newer techniques, it has subsequently been shown that CLU is significantly increased in the CSF of AD patients (Nilselid et al. 2006; Sihlbom et al. 2008;

Thambisetty et al. 2012) thereby possibly indicating its utility as a diagnostic biomarker. The protective allele of rs11136000 has been shown to be associated with increased cognitive performance in the "oldest old" (92-93 years at recruitment) (Mengel-From et al. 2010). Recently, plasma clusterin levels have been shown to be associated with brain atrophy both in AD (Mengel-From et al. 2010) and in mild cognitive impairment (MCI) (Thambisetty et al. 2012), the latter of which is indicative of an early role for CLU in the neurodegeneration seen in AD patients. CLU expression was also shown to be linked to disease severity (Thambisetty et al. 2010; Schrijvers et al. 2011) and clinical progression (Thambisetty et al. 2010) in AD patients. CLU expression has been shown to be increased in neurons as aging (a major risk factor for AD) occurs (Grassilli et al. 1992). Disregulation of epigenetic control may be central in conditions such as AD (Wang et al. 2008), and as discussed previously, CLU expression may be controlled largely by epigenetic mechanisms, again, linking CLU to potential pathogenic mechanisms in AD. CLU has been shown to be present in the amyloid plaques characteristic of AD (McGeer et al. 1992), but is absent from neurons containing neurofibrillary tangles (Giannakopoulos et al. 1998).

Contradictory findings have been reported with regards to the effect of *APOE* genotype on CLU expression levels. Harr et al. reported a significant decrease in CLU expression in the frontal lobes of AD patients with the *APOE* ε4/ε4 genotype (Harr et al. 1996). However, Bertrand et al. reported that the decreased *APOE* expression in ε4/ε4 AD patients was accompanied by an increase in CLU expression, speculating that this may be some kind of compensatory mechanism, since the two apolipoproteins have overlapping functions (Bertrand et al. 1995). These apparently contradictory findings could be due to a variety of factors. It could reflect the inconsistent roles of different CLU isoforms, or varying effects of CLU at different stages of the disease, but it may simply reflect differences in experimental design.

Although it has been overwhelmingly demonstrated that *CLU* is a genetic risk factor for AD, the exact nature of its relationship with AD, and how this alteration in risk is conveyed remains unclear.

Clusterin and Aβ

Despite a recent shift away from the amyloid cascade hypothesis, it has been the prevailing theory of AD pathogenesis for around two decades, and amyloid plaques constitute one of the two major pathogenic hallmarks of AD. With this in mind, the relationship between any AD associated gene and Aβ cannot be ignored.

There is a wealth of evidence linking CLU and Aβ. CLU is present in amyloid plaques (McGeer et al. 1992), CLU can bind to both Aβ peptides and fibrils in CSF (Ghiso et al. 1993), and CLU can interact with Aβ-40 and Aβ-42 in vitro (Matsubara et al. 1996). CLU and ApoE together have been shown to be capable of suppressing Aβ plaque formation, and levels of soluble and

insoluble Aβ in the brain (DeMattos et al. 2004). Recently, increased plasma CLU levels have been shown to be positively associated with the burden of fibrillar Aβ in the entorhinal cortex (Thambisetty et al. 2010). It has been suggested that CLU may be capable of masking early Aβ aggregates from recognition by the immune system, which could minimise the potentially harmful effects of invoking an immune response against protein clusters within the brain (Nuutinen et al. 2009).

Many studies have been published which consider the effect of CLU on Aβ solubility and aggregation, with some apparently contradictory findings. Much evidence suggests CLU can enhance the solubility of Aβ, preventing its oligomerisation and inhibiting the formation of fibrillar structures (Oda et al. 1994; Matsubara et al. 1996). However, DeMattos et al. demonstrated using a mouse model that CLU can actually enhance Aβ aggregation and plaque formation (DeMattos et al. 2002). It is thought the effect may be dependent on the relative proportions of CLU and Aβ present, with low CLU:Aβ ratios leading to a promotion of aggregation and plaque formation, and a high CLU:Aβ ratio decreasing aggregation and maintaining solubility (Yerbury et al. 2007).

It has been speculated that CLU may be involved in the clearance of Aβ from the brain. The first way in which this could occur is via endocytosis. Hammad et al. demonstrated that complexes of CLU and Aβ can be internalised by cells, dependent on megalin, and that the Aβ can then be broken down via lysosomal degradation (Hammad et al. 1997). It has also been shown that accumulation of fibrillar Aβ increases CLU expression, and this is accompanied by an increase of endocytosis of fibrillar Aβ in astrocytic cell lines, although CLU is not itself necessary for this phagocytosis (Nuutinen et al. 2007). Endocytosis and Aβ degradation would reduce the overall levels of Aβ within the brain, potentially having a protective effect on cells. Secondly, CLU may be able to facilitate clearance of Aβ across the blood brain barrier (BBB). It has been demonstrated that CLU-Aβ complexes can cross the BBB (Zlokovic 1996). Bell et al. found that administering CLU-Aβ42 complexes to mouse brains gave an almost two-fold increase in clearance rate across the BBB compared to Aβ42 alone, an effect that was disrupted by antibodies against megalin (Bell et al. 2007). Interestingly, the known AD risk allele of *APOE*, ε4, has been shown to be less efficient in Aβ clearance via this mechanism (Bell et al. 2007). Taken together, this may suggest clearance of Aβ across the BBB, aided by CLU, represents an important means of reducing the brain burden of Aβ, resulting in a protection of the brain from the neurotoxic effects of Aβ.

Of course, CLU's relationship with Aβ could simply be a reflection of its capacity to act as a chaperone. It interacts with a plethora of other molecules but because Aβ is so strongly linked to AD, this particular relationship is subject to intense scrutiny, perhaps without being directly relevant to AD development at all.

Clusterin, the cell cycle and apoptosis

The symptomatic changes seen in AD occur as a result of the massive neuronal loss associated with the condition, although the cause of this loss remains to be elucidated. Regulators of the cell cycle and apoptosis could affect the way in which cells cope with stress, and thus mediate the extent of the neuronal destruction incurred when neurons are exposed to these unknown AD triggers. DNA damage and apoptotic features have been linked to AD for many years, with speculation that neurons in AD-affected regions may be in a struggle between apoptosis and repair (Cotman and Su 1996).

The major, secreted, form of CLU is known to have largely pro-survival functions, which could be of great significance to AD pathology, since the symptoms stem from cell death. Alteration of CLU's pro-survival properties could affect the survival capacity of neurons, and thus affect how resilient a brain will be to AD type changes.

Much research has been done on the effect of CLU on the cell cycle, but it is important to remember neurons are terminally differentiated, and thus post-mitotic, therefore such effects are likely to be irrelevant to neuronal survival. However, it may be that CLU's effect on the cell cycle can indirectly enhance neuronal survival. It has been demonstrated that CLU can increase the proliferation of primary astrocytes in culture (Shin et al. 2006; Shim et al. 2009). Astrocytes in affected areas of AD brains have been shown to have up-regulated CLU expression, which, if it causes similar proliferative effects *in vivo* as *in vitro*, could create a pro-survival feedback mechanism, supporting neuronal survival (Nuutinen et al. 2009).

Clusterin and lipid metabolism

Deregulation of processes involved in lipid metabolism and transport are increasingly being seen as potential causes of the pathogenic features seen in AD. It has long been observed that higher cholesterol levels in middle age are linked to an increased incidence of AD later in life, and that use of statins, which lower cholesterol, reduce AD risk (Jick et al. 2000). *APOE*, the longest established genetic risk factor for AD, is known to participate in lipid trafficking, and CLU has a similar role in this process, reflected in its alternative name of ApoJ. The ε4 allele of ApoE has been shown to be less efficient at transporting cholesterol (Gong et al. 2002), which may indicate impaired lipid transport is of importance in the development of AD.

The brain is an organ rich in insoluble lipids. In order to be transported between cells, these lipids must be solubilised, which is achieved via the binding of various proteins, forming lipoprotein particles. ApoE and ApoJ are two of the main cholesterol transporting molecules within the brain (Beffert et al. 1998). CLU has also been shown to be present in lipoprotein particles in the CSF (Suzuki et al. 2002).

ApoE and CLU are thought to be present in different lipoprotein particles, with ApoE-containing particles being larger, and with higher lipid content than CLU-containing ones. The types of lipid also differ, with approximately equal proportions of phospholipid and cholesterol in ApoE-containing particles, while CLU-containing particles have more phospholipid than cholesterol (DeMattos et al. 2001).

Two studies have previously reported potential associations between polymorphisms within CLU and lipid levels (Nestlerode et al. 1999; Miwa et al. 2005), raising the interesting possibility that CLU could exert its effect on AD indirectly, affecting susceptibility to other AD risk factors such as cardiovascular disease and atherosclerosis (Yu and Tan 2012).

Clusterin and neuroinflammation

It has long been observed that neuroinflammation is a key characteristic of AD. Plaques are commonly surrounded by inflammatory and immune antigens, activated microglia, astrocytes and complement. What is becoming increasingly appreciated is that instead of being incidental bystanders in AD, inflammation and the immune response could be early, possibly causative processes in AD pathology, as discussed in section 1.2. AD in the clinic – Prevention.

There are several links between CLU and inflammation/immunity. CLU is important in the regulation of complement activation (Jenne and Tschopp 1989), it can modulate the membrane attack complex (Kirszbaum et al. 1992) and can activate microglia (Xie et al. 2005). It can also regulate important modulators of the immune response, such as NFκB (Takase et al. 2008), and its own expression in astrocytes can be regulated by cytokines such as IL-1β and IL-2 (Zwain et al. 1994). As mentioned previously, CLU is thought to be able to mask growing Aβ plaques from immune recognition (Nuutinen et al. 2009).

It seems from the NSAID's evidence that reducing immune responses within the brain is protective against AD risk. CLU is clearly able to limit the immune response, directly (e.g. by preventing complement activation) and indirectly (e.g. by masking Aβ aggregates), raising the possibility that CLU's effect on AD risk is via its involvement in inflammation and immune responses.

Clusterin as a neuroprotective guardian

Many of CLU's functions suggest a largely neuroprotective role for the protein. Neurofibrillary tangle-free neurons which express CLU have been shown to be resistant to cell death (Giannakopoulos et al. 1998). CLU is up-regulated following many types of brain injury, including pathogenic conditions such as AD, and experimental lesions (May et al. 1990), implying a protective role. It has already been discussed that following experimental induction of ischemic cerebral injury, CLU knockout mice showed impaired tissue remodelling and had delayed recovery when compared to wild type mice, again supporting a protective role for CLU when faced with tissue

damage (Imhof et al. 2006). Interestingly, this protection could stem from CLU's pro-survival functions, reducing apoptotic cell death, from its lipid transporting capacity, since cellular damage requires lipids for repair and remodelling, or from its relationship with the immune response, limiting inflammation and modulating damage in this way.

However, CLU's role as a neuroprotective factor is not undisputed. CLU has been shown to accumulate in dying neurons following seizures and neonatal hypoxic ischemia (a model for cerebral palsy), leading Han et al. to investigate the role of CLU in this cell death. The group found CLU knockout mice incurred around 50% less neuronal injury than WT mice following neonatal hypoxic ischemia, implying that CLU normally exacerbates cell death, a finding that was confirmed by CLU increasing cell death in response to oxygen/glucose starvation *in vitro* (Han et al. 2001). This was shown to be independent of caspase-3, a key protein in apoptosis (Han et al. 2001).

Clusterin as a therapeutic target

With an unequivocal role in AD, CLU must surely be considered as a potential target for therapeutic intervention. CLU based therapies are already under development for cancer (e.g. OGX-011). The major aim of these is to reduce levels of CLU, which has largely pro-survival properties, in the hope this will render cancer cells more susceptible to treatment. However, it is likely that in AD it is the pro-survival functions that would need to be enhanced, not diminished, so this approach is unlikely to be beneficial here.

There has been some evidence of CLU based therapies having beneficial effects in treating atherosclerosis and peripheral neuropathies in animals (Navab et al. 2005; Dati et al. 2007), but AD presents the additional challenge of requiring a method of delivery that can traverse the blood brain barrier. Since it remains unclear which of CLU's functions are important in the development of AD, it is difficult to know what aspect of its action to target, and how this would affect its other functions, perhaps leading to unacceptable side effects. The area is further complicated by the existence of different CLU isoforms with apparently opposing functions. Significant further research is needed in the area to explore the full therapeutic potential of CLU in AD.

1.10. *PICALM*

An introduction to *PICALM*

The gene encoding phosphatidylinositol binding clathrin assembly protein (PICALM) has sometimes been overlooked by the scientific community, regarded as a homologue of the neuron specific Adaptor Protein 180 (AP180) with a more widespread expression but equivalent function. However, as increasing differences between the two proteins come to light, attention is turning to *PICALM,* whose roles in cancer, growth and development, haematopoiesis and now neurodegeneration make it a fascinating target for

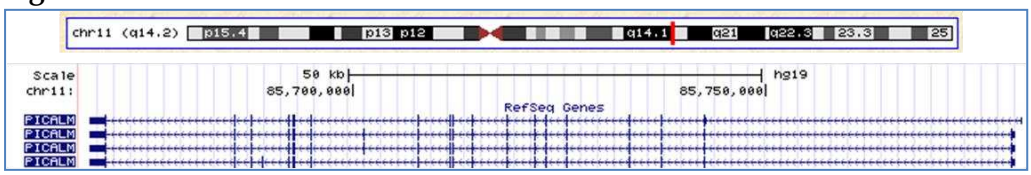
study. One of PICALM's major roles is in CME, an indispensible step in intracellular trafficking of proteins and lipids, to which other AD associated genes (namely *BIN1)* have also been linked.

Despite speculation of late, it is yet to be determined how *PICALM* is mechanistically linked to AD risk, and what the genetic determinants underlying this relationship are.

PICALM – genetics and regulation

The ubiquitously expressed *PICALM* gene (NCBI Gene ID - 8301, Ensembl - ENSG00000073921), whose protein product is also known as clathrin assembly lymphoid myeloid leukaemia protein (CALM), is a ~112kb gene situated on chromosome 11q14 (see Figure 1.8). It was first identified in 1996 when it was found to be involved in a rare but recurrent translocation (t(10;11)(p13;q14)), creating a PICALM/AF10 fusion gene in acute myeloid leukaemia and acute lymphoblastic leukaemia patients (Dreyling et al. 1996).

Figure 1.8 – Genetic location of *PICALM*



Location of *PICALM* on chromosome 11q14, with the gene transcripts according to RefSeq shown below. Image taken from the UCSC Genome Browser (Kent et al. 2002) (http://genome.ucsc.edu/).

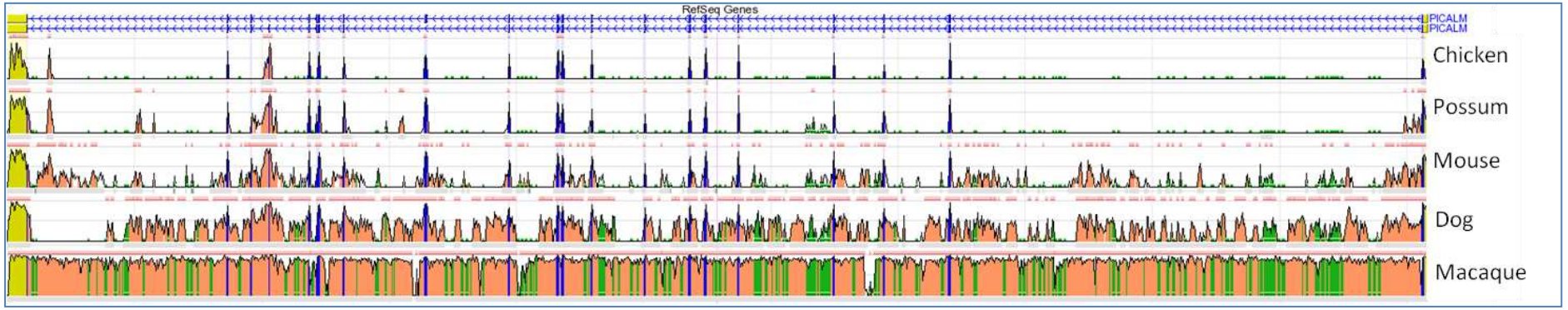
At least three protein isoforms of PICALM exist (Baig et al. 2010), and debatably more. The generally uncontested isoforms are the full length protein at 652aa and a shorter isoform of 610aa. In the literature, a 632aa isoform is also documented (Baig et al. 2010). In online databases (Ensembl and NCBI's RefSeq), this 632aa isoform is not included, and instead a 645aa isoform and a 551aa isoform are reported. The RefSeq transcripts that give rise to these isoforms are shown in the lower panel of Figure 1.8. Clearly the discrepancies between the database and literature are in need of resolution. Table 1.4 shows the isoforms of *PICALM* as per the online databases. Figure 1.9 shows the level of conservation in the *PICALM* region across multiple vertebrate species.

Table 1.4 – *PICALM* isoforms

Isoform*	Encoded Protein Length (aa)	RefSeq Transcript ID	RefSeq Protein ID	Ensembl Transcript ID	Ensembl Protein ID
1	652	NM_007166.3	NP_009097.2	ENST00000393346	CCDS8272
2	610	NM_001008660.2	NP_001008660.1	ENST00000532317	CCDS31653
3	645	NM_001206946.1	NP_001193875.1	ENST00000526033	CCDS55784
4	551	NM_001206947.1	NP_001193876.1	ENST00000528398	CCDS55783

Isoforms of *PICALM*, as reported in online databases, RefSeq and Ensembl.
*Numbered consistent with NCBI RefSeq.

Figure 1.9 - Conservation in the *PICALM* region



Genetic conservation across selected vertebrate species in *PICALM* and the surrounding chromosomal region, taken from ECR browser (Ovcharenko et al. 2004) (http://ecrbrowser.dcode.org/). Blue – exons, pink – introns, yellow – untranslated regions, red – intergenic regions, green – repetitive DNA elements. Height of graph proportional to level of conservation between human and that particular species (shown on right).

PICALM is ubiquitously expressed, unlike its neuronal cousin, AP180. Recent research has suggested that *PICALM*'s major site of expression within the brain is in the endothelial cells of vessel walls, with weak labelling in neurons and glial cells (Baig et al. 2010). Xiao et al., when considering *PICALM* expression in the brains of APP/PS1 transgenic mice found neurons were the major site of expression in the hippocampus and cortex, with no labelling of PICALM in astrocytes or microglia (Xiao et al. 2012). Yao et al. found expression of PICALM in hippocampal and cerebella neurons, dispersed in and around synapses, particularly at the sites of clusters of SVs, while expression of AP180 was largely restricted to the pre-synaptic region (Yao et al. 2005).

Schwartz et al. found that in the rat, *PICALM* expression began as early as the twelfth day of development, in undifferentiated embryonic stem cells, neural stem cells and in post mitotic neurons, implying a role in the development of the nervous system (Schwartz et al. 2010). Indeed, Bushlin et al. had previously provided evidence that PICALM and AP180 were involved in the normal development and growth of hippocampal neurons (Bushlin et al. 2008). Schwartz's group also looked at the expression of both long and short isoforms of *PICALM,* and found expression of each, but following opposite trends. The long isoform increased in expression while the short isoform decreased in expression as the brain developed (between the twelfth and eighteenth days of gestation) (Schwartz et al. 2010). During this time there is a transition from neural progenitors being the predominant cell type to post-mitotic neurons being most abundant. This may indicate two things. Firstly, that in these two cell types, PICALM plays differing roles, and secondly, that the two PICALM isoforms observed have themselves different roles, which should be considered when investigating the function of the protein. Only two transcripts were detected, despite the presence of a third PICALM protein with a higher molecular weight, speculated to be due to post translational modifications.

PICALM - the protein

The main function of PICALM appears to be in CME, a process which over expression of PICALM can inhibit (Tebar et al. 1999). CME allows the internalisation of surface bound ligands, such as proteins and lipids, facilitating their intracellular trafficking (Hollingworth et al. 2010). It has been reported to be involved in regulating the protein content of the cell membrane, managing the insertion and removal of receptors, which could be particularly important in neurons where it may provide a mechanism for modulating synaptic strength (Man et al. 2000; Wang and Linden 2000). It is also important in maintaining sustained neurotransmission, as CME is a mechanism for recycling SVs following neurotransmitter release (Jung and Haucke 2007). CME mainly traffics its cargo via clathrin coated vesicles (CCVs), which can transport molecules from the cell membrane to early endosomes, or from the trans golgi network to late endosomes (Nordstedt et al. 1993). These processes

are under tight regulation by a variety of factors, in order that transportation occurs efficiently and accurately.

The clathrin coats of CCVs are formed from networks of clathrin triskelions, which consist of three clathrin heavy chains and three clathrin light chains (Wu and Yao 2009). The C-terminal region of PICALM can bind to the clathrin heavy chain, and to AP2, while the N-terminal region binds to phosphatidylinositol-4,5- bisphosphate, which is present in the plasma membrane, and thus PICALM may have a role in recruiting clathrin and AP2 to the cell membrane (Ford et al. 2001). It is the N-terminal, membrane binding, region of PICALM which shares the greatest homology with AP180 (~82%) (Miller et al. 2011). This is termed the ANTH (AP180 N-terminal homology) domain.

Kim et al. showed that PICALM purified from rat livers was able to promote the assembly of clathrin triskelia into clathrin cages, a function it bears in common with AP180 (Kim and Kim 2000). PICALM is able to interact with different sites of the clathrin heavy chain, allowing it to regulate the size and shape of the budding CCV by dictating the degree of curvature in the clathrin coat (Tebar et al. 1999). Meyerholz et al. demonstrated that knockdown of *PICALM* expression via RNA interference (RNAi) lead to an excess of particularly small vesicles forming, and the normally uniformly round vesicles showing a tendency to elongate and form tubular structures (Meyerholz et al. 2005). In *Drosophila*, deletion of an AP180/PICALM homologue was shown to lead to disruption of the normal localisation of clathrin in nerves, and reduced endocytosis of SVs (Zhang et al. 1998).

Klebig et al. reported fit1 mice, with mutations in *PICALM*, showed a decreased lifespan and growth retardation, along with numerous haematopoietic abnormalities, manifesting in severe anaemia and a decreased white blood cell count (Klebig et al. 2003), suggesting potential roles for PICALM in growth, haematopoiesis and iron metabolism. There are a number of different fit1 mice, each with a different mutation within *PICALM*, and with phenotypic severity related to the severity of the mutations (Klebig et al. 2003).

Suzuki et al. recently found similar results in their study investigating PICALM deficiency in murine development (Suzuki et al. 2012). While mice with heterozygote PICALM deficiency exhibited no discernable phenotype, the vast majority of homozygotes died between birth and weaning, with those that survived longer still having a shortened lifespan relative to wild type mice. PICALM deficient embryos were found to weigh just 74% of those with normal PICALM expression, and by 28days, they weighed just 30-40% of their wild type littermates, indicating PICALM is important for growth both *in utero*, and after birth. In PICALM deficient mice, there was evidence of cortical atrophy and enlargement of the ventricles, although the hippocampus appeared unchanged (Suzuki et al. 2012). The group also found an effect on

haematopoiesis – the mice were severely anaemic, with fewer red blood cells, and lower than normal haemoglobin levels (Suzuki et al. 2012).

There is evidence that PICALM is important in early development and growth of neurons, since PICALM deficient cells have been shown to lack normal dendrite structures (Bushlin et al. 2008; Schwartz et al. 2010). Conversely, AP180 lacking cells develop normal dendrites but do not show normal axonal development (Schwartz et al. 2010). Neither of the genes appear to have an influence on the proliferation of neuronal progenitors, rather on their correct development and morphology (Schwartz et al. 2010). Whether this effect occurs as a result of the disruption of CME, or through some alternative mechanism, remains to be elucidated.

Potential role in AD

Since *PICALM* has been implicated in AD risk, a number of studies have been published which have looked for links between SNPs in *PICALM* and various aspects of the disease.

Biffi et al. found significant associations between the GWAS SNP rs3851179 and neuroimaging measures ascertained by MRI scan (Biffi et al. 2010). Both overall hippocampal volume and entorhinal cortex thickness were associated with the SNP. Indeed, this finding has been corroborated by a study in which Furney et al. found that the protective allele of the *PICALM* GWAS SNP was related to an increased thickness of the entorhinal cortex (Furney et al. 2010).

A number of studies have looked for a relationship between *PICALM* SNPs and CSF biomarker levels. Schjeide et al. found that the risk allele of *PICALM* SNP rs541458 was associated with a dose dependent decrease in levels of CSF Aβ-42 (Schjeide et al. 2011), with homozygotes for the risk allele showing around a 20% reduction in CSF Aβ42 levels, which they speculated could give a clue as to the pathogenic mechanism by which *PICALM* is linked to AD (Schjeide et al. 2011). Kauwe et al., however, failed to find any associations between *PICALM* SNPs and CSF levels of Aβ-42 (Kauwe et al. 2011), and although there was some suggestion of an association between *PICALM* SNPs and levels of tau in the CSF, this was not strong enough to withstand correction for multiple testing (Kauwe et al. 2011). Kok et al. reported that rs3851179 was significantly associated with plaque load in post-mortem brains, with the allele associated with a lower risk of AD also emerging as protective against amyloid plaques (Kok et al. 2011). There is some evidence that the SNP rs3851179 may also be associated with cognitive function. Mengel-From et al. found the protective allele of the SNP was associated with better cognitive function in the "oldest old" (92-93 years of age at the time of enrolment), but in male subjects only (Mengel-From et al. 2010).

Disruption to the endocytic pathway has been reported as one of the earliest detectable changes in AD, preceding the initiation of plaque deposition (Cataldo et al. 2000). Whilst there is no evidence that PICALM is present in

plaques and tangles (Baig et al. 2010), there is some evidence that PICALM expression is increased in the frontal cortex in AD (Baig et al. 2010). It was speculated this up regulation could be as a result of increased Aβ, but in that case, there would also be an increase in expression expected in the temporal cortex, which was not observed (Baig et al. 2010). Thomas et al. reported an approximately 2.4 fold increase in PICALM expression (along with increases in other CME related proteins, clathrin and dynamin) in the cortex of mice expressing the Swedish mutation form of human APP compared to wild type littermates (Thomas et al. 2011). In contrast to this, it has been shown that PICALM can be cleaved and degraded by calpain, a protease which is elevated and activated in AD brains, and has been shown to be able to block CME (Kim and Kim 2001; Rudinskiy et al. 2009).

Disruption to APP processing

Jun et al., in their meta-analysis (see Table 1.3), found that when data was adjusted for the presence of at least one *APOE* ε4 allele, the evidence for association between *PICALM* SNPs and AD was greatly reduced (Jun et al. 2010). *PICALM* was seen to affect AD risk largely in ε4 positive subjects alone, leading to the speculation that *APOE* and *PICALM* may interact synergistically (Jun et al. 2010). It is worth noting that other groups seeking epistatic interactions between *PICALM* and ApoE have failed to detect an effect (Belbin et al. 2011; Lambert et al. 2011). However, if there is a genuine interaction between the genes, it suggests they both participate in a common pathway that contributes to the development of AD. Since there is compelling evidence linking each of the two genes to Aβ production and metabolism, this could constitute said pathway. As mentioned above, Schjeide et al. found a link between *PICALM* SNPs and levels of Aβ42 in CSF (Schjeide et al. 2011), while Kok et al. demonstrated a link between *PICALM* SNPs and plaque load (Kok et al. 2011). Both of these findings strengthen the evidence that *PICALM*'s effect on AD risk might arise through its relationship with Aβ metabolism. Some of the evidence linking PICALM to APP processing and the production of Aβ is documented below.

The production of Aβ, generated by the cleavage of APP with β- and γ-secretases, is reliant on the endocytic pathway and internalisation of APP (Koo and Squazzo 1994; Vetrivel and Thinakaran 2006). There is evidence that APP is subject to CME (Nordstedt et al. 1993), immediately linking the protein mechanistically with PICALM. Mutational analysis of the cytoplasmic domain of APP, thought to contain an internalisation signal (Chen et al. 1990), leads to decreased endocytosis of the protein, and consequentially, reduced Aβ release (Koo and Squazzo 1994). A number of other studies have also demonstrated decreasing endocytosis can decrease Aβ production or release (Carey et al. 2005; Cirrito et al. 2008; Xiao et al. 2012), while the converse is also true, with increased levels of endocytosis increasing Aβ levels (Grbovic et al. 2003; Cirrito et al. 2008; Xiao et al. 2012).

Alterations in PICALM which affect endocytosis may affect the subcellular distribution of APP, or the secretase enzymes that process it, potentially leading to disturbances in Aβ production (Miller et al. 2011), but while endocytosis generally is clearly liked to APP processing and Aβ production, PICALM's involvement in this remains more controversial.

Xiao et al. looked at the relationship between PICALM, APP processing and plaque pathogenesis in a cell culture model of APP processing (neuroblastoma cells over expressing APP) and in APP transgenic mice (Xiao et al. 2012). In the cell line, the group found prior to the initiation of endocytosis, APP was largely confined to the cell membrane, and PICALM to cytosolic vesicles, but once endocytosis was initiated, APP and PICALM co-localised to intracellular vesicles. Similarly, in APP/PS1 transgenic mice, PICALM expression was detected in neurons, co-localising with APP in the hippocampus and cortex (Xiao et al. 2012). As mentioned above, the group were able to show in both their *in vitro* and *in vivo* systems that altering levels of PICALM would alter APP internalisation, and Aβ production and release, and that this was at least partly specific, since uptake of transferrin, also subject to CME, remained unchanged (Xiao et al. 2012). When the group altered the expression of PICALM in six month old mice, and investigated the effects on the brain four months later, they found that decreasing PICALM expression (by ~50% in the hippocampus) decreased levels of soluble and insoluble Aβ in the brain, and caused a trend towards non-amyloidogenic APP processing, while increasing PICALM expression increased hippocampal Aβ, and lead to a shift towards amyloidogenic APP processing. Levels of full length APP were consistent regardless of treatment, indicating that while PICALM affects processing of APP, it does not affect its production (Xiao et al. 2012). A failure to co-immunoprecipitate PICALM and APP may indicate that any interaction between the two is either weak or indirect. The study also reported an effect of PICALM expression on Aβ plaque load in the hippocampus, with decreased expression leading to a decreased plaque load, and *vice versa*, likely due to the effects on Aβ levels (Xiao et al. 2012). Wu et al., however, found that while RNAi knock down of AP180 expression reduced the production of Aβ, knock down of PICALM did not, suggesting it may not have a direct role in the generation of Aβ (Wu et al. 2009). This could potentially have been due to the use of a cell line expressing the Swedish mutant form of APP, which may be processed differently to the wild type form (Wu et al. 2009; Xiao et al. 2012).

Treusch et al. recently conducted a comprehensive study investigating modifiers of Aβ toxicity in yeast, and found twelve yeast genes notably affected Aβ toxicity and had clear human homologues. Three of these were involved in CME, including *YAP1802*, the yeast homologue of human *PICALM* (Treusch et al. 2011). Following up on this finding, the group investigated the *C.elegans* homologue of yeast *YAP1802* and human *PICALM*; *unc-11*. Wild type *C.elegans* have five glutamatergic neurons in their tails, but when modified to express Aβ, there is an age dependent loss of these cells, with only 25% of worms having five intact neurons by day seven.

Simultaneous expression of *unc-11* (the *PICALM* homologue) was shown to increase the number of *C.elegans* which had five intact neurons (Treusch et al. 2011). The group additionally considered the toxic effect of Aβ on cultured cortical rat neurons, and found those containing a lentivirus engineered for *PICALM* expression were partially rescued from cell death caused by Aβ (Treusch et al. 2011). This group thus provided three separate lines of evidence, in three separate model systems, that *PICALM* is able to modulate Aβ toxicity. It was speculated that this may be due to PICALM targeting harmful Aβ for degradation, however, in yeast, no decrease in Aβ levels was detected in cells expressing *YAP1802*, rendering this unlikely. Aβ was found to affect the distribution of clathrin, decreasing the size of clathrin foci at cell membranes, but increasing the number and intensity of these (Xiao et al. 2012), an effect which may be linked to PICALM, given its proposed ability to recruit clathrin to the cell membrane (Ford et al. 2001) .

It has also been commented that since *PICALM*'s expression in the brain may be predominantly in the endothelial cells of vessel walls, it is perfectly situated for a role in the clearance of Aβ across the blood brain barrier (Baig et al. 2010). This is consistent with the finding of Schjeide et al. that the risk allele of *PICALM* SNP rs541458 was associated with decreased levels of Aβ in CSF, perhaps implying the AD risk associated allele is poorer in clearing Aβ from the brain to the CSF (Schjeide et al. 2011).

APP independent links with AD

Although there is strong evidence PICALM may play a role in APP metabolism and transport, APP is just one of a wide range of molecules which are subject to CME. Because APP has been so intrinsically linked to AD historically, it is easy to see why such efforts have been made to characterise its relationship with PICALM. However, we are a long way from knowing all of the molecules with which PICALM interacts and affects, so it is impossible to say which might be involved in the development of AD, and how that involvement comes about.

Perturbations of endocytosis could easily upset the homeostasis of any type of cell, but even more so for neurons, which must continually recycle receptors and neurotransmitters to maintain long term signalling and function (Jung and Haucke 2007).

A number of molecules other than APP display disrupted endocytosis when expression levels of PICALM are altered. The GluR2 subunit of the AMPA receptor shows a small but significant increase in its cell surface levels when PICALM expression is repressed using RNAi (Harel et al. 2011). This had previously been implicated in AD since Aβ can increase AMPA's rate of endocytosis, decreasing surface AMPA receptor presence, and leading to signalling abnormalities and structural changes in neurons (Hsieh et al. 2006). RNAi knockdown of PICALM also affects the endocytosis of EGFR (Huang et al. 2004) and R-SNARE proteins (Harel et al. 2008; Koo et al. 2011; Miller et al.

2011). Changes in PICALM expression levels can also alter the intracellular distribution of many other molecules, such as AP1, mannose-6-phosphate receptor, and transferrin (Meyerholz et al. 2005).

As mentioned above, several studies have been published investigating the relationship between PICALM and SNARE proteins (soluble N-ethylmaleimide-sensitive-factor attachment protein receptor). The SNAREs are a family of proteins, all containing the conserved 60-70 amino acid SNARE motif. They are generally membrane bound, tetramer complexes which are key in mediating the fusion of vesicles, organelles and membranes. There are multiple SNARE proteins in mammalian cells and in order for trafficking to occur accurately, it is imperative that the correct SNARE proteins are present both in the vesicle and the organelle membrane to which it needs to fuse (Miller et al. 2011). As there are a finite number of combinations of SNARE proteins, given the complexity of the sorting task, it is thought regulation of the localisation of the specific SNARE proteins is important in the regulation of the transport process as a whole. They are also crucial in mediating neurotransmitter release, allowing fusion of SVs with pre-synaptic membranes to facilitate signalling, thought to be important in memory formation (Yao 2004). Efficient recycling and sorting of SNAREs with high accuracy is crucial to ensure prolonged neurotransmitter release is possible. It was unclear how these SNARE proteins are endocytosed and sorted with such specificity, but it has recently reported that PICALM may be a key player in these processes (Koo et al. 2011).

Harel et al. first reported a link between R-SNARE protein VAMP2 (vesicle associated membrane protein 2, also known as synaptobrevin 2), the most abundant synaptic vesicle protein (Koo et al. 2011), and PICALM when they demonstrated that over expression of PICALM lead to a reduction in surface VAMP2 by around 20%, while PICALM knockdown using siRNA increased the presence of surface VAMP2 by around 30% (Harel et al. 2008). They did not find co-localisation of the two molecules, leading to speculation that any interaction must be weak or indirect (Harel et al. 2008), although later studies have indicated that PICALM and VAMP2 (as well as VAMP3 and VAMP8) do physically interact via the N-terminal region of the SNARE motif in VAMP2 and the ANTH domain common to both PICALM and AP180 (Koo et al. 2011; Miller et al. 2011). Koo et al. similarly reported a link between PICALM expression and surface VAMP2 – again, suppression of PICALM (and AP180) expression was shown to lead to an increase in VAMP2 present at the neuronal surface, indicative that it is failing to be retrieved and recycled effectively; an effect which seemed specific to VAMP2, since other SV proteins were unaffected (Koo et al. 2011). The effect was more pronounced when both PICALM and AP180 expression were suppressed, perhaps suggesting overlapping functionality in this context (Koo et al. 2011). Miller et al. found the interaction was strongest between PICALM and VAMP8, which is consistent with its higher rate of internalisation (Miller et al. 2011). VAMP8 and VAMP3 are thought to be important in the fusion of endocytic vesicles

with the cell's limiting membrane and early endosomes, while VAMP2 is involved in the rapid fusion and recycling of SVs with the plasma membrane (Antonin et al. 2000; Miller et al. 2011). Knockdown of PICALM was found to cause surface accumulation of all three of the highly related SNARE proteins considered (Miller et al. 2011). The group found PICALM can bind simultaneously to both VAMP8 (and so presumably VAMP2 and 3) and its other established binding partner, phosphatidylinositol-4,5-bisphosphate (Miller et al. 2011).

These studies have provided compelling evidence that PICALM is involved in the endocytosis of at least three SNARE proteins, VAMP2, VAMP3 and VAMP8. Could disrupted endocytosis of these molecules be the underlying cause of PICALM's involvement in AD development? The correct localisation of such molecules is pivotal in ensuring accurate transport of cargoes about the cell, and in facilitating neurotransmitter release.

As already discussed, alterations in PICALM expression levels can upset the intracellular distribution of a variety of molecules, which could be due to deficiencies in PICALM mediated SNARE endocytosis leading to incorrect localisation of SNAREs, disrupting normal transportation, which could contribute to AD pathogenesis.

Disruption of neurotransmitter release and normal synaptic function could also play a major role in the degeneration seen in AD. It has been observed that AD brains have fewer synapses than controls; that synaptic density actually correlates better with cognitive decline in AD patients than does plaque burden; and that synaptic dysfunction may begin in the AD brain at an early stage, even before the loss of synapses and neurons occurs (Fitzjohn et al. 2001; Masliah et al. 2001; Yao 2004). Schoch et al. studied the fusion of SVs with the pre-synaptic membrane in VAMP2 knockout mice, and found there was roughly a tenfold decrease in spontaneous and sucrose stimulated fusion in the absence of VAMP2, and a 100-fold decrease in Ca^{2+} stimulated fusion (Schoch et al. 2001). This highlights the importance of VAMP2 in facilitating neurotransmitter release. If PICALM is a major player in determining the endocytosis of VAMP2, as appears to be the case, genetic changes which alter its function or regulation could affect VAMP2, interfering with normal neurotransmitter release, disruption of which could result in failed communication between neurons, leading to issues with learning and memory, as are seen in AD (Yao 2004). The strongest phenotype in Fit1 mice is seen when there are nonsense mutations in PICALM's ANTH domain – the very domain underlying its interaction with the SNARE proteins (Klebig et al. 2003), perhaps indicative that it is the disruption of the PICALM/VAMP2 interaction that so disrupts normal development and function (Miller et al. 2011). Synaptic dysfunction could also underlie the observed relationship between *PICALM* and cognitive ability (Mengel-From et al. 2010).

1.11. *CR1*

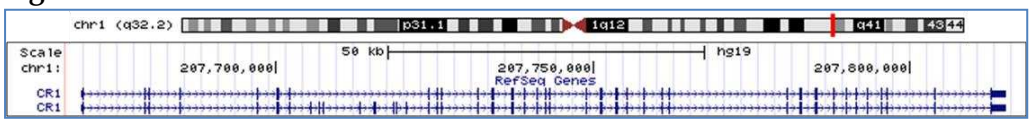
An introduction to *CR1*

Complement Component (3b/4b) Receptor 1 (CR1) is a single chain type I transmembrane glycoprotein. Its main roles are in the regulation of the complement cascade, and in transporting opsonised immune complexes for removal from the circulatory system. It has been extensively studied due to its known genetic polymorphisms, different protein allotypes, and its significant number of disease associations, largely with autoimmune, infectious and inflammatory conditions. Variations within the CR1 protein also form the basis of the Knops blood group system. When *CR1* was first implicated in AD risk by the 2009 GWAS published by Lambert et al., attention turned to its possible role in neurodegeneration (Lambert et al. 2009). Neuroinflammation has long been implicated in AD, often regarded as a harmless bystander. However, the identification of multiple genetic risk factors for AD that are related to immunity and inflammation may suggest inflammation plays a more sinister role in the neurodegenerative process, which genetic variation in *CR1* could perhaps contribute to. The protein's role in Aβ clearance has also emerged as a potential explanation for why genetic variation in *CR1* affects AD risk.

CR1 – genetics and regulation

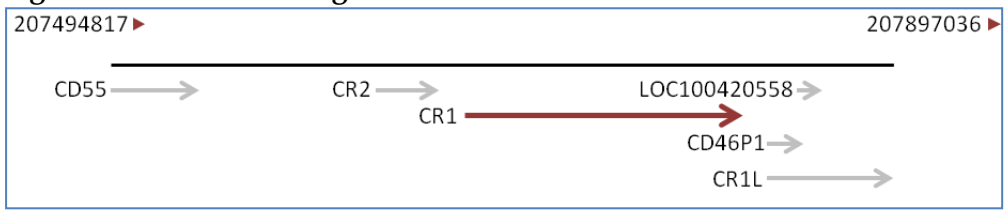
The gene encoding *CR1* (also known as *CD35*) is located on chromosome 1q32 (see Figure 1.10), amidst a cluster of complement related genes, often termed the regulators of complement activation (RCA) gene cluster, whose protein products belong to the RCA family. The genes surrounding *CR1* are shown in Figure 1.11. The pattern of conservation at the region of the *CR1* gene across various mammalian species is displayed in Figure 1.12.

Figure 1.10 – Genetic location of *CR1*



Location of the *CR1* gene on chromosome 1q32 (above) and transcripts (blue tracks) of the gene according to RefSeq (below) – F allele is upper transcript, S allele below. Image taken from the UCSC Genome Browser (Kent et al. 2002) (http://genome.ucsc.edu/).

Figure 1.11 – Genetic neighbours of *CR1*



Locations of *CR1*'s nearest genetic neighbours on chromosome 1, all members of the RCA gene cluster, taken from the NCBI website (http://www.ncbi.nlm.nih.gov/gene/1191). Chromosomal co-ordinates listed at the top of the figure, with genes and orientations displayed below.

Figure 1.12 – Conservation in the *CR1* gene region

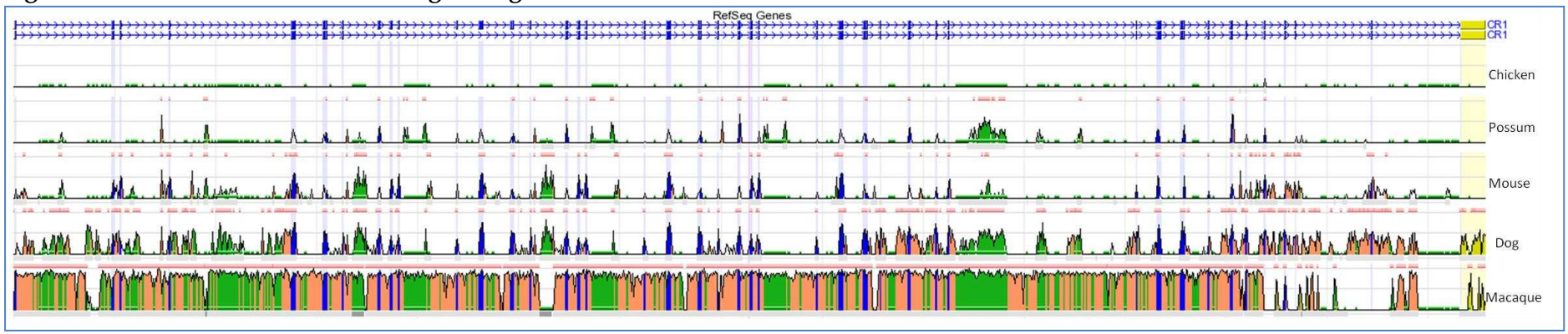


Image to show genetic conservation across selected vertebrate species in *CR1*, taken from the ECR browser (Ovcharenko et al. 2004) (http://ecrbrowser.dcode.org/). Blue – exons, pink – introns, yellow – untranslated regions, red – intergenic regions, green – repetitive DNA elements. Height of graph proportional to level of conservation between human and that particular species (shown on right).

CR1 was first identified as a membrane bound protein on the surface of erythrocytes, and is widely expressed on a number of blood cells, including neutrophils, eosinophils, monocytes, macrophages, B-lymphocytes and a sub-population of CD4-positive T cells. Aside from these peripheral blood cells, CR1 is also expressed on lymph node follicular dendritic cells, Langerhan cells in the skin and glomerula podocytes (Yoon and Fearon 1985; Liu and Niu 2009; Crehan et al. 2012). Expression of CR1 has also been reported on human astrocytes (Gasque et al. 1996) and neurons (Zanjani et al. 2005; Hollingworth et al. 2010), although elsewhere CR1 was not detected on these cells, and it was stated that CR1 expression in the brain was likely to be low, and potentially restricted to the phagocytic Kolmer cells of the choroid plexus (Singhrao et al. 1999). In addition to the membrane confined versions of the protein, a soluble form of CR1 (sCR1) exists at low levels (~30ng/ml) in the blood (Yoon and Fearon 1985), as well as a form of the protein found in urine, thought to be derived from vesicles from glomerula podocytes (Pascual et al. 1994).

The level of expression of CR1 varies between different cell types, and indeed shows vast variation in the figures reported in the literature, depending on the method of detection used (Moulds 2010). Expression levels of CR1 on erythrocytes show huge variation between healthy individuals (up to ten fold in Caucasians), an affect which is largely due to different expression level alleles associated with a *Hind III* restriction fragment length polymorphism (RFLP) site (Wilson et al. 1986). Although erythrocytes generally have lower expression levels than other cell types (e.g. leukocytes, with between 10,000 and 30,000 molecules per cell (Moulds 2010), B cells and monocytes with around 20,000-40,000 molecules per cell (Krych-Goldberg and Atkinson 2001), and resting neutrophils, with around 5,000 molecules per cell, which can increase up to ten fold when stimulated (Fearon and Collins 1983)), because of their relative abundance in the circulation, the majority (>85%) of CR1 in the circulatory system is erythrocyte bound CR1 (E-CR1) (Moulds 2010).

Due to the different expression levels of CR1, Kim et al. (Kim et al. 1999) sought to identify the regulatory elements which may control this expression within the promoter of the *CR1* gene. They studied a region of ~2kb 5′ of the gene, and found no evidence of a typical TATA type promoter sequence, but did find a CAAT-box type sequence (TCAAAA, which had previously been shown to be capable of acting as a CAAT-box (Kunz et al. 1989)), which was observed around position -54 to -49. The 5′ flanking region was also found to contain a GC-rich region, particularly high in CpG dinucleotides (Kim et al. 1999).

As well as variation in expression in CR1 across different cell types, there is variation in the glycosylation levels of CR1, such that the molecular weight of CR1 can differ by around 6kDa between erythrocytes and neutrophils or T cells (Wong 1990; Crehan et al. 2012). Between different cell types on which

CR1 is expressed, its function varies, perhaps partially dependent on these different glycosylation patterns, and not all of the roles of CR1 have been fully elucidated yet.

CR1 – protein structure and function

CR1 is a single chain type I transmembrane glycoprotein. Four protein allotypes of CR1 exist, with varying molecular weights, showing codominant inheritance. The four alleles, termed CR1-A (sometimes referred to as the F allele), CR1-B (sometimes referred to as the S allele), CR1-C and CR1-D encode proteins of 190kDa, 220kDa, 160kDa and 250kDa respectively. The F/S allele naming system of the two most common isoforms, CR1-A and CR1-B, is a reflection of their motility in gel electrophoresis (Fast and Slow moving). In all populations, CR1-C and CR1-D are rare, perhaps indicative of a selective advantage of the two intermediately sized isoforms. CR1-A and CR1-B have frequencies of approximately 0.87 and 0.11 in Caucasian individuals (Crehan et al. 2012), frequencies which are relatively consistent across populations studied (see table 1.5 for population frequencies in different ethnic groups, and for a summary of the characteristics of the four different protein allotypes).

Table 1.5 – CR1 isoform properties

Allele	Protein Size (non-reducing) (kDa)	SCR Number	LHR Number	Frequency (Caucasian)	Frequency (African American)
CR1-C	160	23	3	Rare	Rare
CR1-A (F)	190	30	4	0.87	0.82
CR1-B (S)	220	37	5	0.11	0.11
CR1-D	250	44	6	Rare	Rare

Characteristics of CR1's four protein allotypes, adapted from Crehan et al. (Crehan et al. 2012) and Krych-Goldberg and Atkinson (Krych-Goldberg and Atkinson 2001). Numbers of short consensus repeats (SCRs) and long homologous repeats (LHRs) are given for each allele.

The CR1 protein has four main structural domains: a 41aa signal peptide; the extracellular domain; a 25aa transmembrane domain; and a 43aa cytoplasmic domain (Klickstein et al. 1987). The differences between the four CR1 allotypes lie within the extracellular domain, which is comprised of multiple short consensus repeats (SCRs), also known as complement control protein repeats (CCPs) or sushi domains. This type of motif is common to the extracellular regions of the RCA protein family, with varying numbers of SCRs in different proteins, ranging from just four in CR1's genetic neighbour CD55, to 44 in the longest isoform of CR1. The 59-72aa SCRs have four common conserved cysteine residues, responsible for the formation of two disulphide bridges, and one conserved tryptophan, with looser conservation in the rest of the repeat, although a core of hydrophobic residues is also common to all SCRs (Klickstein et al. 1987; Crehan et al. 2012). The disulphide bridges flank an elongated region, featuring β-pleated sheets, and connecting loops (Liu and Niu 2009). In CR1, unlike in the other RCA family members, the SCRs are

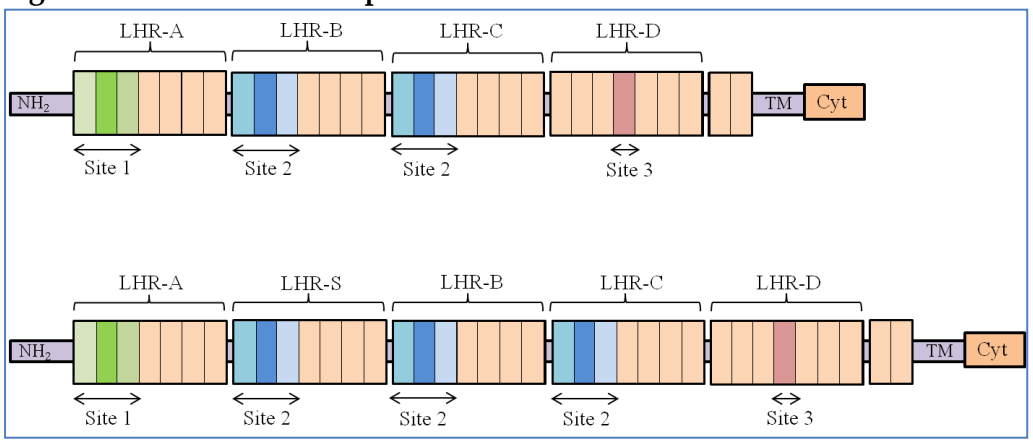
grouped in to long homologous repeats (LHRs). Each LHR is comprised of seven SCRs, with every eighth SCR being highly homologous (such that SCR 1, 8, 15; 2, 9, 16 etc. are 65 - 100% identical) (Klickstein et al. 1987). The two SCRs proximal to the protein's transmembrane domain are not included in the LHR structures. SCRs are also found in other, non-complement related proteins, indicating that although they can play a major role in the formation and function of complement related proteins, they are not restricted to this role (Klickstein et al. 1988).

It is thought that the differences in the alleles arose from unequal crossover events during replication, that lead to deletions or duplications of the highly repetitive section of DNA encoding the LHR, such that the size difference between the alleles is equivalent to one LHR; around 18kb at the genetic level, and 1.4kb at the transcript level (Holers et al. 1987; Wong et al. 1989; Bettens et al. 2012). It is speculated that the crossover events resulting in the creation/deletion of the highly homologous LHRs occurred relatively recently in our evolutionary history, while duplication of the SCRs, which show looser conservation and are found throughout the RCA family, as well as other, non-complement related proteins, arose through much older genetic events (Holers et al. 1987).

The series of duplications and subsequent divergence has brought about the structure of the human CR1 protein – a large, multifunctional molecule. The LHR regions which define the different isoforms of CR1 encode a binding site for complement component C3b/C4b, sometimes termed site 2, so the larger isoforms have more copies of this binding site (one in CR1-C, two in CR1-A, three in CR1-B and four in CR1-D), the specificity of which is conferred by the NH_2 terminal SCRs of LHR-B and C in the F allele (Klickstein et al. 1988). Each isoform also includes a C4b binding site in LHR-A, sometimes termed site 1 (again, with specificity conferred by the NH_2 terminal SCRs) and an additional active site in the centre of LHR-D (site 3), which is responsible for the protein's ability to bind to mannan-binding lectin and complement protein C1q (Tas et al. 1999; Ghiran et al. 2000).

A diagram of the structures of the two most common isoforms (CR1-A and CR1-B) is provided in Figure 1.13, demonstrating the locations of CR1's binding sites.

Figure 1.13 – CR1 common protein isoforms



Structure of the two most common CR1 isoforms, CR1-A (the F allele) above, and CR1-B (the S-allele) below. The active sites highlighted show the SCRs involved in the interaction between CR1 and its various binding partners. Adapted from figures in Crehan et al. (Crehan et al. 2012) and Liu and Niu (Liu and Niu 2009).

According to information assimilated in Krych-Goldberg and Atkinson's 2001 paper (Krych-Goldberg and Atkinson 2001), site 1, in LHR-A, binds to C4b, weakly to C3b, has low cofactor activity for factor 1 mediated cleavage of C3b and C4b and has high decay accelerating activity for C3 convertases (Krych et al. 1994; Krych-Goldberg et al. 1999). Conversely, site 2, in LHR-B and C (in the F allele, with an additional LHR giving an additional copy of site 2 in the S allele) binds relatively strongly to C3b, more weakly to C4b (although affinity is comparable to that of site 1), has high cofactor activity for factor 1 mediated cleavage of C3b and C4b and has low decay accelerating activity for C3 convertases (Krych et al. 1994; Krych et al. 1998; Krych-Goldberg et al. 1999).

CR1 is the main receptor for complement components C3b (an inflammatory protein activated in AD (Bertram and Tanzi 2010)) and C4b, and is an important regulator of the classical and alternate complement cascades. C3b and C4b are thought to have arisen through a gene duplication event, and share around 29% homology (Krych-Goldberg and Atkinson 2001). It is thought the interaction between the two complement proteins and CR1 occurs through a highly conserved region of 27 amino acids proximal to the amino terminal end of the α-chain (Taniguchi-Sidle and Isenman 1994).

On erythrocytes, CR1's main function is in the transportation of opsonised immune complexes. Immune complexes present in the circulatory system which have activated complement are bound by C3b and C4b, for which CR1 has multiple binding sites. C3b/C3b and C3b/C4b complexes are produced when the alternative and classical pathways (respectively) are activated. Although individually, each CR1 binding site has a low affinity for its target molecule (Arnaout et al. 1983), collectively, the presence of multiple CR1 molecules, each with multiple ligand binding sites, allows interaction with complexes containing multiple C3b and C4b molecules to occur relatively strongly as the sites can act synergistically (Krych-Goldberg and Atkinson

2001). This synergistic binding facilitates the transportation of immune complexes from the circulation to the liver and spleen, where they can be removed and degraded by fixed macrophages (Krych-Goldberg and Atkinson 2001), eliminating the factor that triggered the complement response initially.

Since different isoforms of the CR1 protein have different numbers of C3b binding sites, it has been postulated that the different isoforms may show differences in their capacity to clear immune complexes efficiently, with individuals in possession of smaller CR1 isoforms (fewer C3b binding sites) postulated to be worse at this clearance. Wong et al. demonstrated different allotypes of CR1, with different numbers of active sites varied up to 100 fold in their capacity to bind dimeric C3b (Wong and Farrell 1991).

CR1 can act as a versatile inhibitor of the complement cascade, dampening immune responses. It is able to impair the function of C3 and C5 convertases, which feature in both the classical and alternative complement pathways, via its decay accelerating activity (Krych-Goldberg and Atkinson 2001). It can also act as a cofactor for Factor 1, facilitating the irreversible cleavage and inactivation of C3b and C4b (Krych-Goldberg and Atkinson 2001).

As mentioned before, the function of CR1 differs between the different cell types on which it is expressed. E-CR1 is by far the most extensively studied, but other roles on other cell types have also been identified.

On B-cells, CR1 is involved in proliferation and differentiation (Fingeroth et al. 1989). When B-cell surface CR1 is bound by ligands, it appears to prevent B-cell proliferation (Jozsi et al. 2002). This is suggested as a mechanism by which CR1 is involved in autoimmune disorders (Khera and Das 2009).

On neutrophils and monocytes, particularly when these cells are activated (e.g. by cytokines), CR1 mediates phagocytosis (Wright and Silverstein 1982), and can stimulate the release of interleukins, indicating another mechanism by which CR1 may help mediate the immune response (Bacle et al. 1990). The role of CR1 expressed on T cells remains unclear (Khera and Das 2009).

sCR1

It was first discovered in 1985 by Yoon et al. that there existed a soluble form of CR1, termed sCR1, free in circulation, as well as the membrane confined forms of the protein (Yoon and Fearon 1985). sCR1 is present in serum at low concentrations, and plasma and serum levels of sCR1 are identical, indicating the protein is not lost during the clotting process (Pascual et al. 1993). sCR1 has been shown to be derived from the proteolytic cleavage of leukocyte membrane CR1, either during its transition through the golgi apparatus or at the cell membrane itself, giving a form of the protein which is around 5kDa smaller and lacks the intracellular domain of the complete protein (Danielsson et al. 1994; Hamer et al. 1998).

sCR1 is a potent local inhibitor of the classical, lectin and alternative complement pathways (Ramaglia et al. 2008). Its mechanism of action appears to be two-fold: firstly, it aids in the dissociation of C3 convertases, and secondly, it targets C3b and C4b for degradation, preventing excessive activation of the complement cascade. Ramaglia et al. looked at the effect of sCR1 treatment on rats with mechanical peripheral nerve crush injuries (Ramaglia et al. 2008). The group found that complement activation in the damaged nerve was almost completely inhibited by sCR1 treatment; deposition of the membrane attack complex was inhibited, as were deposition of C4c (an activation product of the classical complement pathway) and C3c (an activation product common to all complement pathways). The affected nerves were protected from axonal loss and myelin breakdown in the early stages following the trauma, demonstrating the protective capacity of the molecule, however, the effects were relatively short lived, with nerve damage becoming apparent around 7 days after the initial assault (Ramaglia et al. 2008). Whether this phenomenon is relevant to AD pathology remains to be established. Indeed, at present it is unclear whether sCR1 is even present in the brain, and what effects it may have if it is.

CR1 polymorphisms

There are three types of well documented variation associated with *CR1*. The first of these is the structural polymorphism generating the different protein isoforms, as discussed above. Secondly, there are polymorphisms that alter the expression level of E-CR1. As mentioned previously, the number of CR1 molecules per erythrocyte can vary 10-fold among healthy individuals, and one reason for this is the high (H) and low (L) expression alleles, which are associated with a *Hind III* RFLP site within intron 27 of the gene (reportedly due to a SNP, T520C (Liu and Niu 2009)), but the causative genetic basis of the differing expression remains unknown (Weis et al. 1987; Cockburn and Rowe 2006). Different levels of expression of different allotypes of the protein have been observed on the erythrocytes of heterozygote donors (Dykman et al. 1983; Wong et al. 1983), indicating that the variance stems from some genetic factor within those alleles, and not from some trans-acting genetic or global regulatory mechanism. It has also been demonstrated that the variance does not stem from polymorphisms within the 3' untranslated region, or promoter of the gene (Cockburn and Rowe 2006).

The RFLP site generates two fragments of different lengths – a 6.9kb fragment linked to the low expression (or L) allele, and a 7.4kb fragment linked to the high expression (or H) allele (Liu and Niu 2009). According to Krych-Goldberg and Atkinson, LL homozygotes typically display ~100 CR1 molecules per erythrocyte; for HH homozygotes, this figure is ~1000; while heterozygotes show an intermediate number (Krych-Goldberg and Atkinson 2001). Liu and Niu, however, reported these figures to be <200 per erythrocyte for LL homoxygotes, with HH genotype individuals possessing erythrocytes with 3-4 times more than this, and again, heterozygotes being in between the two (Liu and Niu 2009). It is likely that the discrepancies between the figures

reported are largely due to methodological differences. The frequencies of the alleles are reported to be 0.73 for the H allele and 0.27 for the L allele (Krych-Goldberg and Atkinson 2001) in Caucasians, and 0.51 and 0.49 in Indian subjects (Katyal et al. 2003): both populations in which there is an association between the RFLP site and expression (Xiang et al. 1999; Katyal et al. 2003). The RFLP site is not associated with expression levels of the protein in African populations (Xiang et al. 1999; Rowe et al. 2002).

It is proposed that polymorphisms linked to the RFLP site may affect the stability of CR1 (Liu and Niu 2009), with L allele polymorphisms producing a protein more prone to degradation, and therefore resulting in a reduced quantity of the protein reaching the cell membrane of erythrocytes (Crehan et al. 2012).

The biological consequence of the lower levels of expression is that there are fewer CR1 molecules available to fulfil E-CR1's normal physiological function. Individuals with low expression levels are poorer at removing complement opsonised immune complexes from the circulatory system than are high expressing individuals (Gibson and Waxman 1994; Crehan et al. 2012). Complement activation in such individuals is likely to be consistently higher than for high expressing individuals, since immune complexes, bound by C3b and C4b will persist for longer in the circulation.

Expression levels below ~100/cell, are termed the Helgeson phenotype (Moulds et al. 1992) (also referred to in the literature as Hegelson (Krych-Goldberg and Atkinson 2001)), and are not associated with any overt disease phenotype (Krych-Goldberg and Atkinson 2001), in fact, it has been implicated in protection from severe malaria (Cockburn et al. 2004).

The final group of polymorphisms associated with the *CR1* locus are those that comprise the Knops blood group system. The antibodies which are formed against antigens in this system were previously classed as HTLA (high-titre, low avidity) antibodies, but as research progressed it became apparent that a group of these HTLA antibodies, sharing similar specificities and molecular origins actually belong to a discrete group – termed the Knops blood group system, the 22nd blood group system to be recognised by the ISBT (International Society of Blood Transfusion) Committee on Terminology for Red Cell Surface Antigens (Daniels et al. 1995). The antibodies themselves, raised against the Knops antigens are not regarded as clinically significant, as they do not cause adverse reactions following blood transfusions, or create haemolytic disease in babies (Moulds 2010).

The Knops blood group antigens are named Kn^{a} and Kn^{b}; McC^{a} and McC^{b}; S11 and S12 (also known as Sl^{a} and Vil); and S13, Yk^{a} and KCAM (previously known as KAM). Back in 1991, two groups (Moulds et al. 1991; Rao et al. 1991) identified the CR1 protein as the origin of the Knops blood group antigens, Kn^{a}, McC^{a} and S11/Sl^{a}, and subsequently the other antigens of the system have

been attributed to the same protein. The molecular basis giving rise to the antigens at the genetic level have been identified, with 8 of the 9 antigens arising as a result of mutations in exon 29 (SCR 25), and Yk^a attributed to a SNP within exon 26 (SCR 22), both in LHR-D of CR1 (Moulds et al. 2001; Veldhuisen et al. 2011). Most of the antigens are generated by a single polymorphism, but S13 appears to be a conformational epitope which is formed by the combination of changes at amino acid positions 1601 and 1610 (Moulds et al. 2002). There are also a postulated S14 and S15, which are not yet officially recognised (Moulds et al. 2002; Covas et al. 2007). All the antigens occur in exposed parts of the CR1 protein, where they are accessible by antibodies. Table 1.6 shows the properties of the different Knops blood group antigens, including ISBT number, molecular basis and frequency in Caucasian populations.

Table 1.6 – Properties of the Knops Blood Group System

Antigen	ISBT Number	Nucleotide*	Amino Acid	SNP	Frequency (Caucasian) %
Kn^a	KN1	4681G	1561V	rs41274768	98
Kn^b	KN2	4681A	1561M		4.5
McC^a	KN3	4768A	1590K	rs17047660	100
McC^b	KN6	4768G	1590E		0
S11/Sla	KN4	4801A	1601R	rs17047661	100
S12/Vil	KN7	4801G	1601G		0
S13 +	KN8+	4801A, 4828T	1601R, 1610S	rs17047661, rs4844609	100
-	KN8-	4801A, 4828A	1601R, 1610T		0
Yk^a +	KN5+	4223C	1408T	rs6691117	92
-	KN5-	4223T	1408M		8
KCAM +	KN9+	4843A	1615I	rs3737002	95
-	KN9-	4843G	1615V		5

Information about the Knops blood group Antigens, adapted from Moulds (Moulds 2010) and Veldhuisen et al. (Veldhuisen et al. 2011).
*Numbered from translation start site

Ethnic differences in the frequencies of Knops antigens have long been recognised, with some showing vastly different frequencies between Caucasian and African populations (e.g. McCb is virtually absent in Caucasians but is found in around 50% of West Africans) (Moulds 2010). It is widely thought that this variance is due to differing selective pressures in different geographical regions. CR1 has numerous disease associations (discussed in further detail below), including certain polymorphisms within the gene conferring protection from conditions such as malaria and *M.tuberculosis* infection. Knops antigens associated with such a selective advantage are likely to have become significantly more frequent in African

populations, where such diseases are a serious threat to health, rather than in regions where these are not endemic.

While the molecular basis of all the known Knops antigens has now been established, there are multiple other known polymorphisms within the *CR1* gene, suggesting the possibility that more, as yet undiscovered Knops antigens exist (Moulds 2010).

CR1 – Other Disease Associations

CR1 is largely involved in autoimmune and inflammatory disorders, as well as infectious diseases such as malaria and HIV. sCR1 has been linked to renal and hepatic failure, multiple cancers of the blood (Pascual et al. 1993) as well as SLE (Khera and Das 2009).

It has been postulated that CR1 may play a role in preventing the inappropriate recognition of "self" antigens as foreign by B-cells, a mechanism which if disturbed, could lead to the development of autoimmune disorders (Khera and Das 2009).

Acquired reduction in E-CR1 levels is observed in patients with systemic lupus erythematosis (SLE) (Walport et al. 1987; Kumar et al. 1995), rheumatoid arthritis (Kumar et al. 1994) and insulin dependent diabetes mellitus (Ruuska et al. 1992), and decreased CR1 expression correlates with disease severity in HIV patients (Jouvin et al. 1987).

CR1 is known to be associated with rosetting behaviour in *P.falciparum* malaria, facilitating the invasion of erythrocytes, and thus spreading the infection within an individual (Rowe et al. 1997). Low E-CR1 expression levels appear to be protective against severe malaria (Cockburn et al. 2004). Infection of monocytes and macrophages by *M. tuberculosis* and a variety of *Leishmania* species is also thought to be linked to or mediated by CR1 on the surface of these cells (Moulds 2010).

More recently, the SNPs rs6656401 and rs3818361, implicated in AD, have also been linked to susceptibility to depression (Hamilton et al. 2012).

CR1 and AD

Since *CR1* was first implicated in AD risk, numerous studies have been conducted seeking links between SNPs within the gene and various aspects of AD. Two independent studies seeking potential links between *CR1* SNPs and levels of CSF biomarkers (Aβ and tau) failed to detect any significant associations (Kauwe et al. 2011; Schjeide et al. 2011). Brouwers et al. however, did find some evidence for an association between *CR1* SNPs and levels of $A\beta_{1\text{-}42}$ (Brouwers et al. 2012).

When looking for a relationship between *CR1* SNP rs6656401 and cognitive function in extremely old individuals (92-93 at age of intake), Mengel-From

found no significant association (Mengel-From et al. 2010). They did however find a suggestive association between the GWAS risk allele and poorer cognitive performance, in male subjects only (Mengel-From et al. 2010).

Kok et al. found some evidence that *CR1* rs1408077 may be linked to plaque load in the brain, since the CC genotype was found to be more likely than the AA genotype to have sparse senile plaques, rather than no senile plaques (OR 2.1 (95% CI 1.01-4.43) p=0.048) (Kok et al. 2011). This is strengthened by reports from Chibnik et al., who found that the risk allele of SNP rs6656401 was associated with AD pathological traits (mainly neuritic amyloid plaque load, as well as diffuse plaque load, although not with neurofibrillary tangles) (Chibnik et al. 2011). Additionally, each risk allele at the SNP was linked to increased cognitive decline, both generally, and specifically with episodic and semantic memory, perceptual speed and visuospatial ability (Chibnik et al. 2011). This is in agreement with the tentative link between memory and *CR1* risk SNPs observed by Mengel-From et al. (Mengel-From et al. 2010).

Furthermore, Keenan et al. identified a coding SNP, rs4844609, within LHR-D of *CR1* (in strong LD with rs6656401 (D' 1)) which was associated with decline in episodic memory, accompanied by increased AD neuropathological features, and that the effect showed an interaction with *APOE* genotype (Keenan et al. 2012).

Biffi et al. reported an association between *CR1* SNP rs1408077 and entorhinal cortex thickness (p=0.03) (Biffi et al. 2010). However, when Furney et al. conducted a similar study investigating the effect of AD risk genes on various neuroimaging measures, no associations between *CR1* SNPs and any of the parameters they considered were detected (including entorhinal cortex thickness) (Furney et al. 2010). This inconsistency highlights the requirement for further studies of sufficient power to detect what aspects of AD are affected by *CR1* genotype.

CR1 and Aβ

Given the evidence for links between *CR1* genotype and CSF Aβ levels (Brouwers et al. 2012), as well as brain plaque load (Chibnik et al. 2011; Kok et al. 2011; Keenan et al. 2012), it may be that the relationship between *CR1* and AD stems from altered Aβ metabolism or clearance.

Complement can be activated by Aβ, particularly oligomeric forms, in AD affected brains (Rogers et al. 1992), and complement opsonins (such as C3b) become bound to Aβ in an antibody independent fashion (Bradt et al. 1998; Rogers et al. 2006).

CR1 may affect Aβ clearance in the brain, either directly, promoting removal of Aβ within the brain (e.g. by mediating phagocytosis (Brouwers et al. 2012)), or indirectly (clearing Aβ from the periphery (Rogers et al. 2006)).

It has already been discussed that a major function of E-CR1 is in the removal of opsonised immune complexes from the circulatory system, and Aβ, bound by complement component C3b, constitutes one such complex that can be removed in this way (Rogers et al. 2006). Individual differences in the CR1 protein, either in its structure or expression levels, could therefore render individuals more or less capable of conducting such clearance (Rogers et al. 2006).

It has been suggested that the different alleles of *CR1* may affect this process, with smaller isoforms in possession of fewer C3b binding sites perhaps being less efficient at this clearance. However, this is inconsistent with Brouwers et al.'s finding that the longer, S allele of *CR1* is linked to an increased risk of AD (Brouwers et al. 2012). According to the above hypothesis, this longer form should be more capable of clearing circulating opsonised Aβ from the peripheral blood, and would be expected, therefore, to be linked to a reduced risk of AD (Brouwers et al. 2012). This casts some doubt on the "peripheral sink" theory, and perhaps suggests that *CR1*'s relationship with AD is not directly derived from its relationship with Aβ clearance.

CR1 and neuroinflammation

CR1 is involved in the regulation of the complement cascade on many levels, mainly acting to reduce activation of complement by a variety of mechanisms, so could prevent damage due to inflammation in the brain occurring. Larger forms of CR1, with increased numbers of active sites, would be expected to be more efficient at reducing complement activation, and would therefore be protective. However, this is contradicted by the findings of Brouwers et al. that the larger S allele of CR1 was associated with increased AD risk (Brouwers et al. 2012).

There is evidence from mouse models, however, that inhibiting complement activation actually increases plaque deposition and neurodegeneration, while increasing complement C3 reduces plaque load, suggesting complement activation might actually be protective (Wyss-Coray et al. 2002), which is consistent with Brouwers et al.'s findings. Similarly, if clearance of Aβ from the brain relies on its opsonisation with C3b, and longer CR1 isoforms limit the availability of active C3b, the brain's ability to clear Aβ could be compromised, exposing it to greater damage than do the shorter isoforms (Brouwers et al. 2012).

Keenan et al.'s findings that a SNP within LHR-D of *CR1* is linked to increased AD pathology and cognitive decline perhaps suggests an alternative mechanism for the relationship between *CR1* and AD (Keenan et al. 2012). LHR-D is the domain responsible for CR1's interactions with C1q and mannan-binding lectin, perhaps indicating that it is CR1's interaction with one of these that mediates its relationship with AD risk. Although little is known about mannan-binding lectin in relation to AD, C1q has been shown to be present at high levels in AD brains relative to controls, particularly in areas

with a high predominance of pathological hallmarks (Yasojima et al. 1999). C1q is has been shown to associate with fibrillar Aβ and amyloid plaques (Fonseca et al. 2004), and when knocked out in transgenic animals, is shown to reduce deposition of Aβ, limiting plaque formation (Fonseca et al. 2011).

Because inflammation in the brain is so closely linked to Aβ, disentangling the exact relationship between *CR1* and AD – whether effects are exerted directly on Aβ, with knock-on effects on inflammation, or whether *CR1*'s relationship is with complement, independent of Aβ – will take extensive further research.

CR1 and Brain Structure

A final intriguing possibility is raised by a study in which Bralten et al. reported an association between *CR1* genotype and entorhinal cortex volume in young, healthy adults (Bralten et al. 2011), using high resolution MRI technology to assess brain structures. In the discovery cohort, there was an association between *CR1* genotype at SNP rs6656401 and gray matter volume in both the hippocampus and entorhinal cortex, with carriers of the risky A allele displaying lower gray matter volume in these regions. Although the hippocampal finding was not replicated in the second cohort, the entorhinal cortex volume link was replicated, with evidence that the effect may be dose dependent. Exploratory analysis of other brain regions found evidence the link between *CR1* genotype and brain structure may extend into areas such as the amygdala, anterior medial temporal lobe and collateral sulcus, although further study will be necessary to confirm these suggestive findings (Bralten et al. 2011).

The research suggests that variance in *CR1* at the genetic level affects the structure of the brain, even in young healthy adults, perhaps making carriers of the risk allele more susceptible to AD in later life. As mentioned above, Biffi et al. (Biffi et al. 2010) reported an association between *CR1* genotype and entorhinal cortex thickness in AD and MCI patients, in agreement with the findings of this study. In light of Bralten et al.'s findings, however, perhaps the effect seen by Biffi et al. was an almost "end stage" snapshot of a phenomenon which arose in these patients much earlier in their lives.

APOE genotype has also been shown to have an effect on brain structure in young, healthy individuals, with ε4 allele carriers showing reduced entorhinal cortex thickness (Shaw et al. 2007). It was speculated that a smaller entorhinal cortex volume could leave individuals more prone to displaying symptoms of cognitive decline, while those with larger brain volumes may have an inherent resistance to neurodegenerative processes. It is a fascinating concept that perhaps rather than being mechanistically involved in AD per se, *CR1* risk alleles may contribute to the formation of a neural environment more susceptible to the changes which bring about AD, while the protective alleles may contribute to the formation of brain structures more able to withstand the assaults of the disease.

1.12. Finding Causal Variants

It is generally accepted that the SNPs identified as GWAS "hits" are not themselves the causative variants, but rather are tagging a variant or variants that are causal. All of the studies that have been conducted investigating the GWAS hits and various aspects of AD pathology are forced to use surrogate SNPs, since the true causative functional variants underlying the GWAS signals are not yet known, and this could undermine attempts disentangle AD aetiology. Knowing the true functional variant or variants underlying the association would not only give increased power to these types of study, but it could also give valuable insight into how this involvement in AD risk might come about. For example, if the causative variants lie in regulatory regions, the alteration in AD risk may stem from under or over expression of the gene, while non-synonymous coding changes could alter the biochemical properties of the protein molecule.

Finding the causal variant(s), however, may not prove to be simple. There is currently debate over two competing hypotheses – the common disease, common variant hypothesis (CD/CV), and a hypothesis centred on rare SNPs being the actual causal variants underlying GWAS "hits". It is common variants (>5% minor allele frequency (MAF)), with a small effect on disease risk (ORs ~1.2) that GWAS were designed to detect. It would be expected that if the functional variants behind GWAS associations were common, they should be relatively easy to identify, since their high MAF would allow them to be seen by sequencing a small number of cases and controls. Yet, despite hundreds of disease associated loci being found by GWAS for various complex disorders, almost no causative variants responsible for these association signals have been found. This lends support to the hypothesis that rare variants (MAF <5%) with greater odds ratios may be the true variants underlying GWAS hits, an effect which has been termed "synthetic association" (Goldstein 2009), and occurs when, by chance, more rare causative variants are associated with one allele of the common tag SNP than the other. If this is found to be correct in AD genetics, it could go some way to explaining the "missing heritability"(Manolio et al. 2009) of the condition, since the ORs of the rare variants could be significantly higher than those of the common SNPs with which they are linked (Wang et al. 2010).

Several groups have attempted to pinpoint the true causative variants underlying the observed GWAS signals within *CLU*, *PICALM* and *CR1*, and the findings of some of these studies are reported below.

CLU

All the common variants at the *CLU* locus which have been found to be associated with AD fall within the same ~13.4kb LD block, which is entirely contained within the boundaries of *CLU*. Figure 1.14 shows the pattern of LD in the *CLU* area, and indicates the location of the original GWAS SNPs. This

provides compelling evidence that *CLU* itself is the source of the association with AD, particularly given its strong biological candidacy.

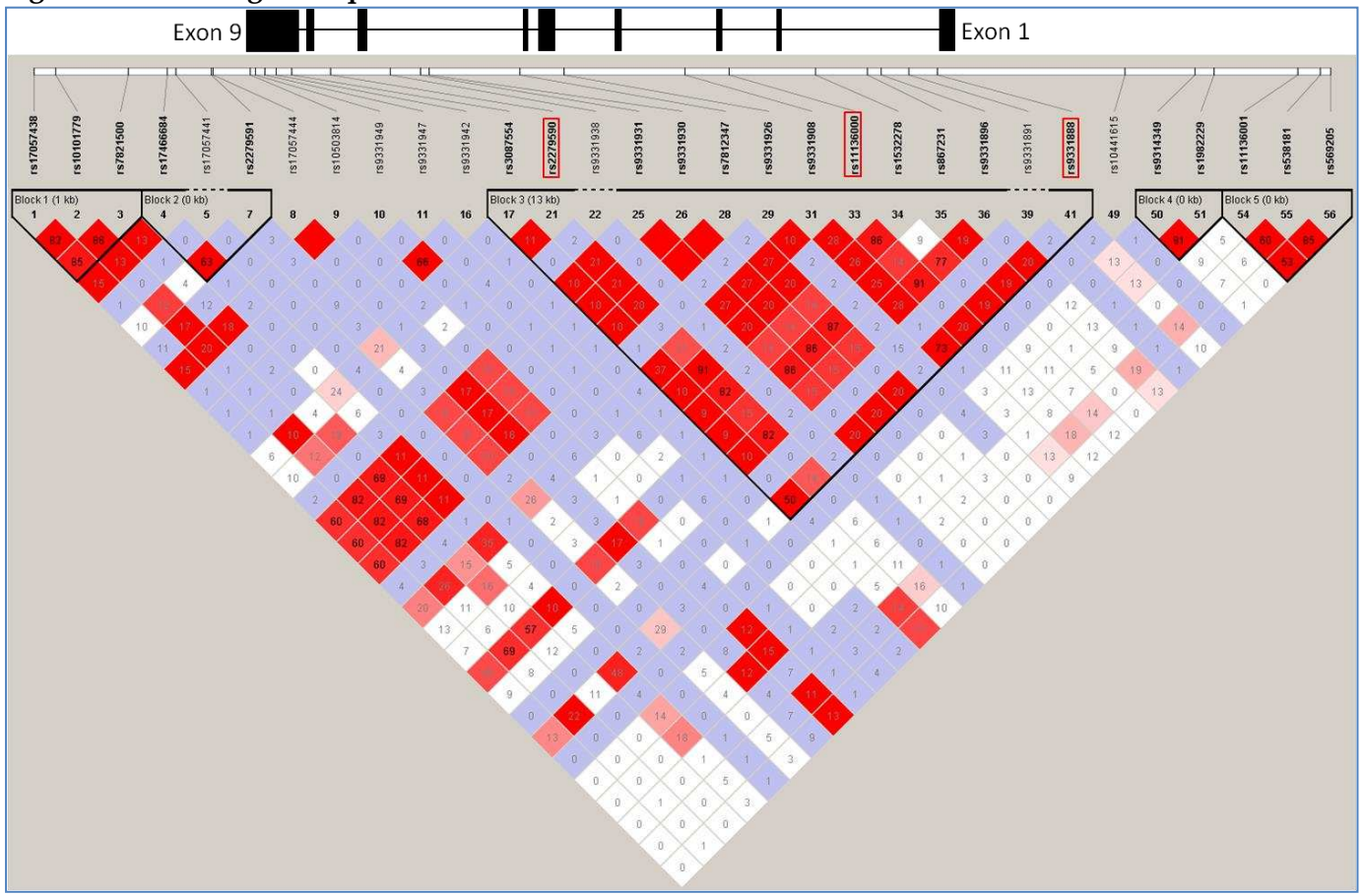
The *CLU* SNP rs11136000 which has been reported so extensively to be associated with AD risk is intronic, and is not thought to have an effect on the function of the gene.

In 1996, far before the association of the *CLU* gene and AD had been highlighted by GWAS, Tycko et al. conducted a study to discover polymorphisms within *CLU* and test these for association with AD, based on functional evidence implicating *CLU* in AD risk. Seven variants were discovered within the gene, including two non-synonymous changes, but none showed association with the disease in their sample set (Tycko et al. 1996).

Harold et al. made initial attempts to discover the underlying causal variants by looking for SNPs showing strong linkage disequilibrium (LD) with the GWAS SNP, and potentially functional variants within the gene from publically available data. The synonymous SNP rs7982 in exon 5 was found to be in strong LD with rs11136000 and showed a similarly significant association with AD (Harold et al. 2009). As no amino acid change is evoked by the polymorphism, it was speculated that the activity of a splicing enhancer signal could be affected by the variant, although no bioinformatics or experimental evidence for this was provided (Harold et al. 2009).

To search for common coding variants that might explain the association signal with rs11136000, Guerreiro et al. sequenced the entire coding region of *CLU* in 495 cases and 330 controls, and exon 5, or exons 5 and 6, were sequenced in additional samples. The group found 24 variants in total, and the 14 of those which occurred in more than one individual were tested for association with AD, but no significant associations were detected, although there was suggestive significance (uncorrected p=0.04) for rs3216167, a SNP which had previously been reported to be associated with cholesterol levels in serum (Miwa et al. 2005; Guerreiro et al. 2010). The non-synonymous variants found were also assessed in terms of likely functionality using bioinformatics tools, and several were deemed likely to be deleterious. However, such analyses should be interpreted with an element of caution since predicted effects are not always reliable, and even seemingly severe mutations can have little phenotypic effect. The study identified a nonsense mutation, which would be expected to obliterate the expression of CLU from that allele, yet the subject in which it was discovered was a 69 year old healthy control (Guerreiro et al. 2010).

Figure 1.14 - Linkage disequilibrium around GWAS SNPs in *CLU*



Pattern of linkage disequilibrium (LD) at the *CLU* locus. Data from HapMap (HapMap 2003) release #28, image created using Haploview (Barrett et al. 2005). LD values are shown as r^2. SNPs which were found to be significantly associated with AD in the Harold et al. and Lambert et al. GWAS are highlighted in red (rs11136000 - Harold et al. OR=0.86, p=8.5x10-10; Lambert et al. OR=0.86, p=7.5x10-9. rs9331888 - Lambert et al. OR=1.16, p=4x10-8. rs2279590 - Lambert et al. OR=0.86, p=8.9x10-9) (Harold et al. 2009; Lambert et al. 2009).

18 variants were detected in *CLU* by Ferrari et al. who sequenced the coding region of the gene in 342 AD patients and 277 controls (Ferrari et al. 2012). The 18 variants included 10 missense mutations, 6 synonymous changes, a nonsense mutation and an intronic SNP. When analysed *in silico*, several of the changes were predicted to be damaging to the structure of the protein (Q15R, S16R, R234H, P286S, M302V, R338Q, N369H and T428M). Three variants (the nonsense mutation, E14X; Q15R and P265S) were found only in cases and not controls in this study (Ferrari et al. 2012), although the nonsense mutation had previously been reported by Guerreiro et al., as discussed above (Guerreiro et al. 2010).

Bettens et al. conducted a comprehensive screen for rare variants in *CLU*, sequencing all coding regions in 1930 individuals (cases and controls). Exons 5-8, which encode CLU's β chain, were found to harbour a significant excess of rare variants in AD patients compared to controls, including a number of predicted deleterious changes, and so were sequenced in up to 2755 further samples (Bettens et al. 2012). Association was seen between rs11136000 and AD, which persisted even when these rare coding variants were excluded from analysis. This indicates that the association of the common and rare SNPs with AD are independent of each other, and the GWAS signal cannot be explained by these rare variants (Bettens et al. 2012). It remains, therefore, to be established what is, or are, the underlying variants generating this association.

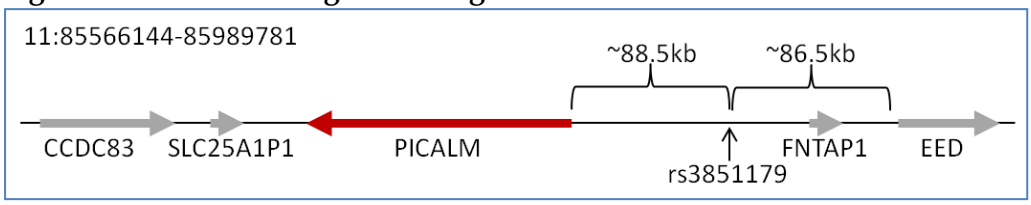
It is noteworthy that these studies have all focussed on coding variants; sequencing exonic regions alone. Variants outside of these areas may affect regulation or expression of the gene, but would have been missed by the studies published to date, and could play a major part in the involvement of CLU in AD risk, although their functional consequences would be even more difficult to establish.

PICALM

Although the gene *PICALM* is often said to be associated with AD, the SNP which first implicated the gene in AD risk (rs3851179) actually resides 88.5kb 5'of the gene. Figure 1.15 shows the chromosomal location of *PICALM*, the SNP rs3851179, and the other genes which are in the vicinity. The nearest genetic feature to rs3851179 is the pseudogene, farnesyltransferase, CAAX box, alpha pseudogene 1 (*FNTAP1*). The active form of this gene, on chromosome 8, encodes the α-subunit of CAAX geranylgeranyltransferase and CAAX farnesyltransferase. Another related pseudogene resides on chromosome 13. rs3851179 actually falls approximately equidistant between *PICALM* and *EED* (embryonic ectoderm development), a gene which encodes a member of the Polycomb-group family, involved in maintaining transcriptional repression of genes across generations. Despite the two genes being almost the same distance from the original SNP found to be associated

with AD, *PICALM* presents a stronger biological candidate for involvement in AD pathogenesis than *EED*, and indeed, other SNPs close to and within *PICALM* have subsequently been shown to associate with AD risk, so it is unlikely *EED* could be the true source of the association despite its equivalent proximity.

Figure 1.15 – *PICALM*'s genetic neighbours

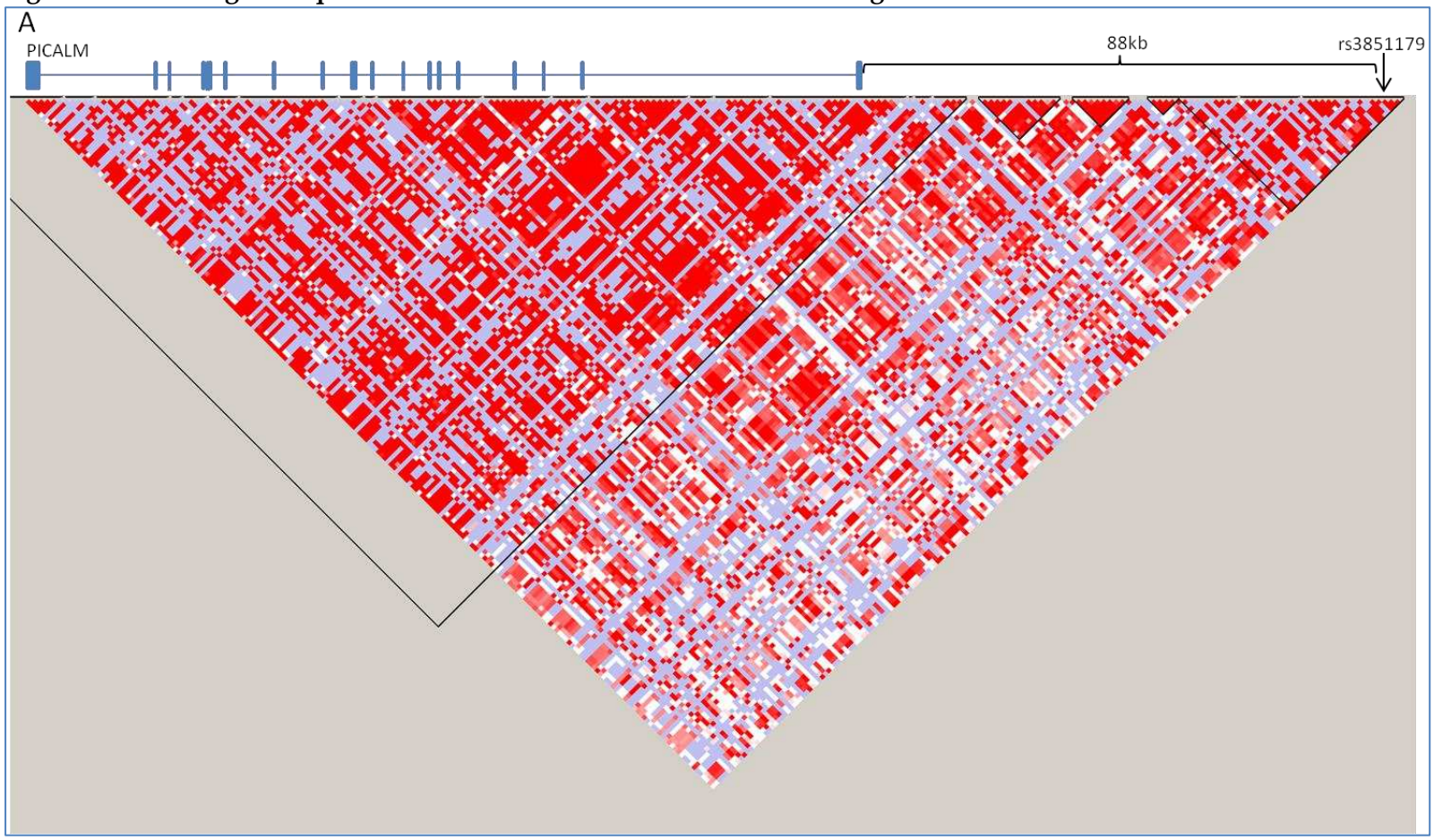


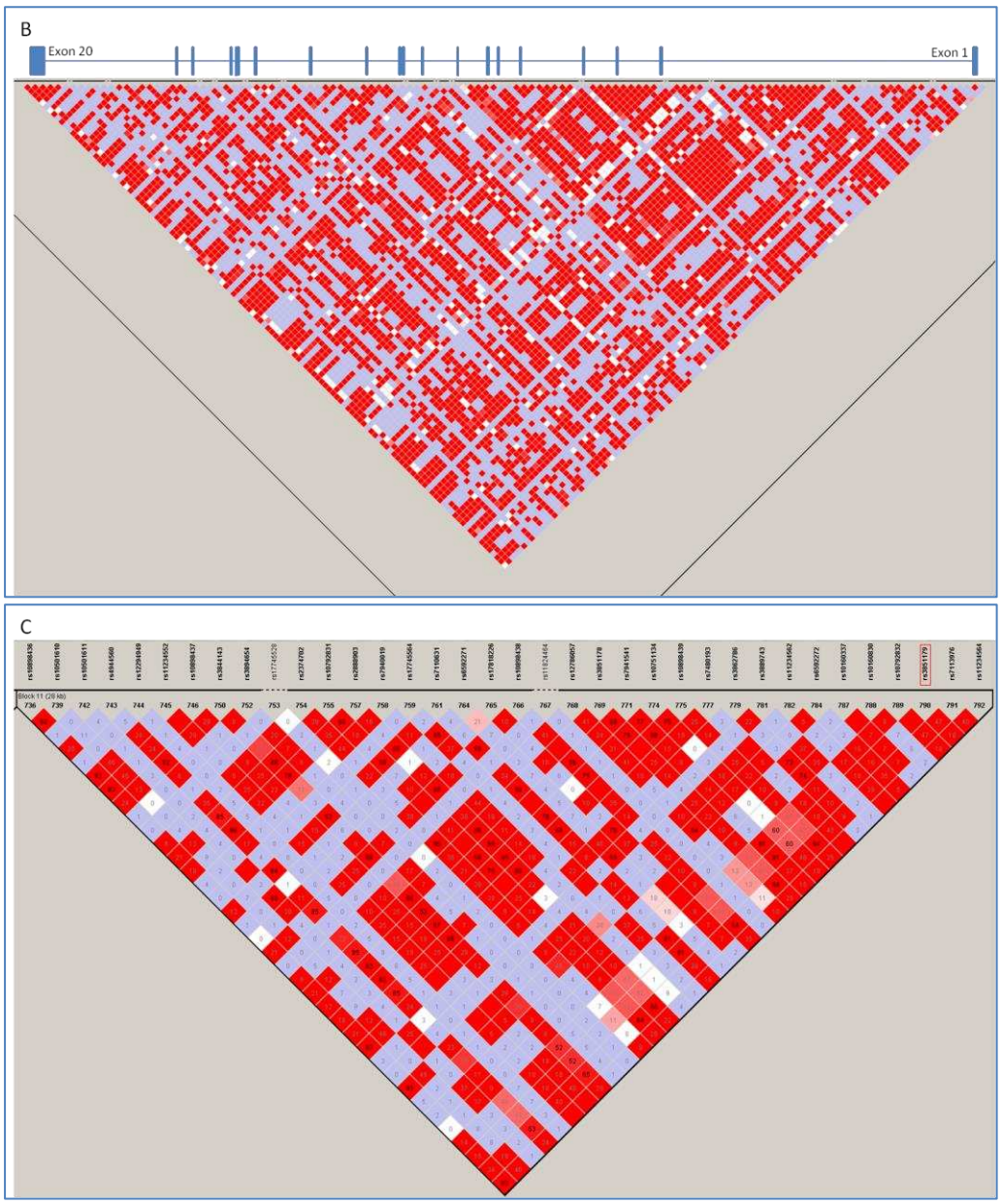
Locations of genes near to rs3851179 on chromosome 11 and relative distances to these, adapted from the NCBI website (http://www.ncbi.nlm.nih.gov/gene/8301). Chromosomal coordinates given above, with genes and orientations demonstrated with arrows.

Figure 1.16 shows the LD block surrounding rs3851179, as well as the pattern of LD within the *PICALM* gene itself. The rs3851179 SNP is contained within a tight LD block of around 28kb, and is therefore not in strong LD with SNPs within *PICALM*. Harold et al. (Harold et al. 2009) however, did state that the SNP rs541458 which resides just 8kb from *PICALM* is in LD with rs3851179, and showed that this SNP was indeed strongly associated with AD with the genotyping of additional samples (as discussed above).

Harold et al. made some preliminary attempts to establish the actual causative variants, looking at potentially functional SNPs at the *PICALM* locus (in putative transcription factor binding sites and synonymous exonic SNPs which could affect splicing regulation), but found these to be more weakly associated with AD, and so unlikely to be altering AD risk themselves (Harold et al. 2009).

Figure 1.16 - Linkage disequilibrium around rs3851179 and the *PICALM* gene





Patterns of linkage disequilibrium (LD) surrounding rs3851179 and the *PICALM* gene. Data from HapMap (HapMap 2003) release #28, images created using Haploview (Barrett et al. 2005). LD values are shown as r^2. A. LD in the full region, including *PICALM*, the ~28kb LD block within which the GWAS SNP rs3851179 falls, and the intervening region. B. LD shown within *PICALM* alone. C. Pattern of LD immediately surrounding the GWAS SNP rs3851179, showing the boundaries of the ~28kb LD block, with the location of rs3851179 highlighted in red (OR=0.86, p=1.3 x 10^{-9} (Harold et al. 2009)).

Uncovering the actual causative variants underlying altered AD risk could give vital clues as to how *PICALM* is aetiologically involved in AD risk. Schnetz-Boutaud et al. sequenced the coding region of *PICALM* in 48 cases and 48 controls in the quest for causal variation, but failed to discover any novel variants (Schnetz-Boutaud et al. 2012). They did, however, comment that synonymous SNP rs592297, in LD with GWAS SNP rs3851179, falls within a potential exonic splicing enhancer site within exon 5 of the gene, which could affect the splicing, and thus expression and function of *PICALM*. Whether this does indeed affect the splicing of the gene, and how this relates to AD remains to be determined. Ferrari et al. also sequenced the coding region of *PICALM*, this time in 342 LOAD and 277 control subjects (Ferrari et al. 2012). 16 variants (3 synonymous (Q174Q, T586T and A590A); 2 missense (A411P, H465R); and 11 non-coding) were detected within *PICALM*, however, all were found in both cases and controls, and none of the variants were likely to be damaging when assessed with *in silico* prediction programs, so it is unlikely these are relevant to AD pathology.

Since the strongest signals of association are at the 5' end of the gene, it has been speculated that the association with AD could be due to variants affecting regulation of gene expression (Sleegers et al. 2010), rather than being coding changes altering protein structure or function.

CR1

Both of the SNPs identified as significant GWAS hits by Lambert et al. (Lambert et al. 2009) fall within a ~127kb block of LD which is entirely encompassed by the *CR1* gene, as shown in figure 1.17, strongly implying that the true causative variants underlying the signal reside within this large gene itself.

Figure 1.17 – Linkage disequilibrium around *CR1* GWAS SNPs

Pattern of linkage disequilibrium (LD) at the *CR1* locus. Data from HapMap (HapMap 2003) release #28, image created using Haploview (Barrett et al. 2005). LD values are shown as r^2. Locations of the SNPs which were found to be significantly associated with AD in the Lambert et al. (Lambert et al. 2009) GWAS combined data set are highlighted by red arrows (rs6656401 OR=1.21 (95% CI 1.14-1.29), *p*-value of 3.5×10^{-9}; rs3818361 OR=1.19 (95% CI 1.11-1.26), *p*-value of 8.9×10^{-8}).

Brouwers et al. sought to fine map the association observed between AD and the *CR1* SNPs from the Lambert et al. GWAS, looking at 26 SNPs spanning the *CR1* locus, all falling outside of the gene's repetitive regions, which were able to capture an estimated 87% of the total genetic variability at the locus (Brouwers et al. 2012). The SNPs associated with AD in the GWAS (rs6656401 and rs9818361) were not genotyped as part of this study as they fall within *CR1*'s repetitive regions/LCRs. Two other SNPs (rs4844610 and rs1408077) however, in strong LD with each other, and the SNPs from the GWAS, showed significant allelic association with AD risk in the Flanders-Belgian cohort of 1883 individuals. The strongest association was seen in *APOE* ε4 allele carriers. The group, as well as analysing the variants singly, combined their analyses to consider haplotypes, grouping the SNPs in to five LD blocks. Only the fourth LD block (spanning a region of around 130kb, including almost the entire gene, excepting the first and last exons) showed association with AD risk, again, strongest in the *APOE* ε4 carriers. Given the potential links between CR1 and Aβ metabolism/clearance, the group also considered the effect of the SNPs analysed with the levels of biomarkers ($A\beta_{1\text{-}42}$, total tau and $ptau_{181}$) in CSF. The SNPs associated with AD (rs4844610 and rs1408077) did not show any association with any of the tested biomarkers, however there was evidence that the minor alleles of four other SNPs (rs646817, rs1746659, rs11803956 and rs12034383), all within the same LD block, were associated with increased CSF levels of $A\beta_{1\text{-}42}$ (Brouwers et al. 2012).

As discussed previously, the group also used a multiplex amplicon quantification (MAQ) technique to allow them to distinguish the different alleles encoding the CR1 protein isoforms. Quantification of LCR1 copy number allowed the inference of the F- and S- allele genotypes (with one and two copies of this LCR respectively), with the caveat that S allele homozygotes would appear with the same LCR1 copy number as CR1-D/F allele heterozygotes. It was found that those with three copies of LCR1 (i.e. F- and S-allele heterozygotes) had around a 30% increase in AD risk than those with only two copies of LCR1 (i.e. F- allele homozygotes). This CNV was found to be in LD with the two SNPs (rs4844610 and rs1408077) which were also found to be associated with AD risk, suggesting the two actually represent a single common signal of association. The LCR1 CNV association with AD, but not the two SNP associations, were replicated in an independent French cohort (n=2003), with a meta-analysis of both data sets strengthening the evidence that the CNV is a genuine AD risk factor. The inconsistency of the SNP associations in contrast to the strength and replicability of the CNV association may suggest that the CNV itself is the underlying source of the association signal (Brouwers et al. 2012). It may be that the various SNPs at the *CR1* locus which have shown association with AD have actually been tagging this CNV with which they are in LD.

Ferrari et al. sequenced CR1's coding regions in 342 AD patients and 277 controls, and identified a total of 65 variants (39 missense, 15 synonymous variants, 9 intronic and 2 in the 3' UTR) (Ferrari et al. 2012). Six of these

variants were both found only in cases, not controls, and were predicted to be probably or possibly damaging to the protein's structure using *in silico* prediction programs (P110T, I127T, K113E, T1349I, L172M and G2109S). Further work is needed to establish whether these variants do in fact affect protein function, and how this in turn impacts on susceptibility to AD.

Next generation sequencing

Next generation sequencing (NGS) technologies allow an unprecedented way to characterise and catalogue all genetic variation within a given locus. A number of different NGS methods are currently available (Metzker 2010), and more are in the pipeline, leading to fierce competition in terms of cost, throughput and quality of data, driving these methods to become increasingly affordable and efficient. The majority of investigative studies looking for causative variants underling GWAS signals so far have relied on costly and time consuming Sanger sequencing. This has meant that only small regions, usually the protein coding exons of the genes have been included by most studies. The capacity of NGS technologies to produce millions or billions of short sequencing reads in a single run means that these constraints no longer apply. While we are still some way from having whole genome sequencing of individuals affordable in a practical way to researchers, methods of target enrichment allow already identified genetic regions to be deep resequenced, and all genetic variation at that locus within a sample to be detected.

Target enrichment

Target enrichment methods enable the filtering out of regions of interest from genomic DNA as a whole. A number of different techniques can be used to achieve this, from traditional and long range PCR to a number of commercially available strategies, each with its own relative strengths and weaknesses. There are various important parameters to consider when assessing the performance of target enrichment methods. These include how much of the region of interest is able to be targeted by the enrichment strategy, and the proportion of sequencing reads which can be mapped back to the region of interest (specificity), which is important since off target reads reduce the capacity for the production of usable data. The depth of coverage across the region of interest, and the uniformity of this coverage are important when it comes to calling variants. Around 10-20x coverage is generally seen as necessary for confident SNP calls; less than this, and variants may not pass QC filters; significantly more than this can be a waste of resources and sequencing capacity. Sensitivity in this context is the proportion of the region of interest for which sequence data is obtained. Also of importance when comparing different methods are the cost of reagents and any specialist equipment required, throughput, ease of use, and timescale of processing.

At the time of the design of this experiment, two of the major commercially available target enrichment strategies were Agilent's SureSelect (SS) solution based hybridisation, and Nimbegen's (NG) array based hybridisation method. The NG method works by the synthesis of oligonucleotides complimentary to

the genomic region of interest directly on to an array, to which the prepared genomic library can then be hybridised, while the SS method uses biotinylated RNA baits, again, complimentary to the genomic region of interest, which hybridise with target sequences in solution, and can then be pulled out using streptavidin coated beads. With each method captured DNA undergoes an amplification step prior to sequencing.

A number of papers were available which sought to compare the two methods on a number of different parameters. Practically, SS is faster, requires less in the way of specialist equipment and has lower DNA requirements than its array based counterpart (Mamanova et al. 2010; Teer et al. 2010). In terms of performance, SS's recurrent downfall is that, due to the stringent repeat masker (discussed in greater detail in section 2.4) used in the generation of target regions for bait design, often less of the region of interest is targeted by this method when compared to NG, which uses its own, seemingly less conservative, repeat masking software. As a result of this, often a lower percentage of the region of interest is able to be sequenced using SS (Teer et al. 2010; Hedges et al. 2011; Kiialainen et al. 2011). However, SS has been consistently found to have a higher proportion of reads mapping to the region of interest (Teer et al. 2010; Hedges et al. 2011; Kiialainen et al. 2011), indicating it is more specific than the NG method. In one study, SS was found to have an inferior read depth when compared to NG (Teer et al. 2010), but in two others (Hedges et al. 2011; Kiialainen et al. 2011), the depth of coverage was found to be greater when SS was used. SS was also found to give a greater level of consistency between samples than NG, demonstrating a greater reproducibility with Agilent's method (Hedges et al. 2011; Kiialainen et al. 2011). When sequenced regions are compared like for like, SS libraries have been shown to yield more SNP calls compared to NG, and these have been found generally to be more accurate (Kiialainen et al. 2011). For these reasons, SS was chosen for this project due to its cost efficiency and overall strengths in terms of specificity, coverage and reproducibility.

Simply finding variants within a known associated locus, however, is insufficient. Countless SNPs exist within any given individual, harmful, protective and benign, and the effects of these are not easy to deduce. There are a plethora of bioinformatics tools available to predict the functions of variants, but these are never flawless, so functional characterisation is a must when determining whether any discovered variant is causative. Furthermore, a cautious attitude must be adopted since even when the effect of the variant on the gene or protein is known, the effect *in vivo* may not be clear – as shown by the non-sense mutation detected by Guerreiro et al., which may have been assumed to be strongly linked to the condition, had that subject not have been a healthy control (Guerreiro et al. 2010).

The beauty of deep resequencing is that it offers an almost unparalleled opportunity to elucidate and begin to understand the true causative genetic basis of complex disorders, such as AD, at the very base level. This kind of

knowledge is likely, in time, to bring a better fundamental understanding of the aetiology of such conditions, which in turn could lead to huge progress in terms of diagnosis, treatment, and ultimately cures for some of these devastating conditions that are proving such a challenge to public health in modern society.

1.13. Project Statement

With the aim of detecting and cataloguing rare variation at the loci implicated in AD risk by the first two major AD GWAS (Harold et al. 2009; Lambert et al. 2009), an NGS project was undertaken, using Agilent's SS system to specifically target the *CLU*, *PICALM* and *CR1* loci. This enrichment was designed to capture the whole locus, rather than just the coding regions on which previous studies had been based, since exonic sequencing alone has so far not yielded many answers as to the causative variants underlying GWAS signals. A number of different types of analysis software for NGS data were utilised, enabling an assessment of the relative strengths and weaknesses of each, and the development of a definitive pipeline using the best tools tested (Chapter 3). Once variants within the region were detected, Sanger sequencing was used to validate the methodology and highlight some issues arising from indels and mononucleotide repeats when using NGS technologies (Chapter 4). The exonic variants detected in the three genes were prioritised for further analysis as there are more reliable tools for assessing coding variants functionally, and it provided a modest number to focus on, minimising the necessary correction for multiple tests needed when testing for association. Various bioinformatic resources were utilised to assess likely functionality of the exonic variants, which were also tested for association with AD in a large imputed dataset, giving two independent methods of assessing each variant's likely contribution to AD pathogenesis, and highlighting those warranting further, functional investigation (Chapter 5). As for the exonic variants, the non-coding variants found were also assessed using bioinformatic resources and association testing in the imputed data set where evidence from the tools used suggested functional consequences (Chapter 6).

2. Methods

Sequencing Project One

2.1. Patient Demographics and Sample Preparation

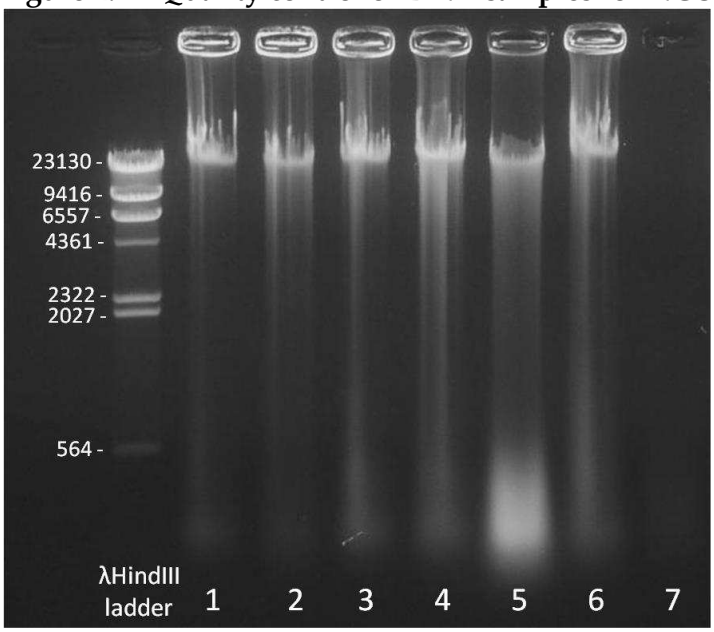
96 Alzheimer's disease (AD) samples were obtained from two UK centres – The University of Nottingham Brain Bank and Manchester Brain Bank (48.4% female, 51.6% male; mean age at onset 70.4 years; standard deviation 11.8; range 56-87. *APOE* alleles; ε2 - 5.7%; ε3 - 63.5%; ε4 - 30.8%). Both of these resources comprise part of the Alzheimer's Research UK (ARUK) brain bank, which has been used in such projects as the GWAS that first implicated *CLU* and *PICALM* in AD risk (Harold et al. 2009). Cases with age of clinical onset over 56 and a confirmed or probable AD diagnosis, based on NINCDS-ADRDA classification, were recruited. All subjects gave informed consent to be included in the study, which was granted approval by the local ethics committee.

High quality, undegraded genomic DNA was required for the next generation sequencing (NGS) project. DNA was prepared from brain tissue samples, using a phenol chloroform based extraction procedure. Approximately 0.5cm^3 of brain tissue was manually chopped on dry ice, to prevent the tissue from thawing. This was incubated overnight (~18 hours, shaking at 380rpm, 50^{0}C) with 500µl AL lysis buffer and 50µl proteinase K (both Qiagen). Next, 500µl of refrigerated phenol chloroform was added to the sample (Sigma), which was mixed by inverting before being subjected to centrifugation for 5 minutes at 13,000rpm. The top phase of the resultant sample was then removed to a clean eppendorf, and the addition of phenol chloroform, mixing and centrifugation was repeated. The top phase of this was removed, and had 3M sodium acetate (pH 5.2) added to it in a 1:9 ratio of sodium acetate to sample. Chilled 100% ethanol was then added, at an equal volume to the sample, to precipitate out the DNA. Following centrifugation at 13,000rpm for 15 minutes, a wash was performed using 200µl 70% ethanol, before another centrifugation step, this time for 10 minutes (13,000rpm). The remaining ethanol was then discarded and the pellet air dried, before resuspension in 100µl AE buffer (Qiagen), heating to 50^{0}C for one hour. A nanodrop spectrophotometer was used to assess the concentration and purity of the extracted DNA. The samples were assessed for degradation using gel electrophoresis (1% agarose gel). An example gel is shown in figure 2.1.

Quantification of the 96 DNA samples was conducted using the Quant-iT™ dsDNA Broad Range Assay Kit from Invitrogen, following standard manufacturer's method, with all samples run in triplicate. Pooling was conducted such that samples of similar concentration were grouped together (see Appendix section 2.1 for sample pools). For each pool of 12, 600ng of

DNA per sample were combined, giving an overall DNA quantity of 7.2µg per pool.

Figure 2.1 – Quality control of DNA samples for NGS



Representative gel from phenol chloroform DNA extractions, run against λHindIII ladder (sizes of marker in bp shown). Samples 1-4 and 6 show successful extractions with minimal degradation, exactly as needed for NGS experiments. Sample 5 shows a significant amount of degradation, and would not have been accepted in to the project. The extraction of sample 7 failed.

2.2. Power

The predominant aim of this study was to discover novel rare single nucleotide polymorphisms (SNPs), so it was important to ensure a sufficient sample size was used to give adequate power to do this. Typically, common SNPs are viewed as having a minor allele frequency (MAF) greater than 0.05, while rare SNPs have MAFs between 0.01 and 0.05. Anything below 0.01 is very rare. The following equation was used for power calculations, where n is the number of chromosomes:

$$n = [\log(1-power)] / [\log(1-MAF)]$$

Table 2.1 shows the power this study had to detect SNPs of various MAFs, given a sample size of 96. It was calculated that the study had 80% power to detect SNPs with a MAF down to ~0.85%, thus this study had sufficient power to fulfil its aim to discover rare novel SNPs.

Table 2.1 – Power Calculations

MAF	Power (%)
0.001	17.5
0.005	61.8
0.01	85.5
0.02	97.9
0.03	99.7
0.04	~100
0.05	~100
0.10	~100
0.20	~100
0.30	~100
0.40	~100
0.50	~100

Table to show the power this study has to detect SNPs of varying MAFs based on a sample size of 96 individuals, or 192 alleles.

2.3. Defining Regions to Sequence

The UCSC genome browser (Kent et al. 2002) was used to obtain the basic coordinates for the genes of interest (*CLU, PICALM* and *CR1*). These coordinates were then expanded to encompass any areas of notable conservation across vertebrate species, assessed by eye using ECR browser (Ovcharenko et al. 2004), since evolutionary constraint may suggest functional regions of DNA, such as gene regulatory elements.

In addition to the three genes, a fourth region was targeted in the study; the area in which SNP rs3851179 (the *PICALM* GWAS SNP) is located. To visualise the pattern of linkage disequilibrium (LD) in the region, Haploview (Barrett et al. 2005) was used, with SNP genotype data downloaded from HapMap (HapMap 2003) (on 05.11.2010). A region of 500kb surrounding the GWAS SNP was downloaded, to ensure no variants in strong LD with the GWAS SNP would be missed (since it is unlikely anything exceeding this distance would be in LD with the variant). When defining LD blocks, an r^2 of 0.8 was selected as the LD parameter (as opposed to D'). Only SNPs with a MAF lower than 0.01 were excluded (default 0.05), and all other Haploview settings were left as default. The entire LD block was targeted (~29kb).

An additional 150bp was added on either side of the final genomic coordinates to be sequenced, ensuring that the ends of the region of interest would be covered by the full 5x tiling used in the bait design (see below). The coordinates and sizes of the genes/regions targeted are given in table 2.2.

Table 2.2 – Target regions for NGS

Gene	Coordinates (hg19)	Size (kb)
CLU	8:27450849-27475277	24.43
CR1	1:207667495-207816719	149.22
PICALM	11:85665237-85783519	118.28
rs3851179 LD block	11:85840998-85870094	29.10
Total:		321.03

Coordinates and sizes of the regions targeted by SureSelect baits for the first sequencing project.

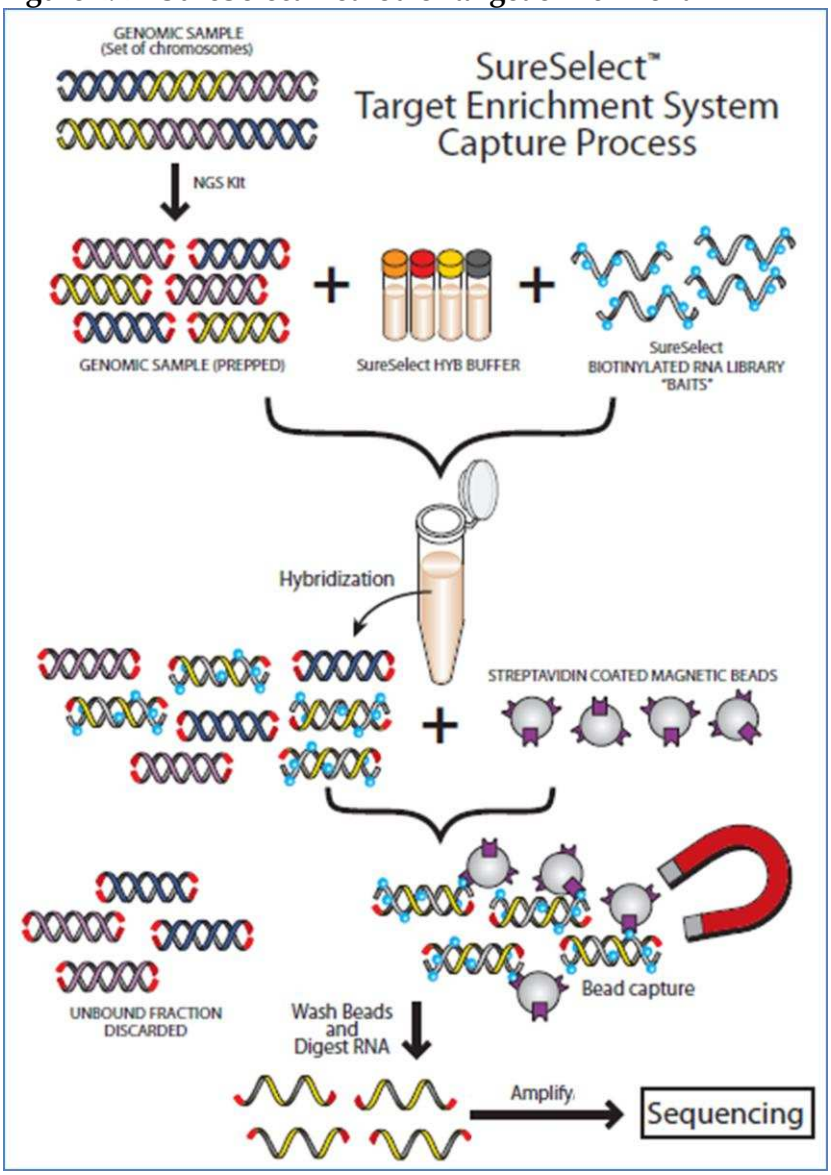
2.4. Enrichment and Sequencing

Agilent's SureSelect

Agilent's SureSelect (SS) system was the chosen method of target enrichment for the project, given its reported benefits over other commercially available target enrichment methods available at the time of study design (see Introduction section 1.12. Finding causal variants – Target enrichment) (Mamanova et al. 2010). SS is a method of target enrichment utilising a solution based hybridisation approach, whereby 120 base biotinylated RNA "baits", complementary to the genomic regions of interest are designed and hybridised with the desired targets in a library of fragmented whole genomic DNA. Using streptavidin coated beads, these baits and their bound complimentary DNA can be extracted from the noise of whole genomic DNA using a magnetised system. This process is summarised in Figure 2.2.

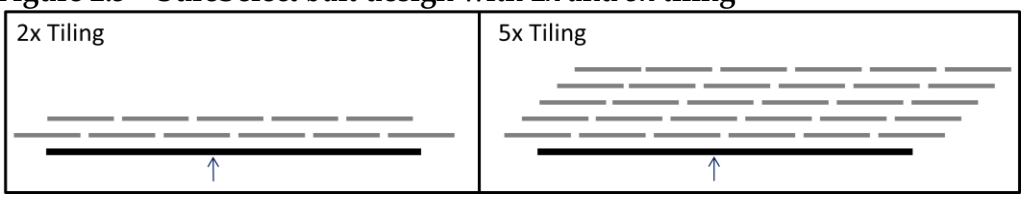
Once the genomic regions to be sequenced had been defined, the SS baits could be designed using Agilent's online eArray program (https://earray.chem.agilent.com/earray/). In this, a number of different parameters can be specified, such as the sequencing platform to be used (in this case, Illumina single end short read), and the genomic coordinates (in hg19) of the regions of interest specified, in order for baits to be designed against them. Agilent specify a set of optimised parameters, and for the most part our study design adhered to these, however, instead of the default 2x tiling, we opted for 5x tiling (see figure 2.3), as it was hoped this would give a better enrichment of the target region, and was recommended by an Agilent eArray specialist (personal correspondence).

Figure 2.2 – SureSelect method of target enrichment



Schematic diagram to show the processes involved in the SureSelect method of target enrichment, taken from Agilent's SureSelect protocol v1.2.

Figure 2.3 – SureSelect bait design with 2x and 5x tiling



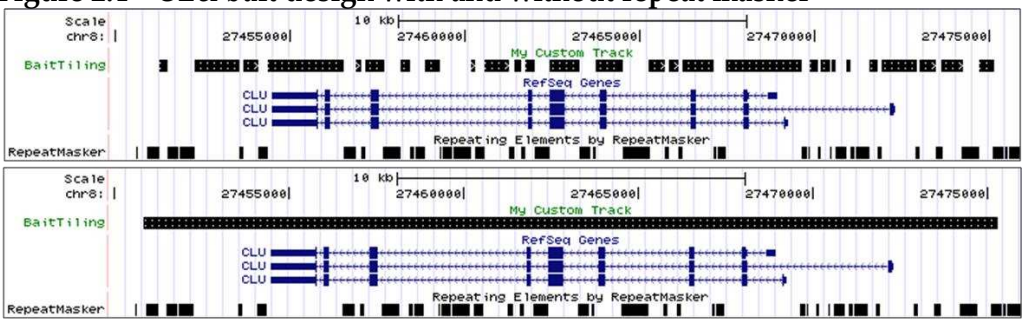
Genomic DNA represented by black bar, 120mer baits represented by gray bars. Arrow indicates a specific position in the region of interest. Under 2x tiling, each bait overlaps by 50% (60bp), meaning any given position in the target region is covered by two baits. For 5x tiling, the baits overlap by 24 bases, resulting in any given position in the DNA being represented by 5 baits.

As default in eArray, a repeat masker is used in the design of SS baits (Repeatmasker.org (Smit 1996-2010), which is based on the RepBase (Jurka et al. 2005) library of repeats). This is the same repeat masker upon which the UCSC Genome Browser's RepeatMasker track is based. In order to investigate the proportion of the region of interest that would be not be targeted by baits due to this repeat masking software, repeatmasker.org (Smit 1996-2010) was used. The gene sequences for the regions of interest were uploaded to this website, and submitted to the program for analysis, using its default settings (on 11.01.11).

When the SS baits were designed with the repeat masker on, the proportion of each of the genes that would not be targeted by baits seemed very high, leading to questions as to whether the repeat masker used by eArray was overly conservative. Using repeatmasker.org the percentage of each gene that would not be targeted by baits were the repeat masker used was quantified, and was 34% for *CLU*, 48% for *CR1*, 34% for *PICALM* and 42% for the rs3851179 LD block. This would have meant that between a third and half of each gene we were aiming to acquire complete sequence data on would not even have been targeted. Figure 2.4 shows the .bed file results of the *CLU* bait design process when uploaded to UCSC's custom tracks, both with and without the repeat masker enabled in the design process (the other genes, and the rs3851179 LD block showed similar results). It was therefore decided that the repeat masker would not be utilised in the design of the baits for this project, the hope being that only truly repetitive regions would fail to be sequenced, giving an obvious drop out in coverage at these regions, and this would be less in reality than with the repeat masker utilised.

The process of enrichment was outsourced to Source Bioscience (http://www.sourcebioscience.com/) for financial and practical regions, and was conducted by them, following Agilent's standard protocol.

Figure 2.4 – *CLU* bait design with and without repeat masker



The upper panel shows *CLU's* bait design conducted using the repeat masker, while the lower panel shows the targeted region when the repeat masker is not utilised. Genomic coordinates are shown at the top of each panel, with targeted region indicated with the custom "BaitTiling" track. RefSeq transcripts of the gene are shown in blue, while the lower region of each panel shows the locations of repeat regions according to RepeatMasker.org (Smit 1996-2010).

Illumina GaIIX Sequencing by Synthesis

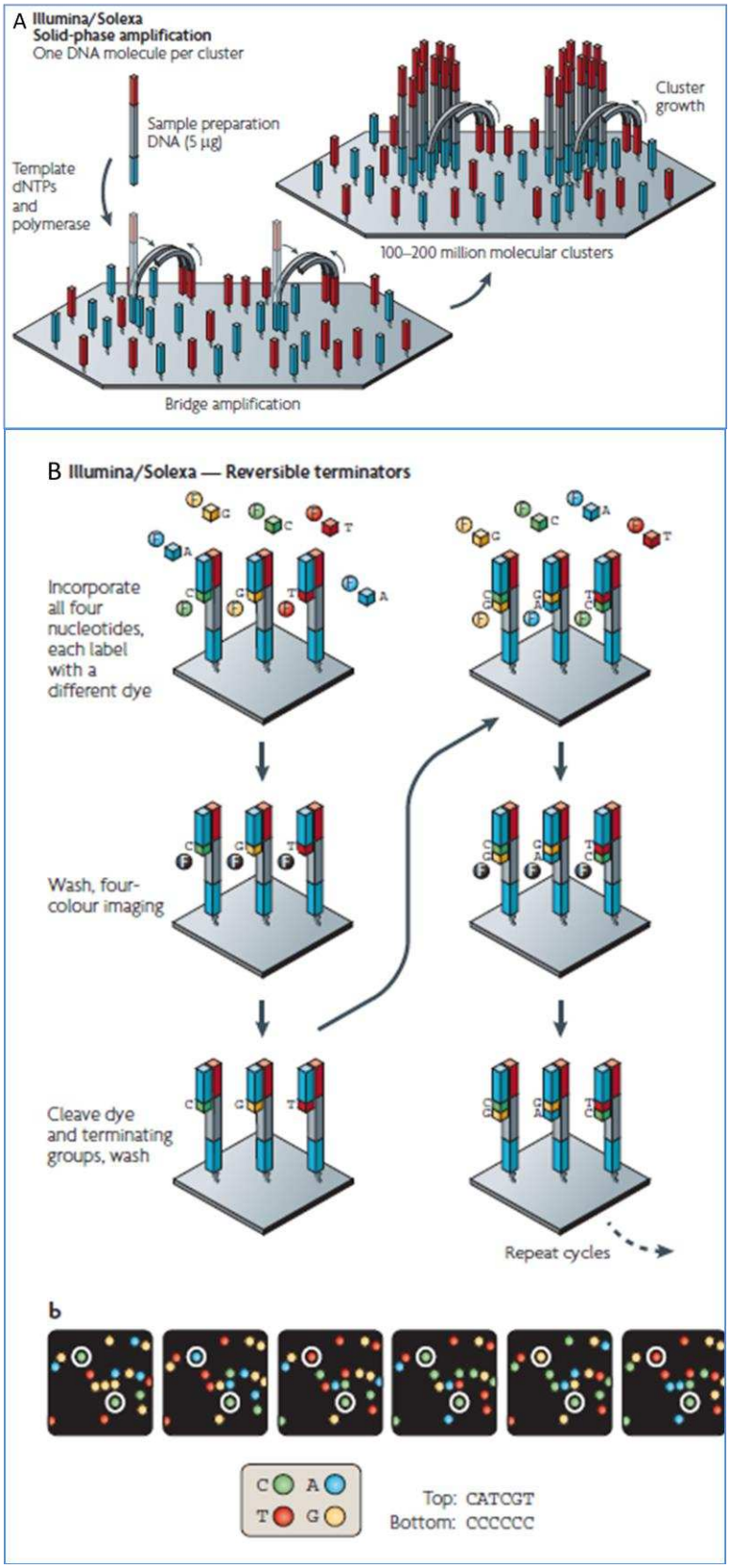
Illumina's GaIIX was selected as the method of NGS to be utilised by the project, based on practical and financial reasoning, in combination with Illumina being one of the major players in NGS technologies at the time, offering many widely used and well respected platforms (Metzker 2010).

Illumina's sequencing technologies are based on the sequencing by synthesis method. The initial stage involves solid-phase amplification of single DNA molecules (in our case, fragments of DNA from loci of interest) via bridge amplification, generating 100-200 million molecular clusters of DNA, each of a single template DNA fragment. The free ends generated can then be hybridised with universal sequencing primers to enable NGS to be performed. The actual sequencing is conducted using a four-colour cyclic reversible termination method, with total internal reflection fluorescence imaging facilitated by two lasers allowing the distinction of the different colours, each associated with a different base. The fluorescently labelled bases, with terminal 3' blocker, are added simultaneously, and DNA polymerase bound to the template adds a single specific modified base in each cycle. The base is then imaged, and unbound nucleotides are washed away, before the fluorescent dye and terminating group attached to the nucleotide are cleaved, leaving a base which can then be added to in subsequent cycles. These processes are depicted in figure 2.5.

Illumina's software, CASAVA (Consensus Assessment of Sequence and Variation) was used to convert the images read by the machine into intensity scores and base calls with quality scores.

The sequencing was conducted by Source Bioscience, following standard manufacturer's protocols.

Figure 2.5 – Illumina's sequencing by synthesis method



Illumina's sequencing by synthesis method. Panel A shows the bridge amplification step, used to create clusters of template molecules. Panel B shows the four-colour cyclic reversible termination method used to obtain sequence information.

2.5. Data Analysis

Data files for each pool were received from Source Biosciences in zipped .fastq format, which contains the nucleotide sequence and ASCII coded quality score for each of the sequencing reads produced. These files were processed along a pipeline comprised of various bioinformatic tools, designed to assess the quality of raw reads, align the data, assess the quality of the alignment and call variants. At the time, the pipeline was experimental, so several different software packages were utilised for many of the stages, but based on the data here presented, this has since been refined to a definitive pipeline (which is presented in the Section 3.4 - Defining the Pipeline).

An initial assessment of the quality of the data was performed using the program FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), which analyses data (in Sam, Bam or FastQ format) for a number of parameters (e.g. read quality, GC content and over represented sequences), which may be indicative of issues with the data as a whole, such as poor quality of reads or systematic biases.

Example command to run FastQC:

```
$ ./fastqc 1_NoIndex_L001_R1.fastq.gz
```

Where 1_NoIndex_L001_R1.fastq.gz is the zipped fastq file containing the data from the first lane of the flow cell used, i.e. pool 1 samples.

Alignment

Alignment is the process by which the millions (or billions) of short sequencing reads generated by NGS technologies are mapped to a reference sequence. It is typically the first stage in processing NGS data, on a pathway which will eventually lead to the identification of variant sites. There are a plethora of programs available to conduct this function, each with individual strengths and weaknesses in terms of speed, sensitivity and accuracy. For our project, two alignment programs were selected, MOSAIK (highly user friendly, and had been used by the 1000 genomes project) and BFAST (which was the best performing alignment algorithm in a comparison of programs from a previous NGS study conducted within our laboratory (data unpublished)).

MOSAIK

MOSAIK (http://code.google.com/p/mosaik-aligner/) produces gapped alignments based on the Smith-Waterman algorithm. The multistage program is intuitive to use, with a comprehensive user guide and sensibly chosen default values to minimise the necessary amount of user input. It is multithreaded, so can run simultaneously on multiple processors, speeding

up the process of alignment, and is suitable for use with a number of sequencing technologies.

The following steps are needed to generate alignments:

MosaikBuild: This module converts various types of input files to a binary format (.dat), allowing efficient data processing. Both the reference and the sequencing reads need to be processed in this way, with the sequence technology (-st) used to generate the data being specified for the reads.

MosaikJump: MOSAIK uses a hashing strategy, comparing reference and read hashes to compute alignments. Creating and utilising a hash map of a mammalian genome, such as the human one, would be prohibitively memory intensive; therefore MosaikJump is used to create a Jump Database (-j). This is comprised of "positions" and "keys" files, storing the same information as a conventional hash map, but in a much more efficient format, allowing for faster and less computationally intensive mapping. The selected hash size (-hs) used requires a trade off between the greater speed provided by larger hashes, and the greater sensitivity provided by smaller ones. MOSAIK's manual recommends a hash size of 15 as a good compromise between the two.

MosaikAligner: This component performs the pairwise alignment between the short sequencing reads (converted to a hashed format) and the reference file (stored in the jump database). "Hits" between the read and reference hashes are clustered, and any clusters above the user defined alignment candidate threshold (-act) are submitted for analysis by the Smith-Waterman algorithm and filtered to generate the final alignment file. Several parameters can be altered at this stage to tailor the output of the alignment software or the running speed. This includes, the maximum number of mismatches allowed (-mm) between the read and the reference, which is set as default to 4. Two parameters can be specified which reduce the time taken to perform the alignment. The maximum hash positions (-mhp) can be set (-mhp 100 recommended by manual), which means only 100 random potential hash matches are stored within the program, speeding up alignment with little cost to alignment accuracy. As mentioned above, the alignment candidate threshold (-act) can also be set, defining the minimum cluster size to be submitted to the Smith-Waterman algorithm, again speeding up alignment without compromising accuracy (manual suggests –act 20). The number of processors (-p) available to conduct the analysis can also be specified, allowing full utilisation of the available computational resources.

MosaikText: This utility allows the conversion of the alignment .dat file to the more universally accepted .bam format, allowing the output to be used with other data analysis and handling software.

Example commands to run Mosaik:

Convert reference file to .dat format:

```
$ MosaikBuild –fr
Homo_sapiens.GRCh37.59.dna.toplevel.chr_only.valid.fa –oa
HG19_ref.dat
```

Where -fr is the reference file to which alignment will be conducted (Homo_sapiens.etc.valid.fa), and -oa specifies the name of the output file.

Convert sequencing reads to .dat format (run for each sequence file):

```
$ MosaikBuild –q 1_NoIndex_L001_R1.fastq.gz –out
sample1.dat –st illumina
```

Where –q is the input file to be aligned and –st specifies the sequencing technology used.

Create Jump Database:

```
$ MosaikJump –ia HG19_ref.dat –out HG19_jump –hs 15
```

Align reads with MosaikAlign (run for each sequence file):

```
$ MosaikAligner –in sample1.dat –out sample1_aligned.dat
–ia HG19_ref.dat –hs 15 –mm 4 –mhp 100 –act 20 –j
HG19_jump –p8
```

Convert .dat to .bam:

```
$ MosaikText –in sample1_aligned.dat –bam
sample1_aligned.bam
```

BFAST

BFAST (BLAT-like Fast Accurate Search Tool) (Homer et al. 2009) aims to provide a fast alignment program, without compromising on accuracy. Again, the program has multiple stages, with capability for multithreaded processing, allowing the full utilisation of available computational resources, and the benefits of increased speed which comes with this. The program works by creating indexes of the desired reference sequence, allowing the rapid generation of candidate alignment locations (CALs), with gapped local alignment then occurring according to a gapped Smith-Waterman algorithm, facilitating the selection of optimal alignments based on user specified parameters. This allows customisation of speed, sensitivity, and accuracy, depending on the needs of the user.

The following steps are needed to generate the alignment:

bfast fasta2brg: Generates a compressed binary version of the reference genome from a specified FASTA file. The only required inputs are the reference file (-f) and specification of –A 0 to indicate that the reads are not coded in colour space (-A 1 specifies colour space).

bfast index: Creates indexes of the reference file (-f) and stores these in a compressed binary format. The command –m specifies the mask to use for this

index. The masks are strings of 1 and 0 which specify where mismatches are "allowed" to occur between the reads and reference, and vary depending on the sequencing technology and read length used. Figure 2.6 shows the indexes recommended for Illumina reads with a length less that 40bp. The command should be run 10 times, once for each of the masks, with the index number each time being specified by the -i command (-i 1 for first index, -i 2 for second, etc.). Lookup time is minimised via the incorporation of a hash into the index, the size of which is specified by -w (14 is recommended for short Illumina reads). Specifying the available number of processers (-n) allows the program to make full utilisation of the available resources.

Figure 2.6 – Indexes for BFAST alignment of reads <40bp

```
11111111111111111
111010001110001110100011011111
111101001101111010101101111
111111111111111001111
111101110110010100111111111
111101110001010100000101011101111
101100110101111010011001001011111
111011001010000100010110011100111111
111101111111111111111
110111111000101101111101101
```

Indexes for BFAST indexing step when using Illumina data with a read length less than 40bp.

bfast match: Searches the produced set of reference indexes for CALs for a set of reads (-r, with additional command –z if file is zipped).

bfast localalign: Uses a Smith-Waterman algorithm to perform a local alignment of each read to the reference sequence, and assign each a quality score, based on the list of CALs, allowing for mismatches and gaps. Because this stage can be time consuming, it can be specified that reads with an excessive number of CALs be disregarded (-M 500 tells the program to ignore any reads with more than 500 CALs).

bfast postprocess: Converts the file format to an output format which can be taken forward to use with other programs, in our case, the .sam format, which can easily be converted to the more universally accepted .bam format. This stage of processing can also be used to filter the alignments.

Example commands to run BFAST:

Make BFAST reference file:

```
$ bfast fasta2brg –f
Homo_sapiens.GRCh37.59.dna.toplevel.chr_only.valid.fa –A
0
```

Index reference file:

```
$ bfast index
Homo_sapiens.GRCh37.59.dna.toplevel.chr_only.valid.fa -m
111111111111111111 -w 14 -i 1 -n 8
```

Find CALs:

```
$ bfast match -f
Homo_sapiens.GRCh37.59.dna.toplevel.chr_only.valid.fa -r
1_NoIndex_L001_R1.fastq.gz -z -A 0 -n 8 > Sample1_CAL
```

Run local alignment:

```
$ bfast localalign -f
Homo_sapiens.GRCh37.59.dna.toplevel.chr_only.valid.fa -m
Sample1_CAL -A 0 -M 500 -n 8 > Sample1_aligned.baf
```

Prioritise final alignments with bfast postprocess:

```
$ bfast postprocess -f
Homo_sapiens.GRCh37.59.dna.toplevel.chr_only.valid.fa -I
Sample1_aligned.baf -A 0 -n 8 > Sample1.sam
```

Data Manipulation and Analysis

Samtools

Samtools (Li et al. 2009) is arguably one of the most crucial pieces of software when dealing with NGS data. The majority of downstream analysis programs for NGS data require that aligned files are sorted (by position) and indexed before use. Samtools allows you to do this. It also enables the splitting of files from whole genomic alignments in to sections of interest to facilitate further analysis in a gene specific way. It can also be used to visualise and obtain basic statistics on data, and to convert between commonly used data formats such as SAM (Sequence Alignment/Map) and BAM (a binary form of SAM files). Samtools can be used for variant calling, however, it was not utilised in this way for this project, as our pooled sequencing data required programs specifically designed to handle variant calling in data where multiple individual's genotypes were present.

After the completion of data alignment, Samtools was used to sort and index the data, then divide it in to regions of interest. Example commands to do this are given below:

```
$ samtools sort sample1_aligned.bam
sample1_aligned_sorted.bam
```

Where sample1_aligned.bam is the input file, and sample1_aligned_sorted.bam is the sorted output file.

```
$ samtools index sample1_aligned_sorted.bam
```

This creates a .bai index file specific to that bam file. This allows the retrieval of reads in any given region quickly and efficiently.

```
$ samtools view sample1_aligned_sorted.bam 8:27450349-
27475777 -bo CLU_1.bam
```

Where 8:27450349-27475777 are the genomic coordinates for *CLU*, and -bo specifies that the output file should be in .bam format. This creates a file of all the aligned data within the *CLU* locus. Coordinates for the other regions were 1:207666995-207817219 for *CR1*, 11:85664737-85784019 for *PICALM* and 11:85840498-85870594 for the rs3851179 LD block. These coordinates include the targeted region, plus 500bp either side, to ensure the maximum amount of useful data were included in analyses.

The alignment from BFAST was outputted in .sam, not .bam format, therefore it was necessary to convert to .bam before these steps could be performed:

```
$ samtools view -S sample1_bfast.sam -bo
sample1_bfast.bam
```

Where -S identifies the input file is in .sam format, and -bo specifies that the output file should be in .bam format.

The split files for each of the gene regions were also sorted and indexed before utilisation in other programs (as above).

SamStat

SamStat (Lassmann et al. 2011) is a program that gives statistics on mapped NGS data files, allowing the quality of reads and alignments to be assessed, and any major issues or abnormalities within the data set to be identified. Usage is very simple:

```
$ samstat CLU_1.bam CLU_2.bam ...
```

The output includes information on base quality distribution, read lengths, mapping quality, nucleotide composition and di-nucleotide over-representation.

Integrative Genomics Viewer (IGV)

IGV (Robinson et al. 2011) is a simple and intuitive high-performance data visualisation package from the Broad Institute. It can be used for a variety of different types of data, including NGS and array-based data, and enables the researcher to "see" their data on various scales, across the whole genome or at the single base pair level. It is a Java program which can be installed on Windows, Linux or Mac. Usage is very simple - the desired file is opened within the program (in our case, indexed .bam files) and coordinates can be specified to focus on the regions of interest.

Variant Calling

In this project, the main aim of the study was the detection of variant sites within the sequenced regions. A wide range of variant calling software exists, both commercial and freeware, and each with its own relative strengths and

weaknesses in speed and accuracy. For individual sequencing data, the genome analysis toolkit (GATK) (McKenna et al. 2010) is emerging as the "gold standard" for variant calling, but this method (at the time of writing) was not applicable to pooled sequencing data. Two methods were used for variant calling in our data, both specifically designed with pooled sequencing designs in mind – Syzygy and CRISP.

Syzygy

Syzygy (Rivas et al. 2011) is a piece of software developed at the Broad Institute for the identification of variants in either pooled or individual sequencing data. In addition to simple identification of variants, it offers information on a number of other parameters, such as estimation of allele frequencies, power evaluation, single and group-wise marker association testing and basic annotation of detected variants.

Syzygy requires a number of input files, additional to the aligned sequencing data. A dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) file containing known variants within the region was downloaded from the Tables section of the UCSC genome browser (Kent et al. 2002). A separate reference file for each chromosome was needed (for *CLU, CR1* and *PICALM*; chromosomes 8, 1 and 11 respectively), which again was downloaded from UCSC. For each gene, a manually created tab delimited .pif and .tgf file was also required, following the format shown in figure 2.7. The .pif file provides Syzygy with information on the pools/samples being analysed, while the .tgf file contains information on the targeted region.

Figure 2.7 – Format of Syzygy's .pif and .tgf input files

CLU.pif

PoolBAM	Phenotype	Inds	Chroms	
CLU_s1_bfast.bam		0	12	24
CLU_s2_bfast.bam		0	12	24
CLU_s3_bfast.bam		0	12	24
CLU_s4_bfast.bam		0	12	24
CLU_s5_bfast.bam		0	12	24
CLU_s6_bfast.bam		0	12	24
CLU_s7_bfast.bam		0	12	24
CLU_s8_bfast.bam		0	12	24

CLU.tgf

FEATURE_NAME	CHR	START_POSITION	END_POSITION	LENGTH	GENOME_BUILD
CLU	8	27450349	27475777	25428	19

The format for the .pif file is shown in the top panel. This gives Syzygy information on the pooling design used, firstly specifying the names and location of the .bam files, giving the phenotype (in our case, irrelevant, as only case samples were utilised), along with the number of individuals (12) and chromosomes (24) per pool. Below is the format for the .tgf file. This gives Syzygy information on the regions which have been sequenced (including gene/feature name, chromosomal coordinates, size, and which genome build these numbers refer to).

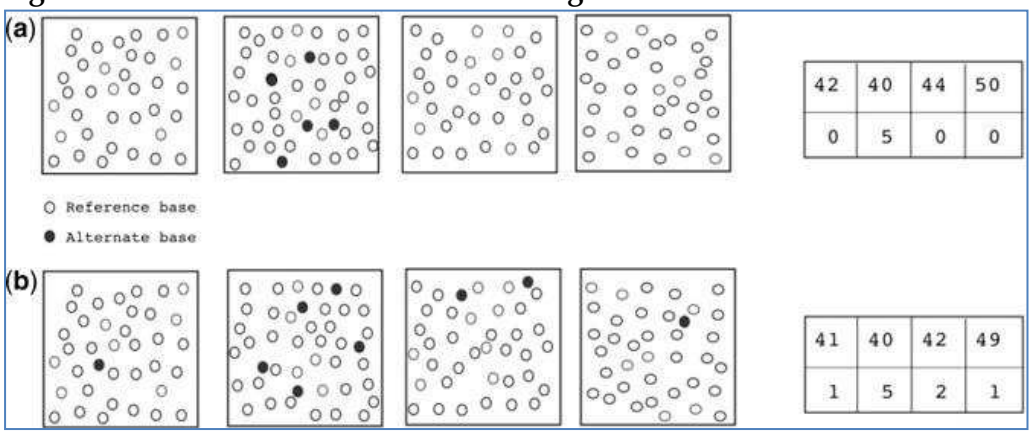
Example command to use Syzygy:

```
$ syzygy --pif CLU.pif --tgf CLU.tgf --outputdir ./ --ref
chr8.fa --hg19 --dbsnp CLU.dbsnp --power
```

CRISP

CRISP (or Comprehensive Read analysis for Identification of SNVs (and short indels) from Pooled sequencing data) (Bansal 2010) is a program designed for the identification of both common and rare variants from at least two pools of sequencing data, although at least five pools are recommended for optimal results. The principle of the program is centred around the concept of comparing variant distribution across multiple sequencing pools to distinguish genuine rare variants from sequencing errors. For common variant identification, individual base quality scores are utilised to evaluate the likelihood of observing that number of non-reference base calls due to sequencing errors alone. Rare variants are assessed by contingency table, comparing the distribution of variant calls across DNA pools, which also reduces the number of false positive variant calls occurring due to sequence context. A diagrammatical explanation of this concept is presented in figure 2.8, taken from Bansal 2010 (Bansal 2010). The program also takes in to account the size of the pool used, and any biases in the distribution of reads on the forward and reverse strands.

Figure 2.8 – Detection of rare variants using CRISP



Each box represents a sequencing pool, with the dots representing individual sequencing calls (white – reference, black – alternative). In scenario A, all alternative calls occur within a single pool. The *p*-value for the contingency table here would be 0.02, indicating it is likely that this data did not arise from random chance alone, and a variant should be called. In scenario B, although there are the same number of alternative calls in the second pool, there are also alternative calls in other pools. The *p*-value for the contingency table in this instance would be 0.24, suggesting that this could have arisen through sequencing errors alone, and therefore should not be identified as a potential variant. Image taken from Bansal 2010 (Bansal 2010).

During the project, several different releases of CRISP were utilised for variant calling (v5, release 071812 and release 082412). The usage instructions below and the variants presented in this document pertain to the 082412 release of CRISP.

Example command to run CRISP:

```
$ ./CRISP_092412 --bam CLU_1.bam --bam CLU_2.bam --bam
CLU_3.bam --bam CLU_4.bam --bam CLU_5.bam --bam CLU_6.bam
--bam CLU_7.bam --bam CLU_8.bam --ref
Homo_sapiens.GRCh37.59.dna.toplevel.chr_only.valid.fa --
poolsize 24 --VCF CLU_variantcalls.vcf --qvoffset 33 --
regions 8:27450849-27475277 > CLU_variantcalls.log
```

The poolsize specified is the haploid number of genomes per pool; in our case 12 individuals = 24 haploid genomes. The command qvoffset specifies the quality value offset to be used in analysis, which should be 33 for data encoded in Sanger format, such as ours. This instructs the program how the quality scores of the data are coded, which varies across different sequencing platforms. Other commands can be incorporated to customise the algorithm, including altering the *p*-value thresholds for reporting SNPs, adjusting minimum base and minimum mapping quality. These options were left at their default settings.

Coverage Calculations

Average coverage for each of the regions of interest was calculated using the total reads aligned to the region (as outputted by SamStat) and the following equation:

Average coverage = (number of reads * read length) / size of region (bp)

This gave the average coverage for the region per pool, which could then be divided by 12 (i.e. the number of individuals in the sample) to calculate average coverage for the region per individual.

% On Target

The percentage of reads mapped to the target region was calculated for each pool, and an average for the whole experiment established, again using SamStat to obtain information on the number of reads which map to the target region, using the following equation:

% on target = (number of reads on target / total number of reads generated) * 100

Enrichment factor calculations

To assess the efficiency of the enrichment, enrichment factor calculations were performed, using the equation below:

Enrichment factor = average depth of coverage at region of interest (ROI)/ average depth of coverage across the rest of the genome

Where:

Average depth of coverage at ROI = reads mapping to ROI * read length (bp) / size of all ROIs (bp)

Average depth of coverage across the rest of the genome = reads mapping to rest of genome * read length (bp) / size of genome (taken to be ~3,200,000,000 bp)

Ts/Tv Ratios

Each SNP identified by CRISP was classified as a transition (A/G, T/C) or a transversion (A/C, A/T, G/C, G/T), and the ratio of transition to transverstions (Ts/Tv ratio) was calculated by dividing the number of transitions by the number of transversions. This was calculated separately for exonic and non-exonic variants.

Variant Effect Predictor

Basic annotation of the polymorphisms called in the NGS data was conducted using Ensembl's Variant Effect Predictor (VEP) (McLaren et al. 2010) (accessed November 2011), which provided information on where the variants lie in relation to the major transcripts of each gene and whether these were novel or had been documented in dbSNP. Additional information can be included for coding variants (e.g. position of affected amino acid in protein sequence, whether the variant is synonymous or non-synonymous, and SIFT/PolyPhen predictions of functional consequences for missense SNPs).

The web based version of the program was utilised, ensuring the most up to date versions of the databases are interrogated, as the standalone perl script version relies on downloading datasets, which will quickly become dated.

Uploading variants for assessment is simple, requiring a .vcf file of SNPs/indels of interest, such as the one generated by CRISP. The output file is typically a text document which can then be opened in Microsoft Excel on Windows or LibreCalc on Linux operating systems.

The VEP gives the relative position of the variants compared to all of the transcripts Ensembl has in its database for that region. Many of these transcripts are not known and confirmed protein coding transcripts, so for each of the genes, the main transcripts of interest were identified, and data for these alone was utilised in further analysis. The selected transcripts were ENST00000316403 for *CLU*, ENST00000393346 for *PICALM*, and ENST00000400960 plus ENST00000367049 for CR1 (encoding the F and S isoforms respectively).

Uploading the variants to UCSC's custom tracks function (Kent et al. 2002) was another way in which the variants detected could be put in context of the genes. This required the data to be in .bed format (a tab delimited format, with column one indicating the chromosome, and columns two and three specifying the start and end coordinates of the variant (for a SNP, these numbers will be the same)). Once the custom track is uploaded, navigating to the region of interest displays the variants where they fall within the gene.

Tabix and the 1000 Genomes Project

In order to obtain the most accurate and up to date frequency estimates from the 1000 genomes project data, a combination of Tabix and an in-house compiled perl script were used. The perl script is given in Appendix 2.2 (written by former colleague, Hui Shi).

The required input files are named Filename1 and Filename2. Filename1 gives details on the sample IDs and the population to which they belong. This is consistent for all genes and was obtained from the 1000 genomes project data site (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/). The second file, specified by filename2 is different for each gene, containing genotype information for a particular region. This was obtained using Tabix (example command shown below):

```
$ ./tabix -hf
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521
/ALL.chr8.phase1_integrated_calls_20101123.snps_indels_sv
s.genotypes.vcf.gz 8:27450849-27475277 > CLU_Tabix_Out
```

Where -hf specifies the location of the data of interest in the 1000genomes data ftp site, and the coordinates of interest are specified, pulling down the data for that region. Data for the four regions were obtained 04.01.2012.

To run the script (example command):

```
$ perl.pl
```

(The input files are specified within the perl script itself)

The output file can then be searched for coordinates of interest, and will give frequencies of variants at those positions for the four major 1000 genomes populations (EUR, AMR, AFR, ASN).

Sequencing Project Two

2.6. *CR1* Sequencing (Take Two)

Due to the poor coverage obtained for the *CR1* region in the first sequencing project, the gene was included in a second sequencing project conducted by the lab. Also included in this project were the other recently identified AD risk genes from GWAS (*ABCA7, BIN1*, the *MS4A* gene cluster, *CD2AP, CD33* and *EPHA1*), in addition to other loci of interest to our group and collaborators (*APOE, VAMP1*, *VAMP2*, and the *IDE/KIF* locus). The details of these regions are presented in table 2.3.

The basic methodology utilised in the second sequencing project was essentially the same as the first, but with a few differences and modifications. Firstly, following analysis of the initial sequencing project data, it was decided to utilise the repeat masker function offered in the design of the SureSelect baits, given that repetitive regions were poorly represented in the initial project, and that the inclusion of such regions was believed to contribute to the lower than expected percentage of reads mapping to the target regions. In the intervening time, a less stringent repeat masker had been incorporated into the eArray system, so this was utilised. Additionally, regions which can be problematic for sequencing, such as GC rich regions, were "boosted" in the design process, an option which Agilent's online e-array system offers. This essentially means problematic regions were targeted by an increased number of baits, increasing the likelihood that these regions would be successfully sequenced.

Given the advances in sequencing technology in the intervening time between the two projects, a different sequencing technology was utilised for the second sequencing project. Illumina's HiSeq was used for the second project. With advances in technology and chemistry, this meant a greater quantity and quality of data were produced. Furthermore, a longer read length was utilised, and paired-end reads instead of single-end reads were used. Both of these factors contribute to more reliable alignment of reads to the reference genome. The use of 100bp reads, rather than 35bp makes alignment easier, with more reads uniquely matching to a single position in the genome. The use of paired-end reads also enhances alignment. This is where a fragment of DNA of a known approximate length is sequenced from both ends, giving 100bp of sequence information for each, as well as positional information. Since it is known how far apart these reads should be it aids the mapping process, and improves the elucidation of structural rearrangements and copy number variation.

Table 2.3 – Regions targeted by second sequencing project

Locus	Coordinates	bp	Target bp Covered	% bp Baited	bp Excised Extra +/- ~150bp
CR1	Chr1:207667495-207816719	149224	110252	73.88	1:207667345-207816869
ABCA7	Chr19:1038952-1066720	27768	21472	77.32	19:1038800-1066850
BIN1	Chr2:127778085-127895723	117638	98261	83.53	2:127777935-127895873
*MS4A*_LD_Locus	Chr11:59856028-60041296	185268	129093	69.68	11:59855878-60041446
VAMP1	Chr12:6566406-6584843	18437	13059	70.83	12:6566256-6584993
VAMP2	Chr17:8057465-8071293	13828	10852	78.47	17:8057315-8071443
TRIM15	Chr6:30132298-30138448	6150	4426	71.96	6:30132148-30138598
SPARCL1	A - Chr4:88442070-88447047	4977	3981	79.97	
	B - Chr4:88426174-88435612	9438	7011	74.28	
	C - Chr4:88403759-88416324	12565	5467	43.51	
CD2AP	Chr6:47427281-47601015	173734	120861	69.56	6:47427131-47601165
CD33	Chr19:51718317-51748546	30229	18690	61.83	19:51718167-51748696
EPHA1	Chr7:143082382-143110385	28003	24275	86.68	7:143082232-143110535
IDE_KIF11_HHEX	Chr10:94192885-94491751	298866	169051	56.56	10:94192735-94491901
APOE	Chr19:45260160-45451160	191000	96074	50.30	19:45260010-45451310
	Total/**Average**	1,267,125	832,825	**69.89**	

Details of the targeted regions from the second sequencing project, including genes and coordinates, the size of the target region, and the amount of the target region and percentage successfully baited. Coordinates in the last column indicate the final region actually targeted. For the SPARCL1 targets, 150bp had already been added to the ends of the region, so this was not repeated.

Data analysis was largely conducted as before. Based on the first project, BFAST was selected as the alignment method of choice (with slight modifications to adapt to the different sequencing technology used), while CRISP was utilised for variant calling (as before). For the justification of these decisions, see Section 3 – defining the pipeline.

BFAST

Because the data for the second project was paired end data, the files had to be interleaved before BFAST could be run. This process involved taking the two separate files (one for the first read set, one for the second read set (the pairs)) and generating a single file with both reads aligned to the same strand (requiring the reverse complement of the second read set to be generated), whilst maintaining the quality score information for both.

The script (shuffle_fastq_new_100412.pl) used to do this was written in-house, with guidance from Nils Homer, the creator of BFAST. The contents of this script are shown in Appendix section 2.3.

Example command to run script:

```
$ perl shuffle_fastq_new_100412.pl
Extracted_Reads/L001_R1.fq Extracted_Reads/L001_R2.fq
Sample1_R1R2.fastq
```

After the interleaving step, alignment was run as documented for project 1, except that the indexes used were different since the reads here were longer (see figure 2.9).

Figure 2.9 – Indexes for BFAST alignment of reads >40bp

```
1111111111111111111111
1111101110111010100101011011111
1011110101101001011000011010001111111
1011100110100110010011110101000101111
1111101101110111011111111
1111111001010010001011111101110111
1111010111001010001010110101010111111
1111011010110110011000001011010010111101
1111011010001000110101100101100110100111
1111010010110110101110010110111011
```

Indexes for BFAST indexing step when using Illumina data with a read length greater than 40bp.

2.7. Prioritisation and Validation

Sanger Sequencing

PCR primers were designed to amplify a region including at least 100bp either side of the positions of interest using Primer3 (Rozen and Skaletsky 2000) v0.4.0 (http://frodo.wi.mit.edu/). Specificity for each primer pair was checked using UCSC's (Kent et al. 2002) Virtual PCR function (http://genome.ucsc.edu/cgi-bin/hgPcr), and the primer binding sites were determined to be free of known polymorphisms using NGRL Manchester's SNPCheck v2.1 (https://ngrl.manchester.ac.uk/SNPCheckV2/snpcheck.htm).

PCR optimisation and amplifications were completed following the standard laboratory protocol (reaction mix: 1xPCR buffer (Roche Diagnostics Corp.); 200μM dNTPs (Fermentas); 1μM of each primer (Eurogentec Biologics); 1 unit Taq DNA Polymerase (Roche Diagnostics Corp.); plus molecular grade water up to a final volume of 30μl. Primer concentrations were halved for 8:27466924 and doubled for 1:207690803 after optimisation. Thermal cycling conditions used were 94°C for two minutes; 30 cycles of 94°C for 30 seconds, appropriate annealing temperature for 1 minute, 72°C for 1 minute; and finally 72°C for 7 minutes). Primer sequences and annealing temperatures for the SNPs validated by this method are shown in Table 2.4. Sequencing was conducted using PCR primers with Applied Biosystems BigDye Terminator v3.1 chemistry, run on the ABI 3130xl (Applied Biosystems). Chromas Lite v2.01 (http://www.technelysium.com.au/chromas_lite.html) was used to visualise electropherograms which were assessed by eye to determine genotype. In each case, one pool of samples (12 individuals) was Sanger sequenced. The pool to be sequenced was selected based on having the highest proportion of alternative reads in the NGS data at the position of interest.

Table 2.4 – PCR primers and annealing temperatures

Variant (Chr:coordinate)	Forward Primer Sequence	Reverse Primer Sequence	Annealing Temperature
8:27452179	GCG-GTG-AGC-TAT-GAT-TCC-AC	GCT-CAG-GTG-CCC-AAT-CCT-AT	64ºC
8:27466924	CTG-CAC-CCT-ACT-GCT-TAG-AAA	TGC-ATT-TGT-CAC-CAG-TGC-TAT	54ºC
8:27473743	ATG-AGG-AAT-CGG-GAA-TGG-AT	GGA-GCG-AGC-TCA-AAA-ACA-AT	60ºC
11:85668102	CAC-CCA-GCT-CCT-TTT-CTG-AT	GGA-TCA-AAA-GCT-TTG-CAT-TGA	58ºC
11:85692077	TGG-AAT-ATG-TCT-GGC-ACA-AAG	GGG-ATC-TAA-CTG-GCA-ACC-AA	58ºC
11:85774424	TGT-CTC-ACA-AAG-CGT-ATG-AAA-G	GGC-AGA-ACA-GAA-TGC-CTG-AG	60ºC
1:207690803	GTG-TGT-GCA-GGA-TTG-CTC-AT	TGT-TAC-ACA-AAT-TGT-TCC-AGA-CA	62ºC

Primers and annealing temperatures for the variants selected for Sanger validation from the NGS data.

Exome Project

Validation by Sanger sequencing is both expensive and time consuming: impractical for the large number of variants detected by our NGS experiments.

Through collaboration with John Hardy's group at UCL (correspondence via Rita Guerriero), we were able to pseudo-validate our coding variants. The group at UCL had exome sequencing data for up to 143 cases and 183 controls, and were able to look up our exonic variants in their data set, providing immediate "validation" for our SNPs, since variants present in both data sets were assumed to be genuine (data as published in (Guerreiro et al. 2013)).

Imputation

Imputation was used to enable us to perform association testing on detected variants within GWAS data available to us. This was in the form of a combined ARUK/Mayo GWAS data set, comprised of 2067 AD cases and 1376 controls, genotyped using Illumina's HumanHap300v1. 6000 control samples from the WTCCC2 project (3000 from the 1958 birth cohort (58C), and 3000 from the National Blood Service (NBS) cohort), genotyped using the Illumina 1.2M (custom) chip were also utilised. These are population controls, so some may develop AD later in life.

Both the merged data set and the WTCCC2 data sets were aligned to hg18, while the reference haplotypes to be used were aligned to hg19. To accommodate this, the GWAS data were converted to hg19 using plink (http://pngu.mgh.harvard.edu/purcell/plink/) (Purcell et al. 2007) and UCSC's liftOver program (available online at http://genome.ucsc.edu/cgi-bin/hgLiftOver or as a downloadable tool at http://hgdownload.cse.ucsc.edu /admin/exe/).

To convert the Mayo/ARUK merged data set:

```
$ ./plink --bfile MERGED_MAYO_ARUK --recode --out
Merged_Mayo_ARUK_hg18 --noweb --allownosex
```

Where MERGED_MAYO_ARUK specified the input files (.bed, .bim, .fam), and the output is in .map and .ped format. --noweb instructs plink not to connect to the web for updates, as this can disrupt the running of the program, and --allownosex allows individuals without gender information to be included.

The input for liftOver required chromosome, start position, end position, and rsID. All of this information was extracted from the .map file using the following awk command:

```
$ awk '{print "chr"$1, "\t", $4, "\t", $4+1 "\t", $2}c'
Merged_Mayo_ARUK_hg18.map > out_chr.bed
```

liftOver was then used to change the coordinates in the file to hg19, using a chain file downloaded from UCSC (hg18Tohg19.over.chain.gz, from http://hgdownload.cse.ucsc.edu/goldenPath/hg18/liftOver/):

```
$ ./liftOver out_chr.bed hg18Tohg19.over.chain.gz
Merged_hg19.bed unMapped
```

The unMapped command means any coordinates that cannot be lifted over are printed to a separate file.

To recode the actual GWAS data, plink required a file that was simply rsID and coordinate, separated by a tab. Again, awk commands were used to obtain this information:

```
$ awk '{print $4, "\t", $2}' Merged_hg19.bed >
hg19_Merged.txt
```

Then plink was used to recode the GWAS data:

```
$ ./plink --bfile MERGED_MAYO_ARUK --update-map
hg19_Merged.txt --make-bed --out Merged_data_hg19 --noweb
--allownosex
```

This gave the .bed, .bim and .fam files for the combined data set in hg19. Regions of interest (in this case individual chromosomes) were then separated out of the whole genomic file:

```
$ ./plink --bfile Merged_data_hg19 --chr8 --make-bed --
out chr8_Merged --noweb --allownosex
```

Giving .bed, .bim and .fam for chromosome 8.

```
$  ./plink --bfile chr8_Merged  --recode --out
chr8_Merged --noweb --allownosex
```

Giving .map and .ped files for the chromosome, which can then be converted via gtool (http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html) to the .gens and .samples files required by the imputation software.

```
$ ./gtool -P --ped chr8_Merged.ped --map chr8_Merged.map
--og chr8_Merged.gens --os chr8_Merged.samples
```

This was done for all of the chromosomes of interest: 8, 1, 11, and 19, where the *APOE* locus lies. The .samples file required some editing (under "phenotype" P was changed to B to demonstrate that it is a binary trait, and plink's 1+2 coding (for cases and controls) was altered to 0+1 coding. A final column was also added as a covariate called "centre", with D on the second row (to demonstrate it was a discrete not continuous covariate). ARUK samples were coded 1, while Mayo samples were coded 4, allowing for correction based on centre when running association tests later. Any missing values were changed from plink's default of -9 to "NA", which is recognised as "value missing" when running the imputation and association testing.

For the WTCCC2 data, preparation was conducted slightly differently: firstly, the data was encrypted, so needed to be decrypted and unpacked using Kleopatra (http://www.kde.org/applications/utilities/kleopatra/) and the encryption key supplied. This data was also different in that each chromosome had a separate .gens file. To facilitate the alteration of the data from hg18 to hg19, a colleague, Christopher Medway, wrote a number of perl scripts.

The first step was to unzip the .gens data file, then similar to above, use awk commands to extract the required information for the liftover step:

```
$ awk '{print "chr8", "\t", "$3", "\t", $3+1, "\t", $1}'
58C_chr8.gens > 58C_chr8.bed
```

The chromosome number needed to be adjusted so that all chromosomes of interest, for both the NBS and 58C data were processed.

This .bed file was then used in the liftOver process:

```
$ ./liftOver 58C_chr8.bed hg18Tohg19_over.chain.gz
chr8_hg19_58C.bed unMapped
```

The output from this, plus the original .gens file (in this example, 58C_chr8.gens) could then be called in to the perl script liftOverGen.pl (see Appendix section 2.4), which replaces the coordinate in the .gens file to match hg19 numbering:

```
$ perl liftOverGen.pl chr8_58C.gens chr8_hg19_58C.bed
```

N.B. The output file each time was named OUTPUT.txt, so needed renaming (in this case, to chr8_hg19_58C.gens) before running the program again, to ensure this is not overwritten.

The WTCCC2 .samples files also required some recoding (e.g. recode phenotype, add centre information), which again, Christopher Medway compiled a perl script to complete - recode_WTCCC2.sample.pl (see Appendix section 2.5).

To run the script:

```
$ perl recode_WTCCC2.sample.pl WTCCC2_NBS.sample
```

Once all the files were prepared, the imputation itself was run. Imputev2.2.2 (http://mathgen.stats.ox.ac.uk/impute/impute_v2.html) (Howie et al. 2009) was used - with each data file requiring separate processing (i.e. each of the chromosomes of interest, for each of the three (ARUK/Mayo Merged, 58C and NBS) datasets).

Example command:

```
$ ./impute2 -m genetic_map_chr8_combined_b37.txt -h
ALL_1000G_phase1integrated_v3_chr8_impute.hap.gz -l
ALL_1000G_phase1integrated_v3_chr8_impute.legend.gz -g
chr8_58C_hg19_58C.gens -align_by_maf_g -int 26450849
28475277 -Ne 20000 -o CLU_58C_phased.impute2
```

Where -m gives a fine scale recombination map for the region to be analysed (in this case, chromosome 8 (downloaded from impute2 website)), -h is a file of known haplotypes and -l is the corresponding legend file (both also downloaded from impute2 website). -g is the data file containing genotypes for the GWAS data (in .gens format, as created above), and requires a corresponding .samples file (again, as discussed above). The format of these files is shown in Figure 2.10. -int specifies the region to be imputed (26450849-28475277 for *CLU*, 84665237-86870094 *PICALM*, 206667495-208816719 for *CR1*, and 44394477-46412650 for *APOE*). -Ne specifies the effective size of the population from which the imputed data was sampled. When using the full panel of reference hapotypes (as recommended), an -Ne of 20,000 is suggested. The command -align_by_maf_g was used for the WTCCC2 data, as no strand file was available for this data. This instructs impute2 to deduce the strand for each variant. For the Merged data, the strand file BDCHP-1x10-HumanHap300v1-1_11219278_C-b37-strand.zip was used (downloaded from http://www.well.ox.ac.uk/~wrayner/strand/). This had to be unzipped, then the necessary information extracted (impute2 strand files only need coordinate and a + or - to specify the strand, the file as downloaded contained extra information) using an awk command:

```
$ awk '{print $3, $5}' BDCHP-1x10-HumanHap300v1-
1_11219278_C-b37-strand > HumanHap300v1.stand
```

Figure 2.10 – Format for .gens and .samples files for Impute2

```
SNP1 rs1 1000 A C 1 0 0 1 0 0
SNP2 rs2 2000 G T 1 0 0 0 1 0
SNP3 rs3 3000 C T 1 0 0 0 1 0
SNP4 rs4 4000 C T 0 1 0 0 1 0
SNP5 rs5 5000 A G 0 1 0 0 0 1
```

chr8_Merged.samples

```
ID_1 ID_2 missing sex phenotype centre
0 0 0 D B B
68396_D02_LOAD393281 68396_D02_LOAD393281 0.00121073 2 1 1
68396_E02_LOAD393293 68396_E02_LOAD393293 0.000813769 1 1 1
68397_C05_LOAD392730 68397_C05_LOAD392730 0.000592133 2 1 1
68397_E05_LOAD393458 68397_E05_LOAD393458 0.000615289 1 1 1
```

Above: Format of the .gens file taken from the Impute2 file format website (http://www.stats.ox.ac.uk/~marchini/software/gwas/file_format.html). There is one line of information per SNP, including variant ID, rs ID, coordinate, allele A and allele B. The next numbers are the probabilities of genotypes (homozygous allele A, heterozygous, homozygous allele B) for that individual at the SNPs in question. Below: Format of the .samples file, example shown is for chromosome 8 from the ARUK/Mayo merged data set. The top line is the header, specifying the contents of

each column, followed by a line describing the type of variable in each column where necessary (D for discrete covariates, B for binary phenotypes (0 = controls, 1 = cases)). The following lines contain the information specified in the header for each of the individuals being included in the imputation, with one line per individual.

Once the imputation was run for all regions and all datasets, another perl script written by Christopher Medway (remove_dup_lines.pl) was used to remove duplicate positions. The datasets were generated using different chips, so the command -overlap will need to be used when association testing, but this will not work if the files have more than one variant at a given position. The contents of this perl script are given in Appendix section 2.6, and an example command is given below:

```
$ perl remove_dup_lines.pl CLU_58C_phased.impute2
```

The output from this script was named in the form of CLU_58C_phased.impute2_duplicates_removed. Snptest_v2.4.1 (https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html) (Marchini and Howie 2010) was then used to test the imputed variants for association with AD. An example command is given below:

```
$ ./snptest_v2.4.1 -data
CLU_Merged_phased.impute2_duplicates_removed
chr8_Merged.samples
CLU_NBS_phased.impute2_duplicates_removed
chr8_NBS.samples
CLU_58C_phased.impute2_duplicates_removed
chr8_58C.samples -o CLU_Full_snptest.out -frequentist 1 -
method threshold -pheno phenotype -cov_names centre -
overlap
```

2.8. *In Silico* Functional Analyses of Coding Variants

A number of *in silico* functional analyses were conducted on the coding variants to prioritise those likely to be having functional consequences on protein structure, expression or regulation of the gene.

LD Calculations

Calculating the level of LD between the GWAS SNP and detected variants can help disentangle which variants could be underlying the reported association with AD from GWAS, and which may be independent from this effect.

Christopher Medway wrote a perl program (LD_calculator.pl – script contents shown in Appendix section 2.8), utilising VCFtools (http://vcftools.sourceforge .net/) and Tabix (http://samtools.sourceforge.net/tabix.shtml), to calculate the level of LD between a specified SNP (in our case, the main GWAS SNP(s) for the locus) and a list of variants of interest presented by chromosome number, coordinate and rs number (if available) in a tab delimited text format. The files

Phase1_Samples_GBR_CEU and vcf-subset were needed to filter the samples to give only data for GBR and CEU populations (the most relevant to our study samples).

Example command:

```
$ perl LD_calculator.pl rs11136000 chr8 27464519
CLU_SNPs.txt
```

Where rs___ gives the identity of the reference variant (here, the *CLU* GWAS SNP), followed by the chromosome and coordinate of that variant

Output from the program gives LD scores in both D' and r^2 for the variants, as well as the number of samples on which this has been calculated.

Prediction of Non-synonymous Functional Consequences

Included in the output of Ensembl's VEP (McLaren et al. 2010) were Polyphen (Adzhubei et al. 2010) predictions for the consequence of the variants detected on the structural integrity and function of the proteins.

Splicing Investigations

A number of different programs were utilised to assess any impact on the splicing of the gene likely to arise from the observed variants in our samples.

Preparing the input for the programs involved obtaining the sequence of the affected exon, plus at least 100bp of surrounding intronic sequence, for each of the variants of interest, both synonymous and non-synonymous changes. The wild type and variant forms of the sequence were run through the programs, and any differences between the two recorded. The programs used were all web based interfaces, all functioning in a similar way.

ESEfinder v3.0: ESEfinder (Cartegni et al. 2003) incorporates two different functions for the assessment of splicing variants, both of which were utilised. The first (SpliceSites) predicts the actual donor and acceptor sites within the sequence provided. Mouse splice sites, included as default, were not included in our analysis. The other function, SRProteins predicts binding sites for the serine/arginine-rich splicing regulatory protein family and was used with default settings.

BDGP: The Berkeley Drosophila Genome Project site (BDGP) (Reese et al. 1997) offers a tool for predicting splice donor and acceptor sites. Used "Human/other" option, all other settings as default.

NetGene2: Again, NetGene2 (Hebsgaard et al. 1996) provides a program for the prediction of splice sites (again, used "Human" option, and all settings as default).

UTR variants and miRNAs

Ten of the variants called at the *CLU* locus fell within the gene's 3' untranslated region (UTR). Knowing variants in this region can have an impact on the regulation and expression of a gene, partly through the binding of miRNAs, TargetScan v6.1 (Garcia et al. 2011) was used. This predicts the binding sites of miRNAs by searching for conserved 7-8bp regions that match the seed region of known miRNAs. Entering the desired species (Human) and gene ID (*CLU*) gave the locations of all miRNA binding sites predicted within the UTR of that gene. These were then overlaid with the variants detected to see if any of the variants fell within these sites, and how this may affect the gene.

Any variants which fell within predicted miRNA sites via this method were then followed up in a second program, PITA from the Segal Lab of Computational Biology at the Weizmann Institute (Kertesz et al. 2007). This looks at miRNA binding sites in a given UTR sequence, so it was possible to run predictions for the wild type version of the sequence, as well as the sequence of the UTR with the variant positions included individually. The "Predict Your UTR" function was utilised, with default settings, for each of the UTR sequences.

2.9. *In Silico* Functional Assessment of Non-coding Variants

Due to the large number of detected non-coding variants, a number of *in silico* resources were utilised to assess which of the variants could be having functional consequences on gene regulation, and thus should be prioritised for further study.

As above, levels of LD with the reference SNP were calculated.

Conservation

Conservation at the site of variants of interest could highlight potentially interesting SNPs, since those under strict genetic conservation are likely to be more damaging when altered. To assess the conservation at each of the variant sites, a perl script (conservation.pl, see Appendix section 2.7) was compiled by Christopher Medway to extract conservation information downloaded from the UCSC genome browser (tables > comparative genomics > conservation) for relevant positions. Two types of conservation score are contained within UCSC - Phastcons and PhyloP - mammalian conservation scores for each for the regions of interest were downloaded (Siepel et al. 2005; Pollard et al. 2010). To run the script, this data, plus a SNP list (bp coordinate only, in a text document) for the variants of interest were needed. An example command to run the script is shown below:

```
$ perl conservation.pl CLU.phastcon CLU_SNPs.txt >
CLU_phast.txt
```

ENCODE Data

The UCSC genome browser contains a wealth of information from the Encyclopaedia of DNA Elements (ENCODE) project (ENCODE 2011). Colleague, Christopher Medway, wrote a perl script to extract information of interest (DNAseI hypersensitivity clusters, transcription factor binding sites (from ChIP-Seq) and acetylation/methylation data, all of which can imply functional activity) from the UCSC genome browser, based on Ensembl's VEP. This program is pending publication, so the perl code is not given.

TaqMan Assays

TaqMan assays were used to genotype two potentially functional SNPs in the rs3851179 LD block region in an independent case-control cohort to test for association with AD.

The TaqMan assays for the two variants were designed and synthesised by Applied Biosystems. Table 2.5 gives the details of these assays (including primer and probe sequences, and the number of individuals (cases and controls) genotyped using each assay). All genotyping was carried out by Ng See May, as the basis of her dissertation for her MSc in Molecular Diagnostics. TaqMan reactions were conducted in a total volume of 20µl, with 0.9xTaqMan Universal PCR Master Mix, 1xTaqMan SNP Genotyping Assay Mix (both from Applied Biosystems), plus 10ng of genomic DNA and nuclease free water to reach the final volume. No-template controls (NTCs) were included in each run to test for contamination. An MX3000P Real-Time PCR Thermocycler from Stratagene was used for the experiment, with cycling conditions of 50°C for 2 mins, followed by a denaturation cycle at 95°C for 10 mins, and 50 cycles of denaturation and annealing (respectively, 92°C for 15 seconds, 60°C for 1 minute). Genotypes were analysed using Agilent's MxPro™ software.

Association testing was performed by Fisher's Exact Test using SPSSv19, and the power of the study was assessed using Quanto.

Table 2.5 – TaqMan assay design for rs3851179 LD block variants

SNP	Subject (n)		Primer sequence (5'->3')	Probes
	Case	Control		
11:85862491	247	215	F: GAACCCTGAGTCTCCCAGATACT R: TCCAGCCAGCCCAAATCC	VIC-5'-CCCTGTAGCAATCAA-3'-NFQ FAM-5'-CTGTGGCAATCAA-3'-NFQ
11:85862739	237	205	F:TGAAACAGACCTGTTGCTATTCTAAGG R: GCCTGAAGCTGGCATGTTT	VIC-5'-AGTCTCACAAGTCACCATAT-3'-NFQ FAM-5'-AGTCTCACAAATCACCATAT-3'-NFQ

Information on TaqMan assay design, and the number of individuals genotyped with each assay, conducted by Ng See May.

Sanger validation of the two variants was also conducted, as described above. As the variants were only ~250bp apart, the same primers could be used in the sequencing of the two variants. Details of the PCR and sequencing primers used are given in Table 2.6. The sequencing provided positive control samples to use in the TaqMan genotyping assays where appropriate.

Table 2.6 – Sanger sequencing rs3851179 LD block variants

SNP coordinate	Alleles	Location in gene	MAF	PCR primer sequence (5'->3')	Sequencing primer sequence (5'->3')
11:85862491	A>G	82.4kb upstream	2.9% 0.5%[a]	F2:CAGTCCCAGCCCCTATAAAATAAGG	F2:CAGTCCCAGCCCCTATAAAATAAGG
11:85862739	G>A	82.6kb upstream	1.0%	R2:TGGCCTCATGGGTGGGAACA	NR2:GTACCAGCCAGCCACATCATTCA

Information on the variants and primers for the SNPs Sanger sequenced by Ng See May in the LD block containing rs3851179 near *PICALM*.

3. Data Analysis

This chapter describes in detail the processing of the NGS data, from raw reads in FastQ format to identified variants. A number of different programs were utilised and compared, enabling the development of a pipeline of programs best suited to handling data of this type.

3.1. Next Generation Sequencing

From the first sequencing project, a total of 380,589,701 38bp single end reads were obtained from Source Biosciences from the Illumina GaIIX. Each pool of samples was run on a separate lane of the flow cell. A breakdown of the reads and qualities per pool is given in Table 3.1.

Table 3.1 – Sequencing statistics from NGS project 1

Pool	Yield (Mbases)	% Clusters that passed filtering	Number of reads	% of >=Q30 bases	Mean quality score
1	1418	80.73	46,208,988	94.57	37.48
2	1405	82.77	44,684,019	95.13	37.66
3	1478	75.79	51,324,331	93.27	37.02
4	1227	61.80	52,242,572	87.86	35.35
5	1385	67.04	54,350,139	88.23	35.48
6	1392	83.13	44,061,669	95.04	37.62
7	1254	85.18	38,748,346	96.04	38.00
8	1467	78.83	48,969,637	94.09	37.30

Information on data produced by Source Biosciences from NGS project 1, including information on the number and quality of reads produced.

From the second sequencing project, 2,170,649,020 100bp paired end reads were obtained from Source Biosciences from Illumina's HiSeq. Again, each pool of samples was run on a separate lane of the flow cell, and a summary of reads and qualities per pool is provided in Table 3.2.

Table 3.2 – Sequencing statistics from NGS project 2

Pool	Yield (Mbases)	% Clusters that passed filtering	Number of reads	% of >=Q30 bases	Mean quality score
1	21,981	92.48	237,697,912	86.02	34.26
2	25,032	90.47	276,674,942	84.15	33.69
3	22,541	92.44	243,847,388	85.89	34.23
4	25,668	89.86	285,634,842,	84.56	33.80
5	24,206	91.22	265,350,864	85.57	34.11
6	27,729	87.26	317,781,166	81.25	32.81
7	23,356	91.60	254,987,838	85.19	34.01
8	25,850	89.55	288,674,068	83.13	33.38

Information on data produced by Source Biosciences from NGS project 2, including information on the number and quality of reads produced.

3.2. FastQC

An initial assessment of the quality of the data produced by each NGS run was conducted using FastQC.

The program gives summary statistics for multiple parameters, classifying each file with a pass, warning or fail, designed to highlight any potential issues or biases with the data produced when compared to a "normal" NGS dataset. The summary information from FastQC for the first sequencing project is given in Table 3.3, with the same information for the second sequencing run presented in Table 3.4. The program also provides graphical representations of these parameters. The graphs for the per-base sequence quality scores for each sample, which provides the most insight in to the overall quality of the data, are presented in Figure 3.1 and Figure 3.2 for NGS projects 1 and 2 respectively.

Table 3.3 – FastQC summary statistics for NGS project 1

Sample	Per base sequence quality	Per base sequence quality scores	Per base sequence content	per base GC content	Per sequence GC content	Per base N content	Sequence length distribution	Sequence duplication levels	Over-represented sequences
1	Pass	Pass	Pass	Pass	Warning	Pass	Pass	Pass	Warning
2	Pass	Pass	Pass	Pass	Warning	Pass	Pass	Pass	Warning
3	Pass	Pass	Pass	Pass	Warning	Pass	Pass	Pass	Warning
4	Fail	Pass	Pass	Pass	Warning	Pass	Pass	Pass	Warning
5	Fail	Pass	Pass	Pass	Warning	Pass	Pass	Pass	Warning
6	Pass	Pass	Pass	Pass	Warning	Pass	Pass	Pass	Warning
7	Pass	Pass	Pass	Pass	Warning	Pass	Pass	Pass	Warning
8	Pass	Pass	Pass	Pass	Warning	Pass	Pass	Pass	Warning

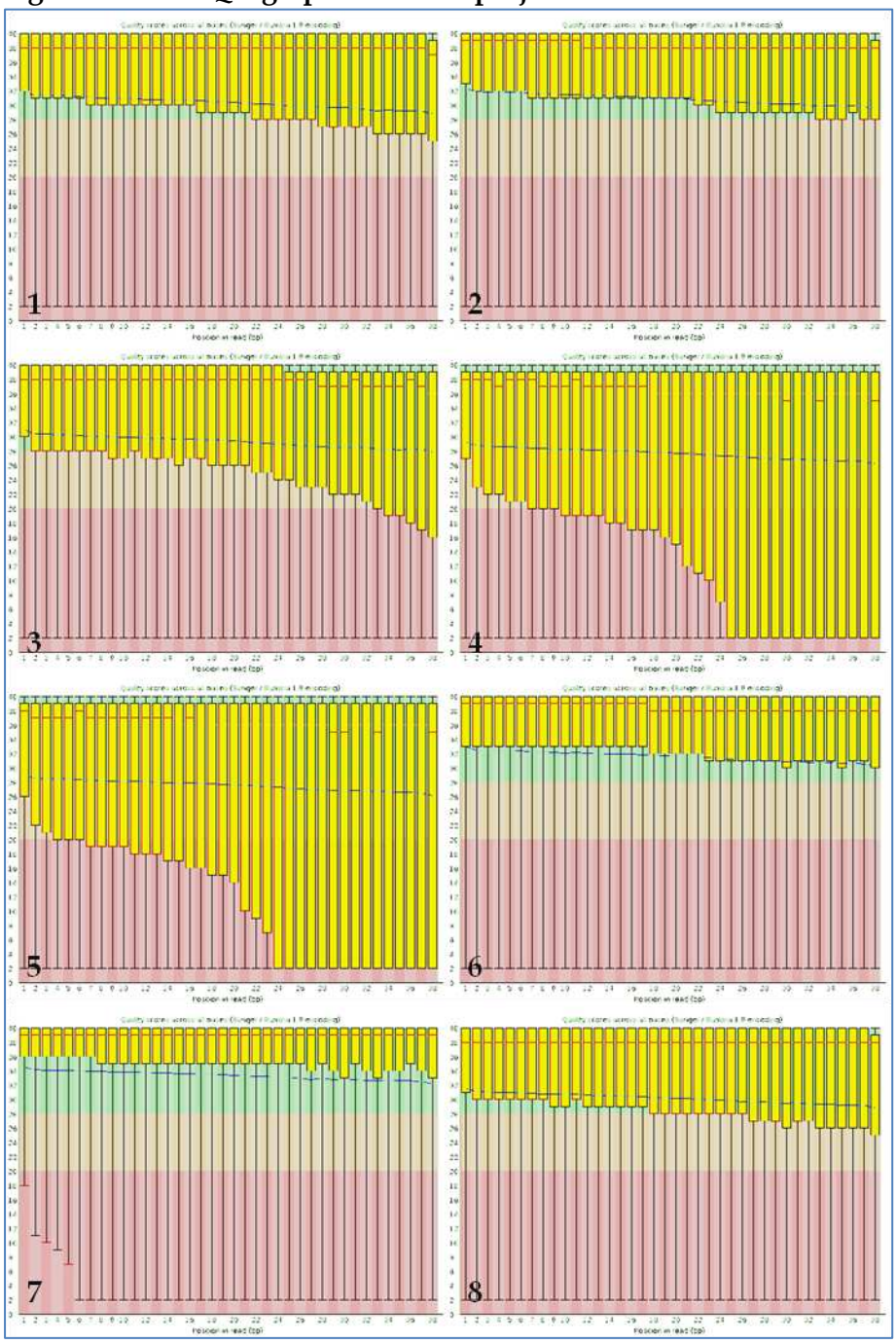
FastQC summary statistics on raw data quality from NGS project 1

Table 3.4 – Fast QC summery statistics for NGS project 2

Sample	Per base sequence quality	Per base sequence quality scores	Per base sequence content	per base GC content	Per sequence GC content	Per base N content	Sequence length distribution	Sequence duplication levels	Over-represented sequences
1 (R1)	Fail	Pass	Warning	Pass	Warning	Pass	Pass	Fail	Pass
1 (R2)	Fail	Pass	Warning	Pass	Warning	Pass	Pass	Fail	Pass
2 (R1)	Fail	Pass	Warning	Pass	Warning	Pass	Pass	Fail	Pass
2 (R2)	Fail	Pass	Warning	Pass	Warning	Pass	Pass	Fail	Pass
3 (R1)	Fail	Pass	Warning	Pass	Warning	Pass	Pass	Fail	Pass
3 (R2)	Fail	Pass	Warning	Pass	Warning	Pass	Pass	Fail	Pass
4 (R1)	Fail	Pass	Warning	Pass	Warning	Pass	Pass	Fail	Pass
4 (R2)	Fail	Pass	Warning	Pass	Warning	Pass	Pass	Fail	Pass
5 (R1)	Fail	Pass	Warning	Pass	Warning	Pass	Pass	Fail	Pass
5 (R2)	Fail	Pass	Warning	Pass	Warning	Pass	Pass	Fail	Pass
6 (R1)	Fail	Pass	Warning	Pass	Warning	Pass	Pass	Fail	Pass
6 (R2)	Fail	Pass	Warning	Pass	Warning	Pass	Pass	Fail	Pass
7 (R1)	Fail	Pass	Warning	Pass	Warning	Pass	Pass	Fail	Pass
7 (R2)	Fail	Pass	Warning	Pass	Warning	Pass	Pass	Fail	Pass
8 (R1)	Fail	Pass	Warning	Pass	Warning	Pass	Pass	Fail	Pass
8 (R2)	Fail	Pass	Warning	Pass	Warning	Pass	Pass	Fail	Pass

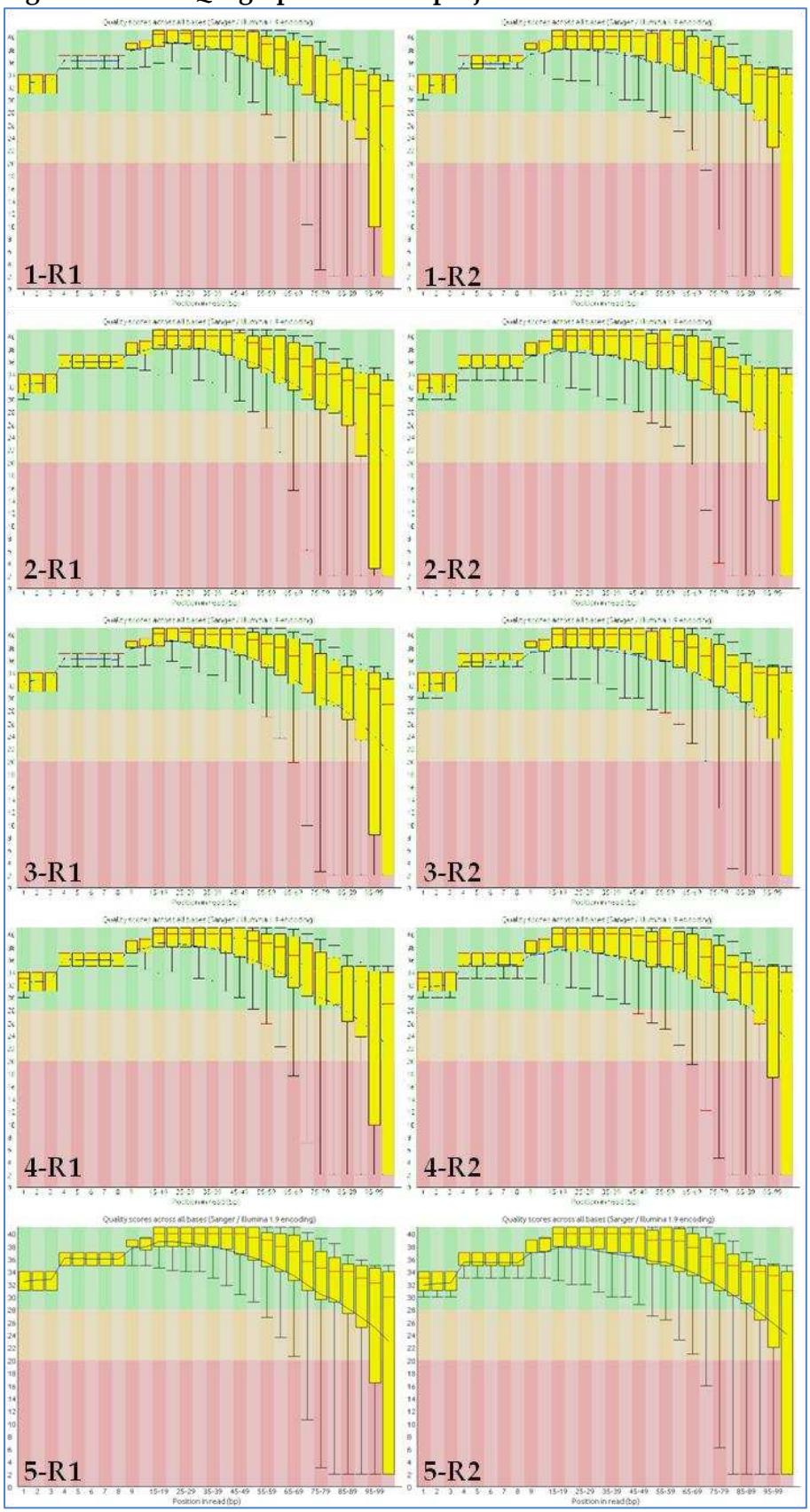
FastQC summary statistics on raw data quality from NGS project 2. Each sample has two data files this time – the forward and reverse reads generated by the use of paired end reads, indicated by R1 and R2.

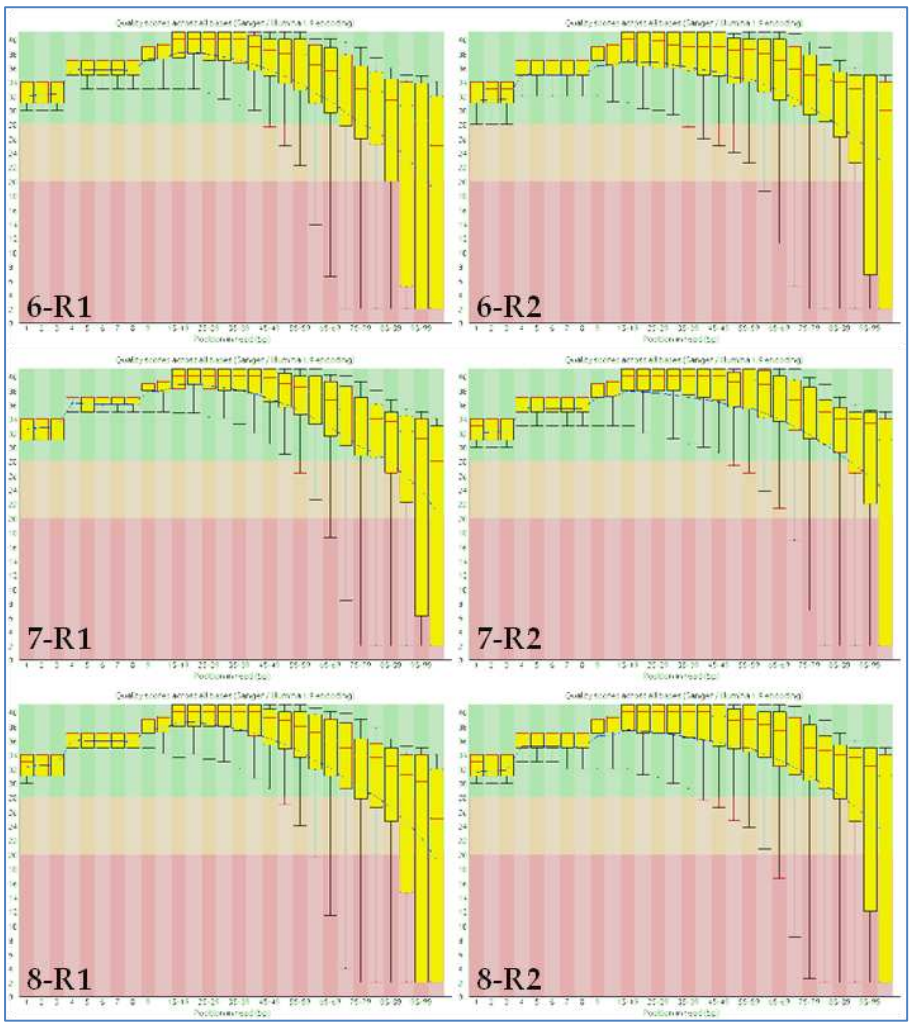
Figure 3.1 – FastQC graphs for NGS project 1



Graphical representation of the per-base sequence quality from FastQC for each of the 8 pools of samples. The red line and blue line represent the median and mean quality values respectively, with the yellow box showing the interquartile range and the whiskers showing the 10% and 90% values. The background colours represent different quality scores, with green indicating very good quality, orange representing reasonable quality and red showing poor quality.

Figure 3.2 – FastQC graphs for NGS project 2





Graphical representation of the per-base sequence quality from FastQC for the forward and reverse reads of each of the 8 sample pools. The red line and blue line represent the median and mean quality values respectively, with the yellow box showing the interquartile range and the whiskers showing the 10% and 90% values. The background colours represent different quality scores, with green indicating very good quality, orange representing reasonable quality and red showing poor quality.

3.3. Discussion of NGS and FastQC data

The difference between the number of reads generated by the first and second sequencing projects (~380.6 million vs. ~2,170.6 million) is a direct reflection of the advancements in sequencing technology in the intervening time (around 9 months). The improvements in chemical and engineering technologies utilised by the later Illumina HiSeq model resulted in the generation of around 5.7x more reads from the latter NGS project, and those reads of a longer read length.

The first parameter assessed by FastQC is the per base sequence quality. This gives an overview of the range of quality scores across all positions of a read.

Graphical representations of these are shown in Figures 3.1 and 3.2. From the summary information given in Tables 3.3 and 3.4 it can be seen that samples 4 and 5 from NGS project 1 and all of the samples from NGS project 2 failed on this parameter. A fail is given if the lower quartile for any position is below 5, or the median for any position is below a score of 20. For samples 4 and 5 from project 1, it can be seen in Figure 3.1 that the quartile does indeed dip below a score of 5, hence why a warning has been provided. However, the median score for these reads remains reasonably high, indeed, comparable to the median for the other samples which have not failed this measure. Taken together, this suggests that there are a greater number of poor quality reads within these sample pools, however, the median quality is still high, and particularly poor reads will simply not be aligned, so should not affect the quality of data in further processing.

The second project featured a much longer read length than did the first (100bp rather than 38bp). It is expected that per base quality will decrease across the length of a read, which is more markedly apparent when using longer read lengths. The "fails" have been given because the lower quartiles fall below a quality score of 5. However, for the most part, the mean read qualities remain in the green section (indicative of good quality) until around the 80base mark, while the median remains in the green region for the full length of the read. Again, this shows that for the majority of the reads, the base quality is good across the full read, and the fail is more an indicator of a wider spread of qualities. As stated before, any particularly poor quality reads will simply fail to be aligned, so will not have an impact on data quality at further processing stages.

The next parameter the program reports is the per sequence quality score, which identifies any subsets of reads that have unusually low quality scores. No warnings or failures were given for this parameter for any of the data files. Next is the per base sequence content, which all of the samples from the first project passed, and all of the files from the second project were given warnings for. This shows the proportion of each base at each position across the read, with strong differences potentially indicative of biases in the data. The score a sample is given for this can be biased by the sequencing primers that are added to each read, which are GC rich, and therefore the warnings given for the second sequencing project are not a cause for concern.

The per base GC content shows the GC content for each position across the length of a read, which ideally should be consistent throughout. This was the case for our data, so no warnings or failures were issued for any of the samples. The per sequence GC content score compares the GC content across the sequence with a modelled normal distribution of GC content. All sample files for both sequencing projects received a warning for this parameter. Although unusually shaped distributions can be indicative of contamination of a sample, it is likely that in this case the deviation arose from the inclusion of the GC rich primer sequences, as mentioned above. The per base N content

shows the number of "N" calls at each position in the read, and should not exceed a few percent, which was the case for all samples from both projects. The sequence length distribution gives a score based on the consistency of read lengths within a file. For each of the projects, these were exactly as expected (38bp for project 1, 100bp for project 2). The sequence duplication levels for the first sequencing project all passed, but all files containing the second project's data failed this parameter. This is a measure of the degree of duplication for each of the sequences in the set, low levels of which can indicate good coverage. However, when enriching for a particular region of the genome, such has been done for these projects, and multiple individuals are included in each sample pool, a reasonable level of sequence duplication is expected. Due to the vast increase in the number of reads in the second project over the first, it would be expected that the duplication levels would be higher. It may also be indicative that the enrichment process worked better for the second than first sequencing project. All of the second project's samples passed the over-represented sequences tests, but the first project's all received warnings. This test lists all sequences that make up more than 0.1% of the total, and can be indicative of contamination. However, the only sequence reported as featuring in this many reads was a string of "N"s, so actually indicates a quality issue with these reads. All this means is the base qualities in >0.1% of reads were not high enough to definitively indentify which bases were present, and were therefore called as "N"s. However, as discussed above, poor quality reads will not be aligned, and so will not affect downstream processing.

Although sample files from both projects received warnings or failures for various parameters, these were largely for expected or justifiable reasons. FastQC is not designed to give an exact representation of the quality of the data, but rather a comparison to a standard NGS dataset, which ours, with its target enrichment and sample pooling is not.

3.4. Defining the pipeline

A number of different programs were utilised in processing the data from NGS project 1 to allow an assessment of the best available processing pipeline, given that there was no "gold standard" for handling pooled sequencing data. The optimal set of programs to meet our requirements was determined based on the results and analyses presented below.

Alignment

Both BFAST and MOSAIK took around six weeks to align the 8 sets of NGS data to the reference human genome, requiring approximately equivalent processing power. After the alignment, the files were split in to regions corresponding to the four loci of interest (*CLU, CR1, PICALM* and the rs3851179 LD block). Testing of programs presented from here on mainly utilises the *CLU* data, given that this was the smallest of the regions sequenced

and would therefore allow a quicker assessment of programs, with the best being applied to the other regions in due course.

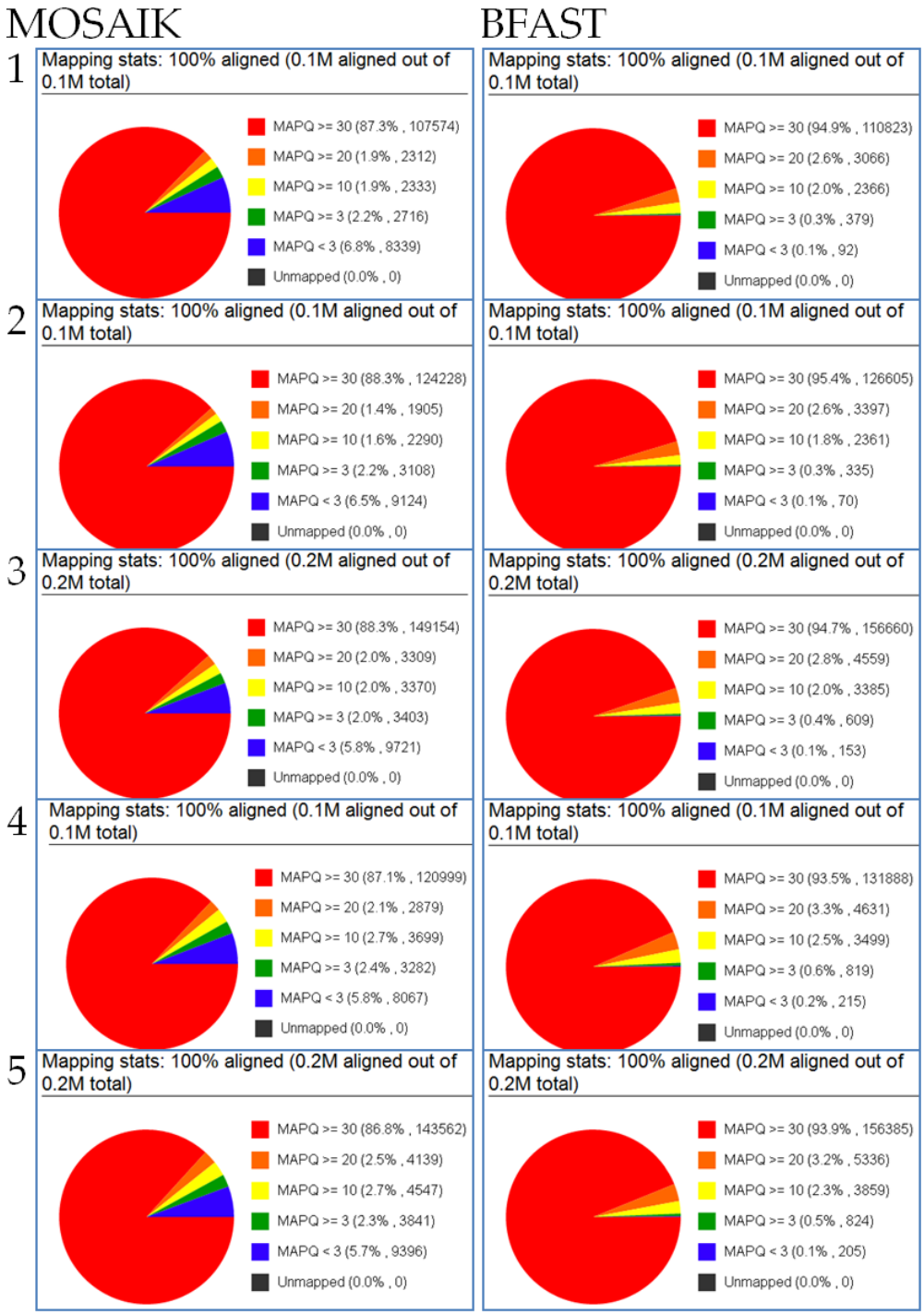
SamStat was used to allow a comparison of the quality of the alignments produced by BFAST and MOSAIK. The aligned data pertaining to the four targeted loci for each of the 8 aligned samples were analysed using SamStat. Graphical information on the quality of the alignments produced by SamStat for the *CLU* region is provided in Figure 3.3, and is representative of the data as a whole. Tables 3.5 and 3.6 show the data from SamStat's output for each of the four regions targeted in project 1 for MOSAIK and BFAST respectively.

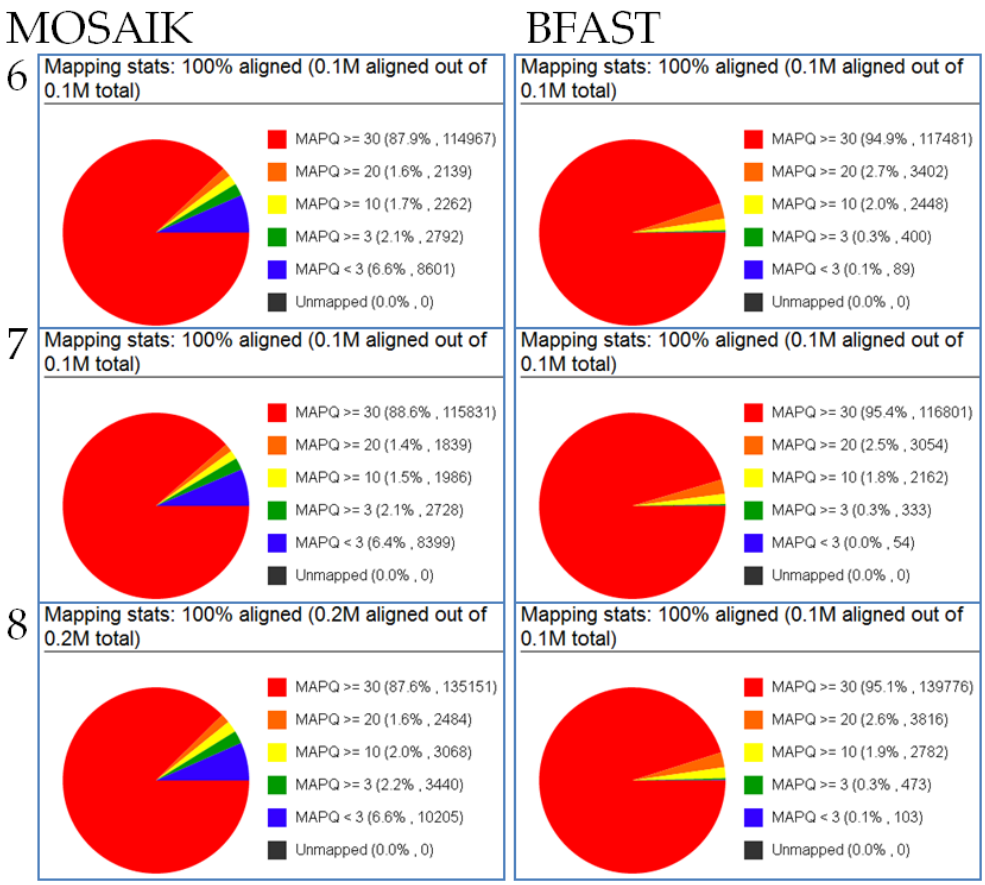
From the pie charts shown in Figure 3.3, it is obvious that BFAST consistently has a higher proportion of reads achieving the greatest quality score, and a lower proportion of reads attaining the lower quality scores. Whilst this immediately suggests BFAST is producing better alignment, the data contained within Tables 3.5 and 3.6 gives a more complete picture of the situation.

MOSAIK actually aligned more of the data from the files in to the regions of interest (33.4 million reads vs. 14.4 million reads). The BFAST alignment consistently gave a higher percentage of reads achieving the maximum mapping quality score (>=30) than did MOSAIK. In terms of the actual number of reads, the *CLU*, *PICALM* and rs3851179 LD block all had a greater number of reads reaching the highest quality score in the BFAST alignment, although the *CR1* region had a greater number of reads achieving quality >=30 in the MOSAIK rather than BFAST alignment. The BFAST alignment had a lower percentage and number of reads achieving the lower quality scores (3 or lower) in all four regions of interest than MOSAIK.

One thing that is notable from the data within the tables is that the alignment of reads to the *CR1* region was poorer with both alignment programs than for the other regions. While the other regions for each alignment had 84-89% (MOSAIK) and 94-95% (BFAST) of reads achieving the highest mapping quality score, for *CR1* this figure was only 58.5% and 67.6% from the MOSAIK and BFAST alignments respectively.

Figure 3.3 – Quality of alignment: BFAST vs. MOSAIK





Graphs from SamStat showing pie charts of data mapping statistics for MOSAIK on the left and BFAST on the right.

Table 3.5 - SamStat mapping quality assessment for MOSAIK alignment of project 1 data

	Total reads (all samples)	Average reads per sample pool	Average reads MAPQ>=30	Average reads MAPQ>=20	Average reads MAPQ>=10	Average reads MAPQ>=3	Average reads MAPQ<3
CLU	1153189	144148.6	126433.3 (87.7%)	2625.8 (1.8%)	2944.4 (2.0%)	3163.8 (2.2%)	8981.5 (6.23%)
PICALM	6569781	821222.6	731134.5 (89.0%)	15880.9 (1.9%)	18748.6 (2.3%)	16612.9 (2.0%)	38845.8 (4.7%)
rs3851179 LD block	1732734	216591.8	182499.9 (84.3%)	3602.4 (1.7%)	6461.9 (3.0%)	5864.4 (2.8%)	18063.3 (8.3%)
CR1	23925338	2990667.3	1750134.3 (58.5%)	199074.1 (6.7%)	359176.5 (12.0%)	224171.1 (7.5%)	458111.3 (15.3%)
Combined	33381042						

Information from SamStat assessing the quality of the alignment by MOSAIK. Total reads gives all the reads aligned to the region by the program for all sample pools. The other columns provide a breakdown on the mapping qualities of these reads.

Table 3.6 - SamStat mapping quality assessment for BFAST alignment of project 1 data

	Total reads (all samples)	Average reads per sample pool	Average reads MAPQ>=30	Average reads MAPQ>=20	Average reads MAPQ>=10	Average reads MAPQ>=3	Average reads MAPQ<3
CLU	1115695	139461.9	132052.4 (94.7%)	3788.4 (2.7%)	2552 (2.0%)	521.5 (0.4%)	122.625 (0.1%)
PICALM	6437215	804651.9	753737.5 (93.7%)	29187.3 (3.6%)	18143 (2.2%)	2821.3 (0.4%)	762.5 (0.1%)
rs3851179 LD block	1597505	199688.1	188102.4 (94.2%)	6569.8 (3.3%)	4159.5 (2.1%)	691.8 (0.3%)	164.8 (0.1%)
CR1	5252304	656538.0	443405.5 (67.6%)	93126.3 (14.2%)	111849.3 (17.0%)	6989.9 (1.0%)	1167.1 (0.2%)
Combined	14402719						

Information from SamStat assessing the quality of the alignment by BFAST. Total reads gives all the reads aligned to the region by the program for all sample pools. The other columns provide a breakdown on the mapping qualities of these reads.

Variant calling

Two different variant calling platforms were utilised in the processing of the data from project 1 - Syzygy and CRISP. Again, *CLU*, the smallest of the genes targeted, was used as an example to determine which of these programs was most appropriate for our data.

In the data, Syzygy identified a total of 73 variants within the BFAST aligned data, and 224 in the MOSAIK aligned data. CRISP identified 100 variants in the BFAST alignment, and 131 in the MOSAIK alignment.

One way to compare the performance of the different variant calling algorithms is to look at the false negative and false positive rates in the variants identified. To look at the false negative rate, a list of known SNPs is needed. Although we cannot know which variants will be present within the samples sequenced, variants which are common within a similar population would be expected to be present, and so can be used to assess the false negative rate. A total of 24 variants with a frequency above 1% in the CEU population were contained within the dbSNP database at the time of analysis (22.09.11). The number of these SNPs which were identified in the aligned data for the *CLU* region using Syzygy and CRISP is presented in Table 3.7. Assessing the false positive rate is more problematic, as it is difficult to distinguish spurious from genuine variants without validation via an independent method. Table 3.7 also contains information on the number of novel variants (SNPs and indels) identified by each of the sequencing programs, which are necessarily genuine variants or false positive calls.

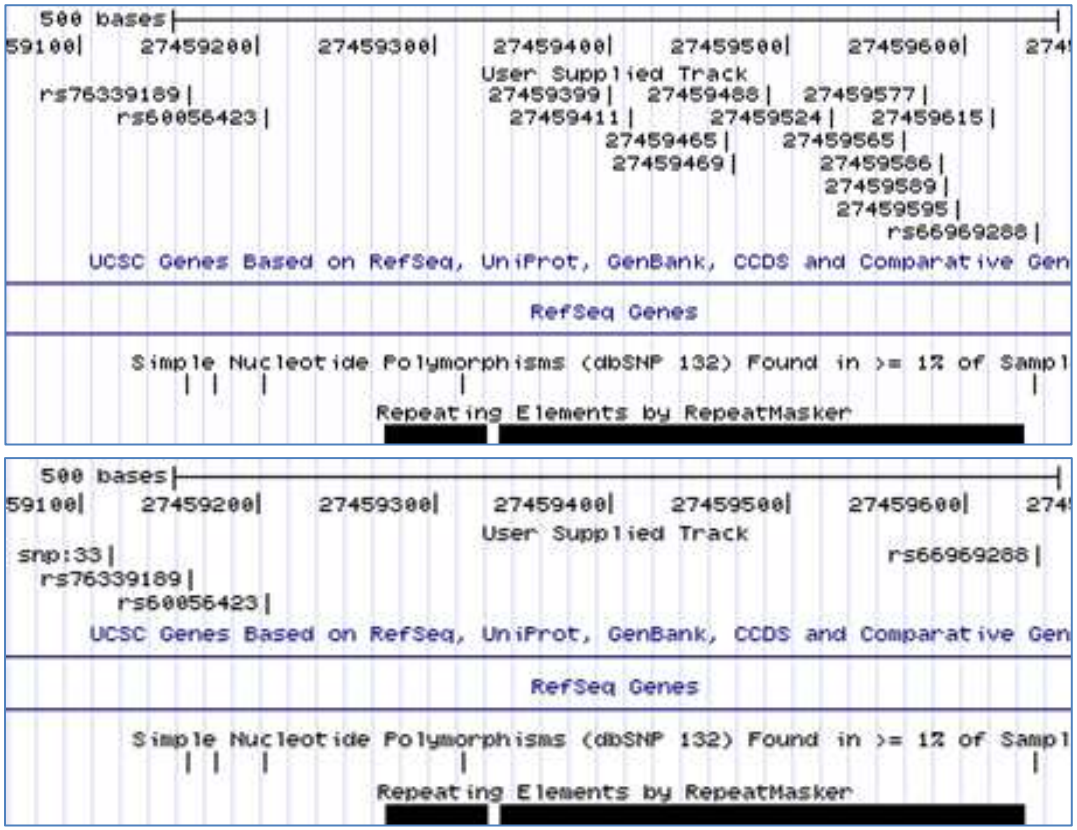
Table 3.7 – Assessing false negative and positive variant call rates

		No. of dbSNPs with MAF >1% detected	No. of novel SNPs	No. of novel indels
MOSAIK	Syzygy	20	142	23
	CRISP	21	20	44
BFAST	Syzygy	19	9	9
	CRISP	21	17	14

Information on the number of false negative variant calls, and the numbers of novel variants detected by each of the variant calling algorithms. NB – the version of CRISP used in this analysis was the latest available at the time of processing. Subsequently a new version has been used, so the numbers of SNPs quoted here do not match the ones in subsequent analyses, which were run using the most up to date version of CRISP in November 2012.

The locations of these variants were established by uploading the .bed file of the variants to the custom tracks section of the UCSC genome browser (Kent et al. 2002). Examples of the resultant images (MOSAIK aligned data, with variant calling by Syzygy and CRISP) are shown in Figure 3.4.

Figure 3.4 – Locations of variants called by Syzygy and CRISP



Images to show the locations of some of the variants identified by Syzygy and CRISP in the MOSAIK alignment using the UCSC genome browser's custom tracks (Kent et al. 2002). The two panels show the same genomic location, featuring a repetitive stretch of DNA, and the variants called within this region by Syzygy (top) and CRISP (below) are displayed under the heading "User Supplied Track". SNPs identified by an rs number are known variants which have been detected within our data. Any identified by chromosomal coordinate (Syzygy) or snp:# (CRISP) are novel.

3.5. Discussion of the Pipeline

The world of NGS has progressed rapidly in the past decade, and the advancements in technology have been accompanied by an explosion in the software available to process the generated data. While GATK has been widely accepted as the "gold standard" set of programs for processing NGS data for single individuals, it does not have the capacity to handle pooled data. It was therefore decided to try a number of different combinations of programs and establish what the best, most reliable combination of these was. Two different alignment programs and two different variant calling programs were utilised, and the best combination of programs was selected, as explained below.

The first stage of processing NGS data is to align the data to a reference genome. BFAST and MOSAIK were the two programs selected for this stage. Each was comparable in terms of the processing power required and the time taken to perform the alignment, so this did not contribute to the decision between them.

The number of dbSNP variants with MAFs >1% in the CEU population identified by CRISP was unaltered by the alignment program used, while Syzygy detected one less variant within the BFAST data than the MOSAIK data. In terms of the false negative rate, it was apparent from the data in Table 3.7 that CRISP was able to identify a slightly greater number of the variants than Syzygy was. The maximum number of the 24 dbSNP variants detected was 21. It can be assumed that these are genuinely present within the samples sequenced; therefore anything below 21 is missing variants known to be present. Although the difference between variants identified by CRISP and Syzygy was small in the *CLU* region data, the same pattern was found in the data from the other three regions sequenced, with Syzygy consistently identifying fewer variants than CRISP. From this, it can be concluded that CRISP has a lower false negative rate than Syzygy.

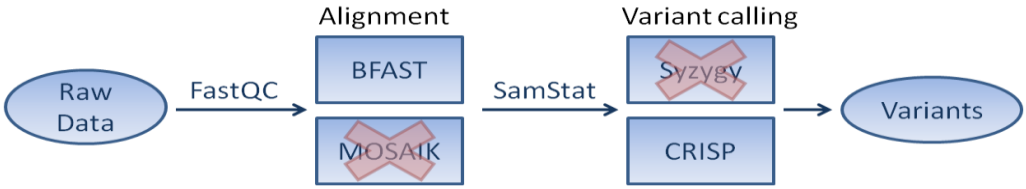
It is more difficult to assess the number of false positives found, since without comprehensive independent validation it cannot be known which variants are genuine and which are not. Although it is unknown how many genuine novel variants lay within the area, the vastly different numbers of variants called by the different combinations of alignment and variant calling algorithms highlighted that there were significant differences in performance, since only one (unknown) value represents the actual number of variants. The images presented in Figure 3.4 help explain these discrepancies.

It can be seen from the top panel of Figure 3.4 that the combination of programs giving the highest number of novel variants (MOSAIK and Syzygy) is calling a vast excess of novel variants around the sites of repetitive regions within the gene when compared with CRISP (one example of many instances shown). The majority of these variants are likely to be false positive calls as a

result of difficulties aligning data around repetitive DNA (Treangen and Salzberg 2012). Although a much more modest number of putative SNPs were identified using the CRISP and MOSAIK combination, 44 novel indels were called. Such a high number within a small region is likely to be indicative of issues with alignment, rather than there being that number of genuine insertions and deletions present.

It was therefore decided that BFAST and CRISP would be the programs of choice for both this NGS project, and for others going forward in the lab. The pipeline is displayed in Figure 3.5.

Figure 3.5 – Pipeline for alignment and variant calling in NGS data



Pipeline for processing NGS data as decided by a comparison of results from the different combinations of programs used.

3.6. Applying the pipeline

Once it had been decided which combination of programs would give the most reliable variant calls from the data we had, this was applied to the four loci sequenced in project 1, as well as the *CR1* data from the second NGS project.

Quality, coverage and enrichment

The quality scores given by SamStat for the BFAST alignment of data from NGS project 1 were presented in Table 3.6. The same SamStat analysis was run when the *CR1* locus data from the second NGS project was received. Table 3.8 provides the mapping quality data for the *CR1* locus from SamStat for both the first and second NGS projects.

The data from SamStat enabled the calculation of the average depth of coverage per individual for each of the regions targeted. This information is presented in Table 3.9.

However, although the average depth is reported here, the consistency of depth across the regions sequenced was by no means uniform. As mentioned previously, the repeat masker offered by Agilent was not utilised in the design of the first sequencing project but was in the second. However, in both sequencing projects, repetitive areas received virtually no coverage, even when directly targeted. Figure 3.6 shows images from the Integrated Genomics Viewer (IGV) displaying dropout in coverage around repetitive regions of DNA. This issue was particularly severe for CR1, which has a highly repetitive structure.

Table 3.8 - SamStat mapping quality assessment for BFAST alignment of *CR1* locus from the two NGS projects

	Total reads (all samples)	Average reads per sample pool	Average reads MAPQ>=30	Average reads MAPQ>=20	Average reads MAPQ>=10	Average reads MAPQ>=3	Average reads MAPQ<3
Project 1	5252304	656538	443405.5 (67.6%)	93126.3 (14.2%)	111849.3 (17.0%)	6989.9 (1.0%)	1167.1 (0.2%)
Project 2	148300000	18537500	7299770.7 (39.4%)	2702092.9 (14.6%)	3789404.8 (20.7%)	4124882.3 (22.3%)	604474.6 (3.3%)

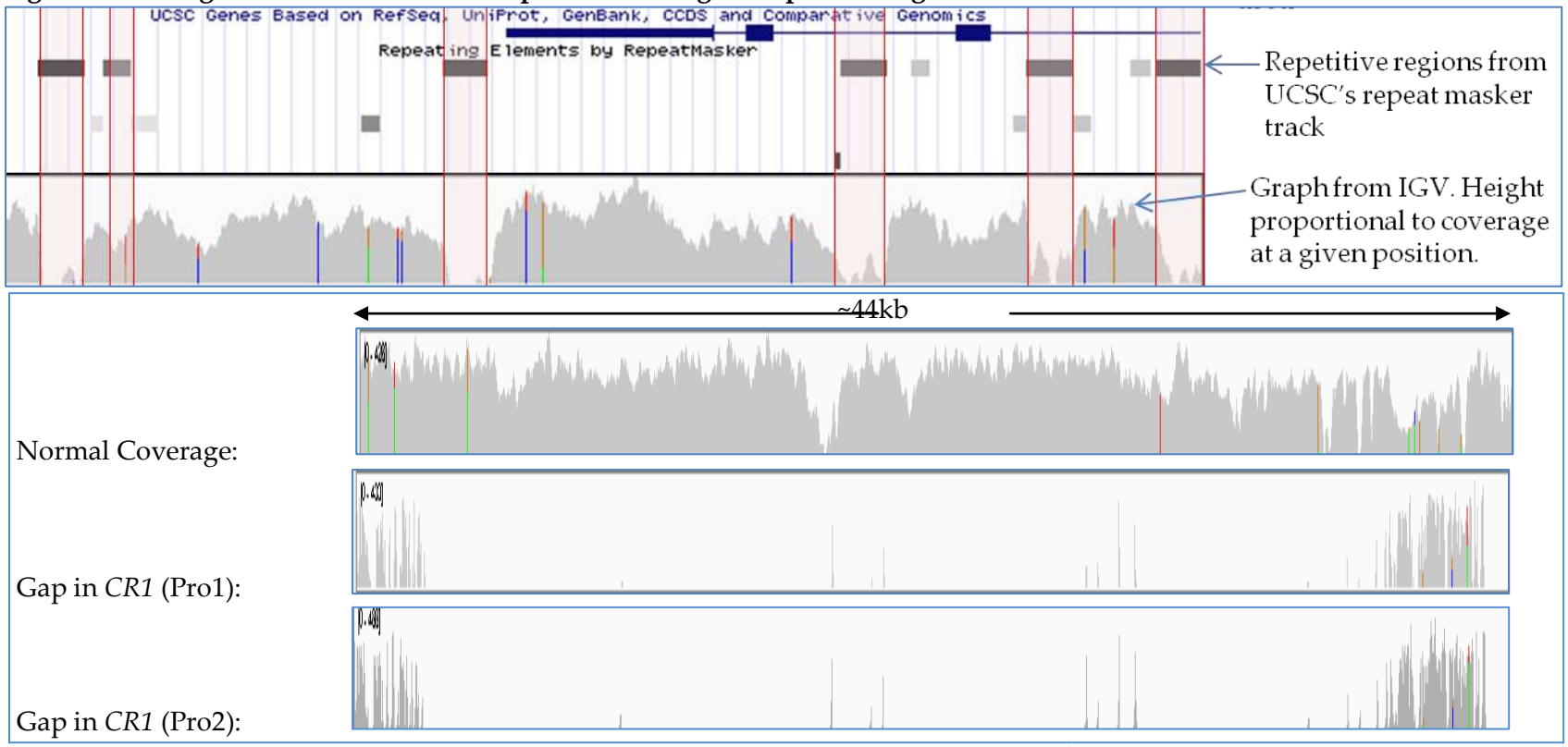
SamStat mapping quality score breakdown for the *CR1* region from the first and second NGS projects. Total reads gives all the reads aligned to the region by BFAST for all sample pools. The other columns provide a breakdown on the mapping qualities of these reads.

Table 3.9 – Average depth of coverage per individual

	CLU	*PICALM*	rs3851179 LD block	*CR1* (project 1)	*CR1* (project 2)
Coverage	18.1x	21.5x	21.7x	13.9x	1034.3x

Average depth of coverage per individual across the entire targeted loci.

Figure 3.6 – Images from IGV to show dropout of coverage at repetitive regions



Images to show the dropout in coverage at repetitive regions. Top panel: representative section of *CLU* locus. Repetitive regions clearly coincide with reduced coverage in the IGV graph. Red boxes highlight these regions. Lower panel: three IGV graphs, the top showing a normal region of genomic DNA, with good coverage in general, but some areas of lower coverage where repetitive regions fall. The lower two show a region of *CR1* coinciding with the tandem repeat responsible for the different isoforms of *CR1* where coverage was particularly poor in each of the sequencing projects.

The information on reads from SamStat allowed for the analysis of the efficiency of the enrichment process. Firstly, the enrichment factor could be calculated. This is a comparison of the depth of coverage at the targeted regions versus the depth of coverage across the genome as a whole. For the first sequencing project, the enrichment factor was 321.9x, while for the second sequencing project, it was 6271.5x (data to calculate this provided by colleague James Turton: total reads 1949915034; reads mapping to regions of interest 1209121676; combined size of regions of interest 832825bp). The percentage of reads mapping to the target region also signals how well the enrichment stage of the experiment worked. From the first sequencing project, 3.9% of the total reads produced mapped to the targeted regions. For the second project, this figure was 62.0%.

Variants identified

The number and types of variants detected by CRISP in each of the regions sequenced are presented in Figure 3.7, divided in to exonic and non-exonic variants, and containing details on the number of novel vs. known variants encountered. Excluding the 3'UTR SNPs, only 10% of the detected exonic variants were novel, while this figure was 23.8% for the non-coding ones.

The numbers of HapMap SNPs with frequencies >5% identified in the regions is presented in Table 3.10.

Table 3.10 – HapMap SNPs detected in NGS data

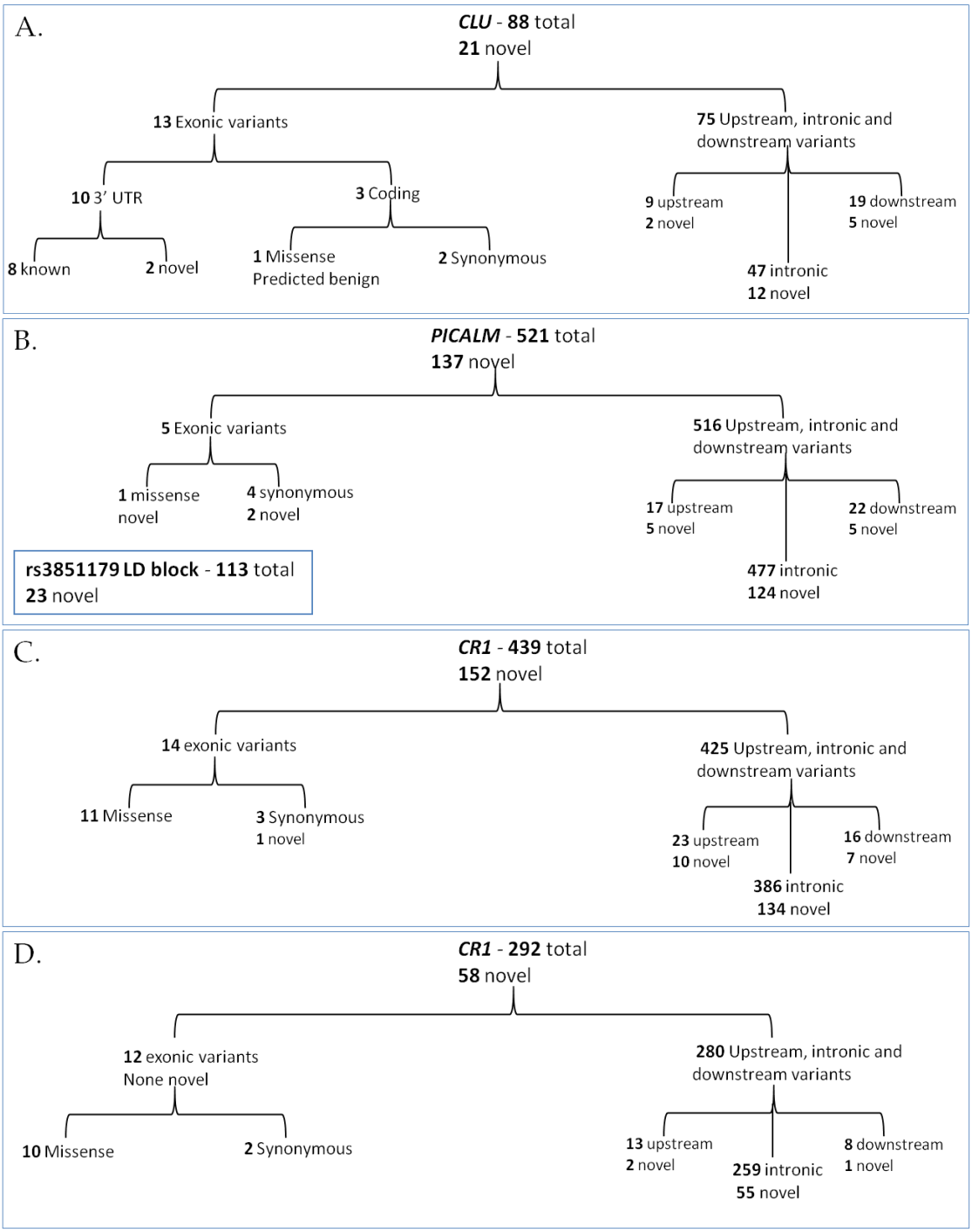
	CLU	*PICALM*	rs3851179 LD block	*CR1* project 1	*CR1* project 2
No. of HapMap SNPs in region	20	117	34	75	75
No. detected in NGS data	20	117	34	70	67

Numbers of HapMap SNPs with frequencies >5% in the EUR population within the targeted loci, and how many of these were detected within the NGS data for those regions.

Ts/Tv Ratios

The Ts/Tv ratio was calculated separately for exonic and non-exonic regions. Across the four sequenced loci, there were 27 exonic transitions and 5 exonic transversions, while there were 555 non-exonic transitions and 352 non-exonic transversions. This gives a Ti/Tv ratio of 5.4 for exonic areas and 1.58 for non-exonic ones.

Figure 3.7 – Variants identified at each locus sequenced



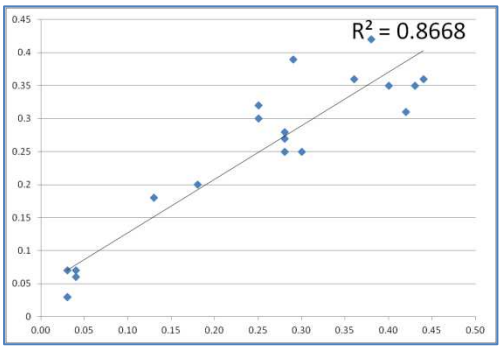
Breakdown of the variants detected in the NGS data after the application of the data analysis pipeline. Panel A shows the variants in *CLU*, panel B shows the breakdown of variants at the *PICALM* locus, as well as the rs3851179 LD block (all variants non-coding, as the region is ~88.5kb upstream of the gene). Panel C shows the results for the *CR1* region from the first sequencing project, while panel D shows the variants detected in the data from the second NGS project. Any variants described as "novel" were found to have no co-located variants by Ensembl's VEP.

HapMap Correlation

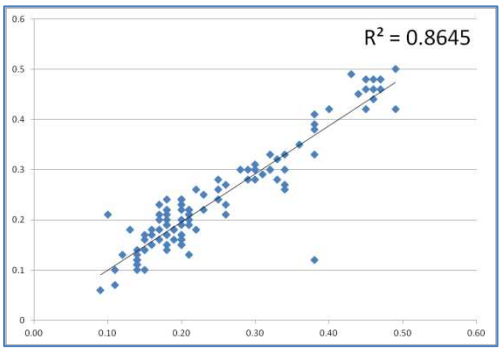
As mentioned previously, a good way to assess the reliability of NGS data is to compare it to known variants. In order to assess the accuracy of the frequency estimates from CRISP, MAF estimates (based on percentage of alternative reads – a surrogate for MAF) from CRISP for the common SNPs called in the three genes were compared with frequencies from HapMap (EUR population). The relationships between these frequencies for each of the targeted areas are shown in Figure 3.8.

Figure 3.8 – Correlation between CRISP frequency estimates and HapMap frequency data

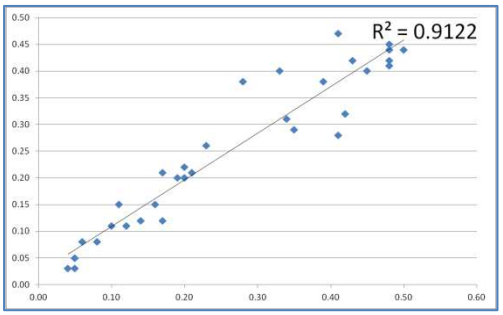
CLU



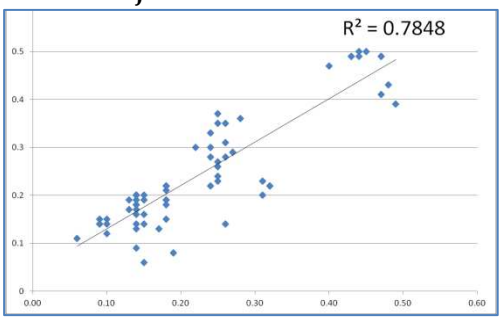
PICALM



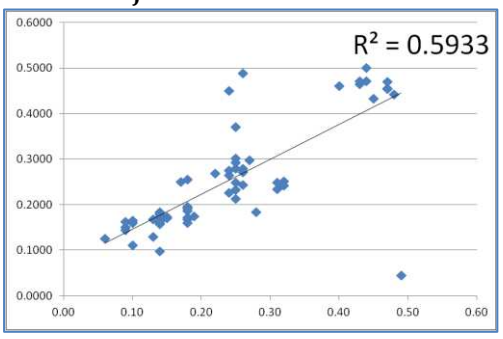
rs3851179 LD block



***CR1* – Project 1**



***CR1* – Project 2**



Comparison of frequency estimates from CRISP with MAF data from the HapMap project (EUR population) for common SNPs within targeted loci. On each graph, the x axes show the variant's frequencies in CRISP data, while the y axes show the HapMap CEU allele frequencies.

MAFs from the 1000 genomes project (European population) were also compared to the frequency estimates from CRISP for all of the detected variants in the targeted loci (data from NGS project 2 used for *CR1*). Due to the high number of variants this entails, the data is not shown here, but upon analysis in SPSS, a strong, significant positive correlation was observed between the frequencies of the two datasets (Spearman Correlation Coefficient=0.964, p=<0.001).

3.7. Discussion of applied pipeline

The major aim of this stage of the study was to identify variants, an aim which was clearly met with the detection of over 1000 variants in the four targeted loci, 239 of which were potentially novel variants. The inclusion of the *CR1* target locus in two separate sequencing project allowed an assessment of the advances in technology which had been achieved in just a small period (approximately 10 months between the running of the two projects).

One of the major differences in the *CR1* data from NGS projects 1 and 2 is in the sheer quantity of data generated. The total number of reads generated by the first project was ~380 million, while the second project generated ~1209 million reads. When considering the aligned data, the 5.25 million 38bp reads mapped to the *CR1* region from the first sequencing project are eclipsed by the 148.3 million 100bp reads mapped to the equivalent region in the second project. The *CR1* data from the first project had a greater proportion of reads awarded the highest quality scores than did the second project, but because the second generated such an increased quantity of data, far more reads in total from the second project achieved the highest mapping quality (>=30).

From the first sequencing project, *CLU*, *PICALM* and the rs3851179 LD block received a reasonable level of coverage. Generally, between 20 and 30x per individual is considered to be adequate for reliable variant identification. *PICALM* and the rs3851179 LD block both exceeded this lower recommended limit of 20x per individual, while *CLU* was close to this (18.1x). *CR1* however fell short of this target, receiving an average of just 13.9x per individual. Because of this particularly poor coverage, it was decided that *CR1* would be included in the second NGS project being conducted by a colleague in the lab. It was hoped that the issues of the first project would not be repeated in the second (this time the repeat masker was utilised, the enrichment process was conducted by Agilent themselves, and a newer sequencing technology with far superior capacity was utilised, with a longer read length and paired-end reads). This was indeed shown to be the case, with *CR1* receiving an average coverage of 1034.3x per individual in the second sequencing project.

However, the second sequencing project did not successfully address the other coverage issue with *CR1*. This issue stems from the structure of the gene, and its highly repetitive nature. As the top panel of Figure 3.6 shows, even in

the first project where repetitive regions were included in the design of the baits, repetitive areas received virtually no coverage. The inclusion of such regions in the experimental design is likely to be in part responsible for the poorer than expected coverage of the targeted areas obtained, and the fact that only 3.9% of reads mapped to the target regions. This figure is far less than Agilent's estimated on target return of 60% (from the SureSelect product information) or 40-50% reported by Gnirke et al. (Gnirke et al. 2009), the publication of the SureSelect technique. Another study published in 2010 explored the issue of the repeat masker. As with our explorations, this group also found Agilent's repeat masking strategy to be overly conservative, so opted for their own custom repeat masking approach, and reported around 20% of reads mapping to target regions (Kenny et al. 2010).

The inclusion of repetitive sequences in the bait design essentially wastes sequencing capacity. A run on an NGS machine will generate a specific, finite quantity of data. Whatever proportion of the reads are generated from repetitive regions of the genome is essentially a proportion of the reads that are lost. Reads with multiple matches across the genome during the alignment process cannot be successfully mapped, and will therefore be lost, reducing the sequencing capacity available for reads which can be mapped. Additionally, allowing these repetitive regions to be included in the bait design will have reduced the specificity of the capture, as non-target, similar DNA will also have been pulled down, reducing the recovery of the true target region, and limiting the number of reads which could be uniquely mapped.

The utilisation of the repeat masker in the second sequencing project meant that a minimal amount of reads were lost in this way, which is reflected in the much more respectable percentage of reads on target of 62%.

However, there was still a major issue with the coverage of *CR1*, which is shown in the lower panel of Figure 3.6. The IGV graph at the top of this panel shows what a normal region of the genome would be expected to look like when sequenced in this way, with a high level of coverage across the region in general, along with a few areas of lower coverage, likely to be due to repetitive sequence. The lower two IGV graphs show a region of *CR1*, spanning a region of around 44kb from each of the two NGS projects. Here there was virtually no coverage, with very few reads at all mapping to the area. This problematic region stems from the structure of the CR1 gene, which, believed to have arisen through a sequence of segmental duplication events over time, is highly repetitive. The *CR1* protein isoforms differ in the number of C3b binding sites present, and the different alleles are thought to have arisen from unequal crossover events involving a stretch of DNA encoding this, making the region problematic for sequencing.

This problem is only exacerbated by the utilisation of the pooled sequencing strategy. Given the frequencies of the F and S alleles in European populations

(0.83 and 0.15 respectively), it is highly likely that each pool of 12 individuals (and thus 24 alleles) contains both the F and S alleles of the gene. Across the experiment as a whole (96 individuals, or 192 alleles), it is likely that the rarer genotypes will also feature. There is currently no way to disentangle which individuals have which genotypes from the pooled data based on the NGS data alone, and no way to tell which alleles are present in the sample pools. A recent study which used multiplex amplicon quantification to distinguish F- and S- alleles found an association between the S-allele (with an extra C3b binding site) and increased AD risk (Brouwers et al. 2012), however, with our methodology it was not possible to determine genotypes for this polymorphism within sample pools, let alone within individual samples.

When repetitive DNA comprises such a large proportion (~50% (Treangen and Salzberg 2012)) of the human genome, these regions cannot simply be ignored, but nor can they be accurately sequenced, given the current technology and data analysis methodologies available. The CR1 example above demonstrates that repetitive DNA can be biologically important and disease relevant, but whether this is a common phenomenon or an isolated example remains to be seen.

Aside from this region, the performance of the second project was significantly greater than the first, with a much greater volume of data generated; well over half of that data mapping to the regions of interest; a robust enrichment factor (almost 20x greater than that of the first project); and a high level of coverage across the targeted loci (with the exception of the *CR1* repeat region). It is likely that the reasons for these improvements are three fold. Firstly, the huge increase in data produced is a reflection of the vast advances in NGS in just a small amount of time (less than one year). Better technology and improvement in the chemistry of the sequencing reactions in a short space of time lead to a leap from a single sequencing run producing ~380 million 38bp single-end reads to ~1209 million 100bp paired-end reads. Secondly, the enrichment performed by Source for the first sequencing project did not seem to have worked as effectively as the one performed by Agilent for the second sequencing project. Finally, the repeat masker not being utilised in the design of the first project almost certainly had a negative effect on the specificity of the enrichment, and on the following data analysis.

Over 1000 variants were identified by CRISP in the four target loci. Given that a much smaller proportion of the targeted regions were coding than non-coding, it is unsurprising that the majority of variants identified by CRISP fell either up- or down-stream of the genes, or in intronic regions. Only 10% of the coding variants detected were determined to be novel by the VEP, whilst almost a quarter of the non-exonic ones were. There are likely to be a number of reasons for this. Historically, and even now, reflected in the popularity of whole exome sequencing, non-coding regions have been far less extensively investigated than coding ones. Coding regions are more likely to have been subjected to sequencing, and thus coding variants are more likely to already

be known. By the same logic, non-coding variants are less likely to be known, and perhaps also suffer from a reporting bias, with coding variants more likely to have been recorded and reported in the literature. However, it is also likely that non-coding regions contain genetic features which predispose to false-positive variant calls, such as indels and mononucleotide repeats, which will be discussed in further detail in Chapter 4 – Sanger Validation.

The Ts/Tv ratios calculated for the variants detected were 5.4 for exonic regions and 1.58 for non-exonic regions. The estimates from 1000 genomes data puts the genome average to be ~2.1, although exonic regions tend to be higher (Le and Durbin 2011). Our variants show a trend in the right direction, with exonic regions having a higher ratio than non-exonic regions, but the non-exonic rate is lower than expected, and the exonic rate higher than expected (although the latter was based on a small number of sites). Ts/Tv ratios deviating significantly from the expected can be an indicator of unreliable variant calls (e.g. high false-positive or false-negative rates), which should be kept in mind when assessing the reliability of our variants.

When considering the amount of known HapMap SNPs with frequencies greater than 5% in the European population, all variants were detected in the NGS data for *CLU*, *PICALM* and the rs3851179 LD block. For *CR1*, the first project missed five of 75 variants, while the second missed eight. Since the extra three of these variants were identified in the first project, it can be assumed they are present in the samples sequenced. Two of the missing three fell close together within a repetitive area which was not targeted by the second project due to the repeat masker. The third of these was in a targeted area, but was not identified. The numbers of exonic variants identified in *CR1* was very similar between the two projects. The major difference in variant calls between the first and second projects was in the number of non-coding variants, and particularly in the novel ones. In the first project's data, out of the 425 variants identified in non-coding regions, 151 were novel, meaning 274 known variants were also found. In the second project, out of the 280 non-coding variants, just 58 were novel, with 222 known. This gives a percentage of novel variants of 35.5% for project 1 and 20.7% for project 2.

The reduction in known variants found between the two projects stems mainly from the inclusion of the repeat masker in the second project – many of the variants found in project 1 but missed by project 2 were in repetitive regions, and therefore were not targeted in the second. This does provide an argument in favour of not using the repeat masker in the design of baits, since more variants were found when the repetitive areas were included, although the necessary trade off in reducing the specificity of the enrichment may not be worth the gain.

Less novel variants being detected in *CR1* by the second project is likely to signify a decrease in the number of false positive variant calls being made. The 35.5% of variants identified in the non-coding region for the first sequencing

attempt of *CR1* is much higher than for the other regions sequenced in that study (*CLU* – 25.3%, *PICALM* – 26.0%, rs3851179 LD block – 20.4%), while the figure of 20.7% of variants detected being novel from the second project is more consistent with these. The major difference in the design of the projects is in the inclusion of the repeat masker in the second project, which would be expected to decrease the novel and known SNPs identified approximately equally, so the percentage of novel variants identified should be approximately equal between the two projects. Since this is not the case, and the percentage of novel variants in the second project is much lower than that from the first, the most probable explanation for the pattern of data observed is that the second project gave less false positive calls.

Although information was lost, in that more known common dbSNP variants were missed in the calling of the second project's variants, and around 50 known variants were missed, the false positive rate being lower is reflective of the better quality enrichment, greater coverage, increased quantity of data and more robust sequencing technology. An average coverage of 13.9x (as from the first project) was bound to give false positive calls, as it is far below the recommended depth for accurate calls of 20-30x. It is arguably better to sacrifice the identification of a number of known variants, than to have a huge quantity of novel variants, if many of those are likely to be false positive calls.

It has previously been reported that frequency estimates from pooled NGS data can be generally regarded as reliable (Ingman and Gyllensten 2009; Bansal et al. 2011; Day-Williams et al. 2011). Frequency estimates from CRISP rely on certain assumptions being made. Largely, it is assumed that the number of reads is a reliable reflection of the frequency of variants being detected. This in turn is based on the assumption that each individual's DNA has equal representation within the pool. Whilst steps were taken to ensure that was the case (e.g. quantification using QuantIT, pooling samples of similar concentrations together, and maintaining a minimum pipetting volume of 1µl to minimise pipetting errors), it cannot be guaranteed. Unreliable quantification or inaccurate pool preparation could lead to some individual's DNA samples being over- or under-represented in the initial sample pool. Even given equality at this stage, there is the possibility that naturally occurring variation may affect the efficiency of the enrichment process. The baits designed to capture the regions of interest by Agilent are complementary to the standard reference genome. Although their large size (120mer) is designed to minimise the chance of this happening, it is possible that deviance from the standard reference genome could destabilise the coupling of the RNA baits with the desired DNA targets, creating an enriched pool of DNA biased towards the wild type. While single SNPs are unlikely to have this effect in such large baits, insertions, deletions, or the occurrence of multiple variants within a specific bait's target could cause issues, and result in unequal representation of samples within the pools. Another potential source of such bias is differential amplification at the PCR stage of the enrichment process, which could lead to certain templates being under- or

over-represented. Any deviance from equal representation could skew the frequency estimates from CRISP, which are based on the number of reads in possession of alternate allele calls. In a pool of 12 individuals, or 24 alleles, each allele should contribute to 4.2% of reads. If the DNA of an individual with a specific variant is not equally represented within the pool, there will be deviance from this figure.

Strong correlation scores were found between the frequencies of the common HapMap SNPs for *CLU, PICALM* and the rs3851179 LD block. The frequency correlation is stronger for the *CR1* project one data than for the second project's data. However, the second project is based on a smaller number of data points, since less common variants were found in the second project's data. It is also likely that any biases in the pooling or enrichment in the second project would have been amplified by the vast quantity of data produced, explaining why the frequency estimates may be less reliable. Since reliable estimation of frequency was not the aim of this project, the lower reliability is not an important factor. The data from the second project appeared more reliable for variant identification, which was the aim of the study.

The HapMap correlations are only based on a small number of variants, all of which can be classed as common. For a more complete picture, the frequencies of all known variants detected in the NGS data were compared with the frequency data from the 1000 genomes project. Here, a strong, positive, significant correlation was found, indicating that the CRISP frequency estimates can be regarded as reliable. A further assessment of the accuracy of CRISP frequency estimates is given in the discussion of Chapter 4 – Sanger validation.

It is also worth noting that while a good correlation in frequencies is indicative of reliable frequency estimates, deviance from correlation can be not only an indicator of unreliable frequency estimates, but of association with AD. Since the 1000 genomes project used population controls for their samples and the NGS projects used AD samples only, differences in allele frequencies between the two could be an indicator of involvement in AD risk. However, in just 96 samples, there is insufficient power to deduce any real information on this.

To summarise, this chapter has explored the process of analysing pooled NGS data. A variety of programs were utilised and their performance compared to enable the development of a robust pipeline for analysing such data. When the pipeline was applied, over 1000 variants were detected within the data. The quality of the raw data, the alignments and the variant calls were assessed and were generally found to be good, although coverage in the *CR1* region was extrememly uneven, leaving large areas of the gene effectively not sequenced.

4. Sanger validation

Given the current high error rates of NGS technologies and issues achieving accurate alignment, particularly around repetitive regions (Treangen and Salzberg 2012), validation of putative variants detected in NGS data via an independent method is important. With over 1000 variants detected, independent validation of all variants would be prohibitively time consuming and expensive. This chapter explores the Sanger sequencing of certain variants, which were selected for validation based on the possession of unusual characteristics.

4.1. Results of Sanger validation

Within the data, there were a number of variants identified which were classed as novel by Ensembl's VEP, but were identified within all or most of our sequencing pools, with MAF estimates ranging from 5% to 24%. It seemed highly unlikely that variants with such frequencies would not have been previously documented were they genuine SNPs, so they were highlighted as being potential false positive calls, and validation via Sanger sequencing was sought.

Seven such variants were identified (those who's inclusion in the project within Table 4.1 is classed as targeted). Since several other variants identified by CRISP fell within the region sequenced, these were also analysed, taking the total number of SNPs considered to 12. Information on all of these variants is provided in Table 4.1.

Table 4.1 - Information on SNPs for validation by Sanger sequencing

Gene	Chr	Coordinate	Alleles	Frequency Alternative Calls	Fold Coverage per individual	Location in gene*	rs number	Inclusion in project
CLU	8	27452179	G/T	0.07	23.69	3.5kb downstream		**Targeted**
	8	27452243	A/T	0.05	24.13	3.5kb downstream		Incidental
	8	27466924	C/A	0.11	17.89	Intron 2		**Targeted**
	8	27473743	T/A	0.19	18.11	1.5kb upstream		**Targeted**
PICALM	11	85668102	G/A	0.11	16.74	1.5kb downstream		**Targeted**
	11	85668163	G/A	0.27	19.11	1.5kb downstream	rs622110	Incidental
	11	85692077	C/T	0.05	22.90	Intron 18	rs139710547	**Targeted**
	11	85692181	A/C	0.63	18.46	Exon 18 (synonymous)	rs76719109	Incidental
	11	85774424	T/G	0.24	16.41	Intron 2		**Targeted**
	11	85774562	T/G	0.46	18.55	Intron 2	rs3016786	Incidental
CR1	1	207690803	T/C	0.07	19.57	Intron 4	rs144047769	**Targeted**
	1	207690871	G/C	0.19	18.40	Intron 4	rs10863358	Incidental

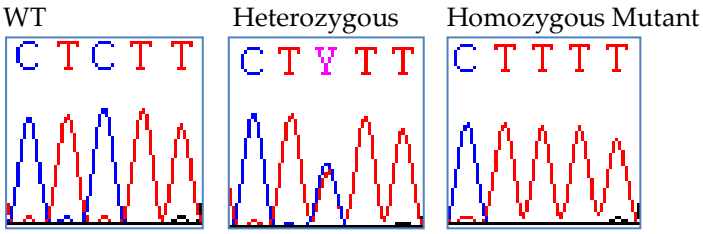
Information on all of the SNPs for which validation by Sanger sequencing was attempted. Coordinates stated give genomic position in hg19. *Relative to *CLU* transcript ENST00000316403, *PICALM* transcript ENST00000447890, *CR1* transcript ENST00000367049. Distances stated are approximate.

Of the 12 putative SNPs included in Sanger sequenced regions, six were found to be genuine (see Figure 4.1 and Table 4.2). Reliability of frequency estimates could also be assessed once the number of actual alternative alleles in a pool was established by Sanger sequencing. Each alternative allele should contribute ~4.2% of reads to the pool total, assuming equal representation. The relationship between the number of actual alternative alleles and the proportion of NGS reads they make up is shown in Table 4.2.

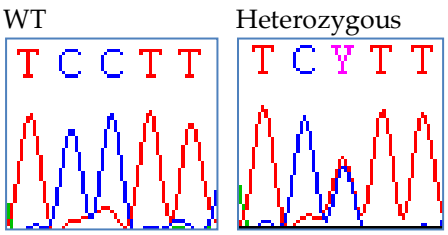
Figure 4.1 – Sanger validated SNPs

PICALM

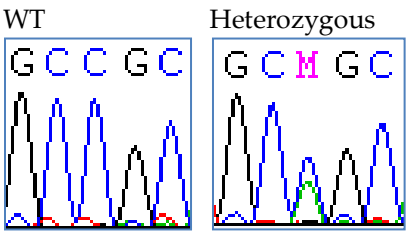
11:85668163 – C/T



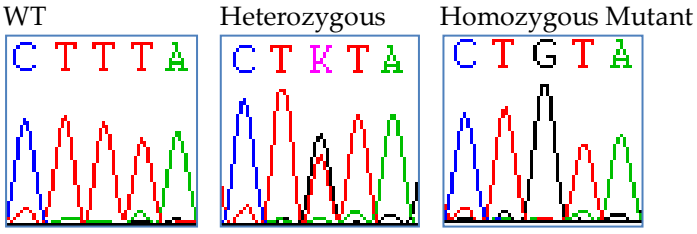
11:85692077 – C/T



11:85692181 – A/C

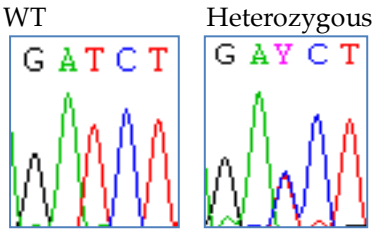


11:85774562 – T/G

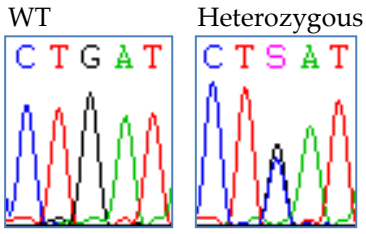


CR1

1:207690803 – T/C



1:207690871 – G/C



Images showing electropherogram traces from the successfully Sanger validated SNPs in *PICALM* and *CR1*. Where all three allelic combinations were found, all three are shown.

Table 4.2 – Successfully validated SNPs

Chr	Coordinate	Alleles	Alternative Allele Call Frequency (CRISP - all pools)	1000 genomes MAF (EUR)	Alternative Allele Call Frequency (CRISP - sequenced pool)	Alternative Alleles (Sanger - sequenced pool)	Alternative Allele Frequency (Sanger – sequenced pool)
11	85668163	G/A	0.268	0.244	0.707	10	0.417
11	85692077	C/T	0.046	0.016	0.103	3	0.125
11	85692181	A/C	0.632	0.583	0.929	19	0.792
11	85774562	T/G	0.460	0.422	0.362	10	0.417
1	207690803	T/C	0.067	0.021	0.333	3	0.125
1	207690871	G/C	0.194	0.214	0.050	4	0.167

SNPs which were successfully validated by Sanger sequencing. The number of alternative alleles within the pool sequenced facilitated the determination of the genuine MAF within that pool (assuming Sanger results reflect true allelic counts). This was then compared to the alternative allele call frequency from the same pool in CRISP, giving a reflection of the accuracy of the CRISP frequency estimates at a much finer level than the total 96 samples allows.

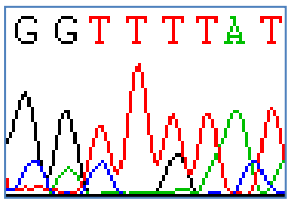
Three of the remaining SNPs were not found to be present in the samples Sanger sequenced, but instead small indels were found at the suggested variant sites (all which were also called by CRISP). These variants are summarised in Table 4.3, with examples from the Sanger sequencing shown in Figure 4.2.

Figure 4.2 – Sanger confirmed indels miscalled as SNPs

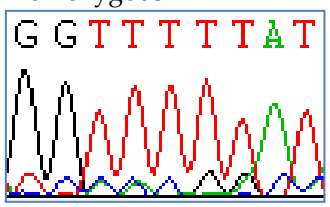
CLU

8:27452179 – Called as G/T SNP, actually +T indel

WT

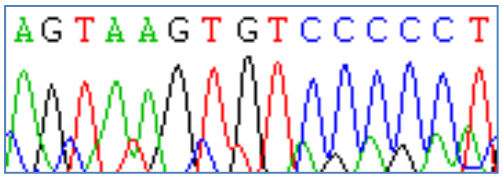


Homozygous +T

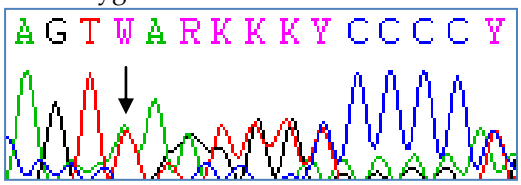


8:27452243 – Called as A/T SNP, actually +T indel

WT



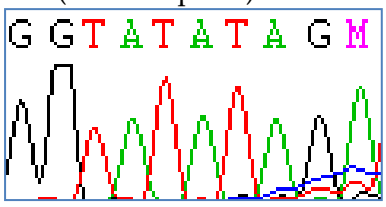
Heterozygous +T



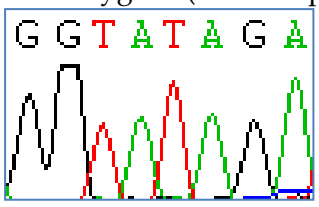
PICALM

11:85774424 – Called as T/G SNP, actually -TA deletion

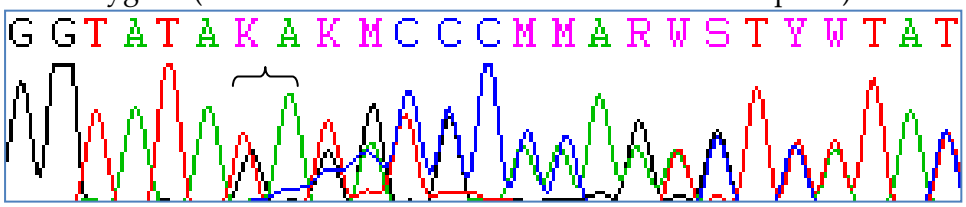
WT (3x TA repeats)



Homozygous (2x TA repeats)



Heterozygous (one allele with 2x TA and one with 3x TA repeats)



Electropherograms showing the absence of the SNPs predicted by CRISP, but featuring genuine indels at the same site. Where individuals heterozygous for the indels are shown, the sequence traces for each allele following the site of the indel will be offset by the number of bases the indel contains.

Table 4.3 – Sanger validated indels

Chr	Coordinate	CRISP called SNP	Actual Variant	Indel coordinate	Indel MAF (CRISP)	1000 genomes MAF (EUR)	rs number	Indel MAF (CRISP - sequenced pool)	Alternative Alleles (Sanger - sequenced pool)	Indel MAF (Sanger – sequenced pool)
8	27452179	G/T	+T ins	27452180	0.145	0.26	rs146954978	0.232	3	0.125
8	27452243	A/T	+T ins	27452242	0.099	0.17	rs35598594	0.115	1	0.042
11	85774424	T/G	-TA del	85774420	0.42	Not available	rs112671434	0.455	17	0.708

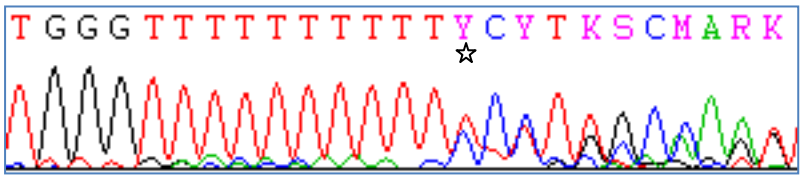
Details of the indels discovered at the sites of false positive SNP calls, all of which were also identified by CRISP and had been previously recorded in dbSNP.

The final three SNPs (8:27466924, 8:27473743 and 11:85668102) were not validated by Sanger Sequencing. Figure 4.3 shows examples of some of the sequencing traces from these variants (NB – reverse primer used in sequencing of 8:27466924 and 11:85668102, so sequence shown is the reverse complement to the standard genome oriented sequence). All three of these putative variants occurred adjacent to mononucleotide polyA repeats, with numerous other potential variants in the immediate area, called by CRISP or present in dbSNP 134 (see Figure 4.4). Within each of the polyA repeats, CRISP also called a +A insertion. There are also multiple +A insertions reported in dbSNP for each of the mononucleotide repeat sites, which suggests these may be genuinely variant. Due to the issues even Sanger sequencing has in dealing with mononucleotide repeats, our findings were inconclusive as to whether there were genuine variations in the number of A nucleotides present at these sites.

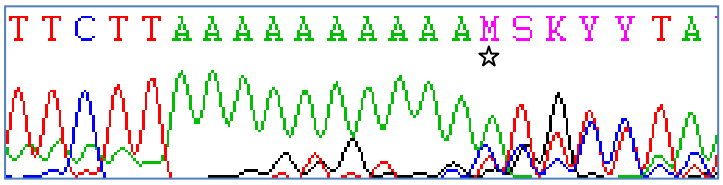
Figure 4.3 – Electropherograms showing mononucleotide repeat regions

CLU

8:27466924 – Called as G/T SNP

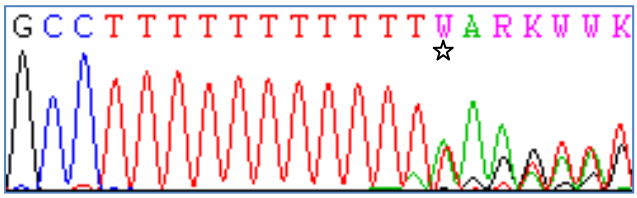


8:27473743 – Called as T/A SNP



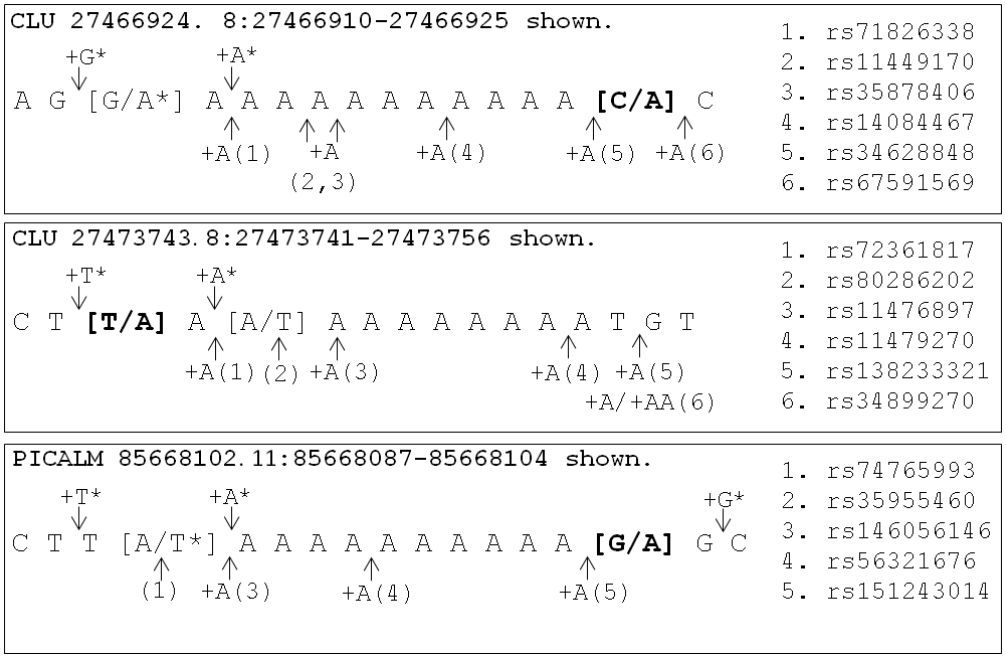
PICALM

11:85668102 – Called as C/T SNP



Electropherograms to demonstrate the failure to validate SNPs identified in the data by CRISP. In each of the instances, the false positive SNP call occurred at the location of a stretch of mononucleotide repeats. The stars in each case mark positions which are potentially heterozygous, given the offsetting of the sequence for the two alleles beyond these points, but may just be due to slippage in the sequencing because of the repetitive nature of the region, making homozygous DNA appear heterozygous. NB – sequencing of variants 8:27466924 and 11:85668102 was conducted using the reverse sequencing primer as this gave a more clear picture of the sequence at the site of the variant (because the variant fell before the mononucleotide repeat stretch when sequenced from this direction) so the sequence shown for these two is the reverse complement of the sequences displayed in Figure 4.4.

Figure 4.4 - Sequence context of spurious SNPs called next to mononucleotide repeats



Sequence context of the three SNPs next to mononucleotide repeats which failed validation by Sanger sequencing. The SNP shown in bold is the variant Sanger sequencing was designed to validate. Variants shown below the sequence are all present in dbSNP. **A.** False C/A variant call within *CLU* at position 8:27466924 (8:27466910-27466925 shown). **B.** Spurious T/A SNP call in *CLU* at 8:27473743 (8:27473741-27473756 shown). **C.** False positive SNP call at position 11:85668102 within *PICALM* (11:85668090-85668104 shown). *Other variants called by CRISP.

4.2. Discussion of Sanger validation

Despite the immense power of NGS, and the capacity it has to revolutionise genetics, there is still a limitation in terms of the error rate, which compared to Sanger sequencing, is still very high. Although great progress has been made in terms of decreasing this error rate through technological and chemical advances, as well as through improvements to alignment and variant calling softwares, there still remains an uncertainty in variant calls, which can only really be alleviated by validating these variants via an independent method. Because even modestly sized NGS experiments can generate a huge quantity of variant calls, the cost and time involved in validating each by Sanger sequencing is prohibitively high. Whilst the comparison with the data from the exome project (discussed in Chapter 5 – Exonic variants) allowed a pseudo-validation of the majority of variants detected within the exons of the genes of interest, there were still a high number of unvalidated variants.

Of the variants identified in the first sequencing project, a number stood out as having unusual characteristics. These were variants which Ensembl's VEP identified as being novel (although some subsequently proved to have been given rs numbers), yet were identified within multiple pools in our cohort,

suggesting they would be at a high frequency in a comparable population, thus seeming unlikely they would not have been previously identified.

Half of the putative SNPs in the Sanger sequenced regions were validated as being genuine SNPs, but this included only two which were deliberately targeted for validation, and almost all of the variants which were incidentally sequenced as they fell within the amplicons to be sequenced. Bansal *et al.* estimated a false positive rate of <1% using CRISP (Bansal et al. 2011), which is significantly lower than our 50% error rate. However, our rate would be expected to be higher than this, since many of the SNPs selected for validation were in possession of unusual characteristics (e.g. being present in all pools sequenced).

The location of these variants relative to the major transcripts of their respective genes is given in Table 4.1. The majority are deeply intronic, so are unlikely to be affecting splicing activity. One variant (11:85668163) falls around 1.5kb downstream of *PICALM*, where it is not likely to be affecting the gene's function or regulation. None of the variants show a high degree of conservation. The other SNP, at 11:85692181, is a synonymous exonic change, which whilst not affecting the primary sequence of the protein, could be having an effect on splicing regulatory elements. This will be discussed in further detail in Chapter 5 – Exonic variants.

The remainder of the potential SNPs sequenced did not turn out to be genuine. These constitute false positives. However, three of these transpired to be next to genuinely variant sites of small indels; two +T insertions and a –TA deletion, all of which had been previously documented and were identified by CRISP.

The final three putative SNPs which were not validated were each next to a string of polyA mononucleotide repeats. These repeats are too small for the issue to be solved by masking repeats in the design of SureSelect baits. Unlike longer repeats, short stretches of mononucleotide repeats do not make alignment of reads impossible, as enough unique sequence is present, even in short 38bp reads to allow mapping. It is likely that these are genuinely variant sites, since CRISP calls +A insertions within each of them, and all have multiple rs number +A insertions falling within the repeat region. However, the problems presented by repetitive DNA even to the relatively robust Sanger sequencing meant it was not possible to tell whether these sites were polymorphic in our samples, or whether apparent variance was simply an artefact of slippage during Sanger sequencing or PCR amplification (Clarke et al. 2001)

The main message from these findings is that it is important to be aware of documented indels and mononucleotide repeats within regions being sequenced, as these could cause potential issues with accurate identification of variants, and any variants called at such sites should be pursued with caution

and independently validated before making any assumptions about their existence.

Although the evidence from the correlations between our variant's MAF estimates and those from the HapMap and 1000 genomes projects suggests in general our method was reasonably accurate at establishing the frequency of variants, Sanger sequencing allowed the determination of the exact number of alternative alleles within a pool. This meant that the actual percentage of variant reads determined by Sanger validation could be compared to CRISP's estimate for that pool. Frequency estimates from pooled data are based on the assumption that each allele contributes equally to the total number of reads. If this assumption is incorrect, frequency estimations will not be accurate. For the majority of the variants considered in this study, there is a discrepancy between the number of alternative alleles within the Sanger sequenced pools, and the frequency estimation from CRISP, which could indicate the assumption is invalid and alleles are not equally represented. This could occur as a result of inaccurate pooling, resulting in DNA from certain subjects being over or under represented. Alternatively, it may reflect inherent biases in the target enrichment or NGS processes if DNA from certain individuals is captured or sequenced to a lesser extent. The more samples included in a study, the more accurate estimations of frequency will become (Ingman and Gyllensten 2009), so when the full 96 are considered, frequency estimates improve, but this does not mean samples are being equally represented, which should be acknowledged during analysis. Although keeping the number of samples per pool small maximises the chance of identifying variants, it reduces the reliability of frequency estimates for those pools.

With a sample size of 96, no meaningful association testing can be conducted using the NGS data alone, as there would simply be insufficient power. A much larger cohort is needed to facilitate meaningful association testing. This could be achieved by genotyping detected variants in a large case-control cohort. However, given the financial and time costs of such a study, it is impractical to conduct this for all of the variants detected. Methods for prioritising variants, and of association testing without extensive genotyping are explored in Chapters 5 and 6.

This chapter has detailed the Sanger sequencing experiments done to validate variants identified by CRISP within the NGS data. While six variants were established to be genuine SNPs, the majority of those with unusual characteristics (common in our samples, but novel) were revealed to be false positive SNP calls at the sites of either small indels or strings of mononucleotide repeats, highlighting the importance of awareness of such genetic features when conducting NGS experiments.

5. Exonic variants

This chapter explores the exonic variants detected within the NGS data. 32 such variants, falling within coding regions or UTRs of *CLU, PICALM* and *CR1* were identified. These variants were investigated for any potential relationship with AD using a two-pronged approach, firstly, testing them for association with AD in an imputed data set, and secondly, exploring potential functional implications of the variants using various *in silico* resources.

5.1. Identification of exonic variants

Ensembl's VEP was used to determine which of the identified variants within the three targeted genes fell within exonic regions (coding variants, and those in the 3′ and 5′ UTRs). These variants are given in Table 5.1, along with information on their frequencies, validation status and LD. Where the variants lie in relation to the genes is depicted in Figure 5.1. Testing the variants for association with AD was facilitated by imputing the variants in a GWAS dataset. The results of this are shown in Table 5.2. Table 5.3 is provided as a "proof of principle" that the imputed data set was capable of detecting genuine associations with AD, showing imputation and association testing outcomes for the three GWAS SNPs which first implicated *CLU, PICALM* and *CR1* in AD risk, along with findings for the *APOE* SNPs rs429358 and rs7412 which dictate the different alleles of *APOE*. A summary of the investigations of the variant's likely functionality using *in silico* resources are presented in Table 5.4, followed by more detailed explanations of some of these.

Table 5.1 - Exonic variants detected in NGS data

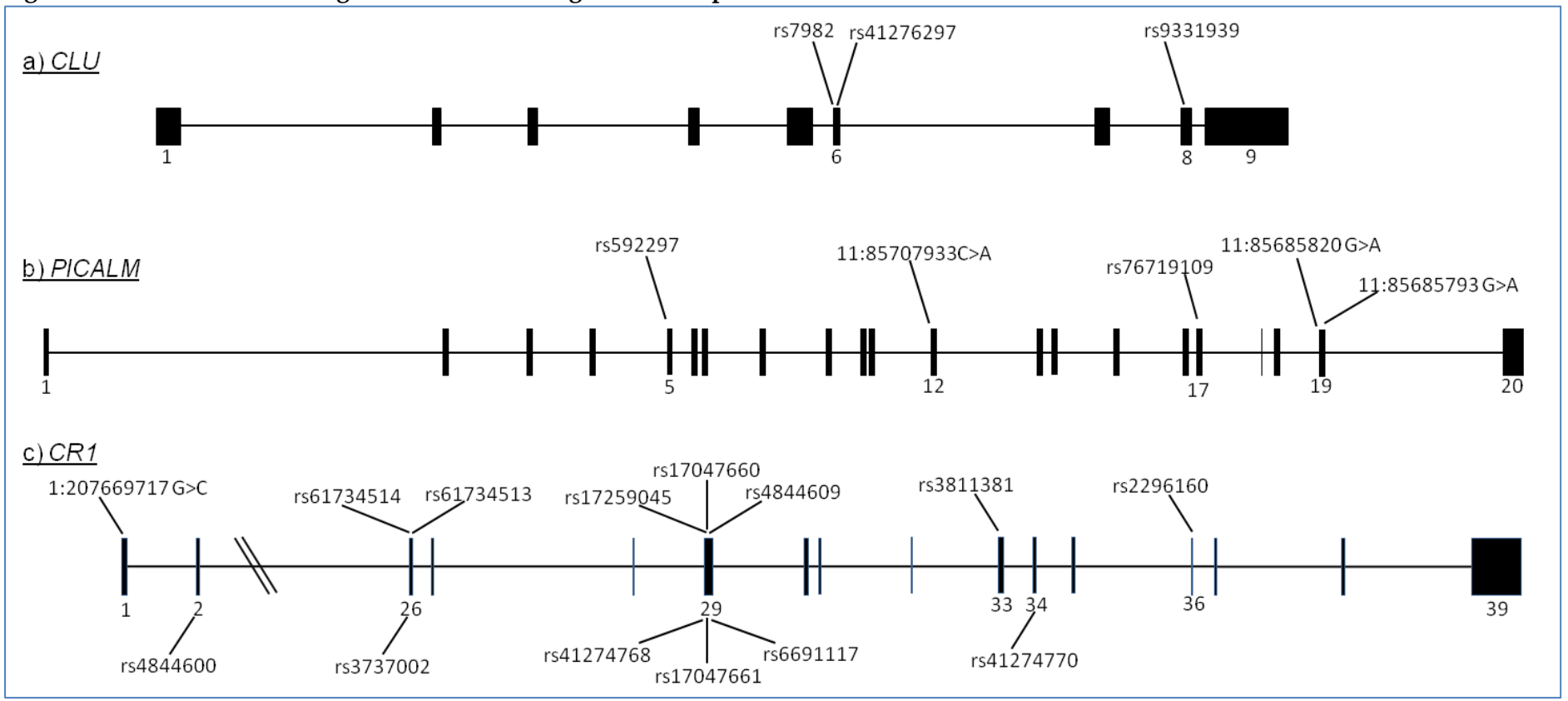
Variants	Coordinate	Ref/Alt	rsID	CRISP freq.	1kg freq.	Exome validated	LD with GWAS SNP (r^2)	LD with GWAS SNP (D′)
CLU coding	27457477	G/A	rs9331939	0.017	0.007	Yes	0.018	1
	27462481	A/G	rs7982	0.609	0.606	Yes	0.976	0.998
	27462662	G/A	rs41276297	0.005	0.004	Yes	0.006	1
CLU 3′ UTR	27454493	T/C	-	0.010	-	Yes	-	-
	27454575	C/T	rs10503814	0.047	0.042	Yes	0.024	0.852
	27454682	G/A	rs9331950	0.207	0.219	Yes	0.171	1.000
	27454686	T/C	rs9331949	0.030	0.033	Yes	0.015	0.790
	27454730	A/G	rs150082283	0.013	0.008	Yes	0.004	1.000
	27454877	A/G	rs9331947	0.026	0.050	Yes	0.055	0.816
	27454957	A/G	rs9331945	0.005	0.013	Yes	0.008	1.000
	27455114	A/G	rs9331942	0.027	0.038	Yes	0.017	0.806
	27455442	T/C	rs3087554	0.131	0.161	Yes	0.134	1.000
	27455570	C/T	-	0.004	-	Yes	-	-
PICALM coding	85685793	G/A	-	0.008	-	Yes	-	-
	85685820	T/C	-	0.005	-	No	-	-
	85692181	A/C	rs76719109	0.630	0.583	Yes	0.005	0.076
	85707933	C/A	-	0.004	-	Yes	-	-
	85725937	C/T	rs592297	0.838	0.818	Yes	0.812	0.259
CR1 coding	207669717	G/C	-	0.508	-	*	-	-
	207679307	A/G	rs4844600	0.371	0.784	Yes	0.828	1
	207760772	A/G	rs61734514	0.038	0.016	*	0.003	1
	207760773	C/T	rs3737002	0.269	0.269	*	0.084	1
	207760906	T/C	rs61734513	0.021	0.004	*	0.002	1
	207782707	A/G	rs17259045	0.111	0.092	*	0.019	1

	207782769	G/A	rs41274768	0.017	0.037	*	0.010	1
	207782856	A/G	rs17047660	0.006	0.000	*	-	-
	207782889	A/G	rs17047661	0.005	0.003	*	-	-
	207782916	A/T	rs4844609	0.955	0.980	*	0.120	1
	207782931	A/G	rs6691117	0.196	0.216	*	0.063	1
	207790088	C/G	rs3811381	0.111	0.174	Yes	0.049	1
	207791434	A/G	rs41274770	0.060	0.020	*	0.005	1
	207795320	A/G	rs2296160	0.489	0.805	*	0.742	0.908

Information on the exonic variants detected in the NGS data.

*CR1 received poor coverage in the Exome Project, so only a small number of our variants could be validated in this way. However, with the exception of one, all have rsIDs and so are likely to be genuine variants.

Figure 5.1 - Locations of coding variants relative to gene transcripts



Ribbon diagram to show the locations of the detected variants within *CLU, PICALM* and *CR1,* relative to transcripts ENST00000316403, ENST00000393346, and ENST00000400960 respectively (Hubbard et al. 2002). *CLU*'s 3'UTR is entirely contained within exon 9, thus all of the variants in this region fall within exon 9.

Table 5.2 – Association testing of exonic variants using imputed data

Variants	Coordinate	Ref/ Alt	rsID	Info	AD Freq	Control Freq	OR (95% CI)	*p*-value
CLU coding	27457477	G/A	rs9331939	0.924	0.0017	0.0076	0.224 (0.104-0.482)	0.85174
	27462481	A/G	rs7982	0.995	0.3582	0.3979	1.184 (1.101-1.273)	0.00065
	27462662	G/A	rs41276297	0.846	0.0007	0.0038	0.193 (0.060-0.617)	0.34400
CLU 3' UTR	27454493	T/C	-	-	-	-	-	-
	27454575	C/T	rs10503814	1.000	0.0375	0.0415	0.900 (0.751-1.078)	0.62956
	27454682	G/A	rs9331950	0.980	0.2428	0.1969	1.308 (1.199-1.425)	0.00269
	27454686	T/C	rs9331949	0.942	0.0144	0.0227	0.629 (0.470-0.841)	0.02268
	27454730	A/G	rs150082283	0.780	0.0007	0.0036	0.205 (0.064-0.656)	0.07772
	27454877	A/G	rs9331947	0.955	0.0165	0.0399	0.402 (0.308-0.526)	0.06516
	27454957	A/G	rs9331945	0.891	0.0119	0.0110	1.083 (0.779-1.506)	0.07940
	27455114	A/G	rs9331942	0.934	0.0150	0.0232	0.642 (0.482-0.855)	0.01555
	27455442	T/C	rs3087554	0.980	0.1015	0.1673	0.562 (0.492-0.642)	0.76905
	27455570	C/T	-	-	-	-	-	-
PICALM coding	85685793	G/A	-	-	-	-	-	-
	85685820	T/C	-	-	-	-	-	-
	85692181	A/C	rs76719109	0.999	0.4062	0.4125	1.026 (0.956-1.102)	0.36776
	85707933	C/A	-	-	-	-	-	-
	85725937	C/T	rs592297	0.999	0.1771	0.1947	1.123 (1.027-1.230)	0.65868
CR1 coding	207669717	G/C	-	-	-	-	-	-
	207679307	A/G	rs4844600	0.958	0.1947	0.1837	0.931 (0.846-1.024)	0.06633
	207760772	A/G	rs61734514	0.278	0.0000	0.0005	-1.000	0.07874
	207760773	C/T	rs3737002	1.000	0.2554	0.2638	0.957 (0.884-1.036)	0.28854
	207760906	T/C	rs61734513	0.766	0.0027	0.0014	1.877 (0.898-3.920)	0.69471
	207782707	A/G	rs17259045	0.984	0.0495	0.1023	0.457 (0.387-0.540)	0.65984

	207782769	G/A	rs41274768	0.991	0.0221	0.0255	0.864 (0.685-1.091)	0.69434
	207782856	A/G	rs17047660	0.954	0.0024	0.0011	2.137 (0.969-4.713)	0.51066
	207782889	A/G	rs17047661	0.975	0.0041	0.0029	1.419 (0.805-2.500)	0.64195
	207782916	A/T	rs4844609	0.975	0.0041	0.0029	1.419 (0.805-2.500)	0.64195
	207782931	A/G	rs6691117	1.000	0.1997	0.1933	1.041 (0.955-1.136)	0.12276
	207790088	C/G	rs3811381	0.995	0.1691	0.1637	1.040 (0.947-1.142)	0.09718
	207791434	A/G	rs41274770	0.382	0.0000	0.0027	-1.000	0.04293
	207795320	A/G	rs2296160	0.998	0.2006	0.1811	0.881 (0.807-0.962)	0.22102

Association testing of exonic variants detected in the NGS data using imputed data. The OR and 95% CI given correspond to the allele listed as "alt", but the MAF refers to which ever allele has the lower frequency, so these do not always correspond. Yellow highlighting indicates statistically significant finding, while orange highlighting indicates significance at the $p<0.05$ level.

Table 5.3 – Imputation and association testing findings of GWAS and *APOE* SNPs

Gene	SNP	Ref/Alt	Info.	MAF AD	MAF Controls	OR (95% CI)	*p*-value
CLU	rs11136000	T/C	0.9999	0.3616	0.3990	1.172 (1.090-1.260)	0.00089
PICALM	rs3851179	T/C	0.9999	0.3313	0.3731	1.201 (1.116-1.292)	0.01913
CR1	rs6656401	A/G	0.9824	0.1766	0.1628	0.907 (0.824-0.998)	0.04176
APOE	rs429358	T/C	0.9745	0.3351	0.1466	2.934 (2.691-3.198)	1.93×10^{-71}
	rs7412	C/T	0.9755	0.0323	0.0795	0.386 (0.317-0.470)	1.13×10^{-9}

Findings from imputation and association testing of the originally identified AD GWAS risk variants, plus variants from the *APOE* locus which dictate *APOE* ε2, ε3 and ε4 alleles. The OR and 95% CI given correspond to the allele listed as "alt", but the MAF refers to which ever allele has the lower frequency, so these do not always correspond.

Table 5.4 – Functional assessment of exonic variants using *in silico* resources

Variants	Coordinate	Ref/ Alt	rsID	Consequence	Exon	AA change	Polyphen-2
CLU coding	27457477	G/A	rs9331939	Synonymous	8	-	-
	27462481	A/G	rs7982	Synonymous	6	-	-
	27462662	G/A	rs41276297	Missense	6	T255I	Benign(0.013)
CLU 3′ UTR	27454493	T/C	-	3′ UTR	9	-	-
	27454575	C/T	rs10503814	3′ UTR	9	-	-
	27454682	G/A	rs9331950	3′ UTR	9	-	-
	27454686	T/C	rs9331949	3′ UTR	9	-	-
	27454730	A/G	rs150082283	3′ UTR	9	-	-
	27454877	A/G	rs9331947	3′ UTR	9	-	-
	27454957	A/G	rs9331945	3′ UTR	9	-	-
	27455114	A/G	rs9331942	3′ UTR	9	-	-
	27455442	T/C	rs3087554	3′ UTR	9	-	-
	27455570	C/T	-	3′ UTR	9	-	-
PICALM coding	85685793	G/A	-	Synonymous	19	-	-
	85685820	T/C	-	Synonymous	19	-	-
	85692181	A/C	rs76719109	Synonymous	17	-	-
	85707933	C/A	-	Missense	12	Q398H	Probably damaging(0.987)
	85725937	C/T	rs592297	Synonymous	5	-	-
CR1 coding	207669717	G/C	-	Synonymous	1	-	-
	207679307	A/G	rs4844600	Synonymous	2	-	-
	207760772	A/G	rs61734514	Missense	26	T1408A	Probably damaging(0.974)
	207760773	C/T	rs3737002	Missense	26	T1408M	Probably damaging(0.997)
	207760906	T/C	rs61734513	Synonymous	26	-	-
	207782707	A/G	rs17259045	Missense	29	N1540S	Benign(0.001)

	207782769	G/A	rs41274768	Missense	29	V1561M	Benign(0.039)
	207782856	A/G	rs17047660	Missense	29	K1509E	Probably damaging(0.988)
	207782889	A/G	rs17047661	Missense	29	R1601G	Possibly damaging(0.876)
	207782916	A/T	rs4844609	Missense	29	T1610S	Benign(0)
	207782931	A/G	rs6691117	Missense	29	I1615V	Benign(0)
	207790088	C/G	rs3811381	Missense	33	P1827R	Benign(0.033)
	207791434	A/G	rs41274770	Missense	34	K1853R	Possibly damaging(0.823)
	207795320	A/G	rs2296160	Missense	36	T1969A	Benign(0.001)

Information on the exonic variants detected in the NGS data from various bioinformatic resources used to prioritise variants on likely functionality, including prediction of consequences in terms of the protein sequence and structure. Possibly and probably damaging effects on the protein predicted by Polyphen-2 highlighted in orange and yellow respectively.

Functional investigations

Polyphen-2 was used to predict whether any of the 13 missense mutations called by CRISP within the NGS data would have detrimental effects on the structure and/or function of the encoded protein, since some amino-acid changes can be tolerated while others can have a significant deleterious impact. As can be seen from Table 5.4, one variant within *PICALM* and three within *CR1* were predicted to probably be damaging to the structure of the protein, while a further two in *CR1* were classified as possibly damaging.

Splicing

As well as disrupting the sequence of the protein, non-synonymous SNPs may be affecting the splicing of the genes, while synonymous variants may be exerting an effect on the regulation of the genes via this mechanism. Three programs with four functions were used to investigate the effect the detected variants may be having on splicing, with the results of these investigations presented in Table 5.5. Further information on variants predicted to be affecting splicing is given in Figure 5.2. No significant effects on splicing were predicted by any of the programs, other than ESEfinder's SRProtein prediction function.

Table 5.5 – Splicing investigations of variants detected by NGS

Gene	Variant	rsID	ESEfinder - SpliceSites	ESEfinder - SRProteins	BDGP	NetGene2
CLU	27457477	rs9331939	-	-	New (weak) acceptor site introduced by variant	-
	27462481	rs7982	-	-	-	-
	27462662	rs41276297	-	Yes	-	-
PICALM	85685793	-	-	-	-	Acceptor site slightly weaker in variant sequence
	85685820	-	-	-	-	Acceptor site slightly weaker in variant sequence
	85692181	rs76719109	Donor site slightly stronger in variant sequence	Yes	-	-
	85707933	-	Acceptor site slightly weaker in variant sequence	Yes	-	-
	85725937	rs592297	-	Yes	-	-
CR1	207669717	-	-	Yes	-	New (weak) acceptor site introduced by variant
	207679307	rs4844600	-	-	-	-
	207760772	rs61734514	-	Yes	-	-
	207760773	rs3737002	-	Yes	-	-
	207760906	rs61734513	-	-	-	-
	207782707	rs17259045	-	-	-	Acceptor site slightly weaker in variant sequence
	207782769	rs41274768	-	-	-	-
	207782856	rs17047660	-	-	-	-
	207782889	rs17047661	-	-	-	Donor site slightly stronger in

						variant sequence
	207782916	rs4844609	-	-	-	Donor site slightly stronger in variant sequence
	207782931	rs6691117	-	-	-	Donor site slightly stronger in variant sequence
	207790088	rs3811381	-	Yes	-	Donor site slightly stronger in variant sequence
	207791434	rs41274770	-	-	-	-
	207795320	rs2296160	-	Yes	New (weak) acceptor site introduced by variant	Acceptor site slightly stronger in variant sequence

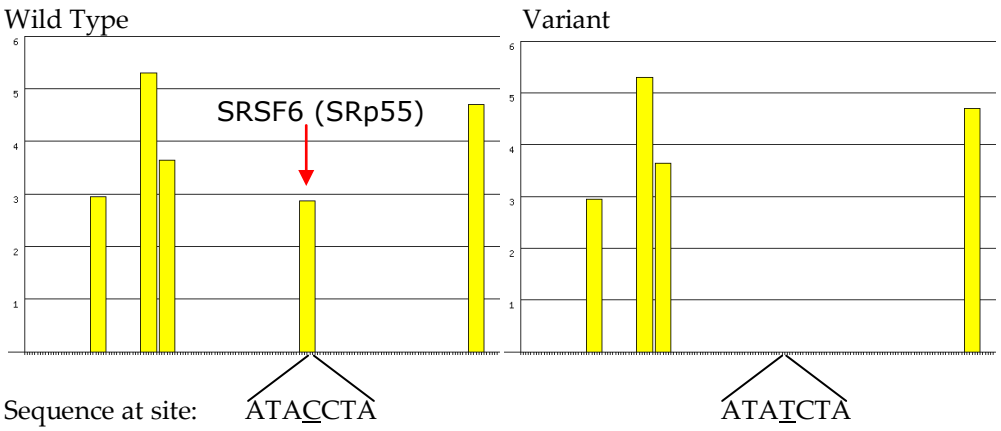
Results from the *in silico* splicing investigations conducted on the coding variants detected in the NGS data. A dash (-) in the box signifies no notable differences between the wild type and variant versions of the sequence. Where strong differences were predicted, the results are highlighted in yellow. Where a "yes" is recorded, see Figure 5.2 for further explanation of the findings.

Figure 5.2 – Splicing results from ESEfinder's SRProteins function

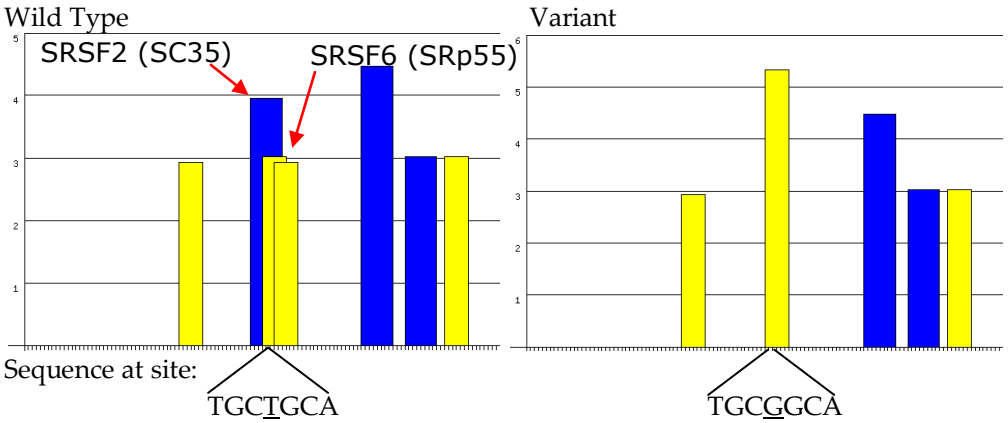
Key -

SRSF1
SRSF2
SRSF5
SRSF6

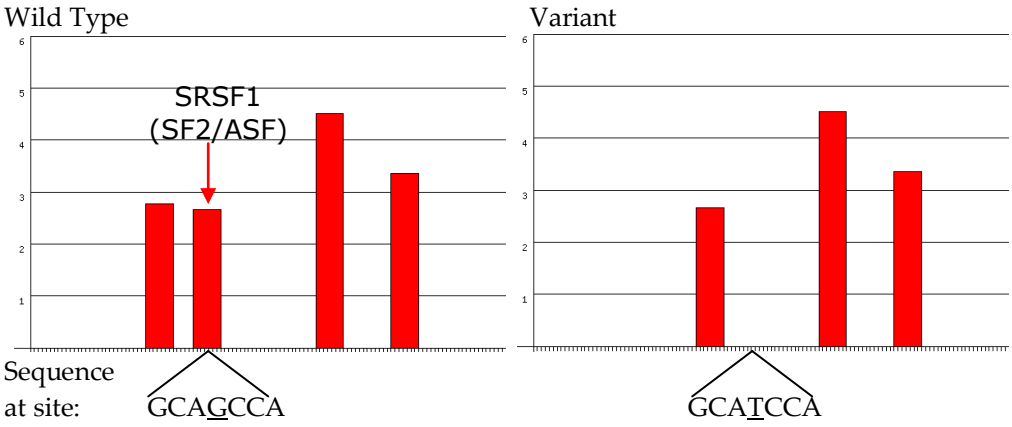
CLU - rs41276297



PICALM - rs76719109

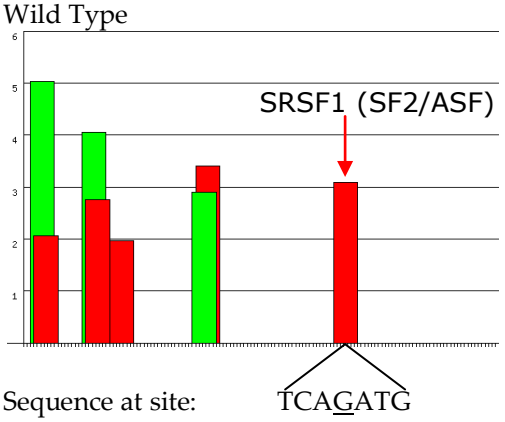


11:85707933 C>A

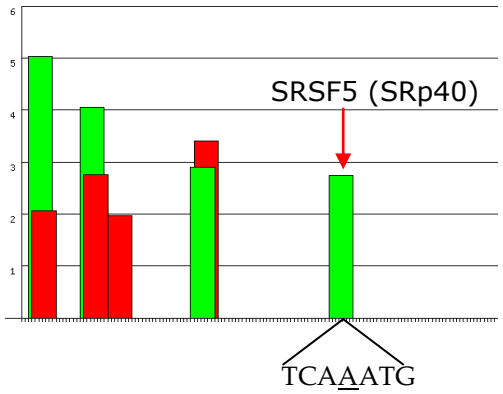


rs592297

Wild Type



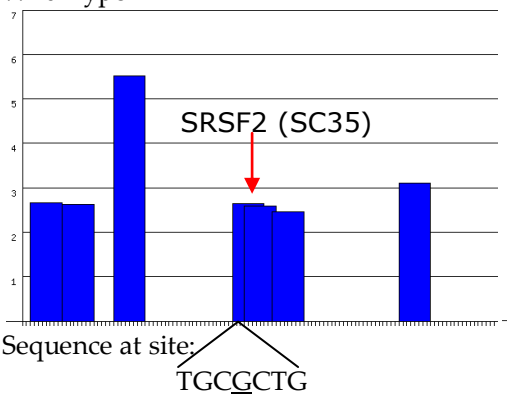
Variant



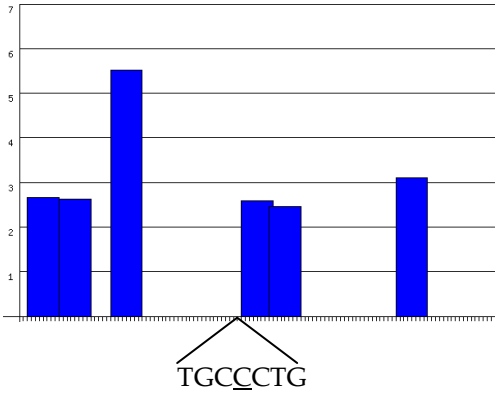
Sequence at site: TCAGATG TCAAATG

CR1 - 1:207669717 G>C

Wild Type



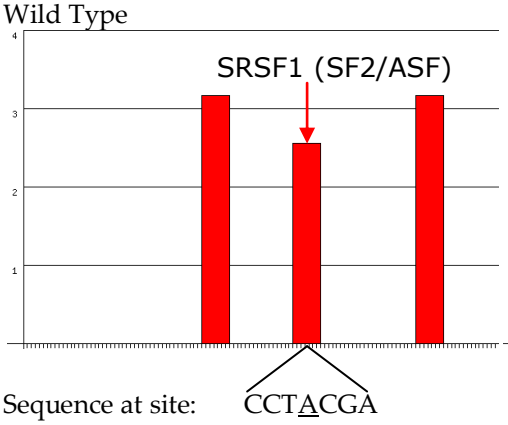
Variant



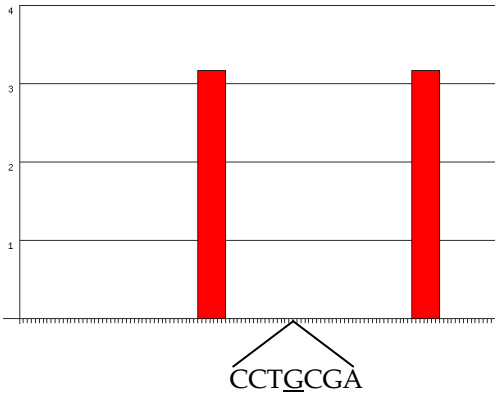
Sequence at site: TGCGCTG TGCCCTG

rs61734514

Wild Type



Variant



Sequence at site: CCTACGA CCTGCGA

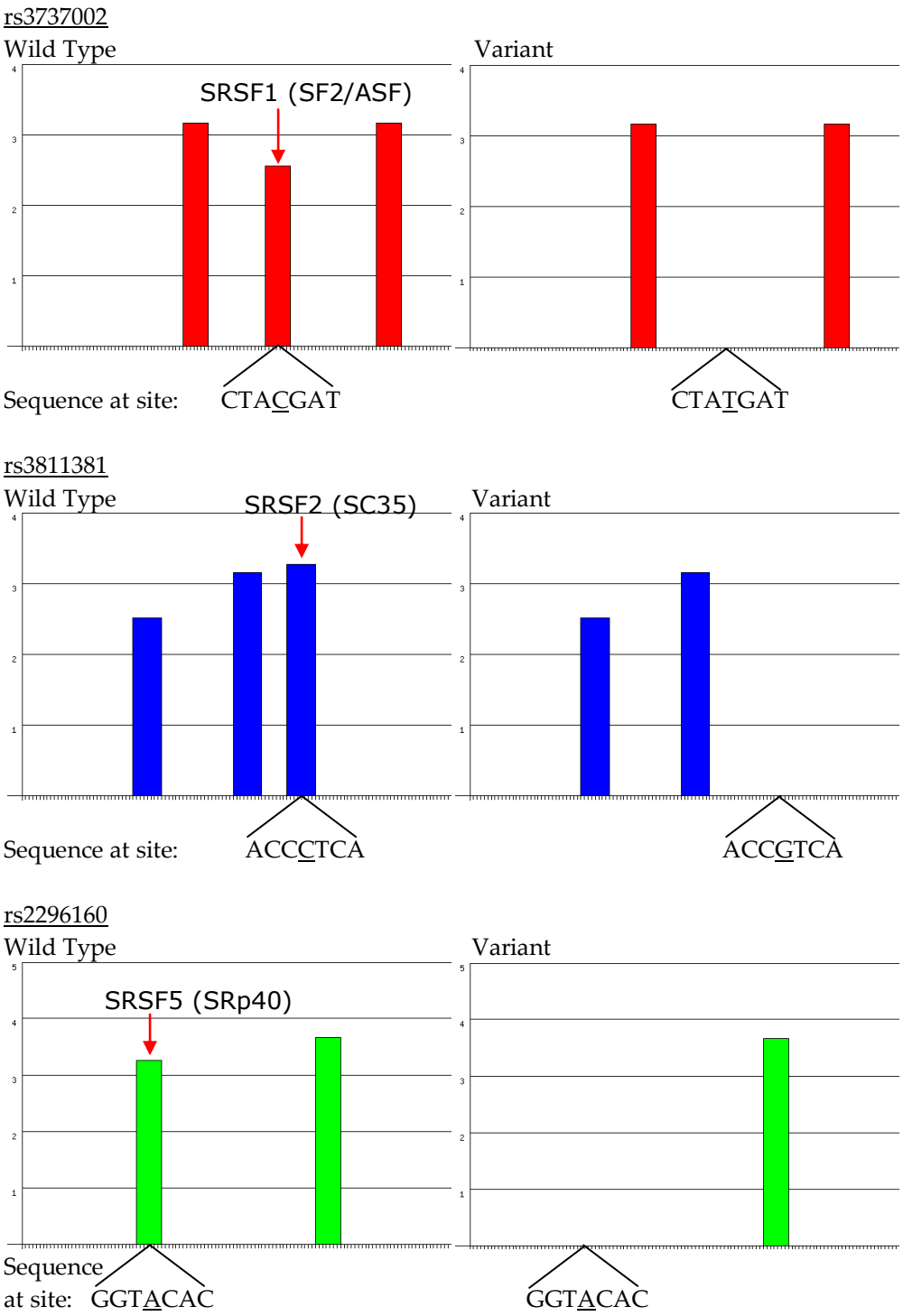


Figure showing the results from the splicing investigations conducted on the coding variants in NGS data using ESEfinder's SRProtein function. For each variant, the SRProtein prediction for the wild type sequence is shown on the left, while the variant sequence's prediction is shown on the right. The colours of the bar indicate the particular SRProtein(s) affected (see key at top of figure), while the height of the bars are proportional to the expected strength of the predicted sites. Differences are highlighted by arrows, and the sequences of the wild type and variant sites are shown below the graphs.

3' UTR SNPs

Ten of the variants detected in the NGS data at the *CLU* locus fell within the gene's 3'UTR. To determine whether any of these variants were likely to be affecting the regulation of the gene by altering the binding sites of miRNAs, it was ascertained whether any of the variants fell within the predicted binding sites of any miRNAs listed in the TargetSan (www.targetscan.org) database (Garcia et al. 2011). The website listed 12 miRNA binding sites within the 3'UTR of *CLU* on the date of analysis (16.04.13). When the locations of these were overlaid with the locations of the 3'UTR variants, two SNPs (rs9331945 and rs9331947) were found to fall within miRNA sites, with rs9331947 intersecting four potential binding sites. PITA enabled a comparison of the miRNA sites predicted for the wild type and variant sequences for each of the SNPs, using the "Predict Your UTR" function. A summary of the results of these investigations is given in Table 5.6.

Table 5.6 – Investigations of miRNA binding sites in *CLU*'s 3'UTR

rsID	miRNA (TargetScan)	PITA ΔΔG (WT)	PITA ΔΔG (Variant)
rs9331945	hsa-miR-3138	Not found	Not found
rs9331947	hsa-miR-1184	-1.62	-1.05
	hsa-miR-4418	Not found	Not found
	hsa-miR-509-3-5p	-1.12	-1.02
	hsa-miR-509-5p	-0.51	-0.41

Co-located SNPs and predicted miRNA sites in *CLU*'s 3' UTR region. The ΔΔG score from PITA is a measure of how strong the binding between the miRNA and the binding site would be expected to be.

5.2. Discussion of exonic variants

It is estimated that around 85% of variants with large effects on disease related traits are exonic, despite coding regions only comprising around 1% of the total genome (Choi et al. 2009). When even relatively small scale sequencing projects such as this can identify a large number of variants, prioritisation is crucial to ensure only variants likely to be affecting the phenotype in question are pursued. This ensures minimal time and resources are consumed investigating variants unlikely to be contributing to the disease process. Concentrating on coding variants is a simple way to reduce the total number of variants being investigated, while maintaining those more likely to be disease relevant. Exonic variants are simpler to predict and verify the functional consequences of. This chapter explores the exonic variants found in the NGS data.

A two-pronged approach was adopted in the further prioritisation of coding variants in this study. Firstly, imputation was used to predict genotypes of SNPs of interest in a GWAS data set, enabling testing for association with AD. Secondly, likely functional consequences of the variants were assessed using

in silico prediction programs, which highlighted variants likely to be affecting the normal functioning of the gene.

13 exonic variants were called within the NGS data for *CLU*, with three falling in coding exons, and the remaining ten in the gene's 3'UTR. Five exonic variants were called within *PICALM*, while 14 were detected in *CR1*. In each of these genes, all of the detected exonic variants fell within coding regions.

NGS notoriously has a high rate of false positive variant calls, due to the nature of the technology used, so validation of variants via an independent method is imperative (Lord et al. 2012). All of the variants bar one in the *CLU* and *PICALM* genes were confirmed as being genuine by comparison with data from the Exome Project at UCL. The *PICALM* SNP not validated, at position 11:85685820, is a putative synonymous variant which showed no evidence of functional effect on the gene, so was not deemed to be worth validating by Sanger sequencing.

Validation via comparison with the Exome Project data was not possible for the majority of variants called in the NGS data in the *CR1* region because of coverage issues they experienced in the area (again, reiterating how problematic the highly repetitive gene structure is for NGS). However, all but one of the *CR1* coding variants identified already had rs numbers assigned, so can be assumed to be genuine, given their prior documentation. The remaining variant, at position 1:207669717 is a putative synonymous variant, which does show some evidence of affecting the activity of a splicing enhancer signal. However, despite being apparently novel, the alternative allele had an estimated frequency of 0.508 in the CRISP data, which as discussed in the previous chapter, is indicative of a false positive call. Upon closer inspection of the CRISP data, the average coverage at the position per individual was just 0.28. These pieces of evidence combine to make it highly likely this is a false positive call, and thus not worthy of pursuit.

In terms of the locations of the detected variants relative to the three gene's transcripts, *CLU*'s three exonic variants all falling within exons 6-8 is consistent with Bettens's findings that variants cluster within the region encoding the *CLU* β-chain (exons 5-8), although three variants is too small a number to draw any meaningful conclusions from this (Bettens et al. 2012). *CLU*'s exon 9 is home to the gene's 3'UTR, so all ten of the 3'UTR variants fall within this exon. Clustering in to a small number of exons in this way can be indicative of involvement in a phenotype – could the 3'UTR of *CLU* be harbouring multiple risk affecting variants? Association testing in the imputed data and investigations of functionality using *in silico* resources of these variants is discussed later. The exonic variants detected within *PICALM* show a reasonable spread throughout the gene's exons. In CR1, two variants were identified within the first two exons of the gene (although as discussed, one of these is not believed to be genuine), while the remainder fell between exons 26 and 36, with the majority in exons 26 and 29. While such clustering can be

indicative of functional involvement of the gene in the disease process (e.g. the exons encode a particular domain of importance), it is likely in this case the apparent clustering is an artefact of the poor sequencing coverage of a large proportion of the *CR1* gene. The ~44kb gap in sequencing spans the same region where a distinct lack of variants were detected, providing a much more mundane explanation for the apparent clustering.

Imputation allowed a method for the variants to be tested for association with AD, without needing costly and time consuming direct genotyping of a large number of variants in a large sample cohort. The utilisation of our own combined Mayo/ARUK GWAS data, in combination with publically available control data from the WTCCC2 project gave a combined sample set of 2067 AD cases and 7376 controls.

There are certain caveats when using imputation in this way, which should be considered when analysing data from this methodology. The use of imputation reduces the power of association testing when compared to directly genotyped data, because of the inherent uncertainty in genotypic calls. This issue is particularly pronounced when imputing rare variants, since there are less of the variants present in the reference set on which to base the imputation, giving a higher level of uncertainty in the genotypic calls. Association testing has a lower power for rarer variants anyway (less variants to make comparisons between), which is only exacerbated by this uncertainty. Another potential issue with this particular methodology is the inclusion of the WTCCC2 project data, which are not ideal control samples for AD since the age of recruitment for such control cohorts are typically earlier than the average age at onset for AD. It is inevitable that some individuals being classed as controls here will actually develop AD in later life. This misclassification in phenotype is another source of loss of power in association studies. This is more likely to be a source of error in rarer variants than common, since small numbers of misclassifications will have more pronounced effects when the variants in question are rare.

While it is not possible to know how these factors will affect our association testing, it was possible to look for positive controls to establish whether associations could be found via this methodology in a sample set of this size.

Imputation using the same GWAS sample set was also conducted on the two *APOE* allele defining SNPs. The *APOE* locus is the longest known genetic risk factor for late-onset AD because of its strong effect size. As such, these SNPs should be clearly associated with AD in the imputation sample set.

The three alleles of *APOE* are defined by genotype at two variant sites – rs429358 and rs74112. The protective ε2 allele is defined by T bases at each of the two positions; the neutral ε3 allele is defined by a T at the first variant position, and a C at the second position; while the other, risky, ε4 allele is defined by Cs at each position. Thus, within our imputed data, the C allele at

rs429358 would be expected to have an OR showing increased risk of AD (OR of 2.9), while the T allele at the second position, rs7412, would be expected to show a protective effect on AD risk (OR of 0.4). This was indeed found to be the case, with highly significant associations with AD in the expected directions for both variants (p=1.93x10^{-71} and p=1.13x10^{-9}, respectively).

Once it had been established that the imputation methodology was capable of finding associations of the magnitude of *APOE*, the original GWAS SNPs were also tested for association with AD. For each of the three genes, the original GWAS SNP did show association with AD in the imputed data (with nominal p-values of 0.0009, 0.0191, and 0.04176 for *CLU*, *PICALM* and *CR1* respectively). The directions of effect were consistent with those reported in the Harold et al. and Lambert et al. GWAS for all three SNPs (although the reporting in the GWAS papers was for the opposite allele to the one tested in our association test – our OR indicates the C allele at rs11136000 is risky, the GWAS papers reported the T allele at the same SNP to be protective. These equate to the same thing, thus our directions are consistent with the published data) (Harold et al. 2009; Lambert et al. 2009). This is good evidence that this methodology is capable of detecting associations of lower magnitudes. The GWAS variants, however, are all common, with MAFs >0.16, so it remained to be seen whether effects of this magnitude in rarer SNPs could be identified.

The two coding variants successfully imputed within the *PICALM* region showed no evidence of association with AD. Just one out of the 13 successfully imputed coding SNPs within the *CR1* locus (rs41274770) was found to be associated with AD at the p<0.05 level. No OR could be calculated for this variant given that no alternative alleles were imputed in the case samples. Another variant, rs61734514, which was only imputed in control samples and not cases also showed a suggestive association with the condition (p=0.0787). With 13 coding variants successfully imputed, 13 multiple tests of association are being conducted. A Bonferroni correction for this number of tests would require a p-value <0.0038 for statistical significance, which neither of these variants achieves, so they are at best suggestive. However, both of these are rare variants, with 1000 genomes project EUR MAFs of 0.020 and 0.016 respectively. The known limitations of imputation of low frequency variants may be holding these SNPs back from achieving statistical significance. Although no alternative alleles of these variants were imputed in case samples, it is clear that these variants do occur in AD patients since they were found in the NGS data generated from AD patient samples only, which calls into question the apparent protective effect of the variants implied by the imputed data. To establish whether these variants are involved in AD risk, direct genotyping in a case-control cohort large enough to give sufficient power would be required.

These two variants are in strong LD with the *CR1* GWAS SNP (D' is 1 for both). rs41274770 falls within exon 34 and rs61734514 lies within exon 26 (in the transcript of the common F allele). Both of these SNPs are missense

mutations, causing changes to the encoded protein of K1853R and T1408A, which are predicted by Polyphen-2 to be possibly and probably damaging respectively (0.823 and 0.974). Again, this renders the implied protective effect of the variants from the imputed data unlikely. There is additionally some evidence that rs61734514 may disrupt a binding site for the splicing SR-protein SRSF1 (SF2/ASF). If this is indeed the case, it could offer another explanation as to how the variant may be causatively linked to AD, by disrupting the normal splicing of the gene. Given the known limitations of *in silico* functional prediction programs, it would be necessary to characterise these effects experimentally to assess whether the variants are indeed pathogenic to the structure of the protein or the expression of the gene, and how this could be related to AD pathology.

None of the other variants predicted to be damaging to the protein structure in either *PICALM* or *CR1* were suggestive of association with AD, so whether they are indeed having an effect on protein structure, and whether this is relevant to AD pathology remains unclear.

Although none of the *PICALM* variants were found to be associated with AD in this dataset, only two of the variants were actually imputed; those which were not imputed successfully showed some evidence of potential functionality, so should also be considered. These are all variants without rs numbers. The unvalidated variant, at position 11:85685820 had no evidence of functional effects, and this is also true of the variant at 11:8568793 – both were synonymous with no evidence of effects on splicing. The other variant, at position 11:85707933, is a missense mutation within exon 12 of the gene, encoding the amino acid change Q398H, which Polyphen predicted would probably be damaging (0.987). The variant also showed evidence of affecting the binding site of splicing SR-protein SRSF1, so could potentially be influencing the splicing pattern of the gene. The variant had a frequency of 0.004 in the CRISP data, which is indicative that just one alternate allele was present in the whole pool of 192 alleles. Without a listed frequency in the 1000 genomes project data, it is likely that this allele is highly rare. Genotyping in a very large case-control cohort would be needed to establish if there was any association with AD, given the low frequency of the variant. Experimental characterisation of the predicted functional effects to the protein structure and splicing patterns could help ascertain whether the variant was likely to have a pathological role in the gene, and how this may related to AD pathogenesis.

Although it has already been stated that *PICALM* SNP rs592297 was not associated with AD in the imputed data set, there is the suggestion that it may be having an effect on the splicing of the gene, with evidence that the variant may alter a splicing SR-protein binding site within exon 5 of the gene, such that an SRSF1 site is abolished, and a SRSF5 site is introduced. During the course of this project, another paper reported this as a possible splicing mutation relevant to AD based on *in silico* evidence, but again, found no association between the SNP and the condition (Schnetz-Boutaud et al. 2012).

Similarly, although rs76719109 showed no evidence of association with AD, it did show evidence of an effect on splicing, potentially altering a SRSF2 site within the gene's 17th exon. Functional characterisation of the effects of these variants on splicing would allow it to be established whether these SNPs do indeed affect the splicing of the gene, and may help to determine whether and how this is of consequence in AD development.

One of the three imputed coding variants within *CLU* (rs7982) showed association with AD in our imputed dataset (p=0.00065). Harold et al. in their GWAS paper had reported the variant as being in strong LD with rs11136000 (r^2 0.976, D' 0.998), which showed a similar level of association with AD as rs11136000 in their data (Harold et al. 2009). In other studies, the SNP has not been found to be significantly associated with AD, perhaps due to insufficient sample sizes to give the power required (Guerreiro et al. 2010; Bettens et al. 2012). In the imputed data, rs7982 showed a more significant association with AD than did rs11136000. This difference, although only marginal, may suggest that the variant is causative. As a synonymous variant, any effect on phenotype is not via the alteration of the amino acid sequence of the protein, and the *in silico* resources used did not suggest the variant was likely to be affecting the splicing of the gene, so it is unclear from these investigations how the variant could be related to AD pathogenesis. The greater significance of this SNP over the GWAS SNP in the imputed data set is only slight, so may simply be a quirk, and the SNP may simply be tagging the same unknown causative factor(s) tagged by rs11136000.

Of the ten SNPs found in *CLU*'s 3'UTR in the NGS data, eight were successfully imputed, and three (rs9331950, rs9331949 and rs9331942) showed evidence of association with AD at the $p<0.05$ level. When corrected for the eight multiple tests being conducted, statistical significance would require $p<0.0065$, which only SNP rs9331950 surpassed. However, whilst rs9331950 is common (1000 genomes EUR MAF 0.197), the other two are rare, both with 1000 genomes EUR MAFs of 0.023, so as before, it could be the limitations of imputing rare variants bringing the power of the association testing down, and preventing the variants from reaching significance. All three of these variants are in strong LD with rs11136000 (D' scores all above 0.79), so may be further tags for the same source, or may themselves be causal in some way.

To investigate the function of these 3'UTR variants, *in silico* resources were again utilised. The 3'UTRs of genes are known to harbour regulatory elements implicated in multiple regulatory process, including control of translational efficiency, transcript cleavage and stability, as well as polyadenylation (Barrett et al. 2012; Pichon et al. 2012). One method by which post-transcriptional regulation of gene expression is mediated is via the binding of miRNAs to sequences within a gene's 3'UTR (Barrett et al. 2012). *In silico* prediction programs were used to determine whether any of the *CLU* 3'UTR SNPs identified in the NGS data could be having an effect on the expression of the gene via the disruption of miRNA binding.

Two of the variants (rs9331945 and rs9331947), which were not associated at the $p<0.05$ level, but were suggestive of association (with p-values of 0.079 and 0.065 respectively) were found to fall within potential miRNA binding sites. Both of these are relatively rare variants (1000 genomes MAFs of 0.013 and 0.050 respectively), so again the weakness of imputation in rarer variants may have affected this, and genotyping in a case-control cohort would be needed to clarify any association. The 3'UTR variants which were associated with AD at the $p<0.05$ level did not fall within any miRNA sites.

For the two variants which did fall in miRNA binding sites according to Target Scan, PITA enabled the comparison of the wild type and variant UTR sequences in terms of miRNA binding site strengths. Two of the five sites predicted to be affected by Target Scan were not detected by PITA. This lack of concordance between programs is likely to indicate they are not genuine binding sites. The other three potential sites affected by the SNPs were recognised by PITA, and in each case, the $\Delta\Delta G$ score was weakened by the presence of the variant allele. However, all of the $\Delta\Delta G$ scores were low. The more negative the $\Delta\Delta G$ score is, the stronger the binding of the miRNA to the target site is expected to be, with -10 recommended as a rough cut-off point – any sites with a score more negative than -10 are likely to be functional at endogenous miRNA expression levels. None of the sites within the *CLU* 3'UTR showed scores even approaching this cut-off, thus the sites are not likely to be actively used in the gene's regulation, and the slight differences in $\Delta\Delta G$ scores between the sequences tested is unlikely to be of any consequence.

This chapter has explored the exonic variants detected within the NGS data, with the aim of prioritising those variants most worthy of follow up research based on functional *in silico* investigations and testing for association with AD in an imputed data set. Within *CLU*, four variants either significantly or suggestively associated with AD were identified, but a lack of functional evidence from the *in silico* resources renders these a low priority for follow up. *PICALM* contained two variants with evidence of affecting splicing, but were not associated with AD in the imputed data set. A third variant within *PICALM*, a novel SNP which could not be imputed, was predicted to both affect splicing, as well as being a missense mutation, which Polyphen-2 predicted would probably be damaging, rendering it a high priority for further study. Two missense variants within *CR1*, which were predicted to be damaging by Polyphen-2 also showed suggestive association with AD, considering their rarity (with p-values of 0.04 and 0.08 and frequencies of around 1-2%), and so present strong candidates for further research.

6. Non-coding variants

This chapter explores the variants detected within the NGS data that did not fall within exonic regions. While non-coding variants are often overlooked, they can affect the regulation and expression of genes, and may well have an effect on disease risk. It is particularly important to consider such variants when there is evidence that the GWAS signal does not stem from coding changes, such as is the case for *CLU*. As with the coding variants discussed in the previous chapter, a two-pronged approach was adopted, combining data on potential functional activity from *in silico* resources with AD association testing in an imputed data set.

6.1. Identification of non-coding variants

When Ensembl's VEP was used to ascertain the locations of the variants detected relative to the genes, it was established that over 850 of the variants detected were non-coding. It was decided that *CLU* and the region to the 5' of *PICALM*, along with the rs3851179 LD block were likely to be the most interesting in terms of non-coding variation (see discussion 6.4 for details). The results of the investigations of these regions are presented below.

6.2. *CLU*

75 non-coding variants were detected within the targeted *CLU* locus, falling in the upstream, downstream and intronic regions. Of these, nine fell upstream of the gene (two of which were novel), 19 fell downstream (five novel), while the remaining 47 (including 12 novel) were situated in the intronic regions of the gene. It was necessary to prioritise these variants, so those in areas with multiple lines of evidence of functionality from ENCODE were selected, and are presented in Table 6.1. Table 6.2 shows the ENCODE data for these variants, upon which the prioritisation was based. Association testing of these variants in imputed data was performed, with the results presented in Table 6.3.

Table 6.1 – Summary of *CLU* non-coding SNPs of interest

Coordinate	Ref/ Alt	rsID	CRISP freq.	1kg freq.	PhyloP	Phastcons	Location relative to *CLU*	LD with GWAS SNP (r^2)	LD with GWAS SNP (D')
27466157	T/C	rs1532276	0.593	0.603	-0.7211	0	Intronic	0.976	0.988
27466181	T/C	rs1532277	0.592	0.603	0.0564	0	Intronic	0.976	0.988
27466315	T/C	rs1532278	0.573	0.606	0.3340	0	Intronic	0.976	0.988
27468503	C/A	rs867230	0.597	0.591	0.0937	0	Intronic	0.930	0.988
27469971	G/C	rs9331883	0.007	0.003	-1.8881	0	Intronic	0.004	1.000
27470010	G/T	-	0.006	-	0.0037	0	Intronic	-	-
27470597	G/A	rs34109053	0.302	0.272	0.5443	0.0215	Intronic	0.178	0.887
27471673	C/T	-	0.004	-	2.0935	0.6673	Intronic	-	-
27471748	C/A	-	0.053	-	-0.0156	0	Intronic	-	-
27472859	C/A	rs76646010	0.022	0.033	0.1443	0	Upstream (311bp)	0.019	1.000
27474202	A/G	rs9314349	0.379	0.393	-0.5716	0	Upstream (1654bp)	0.099	0.467
27474541	G/A	rs117148275	0.009	0.007	-0.7374	0	Upstream (1993bp)	0.001	0.370
27474587	C/G	rs56025648	0.050	0.055	0.0914	0	Upstream (2039)	0.036	0.605
27474599	C/T	rs1982229	0.358	0.368	-0.0191	0	Upstream (2051bp)	0.108	0.503
27474871	G/A	rs77336101	0.010	0.021	-0.8479	0	Upstream (2323bp)	0.013	1.000

Information on the non-coding SNPs identified in the NGS data for the *CLU* region. Variants shown are those filtered based on multiple lines of evidence from ENCODE (see Table 6.2) that the region in which they fall is functionally active. Variants showing conservation scores above 0.5 (by either measure) are highlighted in yellow.

Table 6.2 – ENCODE data on *CLU's* non-coding variants

Coordinate	Ref/Alt	rsID	TFBS	DNASE	NHA_H3K4me1	NHA_H3K4me3	NHA_H3K27ac
27466157	T/C	rs1532276	* See Footnote	33	5.724864		
27466181	T/C	rs1532277	* See Footnote	33	5.724864		
27466315	T/C	rs1532278	* See Footnote	33	5.724864		
27468503	C/A	rs867230		72	5.724864	4.907665	
27469971	G/C	rs9331883	* See Footnote	11	5.724864		
27470010	G/T	-	* See Footnote	11	5.724864		
27470597	G/A	rs34109053		4	5.724864		7.681592
27471673	C/T	-	* See Footnote	135	5.724864	11.0377	7.681592
27471748	C/A	-	* See Footnote	135	5.724864	11.0377	7.681592
27472859	C/A	rs76646010			5.724864	11.0377	7.681592
27474202	A/G	rs9314349	* See Footnote		5.724864		7.681592
27474541	G/A	rs117148275	* See Footnote	14	5.724864	8.594284	7.681592
27474587	C/G	rs56025648	* See Footnote		5.724864		7.681592
27474599	C/T	rs1982229	* See Footnote		5.724864		7.681592
27474871	G/A	rs77336101	* See Footnote	133	5.724864		7.681592

ENCODE data for non-coding variants within the targeted *CLU* locus for parameters suggestive of regulatory activity. In the ENCODE data, TFBS is transcription factor binding sites, DNASE is DNaseI hypersensitivity clusters, NHA refers to normal human astrocytes (the most relevant cell type to AD available), with H3K4me1, H3K4me3 and H3K27a all being histone modifications indicative of active or potentially active DNA. A value for any of these may indicate the DNA position is in some way functionally active. *Each of these variants fell within the locations of multiple putative transcription factor binding sites. For details of which SNPs occurred at which TFBS, see Appendix section 6.1.

Table 6.3 – Association testing of non-coding variants at the *CLU* locus in imputed data

Coordinate	Ref/Alt	rsID	Info	AD MAF	Control MAF	OR (95% CI)	*p*-value
27466157	T/C	rs1532276	0.995	0.360	0.395	1.164 (1.083-1.252)	0.0009
27466181	T/C	rs1532277	0.993	0.352	0.393	1.194 (1.110-1.285)	0.0005
27466315	T/C	rs1532278	0.994	0.353	0.393	1.184 (1.101-1.274)	0.0008
27468503	C/A	rs867230	0.966	0.364	0.411	1.220 (1.126-1.321)	0.0094
27469971	G/C	rs9331883	0.193	0.001	0.001	0.988 (0.205-4.756)	0.0553
27470010	G/T	-					
27470597	G/A	rs34109053	0.952	0.267	0.244	1.129 (1.027-1.242)	0.1154
27471673	C/T	-					
27471748	C/A	-					
27472859	C/A	rs76646010	0.725	0.004	0.007	0.597 (0.345-1.034)	0.4942
27474202	A/G	rs9314349	1.000	0.371	0.382	0.954 (0.888-1.025)	0.4274
27474541	G/A	rs117148275	0.938	0.009	0.007	1.324 (0.909-1.929)	0.2835
27474587	C/G	rs56025648	0.941	0.031	0.040	0.766 (0.626-0.936)	0.9849
27474599	C/T	rs1982229	0.988	0.356	0.367	0.952 (0.884-1.025)	0.4809
27474871	G/A	rs77336101	0.939	0.014	0.018	0.780 (0.585-1.040)	0.7197

Results of association testing for the variants showing evidence of functionality in the targeted *CLU* locus using imputed data. The OR and 95% CI given correspond to the allele listed as "alt", but the MAF refers to which ever allele has the lower frequency, so these do not always correspond. Statistically significant variants are highlighted in yellow, one variant significant at the $p<0.05$ level but not withstanding correction for multiple testing highlighted in orange.

6.3. *PICALM* and the rs3851179 LD block

Given the evidence that the association signal within the *PICALM* gene tracks to the 5' and upstream region, this was the focus of the search within *PICALM*. Of the 516 non-coding variants detected within the *PICALM* target locus, 14 fell upstream of the gene and had multiple lines of evidence of falling in a functional region from the ENCODE data. These spanned a region from 102bp away from the transcription start site, to 3.25kb away. Nine of these had rs numbers already assigned, while the remaining five were novel. Details on these 14 variants and the investigation of their functionality are presented in Table 6.4, with data from the ENCODE project reported in Table 6.5. Results from the association testing using imputed data are given in Table 6.6.

The other prioritised area of interest was the rs3851179 LD block. In this area, 113 variants were identified, with 23 of those determined to be novel by Ensembl's VEP. The eight variants with multiple lines of evidence for functionality are presented in Table 6.7, with data from the ENCODE project on these variants presented in Table 6.8. Highlighted within these tables are two variants of particular interest, which fell within a small region with strong evidence of regulatory activity.

It was decided to validate these variants, and determine whether there was evidence of suggestive association with AD using TaqMan genotyping assays in a cohort of AD cases and controls. Figure 6.1 shows the results from the Sanger sequencing validation of the variants along with TaqMan genotyping assay output. Table 6.9 shows the results from the TaqMan genotyping assays, including the numbers of samples genotyped, and association testing for one of the SNPs with AD using Fisher's exact test in the TaqMan data with 1000 genomes data included to give additional control samples.

Association testing in the imputed data set was also performed on all eight variants falling in regions with the best evidence of functionality or regulatory capacity. This data is provided in Table 6.10.

Table 6.4 – Summary of variants upstream of *PICALM*

Coordinate	Ref/Alt	rsID	CRISP freq.	1kg freq.	PhyloP	Phastcons	Distance from *PICALM* (bp)	LD with GWAS SNP (r^2)	LD with GWAS SNP (D')
85780073	G/T	rs3016326	0.9999*	1.0000	-0.1086	0.9843	102	-	-
85780448	T/C	rs3016327	0.8419	0.7889	-0.2136	0	477	0.199	0.676
85780582	G/T	rs10898433	0.1006	0.1504	-1.5252	0	611	0.013	0.337
85780924	A/G	-	0.0050	-	0.9596	1.0000	953	-	-
85780962	T/C	rs188367538	0.0024	0.0013	2.1640	1.0000	991	-	-
85781279	C/G	-	0.0048	-	0.7187	0.0157	1308	-	-
85781322	C/T	rs669556	0.8563	0.8166	0.1165	0.8346	1351	0.268	0.833
85781523	C/A	rs75172533	0.0608	0.0567	-0.6663	0	1552	0.031	0.795
85781597	CA/C	rs5793180	0.8113	0.8166	-0.3652	0.0236	1627	0.268	0.833
85781599	A/G	rs11304990	0.2446	-	0.2370	0.0157	1628	-	-
85781600	G/T	-	0.2316	-	0.5381	0.0157	1629	-	-
85781600	GT/G	-	0.2699	-	0.5381	0.0157	1630	-	-
85781634	C/G	rs55886146	0.0546	0.0172	0.8994	0.9921	1663	0.015	1
85781856	C/G	-	0.0049	-	-1.4302	0	1885	-	-

Information on the variants upstream of *PICALM* including frequency estimates from CRISP and the 1000 genomes project, conservation scores from both PhyloP and Phastcons, and the distance from the transcription start site. Variants showing conservation scores above 0.5 (by either measure) are highlighted in yellow.

Table 6.5 – ENCODE data on variants upstream of *PICALM*

Coordinate	Ref/Alt	rsID	ENC_TFBS	ENC_DNASE	NHA_H3K4me1	NHA_H3K4me3	NHA_H3K27ac
85780073	G/T	rs3016326	See footnote*			28.941935	13.490137
85780448	T/C	rs3016327	See footnote*		8.547032	28.941935	13.490137
85780582	G/T	rs10898433	See footnote*	18	8.547032	28.941935	13.490137
85780924	A/G	-	See footnote*	136	8.547032	28.941935	13.490137
85780962	T/C	rs188367538	See footnote*	136	8.547032	28.941935	13.490137
85781279	C/G	-			8.547032	28.941935	13.490137
85781322	C/T	rs669556		4	8.547032	28.941935	13.490137
85781523	C/A	rs75172533			8.547032	28.941935	13.490137
85781597	CA/C	rs5793180		37	8.547032	28.941935	13.490137
85781599	A/G	rs11304990		37	8.547032	28.941935	13.490137
85781600	G/T	-		37	8.547032	28.941935	13.490137
85781600	GT/G	-		37	8.547032	28.941935	13.490137
85781634	C/G	rs55886146		37	8.547032	28.941935	13.490137
85781856	C/G	-		37	8.547032	28.941935	13.490137

ENCODE data for variants upstream of the *PICALM* gene for parameters suggestive of regulatory activity In the ENCODE data, TFBS is transcription factor binding sites, DNASE is DNaseI hypersensitivity clusters, NHA refers to normal human astrocytes (the most relevant cell type to AD available), with H3K4me1, H3K4me3 and H3K27a all being histone modifications indicative of active or potentially active DNA. A value for any of these may indicate the DNA position is in some way functionally active. *Each of these variants fell within the locations of multiple putative transcription factor binding sites. For details of which SNPs occurred at which TFBS, see Appendix section 6.2.

Table 6.6 – Association testing of variants upstream of *PICALM* in imputed data

Coordinate	Ref/Alt	rsID	Info	AD MAF	Control MAF	OR (95% CI)[b]	*p*-value
85780073	G/T	rs3016326	0.9037	0.0000	0.0075	-1.0000	0.3091
85780448	T/C	rs3016327	0.9781	0.1918	0.2044	1.0827 (0.9869 - 1.1878)	0.9138
85780582	G/T	rs10898433	0.9975	0.1329	0.1385	0.9530 (0.8603 - 1.0558)	0.4773
85780924	A/G	-	-	-	-	-	-
85780962	T/C	rs188367538	0.0413	0.0000	0.0000	-1.0000	-1.0000
85781279	C/G	-	-	-	-	-	-
85781322	C/T	rs669556	0.9994	0.1782	0.1951	1.1184 (1.0218 - 1.2241)	0.7498
85781523	C/A	rs75172533	0.9929	0.0659	0.0552	1.2069 (1.0458 - 1.3928)	0.1533
85781597	CA/C	rs5793180	0.9971	0.1744	0.1926	1.1298 (1.0318 - 1.2372)	0.7721
85781599	A/G	rs11304990	-	-	-	-	-
85781600	G/T	-	-	-	-	-	-
85781600	GT/G	-	-	-	-	-	-
85781634	C/G	rs55886146	0.9906	0.0292	0.0274	1.0663 (0.8661 - 1.3128)	0.0959
85781856	C/G	-	-	-	-	-	-

Results of association testing of the variants showing evidence of functionality upstream of *PICALM* in imputed data. The OR and 95% CI given correspond to the allele listed as "alt", but the MAF refers to which ever allele has the lower frequency, so these do not always correspond.

Table 6.7 – Variants in areas with multiple lines of evidence of functionality in the rs3851179 LD block

Coordinate	Ref/Alt	rsID	CRISP freq.	1kg freq.	PhyloP	Phastcons	LD with GWAS SNP (r^2)	LD with GWAS SNP (D')
85859598	G/A	-	0.0041	-	0.7114	0.1012	-	-
85862491	A/G	rs187016120	0.0261	0.0053	-0.5417	0	0.0090	1.0000
85862739	G/A	-	0.0054	-	0.5737	0	-	-
85863014	C/G	rs3862786	0.1233	0.1781	1.3011	1.0000	0.0475	1.0000
85863080	A/C	rs56157503	0.1306	0.1306	0.2342	0.9843	0.1051	1.0000
85863473	G/A	rs34731047	0.1004	0.0594	0.2423	0.0236	0.0350	1.0000
85863683	C/G	rs3889743	0.0429	0.0343	0.5844	0.0866	0.0249	1.0000
85863769	T/G	rs11234562	0.2100	0.203	-2.5515	0	0.1823	1.0000

Information on the variants upstream of *PICALM* including frequency estimates from CRISP and the 1000 genomes project, conservation scores from both PhyloP and Phastcons, and the distance from the transcription start site. The two yellow highlighted variants are those TaqMan genotyped to test for association with AD. . Variants showing conservation scores above 0.5 (by either measure) are highlighted in yellow.

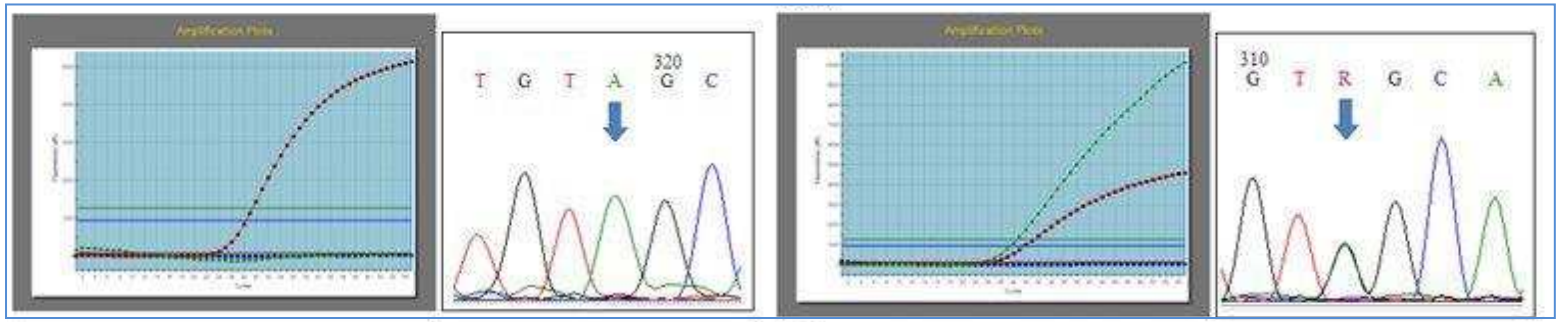
Table 6.8 – ENCODE data on variants in the rs3851179 LD block

Coordinate	Ref/Alt	rsID	ENC_TFBS	ENC_DNASE	NHA_H3K4me1	NHA_H3K4me3	NHA_H3K27ac
85859598	G/A	-	See footnote*	4	4.203475		
85862491	A/G	rs187016120	See footnote*	116		7.903122	5.130389
85862739	G/A	-	See footnote*	116	4.483236	7.903122	5.130389
85863014	C/G	rs3862786		116	4.483236	7.903122	5.130389
85863080	A/C	rs56157503	See footnote*	116	4.483236	7.903122	5.130389
85863473	G/A	rs34731047	See footnote*	116	4.483236	7.903122	5.130389
85863683	C/G	rs3889743	See footnote*			7.903122	5.130389
85863769	T/G	rs11234562	See footnote*			7.903122	5.130389

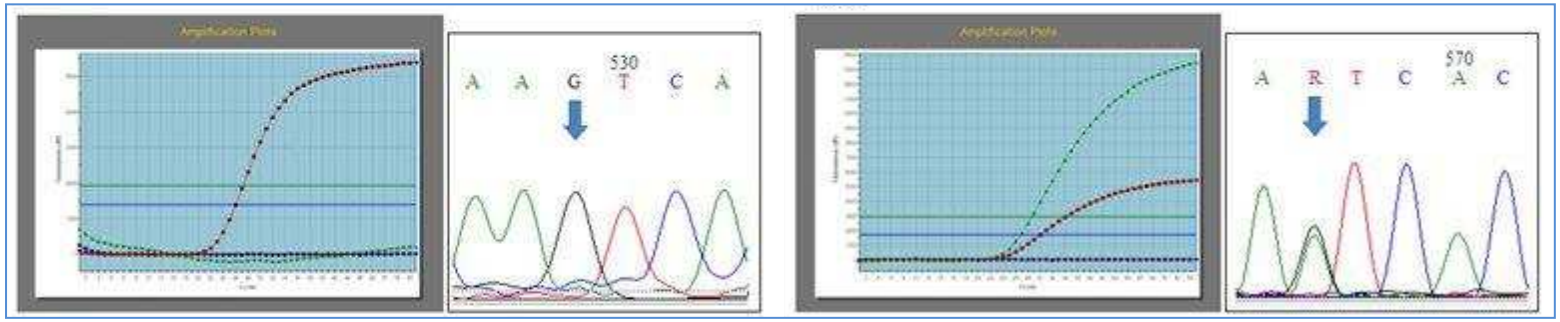
ENCODE data for variants in the rs3851179 LD block for parameters suggestive of regulatory activity. In the ENCODE data, TFBS is transcription factor binding sites, DNASE is DNaseI hypersensitivity clusters, NHA refers to normal human astrocytes (the most relevant cell type to AD available), with H3K4me1, H3K4me3 and H3K27a all being histone modifications indicative of active or potentially active DNA. A value for any of these may indicate the DNA position is in some way functionally active. *Each of these variants fell within the locations of multiple putative transcription factor binding sites. For details of which SNPs occurred at which TFBS, see Appendix section 6.3. The two yellow highlighted variants are those TaqMan genotyped to test for association with AD.

Figure 6.1 – Sanger validation and TaqMan genotyping assays of rs3851179 LD block variants of interest

rs187016120



11:85872739



TaqMan readouts and Sanger sequencing validation of SNPs at rs187016120 and 11:85872739 in each case showing wild type homozygous samples on the left, and heterozygous samples on the right. No individuals homozygous for the minor allele were detected for either SNP.

Table 6.9 – TaqMan genotyping assay results for rs3851179 LD block SNPs

		TaqMan genotype distribution		TaqMan and 1kg genotype distribution		MAF			
Variant	Genotype	Case	Control	Case	Control	Case	Control	OR (95% CI)	*p*-value
rs187016120	AA	254	204	254	579	0.78%	0.34%	2.27 (0.57-9.11)	0.259
	AG	4	0	4	4				
11:85862739	GG	246	194						
	GA	1	0						

Results from the TaqMan genotyping assay for variants 11:85862739 and rs187016120. As no minor alleles were found in the controls for either variant, association testing was not possible on this data alone. For the variant at rs187016120, 1000 genomes data was available and was combined with the genotyping data to allow a Fisher's exact test to be conducted, the odds ratio, 95% CI and *p*-value for which are shown in the table above. As the variant at 11:85862739 was novel, no extra control data was available for association testing.

Table 6.10 – Association testing of variants in areas with good evidence of functionality in the rs3851179 LD block in imputed data

Coordinate	Ref/Alt	rsID	Info	AD MAF	Control MAF	OR (95% CI)[b]	*p*-value
85859598	G/A	-					
85862491	A/G	rs187016120	0.8133	0.0024	0.0023	1.0613 (0.5213 - 2.1607)	0.7442
85862739	G/A	-					
85863014	C/G	rs3862786	0.9917	0.1697	0.1598	1.0746 (0.9784 - 1.1803)	0.4285
85863080	A/C	rs56157503	0.9880	0.1070	0.1102	0.9668 (0.8629 - 1.0833)	0.7670
85863473	G/A	rs34731047	0.9749	0.0641	0.0662	0.9675 (0.8369 - 1.1184)	0.7527
85863683	C/G	rs3889743	0.9897	0.0363	0.0315	1.1573 (0.9580 - 1.3981)	0.7080
85863769	T/G	rs11234562	0.9897	0.1850	0.1941	0.9423 (0.8597 - 1.0329)	0.7740

Results of association testing of the variants with evidence of functionality in the rs3851179 LD block in imputed data. The two yellow highlighted variants are those TaqMan genotyped to test for association with AD. The OR and 95% CI given correspond to the allele listed as "alt", but the MAF refers to which ever allele has the lower frequency, so these do not always correspond.

6.4. Discussion of non-coding variants

Non-coding variants are sometimes overlooked in the search for causative mutations for a given genotype, and indeed, the vast majority of known detrimental variants are coding ones. It is harder both to predict and to prove the functional consequences of non-coding variants, and as there is such a vast quantity in the human genome, it is harder to know where to begin the search.

Over 850 non-exonic variants were detected in the NGS data of the four targeted loci (taking *CR1* variants from project two). To narrow down the quantity of variants to focus on, it was decided that only the non-exonic variants within *CLU*; to the 5' region of the *PICALM* gene; and in the rs3851179 LD block would be investigated.

Several published studies have sought to characterise the source of the association signal within *CLU*, generally focussing on exonic regions, and consistently, no coding variants which can explain the GWAS signal have been found (Guerreiro et al. 2010; Bettens et al. 2012; Ferrari et al. 2012). Given the assumption that the GWAS SNP is tagging a variant or variants implicated in AD risk at the *CLU* locus, it is logical to turn the search to non-coding variants.

Attempts at tracking down the source of *PICALM*'s association signal have been less extensive than those for *CLU*, but again, no coding variants to date have been found to explain *PICALM*'s involvement in AD, so again, perhaps the regulatory regions hold the answers. The fact that the initial GWAS SNP fell so far from the actual gene (~88.5kb upstream) (Harold et al. 2009) implies the alteration in AD risk attributable to this region might be regulatory in nature, particularly from the 5' end of the gene (Sleegers et al. 2010).

For the *CR1* association signal, there is evidence that suggests the alteration in AD risk may stem from the different isoforms of the protein, with the S allele associated with an increased risk of AD (Brouwers et al. 2012). Given that there are also many coding variants within *CR1* which have not been thoroughly investigated for their potential functional role in the condition, *CR1*'s non-coding variants were not seen as a priority for pursuit. Since large quantities of non-coding DNA in the *CR1* region were not adequately covered by the sequencing, any analysis at this stage would be piecemeal at best.

As was done for the assessment of coding variants, a two pronged strategy was adopted, using evidence of functionality, as well as association testing in the imputed case-control cohort. This time, however, only variants which had three separate lines of evidence of functionality from ENCODE's data were tested for association, since the

large number of non-coding variants would have made the necessary correction for multiple testing prohibitively harsh, given the size of the cohort and the limitations of imputation, as discussed in the previous chapter.

Five different parameters from ENCODE were used in this assessment of functionality – TFBS, DNASE, and three histone modifications, H3K4me1, H3K4me3 and H3K27ac. The data given for these latter three is based on the normal human astrocyte cell line, since this was the most relevant to AD of all the available cell lines. The first, TFBS, or transcription factor binding sites, gives details of any conserved TFBS, as determined by ChIP-seq, at the position of each variant. Any SNPs falling within these binding sites may be affecting the affinity of the sites, so could potentially disrupt the regulatory role the transcription factor normally fulfils. The binding motifs for transcription factors are normally degenerate, so the variants may or may not affect the site, but this would need to be determined. The next, DNASE, refers to DNAseI hypersensitivity clusters, based on 125 cell types. Regulatory regions and particularly promoters tend to be DNAseI hypersensitive, as the DNA here is not tightly wound around histones and being exposed, it is vulnerable to enzymatic cleavage. Variants falling within hypersensitive sites may therefore be falling in regions of regulatory activity, and may potentially be disrupting the normal regulatory function of the region, so could be of interest. Histone marks are specific modifications to histone proteins which contribute to gene regulation by altering the accessibility of the region to transcriptional activity. The levels given are of specific histone marks (H3K4me1, which is associated with enhancers; H3K4Me3, which is associated with active (or potentially active) promoters; and H3K27Ac, which is associated with active transcription, possibly via preventing the spread of another, repressive histone mark, H3K27Me3), at the sites of the variants detected, determined by ChIP-seq in the ENCODE data.

None of these parameters individually are conclusive evidence a variant could be functional, however, when multiple lines of evidence converge at the position of a certain variant, this can suggest functionality, and can flag up those variants worthy of further investigation.

Out of the 75 non-coding variants detected within the targeted *CLU* locus, 15 met the filtering criteria ascribed, having at least 3 lines of evidence of function from ENCODE. 12 of these 15 variants were imputed in the GWAS dataset, although one (rs9331883) had a poor info score and a vast range in its 95% CI, thus despite a seemingly suggestive *p*-value of 0.055, it will be disregarded as a potential artefact of imputation.

Three variants (rs1532276, rs1532277 and rs1532278) showed an association with AD (*p*-values of 0.0009, 0.0005, and 0.0008 respectively), which remained significant after Bonferroni correction for the twelve multiple tests being conducted (0.05/12 = 0.0042). rs867230 fell short of this stringent significance threshold, but at p = 0.0094, it is certainly suggestive of significance. All four had strong info scores (>0.96), and all were common variants (MAFs in controls all around 0.4). Each of the variants appeared to convey a modest increase in AD risk, with ORs ranging from 1.16 to 1.22. All of these variants are in strong LD with the original GWAS SNP, rs11136000, which itself had a *p*-value of 0.00089, and an OR of 1.172 (95% CI 1.090-1.260). This signifies these variants do not represent new, independent association signals. Could any of these variants explain the association signal in terms of function, or are these merely further tag SNPs for the same elusive source of the original signal?

Two of the variants (rs1532277 and rs1532278) had more significant *p*-values than the GWAS SNP, which could suggest a causative role in the GWAS signal, but the difference was only marginal. That said, the GWAS SNP was directly genotyped in the majority of samples used for imputation, while the other variants were purely imputed, which despite the strong info scores, has an inherently lower power to detect associations, so perhaps with direct genotyping these variants would further exceed the significance of rs11136000.

The three significantly associated variants, rs1532276, rs1532277 and rs1532278 all fall within the third intron of the gene, while the suggestive variant, rs867230, falls within intron 1. None of the variants showed strong evidence of being in a particularly conserved region, from the PhyloP and Phastcons scores. These are measures of conservation at a specific position (PhyloP) or across a small region, taking in to account neighbouring sites (Phastcons). The scores given for PhyloP reflect –log *p*-values under the null hypothesis of neutral evolution, with positive scores reflecting conservation. The score given for Phastcons is a compressed conservation score reflecting the likelihood of conservation in an area (on a scale of 0-1, with 1 being conserved). A combination of a large positive PhyloP score and a Phastcons score approaching 1 therefore gives good evidence a variant is affecting an area of high conservation. Such variants are more likely to be damaging functionally, as conservation implies evolutionary restraint and thus potential functional significance.

The data from ENCODE for the four variants suggests they all fall within transcriptionally active DNA regions. All show moderate DNASE scores, and all are positive for H3K4me1 histone marks. The suggestive variant, rs867230 is also positive for the H3K4me3 histone mark. rs1532276, rs1532277 and rs1532278 all fall within potential TFBS

according to the ENCODE data. Experimental characterisation would be necessary to confirm whether these TFBS are actually actively utilised in the regulation of the *CLU* gene, and whether the presence of the detected variants has a detrimental effect on this process. The variants which show both evidence of association and evidence of functionality are the ones with highest priority for further pursuit.

Of the 14 variants in the region immediately 5' of the *PICALM* gene with multiple lines of evidence of functionality from ENCODE, eight were successfully imputed. None of the variants showed a significant association with AD at the $p<0.05$ level, let alone a Bonferroni corrected level of significance. One variant showing evidence of suggestive significance was rs55886146 ($p=0.0959$). This is a rare variant (1000 genomes project EUR MAF 0.017), with strong evidence of conservation (PhyloP 0.8994, Phastcons 0.9921), showing a high level of LD with the original GWAS SNP (D' 1). Because of the low MAF, it may be that the association test in the imputed data lacked power (due to the issues imputing rare variants covered in the previous chapter). The variant showed evidence of falling within a region of DNAseI hypersensitivity, and was positive for all three histone marks assessed, so is potentially falling within a regulatory region, although without falling in a predicted TFBS, how this regulatory effect could be mediated is unclear.

There were two further variants of potential interest. These fell within multiple TFBS in the ENCODE data, had high evidence of falling in a DNAseI hypersensitive site, and were positive for all three histone marks. These two variants both fell in conserved DNA (PhyloP 2.1640 and 0.9596 respectively, Phastcons 1 for both), so had strong evidence suggesting regulatory functionality. One of the variants, rs188367538, was classed as being imputed, but had a poor info score, so was not adequately tested for association with AD by our method. The other, at position 11:85780924, was a novel variant so could not be imputed. Direct genotyping of the variants would be necessary to adequately address whether they are associated with AD, while further *in silico* and experimental functional investigations would be needed to elucidate if and how they are related to AD pathology.

Eight variants within the rs3851179 LD block showed three or more lines of evidence of functionality in the ENCODE data. When the project began, a choice was made whether to include simply *PICALM*, the gene tagged by the GWAS association signal, or whether to also include the region in which the GWAS SNP fell, in an LD block ~88.5kb upstream of the gene. Analysis of the information available, particularly ENCODE data, for the LD block suggested that the region may contain some regulatory activity. Of particular interest was a ~500bp span of DNA with high DNAseI hypersensitivity, multiple TFBS, and evidence of high levels of the three histone marks associated with regulatory activity

considered (H3K4me1, H3K4me3 and H3K27ac). For these reasons, the LD block was also targeted in the NGS project. When the variants were identified in the NGS data, two (rs187016120 and the variant at position 11:85862739) fell within this ~500bp region, with high levels of evidence of functionality. An MSc student in Molecular Genetics and Diagnostics, Ng See May, undertook validation of these two variants via Sanger sequencing, and used TaqMan assays to genotype the variants in a case-control cohort, to facilitate association testing. The novel variant, at position 11:85862739 was successfully validated in the sample from the NGS data, but no further alternative alleles were detected in the 440 case and control samples genotyped. Since it has also not been reported in the 1000 genomes project data, the variant is certainly rare, and may be a private mutation, whose relationship with AD would be difficult to assess. The other variant, rs187016120, was also successfully validated, and was found in four case samples out of 258 in the TaqMan data, and in none of the 204 control samples. Since no alternative alleles were found in the control samples, it was not possible to test for association with AD in the TaqMan data alone, but combining the TaqMan case data with 1000 genomes data for the variant enabled this. 379 control samples from the 1000 genomes EUR population had data available for this variant, including four heterozygotes. When this was tested for association with AD in the TaqMan and 1000 genomes data, no evidence of association was found, and the imputed data showed no suggestion of association either. Despite all of the suggestions of functional activity, this SNP, and indeed the other potentially functional SNPs within the region showed no evidence of association with AD, even when in strong LD with the GWAS SNP.

This chapter has detailed the prioritisation of non-coding variants for follow up research. Those falling within the areas thought most likely to harbour functional variants (*CLU*, the region to the immediate 5′ of *PICALM*, and the rs3851179 LD block) were assessed for likely functionality using *in silico* resources, and any with multiple lines of evidence of functionality were tested for association with AD in the imputed data set. Three intronic variants within the targeted *CLU* region occurred at the sites of ENCODE database TFBS, and were significantly associated with AD, so were classed as a high priority for further research. A fourth showed suggestive significance but did not withstand correction for multiple testing, and had less compelling evidence of functionality, so was classed as a low priority for follow up. Only one variant in both of the regions 5′ of *PICALM* showed suggestive significance, but had limited evidence of functionality. Two variants upstream of *PICALM* coincided with the sites of ENCODE database TFBS. These were both inadequately tested for association with AD (one was novel so could not be imputed, and the other was poorly imputed), so were determined to be a high priority for future work.

7. General Discussion

7.1. Summary of main findings

There are take home messages from each of the chapters presented here, as well as from the project as a whole. Chapter 3 – Data analysis provided a robust pipeline for the analysis of pooled sequencing data. The development of the pipeline during the first sequencing project was crucial for the handling of the second, which had a significantly increased volume of data and thus was extremely time consuming to process. It also allowed the assessment of the quality of the data produced, and the reliability of variant calls, which was generally good, although problems of coverage in the *CR1* region were identified. Chapter 4 – Sanger validation was based on a publication which highlighted some of the issues caused by indels and mononucleotide repeats for NGS (Lord et al. 2012). Chapters 5 and 6 (Exonic and Non-exonic variants) present a new method for prioritising variants based on combined evidence of functionality from *in silico* resources, and evidence of association from imputed data. The utilisation of these steps ensure only variants with the best evidence of involvement in AD risk are pursued with expensive and time consuming functional experimentation and direct genotyping. Several coding and non-coding variants within the three genes were identified as worthy of follow up research, which will be discussed in section 7.2. Next steps. The methods used in general provide a cost-effective framework for following up GWAS identified risk loci, in AD as well as other conditions. The utilisation of pooled targeted sequencing has been shown to be capable of identifying variants, significantly reducing the costs involved in individual sequencing. Whilst this project cannot claim to give a comprehensive catalogue of rare variation in the loci targeted, given the modest sample size, it contributes to the general body of knowledge of AD genetics, and may be a small stepping stone on the long path to understanding AD aetiology, and turning that understanding into advances for those affected by the condition.

7.2. Next steps

Out of more than 1000 variants detected within the NGS data, Table 7.1 shows the 16 variants deemed most worthy of further research, the reasons behind this prioritisation, and the vein in which follow up research would proceed. Each of the variants was given a level of priority for follow up research. Those with no evidence of functionality in the investigations conducted were deemed to be low priority, irrespective of their level of association with AD. Variants which did show evidence of functionality, but had been successfully imputed and not been found to be associated with AD were classed as medium priority. The SNPs were classed as high priority if they showed evidence of functional effects, directing the course of follow up studies, as well as evidence of association (significant or suggestive) in the imputed data. Rare variants showing suggestive association may be particularly promising

candidates for involvement in AD, given rarity is known to reduce power in imputed data. Also classed as high priority were those that showed evidence of functionality but had not been adequately tested for association with AD, either because they could not be imputed or were poorly imputed.

The follow up research for the variants depends on what type of functional evidence was found for them. The list included four missense mutations, all of which were predicted by Polyphen-2 to be possibly or probably damaging. While Polyphen-2 normally compares favourably with other mutation prediction programs when tested alongside experimental functional data (Di et al. 2009; Adzhubei et al. 2010; Wei et al. 2010), no *in silico* prediction programs are 100% accurate. With estimated correct prediction rates ranging from 91.7% (Zou et al. 2011) to 66.7% (Di et al. 2009), a Polyphen-2 deleterious prediction alone is not confirmation a variant is functional. Experimental confirmation is necessary to clarify the accuracy of the predictions for these SNPs. Often, to achieve this validation, variant and wild type versions of the protein are expressed in a relevant cell line, and the function of the protein (e.g. enzymatic activity) is measured (Brunham et al. 2005; Di et al. 2009; Zou et al. 2011). However, for *PICALM* and *CR1*, it is not known what aspect of the protein's function is actually implicated in AD pathology, so it would not be obvious what parameter to measure. Levels of extracellular Aβ may provide a useful measure, but it would be unclear whether a lack of effect indicated that the wild-type and variant sequences had the same effect on this level, or whether it indicated that the *PICALM* or *CR1* proteins do not affect AD risk via this mechanism. Without a specific parameter to measure, a more generalised approach could be adopted. Transgenic mice can be created with specific mutations using homologous recombination. Replacing the wild-type gene with sequence containing the variant of interest, and monitoring the effect on the resultant animals in terms of brain development and cognitive function when compared to the wild type would allow a method to assess the consequences of the variants without knowing what specific molecular mechanism is involved. However, such experimentation is highly specialised and expensive to conduct, thus it could not be justified without stronger evidence supporting these variants roles in AD.

Table 7.1 – Variants of interest and future research directions

Type	Gene	Variant	Associated	Functional	Priority	Follow up functional research
Exonic	*CLU*	rs7982	Yes	No	Low	
		rs9331950	Yes	No	Low	
		rs9331949	Suggestive (rare)	No	Low	
		rs9331942	Suggestive (rare)	No	Low	
	PICALM	11:85707933	Not imputed	Missense, probably damaging SRSF1 site disrupted	High	Investigate variant protein Investigate splicing
		rs592297	No	SRSF1 site disrupted	Medium	Investigate splicing
		rs76719109	No	SRSF2 site disrupted	Medium	Investigate splicing
	CR1	rs41274770	Suggestive (rare)	Missense, possibly damaging	High	Investigate variant protein
		rs61734514	Suggestive (rare)	Missense, probably damaging SRSF1 site disrupted	High	Investigate variant protein Investigate splicing
Non-coding	*CLU*	rs1532276	Yes	TFBS	High	Investigate TFBS
		rs1532277	Yes	TFBS	High	Investigate TFBS
		rs1532278	Yes	TFBS	High	Investigate TFBS
		rs867230	Suggestive	No	Low	
	PICALM	rs55886146	Suggestive (rare)	No	Low	
		rs188367538	Low info score (rare)	TFBS	High	Investigate TFBS
		11:85780924	Not imputed	TFBS	High	Investigate TFBS

Summary information on the variants with the strongest evidence of links to AD, either from association testing or the *in silico* functional prediction programs used. Ideas for further research on these variants designed to clarify associations and validate the predictions of the *in silico* resources are listed, along with the level of priority for follow up for each of the variants. For strengths of associations (ORs and *p*-values) see Table 5.2 for exonic SNPs and Tables 6.3 and 6.6 for *CLU* and *PICALM*'s non-exonic variants respectively.

For the four variants which showed evidence of affecting the binding sites of SR-splicing proteins, an experimental approach utilising a mini-gene vector system would be adopted. This involves the introduction of wild-type and variant versions of the affected exon in to mini-gene vectors, and transfection in to a relevant cell line. The mRNA generated from the vectors can then be analysed (e.g. PCR and gel electrophoresis to assess large scale effects, Sanger sequencing to confirm or to check for smaller scale alterations), and any differences in the splicing of the exon would indicate that that variant does indeed have an effect on splicing. These experiments are currently on going.

Five of the variants fell within TFBS according to the ENCODE data. To follow up these variants further *in silico* analyses would be wise to guide experimental characterisation attempts. It would need to be established which of the putative disrupted sites were actually utilised in the regulation of the genes in question, and how the variants would affect these, since many TFBS are degenerate and the variants may be tolerated even if they fall within actively used sites. There are multiple other tools and databases available for the investigation of TFBS, such as TRANSFAC (Matys et al. 2006) and JASPAR (Sandelin et al. 2004), and collation of the predicted binding sites from each of these would strengthen the evidence. Several potential experimental approaches would be appropriate following thorough *in silico* investigations. EMSA (electrophoretic mobility shift assay) allows an *in vitro* method for assessing TF binding. The wild-type and variant sequences would need to be labelled, and incubated with relevant cell nuclear extracts (e.g. neuronal cells). When subjected to electrophoresis, labelled DNA alone will show a different motility than DNA bound by TFs, and any differences in pattern between the sequences could indicate disruption of TFBS. Alternatively, the wild-type and variant sequences could be cloned upstream of a reporter gene in a promoter construct plasmid and be transfected in to a relevant cell line, with any differences in reporter gene expression indicative of differential activity (de Vooght et al. 2009).

Two of the prioritised variants were novel, and thus could not be imputed, and a third was poorly imputed within our dataset, so they were not adequately tested for association with AD via the methodology used. Genotyping of the variants (e.g. using TaqMan and/or KASP assays) could be conducted to facilitate association testing. However, as all of these variants are rare, the number of samples needed to give the study sufficient power would be very high, and this would therefore be very costly. For these variants, experimental characterisation of functional effects would be more cost effective as a first step rather than genotyping all in a large number of samples. Any which showed compelling evidence of functionality could then be followed up by genotyping.

Direct genotyping would also be useful for those variants showing suggestive associations with AD in our imputed data. The greater power afforded by this

method over imputation would establish which variants were genuinely associated with the disease. Particularly for the rarer variants, on which the limitations of imputation would have had the greatest effect, this would allow thorough association testing, although the rarer variants would also require larger sample sizes to give sufficient power. Even for the variants which were associated with AD in the imputed data, replication in an independent cohort is important for any such genetic findings, and even more so when the associations are in imputed data, so genotyping in a case-control cohort would also be useful here, but this is not the top priority for future work.

7.3. AD genetics - update

Our results suggest some exciting potential directions for future work in exploring how these three genes relate to AD, but there are still no real answers as to how these genes are linked to the pathology of the condition, and which variants are responsible for these links. Four years on from the publication of the first two major GWAS, the source of the three association signals detected remains largely elusive.

As mentioned previously, there is some evidence that the *CR1* GWAS association signal may stem from the different alleles of the gene, essentially a copy number variation in the number of binding sites in the protein (Brouwers et al. 2012). The larger S allele is associated with an increased risk of the condition, while the smaller F allele appears to be protective. This is a contradiction to many of the hypothesised ways in which *CR1* was thought to be related to AD pathology, such as the peripheral sink hypothesis. If *CR1* was related to AD through its capacity to act as a peripheral sink, removing Aβ from the circulation, facilitating further removal of Aβ from the brain, it would be expected that the larger isoforms would be more efficient at this process, and thus would be protective. Since this is not the case, it strengthens arguments for *CR1*'s relationship with AD stemming from effects on neuroinflammation, or perhaps, as suggested by Bralten et al.'s research, by affecting the structure of the brain much earlier in life, affecting the brain's ability to cope with AD processes when they occur later on (Bralten et al. 2011).

For *CLU* and *PICALM*, the causal source underlying the GWAS signals remains a mystery. Although attempts to track down the underlying causative variants have been more extensive for *CLU* (Harold et al. 2009; Guerreiro et al. 2010; Bettens et al. 2012; Ferrari et al. 2012) than for *PICALM* (Harold et al. 2009; Ferrari et al. 2012; Schnetz-Boutaud et al. 2012), neither have been particularly fruitful, with no real explanatory causal variation found.

The lack of success in tracking down the source of the original three association signals is not a reflection of a lack of progress in the AD genetics field in general in the past four years. Indeed, the extensive exome and other next generation sequencing projects embarked on in the last three years are

beginning to reach fruition, and publication of their results in AD and other relevant traits are beginning to appear, identifying new variants and new genes which appear to be involved in AD risk. Pottier et al. conducted exome sequencing in 14 autosomal dominant early onset AD cases with no known *APP*, *PSEN1* or *PSEN2* mutations, as well as a further 15 replication cases (Pottier et al. 2012). 29 previously unknown variants were identified within the cases in the *SORL1* gene, 7 of which were predicted to have pathogenic effects. *SORL1*'s protein product has been shown to be involved in Aβ production (Rogaeva et al. 2007). The gene had already been linked to the late onset form of AD (Rogaeva et al. 2007), which has recently received replication in several independent studies (Jin et al. 2013; Lambert et al. 2013; Miyashita et al. 2013; Wen et al. 2013). Using exome sequencing and genotype imputation in MCI patients, Nho et al. sought coding variants associated with rate of hippocampal volume loss in 16 patients, 8 of whom showed rapid rates of atrophy while 8 showed slow or steady rates of atrophy (Nho et al. 2013). Three variants were found that were present in at least 6 of the rapid atrophy samples, but absent from the slow atrophy group in the genes *HYAL4*, *PARP1*, and *CARD10*, the latter two of which were predicted to be damaging by Polyphen-2. Genetic variation within *PARP1* and *CARD10* was found to be associated with the rate of neurodegeneration in the hippocampus in *APOE* ε3/ε3 subjects (Nho et al. 2013). *CARD10* is involved in regulation of apoptosis and inflammation (Wang et al. 2001; Nho et al. 2013), while *PARP1* has roles in a number of cellular processes, such as DNA repair, cell proliferation and cell death (Menissier de Murcia et al. 2003; Nho et al. 2013). Its protein product had previously been reported to show enhanced activity in AD affected brains (Love et al. 1999).

Following extensive exome and genome sequencing, imputation and genotyping, two independent studies were simultaneously published, reporting a number of variants within the *TREM2* gene to be associated with AD (Guerreiro et al. 2013; Jonsson et al. 2013). Of particular interest was SNP rs75932628 (encoding the missense mutation R47H, which was predicted to be damaging by Polyphen-2). This was the most significantly associated variant in both studies. Homozygous loss of function of the *TREM2* gene is known to cause the rare Nasu-Hakola disease, a severe condition characterised by bone cysts and early onset dementia resulting in premature death (Bianchin et al. 2004). Within the brain, *TREM2*'s major site of expression is on microglia (Sessa et al. 2004), it is thought to be a mediator of inflammation in the brain, and may be involved in Aβ clearance (Piccio et al. 2007; Takahashi et al. 2007; Frank et al. 2008).

Recently, the largest AD GWAS to date was published, comprised of a meta-analysis of GWAS in individuals of European ancestry (Lambert et al. 2013). The two stage approach saw 7,055,881 SNPs genotyped or imputed in 17,008 AD cases and 37,154 controls in stage one, with 11,632 SNPs genotyped in a second independent cohort consisting of 8572 AD cases and 11,312 controls. 19 loci reached genome wide significance in the combined data set, of which 11

represented new AD risk loci (*HLA-DRB5–HLA-DRB1*, *PTK2B*, *SORL1*, *SLC24A4-RIN3*, *INPP5D*, *MEF2C*, *NME8*, *ZCWPW1*, *CELF1*, *FERMT2* and *CASS4*) (Lambert et al. 2013). Many of these genes fit with previously identified AD pathways, such as the immune response and Aβ metabolism, with potential new pathways including cytoskeletal function, axonal transport and hippocampal synaptic function (Lambert et al. 2013). Use of the methods presented here, with pooled NGS to detect variants in coding and non-coding regions, and a combination of *in silico* functional analyses and imputed association testing to prioritise variants of interest would be an advisable way to begin to investigate these gene's roles in AD.

So how much do we now know about the genetics of Alzheimer's disease, and what is there still to find out? Recently, Ridge et al. estimated the phenotypic variance in AD explained by common variants in the human genome to be 33.1%, and that 7.78% of that was attributable to *APOE* and the nine AD associated genes from GWAS prior to the Lambert meta-analysis, leaving 25.3% to be explained by as yet unidentified common variants (Ridge et al. 2013). Although the 11 new loci found in the recent meta-analysis (Lambert et al. 2013) will also contribute to the explained variance, the effect sizes of the new variants were generally smaller than those for the nine previously known loci, so a large proportion of missing heritability still remains. Part of this discrepancy is likely to be due to the variants identified by GWAS not being the actual causative variants, but imperfect tags for them, which leads to an underestimation of the effect size. In Crohn's disease (CD), over 70 loci associated with the condition have been found using GWAS, but only around 23.3% of the condition's heritability is explained by these. However, it was reported that at one of the loci (*NOD2)*, the most significant SNP in the GWAS explained around 0.8% of CD risk, but three known coding mutations within the gene accounted for 5% (Franke et al. 2010). This is over six times greater than the value for the GWAS tag SNP alone. If a similar pattern is true in some of the known AD risk loci, this could render the total heritability explained by the known genes higher than current estimates suggest. Identifying the causal variants underlying the GWAS signals would enable a true assessment of the loci's contribution to AD risk, but this is not proving to be an easy task. Other potential sources of missing heritability in AD include: unknown associated variants at known loci; other as yet unknown loci, which could be identified through further GWAS and meta-analyses (such as the 11 new loci Lambert et al. identified (Lambert et al. 2013)) or sequencing based investigations (e.g. the *TREM2* locus); and epistatic interactions between variants or genes.

Despite the vast quantity of disease associated loci that have been detected since the advent of GWAS, few causal variants underlying these association signals have been definitively established. Many of the attempts to track down causal variants underlying the AD GWAS signals have focussed on coding regions only, and have largely proved unfruitful (Guerreiro et al. 2010; Bettens et al. 2012; Ferrari et al. 2012). Bettens et al. did find evidence of rare variants within *CLU* impacting on AD risk, but these were independent of the GWAS

SNP (Bettens et al. 2012). This could be indicative that the GWAS SNPs are actually tagging non-coding, regulatory variants rather than coding ones. In order to find such variants, projects such as the one presented here will be important, focussing not just on coding variants but the entire gene and its surrounding regulatory regions. Assessing the involvement of such variants is less easy than assessing the impact of coding ones, but if coding variants are not responsible, then we have no choice but to pursue non-coding ones. It is also possible that other phenomena, such as the CNV underlying at least in part the *CR1* GWAS signal (Brouwers et al. 2012), may play a greater role than anticipated.

As for the common disease, common variant hypothesis versus the multiple rare variants hypothesis presented in the introduction, which has been an ongoing debate since the inception of GWAS, the answer appears to be both. The most extensively sequenced of the three genes from the first GWAS studies is *CLU*, and within *CLU*, it appears that there are multiple rare variants of large impact which affect AD risk (Bettens et al. 2012), but that these are independent of the common association signal found by GWAS, which is likely to be tagging regulatory variants. Perhaps most disease associated loci contain variants both common and rare, with varying levels of impact. Classing variants as "common" or "rare" creates a false dichotomy, over-simplifying a spectrum of different allele frequencies into two groups, and the same could be said for high and low impact classifications.

The aetiology of AD is clearly very complex, and as it is a disease of old age, this is exacerbated – genes have a long time to interact, environmental factors have longer to act, interactions between genes and environmental factors whose exposure levels may be changing all the time may all contribute. Given this immense complexity, it seems unlikely the full aetiology of AD will ever be completely understood, but progress is being made, and with this increase in knowledge comes an improvement in prospects for treatments that could actually make a difference to individuals suffering from this devastating condition.

URLs:

UCSC Virtual PCR - http://genome.ucsc.edu/cgi-bin/hgPcr

UCSC - http://genome.ucsc.edu/

UCSC liftOver - http://genome.ucsc.edu/cgi-bin/hgLiftOver or downloadable from

http://hgdownload.cse.ucsc.edu/admin/exe/

UCSC Tables - http://genome.ucsc.edu/cgi-bin/hgTables?command=start

ENCODE @ UCSC - http://genome.ucsc.edu/ENCODE/

ECR browser - http://ecrbrowser.dcode.org/

dbSNP - http://www.ncbi.nlm.nih.gov/SNP/

Source Bioscience - http://www.sourcebioscience.com/

eArray - https://earray.chem.agilent.com/earray/

FastQC - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

MOSAIK aligner - http://code.google.com/p/mosaik-aligner/

BFAST - http://bfast.sourceforge.net

SAMtools - http://samtools.sourceforge.net/

SAMStat - http://samstat.sourceforge.net/

IGV - http://www.broadinstitute.org/igv/

Syzygy - http://www.broadinstitute.org/software/syzygy/

CRISP - https://sites.google.com/site/vibansal/software/crisp

Impute2 File Formats:
http://www.stats.ox.ac.uk/~marchini/software/gwas/file_format.html

Impute2 - http://mathgen.stats.ox.ac.uk/impute/impute_v2.html#home

Plink - http://pngu.mgh.harvard.edu/purcell/plink/

Gtool - http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html

VCFtools - http://vcftools.sourceforge.net/

Tabix - http://samtools.sourceforge.net/tabix.shtml

SIFT - http://sift.jcvi.org/

Polyphen - http://genetics.bwh.harvard.edu/pph2/

ESEfinder - http://rulai.cshl.edu/cgi-bin/tools/ESE3/esefinder.cgi?process=home

BDGP - http://www.fruitfly.org/seq_tools/splice.html

NetGene2 - http://www.cbs.dtu.dk/services/NetGene2/

TargetScan - http://www.targetscan.org/

PITA - http://genie.weizmann.ac.il/pubs/mir07/mir07_prediction.html

Quanto - http://hydra.usc.edu/gxe/

HapMap - http://hapmap.ncbi.nlm.nih.gov/

Haploview - http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/haploview

Repeat Masker - www.repeatmasker.org

RepBase - http://www.girinst.org/repbase/

GATK - http://www.broadinstitute.org/gatk/

Ensembl - http://www.ensembl.org/index.html

Ensembl's VEP - http://www.ensembl.org/tools.html

1000 genomes project data site - http://www.1000genomes.org/data (for information), ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/ (for data access)

Primer3 - http://bioinfo.ut.ee/primer3-0.4.0/

NGRL Manchester SNP check - https://ngrl.manchester.ac.uk/SNPCheckV3/snpcheck.htm

Chromas lite - http://technelysium.com.au/?page_id=13

SNPtest - https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html

References

. "Agilent's eArray Online Bait Design Program." from https://earray.chem.agilent.com/earray/.

. "Alzheimer's Society - Alzheimer's Statistics." from http://alzheimers.org.uk/site/scripts/documents_info.php?documentID=100.

Abraham, R., V. Moskvina, R. Sims, P. Hollingworth, A. Morgan, et al. (2008). "A genome-wide association study for late-onset Alzheimer's disease using DNA pooling." BMC Med Genomics **1**: 44.

Adzhubei, I. A., S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, et al. (2010). "A method and server for predicting damaging missense mutations." Nat Methods **7**(4): 248-249.

Aisen, P. S., S. Gauthier, S. H. Ferris, D. Saumier, D. Haine, et al. (2011). "Tramiprosate in mild-to-moderate Alzheimer's disease - a randomized, double-blind, placebo-controlled, multi-centre study (the Alphase Study)." Arch Med Sci **7**(1): 102-111.

Aisen, P. S., K. A. Schafer, M. Grundman, E. Pfeiffer, M. Sano, et al. (2003). "Effects of rofecoxib or naproxen vs placebo on Alzheimer disease progression: a randomized controlled trial." JAMA **289**(21): 2819-2826.

Andersen, C. L., T. Schepeler, K. Thorsen, K. Birkenkamp-Demtroder, F. Mansilla, et al. (2007). "Clusterin expression in normal mucosa and colorectal cancer." Mol Cell Proteomics **6**(6): 1039-1048.

Anderson, R., J. C. Barnes, T. V. Bliss, D. P. Cain, K. Cambon, et al. (1998). "Behavioural, physiological and morphological analysis of a line of apolipoprotein E knockout mouse." Neuroscience **85**(1): 93-110.

Andrews, R. G., B. Torok-Storb and I. D. Bernstein (1983). "Myeloid-associated differentiation antigens on stem cells and their progeny identified by monoclonal antibodies." Blood **62**(1): 124-132.

Antonell, A., A. Llado, J. Altirriba, T. Botta-Orfila, M. Balasa, et al. (2013). "A preliminary study of the whole-genome expression profile of sporadic and monogenic early-onset Alzheimer's disease." Neurobiol Aging **34**(7): 1772-1778.

Antonin, W., C. Holroyd, R. Tikkanen, S. Honing and R. Jahn (2000). "The R-SNARE endobrevin/VAMP-8 mediates homotypic fusion of early endosomes and late endosomes." Mol Biol Cell **11**(10): 3289-3298.

Arnaout, M. A., N. Dana, J. Melamed, R. Medicus and H. R. Colten (1983). "Low ionic strength or chemical cross-linking of monomeric C3b increases its binding affinity to the human complement C3b receptor." Immunology **48**(2): 229-237.

Bacle, F., N. Haeffner-Cavaillon, M. Laude, C. Couturier and M. D. Kazatchkine (1990). "Induction of IL-1 release through stimulation of the C3b/C4b complement receptor type one (CR1, CD35) on human monocytes." J Immunol **144**(1): 147-152.

Baig, S., S. A. Joseph, H. Tayler, R. Abraham, M. J. Owen, et al. (2010). "Distribution and expression of picalm in Alzheimer disease." J Neuropathol Exp Neurol **69**(10): 1071-1077.

Bailey, R. W., A. K. Dunker, C. J. Brown, E. C. Garner and M. D. Griswold (2001). "Clusterin, a binding protein with a molten globule-like region." Biochemistry **40**(39): 11828-11840.

Bansal, V. (2010). "A statistical method for the detection of variants from next-generation resequencing of DNA pools." Bioinformatics **26**(12): i318-324.

Bansal, V., R. Tewhey, E. M. Leproust and N. J. Schork (2011). "Efficient and cost effective population resequencing by pooling and in-solution hybridization." PLoS ONE **6**(3): e18353.

Barrett, J. C., B. Fry, J. Maller and M. J. Daly (2005). "Haploview: analysis and visualization of LD and haplotype maps." Bioinformatics **21**(2): 263-265.

Barrett, L. W., S. Fletcher and S. D. Wilton (2012). "Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements." Cell Mol Life Sci **69**(21): 3613-3634.

Beecham, G. W., E. R. Martin, Y. J. Li, M. A. Slifer, J. R. Gilbert, et al. (2009). "Genome-wide association study implicates a chromosome 12 risk locus for late-onset Alzheimer disease." Am J Hum Genet **84**(1): 35-43.

Beffert, U., M. Danik, P. Krzywkowski, C. Ramassamy, F. Berrada, et al. (1998). "The neurobiology of apolipoproteins and their receptors in the CNS and Alzheimer's disease." Brain Res Brain Res Rev **27**(2): 119-142.

Belbin, O., M. M. Carrasquillo, M. Crump, O. J. Culley, T. A. Hunter, et al. (2011). "Investigation of 15 of the top candidate genes for late-onset Alzheimer's disease." Hum Genet **129**(3): 273-282.

Bell, R. D., A. P. Sagare, A. E. Friedman, G. S. Bedi, D. M. Holtzman, et al. (2007). "Transport pathways for clearance of human Alzheimer's amyloid beta-peptide and apolipoproteins E and J in the mouse central nervous system." J Cereb Blood Flow Metab **27**(5): 909-918.

Bertram, L., C. Lange, K. Mullin, M. Parkinson, M. Hsiao, et al. (2008). "Genome-wide association analysis reveals putative Alzheimer's disease susceptibility loci in addition to APOE." Am J Hum Genet **83**(5): 623-632.

Bertram, L., M. B. McQueen, K. Mullin, D. Blacker and R. E. Tanzi (2007). "Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database." Nat Genet **39**(1): 17-23.

Bertram, L. and R. E. Tanzi (2010). "Alzheimer disease: New light on an old CLU." Nat Rev Neurol **6**(1): 11-13.

Bertrand, P., J. Poirier, T. Oda, C. E. Finch and G. M. Pasinetti (1995). "Association of apolipoprotein E genotype with brain levels of apolipoprotein E and apolipoprotein J (clusterin) in Alzheimer disease." Brain Res Mol Brain Res **33**(1): 174-178.

Bettens, K., N. Brouwers, S. Engelborghs, J. C. Lambert, E. Rogaeva, et al. (2012). "Both common variations and rare non-synonymous substitutions and small insertion/deletions in CLU are associated with increased Alzheimer risk." Mol Neurodegener **7**(1): 3.

Bianchin, M. M., H. M. Capella, D. L. Chaves, M. Steindel, E. C. Grisard, et al. (2004). "Nasu-Hakola disease (polycystic lipomembranous osteodysplasia with sclerosing leukoencephalopathy--PLOSL): a dementia associated with bone cystic lesions. From clinical to genetic and molecular aspects." Cell Mol Neurobiol **24**(1): 1-24.

Biedermann, B., D. Gil, D. T. Bowen and P. R. Crocker (2007). "Analysis of the CD33-related siglec family reveals that Siglec-9 is an endocytic receptor expressed on subsets of acute myeloid leukemia cells and absent from normal hematopoietic progenitors." Leuk Res **31**(2): 211-220.

Biffi, A., C. D. Anderson, R. S. Desikan, M. Sabuncu, L. Cortellini, et al. (2010). "Genetic variation and neuroimaging measures in Alzheimer disease." Arch Neurol **67**(6): 677-685.

Bird, T. D., E. Levy-Lahad, P. Poorkaj, V. Sharma, E. Nemens, et al. (1996). "Wide range in age of onset for chromosome 1--related familial Alzheimer's disease." Ann Neurol **40**(6): 932-936.

Blaschuk, O., K. Burdzy and I. B. Fritz (1983). "Purification and characterization of a cell-aggregating factor (clusterin), the major glycoprotein in ram rete testis fluid." J Biol Chem **258**(12): 7714-7720.

Bradt, B. M., W. P. Kolb and N. R. Cooper (1998). "Complement-dependent proinflammatory properties of the Alzheimer's disease beta-peptide." J Exp Med **188**(3): 431-438.

Bralten, J., B. Franke, A. Arias-Vasquez, A. Heister, H. G. Brunner, et al. (2011). "CR1 genotype is associated with entorhinal cortex volume in young healthy adults." Neurobiol Aging.

Breitner, J. C., B. A. Gau, K. A. Welsh, B. L. Plassman, W. M. McDonald, et al. (1994). "Inverse association of anti-inflammatory treatments and Alzheimer's disease: initial results of a co-twin control study." Neurology **44**(2): 227-232.

Brookmeyer, R., E. Johnson, K. Ziegler-Graham and H. M. Arrighi (2007). "Forecasting the global burden of Alzheimer's disease." Alzheimers Dement **3**(3): 186-191.

Brouwers, N., C. Van Cauwenberghe, S. Engelborghs, J. C. Lambert, K. Bettens, et al. (2012). "Alzheimer risk associated with a copy number variation in the complement receptor 1 increasing C3b/C4b binding sites." Mol Psychiatry **17**(2): 223-233.

Brunham, L. R., R. R. Singaraja, T. D. Pape, A. Kejariwal, P. D. Thomas, et al. (2005). "Accurate prediction of the functional significance of single nucleotide polymorphisms and mutations in the ABCA1 gene." PLoS Genet **1**(6): e83.

Bushlin, I., R. S. Petralia, F. Wu, A. Harel, M. R. Mughal, et al. (2008). "Clathrin assembly protein AP180 and CALM differentially control axogenesis and dendrite outgrowth in embryonic hippocampal neurons." J Neurosci **28**(41): 10257-10271.

Cappelletti, V., M. Gariboldi, L. De Cecco, S. Toffanin, J. F. Reid, et al. (2008). "Patterns and changes in gene expression following neo-adjuvant anti-estrogen treatment in estrogen receptor-positive breast cancer." Endocr Relat Cancer **15**(2): 439-449.

Carey, R. M., B. A. Balcz, I. Lopez-Coviella and B. E. Slack (2005). "Inhibition of dynamin-dependent endocytosis increases shedding of the amyloid precursor protein ectodomain and reduces generation of amyloid beta protein." BMC Cell Biol **6**: 30.

Carrasquillo, M. M., O. Belbin, T. A. Hunter, L. Ma, G. D. Bisceglio, et al. (2010). "Replication of CLU, CR1, and PICALM associations with alzheimer disease." Arch Neurol **67**(8): 961-964.

Carrasquillo, M. M., O. Belbin, T. A. Hunter, L. Ma, G. D. Bisceglio, et al. (2011). "Replication of EPHA1 and CD33 associations with late-onset Alzheimer's disease: a multi-centre case-control study." Mol Neurodegener **6**(1): 54.

Carrasquillo, M. M., F. Zou, V. S. Pankratz, S. L. Wilcox, L. Ma, et al. (2009). "Genetic variation in PCDH11X is associated with susceptibility to late-onset Alzheimer's disease." Nat Genet **41**(2): 192-198.

Cartegni, L., J. Wang, Z. Zhu, M. Q. Zhang and A. R. Krainer (2003). "ESEfinder: A web resource to identify exonic splicing enhancers." Nucleic Acids Res **31**(13): 3568-3571.

Cataldo, A. M., C. M. Peterhoff, J. C. Troncoso, T. Gomez-Isla, B. T. Hyman, et al. (2000). "Endocytic pathway abnormalities precede amyloid beta deposition in

sporadic Alzheimer's disease and Down syndrome: differential effects of APOE genotype and presenilin mutations." Am J Pathol **157**(1): 277-286.

Chang, K. A. and Y. H. Suh (2010). "Possible roles of amyloid intracellular domain of amyloid precursor protein." BMB Rep **43**(10): 656-663.

Chang, W. P., X. Huang, D. Downs, J. R. Cirrito, G. Koelsch, et al. (2011). "Beta-secretase inhibitor GRL-8234 rescues age-related cognitive decline in APP transgenic mice." FASEB J **25**(2): 775-784.

Charnay, Y., A. Imhof, P. G. Vallet, D. Hakkoum, A. Lathuiliere, et al. (2008). "Clusterin expression during fetal and postnatal CNS development in mouse." Neuroscience **155**(3): 714-724.

Chayka, O., D. Corvetta, M. Dews, A. E. Caccamo, I. Piotrowska, et al. (2009). "Clusterin, a haploinsufficient tumor suppressor gene in neuroblastomas." J Natl Cancer Inst **101**(9): 663-677.

Chen, W. J., J. L. Goldstein and M. S. Brown (1990). "NPXY, a sequence often found in cytoplasmic tails, is required for coated pit-mediated internalization of the low density lipoprotein receptor." J Biol Chem **265**(6): 3116-3123.

Chen, Y., A. K. Fu and N. Y. Ip (2012). "Eph receptors at synapses: implications in neurodegenerative diseases." Cell Signal **24**(3): 606-611.

Chi, K. N., E. Eisenhauer, L. Fazli, E. C. Jones, S. L. Goldenberg, et al. (2005). "A phase I pharmacokinetic and pharmacodynamic study of OGX-011, a 2'-methoxyethyl antisense oligonucleotide to clusterin, in patients with localized prostate cancer." J Natl Cancer Inst **97**(17): 1287-1296.

Chia, S., S. Dent, S. Ellard, P. M. Ellis, T. Vandenberg, et al. (2009). "Phase II trial of OGX-011 in combination with docetaxel in metastatic breast cancer." Clin Cancer Res **15**(2): 708-713.

Chibnik, L. B., J. M. Shulman, S. E. Leurgans, J. A. Schneider, R. S. Wilson, et al. (2011). "CR1 is associated with amyloid plaque burden and age-related cognitive decline." Ann Neurol **69**(3): 560-569.

Choi, M., U. I. Scholl, W. Ji, T. Liu, I. R. Tikhonova, et al. (2009). "Genetic diagnosis by whole exome capture and massively parallel DNA sequencing." Proc Natl Acad Sci U S A **106**(45): 19096-19101.

Cirrito, J. R., J. E. Kang, J. Lee, F. R. Stewart, D. K. Verges, et al. (2008). "Endocytosis is required for synaptic activity-dependent release of amyloid-beta in vivo." Neuron **58**(1): 42-51.

Cochrane, D. R., Z. Wang, M. Muramaki, M. E. Gleave and C. C. Nelson (2007). "Differential regulation of clusterin and its isoforms by androgens in prostate cells." J Biol Chem **282**(4): 2278-2287.

Cockburn, I. A., M. J. Mackinnon, A. O'Donnell, S. J. Allen, J. M. Moulds, et al. (2004). "A human complement receptor 1 polymorphism that reduces Plasmodium falciparum rosetting confers protection against severe malaria." Proc Natl Acad Sci U S A **101**(1): 272-277.

Cockburn, I. A. and J. A. Rowe (2006). "Erythrocyte complement receptor 1 (CR1) expression level is not associated with polymorphisms in the promoter or 3′ untranslated regions of the CR1 gene." International Journal of Immunogenetics **33**(1): 17-20.

Coon, K. D., A. J. Myers, D. W. Craig, J. A. Webster, J. V. Pearson, et al. (2007). "A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease." J Clin Psychiatry **68**(4): 613-618.

Corneveaux, J. J., A. J. Myers, A. N. Allen, J. J. Pruzin, M. Ramirez, et al. (2010). "Association of CR1, CLU and PICALM with Alzheimer's disease in a cohort

of clinically characterized and neuropathologically verified individuals." Hum Mol Genet **19**(16): 3295-3301.

Cotman, C. W. and J. H. Su (1996). "Mechanisms of neuronal death in Alzheimer's disease." Brain Pathol **6**(4): 493-506.

Covas, D. T., F. S. de Oliveira, E. S. Rodrigues, K. Abe-Sandes, W. A. Silva, Jr., et al. (2007). "Knops blood group haplotypes among distinct Brazilian populations." Transfusion **47**(1): 147-153.

Crehan, H., P. Holton, S. Wray, J. Pocock, R. Guerreiro, et al. (2012). "Complement receptor 1 (CR1) and Alzheimer's disease." Immunobiology **217**(2): 244-250.

Criswell, T., M. Beman, S. Araki, K. Leskov, E. Cataldo, et al. (2005). "Delayed activation of insulin-like growth factor-1 receptor/Src/MAPK/Egr-1 signaling regulates clusterin expression, a pro-survival factor." J Biol Chem **280**(14): 14212-14221.

Crocker, P. R., J. C. Paulson and A. Varki (2007). "Siglecs and their roles in the immune system." Nat Rev Immunol **7**(4): 255-266.

Dangour, A. D., P. J. Whitehouse, K. Rafferty, S. A. Mitchell, L. Smith, et al. (2010). "B-vitamins and fatty acids in the prevention and treatment of Alzheimer's disease and dementia: a systematic review." J Alzheimers Dis **22**(1): 205-224.

Daniels, G. L., D. J. Anstee, J. P. Cartron, W. Dahr, P. D. Issitt, et al. (1995). "Blood group terminology 1995. ISBT Working Party on terminology for red cell surface antigens." Vox Sang **69**(3): 265-279.

Danielsson, C., M. Pascual, L. French, G. Steiger and J. A. Schifferli (1994). "Soluble complement receptor type 1 (CD35) is released from leukocytes by surface cleavage." Eur J Immunol **24**(11): 2725-2731.

Dati, G., A. Quattrini, L. Bernasconi, M. C. Malaguti, B. Antonsson, et al. (2007). "Beneficial effects of r-h-CLU on disease severity in different animal models of peripheral neuropathies." J Neuroimmunol **190**(1-2): 8-17.

Day-Williams, A. G., K. McLay, E. Drury, S. Edkins, A. J. Coffey, et al. (2011). "An evaluation of different target enrichment methods in pooled sequencing designs for complex disease association studies." PLoS ONE **6**(11): e26279.

de Silva, H. V., J. A. Harmony, W. D. Stuart, C. M. Gil and J. Robbins (1990). "Apolipoprotein J: structure and tissue distribution." Biochemistry **29**(22): 5380-5389.

de Vooght, K. M., R. van Wijk and W. W. van Solinge (2009). "Management of gene promoter mutations in molecular diagnostics." Clin Chem **55**(4): 698-708.

Deane, R., A. Sagare, K. Hamm, M. Parisi, S. Lane, et al. (2008). "apoE isoform-specific disruption of amyloid beta peptide clearance from mouse brain." J Clin Invest **118**(12): 4002-4013.

DeMattos, R. B., R. P. Brendza, J. E. Heuser, M. Kierson, J. R. Cirrito, et al. (2001). "Purification and characterization of astrocyte-secreted apolipoprotein E and J-containing lipoproteins from wild-type and human apoE transgenic mice." Neurochem Int **39**(5-6): 415-425.

DeMattos, R. B., J. R. Cirrito, M. Parsadanian, P. C. May, M. A. O'Dell, et al. (2004). "ApoE and clusterin cooperatively suppress Abeta levels and deposition: evidence that ApoE regulates extracellular Abeta metabolism in vivo." Neuron **41**(2): 193-202.

DeMattos, R. B., A. O'Dell M, M. Parsadanian, J. W. Taylor, J. A. Harmony, et al. (2002). "Clusterin promotes amyloid plaque formation and is critical for neuritic toxicity in a mouse model of Alzheimer's disease." Proc Natl Acad Sci U S A **99**(16): 10843-10848.

Devauchelle, V., S. Marion, N. Cagnard, S. Mistou, G. Falgarone, et al. (2004). "DNA microarray allows molecular profiling of rheumatoid arthritis and identification of pathophysiological targets." Genes Immun **5**(8): 597-608.

Di, Y. M., E. Chan, M. Q. Wei, J. P. Liu and S. F. Zhou (2009). "Prediction of deleterious non-synonymous single-nucleotide polymorphisms of human uridine diphosphate glucuronosyltransferase genes." AAPS J **11**(3): 469-480.

Dosunmu, R., J. Wu, M. R. Basha and N. H. Zawia (2007). "Environmental and dietary risk factors in Alzheimer's disease." Expert Rev Neurother **7**(7): 887-900.

Dreyling, M. H., J. A. Martinez-Climent, M. Zheng, J. Mao, J. D. Rowley, et al. (1996). "The t(10;11)(p13;q14) in the U937 cell line results in the fusion of the AF10 gene and CALM, encoding a new member of the AP-3 clathrin assembly protein family." Proc Natl Acad Sci U S A **93**(10): 4804-4809.

Dubois, B., H. H. Feldman, C. Jacova, S. T. Dekosky, P. Barberger-Gateau, et al. (2007). "Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria." Lancet Neurol **6**(8): 734-746.

DuHadaway, J. B., F. J. Lynch, S. Brisbay, C. Bueso-Ramos, P. Troncoso, et al. (2003). "Immunohistochemical analysis of Bin1/Amphiphysin II in human tissues: diverse sites of nuclear expression and losses in prostate cancer." J Cell Biochem **88**(3): 635-642.

Dustin, M. L., M. W. Olszowy, A. D. Holdorf, J. Li, S. Bromley, et al. (1998). "A novel adaptor protein orchestrates receptor patterning and cytoskeletal polarity in T-cell contacts." Cell **94**(5): 667-677.

Dykman, T. R., J. L. Cole, K. Iida and J. P. Atkinson (1983). "Polymorphism of human erythrocyte C3b/C4b receptor." Proc Natl Acad Sci U S A **80**(6): 1698-1702.

Eikelenboom, P., R. Veerhuis, W. Scheper, A. J. Rozemuller, W. A. van Gool, et al. (2006). "The significance of neuroinflammation in understanding Alzheimer's disease." J Neural Transm **113**(11): 1685-1695.

ENCODE (2011). "A user's guide to the encyclopedia of DNA elements (ENCODE)." PLoS Biol **9**(4): e1001046.

Farrer, L. A., L. A. Cupples, J. L. Haines, B. Hyman, W. A. Kukull, et al. (1997). "Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium." JAMA **278**(16): 1349-1356.

Fearon, D. T. and L. A. Collins (1983). "Increased expression of C3b receptors on polymorphonuclear leukocytes induced by chemotactic factors and by purification procedures." J Immunol **130**(1): 370-375.

Ferrari, R., J. H. Moreno, A. T. Minhajuddin, S. E. O'Bryant, J. S. Reisch, et al. (2012). "Implication of common and disease specific variants in CLU, CR1, and PICALM." Neurobiol Aging.

Ferreira, S. T., M. N. Vieira and F. G. De Felice (2007). "Soluble protein oligomers as emerging toxins in Alzheimer's and other amyloid diseases." IUBMB Life **59**(4-5): 332-345.

Fingeroth, J. D., M. E. Heath and D. M. Ambrosino (1989). "Proliferation of resting B cells is modulated by CR2 and CR1." Immunol Lett **21**(4): 291-301.

Fitzjohn, S. M., R. A. Morton, F. Kuenzi, T. W. Rosahl, M. Shearman, et al. (2001). "Age-related impairment of synaptic transmission but normal long-term potentiation in transgenic mice that overexpress the human APP695SWE mutant form of amyloid precursor protein." J Neurosci **21**(13): 4691-4698.

Fleisher, A. S., R. Raman, E. R. Siemers, L. Becerra, C. M. Clark, et al. (2008). "Phase 2 safety trial targeting amyloid beta production with a gamma-secretase inhibitor in Alzheimer disease." Arch Neurol **65**(8): 1031-1038.

Folstein, M. F., S. E. Folstein and P. R. McHugh (1975). ""Mini-mental state" : A practical method for grading the cognitive state of patients for the clinician." Journal of Psychiatric Research **12**(3): 189-198.

Fonseca, M. I., S. H. Chu, A. M. Berci, M. E. Benoit, D. G. Peters, et al. (2011). "Contribution of complement activation pathways to neuropathology differs among mouse models of Alzheimer's disease." J Neuroinflammation **8**(1): 4.

Fonseca, M. I., C. H. Kawas, J. C. Troncoso and A. J. Tenner (2004). "Neuronal localization of C1q in preclinical Alzheimer's disease." Neurobiol Dis **15**(1): 40-46.

Ford, M. G., B. M. Pearse, M. K. Higgins, Y. Vallis, D. J. Owen, et al. (2001). "Simultaneous binding of PtdIns(4,5)P2 and clathrin by AP180 in the nucleation of clathrin lattices on membranes." Science **291**(5506): 1051-1055.

Francis, P. T., A. M. Palmer, M. Snape and G. K. Wilcock (1999). "The cholinergic hypothesis of Alzheimer's disease: a review of progress." J Neurol Neurosurg Psychiatry **66**(2): 137-147.

Frank, S., G. J. Burbach, M. Bonin, M. Walter, W. Streit, et al. (2008). "TREM2 is upregulated in amyloid plaque-associated microglia in aged APP23 transgenic mice." Glia **56**(13): 1438-1447.

Franke, A., D. P. McGovern, J. C. Barrett, K. Wang, G. L. Radford-Smith, et al. (2010). "Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci." Nat Genet **42**(12): 1118-1125.

Freixes, M., B. Puig, A. Rodriguez, B. Torrejon-Escribano, R. Blanco, et al. (2004). "Clusterin solubility and aggregation in Creutzfeldt-Jakob disease." Acta Neuropathol **108**(4): 295-301.

Fukumoto, H., H. Takahashi, N. Tarui, J. Matsui, T. Tomita, et al. (2010). "A noncompetitive BACE1 inhibitor TAK-070 ameliorates Abeta pathology and behavioral deficits in a mouse model of Alzheimer's disease." J Neurosci **30**(33): 11157-11166.

Furney, S. J., A. Simmons, G. Breen, I. Pedroso, K. Lunnon, et al. (2010). "Genome-wide association with MRI atrophy measures as a quantitative trait locus for Alzheimer's disease." Mol Psychiatry.

Gao, J., X. Huang, Y. Park, A. Hollenbeck and H. Chen (2011). "An exploratory study on CLU, CR1 and PICALM and Parkinson disease." PLoS ONE **6**(8): e24211.

Garcia, D. M., D. Baek, C. Shin, G. W. Bell, A. Grimson, et al. (2011). "Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs." Nat Struct Mol Biol **18**(10): 1139-1146.

Gasque, P., P. Chan, C. Mauger, M. T. Schouft, S. Singhrao, et al. (1996). "Identification and characterization of complement C3 receptors on human astrocytes." J Immunol **156**(6): 2247-2255.

Gatz, M., C. A. Reynolds, L. Fratiglioni, B. Johansson, J. A. Mortimer, et al. (2006). "Role of genes and environments for explaining Alzheimer disease." Arch Gen Psychiatry **63**(2): 168-174.

Gervais, F., J. Paquette, C. Morissette, P. Krzywkowski, M. Yu, et al. (2007). "Targeting soluble Abeta peptide with Tramiprosate for the treatment of brain amyloidosis." Neurobiol Aging **28**(4): 537-547.

Ghiran, I., S. F. Barbashov, L. B. Klickstein, S. W. Tas, J. C. Jensenius, et al. (2000). "Complement receptor 1/CD35 is a receptor for mannan-binding lectin." J Exp Med **192**(12): 1797-1808.

Ghiso, J., E. Matsubara, A. Koudinov, N. H. Choi-Miura, M. Tomita, et al. (1993). "The cerebrospinal-fluid soluble form of Alzheimer's amyloid beta is complexed to

SP-40,40 (apolipoprotein J), an inhibitor of the complement membrane-attack complex." Biochem J **293 (Pt 1)**: 27-30.

Giannakopoulos, P., E. Kovari, L. E. French, I. Viard, P. R. Hof, et al. (1998). "Possible neuroprotective role of clusterin in Alzheimer's disease: a quantitative immunocytochemical study." Acta Neuropathol **95**(4): 387-394.

Gibson, N. C. and F. J. Waxman (1994). "Relationship between immune complex binding and release and the quantitative expression of the complement receptor, type 1 (CR1, CD35) on human erythrocytes." Clin Immunol Immunopathol **70**(2): 104-113.

Gnirke, A., A. Melnikov, J. Maguire, P. Rogov, E. M. LeProust, et al. (2009). "Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing." Nat Biotechnol **27**(2): 182-189.

Goate, A., M. C. Chartier-Harlin, M. Mullan, J. Brown, F. Crawford, et al. (1991). "Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease." Nature **349**(6311): 704-706.

Goedert, M. and R. Jakes (2005). "Mutations causing neurodegenerative tauopathies." Biochim Biophys Acta **1739**(2-3): 240-250.

Goedert, M., M. G. Spillantini, R. Jakes, R. A. Crowther, E. Vanmechelen, et al. (1995). "Molecular dissection of the paired helical filament." Neurobiol Aging **16**(3): 325-334.

Goldstein, D. B. (2009). "Common genetic variation and human traits." N Engl J Med **360**(17): 1696-1698.

Gong, J. S., M. Kobayashi, H. Hayashi, K. Zou, N. Sawamura, et al. (2002). "Apolipoprotein E (ApoE) isoform-dependent lipid release from astrocytes prepared from human ApoE3 and ApoE4 knock-in mice." J Biol Chem **277**(33): 29919-29926.

Grassilli, E., S. Bettuzzi, L. Troiano, M. C. Ingletti, D. Monti, et al. (1992). "SGP-2, apoptosis, and aging." Ann N Y Acad Sci **663**: 471-474.

Grbovic, O. M., P. M. Mathews, Y. Jiang, S. D. Schmidt, R. Dinakar, et al. (2003). "Rab5-stimulated up-regulation of the endocytic pathway increases intracellular beta-cleaved amyloid precursor protein carboxyl-terminal fragment levels and Abeta production." J Biol Chem **278**(33): 31261-31268.

Grehan, S., E. Tse and J. M. Taylor (2001). "Two distal downstream enhancers direct expression of the human apolipoprotein E gene to astrocytes in the brain." J Neurosci **21**(3): 812-822.

Griffin, J. D., D. Linch, K. Sabbath, P. Larcom and S. F. Schlossman (1984). "A monoclonal antibody reactive with normal and leukemic human myeloid progenitor cells." Leuk Res **8**(4): 521-534.

Grunkemeyer, J. A., C. Kwoh, T. B. Huber and A. S. Shaw (2005). "CD2-associated protein (CD2AP) expression in podocytes rescues lethality of CD2AP deficiency." J Biol Chem **280**(33): 29677-29681.

Grupe, A., R. Abraham, Y. Li, C. Rowland, P. Hollingworth, et al. (2007). "Evidence for novel susceptibility genes for late-onset Alzheimer's disease from a genome-wide association study of putative functional variants." Hum Mol Genet **16**(8): 865-873.

Guerreiro, R., A. Wojtas, J. Bras, M. Carrasquillo, E. Rogaeva, et al. (2013). "TREM2 variants in Alzheimer's disease." N Engl J Med **368**(2): 117-127.

Guerreiro, R. J., J. Beck, J. R. Gibbs, I. Santana, M. N. Rossor, et al. (2010). "Genetic variability in CLU and its association with Alzheimer's disease." PLoS ONE **5**(3): e9510.

Hamer, I., J. P. Paccaud, D. Belin, C. Maeder and J. L. Carpentier (1998). "Soluble form of complement C3b/C4b receptor (CR1) results from a proteolytic cleavage in the C-terminal region of CR1 transmembrane domain." Biochem J **329 (Pt 1)**: 183-190.

Hamilton, G., K. L. Evans, D. J. Macintyre, I. J. Deary, A. Dominiczak, et al. (2012). "Alzheimer's disease risk factor complement receptor 1 is associated with depression." Neurosci Lett **510**(1): 6-9.

Hammad, S. M., S. Ranganathan, E. Loukinova, W. O. Twal and W. S. Argraves (1997). "Interaction of apolipoprotein J-amyloid beta-peptide complex with low density lipoprotein receptor-related protein-2/megalin. A mechanism to prevent pathological accumulation of amyloid beta-peptide." J Biol Chem **272**(30): 18644-18649.

Han, B. H., R. B. DeMattos, L. L. Dugan, J. S. Kim-Han, R. P. Brendza, et al. (2001). "Clusterin contributes to caspase-3-independent brain injury following neonatal hypoxia-ischemia." Nat Med **7**(3): 338-343.

HapMap, C. (2003). "The International HapMap Project." Nature **426**(6968): 789-796.

Harding, M. A., L. J. Chadwick, V. H. Gattone, 2nd and J. P. Calvet (1991). "The SGP-2 gene is developmentally regulated in the mouse kidney and abnormally expressed in collecting duct cysts in polycystic kidney disease." Dev Biol **146**(2): 483-490.

Hardy, J. A. and G. A. Higgins (1992). "Alzheimer's disease: the amyloid cascade hypothesis." Science **256**(5054): 184-185.

Harel, A., M. P. Mattson and P. J. Yao (2011). "CALM, A Clathrin Assembly Protein, Influences Cell Surface GluR2 Abundance." Neuromolecular Med.

Harel, A., F. Wu, M. P. Mattson, C. M. Morris and P. J. Yao (2008). "Evidence for CALM in directing VAMP2 trafficking." Traffic **9**(3): 417-429.

Harold, D., R. Abraham, P. Hollingworth, R. Sims, A. Gerrish, et al. (2009). "Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease." Nat Genet **41**(10): 1088-1093.

Harr, S. D., L. Uint, R. Hollister, B. T. Hyman and A. J. Mendez (1996). "Brain expression of apolipoproteins E, J, and A-I in Alzheimer's disease." J Neurochem **66**(6): 2429-2435.

Hebsgaard, S. M., P. G. Korning, N. Tolstrup, J. Engelbrecht, P. Rouze, et al. (1996). "Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information." Nucleic Acids Res **24**(17): 3439-3452.

Hedges, D. J., T. Guettouche, S. Yang, G. Bademci, A. Diaz, et al. (2011). "Comparison of Three Targeted Enrichment Strategies on the SOLiD Sequencing Platform." PLoS ONE **6**(4): e18595.

Hirai, H., Y. Maru, K. Hagiwara, J. Nishida and F. Takaku (1987). "A novel putative tyrosine kinase receptor encoded by the eph gene." Science **238**(4834): 1717-1720.

Holers, V. M., D. D. Chaplin, J. F. Leykam, B. A. Gruner, V. Kumar, et al. (1987). "Human complement C3b/C4b receptor (CR1) mRNA polymorphism that correlates with the CR1 allelic molecular weight polymorphism." Proc Natl Acad Sci U S A **84**(8): 2459-2463.

Hollingworth, P., D. Harold, L. Jones, M. J. Owen and J. Williams (2010). "Alzheimer's disease genetics: current knowledge and future challenges." Int J Geriatr Psychiatry.

Hollingworth, P., D. Harold, R. Sims, A. Gerrish, J. C. Lambert, et al. (2011). "Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease." Nat Genet **43**(5): 429-435.

Holmes, C., D. Boche, D. Wilkinson, G. Yadegarfar, V. Hopkins, et al. (2008). "Long-term effects of A[beta]42 immunisation in Alzheimer's disease: follow-up of a randomised, placebo-controlled phase I trial." The Lancet **372**(9634): 216-223.

Homer, N., B. Merriman and S. F. Nelson (2009). "BFAST: an alignment tool for large scale genome resequencing." PLoS ONE **4**(11): e7767.

Howie, B. N., P. Donnelly and J. Marchini (2009). "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies." PLoS Genet **5**(6): e1000529.

Hruska, M. and M. B. Dalva (2012). "Ephrin regulation of synapse formation, function and plasticity." Mol Cell Neurosci **50**(1): 35-44.

Hsieh, H., J. Boehm, C. Sato, T. Iwatsubo, T. Tomita, et al. (2006). "AMPAR removal underlies Abeta-induced synaptic depression and dendritic spine loss." Neuron **52**(5): 831-843.

Huang, F., A. Khvorova, W. Marshall and A. Sorkin (2004). "Analysis of clathrin-mediated endocytosis of epidermal growth factor receptor by RNA interference." J Biol Chem **279**(16): 16657-16661.

Hubbard, T., D. Barker, E. Birney, G. Cameron, Y. Chen, et al. (2002). "The Ensembl genome database project." Nucleic Acids Res **30**(1): 38-41.

Humphreys, D. T., J. A. Carver, S. B. Easterbrook-Smith and M. R. Wilson (1999). "Clusterin has chaperone-like activity similar to that of small heat shock proteins." J Biol Chem **274**(11): 6875-6881.

Imhof, A., Y. Charnay, P. G. Vallet, B. Aronow, E. Kovari, et al. (2006). "Sustained astrocytic clusterin expression improves remodeling after brain ischemia." Neurobiol Dis **22**(2): 274-283.

Ingman, M. and U. Gyllensten (2009). "SNP frequency estimation using massively parallel sequencing of pooled DNA." Eur J Hum Genet **17**(3): 383-386.

Ishikawa, Y., Y. Akasaka, T. Ishii, K. Komiyama, S. Masuda, et al. (1998). "Distribution and synthesis of apolipoprotein J in the atherosclerotic aorta." Arterioscler Thromb Vasc Biol **18**(4): 665-672.

Janus, C., J. Pearson, J. McLaurin, P. M. Mathews, Y. Jiang, et al. (2000). "A beta peptide immunization reduces behavioural impairment and plaques in a model of Alzheimer's disease." Nature **408**(6815): 979-982.

Jenne, D. E. and J. Tschopp (1989). "Molecular structure and functional characterization of a human complement cytolysis inhibitor found in blood and seminal plasma: identity to sulfated glycoprotein 2, a constituent of rat testis fluid." Proc Natl Acad Sci U S A **86**(18): 7123-7127.

Jick, H., G. L. Zornberg, S. S. Jick, S. Seshadri and D. A. Drachman (2000). "Statins and the risk of dementia." Lancet **356**(9242): 1627-1631.

Jin, C., X. Liu, F. Zhang, Y. Wu, J. Yuan, et al. (2013). "An updated meta-analysis of the association between SORL1 variants and the risk for sporadic Alzheimer's disease." J Alzheimers Dis **37**(2): 429-437.

Jin, G. and P. H. Howe (1997). "Regulation of clusterin gene expression by transforming growth factor beta." J Biol Chem **272**(42): 26620-26626.

Jin, G. and P. H. Howe (1999). "Transforming growth factor beta regulates clusterin gene expression via modulation of transcription factor c-Fos." Eur J Biochem **263**(2): 534-542.

Jin, S. C., P. Pastor, B. Cooper, S. Cervantes, B. A. Benitez, et al. (2012). "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 Res Ther **4**(4): 34.

Jones, S. E. and C. Jomary (2002). "Clusterin." Int J Biochem Cell Biol **34**(5): 427-431.

Jonsson, T., J. K. Atwal, S. Steinberg, J. Snaedal, P. V. Jonsson, et al. (2012). "A mutation in APP protects against Alzheimer's disease and age-related cognitive decline." Nature **488**(7409): 96-99.

Jonsson, T., H. Stefansson, S. Steinberg, I. Jonsdottir, P. V. Jonsson, et al. (2013). "Variant of TREM2 associated with the risk of Alzheimer's disease." N Engl J Med **368**(2): 107-116.

Jouvin, M. H., W. Rozenbaum, R. Russo and M. D. Kazatchkine (1987). "Decreased expression of the C3b/C4b complement receptor (CR1) in AIDS and AIDS-related syndromes correlates with clinical subpopulations of patients with HIV infection." AIDS **1**(2): 89-94.

Jozsi, M., J. Prechl, Z. Bajtay and A. Erdei (2002). "Complement receptor type 1 (CD35) mediates inhibitory signals in human B lymphocytes." J Immunol **168**(6): 2782-2788.

Jun, G., A. C. Naj, G. W. Beecham, L. S. Wang, J. Buros, et al. (2010). "Meta-analysis Confirms CR1, CLU, and PICALM as Alzheimer Disease Risk Loci and Reveals Interactions With APOE Genotypes." Arch Neurol.

Jung, N. and V. Haucke (2007). "Clathrin-mediated endocytosis at synapses." Traffic **8**(9): 1129-1136.

Jurka, J., V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, et al. (2005). "Repbase Update, a database of eukaryotic repetitive elements." Cytogenet Genome Res **110**(1-4): 462-467.

Katyal, M., B. Sivasankar, S. Ayub and N. Das (2003). "Genetic and structural polymorphism of complement receptor 1 in normal Indian subjects." Immunol Lett **89**(2-3): 93-98.

Kauwe, J. S., C. Cruchaga, C. M. Karch, B. Sadler, M. Lee, et al. (2011). "Fine mapping of genetic variants in BIN1, CLU, CR1 and PICALM for association with cerebrospinal fluid biomarkers for Alzheimer's disease." PLoS ONE **6**(2): e15918.

Keenan, B. T., J. M. Shulman, L. B. Chibnik, T. Raj, D. Tran, et al. (2012). "A coding variant in CR1 interacts with APOE-epsilon4 to influence cognitive decline." Hum Mol Genet **21**(10): 2377-2388.

Kenny, E. M., P. Cormican, W. P. Gilks, A. S. Gates, C. T. O'Dushlaine, et al. (2010). "Multiplex Target Enrichment Using DNA Indexing for Ultra-High Throughput SNP Detection." DNA Res.

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, et al. (2002). "The human genome browser at UCSC." Genome Res **12**(6): 996-1006.

Kertesz, M., N. Iovino, U. Unnerstall, U. Gaul and E. Segal (2007). "The role of site accessibility in microRNA target recognition." Nat Genet **39**(10): 1278-1284.

Khera, R. and N. Das (2009). "Complement Receptor 1: disease associations and therapeutic implications." Mol Immunol **46**(5): 761-772.

Kiialainen, A., O. Karlberg, A. Ahlford, S. Sigurdsson, K. Lindblad-Toh, et al. (2011). "Performance of microarray and liquid based capture methods for target enrichment for massively parallel sequencing and SNP discovery." PLoS ONE **6**(2): e16486.

Kim, H. L. and J. A. Kim (2000). "Purification of clathrin assembly protein from rat liver." Exp Mol Med **32**(4): 222-226.

Kim, J., J. M. Basak and D. M. Holtzman (2009). "The role of apolipoprotein E in Alzheimer's disease." Neuron **63**(3): 287-303.

Kim, J. A. and H. L. Kim (2001). "Cleavage of purified neuronal clathrin assembly protein (CALM) by caspase 3 and calpain." Exp Mol Med **33**(4): 245-250.

Kim, J. H., S. Lee and S. Y. Choe (1999). "Characterization of the human CR1 gene promoter." IUBMB Life **47**(4): 655-663.

Kim, W. S., C. S. Weickert and B. Garner (2008). "Role of ATP-binding cassette transporters in brain lipid transport and neurological disease." J Neurochem **104**(5): 1145-1166.

Kirszbaum, L., S. E. Bozas and I. D. Walker (1992). "SP-40,40, a protein involved in the control of the complement pathway, possesses a unique array of disulphide bridges." FEBS Lett **297**(1-2): 70-76.

Kivipelto, M. and A. Solomon (2006). "Cholesterol as a risk factor for Alzheimer's disease - epidemiological evidence." Acta Neurol Scand Suppl **185**: 50-57.

Klebig, M. L., M. D. Wall, M. D. Potter, E. L. Rowe, D. A. Carpenter, et al. (2003). "Mutations in the clathrin-assembly gene Picalm are responsible for the hematopoietic and iron metabolism abnormalities in fit1 mice." Proc Natl Acad Sci U S A **100**(14): 8360-8365.

Klickstein, L. B., T. J. Bartow, V. Miletic, L. D. Rabson, J. A. Smith, et al. (1988). "Identification of distinct C3b and C4b recognition sites in the human C3b/C4b receptor (CR1, CD35) by deletion mutagenesis." J Exp Med **168**(5): 1699-1717.

Klickstein, L. B., W. W. Wong, J. A. Smith, J. H. Weis, J. G. Wilson, et al. (1987). "Human C3b/C4b receptor (CR1). Demonstration of long homologous repeating domains that are composed of the short consensus repeats characteristics of C3/C4 binding proteins." J Exp Med **165**(4): 1095-1112.

Kok, E. H., T. Luoto, S. Haikonen, S. Goebeler, H. Haapasalo, et al. (2011). "CLU, CR1 and PICALM genes associate with Alzheimer's-related senile plaques." Alzheimers Res Ther **3**(2): 12.

Koo, E. H. and S. L. Squazzo (1994). "Evidence that production and release of amyloid beta-protein involves the endocytic pathway." J Biol Chem **269**(26): 17386-17389.

Koo, S. J., S. Markovic, D. Puchkov, C. C. Mahrenholz, F. Beceren-Braun, et al. (2011). "SNARE motif-mediated sorting of synaptobrevin by the endocytic adaptors clathrin assembly lymphoid myeloid leukemia (CALM) and AP180 at synapses." Proc Natl Acad Sci U S A **108**(33): 13540-13545.

Krych-Goldberg, M. and J. P. Atkinson (2001). "Structure-function relationships of complement receptor type 1." Immunol Rev **180**: 112-122.

Krych-Goldberg, M., R. E. Hauhart, V. B. Subramanian, B. M. Yurcisin, 2nd, D. L. Crimmins, et al. (1999). "Decay accelerating activity of complement receptor type 1 (CD35). Two active sites are required for dissociating C5 convertases." J Biol Chem **274**(44): 31160-31168.

Krych, M., L. Clemenza, D. Howdeshell, R. Hauhart, D. Hourcade, et al. (1994). "Analysis of the functional domains of complement receptor type 1 (C3b/C4b receptor; CD35) by substitution mutagenesis." J Biol Chem **269**(18): 13273-13278.

Krych, M., R. Hauhart and J. P. Atkinson (1998). "Structure-function analysis of the active sites of complement receptor type 1." J Biol Chem **273**(15): 8623-8629.

Kujiraoka, T., H. Hattori, Y. Miwa, M. Ishihara, T. Ueno, et al. (2006). "Serum apolipoprotein j in health, coronary heart disease and type 2 diabetes mellitus." J Atheroscler Thromb **13**(6): 314-322.

Kumar, A., A. N. Malaviya and L. M. Srivastava (1994). "Lowered expression of C3b receptor (CR1) on erythrocytes of rheumatoid arthritis patients." Immunobiology **191**(1): 9-20.

Kumar, A., S. Sinha, P. S. Khandekar, K. Banerjee and L. M. Srivastava (1995). "Hind III genomic polymorphism of the C3b receptor (CR1) in patients with SLE: low erythrocyte CR1 expression is an acquired phenomenon." Immunol Cell Biol **73**(5): 457-462.

Kumita, J. R., S. Poon, G. L. Caddy, C. L. Hagan, M. Dumoulin, et al. (2007). "The extracellular chaperone clusterin potently inhibits human lysozyme amyloid formation by interacting with prefibrillar species." J Mol Biol **369**(1): 157-167.

Kunz, D., R. Zimmermann, M. Heisig and P. C. Heinrich (1989). "Identification of the promoter sequences involved in the interleukin-6 dependent expression of the rat alpha 2-macroglobulin gene." Nucleic Acids Res **17**(3): 1121-1138.

Kuznetsova, E. B., T. V. Kekeeva, S. S. Larin, V. V. Zemlyakova, A. V. Khomyakova, et al. (2007). "Methylation of the BIN1 gene promoter CpG island associated with breast and prostate cancer." J Carcinog **6**: 9.

Kyprianou, N., H. F. English, N. E. Davidson and J. T. Isaacs (1991). "Programmed cell death during regression of the MCF-7 human breast cancer following estrogen ablation." Cancer Res **51**(1): 162-166.

Lakins, J. N., S. Poon, S. B. Easterbrook-Smith, J. A. Carver, M. P. Tenniswood, et al. (2002). "Evidence that clusterin has discrete chaperone and ligand binding sites." Biochemistry **41**(1): 282-291.

Lambert, J.-C., S. Heath, G. Even, D. Campion, K. Sleegers, et al. (2009). "Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease." Nat Genet **41**(10): 1094-1099.

Lambert, J. C., C. A. Ibrahim-Verbaas, D. Harold, A. C. Naj, R. Sims, et al. (2013). "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease." Nat Genet.

Lambert, J. C., D. Zelenika, M. Hiltunen, V. Chouraki, O. Combarros, et al. (2011). "Evidence of the association of BIN1 and PICALM with the AD risk in contrasting European populations." Neurobiol Aging.

Lannfelt, L., N. Bogdanovic, H. Appelgren, K. Axelman, L. Lilius, et al. (1994). "Amyloid precursor protein mutation causes Alzheimer's disease in a Swedish family." Neurosci Lett **168**(1-2): 254-256.

Laskin, J. J., G. Nicholas, C. Lee, B. Gitlitz, M. Vincent, et al. (2012). "Phase I/II Trial of Custirsen (OGX-011), an Inhibitor of Clusterin, in Combination with a Gemcitabine and Platinum Regimen in Patients with Previously Untreated Advanced Non-small Cell Lung Cancer." J Thorac Oncol **7**(3): 579-586.

Lassmann, T., Y. Hayashizaki and C. O. Daub (2011). "SAMStat: monitoring biases in next generation sequencing data." Bioinformatics **27**(1): 130-131.

Law, G. L. and M. D. Griswold (1994). "Activity and form of sulfated glycoprotein 2 (clusterin) from cultured Sertoli cells, testis, and epididymis of the rat." Biol Reprod **50**(3): 669-679.

Le, S. Q. and R. Durbin (2011). "SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples." Genome Res **21**(6): 952-960.

Lee, J. H., R. Cheng, S. Barral, C. Reitz, M. Medrano, et al. (2010). "Identification of Novel Loci for Alzheimer Disease and Replication of CLU, PICALM, and BIN1 in Caribbean Hispanic Individuals." Arch Neurol.

Leskov, K. S., D. Y. Klokov, J. Li, T. J. Kinsella and D. A. Boothman (2003). "Synthesis and functional analyses of nuclear clusterin, a cell death protein." J Biol Chem **278**(13): 11590-11600.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, et al. (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Li, H. L., S. S. Shi, Q. H. Guo, W. Ni, Y. Dong, et al. (2011). "PICALM and CR1 variants are not associated with sporadic Alzheimer's disease in Chinese patients." J Alzheimers Dis **25**(1): 111-117.

Li, X., P. E. Massa, A. Hanidu, G. W. Peet, P. Aro, et al. (2002). "IKKalpha, IKKbeta, and NEMO/IKKgamma are each required for the NF-kappa B-mediated inflammatory response program." J Biol Chem **277**(47): 45129-45140.

Liang, Y. and T. F. Tedder (2001). "Identification of a CD20-, FcepsilonRIbeta-, and HTm4-related gene family: sixteen new MS4A family members expressed in human and mouse." Genomics **72**(2): 119-127.

Liao, L., D. Cheng, J. Wang, D. M. Duong, T. G. Losik, et al. (2004). "Proteomic characterization of postmortem amyloid plaques isolated by laser capture microdissection." J Biol Chem **279**(35): 37061-37068.

Lidstrom, A. M., N. Bogdanovic, C. Hesse, I. Volkman, P. Davidsson, et al. (1998). "Clusterin (apolipoprotein J) protein levels are increased in hippocampus and in frontal cortex in Alzheimer's disease." Exp Neurol **154**(2): 511-521.

Ling, I. F., J. Bhongsatiern, J. F. Simpson, D. W. Fardo and S. Estus (2012). "Genetics of Clusterin Isoform Expression and Alzheimer's Disease Risk." PLoS ONE **7**(4): e33923.

Liu, D. and Z. X. Niu (2009). "The structure, genetic polymorphisms, expression and biological functions of complement receptor type 1 (CR1/CD35)." Immunopharmacol Immunotoxicol **31**(4): 524-535.

Logue, M. W., M. Schu, B. N. Vardarajan, J. Buros, R. C. Green, et al. (2011). "A comprehensive genetic association study of Alzheimer disease in African Americans." Arch Neurol **68**(12): 1569-1579.

Lord, J., J. Turton, C. Medway, H. Shi, K. Brown, et al. (2012). "Next generation sequencing of CLU, PICALM and CR1: pitfalls and potential solutions." Int J Mol Epidemiol Genet **3**(4): 262-275.

Love, S., R. Barber and G. K. Wilcock (1999). "Increased poly(ADP-ribosyl)ation of nuclear proteins in Alzheimer's disease." Brain **122 (Pt 2)**: 247-253.

Lovestone, S. (2000). "Fleshing out the amyloid cascade hypothesis: the molecular biology of Alzheimer's disease." Dialogues Clin Neurosci **2**(2): 101-110.

Lund, P., K. Weisshaupt, T. Mikeska, D. Jammas, X. Chen, et al. (2006). "Oncogenic HRAS suppresses clusterin expression through promoter hypermethylation." Oncogene **25**(35): 4890-4903.

Luo, Y., B. Bolon, S. Kahn, B. D. Bennett, S. Babu-Khan, et al. (2001). "Mice deficient in BACE1, the Alzheimer's beta-secretase, have normal phenotype and abolished beta-amyloid generation." Nat Neurosci **4**(3): 231-232.

Lymar, E. S., A. M. Clark, R. Reeves and M. D. Griswold (2000). "Clusterin gene in rat sertoli cells is regulated by a core-enhancer element." Biol Reprod **63**(5): 1341-1351.

Maccioni, R. B., G. Farias, I. Morales and L. Navarrete (2010). "The revitalized tau hypothesis on Alzheimer's disease." Arch Med Res **41**(3): 226-231.

Mahley, R. W., K. H. Weisgraber and Y. Huang (2006). "Apolipoprotein E4: a causative factor and therapeutic target in neuropathology, including Alzheimer's disease." Proc Natl Acad Sci U S A **103**(15): 5644-5651.

Maltsev, A. V., S. Bystryak and O. V. Galzitskaya (2011). "The role of beta-amyloid peptide in neurodegenerative diseases." Ageing Res Rev.

Mamanova, L., A. J. Coffey, C. E. Scott, I. Kozarewa, E. H. Turner, et al. (2010). "Target-enrichment strategies for next-generation sequencing." Nat Methods **7**(2): 111-118.

Man, H. Y., J. W. Lin, W. H. Ju, G. Ahmadian, L. Liu, et al. (2000). "Regulation of AMPA receptor-mediated synaptic transmission by clathrin-dependent receptor internalization." Neuron **25**(3): 649-662.

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

Marchesi, V. T. (2011). "Alzheimer's dementia begins as a disease of small blood vessels, damaged by oxidative-induced inflammation and dysregulated amyloid metabolism: implications for early detection and therapy." FASEB J **25**(1): 5-13.

Marchini, J. and B. Howie (2010). "Genotype imputation for genome-wide association studies." Nat Rev Genet **11**(7): 499-511.

Markesbery, W. R. (1997). "Oxidative stress hypothesis in Alzheimer's disease." Free Radic Biol Med **23**(1): 134-147.

Martin, B. K., C. Szekely, J. Brandt, S. Piantadosi, J. C. Breitner, et al. (2008). "Cognitive function over time in the Alzheimer's Disease Anti-inflammatory Prevention Trial (ADAPT): results of a randomized, controlled trial of naproxen and celecoxib." Arch Neurol **65**(7): 896-905.

Maru, Y., H. Hirai, M. C. Yoshida and F. Takaku (1988). "Evolution, expression, and chromosomal location of a novel receptor tyrosine kinase gene, eph." Mol Cell Biol **8**(9): 3770-3776.

Masliah, E., M. Mallory, M. Alford, R. DeTeresa, L. A. Hansen, et al. (2001). "Altered expression of synaptic proteins occurs early during progression of Alzheimer's disease." Neurology **56**(1): 127-129.

Matsubara, E., C. Soto, S. Governale, B. Frangione and J. Ghiso (1996). "Apolipoprotein J and Alzheimer's amyloid beta solubility." Biochem J **316 (Pt 2)**: 671-679.

Matys, V., O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, et al. (2006). "TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes." Nucleic Acids Res **34**(Database issue): D108-110.

May, P. C., M. Lampert-Etchells, S. A. Johnson, J. Poirier, J. N. Masters, et al. (1990). "Dynamics of gene expression for a hippocampal glycoprotein elevated in Alzheimer's disease and in response to experimental lesions in rat." Neuron **5**(6): 831-839.

McGeer, P. L., T. Kawamata and D. G. Walker (1992). "Distribution of clusterin in Alzheimer brain tissue." Brain Res **579**(2): 337-341.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, et al. (2010). "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." Genome Res **20**(9): 1297-1303.

McKhann, G., D. Drachman, M. Folstein, R. Katzman, D. Price, et al. (1984). "Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease." Neurology **34**(7): 939-944.

McLaren, W., B. Pritchard, D. Rios, Y. Chen, P. Flicek, et al. (2010). "Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor." Bioinformatics **26**(16): 2069-2070.

McLaughlin, L., G. Zhu, M. Mistry, C. Ley-Ebert, W. D. Stuart, et al. (2000). "Apolipoprotein J/clusterin limits the severity of murine autoimmune myocarditis." J Clin Invest **106**(9): 1105-1113.

McLaurin, J., M. E. Kierstead, M. E. Brown, C. A. Hawkes, M. H. Lambermon, et al. (2006). "Cyclohexanehexol inhibitors of Abeta aggregation prevent and reverse Alzheimer phenotype in a mouse model." Nat Med **12**(7): 801-808.

Mengel-From, J., K. Christensen, M. McGue and L. Christiansen (2010). "Genetic variations in the CLU and PICALM genes are associated with cognitive function in the oldest old." Neurobiol Aging.

Menissier de Murcia, J., M. Ricoul, L. Tartier, C. Niedergang, A. Huber, et al. (2003). "Functional interaction between PARP-1 and PARP-2 in chromosome stability and embryonic development in mouse." EMBO J **22**(9): 2255-2263.

Metzker, M. L. (2010). "Sequencing technologies - the next generation." Nat Rev Genet **11**(1): 31-46.

Meyerholz, A., L. Hinrichsen, S. Groos, P. C. Esk, G. Brandes, et al. (2005). "Effect of clathrin assembly lymphoid myeloid leukemia protein depletion on clathrin coat formation." Traffic **6**(12): 1225-1234.

Michel, D., G. Chatelain, Y. Herault and G. Brun (1995). "The expression of the avian clusterin gene can be driven by two alternative promoters with distinct regulatory elements." Eur J Biochem **229**(1): 215-223.

Michel, D., G. Chatelain, S. North and G. Brun (1997). "Stress-induced transcription of the clusterin/apoJ gene." Biochem J **328 (Pt 1)**: 45-50.

Miller, S. E., D. A. Sahlender, S. C. Graham, S. Honing, M. S. Robinson, et al. (2011). "The molecular basis for the endocytosis of small R-SNAREs by the clathrin adaptor CALM." Cell **147**(5): 1118-1131.

Miwa, Y., S. Takiuchi, K. Kamide, M. Yoshii, T. Horio, et al. (2005). "Insertion/deletion polymorphism in clusterin gene influences serum lipid levels and carotid intima-media thickness in hypertensive Japanese females." Biochem Biophys Res Commun **331**(4): 1587-1593.

Miyake, H., I. Hara, S. Kamidono and M. E. Gleave (2001). "Synergistic chemsensitization and inhibition of tumor growth and metastasis by the antisense oligodeoxynucleotide targeting clusterin gene in a human bladder cancer model." Clin Cancer Res **7**(12): 4245-4252.

Miyake, H., C. Nelson, P. S. Rennie and M. E. Gleave (2000). "Testosterone-repressed prostate message-2 is an antiapoptotic gene involved in progression to androgen independence in prostate cancer." Cancer Res **60**(1): 170-176.

Miyashita, A., A. Koike, G. Jun, L. S. Wang, S. Takahashi, et al. (2013). "SORL1 is genetically associated with late-onset Alzheimer's disease in Japanese, Koreans and Caucasians." PLoS ONE **8**(4): e58618.

Monzo, P., N. C. Gauthier, F. Keslair, A. Loubat, C. M. Field, et al. (2005). "Clues to CD2-associated protein involvement in cytokinesis." Mol Biol Cell **16**(6): 2891-2902.

Morgan, D., D. M. Diamond, P. E. Gottschall, K. E. Ugen, C. Dickey, et al. (2000). "A beta peptide vaccination prevents memory loss in an animal model of Alzheimer's disease." Nature **408**(6815): 982-985.

Morgan, K. (2011). "Commentary: The three new pathways leading to Alzheimer's disease." Neuropathol Appl Neurobiol.

Morgan, K. C., Minerva M. (Eds.), Ed. (2013). Genetic Variants in Alzheimer's Disease, Springer New York.

Morgan, T. E., Z. Xie, S. Goldsmith, T. Yoshida, A. S. Lanzrein, et al. (1999). "The mosaic of brain glial hyperactivity during normal ageing and its attenuation by food restriction." Neuroscience **89**(3): 687-699.

Moulds, J. M. (2010). "The Knops blood-group system: a review." Immunohematology **26**(1): 2-7.

Moulds, J. M., J. J. Moulds, M. Brown and J. P. Atkinson (1992). "Antiglobulin testing for CR1-related (Knops/McCoy/Swain-Langley/York) blood group antigens: negative and weak reactions are caused by variable expression of CR1." Vox Sang **62**(4): 230-235.

Moulds, J. M., M. W. Nickells, J. J. Moulds, M. C. Brown and J. P. Atkinson (1991). "The C3b/C4b receptor is recognized by the Knops, McCoy, Swain-langley, and York blood group antisera." J Exp Med **173**(5): 1159-1163.

Moulds, J. M., P. A. Zimmerman, O. K. Doumbo, D. A. Diallo, J. P. Atkinson, et al. (2002). "Expansion of the Knops blood group system and subdivision of Sl(a)." Transfusion **42**(2): 251-256.

Moulds, J. M., P. A. Zimmerman, O. K. Doumbo, L. Kassambara, I. Sagara, et al. (2001). "Molecular identification of Knops blood group polymorphisms found in long homologous region D of complement receptor 1." Blood **97**(9): 2879-2885.

Naj, A. C., G. Jun, G. W. Beecham, L. S. Wang, B. N. Vardarajan, et al. (2011). "Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease." Nat Genet **43**(5): 436-441.

National Institute for Health and Care Excellence (2011). "Donepezil, galantamine, rivastigmine and memantine for the treatment of Alzheimer's disease - Review of NICE technology appraisal guidance 111."

Navab, M., G. M. Anantharamaiah, S. T. Reddy, B. J. Van Lenten, A. C. Wagner, et al. (2005). "An oral apoJ peptide renders HDL antiinflammatory in mice and monkeys and dramatically reduces atherosclerosis in apolipoprotein E-null mice." Arterioscler Thromb Vasc Biol **25**(9): 1932-1937.

Nestlerode, C. S., C. H. Bunker, D. K. Sanghera, C. E. Aston, F. A. Ukoli, et al. (1999). "Apolipoprotein J polymorphisms and serum HDL cholesterol levels in African blacks." Hum Biol **71**(2): 197-218.

Neumann, M., M. Tolnay and I. R. Mackenzie (2009). "The molecular basis of frontotemporal dementia." Expert Rev Mol Med **11**: e23.

Newkirk, M. M., P. Apostolakos, C. Neville and P. R. Fortin (1999). "Systemic lupus erythematosus, a disease associated with low levels of clusterin/apoJ, an antiinflammatory protein." J Rheumatol **26**(3): 597-603.

Nho, K., J. J. Corneveaux, S. Kim, H. Lin, S. L. Risacher, et al. (2013). "Whole-exome sequencing and imaging genetics identify functional variants for rate of change in hippocampal volume in mild cognitive impairment." Mol Psychiatry **18**(7): 781-787.

Nilselid, A. M., P. Davidsson, K. Nagga, N. Andreasen, P. Fredman, et al. (2006). "Clusterin in cerebrospinal fluid: analysis of carbohydrates and quantification of native and glycosylated forms." Neurochem Int **48**(8): 718-728.

Nordstedt, C., G. L. Caporaso, J. Thyberg, S. E. Gandy and P. Greengard (1993). "Identification of the Alzheimer beta/A4 amyloid precursor protein in clathrin-coated vesicles purified from PC12 cells." J Biol Chem **268**(1): 608-612.

Nuutinen, T., J. Huuskonen, T. Suuronen, J. Ojala, R. Miettinen, et al. (2007). "Amyloid-beta 1-42 induced endocytosis and clusterin/apoJ protein accumulation in cultured human astrocytes." Neurochem Int **50**(3): 540-547.

Nuutinen, T., T. Suuronen, A. Kauppinen and A. Salminen (2009). "Clusterin: a forgotten player in Alzheimer's disease." Brain Res Rev **61**(2): 89-104.

Nuutinen, T., T. Suuronen, S. Kyrylenko, J. Huuskonen and A. Salminen (2005). "Induction of clusterin/apoJ expression by histone deacetylase inhibitors in neural cells." Neurochem Int **47**(8): 528-538.

Oda, T., G. M. Pasinetti, H. H. Osterburg, C. Anderson, S. A. Johnson, et al. (1994). "Purification and characterization of brain clusterin." Biochem Biophys Res Commun **204**(3): 1131-1136.

Olson, M. I. and C. M. Shaw (1969). "Presenile dementia and Alzheimer's disease in mongolism." Brain **92**(1): 147-156.

Orgogozo, J. M., S. Gilman, J. F. Dartigues, B. Laurent, M. Puel, et al. (2003). "Subacute meningoencephalitis in a subset of patients with AD after Abeta42 immunization." Neurology **61**(1): 46-54.

Ovcharenko, I., M. A. Nobrega, G. G. Loots and L. Stubbs (2004). "ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes." Nucleic Acids Res **32**(Web Server issue): W280-286.

Ozudogru, S. N. and C. F. Lippa (2012). "Disease modifying drugs targeting beta-amyloid." Am J Alzheimers Dis Other Demen **27**(5): 296-300.

Pant, S., M. Sharma, K. Patel, S. Caplan, C. M. Carr, et al. (2009). "AMPH-1/Amphiphysin/Bin1 functions with RME-1/Ehd1 in endocytic recycling." Nat Cell Biol **11**(12): 1399-1410.

Panza, F., V. Solfrizzi, V. Frisardi, C. Capurso, A. D'Introno, et al. (2009). "Disease-modifying approach to the treatment of Alzheimer's disease: from alpha-secretase activators to gamma-secretase inhibitors and modulators." Drugs Aging **26**(7): 537-555.

Pascual, M., M. A. Duchosal, G. Steiger, E. Giostra, A. Pechere, et al. (1993). "Circulating soluble CR1 (CD35). Serum levels in diseases and evidence for its release by human leukocytes." J Immunol **151**(3): 1702-1711.

Pascual, M., G. Steiger, S. Sadallah, J. P. Paccaud, J. L. Carpentier, et al. (1994). "Identification of membrane-bound CR1 (CD35) in human urine: evidence for its release by glomerular podocytes." J Exp Med **179**(3): 889-899.

Pasinetti, G. M., S. A. Johnson, T. Oda, I. Rozovsky and C. E. Finch (1994). "Clusterin (SGP-2): a multifunctional glycoprotein with regional expression in astrocytes and neurons of the adult rat brain." J Comp Neurol **339**(3): 387-400.

Pericak-Vance, M. A., J. L. Bebout, P. C. Gaskell, Jr., L. H. Yamaoka, W. Y. Hung, et al. (1991). "Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage." Am J Hum Genet **48**(6): 1034-1050.

Piccio, L., C. Buonsanti, M. Mariani, M. Cella, S. Gilfillan, et al. (2007). "Blockade of TREM-2 exacerbates experimental autoimmune encephalomyelitis." Eur J Immunol **37**(5): 1290-1301.

Pichon, X., L. A. Wilson, M. Stoneley, A. Bastide, H. A. King, et al. (2012). "RNA binding protein/RNA element interactions and the control of translation." Curr Protein Pept Sci **13**(4): 294-304.

Pimplikar, S. W. (2009). "Reassessing the amyloid cascade hypothesis of Alzheimer's disease." Int J Biochem Cell Biol **41**(6): 1261-1268.

Pogge, E. (2010). "Vitamin D and Alzheimer's disease: is there a link?" Consult Pharm **25**(7): 440-450.

Pollard, K. S., M. J. Hubisz, K. R. Rosenbloom and A. Siepel (2010). "Detection of nonneutral substitution rates on mammalian phylogenies." Genome Res **20**(1): 110-121.

Poon, S., S. B. Easterbrook-Smith, M. S. Rybchyn, J. A. Carver and M. R. Wilson (2000). "Clusterin is an ATP-independent chaperone with very broad substrate specificity that stabilizes stressed proteins in a folding-competent state." Biochemistry **39**(51): 15953-15960.

Pottier, C., D. Hannequin, S. Coutant, A. Rovelet-Lecrux, D. Wallon, et al. (2012). "High frequency of potentially pathogenic SORL1 mutations in autosomal dominant early-onset Alzheimer disease." Mol Psychiatry **17**(9): 875-879.

Poulakou, M. V., K. I. Paraskevas, M. R. Wilson, D. C. Iliopoulos, C. Tsigris, et al. (2008). "Apolipoprotein J and leptin levels in patients with coronary heart disease." In Vivo **22**(4): 537-542.

Pruitt, K. D., T. Tatusova and D. R. Maglott (2007). "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." Nucleic Acids Res **35**(Database issue): D61-65.

Pucci, S., E. Bonanno, F. Pichiorri, C. Angeloni and L. G. Spagnoli (2004). "Modulation of different clusterin isoforms in human colon tumorigenesis." Oncogene **23**(13): 2298-2304.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, et al. (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." Am J Hum Genet **81**(3): 559-575.

Ramaglia, V., R. Wolterman, M. de Kok, M. A. Vigar, I. Wagenaar-Bos, et al. (2008). "Soluble complement receptor 1 protects the peripheral nerve from early axon loss after injury." Am J Pathol **172**(4): 1043-1052.

Rao, N., D. J. Ferguson, S. F. Lee and M. J. Telen (1991). "Identification of human erythrocyte blood group antigens on the C3b/C4b receptor." J Immunol **146**(10): 3502-3507.

Reddy, K. B., G. Jin, M. C. Karode, J. A. Harmony and P. H. Howe (1996). "Transforming growth factor beta (TGF beta)-induced nuclear localization of apolipoprotein J/clusterin in epithelial cells." Biochemistry **35**(19): 6157-6163.

Redondo, M., E. Villar, J. Torres-Munoz, T. Tellez, M. Morell, et al. (2000). "Overexpression of clusterin in human breast carcinoma." Am J Pathol **157**(2): 393-399.

Reese, M. G., F. H. Eeckman, D. Kulp and D. Haussler (1997). "Improved splice site detection in Genie." J Comput Biol **4**(3): 311-323.

Reiman, E. M., K. Chen, X. Liu, D. Bandy, M. Yu, et al. (2009). "Fibrillar amyloid-beta burden in cognitively normal people at 3 levels of genetic risk for Alzheimer's disease." Proc Natl Acad Sci U S A **106**(16): 6820-6825.

Ridge, P. G., S. Mukherjee, P. K. Crane and J. S. Kauwe (2013). "Alzheimer's disease: analyzing the missing heritability." PLoS ONE **8**(11): e79771.

Rivas, M. A., M. Beaudoin, A. Gardet, C. Stevens, Y. Sharma, et al. (2011). "Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease." Nat Genet **43**(11): 1066-1073.

Rizzi, F. and S. Bettuzzi (2010). "The clusterin paradigm in prostate and breast carcinogenesis." Endocr Relat Cancer **17**(1): R1-17.

Robinson, J. T., H. Thorvaldsdottir, W. Winckler, M. Guttman, E. S. Lander, et al. (2011). "Integrative genomics viewer." Nat Biotechnol **29**(1): 24-26.

Rogaeva, E., Y. Meng, J. H. Lee, Y. Gu, T. Kawarai, et al. (2007). "The neuronal sortilin-related receptor SORL1 is genetically associated with Alzheimer disease." Nat Genet **39**(2): 168-177.

Rogers, J., N. R. Cooper, S. Webster, J. Schultz, P. L. McGeer, et al. (1992). "Complement activation by beta-amyloid in Alzheimer disease." Proc Natl Acad Sci U S A **89**(21): 10016-10020.

Rogers, J., R. Li, D. Mastroeni, A. Grover, B. Leonard, et al. (2006). "Peripheral clearance of amyloid beta peptide by complement C3-dependent adherence to erythrocytes." Neurobiol Aging **27**(12): 1733-1739.

Rosemblit, N. and C. L. Chen (1994). "Regulators for the rat clusterin gene: DNA methylation and cis-acting regulatory elements." J Mol Endocrinol **13**(1): 69-76.

Rowe, J. A., J. M. Moulds, C. I. Newbold and L. H. Miller (1997). "P. falciparum rosetting mediated by a parasite-variant erythrocyte membrane protein and complement-receptor 1." Nature **388**(6639): 292-295.

Rowe, J. A., A. Raza, D. A. Diallo, M. Baby, B. Poudiougo, et al. (2002). "Erythrocyte CR1 expression level does not correlate with a HindIII restriction fragment length polymorphism in Africans; implications for studies on malaria susceptibility." Genes Immun **3**(8): 497-500.

Roy, S., B. Zhang, V. M. Lee and J. Q. Trojanowski (2005). "Axonal transport defects: a common theme in neurodegenerative diseases." Acta Neuropathol **109**(1): 5-13.

Rozen, S. and H. Skaletsky (2000). "Primer3 on the WWW for general users and for biologist programmers." Methods Mol Biol **132**: 365-386.

Rudinskiy, N., Y. Grishchuk, A. Vaslin, J. Puyal, A. Delacourte, et al. (2009). "Calpain hydrolysis of alpha- and beta2-adaptins decreases clathrin-dependent endocytosis and may promote neurodegeneration." J Biol Chem **284**(18): 12447-12458.

Ruuska, P. E., I. Ikaheimo, S. Silvennoinen-Kassinen, M. L. Kaar and A. Tiilikainen (1992). "Normal C3b receptor (CR1) genomic polymorphism in patients with insulin-dependent diabetes mellitus (IDDM): is the low erythrocyte CR1 expression an acquired phenomenon?" Clin Exp Immunol **89**(1): 18-21.

Salloway, S., R. Sperling, R. Keren, A. P. Porsteinsson, C. H. van Dyck, et al. (2011). "A phase 2 randomized trial of ELND005, scyllo-inositol, in mild to moderate Alzheimer disease." Neurology **77**(13): 1253-1262.

Samson, K. (2010). "NerveCenter: Phase III Alzheimer trial halted: Search for therapeutic biomarkers continues." Ann Neurol **68**(4): A9-A12.

Sandelin, A., W. Alkema, P. Engstrom, W. W. Wasserman and B. Lenhard (2004). "JASPAR: an open-access database for eukaryotic transcription factor binding profiles." Nucleic Acids Res **32**(Database issue): D91-94.

Saura, J., V. Petegnief, X. Wu, Y. Liang and S. M. Paul (2003). "Microglial apolipoprotein E and astroglial apolipoprotein J expression in vitro: opposite effects of lipopolysaccharide." J Neurochem **85**(6): 1455-1467.

Scaltriti, M., A. Santamaria, R. Paciucci and S. Bettuzzi (2004). "Intracellular clusterin induces G2-M phase arrest and cell death in PC-3 prostate cancer cells1." Cancer Res **64**(17): 6174-6182.

Schenk, D. (2002). "Amyloid-beta immunotherapy for Alzheimer's disease: the end of the beginning." Nat Rev Neurosci **3**(10): 824-828.

Schenk, D., R. Barbour, W. Dunn, G. Gordon, H. Grajeda, et al. (1999). "Immunization with amyloid-beta attenuates Alzheimer-disease-like pathology in the PDAPP mouse." Nature **400**(6740): 173-177.

Schenk, D., G. S. Basi and M. N. Pangalos (2012). "Treatment strategies targeting amyloid beta-protein." Cold Spring Harb Perspect Med **2**(9): a006387.

Schepeler, T., F. Mansilla, L. L. Christensen, T. F. Orntoft and C. L. Andersen (2007). "Clusterin expression can be modulated by changes in TCF1-mediated Wnt signaling." J Mol Signal **2**: 6.

Schjeide, B. M., C. Schnack, J. C. Lambert, C. M. Lill, J. Kirchheiner, et al. (2011). "The role of clusterin, complement receptor 1, and phosphatidylinositol binding clathrin assembly protein in Alzheimer disease risk and cerebrospinal fluid biomarker levels." Arch Gen Psychiatry **68**(2): 207-213.

Schnetz-Boutaud, N. C., J. Hoffman, J. E. Coe, D. G. Murdock, M. A. Pericak-Vance, et al. (2012). "Identification and Confirmation of an Exonic Splicing Enhancer Variation in Exon 5 of the Alzheimer Disease Associated PICALM Gene." Ann Hum Genet.

Schoch, S., F. Deak, A. Konigstorfer, M. Mozhayeva, Y. Sara, et al. (2001). "SNARE function analyzed in synaptobrevin/VAMP knockout mice." Science **294**(5544): 1117-1122.

Schrijvers, E. M., P. J. Koudstaal, A. Hofman and M. M. Breteler (2011). "Plasma clusterin and the risk of Alzheimer disease." JAMA **305**(13): 1322-1326.

Schwartz, C. M., A. Cheng, M. R. Mughal, M. P. Mattson and P. J. Yao (2010). "Clathrin assembly proteins AP180 and CALM in the embryonic rat brain." J Comp Neurol **518**(18): 3803-3818.

Selkoe, D. J. and M. S. Wolfe (2007). "Presenilin: Running with Scissors in the Membrane." Cell **131**(2): 215-221.

Seshadri, S., A. L. Fitzpatrick, M. A. Ikram, A. L. DeStefano, V. Gudnason, et al. (2010). "Genome-wide analysis of genetic loci associated with Alzheimer disease." JAMA **303**(18): 1832-1840.

Sessa, G., P. Podini, M. Mariani, A. Meroni, R. Spreafico, et al. (2004). "Distribution and signaling of TREM2/DAP12, the receptor system mutated in human polycystic lipomembraneous osteodysplasia with sclerosing leukoencephalopathy dementia." Eur J Neurosci **20**(10): 2617-2628.

Shaw, P., J. P. Lerch, J. C. Pruessner, K. N. Taylor, A. B. Rose, et al. (2007). "Cortical morphology in children and adolescents with different apolipoprotein E gene polymorphisms: an observational study." Lancet Neurol **6**(6): 494-500.

Shih, N. Y., J. Li, V. Karpitskii, A. Nguyen, M. L. Dustin, et al. (1999). "Congenital nephrotic syndrome in mice lacking CD2-associated protein." Science **286**(5438): 312-315.

Shim, Y. J., Y. J. Shin, S. Y. Jeong, S. W. Kang, B. M. Kim, et al. (2009). "Epidermal growth factor receptor is involved in clusterin-induced astrocyte proliferation." Neuroreport **20**(4): 435-439.

Shin, Y. J., S. W. Kang, S. Y. Jeong, Y. J. Shim, Y. H. Kim, et al. (2006). "Clusterin enhances proliferation of primary astrocytes through extracellular signal-regulated kinase activation." Neuroreport **17**(18): 1871-1875.

Siepel, A., G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, et al. (2005). "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes." Genome Res **15**(8): 1034-1050.

Sihlbom, C., P. Davidsson, M. Sjogren, L. O. Wahlund and C. L. Nilsson (2008). "Structural and quantitative comparison of cerebrospinal fluid glycoproteins in Alzheimer's disease patients and healthy individuals." Neurochem Res **33**(7): 1332-1340.

Singhrao, S. K., J. W. Neal, N. K. Rushmere, B. P. Morgan and P. Gasque (1999). "Differential expression of individual complement regulators in the brain and choroid plexus." Lab Invest **79**(10): 1247-1259.

Sleegers, K., J. C. Lambert, L. Bertram, M. Cruts, P. Amouyel, et al. (2010). "The pursuit of susceptibility genes for Alzheimer's disease: progress and prospects." Trends Genet **26**(2): 84-93.

Smit, A., Hubley, R., Green, P. (1996-2010). "RepeatMasker Open-3.0." from http://repeatmasker.org.

So, A., S. Sinnemann, D. Huntsman, L. Fazli and M. Gleave (2005). "Knockdown of the cytoprotective chaperone, clusterin, chemosensitizes human breast cancer cells both in vitro and in vivo." Mol Cancer Ther **4**(12): 1837-1849.

Stewart, W. F., C. Kawas, M. Corrada and E. J. Metter (1997). "Risk of Alzheimer's disease and duration of NSAID use." Neurology **48**(3): 626-632.

Su, A. I., T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, et al. (2004). "A gene atlas of the mouse and human protein-encoding transcriptomes." Proc Natl Acad Sci U S A **101**(16): 6062-6067.

Suzuki, M., H. Tanaka, A. Tanimura, K. Tanabe, N. Oe, et al. (2012). "The clathrin assembly protein PICALM is required for erythroid maturation and transferrin internalization in mice." PLoS ONE **7**(2): e31854.

Suzuki, T., M. Tozuka, Y. Kazuyoshi, M. Sugano, T. Nakabayashi, et al. (2002). "Predominant apolipoprotein J exists as lipid-poor mixtures in cerebrospinal fluid." Ann Clin Lab Sci **32**(4): 369-376.

Swerdlow, R. H. and S. M. Khan (2004). "A "mitochondrial cascade hypothesis" for sporadic Alzheimer's disease." Med Hypotheses **63**(1): 8-20.

Takahashi, K., M. Prinz, M. Stagi, O. Chechneva and H. Neumann (2007). "TREM2-transduced myeloid precursors mediate nervous tissue debris clearance and facilitate recovery in an animal model of multiple sclerosis." PLoS Med **4**(4): e124.

Takase, O., A. W. Minto, T. S. Puri, P. N. Cunningham, A. Jacob, et al. (2008). "Inhibition of NF-kappaB-dependent Bcl-xL expression by clusterin promotes albumin-induced tubular cell apoptosis." Kidney Int **73**(5): 567-577.

Tamboli, I. Y., E. Barth, L. Christian, M. Siepmann, S. Kumar, et al. (2010). "Statins promote the degradation of extracellular amyloid {beta}-peptide by microglia via stimulation of exosome-associated insulin-degrading enzyme (IDE) secretion." J Biol Chem **285**(48): 37405-37414.

Tan, Z. S. and S. Seshadri (2010). "Inflammation in the Alzheimer's disease cascade: culprit or innocent bystander?" Alzheimers Res Ther **2**(2): 6.

Tanaka, N., S. Abe-Dohmae, N. Iwamoto and S. Yokoyama (2011). "Roles of ATP-binding cassette transporter A7 in cholesterol homeostasis and host defense system." J Atheroscler Thromb **18**(4): 274-281.

Taniguchi-Sidle, A. and D. E. Isenman (1994). "Interactions of human complement component C3 with factor B and with complement receptors type 1 (CR1, CD35) and type 3 (CR3, CD11b/CD18) involve an acidic sequence at the N-terminus of C3 alpha'-chain." J Immunol **153**(11): 5285-5302.

Tanzi, R. E. (2012). "The genetics of Alzheimer disease." Cold Spring Harb Perspect Med **2**(10).

Tanzi, R. E. and L. Bertram (2005). "Twenty years of the Alzheimer's disease amyloid hypothesis: a genetic perspective." Cell **120**(4): 545-555.

Tas, S. W., L. B. Klickstein, S. F. Barbashov and A. Nicholson-Weller (1999). "C1q and C4b bind simultaneously to CR1 and additively support erythrocyte adhesion." J Immunol **163**(9): 5056-5063.

Tateno, H., H. Li, M. J. Schur, N. Bovin, P. R. Crocker, et al. (2007). "Distinct endocytic mechanisms of CD22 (Siglec-2) and Siglec-F reflect roles in cell signaling and innate immunity." Mol Cell Biol **27**(16): 5699-5710.

Tebar, F., S. K. Bohlander and A. Sorkin (1999). "Clathrin assembly lymphoid myeloid leukemia (CALM) protein: localization in endocytic-coated pits, interactions with clathrin, and the impact of overexpression on clathrin-mediated traffic." Mol Biol Cell **10**(8): 2687-2702.

Teer, J. K., L. L. Bonnycastle, P. S. Chines, N. F. Hansen, N. Aoyama, et al. (2010). "Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing." Genome Res **20**(10): 1420-1431.

Terry, R. D., E. Masliah, D. P. Salmon, N. Butters, R. DeTeresa, et al. (1991). "Physical basis of cognitive alterations in Alzheimer's disease: synapse loss is the major correlate of cognitive impairment." Ann Neurol **30**(4): 572-580.

Thambisetty, M., Y. An, A. Kinsey, D. Koka, M. Saleem, et al. (2012). "Plasma clusterin concentration is associated with longitudinal brain atrophy in mild cognitive impairment." Neuroimage **59**(1): 212-217.

Thambisetty, M., A. Simmons, L. Velayudhan, A. Hye, J. Campbell, et al. (2010). "Association of plasma clusterin concentration with severity, pathology, and progression in Alzheimer disease." Arch Gen Psychiatry **67**(7): 739-748.

Thinakaran, G. and E. H. Koo (2008). "Amyloid precursor protein trafficking, processing, and function." J Biol Chem **283**(44): 29615-29619.

Thomas-Tikhonenko, A., I. Viard-Leveugle, M. Dews, P. Wehrli, C. Sevignani, et al. (2004). "Myc-transformed epithelial cells down-regulate clusterin, which inhibits their growth in vitro and carcinogenesis in vivo." Cancer Res **64**(9): 3126-3136.

Thomas, R. S., M. J. Lelos, M. A. Good and E. J. Kidd (2011). "Clathrin-mediated endocytic proteins are upregulated in the cortex of the Tg2576 mouse model of Alzheimer's disease-like amyloid pathology." Biochem Biophys Res Commun **415**(4): 656-661.

Treangen, T. J. and S. L. Salzberg (2012). "Repetitive DNA and next-generation sequencing: computational challenges and solutions." Nat Rev Genet **13**(1): 36-46.

Treusch, S., S. Hamamichi, J. L. Goodman, K. E. Matlack, C. Y. Chung, et al. (2011). "Functional links between Abeta toxicity, endocytic trafficking, and Alzheimer's disease risk factors in yeast." Science **334**(6060): 1241-1245.

Triplett, J. W. and D. A. Feldheim (2012). "Eph and ephrin signaling in the formation of topographic maps." Semin Cell Dev Biol **23**(1): 7-15.

Trougakos, I. P., J. Y. Djeu, E. S. Gonos and D. A. Boothman (2009). "Advances and challenges in basic and translational research on clusterin." Cancer Res **69**(2): 403-406.

Tycko, B., L. Feng, L. Nguyen, A. Francis, A. Hays, et al. (1996). "Polymorphisms in the human apolipoprotein-J/clusterin gene: ethnic variation and distribution in Alzheimer's disease." Hum Genet **98**(4): 430-436.

Vakeva, A., P. Laurila and S. Meri (1993). "Co-deposition of clusterin with the complement membrane attack complex in myocardial infarction." Immunology **80**(2): 177-182.

Valent, P. and P. Bettelheim (1992). "Cell surface structures on human basophils and mast cells: biochemical and functional characterization." Adv Immunol **52**: 333-423.

Varki, A. (2009). "Natural ligands for CD33-related Siglecs?" Glycobiology **19**(8): 810-812.

Vasiliou, V., K. Vasiliou and D. W. Nebert (2009). "Human ATP-binding cassette (ABC) transporter family." Hum Genomics **3**(3): 281-290.

Veldhuisen, B., P. C. Ligthart, G. Vidarsson, I. Roels, C. C. Folman, et al. (2011). "Molecular analysis of the York antigen of the Knops blood group system." Transfusion **51**(7): 1389-1396.

Vetrivel, K. S. and G. Thinakaran (2006). "Amyloidogenic processing of beta-amyloid precursor protein in intracellular compartments." Neurology **66**(2 Suppl 1): S69-73.

Villemagne, V. L., M. T. Fodero-Tavoletti, K. E. Pike, R. Cappai, C. L. Masters, et al. (2008). "The ART of loss: Abeta imaging in the evaluation of Alzheimer's disease and other dementias." Mol Neurobiol **38**(1): 1-15.

Walport, M., Y. C. Ng and P. J. Lachmann (1987). "Erythrocytes transfused into patients with SLE and haemolytic anaemia lose complement receptor type 1 from their cell surface." Clin Exp Immunol **69**(3): 501-507.

Walshe, C. A., S. A. Beers, R. R. French, C. H. Chan, P. W. Johnson, et al. (2008). "Induction of cytosolic calcium flux by CD20 is dependent upon B Cell antigen receptor signaling." J Biol Chem **283**(25): 16971-16984.

Walter, R. B., B. W. Raden, R. Zeng, P. Hausermann, I. D. Bernstein, et al. (2008). "ITIM-dependent endocytosis of CD33-related Siglecs: role of intracellular domain, tyrosine phosphorylation, and the tyrosine phosphatases, Shp1 and Shp2." J Leukoc Biol **83**(1): 200-211.

Wang, K., S. P. Dickson, C. A. Stolle, I. D. Krantz, D. B. Goldstein, et al. (2010). "Interpretation of association signals and identification of causal variants from genome-wide association studies." Am J Hum Genet **86**(5): 730-742.

Wang, L., Y. Guo, W. J. Huang, X. Ke, J. L. Poyet, et al. (2001). "Card10 is a novel caspase recruitment domain/membrane-associated guanylate kinase family member that interacts with BCL10 and activates NF-kappa B." J Biol Chem **276**(24): 21405-21409.

Wang, S. C., B. Oelze and A. Schumacher (2008). "Age-specific epigenetic drift in late-onset Alzheimer's disease." PLoS ONE **3**(7): e2698.

Wang, Y. T. and D. J. Linden (2000). "Expression of cerebellar long-term depression requires postsynaptic clathrin-mediated endocytosis." Neuron **25**(3): 635-647.

Wechsler-Reya, R., D. Sakamuro, J. Zhang, J. Duhadaway and G. C. Prendergast (1997). "Structural analysis of the human BIN1 gene. Evidence for tissue-specific transcriptional regulation and alternate RNA splicing." J Biol Chem **272**(50): 31453-31458.

Wei, Q., L. Wang, Q. Wang, W. D. Kruger and R. L. Dunbrack, Jr. (2010). "Testing computational prediction of missense mutation phenotypes: functional characterization of 204 mutations of human cystathionine beta synthase." Proteins **78**(9): 2058-2074.

Weis, J. H., C. C. Morton, G. A. Bruns, J. J. Weis, L. B. Klickstein, et al. (1987). "A complement receptor locus: genes encoding C3b/C4b receptor and C3d/Epstein-Barr virus receptor map to 1q32." J Immunol **138**(1): 312-315.

Wen, Y., A. Miyashita, N. Kitamura, T. Tsukie, Y. Saito, et al. (2013). "SORL1 is genetically associated with neuropathologically characterized late-onset Alzheimer's disease." J Alzheimers Dis **35**(2): 387-394.

Wilson, J. G., E. E. Murphy, W. W. Wong, L. B. Klickstein, J. H. Weis, et al. (1986). "Identification of a restriction fragment length polymorphism by a CR1 cDNA that correlates with the number of CR1 on erythrocytes." J Exp Med **164**(1): 50-59.

Wong, W. W. (1990). "Structural and functional correlation of the human complement receptor type 1." J Invest Dermatol **94**(6 Suppl): 64S-67S.

Wong, W. W., J. M. Cahill, M. D. Rosen, C. A. Kennedy, E. T. Bonaccio, et al. (1989). "Structure of the human CR1 gene. Molecular basis of the structural and quantitative polymorphisms and identification of a new CR1-like allele." J Exp Med **169**(3): 847-863.

Wong, W. W. and S. A. Farrell (1991). "Proposed structure of the F' allotype of human CR1. Loss of a C3b binding site may be associated with altered function." J Immunol **146**(2): 656-662.

Wong, W. W., J. G. Wilson and D. T. Fearon (1983). "Genetic regulation of a structural polymorphism of human C3b receptor." J Clin Invest **72**(2): 685-693.

Wright, S. D. and S. C. Silverstein (1982). "Tumor-promoting phorbol esters stimulate C3b and C3b' receptor-mediated phagocytosis in cultured human monocytes." J Exp Med **156**(4): 1149-1164.

Wu, F., Y. Matsuoka, M. P. Mattson and P. J. Yao (2009). "The clathrin assembly protein AP180 regulates the generation of amyloid-beta peptide." Biochem Biophys Res Commun **385**(2): 247-250.

Wu, F. and P. J. Yao (2009). "Clathrin-mediated endocytosis and Alzheimer's disease: an update." Ageing Res Rev **8**(3): 147-149.

Wyss-Coray, T., F. Yan, A. H. Lin, J. D. Lambris, J. J. Alexander, et al. (2002). "Prominent neurodegeneration and increased plaque formation in complement-inhibited Alzheimer's mice." Proc Natl Acad Sci U S A **99**(16): 10837-10842.

Xiang, L., J. R. Rundles, D. R. Hamilton and J. G. Wilson (1999). "Quantitative alleles of CR1: coding sequence analysis and comparison of haplotypes in two ethnic groups." J Immunol **163**(9): 4939-4945.

Xiao, Q., S. C. Gil, P. Yan, Y. Wang, S. Han, et al. (2012). "Role of phosphatidylinositol clathrin assembly lymphoid-Myeloid Leukemia (PICALM) in intracellular amyloid precursor protein (APP) processing and amyloid plaque pathogenesis." J Biol Chem.

Xie, Z., M. E. Harris-White, P. A. Wals, S. A. Frautschy, C. E. Finch, et al. (2005). "Apolipoprotein J (clusterin) activates rodent microglia in vivo and in vitro." J Neurochem **93**(4): 1038-1046.

Xu, Q., A. Bernardo, D. Walker, T. Kanegawa, R. W. Mahley, et al. (2006). "Profile and regulation of apolipoprotein E (ApoE) expression in the CNS in mice with targeting of green fluorescent protein gene to the ApoE locus." J Neurosci **26**(19): 4985-4994.

Yamazaki, T., J. Masuda, T. Omori, R. Usui, H. Akiyama, et al. (2009). "EphA1 interacts with integrin-linked kinase and regulates cell morphology and motility." J Cell Sci **122**(Pt 2): 243-255.

Yang, C. R., K. Leskov, K. Hosley-Eberlein, T. Criswell, J. J. Pink, et al. (2000). "Nuclear clusterin/XIP8, an x-ray-induced Ku70-binding protein that signals cell death." Proc Natl Acad Sci U S A **97**(11): 5907-5912.

Yao, P. J. (2004). "Synaptic frailty and clathrin-mediated synaptic vesicle trafficking in Alzheimer's disease." Trends Neurosci **27**(1): 24-29.

Yao, P. J., R. S. Petralia, I. Bushlin, Y. Wang and K. Furukawa (2005). "Synaptic distribution of the endocytic accessory proteins AP180 and CALM." J Comp Neurol **481**(1): 58-69.

Yasojima, K., C. Schwab, E. G. McGeer and P. L. McGeer (1999). "Up-regulated production and activation of the complement system in Alzheimer's disease brain." Am J Pathol **154**(3): 927-936.

Yerbury, J. J., S. Poon, S. Meehan, B. Thompson, J. R. Kumita, et al. (2007). "The extracellular chaperone clusterin influences amyloid formation and toxicity by interacting with prefibrillar structures." FASEB J **21**(10): 2312-2322.

Yokoi, H., A. Myers, K. Matsumoto, P. R. Crocker, H. Saito, et al. (2006). "Alteration and acquisition of Siglecs during in vitro maturation of CD34+ progenitors into human mast cells." Allergy **61**(6): 769-776.

Yoon, S. H. and D. T. Fearon (1985). "Characterization of a soluble form of the C3b/C4b receptor (CR1) in human plasma." J Immunol **134**(5): 3332-3338.

Yu, J. T. and L. Tan (2012). "The role of clusterin in Alzheimer's disease: pathways, pathogenesis, and therapy." Mol Neurobiol **45**(2): 314-326.

Zamrini, E., G. McGwin and J. M. Roseman (2004). "Association between statin use and Alzheimer's disease." Neuroepidemiology **23**(1-2): 94-98.

Zanjani, H., C. E. Finch, C. Kemper, J. Atkinson, D. McKeel, et al. (2005). "Complement activation in very early Alzheimer disease." Alzheimer Dis Assoc Disord **19**(2): 55-66.

Zhang, B., Y. H. Koh, R. B. Beckstead, V. Budnik, B. Ganetzky, et al. (1998). "Synaptic vesicle size and number are regulated by a clathrin adaptor protein required for endocytosis." Neuron **21**(6): 1465-1475.

Zlokovic, B. V. (1996). "Cerebrovascular transport of Alzheimer's amyloid beta and apolipoproteins J and E: possible anti-amyloidogenic role of the blood-brain barrier." Life Sci **59**(18): 1483-1497.

Zotova, E., J. A. Nicoll, R. Kalaria, C. Holmes and D. Boche (2010). "Inflammation in Alzheimer's disease: relevance to pathogenesis and therapy." Alzheimers Res Ther **2**(1): 1.

Zou, M., E. Y. Baitei, A. S. Alzahrani, R. S. Parhar, F. A. Al-Mohanna, et al. (2011). "Mutation prediction by PolyPhen or functional assay, a detailed comparison of CYP27B1 missense mutations." Endocrine **40**(1): 14-20.

Zwain, I. H., J. Grima and C. Y. Cheng (1994). "Regulation of clusterin secretion and mRNA expression in astrocytes by cytokines." Mol Cell Neurosci **5**(3): 229-237.

Appendix

2.1 – Sample pooling strategy

Pool No.	Sample No.	ID	Conc. (ng/µl)	Dilution	Volume added to pool (µl)*
1	1	AD219	106.08	-	5.66
	2	AD232	112.08	-	5.35
	3	M547	191.26	-	3.14
	4	AD218	225.07	-	2.67
	5	AD236	269.44	-	2.23
	6	M659	320.08	-	1.87
	7	AD221	320.08	-	1.87
	8	M523	349.18	-	1.72
	9	M551	394.15	-	1.52
	10	M641	400.98	-	1.50
	11	M604	414.92	-	1.45
	12	AD235	417.42	-	1.44
2	13	AD222	425.85	-	1.41
	14	AD224	447.25	-	1.34
	15	AD245	463.41	-	1.29
	16	AD233	479.74	-	1.25
	17	AD197	493.48	-	1.22
	18	M596	503.88	-	1.19
	19	M565	520.05	-	1.15
	20	AD220	524.70	-	1.14
	21	AD238	542.07	-	1.11
	22	M540	543.15	-	1.10
	23	M546	568.51	-	1.06
	24	AD257	612.08	1 in 2	1.96
3	25	AD231	618.17	1 in 2	1.94
	26	M605	644.26	1 in 2	1.86
	27	AD227	650.27	1 in 2	1.85
	28	AD251	667.24	1 in 2	1.80
	29	AD207	670.96	1 in 2	1.79
	30	AD246	686.33	1 in 2	1.75
	31	AD212	698.82	1 in 2	1.72
	32	M651	704.82	1 in 2	1.70
	33	M530	719.72	1 in 2	1.67
	34	AD253	738.23	1 in 2	1.63
	35	AD203	741.81	1 in 2	1.62
	36	AD210	780.08	1 in 2	1.54
4	37	AD255	792.41	1 in 2	1.51
	38	M571	820.01	1 in 2	1.46
	39	M094	821.84	1 in 2	1.46
	40	AD206	821.87	1 in 2	1.46
	41	AD249	886.29	1 in 2	1.35
	42	M589	889.20	1 in 2	1.35
	43	M531	890.77	1 in 2	1.35
	44	M593	895.00	1 in 2	1.34
	45	51/05	978.55	1 in 2	1.23
	46	AD201	1015.50	1 in 2	1.18
	47	AD216	1022.11	1 in 2	1.17
	48	M576	1031.98	1 in 2	1.16

5	49	M643	1034.59	1 in 2	1.16
	50	AD243	1048.06	1 in 2	1.14
	51	M579	1057.61	1 in 2	1.13
	52	M562	1111.66	1 in 2	1.08
	53	M543	1134.91	1 in 2	1.06
	54	M522	1144.07	1 in 2	1.05
	55	AD213	1146.90	1 in 2	1.05
	56	AD244	1170.06	1 in 2	1.03
	57	AD225	1267.00	1 in 3	1.42
	58	AD199	1283.91	1 in 3	1.40
	59	M524	1294.32	1 in 3	1.39
	60	AD242	1301.59	1 in 3	1.38
6	61	AD248	1394.94	1 in 3	1.29
	62	M573	1455.10	1 in 3	1.24
	63	M526	1518.86	1 in 3	1.19
	64	M647	1679.06	1 in 3	1.07
	65	M528	1679.65	1 in 3	1.07
	66	M646	1719.55	1 in 3	1.05
	67	AD117	1768.54	1 in 3	1.02
	68	AD252	1802.82	1 in 3	1.00
	69	M577	1841.86	1 in 4	1.30
	70	M599	1848.20	1 in 4	1.30
	71	M590	2018.11	1 in 4	1.19
	72	AD254	2046.27	1 in 4	1.17
7	73	M637	2073.50	1 in 4	1.16
	74	M645	2124.15	1 in 4	1.13
	75	AD211	2130.62	1 in 4	1.13
	76	M648	2138.16	1 in 4	1.12
	77	M594	2214.87	1 in 4	1.08
	78	AD208	2225.99	1 in 4	1.08
	79	M575	2291.49	1 in 4	1.05
	80	M639	2346.48	1 in 4	1.02
	81	M527	2441.02	1 in 5	1.23
	82	M536	2531.52	1 in 5	1.19
	83	M649	2651.15	1 in 5	1.13
	84	M644	2882.31	1 in 5	1.04
8	85	M529	2935.57	1 in 5	1.02
	86	AD115	3063.75	1 in 6	1.18
	87	M638	3096.46	1 in 6	1.16
	88	M642	3161.98	1 in 6	1.14
	89	AD141	4793.94	1 in 8	1.00
	90	AD146	4833.60	1 in 9	1.12
	91	AD112	5548.82	1 in 10	1.08
	92	AD107	5999.41	1 in 10	1.00
	93	AD123	6342.72	1 in 11	1.04
	94	AD125	7340.78	1 in 13	1.06
	95	AD157	7643.78	1 in 13	1.02
	96	AD109	7937.67	1 in 14	1.06

Details of samples used in NGS project, their concentrations, and the pooling strategy utilised. *Where a dilution is stated in the previous column, this is the volume of diluted DNA added to the pool.

2.2 – Perl script for obtaining 1000 genomes project frequency data

perl.pl

```
#!/usr/bin/perl

use strict;
use warnings;
use modules;

my $filename1 = 'phase1_integrated_calls.20101123.ALL.panel.txt';       # Sample ID and Populations
my $filename2 = 'PICALM_Extra_Variants_160112';                # Data_File, downloaded using 'tabix -hf ...'
my $print_allele_count = "Yes";                # "Yes" - print, "No" - skip
my $print_sample_ID = "No";                    # "Yes" - print, "No" - skip
my $str = modules::readData($filename1,$filename2,$print_allele_count,$print_sample_ID);
```

Contents of the perl script, written by Hui Shi, used for obtaining frequency estimates for variants of interest based on 1000 genomes phase 1 data. Shown above is the version of the script edited to obtain the 1000 genomes frequency data for the PICALM rs3851179 LD block. Toggling between "Yes" and "No" for the sections of the perl script print_allele_count and print_sample_ID determines whether these pieces of information are included in the output file.

2.3 – Script to interleave paired-end reads

*Interleving.pl

```
#!/usr/bin/Perl

$filenameA = $ARGV[0];
$filenameB = $ARGV[1];
$filenameOut = $ARGV[2];

open $FILEA, "< $filenameA";
open $FILEB, "< $filenameB";
open $OUTFILE, "> $filenameOut";

while(<$FILEA>) {
        # $_ holds read name
        $readname1 = $_;
        print $OUTFILE $readname1;

        $read1 = <$FILEA>;
        print $OUTFILE $read1;

        $superfluous_plus = <$FILEA>;
        print $OUTFILE $superfluous_plus;

        $qual1 = <$FILEA>;
        print $OUTFILE $qual1;

        $readname2 = <$FILEB>;
        die("Readnames differ: $readname1 != $readname2") if ($readname1 != $readname2);
        print $OUTFILE $readname2;

        $read2 = <$FILEB>;
        # Reverse complement read2
        chop $read2;

        $read2 = reverse($read2);
        $read2 =~ tr/ACGTacgt/TGCAtgca/;
        print $OUTFILE "$read2\n";

        $superfluous_plus = <$FILEB>;
        print $OUTFILE $superfluous_plus;

        $qual2 = <$FILEB>;
        chop $qual2;

        # Reverse qual
        $qual2 = reverse($qual2);
        print $OUTFILE "$qual2\n";
}
close($OUTFILE)
```

The script opens read file A (i.e. the forward reads) and interrogates read file B for the read with the matching header (i.e. the pairs, aligned to the reverse strand), calculating the reverse complement of the second read and reversing the quality

scores. The output contains the interleaved reads in the format 1Read1, 1Read2; 2Read1, 2Read2, which can then be used to run the BFAST alignment.

2.4 – Perl script for converting WTCCC2 data from hg18 to hg19

*liftOverGen (copy).pl

```
#! /usr/bin/perl
use strict;
use warnings;

my $count2 =0;
my $rs;
my $bp;
my ($gen, $lo) = @ARGV;
open (OUT2, '>WARNING.txt');
open (OUT, '>OUTPUT.txt');
open (GEN, $gen); ##open gen file
while (<GEN>)
        {
        $count2++;
        my $count =0;
        chomp;
        my @line = split(" ",$_);
        if (/^(\S+)\s+\S+\s+(\S+)/)
                {
                $rs = $1;
                $bp = $2;
                open (LO, $lo); ##open liftover data
                while (<LO>)
                        {
                        chomp;
                        if (/^\S+\s+(\d+)\s+\d+\s+(\S+)/)
                                {
                                my $lo_bp = $1;
                                my $lo_rs = $2;
                                if ($rs eq $lo_rs)
                                        {
                                        print "ln$count2 MATCH\n";
                                        $line[2] = $lo_bp;
                                        print OUT "@line\n";
                                        $count++;
                                        }
                                }
                        }
                }
        if ($count == 0)        {print OUT2 "$rs $bp\n";print "ln$count2 NO MATCH\n";}
        }

close;
```

Contents of the perl script liftOverGen.pl used in the preparation of WTCCC2 data for imputation. The script converts the coordinates of variants from hg18 to hg19.

2.5 – Perl script to recode WTCCC2 .samples files

recode_WTCCC2_sample.pl

```
#! /usr/bin/perl
#UCSC_Conservation.pl to convert sample file from WTCCC data
use strict;
use warnings;

my $newpheno;
my ($file) = @ARGV;

open (OUT, '>output.sample');
open (IN, $file);
while (<IN>)
        {
        chomp;
        if (/(\S+\s\S+\s)(\S+\s)\S+\s(\S+\s)(\S+)/)
                {
                my $id = $1;
                my $miss = $2;
                my $sex = $3;
                my $pheno = $4;
                if (/^ID_1/)    {print OUT "ID_1 ID_2 missing sex phenotype centre\n";}
                elsif (/^0\s0\s0/)      {print OUT "0 0 0 D B D\n";}
                elsif (($id =~ m/\S+/))
                        {
                        if ($miss eq '-9 ')     {$miss = 'NA ';}
                        if ($sex eq '-9 ')      {$sex = 'NA ';}
                        $newpheno = '0';
                        print OUT "$id$miss$sex$newpheno 3\n";
                        }
                }
        }
close;
```

Contents of the perl script recode_WTCCC2.sample.pl, utilised in the preparation of WTCCC2 data for imputation. The script recodes phenotype information to a format which Impute2 will recognise, as well as adding centre information and changing the

format of missing values to ones Impute2 can process. The script is altered at line 27 depending on the centre (58C = 2, NBS = 3).

2.6 – Perl script to remove duplicate lines

remove_dup_lines.pl

```
#! /usr/bin/perl
#remove duplicate lines from impute output file so SNPTEST can be run using the -overlap function.
use strict;
use warnings;

my ($file) = @ARGV;
my $memory=0;
open (OUT, ">$file\_duplicates_removed");
open (OUT2,">$file\_filtered_duplicates");
open (IN, $file);
while (<IN>)
        {
        chomp;
        my @array = split(" ",$_);
        my $bp = $array[2];
        if ($bp == $memory)     {print OUT2 "@array\n";$memory = $bp;}
        else {print OUT "@array\n";$memory = $bp;}
        }
close;
```

Contents of the perl script used to remove duplicate positions from the imputed data sets, enabling the –overlap command to be used when running association testing with snptest.

2.7 – Perl script to calculate conservation at positions of interest

*conservation.pl

```
#!/usr/bin/perl
# conservation.pl
use strict; use warnings;

my $bp = 0;
my $cons =0;
my ($uscs, $input) = @ARGV;

open (INPUT, $input);

while (<INPUT>) {

my @array = split(" ",$_);

        open (USCS, $uscs);

        while (<USCS>)  {

                chomp;

                if (/^(\d+)/){($bp, $cons) = split(" ",$_);}

                if ($array[0] eq $bp) {print $array[0];print "\t";print $cons;print "\n";}

                }
        }

        close:
```

Contents of the perl script used in calculating the level of conservation at the positions affected by variants found in the NGS data.

2.8 – Perl script to calculate LD between variants of interest

LD_Calculator.pl

```
#! /usr/bin/perl
#LD_Calculator
use strict;
use warnings;

my ($rs,$chr,$bp,$inputfile) = @ARGV;
my $n;
my $r;
my $d;
if ("-e $rs\_LD_OUTPUT"){system("rm $rs\_LD_OUTPUT")};
open (OUTPUT, ">>$rs\_LD_OUTPUT");
print OUTPUT "SNP1\tSNP2\tN\tR2\tD'\n";
my $startbp = ($bp - 1);
my $endbp = ($bp + 1);

my $tabix = "tabix -hf ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/integrated_call_sets/ALL.chr
$chr.integrated_phase1_v3.20101123.snps_indels_svs.genotypes.vcf.gz $chr:$startbp-$endbp > $rs.vcf";

system($tabix);

system("perl vcf-subset -c Phase1_Samples_GBR_CEU $rs.vcf > $rs\_CEU.vcf");

open(FILE, $inputfile) || die "Cannot open $inputfile";
while (<FILE>)  {

                chomp;
                my ($chr2,$bp2,$rs2) = split("\t",$_);
                if (length($rs2) < 3)   {$rs2 = $bp2;}
                $startbp = ($bp2 - 1);
                $endbp = ($bp2 + 1);

                $tabix = "tabix -hf ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/integrated_call_sets/ALL.chr
$chr.integrated_phase1_v3.20101123.snps_indels_svs.genotypes.vcf.gz $chr2:$startbp-$endbp > $rs2.vcf";

                system($tabix);

                system("perl vcf-subset -c Phase1_Samples_GBR_CEU $rs2.vcf > $rs2\_CEU.vcf");

                mergevcf($rs,$bp,$rs2,$bp2,$chr);

                system ("vcftools --vcf $rs\_$rs2.vcf --hap-r2 --out $rs\_$rs2");

                my $ld = "$rs\_$rs2.hap.ld";
                open (LD, $ld) || die;
                while (<LD>)    {
                                if ((/^$chr\t$bp\t$bp2\t(\d+)\t(\S+)\t\S+\t(\S+)/) || (/^$chr\t$bp2\t$bp\t(\d+)\t(\S+)\t\S+\t(\S+)/))   {

                                                $n = $1;
                                                $r = $2;
                                                $d = $3;

                                                print OUTPUT "$rs\t$rs2\t$n\t$r\t$d\n";
                                                print "\n\n**LD for $rs Vs $rs2 for $n samples**\n";
                                                print "\tR2=$r\n";
                                                print "\tD'=$d\n\n";

                                                }

                                }
                }

system("rm *.vcf");
system("rm *.vcfidx");
system("rm *.log");
system("rm *.ld");
system("rm *.tbi");
close;

#-------------------------------------------
sub mergevcf    {

my ($rs,$bp,$rs2,$bp2,$chr) = @_;
if ("-e $rs\_$rs2.vcf"){system("rm $rs\_$rs2.vcf")};
my $file1 = "$rs\_CEU.vcf";
my $file2 = "$rs2\_CEU.vcf";
open (OUT, ">>$rs\_$rs2.vcf");
open (FILE1, $file1);
while(<FILE1>)  {
                chomp;
                if (/^#/) {print OUT $_;print OUT "\n";}
                elsif (/^$chr\t$bp/)    {print OUT $_;print OUT "\n";}
                }

open (FILE2, $file2);
while (<FILE2>) {
                chomp;
                if (/^$chr\t$bp2/)      {print OUT $_;print OUT "\n";}
                }
close FILE2;
close FILE1;
close OUT;
}
1;
```

Contents of the perl script used to calculate the level of LD between index GWAS SNPs and variants detected by NGS.

6.1 – *CLU* variants at the sites of ENCODE TFBS

Coordinate	rsID	ENC_TFBS
27466157	rs1532276	GATA-2,TAF1
27466181	rs1532277	GATA-2,TAF1,USF-1
27466315	rs1532278	GATA-2,Max,NANOG_(SC-33759),USF2,TAF1,USF-1,USF1_(SC-8983)
27469971	rs9331883	FOXA2_(SC-6554),FOXA1_(SC-101058),FOXA1_(C-20)
27470010	-	FOXA2_(SC-6554),FOXA1_(SC-101058),FOXA1_(C-20)
27471673	-	BATF,eGFP-FOS,CEBPB,c-Fos,GATA-2,JunD,STAT3,c-Jun,RFX5_(N-494)
27471748	-	BATF,eGFP-FOS,FOSL2,eGFP-JunD,c-Fos,GATA-2,JunD,TCF4,STAT3,c-Jun,RFX5_(N-494)
27474202	rs9314349	Pol2(b)
27474541	rs117148275	TAF1,GR,AP-2alpha,AP-2gamma,Pol2,GATA-1
27474587	rs56025648	GR,AP-2alpha,AP-2gamma,Pol2,GATA-1
27474599	rs1982229	GR,GATA-2,AP-2alpha,AP-2gamma,Pol2,GATA-1
27474871	rs77336101	p300_(N-15),Pol2-4H8,GATA-2,c-Fos,p300,YY1_(C-20),RFX5_(N-494),MafK_(ab50322),JunD,STAT3,c-Jun,GATA-1

Full list of the ENCODE TFBS falling at the positions of variants of interest from the NGS project data in the *CLU* region.

6.2 - *PICALM* variants at the sites of ENCODE TFBS

Coordinate	rsID	ENC_TFBS
85780073	rs3016326	Egr-1,RFX5_(N-494),YY1,HEY1,USF2,SP1,HA-E2F1,PU.1,NRSF,GTF2F1_(RAP-74),TCF12,YY1_(C-20),Pol2,NFKB,E2F6,USF1_(SC-8983),Pol2(phosphoS2),ETS1,TAF1,Pol2-4H8,TBP,Pbx3,GATA2_(CG2-96),GABP,SRF,GATA-1,Sin3Ak-20,c-Myc,USF-1,ELF1_(SC-631)
85780448	rs3016327	TAF1,Egr-1,Pol2,TCF12,ZBTB7A_(SC-34508),PAX5-N19,USF1_(SC-8983),IRF4_(M-17),HEY1,RXRA,SP1,NFKB,USF-1,c-Myc,Pol2-4H8
85780582	rs10898433	TAF1,Pol2,TCF12,ZBTB7A_(SC-34508),PAX5-N19,IRF4_(M-17),HEY1,E2F6_(H-50),SP1,NFKB,c-Myc,Pol2-4H8,Oct-2,POU2F2
85780924	-	eGFP-JunD,TBP,c-Jun,TAF1,SRF,Pol2,YY1,Sin3Ak-20,NRSF,CEBPB,p300_(N-15),Pol2-4H8,GATA-2,YY1_(C-20),GABP,ELF1_(SC-631),BCL11A,GATA2_(CG2-96),FOXA1_(C-20),CCNT2,Pol2(b),Mxi1_(bHLH),GATA3_(SC-268),BATF,STAT3,SMC3_(ab9263),Rad21,JunD,GTF2F1_(RAP-74),c-Fos
85780962	rs188367538	eGFP-JunD,TBP,c-Jun,TAF1,SRF,Pol2,YY1,Sin3Ak-20,NRSF,CEBPB,p300_(N-15),Pol2-4H8,GATA-2,NF-YB,YY1_(C-20),GABP,ELF1_(SC-631),BCL11A,PU.1,GATA2_(CG2-96),FOXA1_(C-20),CCNT2,Pol2(b),Mxi1_(bHLH),GATA3_(SC-268),BATF,STAT3,SMC3_(ab9263),Rad21,JunD,GTF2F1_(RAP-74),c-Fos

Full list of the ENCODE TFBS falling at the positions of variants of interest from the NGS project data in the *PICALM* region.

6.3 - rs3851179 LD block variants at the sites of ENCODE TFBS

Coordinate	rsID	ENC_TFBS
85859598	-	MafK_(ab50322),eGFP-GATA2,p300_(F-4),GATA-1,Ini1,Pol2,CCNT2,TAL1_(SC-12984),GATA-2,Brg1
85862491	rs187016120	CEBPB,TAL1_(SC-12984),IRF1,Brg1,STAT2,HDAC2_(SC-6296),Pol2,GATA-2,eGFP-JunD,p300_(F-4),GATA-1,eGFP-GATA2,Mxi1_(bHLH),SMC3_(ab9263)
85862739	-	TAL1_(SC-12984),Brg1,Pol2,eGFP-JunD,GATA-1,eGFP-GATA2
85863080	rs56157503	EBF1_(C-8)
85863473	rs34731047	p300
85863683	rs3889743	TFIIIC-110
85863769	rs11234562	TFIIIC-110

Full list of the ENCODE TFBS falling at the positions of variants of interest from the NGS project data in the rs3851179 LD block.