# Complementary Approaches to Analyse Genetic Data in Late Onset Alzheimer's Disease (LOAD)

**Hui Shi, BSc., MSc**

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

**February 2012**

# Table of Contents

# List of Figures

# List of Tables

## Abstract

Alzheimer's disease is the most common form (~60-80%) of dementia, currently affecting approximately half a million people in the UK and ~30 million people worldwide. The autosomal dominant form of AD represents a small proportion (~1-2%) of AD cases and is genetically well characterised. The vast majority of AD cases that show symptoms later in life (> 65 years of age) are genetically complex. This type of AD, also known as late onset Alzheimer's disease (LOAD) disease, is still highly heritable with an estimated heritability of up to 76% (Gatz et al., 2006).

Unfortunately, there is no cure for this devastating disease. Investigating genetic factors influencing the risk of LOAD is imperative for development of effective therapeutic treatments and more accurate diagnosis.

A cross-platform comparison of four Genome-wide association studies (GWAS) was performed in an effort to identify novel genetic associations with LOAD (**Chapter 3**). A *TRIM15* SNP rs929156 demonstrated significant evidence of association with LOAD with a p-value approaching genome-wide significance ($p = 8.77 \times 10^{-8}$) and an odds ratio that showed consistent effect on risk (OR = 1.1, $p = 0.03$). Within this chapter, a bio-informatic program to automate the process of GWAS meta-analysis taking into account linkage disequilibrium (LD) is also presented. Subsequently two fragments of the *TRIM15* gene (including both 5' and 3' end flanking regions) were sequenced using the ABI SOLiD$^{TM}$ next generation sequencing technology. This was a pilot study using a pooled DNA strategy to determine whether this region harbours multiple rare variants which are associated with the disease (**Chapter 4**).

Lastly, a candidate gene study combined with whole genome analysis was performed in an effort to search for genetic variants influencing human ageing using LOAD GWAS data (**Chapter 5**).

## Declaration

I, Hui Shi, hereby declare that the work presented in this project was performed by myself. Exceptions to this have been clearly stated in the text.

## Acknowledgements

I am grateful to my supervisors Professor Kevin Morgan and Professor Noor Kalsheker, for their support and supervision during the course of the project, and during the writing up of this thesis. I would like to thank both of them for giving their insightful comments on papers that have been published as well as my research.

I want also to acknowledge the additional supervisory support from Dr. Sally Chappell and Dr. Linda Morgan on various aspects of the project.

I would like to express my sincere gratitude to Dr. Kristelle Brown and Dr. Tamar Guetta-Baranes, who have taught me how to perform various laboratory experiments. I wish to give my special thanks to Dr. Kristelle Brown for proof-reading of my thesis and a number of publications.

I would like to thank all my colleagues and friends in the lab - Christopher Medway, James Bullock, James Turton and Jenny Lord. Especially I would like to thank James Turton for giving valuable suggestions to the writing of this thesis.

I want to express my thanks to all funding bodies for their financial support of the project - Alzheimer's Research UK and Big Lottery Fund.

I owe my deepest gratitude to my family, especially mum and dad, for their unwavering and unconditional support.

# Abbreviations

| | |
|---|---|
| α2M | α2-macroglobulin |
| AAD | Age at death |
| AAO | Age at onset |
| Aβ | Beta amyloid |
| ABCA1 | ATP-binding cassette transporter protein 1 |
| ABCA7 | ATP-binding cassette transporter protein 7 |
| AD | Alzheimer’s disease |
| ADRDA | Alzheimer’s Disease and Related Disorders Association |
| AICD | APP intracellular domain |
| ANOVA | Analysis of Variance |
| APH1 | Anterior pharynx defective 1 |
| APOE | Apolipoprotein E |
| APP | Amyloid precursor protein |
| APPCTF | Membrane-bound APP C-terminal fragment |
| ARSB | Arylsulfatase B |
| ARUK | Alzheimer's research UK |
| BACE1 | β-site APP cleaving enzyme 1 |
| BBB | Blood brain barrier |
| BIN1 | Bridging integrator 1 |
| BLAST | Basic Local Alignment Search Tool from NCBI |
| CBD | Corticobasal degeneration |
| CD2AP | CD2 associated protein |
| CD33 | Sialic acid binding immunoglobulin-like lectin 3 |
| CDK5 | Cyclin-dependent kinase 5 |
| CERAD | Consortium to Establish a Registry for Alzheimer's Disease |
| CI | Confidence interval |
| CJD | Creutzfeldt-Jakob disease |
| CLU | Clusterin |
| CNDP1 | Carnosine dipeptidase 1 |
| CNS | Central nervous system |
| CNTN1 | Contactin-1 |
| CNTN2 | Contactin-2 |
| CNVs | Copy number variations |
| CR1 | Complement receptor 1 |
| CT | Computed tomography |
| ddNTP | Dideoxynucleotide triphosphate |
| DHA | Docosahexaenoic acid |
| DLB | Dementia with Lewy-body |
| dNTP | Deoxyribonucleotide triphosphate |
| DSM-IV-TR | Diagnostic and Statistical Manual of Mental Disorders 4th |

| | |
|---|---|
| | Edition Text Revision |
| ECR | Evolutionary conserved regions |
| EOAD | Early onset Alzheimer's disease |
| EPA | Eicosapentaenoic acid |
| EPHA1 | Ephrin receptor A1 |
| EtBr | Ethidium bromide |
| FAD | Familial Alzheimer’s disease |
| FTD | Frontotemporal dementia |
| FTDP-17 | Frontotemporal dementia and parkinsonism linked to chromosome 17 |
| GRC | Genome Reference Consortium |
| GSK3 | Glycogen synthase kinase 3 |
| GSTT1 | Glutathione S-transferase theta 1 |
| GUI | Graphical user interface |
| GWAS | Genome Wide Association Study |
| HLA | Human leukocyte antigen |
| HWE | Hardy-Weinberg Equilibrium |
| IDE | Insulin degrading enzyme |
| IGF1 | Insulin-like growth factor 1 |
| IGF1R | Insulin-like growth factor 1 receptor |
| IL10 | Interleukin 10 |
| IL6 | Interleukin 6 |
| LD | Linkage disequilibrium |
| LOAD | Late onset Alzheimer’s disease |
| LR-PCR | Long range polymerase chain reaction |
| MAF | Minor allele frequency |
| MDS | Multidimensional scaling |
| MRI | Magnetic resonance imaging |
| MS4A6A | Membrane spanning 4 domains, subfamily A, member 6A |
| NCAM-120 | 120kDa isoform precursor of neural cell adhesion molecule 1 |
| NCSTN | Nicastrin |
| NEP | Neprilysin |
| NFTs | Neurofibrillary tangles |
| NGS | Next generation sequencing |
| NIMH | National Institute of Mental Health |
| NINCDS | National Institute of Neurological and Communicative Disorders and Stroke |
| NPR | Neuronal pentraxin receptor |
| NTC | No template control |
| OR | Odds ratio |
| PCs | Principal components |
| PET | Positron emission tomography |
| PHF | Paired helical filaments |

| | |
|---|---|
| PICALM | Phosphatidylinositol-binding assembly protein |
| PKA | Cyclic AMP-dependent protein kinase A |
| PKC | Protein kinase C |
| PON1 | Paraoxonase 1 |
| PSEN1 | Presenilin 1 |
| PSEN2 | Presenilin 2 |
| PSP | Progressive supranuclear palsy |
| QC | Quality control |
| SAD | Sporadic Alzheimer's disease |
| SIRT3 | Sirtuin 3 |
| SMON | Subacute myelo-optic neuropathy |
| SNPs | Single nucleotide polymorphisms |
| SOLiD | Sequencing by Oligonucleotide Ligation and Detection |
| SPARCL1 | Secreted protein acidic and rich in cysteine like protein 1 |
| Th1 | T helper cell |
| TFCP2L1 | Transcription factor CP2-like 1 |
| TLR | Toll-like receptors |
| TRIM15 | Tripartite motif containing 15 |
| UTR | Untranslated region |
| VCF | Variant call file |

## List of Publications

1. **Shi H**, Medway C, Brown K, Kalsheker N and Morgan K. Using Fisher's method with PLINK 'LD clumped' output to compare SNP effects across Genome-wide Association Study (GWAS) datasets. Int J Mol Epidemiol Genet 2011;2(1):30-35.

2. Belbin O, Brown K, **Shi H**, Medway C, Abraham R, Passmore P, Mann D, Smith AD, Holmes C, McGuinness B, Craig D, Warden D, Heun R, Kolsch H, Love S, Kalsheker N, Williams J, Owen MJ, Carrasquillo M, Younkin S, Morgan K and Kehoe PG. A Multi-Center Study of ACE and the Risk of Late-Onset Alzheimer's Disease. J Alzheimers Dis 2011;24(3):587-597.

3. Turton JC, Bullock J, Medway C, **Shi H**, Brown K, Belbin O, Kalsheker N, Carrasquillo M, Dickson D, Graff-Radford N, Petersen RC, Younkin SG and Morgan K. Investigating Statistical Epistasis in Complex Disorders. J Alzheimers Dis 2011; 25: 635-644.

4. **Shi H**, Medway C, Bullock J, Brown K, Kalsheker N and Morgan K. Analysis of Genome-Wide Association Study (GWAS) data looking for replicating signals in Alzheimer's disease (AD) Int J Mol Epidemiol Genet 2010;1(1):53-66.

5. Medway C, **Shi H**, Bullock J, Black H, Brown K, Vafadar-isfahani B, Matharoo-ball B, Ball G, Rees R, Kalsheker N and Morgan K. Using In silico LD clumping and meta-analysis of genome-wide datasets as a complementary tool to investigate and validate new candidate biomarkers in Alzheimer's disease. Int J Mol Epidemiol Gene 2010;1(2):134-144.

6. **Shi H**, Medway C, Brown K, Kalsheker N, Goate A, Proitsi P, Powell J, Lovestone S, Carrasquillio M, Younkin S and Morgan K. Genetic Variants

Influencing Human Longevity from Late-Onset Alzheimer's Disease (LOAD) Genome-Wide Association Studies (GWAS). Alzheimer's and Dementia 2011;7(4):S195

7. Medway C, **Shi H**, Brown K, Wilson R, Blythe M, Aboobaker A, Kalsheker N and Morgan K. Novel Rare Variants in Alzheimer's Disease Candidate Genes Identified Using Next Generation Sequencing. Alzheimer's and Dementia 2011;7(4):S190

8. Morgan K, Medway C, **Shi H**, Turton J, Bullock J, Brown K, Belbin O, Kalsheker N, Carrasquillo M and Younkin S. Differential Epistatic Interactions with Alternative APOE alleles in Late-Onset Alzheimer's Disease. Alzheimer's and Dementia 2011;7(4):S191

# Chapter 1: Introduction

Alzheimer's disease is the most common cause of dementia in the elderly and accounts for more than two-thirds of all dementia cases. Dementia affects ~820,000 people in the UK, and costs the UK economy approximately £23 billion per year (Alzheimer's Research UK). There are ~35 million people worldwide who suffer from Alzheimer's disease, and this figure has been estimated to rise to 65.7 million in 2030 and 115.4 million in 2050 (Ferri et al., 2009). The prevalence of AD ranges from 0.6% in persons aged 65 to 69 years to 22.2% at ages 90 and older (Lobo et al., 2000).

## 1.1 Outline of the project

The following studies have been explored in this thesis (with the general aspects covered by **Chapters 1, 2 and 6**):

- Analysis of Genome Wide Association Study (GWAS) data looking for replicating signals in LOAD – **Chapter 3**
- Next generation sequencing (NGS) of tripartite motif containing 15 (*TRIM15*) gene using pooled DNA samples – **Chapter 4**
- Genetic variants influencing human ageing from LOAD Genome Wide Association Studies (GWAS) – **Chapter 5**

Respective specific aims are described in each of the chapters as appropriate.

## 1.2 Symptoms, diagnosis and treatment of Alzheimer's disease

Symptoms

AD is clinically identified by a progressive loss of cognitive abilities. The symptoms at early stages involve mild memory loss - finding difficulty in remembering recently learned facts such as people, places and meetings. As the disease progresses, various advanced symptoms can occur such as confusion, irritability and aggression, mood swings, language breakdown, long-term memory loss, and sensory decline. Ultimately, the disease causes loss of body functions, and finally death.

AD is pathologically characterized by extracellular deposits of abnormally accumulated β-amyloid (Aβ) peptide in the form of senile plaques in cerebral cortex, and intracellular neurofibrillary tangles (NFTs) of abnormally hyperphosphorylated tau (τ) proteins. Both observations are likely to be caused by misfolding and gradual conversion of highly soluble proteins into insoluble filamentous polymers (Forman et al., 2004). Furthermore, through brain scanning, such as computed tomography (CT) or magnetic resonance imaging (MRI), AD brains demonstrate severe cortical shrinkage, enlarged ventricles and shrinkage of the hippocampus, a region of the brain thought to be responsible for storing and retaining memories.

Diagnosis

Alzheimer's disease (AD) can only be definitely (100%) diagnosed post mortem when an autopsy of the brain is performed (Carrette et al., 2003). However, it has been demonstrated that using a combination of tools, it is possible to estimate and make a probable diagnosis of Alzheimer's disease in a living patient, and the accuracy can range from ~80% to 95% (Ballard et al., 2011; Mucke, 2009).

A number of AD diagnosis criteria have been established to date. The most widely used methods are known as NINCDS_ADRAD, DSM-IV and CERAD. These criteria

involve a number of diagnostic procedures, such as taking history from patients and their families, assessment of their cognitive function by carrying out neuropsychological tests (e.g. mini-mental state examination MMSE (Folstein et al., 1975)), and distinguish Alzheimer's disease from other neurodegenerative dementias. Other forms of dementia include frontotemporal dementia (FTD), dementia with Lewy-body (DLB) and Creutzfeldt-Jakob disease (CJD).

According to these diagnostic criteria, AD patients are assigned into three different risk groups - definite, probable and possible. Definite AD is defined only if histopathological evidence is available (**Table 1.1**) (Dubois et al., 2007).

The NINCDS-ADRDA criteria were established in 1984 by the NINCDS (National Institute of Neurological and Communicative Disorders and Stroke) and ADRDA (Alzheimer's Disease and Related Disorders Association) (McKhann et al., 1984). A similar AD diagnosis criteria DSM-IV TR was published by the American Psychiatric Association in 2000. These criteria are under constant review and take into account technology advances in functional neuroimaging techniques such as PET (positron emission tomography) and SPECT (single photon emission computed tomography) scans. The latest amendment to the NINCDS-ADRDA criteria was carried out in 2007 (Dubois et al., 2007).

**Table 1.1 Summary of NINCDS-ADRDA criteria for diagnosis of AD.** Table lists criteria for the different AD risk groups. Adapted from Yaari and Corey-Bloom, 2007.

| | |
|---|---|
| Possible | • Atypical onset, presentation, or clinical course of dementia<br>• Presence of another illness capable of producing dementia |
| Probable AD | • Deficits in two or more domains of cognition<br>• Progressive decline of memory and other cognitive functions<br>• Preserved consciousness<br>• Onset between ages 40 and 90<br>• Absence of systemic or other brain disease that could account for symptoms |
| Definite AD | • Clinical criteria for probable AD<br>• Tissue diagnosis by autopsy or biopsy |

The severity of AD can be assessed using the MMSE score (Folstein et al., 1975);

- mild AD: MMSE score 21 to 26
- moderate AD: MMSE score 10 to 20
- moderate severe AD: MMSE score 10 to 14
- severe AD: MMSE score less than 10

Aβ42, total tau and hyperphosphorylated tau are well established AD biomarkers. Abnormally low Aβ42 levels and high tau (either total tau or hyperphosphorylated tau) levels in CSF act as an important indicator of AD pathogenesis (Buchhave et al., 2009; Tapiola et al., 2009).

Over the last few years, a number of more distinctive biomarkers of AD have become available from studying cerebrospinal fluid (CSF). The level of biomarkers such as secreted protein acidic and rich in cysteine-like protein 1 (SPARCL1), contactin-1 (CNTN1), contactin-2 (CNTN2), alpha-dystroglcan, neuronal pentraxin receptor (NPR), carnosine dipeptidase 1 (CNDP1) and a 120kDa isoform of the precursor of neural cell adhesion molecule 1 (NCAM-120) have been found to be significantly different in AD CSF compared with normal subjects (Yin et al., 2009).

Treatment

Currently, there is no cure for Alzheimer's disease. The damage to the brain is thought to have occurred as many as 10 to 20 years before any symptoms arise. Therefore, pre-symptomatic diagnosis is considered crucial for early and effective treatment of Alzheimer's disease to halt disease progression and reduce symptoms.

Though the disease is generally believed to be irreversible, it is hoped that interventions preventing neuronal cell death, could activate the self-repair mechanism of the brain, leading to restoration of broken neural circuits, and a functional recovery may become possible (Mucke, 2009). Current therapeutic drugs

are effective in relieving the disease symptoms. However, the efficacy of these drugs in slowing down the disease progression and recovery of the brain is limited.

Five drugs have been approved by EMEA (European Agency for the Evaluation of Medical Products) and FDA (USA Food and Drug Administration) for the treatment of AD. These drugs are Donepezil, ENA-713, Galantamine, Memantine and Tacrine. These medicines can be divided in two major categories according to their targets - cholinesterase inhibitors (Donepezil, ENA-713, Galantamine, Tacrine) and an antagonist for NMDA-type (N-methyl-D-aspartate) glutamate receptor (Memantine).

Acetylcholine is an important neurotransmitter, which has been found to be depleted in AD brains. Antagonising its degrading enzyme acetylcholinesterase increases the acetylcholine level in the brain. Thus it improves neurotransmission and ultimately cognitive function. Cholinesterase inhibitors are prescribed to AD patients with mild to moderate symptoms (Winblad et al., 2001). However, these inhibitors may cause adverse effects such as diarrhea, vomiting, nausea, fatigue, insomnia and anorexia.

Memantine is an uncompetitive antagonist of NMDA-type receptor for glutamate, a main excitatory neurotransmitter in the human central nervous system (CNS). Glutamate is known to play an important role in neural transmission, learning, memory processes and neuronal plasticity (Sucher et al., 1996). The level of glutamate in the brain has important implications in determining synaptic cell survival, where it has been found that excess levels of glutamate are toxic to neurons. This increase in the level of glutamate was found to be caused by over-stimulation of NMDA receptors (Robinson and Keating, 2006). Antagonising NMDA receptor thus formed the biological basis for this drug. A previous clinical trial has suggested a small beneficial effect of Memantine during six month placebo controlled trials in moderate to severe AD. However, it is not yet clear if it has any effect in AD patients with less severe symptoms (Areosa et al., 2005). Memantine has been found well

tolerated in clinical trials, with dizziness and mild headaches reported as the main adverse effects. Interestingly, combined treatment using both Memantine and Donepezil appear to improve cognitive performance over either therapy alone on multiple clinical measures, suggesting a synergistic effect between the two drugs (Ihalainen et al., 2011; Tariot et al., 2004).

Aβ vaccination

Current available drugs for treatment of AD are severely limited in that they are designed to relieve AD symptoms. In order to halt the disease progression, interference of pathogenic events leading to the clinical symptoms is essential.

Since deposition of amyloid plaques is a major clinical feature of AD, removal of these plaques has been thought to be able to block the disease progression. A placebo-controlled clinical trial of Aβ42 immunisation showed that although patients exhibited clear reduction of Aβ plaques from the brain, there was no evidence of either slowing down of disease progression or improving survival (Holmes et al., 2008). Furthermore, full length Aβ vaccination has been shown to elicit strong side effects which can cause over-activation of the innate immune system, which in turn accelerates disease progression rather than slowing it down (Holmes et al., 2008). AD patients treated with AN1792 (an active Aβ vaccine) exhibited a significant increase ($p = 0.02$) in the risk of aseptic meningoencephalitis compared with placebo controlled AD patients (Orgogozo et al., 2003). As a result, Aβ vaccination avoiding these pro-inflammatory responses (mainly anti-Aβ Th1 immune response) is under intensive development. Moreover, Aβ vaccination combined with immuno-suppressive therapy has also been suggested (Cribbs, 2010).

However, previous studies indicate a strong correlation between soluble oligomeric Aβ concentration and AD, whereas a poor correlation has been observed as

compared with Aβ plaques counts (Davis et al., 1999; Lue et al., 1999; Neuropathology Group of the Medical Research Council Cognitive Function and Ageing Study, 2001). This has led to speculations that the plaque form of Aβ is perhaps harmless or even protective. Active removal of Aβ plaques could however elevate the concentration of soluble Aβ oligomers, resulting in acceleration of disease progression (Holmes et al., 2008).

Alternative medications

A number of over-the-counter medications (such as melatonin and Omega-3 fatty acids) are also available to AD patients.

Melatonin is a natural hormone secreted by the pineal gland. It regulates sleeping cycles and has shown putative beneficial effects to people with sleeping disturbance (Brzezinski et al., 2005). Cardinali et al., 2002 suggested it might also be helpful in suppressing agitation and anxiety. However, such beneficial effects have not been consistently observed in other studies. For example, Gehrman et al., 2009 did not find any significant effect of melatonin (including sleep, circadian rhythms or agitation) on AD patients (when compared with randomized AD subjects who take placebos). There is also evidence that melatonin may disrupt, rather than improve sleep if inappropriately used (Arendt et al., 2008).

Omega-3 fatty acid (which mainly consists of eicosapentaenoic acid [EPA] and docosahexaenoic acid [DHA]) is one of the most widely used alternative therapies for treating AD. The popularity of Omega-3 fish oil is probably due to its well recognized effects of protection of heart diseases with no obvious adverse effects. DHA, which is a major constituent of fish oils, is a long-chain polyunsaturated fatty acid (comprising ~12-16% of the total fatty acids) in the brain (Quinn et al., 2010). The level of DHA has been found to be decreased (~30-50%) in AD patients compared with age-matched controls.

A longitudinal study taking 815 individuals (who were unaffected by AD) aged between 65 to 94 years and followed by 3.9 years showed that subjects who consumed fish regularly (at least once a week) exhibited a reduced risk (~60%) of developing Alzheimer's disease (relative risk, 0.4; 95% CI 0.2-0.9) (Morris et al., 2003). This protective effect has led to extensive follow-up studies which aim to test whether Omega-3 fatty acids can also slow down the disease progression. Several studies have shown that the intake of Omega-3 fatty acids only appear to have a small beneficial effect on patients with very mild AD symptoms, and no effects for patients with moderate to severe symptoms (Freund-Levi et al., 2006; Quinn et al., 2010). Interestingly, a recent study has found that the level of plasma DHA is not only proportional to intake of DHA by eating fish, seafood or DHA supplement, but is also associated with *APOE* ε4 genotype. It has been suggested that the absorption of DHA may be impaired in *APOE* ε4 carriers, and therefore these people do not benefit from consumption of DHA (Cunnane et al., 2009).

Other alternative treatments for AD include aromatherapy, music therapy, drinking wine (in moderation) and green tea, as well as taking Vitamin E as dietary supplements. However, these alternative treatments generally lack biological and scientific basis, and their effectiveness are questionable. It has been shown that an intake of high-dosage vitamin E increases the risk of mortality and thus should be avoided (Miller et al., 2005).

It is hoped, with the additional knowledge gained through genetic studies of AD, that more effective therapeutic interventions could be developed in the future by targeting the root cause of AD rather than only its symptoms (Mucke, 2009).

## 1.3 Early onset Alzheimer's disease

Early onset Alzheimer's disease (EOAD) refers to AD cases that develop symptoms early in life (before 65 years of age). This form of AD cases constitutes only a small proportion (~1% to 2%) of all AD patients, and is genetically well characterised (Campion et al., 1999). EOAD exhibits a Mendelian form of inheritance in an autosomal dominant manner.

EOAD is largely caused by fully penetrant mutations in three genes – amyloid precursor protein (*APP*), chromosome 21q21.3; presenilin 1 (*PSEN1*), chromosome 14q24.2 and presenilin 2 (*PSEN2*), chromosome 1q42.1. Mutations in the *PSEN1* gene have been found to account for the majority of familial AD cases with ~170 mutations being identified, compared with only ~30 and ~10 mutations in *APP* and *PSEN2*, respectively (Shepherd et al., 2009).

All of these mutations have been shown to affect APP proteolytic processing, resulting in generation of toxic Aβ peptides (the major component of senile plaques) in the brain. The alterations in APP processing in favour of Aβ production and its accumulation in the brain are key pathogenic events in EOAD (Marzolo and Bu, 2009).

Proteolytic processing of APP

APP is proteolytically processed through one of the two mutually exclusive pathways (amyloidogenic pathway and non-amyloidogenic pathway) via three enzymes - α, β and γ secretase (Haass and Selkoe, 2007).

In the amyloidogenic pathway (**Figure 1.1**), APP is first cleaved by β-secretase (also known as β-site APP cleaving enzyme 1 - BACE1), releasing a soluble APPβ fragment and a membrane-bound APP C-terminal fragment (APPCTFβ). The C-terminal fragment is subsequently cleaved by γ-secretase within the membrane, releasing Aβ peptides and the APP intracellular domain (AICD). This intracellular domain acts as a transcription factor which regulates gene expression (Konietzko et al., 2008). AICD has been shown to induce transcriptional activation of neprilysin (NEP), which in turn plays an important role in degradation of Aβ (Pardossi-Piquard et al., 2005). AICD has a short half-life and is rapidly degraded in the cytosol (Cupers et al., 2001) by insulin-degrading enzyme (IDE) (Edbauer et al., 2002). In addition, APP-binding protein Fe65 stabilizes AICD and stimulates translocation of AICD to the nucleus and binding of histone acetyltransferase TIP60 (Goodger et al., 2009).

A number of cleavage sites of intramembrane proteolysis by γ-secretase have been identified (**Figure 1.1**), each result in production of different sizes of Aβ peptides ranging from 37 to 43 amino acids (e.g. Aβ38, Aβ40 and Aβ42) (Marzolo and Bu, 2009). The precise site of γ-secretase cleavage has important implications for Aβ aggregation, which in turn can affect the downstream disease pathology. The therapeutic modification of the γ-secretase cleavage site to γ38 has been shown to significantly reduce Aβ aggregation propensity (Haass and Selkoe, 2007).

The γ-secretase complex is composed of four proteins including PSEN1 or PSEN2, nicastrin (NCSTN), anterior pharynx defective 1 (APH-1) and presenilin enhancer protein 2 (PEN2) (Haass, 2004). A previous study has shown that a fully active γ-secretase can be reconstituted in yeast when all four components are expressed (Edbauer et al., 2003).

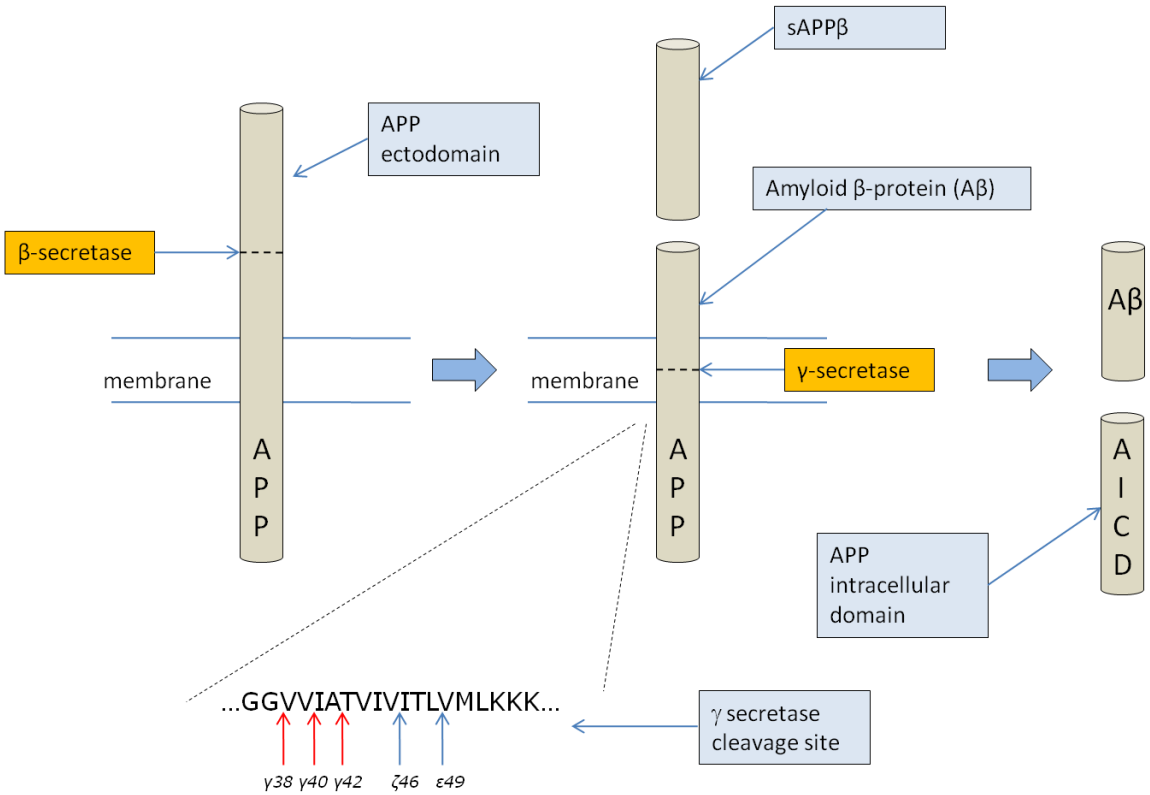



**Figure 1.1 Schematic representation of the amyloidogenic pathway.** APP (grey cylinder) is first cleaved by β-secretase, releasing the soluble β-cleaved APP fragment (sAPPβ). The C-terminal fragment (99 amino acids in length) is subsequently cleaved within the transmembrane domain by γ-secretase, which liberates the Aβ peptide and an APP intracellular domain (AICD). γ-secretase can cleave the APP transmembrane domain at multiple sites – γ, ζ and ε (Adapted from Haass and Selkoe, 2007).

PSEN1 or PSEN2 protein constitute the catalytic core of the γ-secretase complex. Mutations in *PSEN1* and *PSEN2* are thought to influence γ-secretase cleavage events by shifting it two amino acids to the C-terminus, and thus increasing the production of Aβ42 (Haass, 2004). Aβ40 and Aβ42 are the most common isoforms of Aβ peptides (Deane et al., 2009). The longer form (Aβ42) is more fibrillogenic and neurotoxic, and has been shown to be more difficult to clear from the brain compared with Aβ40 (Shepherd et al., 2009).

In the non-amyloidogenic pathway (**Figure 1.2**), APP is first cleaved by α-secretase within the Aβ domain, and thus precludes Aβ production. The cleavage by α-secretase generates a soluble APPα peptide and membrane-bound C-terminal APP fragment (APPCTFα). Subsequent intramembrane cleavage of the APPCTFα by the γ-secretase complex produces a shortened fragment P3 and a cytoplasmic APP intracellular domain (AICD). It is unclear if P3 peptides play any functional role in pathogenesis of AD.

It is noticeable that AICD is produced in both amyloidogenic and non-amyloidogenic pathways. Interestingly, functional active AICD is likely to be generated predominantly through amyloidogenic pathways, where translocation of AICD to the nucleus has been found to be significantly reduced when the endosomal β-cleavage pathway was blocked by pharmacological or genetic inhibitors (Goodger et al., 2009).

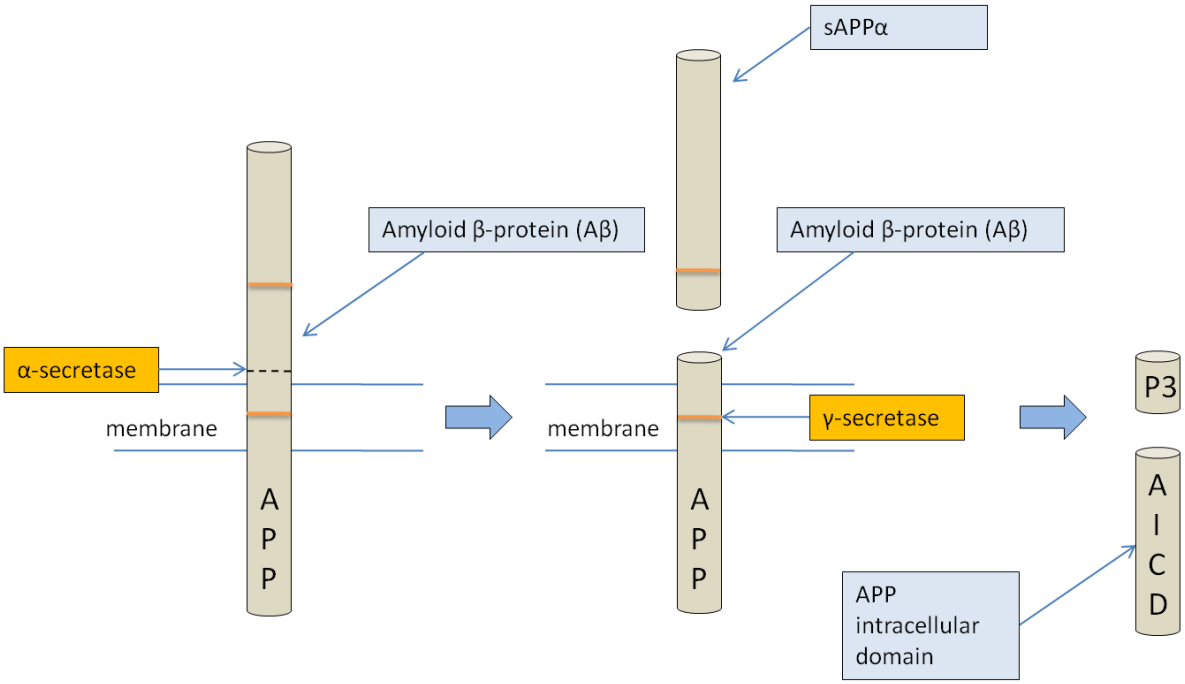



**Figure 1.2 Schematic representation of the non-amyloidogenic pathway.** Schematic structure of APP (grey cylinder) is shown together with Aβ (as shown). In the non-amyloidogenic pathway, APP is first cleaved by α-secretase within the Aβ domain, thus precluding production of Aβ. This cleavage by α-secretase results in release of a soluble APP fragment (sAPPα), and a shortened form of the membrane bound C-terminal fragment (83 amino acids in length). Subsequent cleavage of the C-terminal fragment by γ-secretase within the membrane releases a P3 peptide and a cytoplasmic APP intracellular domain (Adapted from Haass and Selkoe, 2007).

## 1.4 Late onset Alzheimer's disease

Late onset Alzheimer's disease (LOAD), also known as sporadic AD (SAD) represents the majority (~98-99%) of AD cases. LOAD exhibits a complex aetiology with strong genetic and environmental determinants. Like many other complex common diseases, sporadic AD is likely to be governed by an array of common risk alleles across a number of different genes (Bertram et al., 2008).

Mutations in *APP*, *PSEN1* and *PSEN2* responsible for causing EOAD have not been reliably detected in LOAD patients, suggesting a distinct pathogenesis of LOAD exists in comparison to EOAD. Although LOAD does not show Mendelian inheritance, it is still highly heritable with an estimated heritability of up to 76% as determined by studies of monozygotic and dizygotic twins (Gatz et al., 2006).

Although EOAD and LOAD share common neuropathological phenotypes including both extracellular senile plaques and intracellular neurofibrillary tangles (NFTs), the accumulation of Aβ in LOAD is believed to be a result of Aβ clearance deficits or increased Aβ aggregation rather than being causative as in EOAD pathology (Shepherd et al., 2009; Sleegers et al., 2010).

Age is the one of the biggest non-genetic risk factors for LOAD, where the likelihood of developing AD approximately doubles every 5 years after the age of 65 (Feulner et al., 2009). It should be noted that as much as 24% of LOAD risk could be attributable to non-genetic factors, such as diet and lifestyle (Gatz et al., 2006).

## 1.5 The Amyloid Cascade

As the two main features of AD are deposits of senile plaques containing Aβ peptides and intracellular deposits of neurofibrillary tangles containing hyperphosphorylated tau protein, and there is genetic evidence from studies of EOAD, pathways concerning Aβ and tau have been a major focus of AD research.

All identified mutations associated with AD (both EOAD and LOAD) have been found either directly or indirectly linked with the formation, aggregation and removal processes of Aβ. These findings eventually lead to the formation of the amyloid cascade hypothesis. Aβ are small peptides (~4 kDa) and the most common isoforms are Aβ40 and Aβ42 (Deane et al., 2009).

The Amyloid Cascade Hypothesis proposes that progressive cerebral accumulation of beta-amyloid (Aβ) is the central trigger of the pathological changes found in the brain of AD patients. These changes include synapse loss, activation of inflammatory processes, induction of neurofibrillary changes and ultimately neuronal death (Hardy and Higgins, 1992; Selkoe, 1991).

Aβ conformation and toxicity

It has been suggested that different conformations of Aβ could induce neurotoxicity in distinct biological pathways. Aβ40 and Aβ42 exist in different aggregation states from monomers to dodecamers, where oligomer refers to any aggregation state with the exception of a monomer. Insoluble fibrils formed as oligomers grow in size, and accumulation of these ultimately forms Aβ plaques found in the brain of AD patients. In addition, Aβ42 peptide has been found to be more fibrillogenic than Aβ40 (Tanzi and Bertram, 2005).

It has been shown that accumulation of Aβ can cause neurotoxic effects, resulting in the release of reactive oxygen species, loss of calcium homeostasis and activation of

the several kinases including GSK3 - a kinase responsible for phosphorylation of tau protein (Lee et al., 2005).

There is emerging evidence that small soluble Aβ oligomers are more toxic than mature fibrils. This has been supported by numerous studies in biochemistry and histopathology, where biochemically measured levels of soluble Aβ (monomers and oligomers) correlate much better with the extent of synaptic loss and severity of cognitive dysfunction in AD than do simple plaque counts (Holmes et al., 2008; Lue et al., 1999).

Aβ homeostasis

The level of soluble Aβ is homeostatically controlled by its production in neurons and its subsequent clearance. Such homeostasis is thought to be deficient in the brain of an AD patient. Levels of neurotoxic Aβ in the brain have been found elevated in AD contributing to the disease progression and neuropathology (Deane et al., 2008). Furthermore, an increased Aβ42 to Aβ40 radio is a robust indicator of AD (Kuperstein et al., 2010).

Given that the accumulation of Aβ in the brain is determined by the rate of generation versus clearance, both pathways are considered targets for therapeutic interventions. The clearance of Aβ from the brain can be achieved through two biological pathways – proteolytic degradation or receptor mediated transport (Tanzi et al., 2004).

Aβ clearance by receptor mediated transport

The clearance of Aβ from the brain through the blood brain barrier (BBB) is facilitated by lipoprotein receptor-related protein 1 (LRP1) and p-glycoprotein on brain capillaries by binding to chaperones (such as apolipoprotein E (APOE)) and α2-macroglobulin (α2M) (Cirrito et al., 2005). LRP1 antagonists have shown to reduce the efflux of Aβ from brain by up to 90% in mice injected with radiolabeled Aβ40

(Shibata et al., 2000). However, it is still unclear if LRP1 also mediates Aβ influx from the bloodstream. Like Aβ and APOE protein, LRP1 and its ligands are also detected in amyloid plaques in AD brains (Marzolo and Bu, 2009).

It has also been found that LRP1 favours clearance of Aβ40 over the Aβ42 (more amyloidogenic species of the peptide) and this might impede Aβ42 transportation out of the brain through the BBB (Deane et al., 2004). Thus the predominant path of Aβ42 clearance from the brain is thought to be via the proteolytic degradation, as this peptide is not efficiently exported.

Aβ clearance by proteolytic degradation

In the brain, soluble Aβ is degraded by activated microglia. The activation of microglia is likely to be promoted by toll-like receptors (TLR). Frank et al., 2009 found that mRNA which encodes a membrane surface TLR is significantly up-regulated in plaque-associated brain tissues in aged APP23 transgenic mice.

Aβ peptides are proteolytically degraded within the brain principally by neprilysin (NEP) intracellularly and insulin degrading enzyme (IDE) extracellularly (Jiang et al., 2008). Genetic inactivation of these genes or administration of inhibitors of these proteinases in the brain (of non-transgenic mice) leads to substantial elevation of Aβ levels in the brain and induction of plaque deposition (Dolev and Michaelson, 2004). Conversely, overexpression of IDE or neprilysin results in lowered brain Aβ levels and reduced plaque formation (Hemming et al., 2007).

The APOE protein plays a critical role in efficient intracellular degradation of soluble Aβ by microglia. The APOE activity has been shown (using transgenic mice) to be influenced by ATP-binding cassette 1 (ABCA1), which lipidates APOE. Loss of function of ABCA1 impairs Aβ degradation in microglia. ABCA1 null microglia

demonstrates a significantly higher level of intracellular Aβ compared with wild type (microglia in presence of ABCA1) (Jiang et al., 2008).

IDE is secreted by both microglia and astrocytes. It plays an important role in extracellular degradation of soluble Aβ with minor contributions by other secreted proteinases (Qiu and Folstein, 2006; Qiu et al., 1998). In an experiment carried out by Jiang et al. 2008, soluble Aβ was found to be efficiently degraded after addition to an astrocyte-conditioned medium. Addition of insulin (a competitive inhibitor of IDE) prevents this degradation. Interestingly, Aβ clearance by IDE is also influenced by APOE lipidation by ABCA1, where conditioned medium from *ABCA1* deficient astrocytes exhibited significantly higher levels of Aβ compared with medium from wild type astrocytes. In addition, extracellular Aβ clearance has been found more efficient in the presence of both APOE and IDE (Jiang et al., 2008).

Aβ and metals

Aβ aggregation has been found to be facilitated by interaction with metal ions, such as zinc, copper, and other heavy metals. Aberrant metal homeostasis has been observed in AD patients, and it is thought that these ions contribute to AD pathogenesis through enhancing the formation of reactive oxygen species and toxic Aβ oligomers. These metals have been shown to be able to facilitate and stabilize Aβ deposits (Maynard et al., 2005).

Intervention of such interaction using metal-complexing drugs (e.g. clioquinol) has been on clinical trial for treatment of AD (ongoing). A pilot phase II clinical trial using a small number of subjects (n = 36) suggests that clioquinol improves cognition and lowers plasma levels of Aβ42 (Ritchie et al., 2003). However, clioquinol appears to be neurotoxic which induces subacute myelo-optic neuropathy (SMON) syndrome (a syndrome that involves sensory and motor disturbance in the lower limbs and visual changes) (Bareggi and Cornelli, 2011).

## 1.6 Pathways concerning tau protein

Tau protein, also known as microtubule-associated protein tau/saitohin (MAPT/STH), is considered a central mediator of Alzheimer's disease pathogenesis, since one of the clinically observed characteristic of AD is the formation of intracellular neurofibrillary tangles (NFT) mainly composed of abnormally hyperphosphorylated tau proteins.

Mutations in the *MAPT* (which encodes tau protein), although found to result in tau hyperphosphorylation, do not specifically lead to AD symptoms. These mutations have been shown to be a major cause of a different type of dementia - frontotemporal dementia (FTD) (Haberland, 2010). FTD is histologically distinct from AD as the brain is normally free of Aβ plaques (Small and Duff, 2008).

Tau is a phosphoprotein which normally contains one to three moles of phosphate per mole of tau protein in the healthy brain, and the level of which is dramatically increased (three to four folds higher) in the brain tissues of AD patient of similar age (Zhang et al., 2009).

The main recognized function of tau is to promote assembly and stabilization of microtubules in the brain. The binding capacity of tau is highly regulated by protein phosphorylation. The hyperphosphorylated tau has lower binding capacity to microtubules compared with unphosphorylated tau protein, resulting in destabilization of microtubules in the brain (Sato-Harada et al., 1996). Tau protein phosphorylation is achieved by protein kinases: glycogen synthase kinase 3 (GSK3), cyclin-dependent kinase 5 (CDK5), possibly cyclic AMP-dependent protein kinase A (PKA) and protein kinase C (PKC) (Churcher, 2006).

Similar to Aβ, tau can polymerize and form paired helical filaments (PHFs), and accumulation of these leads to formation of neurofibrillary tangles (NFTs). NFTs are

not exclusive to AD however and have been found in many other neurodegenerative diseases: Down's syndrome, progressive supranuclear palsy (PSP), corticobasal degeneration (CBD), frontotemporal dementia and Parkinsonism linked to chromosome 17 (FTDP-17), Pick's disease, and Niemann-Pick type C disease (Avila et al., 2004).

## 1.7 Discussion over current belief on Alzheimer's disease central pathogenesis

In EOAD, it is clear that alteration of APP processing in favour of Aβ production (particularly Aβ42) is sufficient to cause the disease. Given EOAD and LOAD share common pathological features (extracellular Aβ plaques and intracellular hyperphosphorylated tau tangles), biological pathways concerning Aβ and tau are undoubtedly critical pathological events in AD.

There is an increasing volume of evidence that suggests that Aβ and tau may not be the root cause of LOAD pathology (Hardy and Selkoe, 2002) as partial and complete removal of Aβ plaques by immunisation does not show significant effects on cognitive function. Clear end stage dementia has been observed in individuals with almost the complete elimination of plaques (Holmes et al., 2008).

Regarding tau proteins, although it has been found that the number and total length of microtubules were significantly reduced in pyramidal neurons from AD in comparison to controls ($p = 4 \times 10^{-6}$), no significant correlation between the loss of microtubules and PHFs has been observed ($p = 0.8$). Individuals without PHF have often been found with clear microtubule deficits (Cash et al., 2003). In addition, it has been suggested that tau hyperphosphorylation may be neuroprotective in the early stages of disease process, which possibly enables neurons to self-repair. Cells overexpressing hyperphosphorylated tau protein has been shown to be more resistant to apoptosis (Zhang et al., 2009). There is however, a general consensus that a prolonged existence of NFTs is toxic and harmful to neurons.

These studies provide evidence that other biological pathways exist that are crucial to the pathogenesis of LOAD. The recently identified LOAD genes (Harold et al., 2009; Hollingworth et al., 2011; Lambert et al., 2009; Naj et al., 2011; Seshadri et al.,

2010) promise a new era of AD research, and are likely to reveal biological pathways underlying the root cause of the disease.

## 1.8 Genetic risk factors in LOAD

Apolipoprotein E

*APOE* encodes for a 299 amino acid glycoprotein (~34 kDa) in humans. This gene is expressed in several organs, with the highest expression found in the liver and the brain (Bu, 2009).

The APOE protein exists in three isoforms, E2, E3, E4 translated from three specific alleles, ε2, ε3 and ε4, respectively (Mahley, 1988). The three allelic forms of *APOE* are determined by two missense single nucleotide polymorphisms (SNPs) – rs429358 (T/C) and rs7412 (C/T) (**Table 1.2**). These two SNPs cause coding change at amino acid positions 112 and 158, respectively (C112R and R158C).

*APOE* is a major risk factor for LOAD explaining ~25% of the population attributable risk (Lambert et al., 2009). The association of *APOE* ε4 with LOAD was first reported in 1993 through linkage analysis using family pedigrees (Corder et al., 1993; Saunders et al., 1993; Sleegers et al., 2010). This association has been confirmed by numerous genetic association studies (http://www.alzgene.org/).

The presence of *APOE* ε4 greatly increases the risk of AD and reduces the average age at onset (Feulner et al., 2009). Individuals carrying a single copy of ε4 have a ~4-fold higher risk of developing AD in comparison to carrier of the ε3 allele; ~12-16-fold increased risk of AD if with two copies of ε4. *APOE* ε2 is known to engender a reduced risk of AD (Bertram et al., 2010).

**Table 1.2 Summary of *APOE* ε2, ε3 and ε4 allelic status.** The *APOE* ε2, ε3 and ε4 status are determined by two SNPs: rs429358 and rs7412. Individual heterozygous at both SNP loci is an indication of *APOE* ε2/ε4 status (Kim et al., 2009).

| Allelic status | rs429358 | rs7412 | Description |
|---|---|---|---|
| **ε2** | T | T | ε2 allele has a frequency of ~8% in the general population, and is known to elicit a protective effect against AD |
| **ε3** | T | C | ε3 is the most common allele of *APOE* with a frequency of ~77% in the general population |
| **ε4** | C | C | the ε4 allele has a frequency of ~15% in the general population, whereas ~40% in patients with AD. |

Neither of these SNPs that generate the APOE alleles has been genotyped in HapMap project. As existing GWAS chips are designed using SNPs derived from the International HapMap Project, both *APOE* SNPs are therefore not genotyped by current LOAD GWAS (The International HapMap Consortium, 2003). SNP rs2075650, which does present on the chip, is known to be in tight LD with the APOE ε4 allele (Yu et al., 2007). SNP rs2075650 has been reported to be associated with risk of AD with p-value $1.8 \times 10^{-157}$ (Harold et al., 2009).

The principal biological function of APOE is lipid and cholesterol metabolism. In plasma, high density lipoprotein (HDL) contains APOA-1 as its major apolipoprotein, whereas APOE is the most predominant apolipoprotein of HDL in the CNS (Kim et al., 2009).

It has been shown that APOE is involved in both Aβ aggregation and clearance (**Figure 1.3**) (Bu, 2009).

APOE can interact with Aβ either directly or indirectly and promotes Aβ clearance through both receptor mediated transport and proteolytic degradation as described (See **Introduction 1.5**). The E4 isoform of APOE is associated with not only the least efficient transport, but also reduced capability of promoting degradation of soluble Aβ in comparison with E2 and E3 isoforms (Deane et al., 2008; Jiang et al., 2008).

Furthermore, Aβ in the blood stream is transported by cholesterol-rich HDL particles, where APOE is also one of the structural components, prior to elimination by the liver (Koudinov et al., 1998).

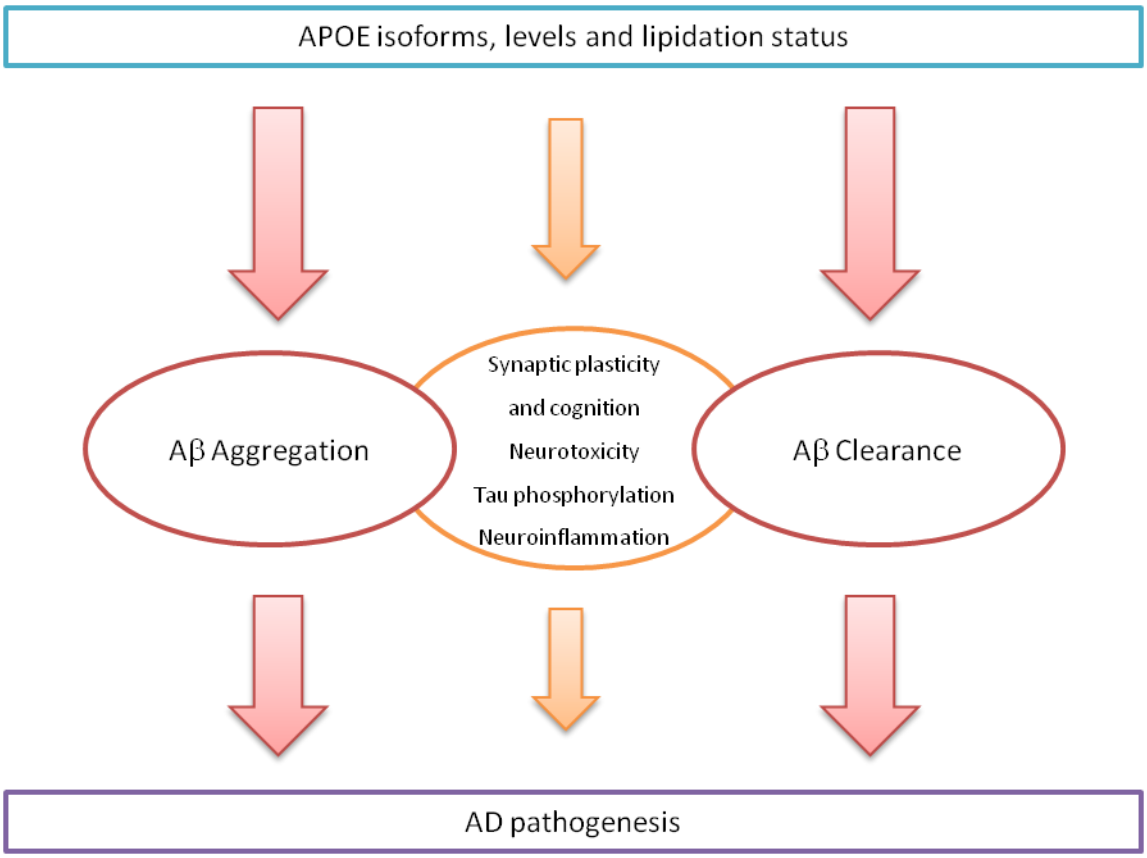



**Figure 1.3 Pathogenic mechanism of APOE in LOAD.** It has been postulated that APOE isoforms influence risk of AD via regulating aggregation and clearance of Aβ. In addition, the different isoforms, levels and lipidataion status of APOE have been proposed as central mediators of LOAD pathology through modulating synaptic functions, Aβ neurotoxicity, tau hyperphosphorylation, and neuroinflammation (Adapted from Kim et al., 2009).

APOE has also been suggested to play a crucial role in Aβ aggregation (Kim et al., 2009). A positive correlation of the *APOE* ε4 allele dosage and increased neuritic plaques in AD has been observed in humans through post-mortem microscopic examination (Tiraboschi et al., 2004). A follow-up imaging study using PET scans also confirmed this association (*APOE* ε4 allele dosage vs. fibrillar Aβ burden) (Reiman et al., 2009). Moreover, a previous study has revealed that *APOE* ε4 allele dosage is associated with a decreased Aβ42/Aβ40 ratio in CSF ($p = 0.0001$), a robust indicator of Aβ levels in the brain (Kauwe et al., 2009).

APOE may also influence AD pathology through pathways not directly linked to Aβ. As a major apolipoprotein in the brain, APOE is known to play a pivotal rule in cholesterol homeostasis by serving as a ligand in receptor-mediated endocytosis of cholesterol-containing lipoprotein particles (Sleegers et al., 2010). Abnormal cholesterol metabolism has been implicated as a key event leading to the pathogenesis of AD (Martins et al., 2006).

LOAD susceptibility genes apart from *APOE*

In 2009 and 2011, a total of nine genes, *CLU, PICALM, CR1, BIN1, ABCA7, MS4A6A, CD33, CD2AP* and *EPHA1*, have been unequivocally identified and confirmed by several large GWAS (each consisting of over 10,000 samples) influencing the risk of LOAD (Harold et al., 2009; Hollingworth et al., 2011; Lambert et al., 2009; Naj et al., 2011; Seshadri et al., 2010).

These genes can be assigned into three biological pathways, with a number of genes involved in multiple pathways. *CLU, CR1, ABCA7, MS4A6A, CD33* and *EPHA1* have putative functions in the immune system; *PICALM, BIN1, CD33* and *CD2AP* are proteins that play a critical role in synaptic cell membrane processes and endocytosis.

*CLU* and *ABCA7* (together with *APOE*) are crucial in lipid and cholesterol metabolism (Hollingworth et al., 2011; Morgan, 2011).

*CLU* (clusterin)

Two independent GWAS (Harold et al., 2009; Lambert et al., 2009) both identified and replicated a SNP rs11136000 located in an intron of the *CLU* gene, giving strong evidence of association with risk of LOAD with genome-wide level of significance ($p = 1.4 \times 10^{-9}$ and $p = 7.5 \times 10^{-9}$, respectively).

*CLU* (also known as *APOJ*) encodes another abundantly expressed apolipoprotein in the human brain. CLU exhibits similar biological functions as APOE, involved in both cholesterol and lipid metabolism and has been shown to promote export of Aβ across the BBB (Guerreiro and Hardy, 2011). CLU was also found in amyloid plaques in the brain (Calero et al., 2000; May et al., 1990). It has been shown that Aβ deposition are cooperatively regulated by APOE and CLU *in vivo*, where *APOE* and *CLU* double gene knockout PDAPP mice exhibit significantly higher Aβ load in comparison to either of the single gene knockout transgenic mice (DeMattos et al., 2004).

CLU is a multifactorial glycoprotein, and one of the described functions is related to inflammation and immunity through regulating activity in the complement pathway (Falgarone and Chiocchia, 2009; Jenne and Tschopp, 1989; Jones and Jomary, 2002).

Furthermore, expression of CLU has been found to be elevated in response to injury and chronic inflammation of the brain, suggesting that CLU may have an important role in preventing possible damage to neurons (Calero et al., 2000).

*PICALM* (phosphatidylinositol-binding clathrin assembly protein)

A SNP rs3851179 located at 5' to the *PICALM* gene was first discovered and reported to be associated with risk of LOAD (OR = 0.85, $p = 1.9 \times 10^{-8}$) by Harold and colleagues (Harold et al., 2009). The effect has been replicated in an independent sample cohort with ~4,000 samples ($p = 0.014$, OR = 0.90). On meta-analysis, the combined datasets showed a strong evidence of association with a p-value that exceeded genome-wide significance ($p = 1.3 \times 10^{-9}$, OR = 0.86).

The association with *PICALM* was further supported by a large GWAS conducted by Lambert and colleagues using samples from a French population; a SNP proxy rs541458 in LD with rs3851179 ($r^2 = 0.622$) has shown a significant evidence of association with LOAD ($p = 0.0028$, OR = 0.88), although it did not reached genome-wide significance.

*PICALM* encodes a protein which plays a critical role in clathrin-mediated endocytosis, a key process involved in regulation of receptors, synaptic transmission and clearance of apoptotic cells (Baig et al., 2010). PICALM may alter the risk of LOAD through regulating synaptic transmission and/or Aβ production by modulating the rate of endocytosis of APP, an essential step preceding APP cleavage by β-secretase (**Figure 1.1**) (Goodger et al., 2009; Tebar et al., 1999).

A recent study using immunolabelling has shown that PICALM is predominately present in endothelial cells, mainly expressed in the endothelium of the blood vessel walls and weakly labelled in neurons. This has led to speculation that PICALM may also be involved in Aβ clearance via the BBB into the blood stream (Baig et al., 2010).

Furthermore, a significant epistatic interaction ($p = 0.0068$; logistic regression using an additive model) was reported between the *APOE* ε4 allele and *PICALM* SNP rs3851179 using 3,055 AD cases and 8,169 age-matched controls. The effect of

PICALM was only observed in samples carrying at least one copy of *APOE* ε4 (Jun et al., 2010); $p = 3.4 \times 10^{-3}$ in presences of *APOE* ε4 and $p = 0.73$ in absence of *APOE* ε4.

*CR1* (complement receptor 1)

SNP rs6656401, present in an intron of CR1, has been shown to be associated with an increased risk of LOAD by the Lambert et al., 2009 GWAS. This association was confirmed by Harold et al., 2009 GWAS, with a p-value of $10^{-6}$ (OR = 1.17). Together these two GWAS comprise over 25,000 samples (Harold et al., 2009; Lambert et al., 2009).

In addition to the association with risk of LOAD, the CR1 SNP rs6656401 was also found to be associated with a faster rate of cognitive decline ($p = 0.011$) and an increased deposition of neuritic amyloid plaques ($p = 0.009$), where the significance was not affected by including the *APOE* ε4 status as a covariate (Chibnik et al., 2011).

*CR1* encodes a major receptor of C3b, a key inflammatory protein involved in AD pathogenesis (Khera and Das, 2009). It has been postulated that CR1 may be involved in the process of Aβ clearance through mediating complement-driven phagocytosis, which may in turn prevent brain damage through reducing Aβ-induced neurotoxicity (Carrasquillo et al., 2010). Using hAPP transgenic mice, it has been shown that mice expressing sCrry (soluble complement receptor-related protein y), an inhibitor of C3 activation, exhibited ~2-3 folds elevated Aβ deposition in the brains compared with mice without such inhibition (Wyss-Coray et al., 2002).

Furthermore, a significant epistatic interaction between the *APOE* ε4 and *CR1* SNP rs6656401 has been reported ($p = 9.6 \times 10^{-3}$), with stronger association observed in carriers of the *APOE* ε4 (Lambert et al., 2009).

*BIN1* (bridging integrator 1)

A three stage meta-analysis consisting of over 35,000 samples (8,371 cases) identified an association of a SNP rs744373 (within ~30kb of the *BIN1* gene) with an increased risk of LOAD ($p = 1.59 \times 10^{-11}$, OR = 1.13) (Seshadri et al., 2010). The effect of this SNP was replicated in an independent Spanish sample cohort (1,140 AD cases and 1,209 controls) with a p-value 0.02 and an odds ratio in the same direction.

BIN1 (also known as amphiphysin-2) is expressed most abundantly in the CNS and muscles, and appears to be involved in the endocytosis of synaptic vesicles (Cousin and Robinson, 2001). A study using transgenic mice found that the amphiphysin 1 knockout mice, which cause reduction of amphiphysin 2 selectively in the brain, exhibited major learning deficits and increased rate of mortality (Di Paolo et al., 2002).

*ABCA7* (ATP-binding cassette transporter protein)

The *ABCA7* SNP rs3764650 was found to be significantly associated with an increased risk of LOAD in a combined sample cohort, consisting of over 60,000 samples (25,900 LOAD cases and 41,584 controls) ($p = 5 \times 10^{-21}$, OR = 1.23) (Hollingworth et al., 2011).

*ABCA7* encodes an ATP-binding cassette (ABC) transporter and is abundantly expressed in the brain (Kim et al., 2006). ABCA7 is involved in the transfer of lipids and cholesterol to lipoprotein particles such as APOE and CLU.

Although no evidence of epistatic interactions between these loci were observed, it does not preclude possibility of biological interactions (Hollingworth et al., 2011).

*MS4A6A* (membrane-spanning 4 domains, subfamily A, member 6A)

A SNP (rs610932) in proximity to gene *MS4A6A* exhibited a significant evidence of association with a reduced risk of LOAD ($p = 1.2 \times 10^{-16}$, OR = 0.91) (Hollingworth et al., 2011). The marker was found within an LD block ~290kb in size. This region comprises six genes of *MS4A* gene family, which includes *MS4A2, MS4A3, MS4A4A*, *MS4A4E, MS4A6A* and *MS4A6E*.

These genes encode for proteins which share structural similarities – all members comprise a transmembrane domain, suggesting that these proteins may be involved in synaptic cell membrane processes (Liang et al., 2001). The exact biological function of MS4A6A has yet to be characterised.

*CD2AP* (CD2 associated protein), *CD33* (sialic acid binding immunoglobin-like lecting) and *EPHA1* (ephrin receptor A1)

Three other genes *CD2AP, CD33* and *EPHA1* have also been implicated in LOAD pathogenesis (Hollingworth et al., 2011). Markers of these genes (rs9349407, rs3865444 and rs11767557) demonstrated strong evidence of association with LOAD which reached genome-wide level of significance ($p < 1 \times 10^{-8}$), albeit at a lower statistical significance than genes mentioned earlier (*CLU, PICALM, CR1, BIN1, ABCA7, MS4A6A*).

Both *CD33* and *CD2AP* encode for proteins important in communication between cells and transduction of molecules across the membrane (Crocker et al., 2007; Lynch et al., 2003). *EPHA1* encodes for an ephrin receptor which has been previously reported to play a role in synaptic development and plasticity (Lai and Ip, 2009).

Summary of genes and pathways in Alzheimer's disease

Identification of these genetic risk factors provides better understanding of underlying biological pathways and mechanisms of LOAD. Future drugs that target pathways highlighted by these genes enable potential development of effective treatments and more accurate diagnosis of AD.

**Figure 1.4** summarises the genes and pathways in Alzheimer's disease implicated from recent large GWAS (Harold et al., 2009; Hollingworth et al., 2011; Lambert et al., 2009; Naj et al., 2011; Seshadri et al., 2010). These pathways include immune system function, lipid and cholesterol metabolism and synaptic cell membrane processes and endocytosis.

Apart from pathways elucidated from recent large GWAS, several other biological pathways have also been implicated in AD, including oxidative stress (Lovell and Markesbery, 2007), mitochondria function (Swerdlow, 2011), the insulin signalling pathway (Liolitsa et al., 2002; Stewart and Liolitsa, 1999) and metal homeostasis (Maynard et al., 2005).

One of the main advantages of GWAS is that the genes selected are not dependent on pre-conceived knowledge about their function, and therefore may be able to highlight a more general picture of AD genetics (Lambert and Amouyel, 2011).

In view of the genes identified by current GWAS results and the amyloid cascade hypothesis proposed by Hardy and Selkoe, 2002, a number of common mechanisms can be implicated (**Figure 1.5**):

- Familial early-onset forms of AD caused by rapid accumulation of Aβ peptides due to overproduction, which is associated with mutations in *APP*, *PSEN1* and *PSEN2*

- Late onset AD caused by subtle and insidious impaired clearance of the Aβ peptides, associated with *APOE*, *CLU*, *CR1*, *PICALM*, *BIN1*, *ABCA7* and potentially as yet unidentified others.

If the hypothesis is true, there is likely an overlap of the two gene categories at intermediate age at onset (Lambert and Amouyel, 2011). This hypothesis requires further investigation.

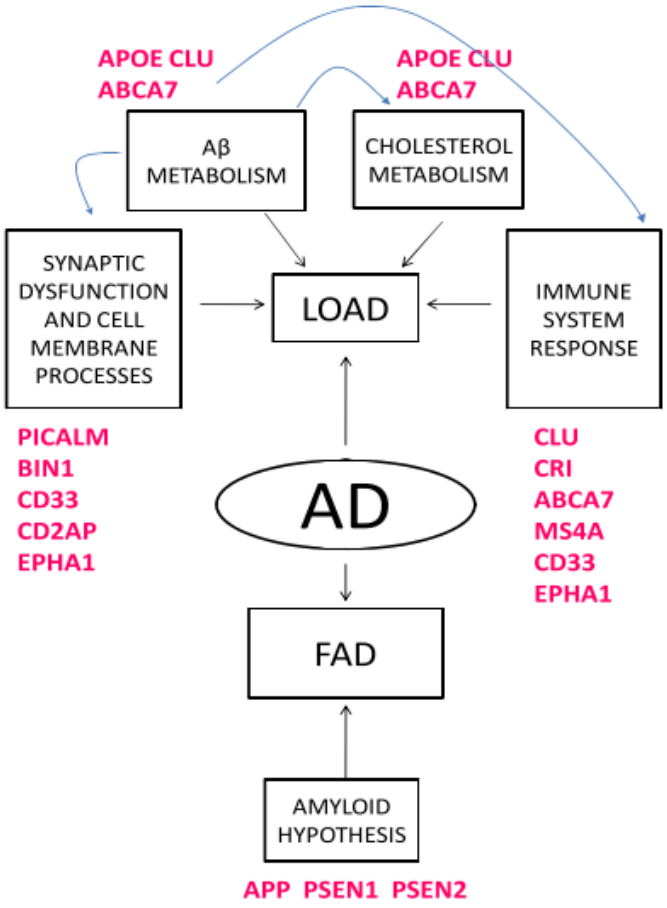



**Figure 1.4 Genetic risk factor and pathways in Alzheimer's disease.** Genes involved in each pathway are shown in red. There is a significant overlap with a number of genes being involved in multiple pathways. Aβ metabolism and homeostasis may have a direct effect on pathways implicated in LOAD (as indicated by blue arrows). The familial form of AD, which occurs before age 65, is predominantly caused by mutations in three genes – *APP*, *PSEN1* and *PSEN2*. (Adapted from Morgan, 2011)

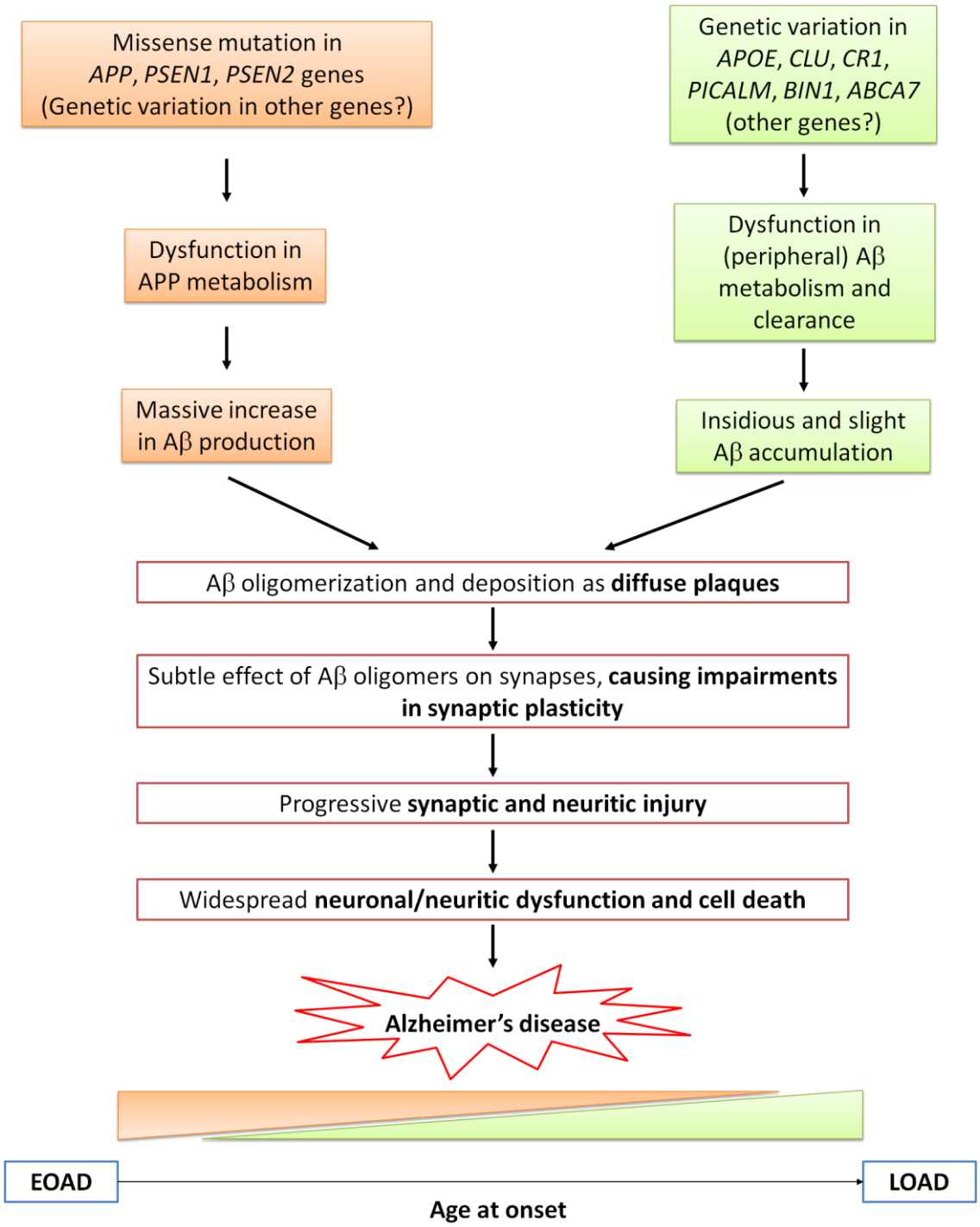



**Figure 1.5 Summary of pathogenic events leading to EOAD and LOAD proposed by Hardy and Selkoe, 2002 considering new AD genes identified by recent large GWAS (Adapted from Lambert and Amouyel, 2011).**

## 1.9 The missing heritability of LOAD

Genome-wide association studies have been successful in identifying hundreds of replicable common genetic variants associated with a variety of complex diseases (Hindorff et al., 2009). These genetic risk factors found by GWAS greatly improve our understanding of the genetic basis of many complex disorders including LOAD.

The Population-attributable fraction (PAF) is defined as the proportion of disease cases in a population that would be prevented if an exposure were eliminated (Bertram et al., 2007). Population-attributable fraction (PAF) is also known as population-attributable fraction of risk or population-attributable risk (Ertekin-Taner, 2010; Lambert and Amouyel, 2011; Lambert et al., 2009).

The Population Attributable fraction (PAF) can be calculated using the formula shown below (Bertram et al., 2007):

$$\mathbf{PAF} = \frac{\mathbf{F} \times (\mathbf{OR} - \mathbf{1})}{\mathbf{F} \times (\mathbf{OR} - \mathbf{1}) + \mathbf{1}}$$

F is the frequency of the risk allele in the general population and OR is the odds ratio of the risk allele (Yang et al., 2003).

The newly found LOAD genes (*CLU, PICALM, CR1, BIN1, ABCA7, MS4A6A, CD33, CD2AP* and *EPHA1*) were estimated to have a PAF range from 2.72% to 5.97% (Naj et al., 2011). The estimation of PAF must be interpreted with caution, as it is based on a number of assumptions, and may vary substantially between studies (Bertram et al., 2007).

It has become increasingly evident that despite expanded GWAS that are capable of capturing most common variants with both moderate and small effects, a substantial fraction of the heritability of LOAD remains unaccounted for (‘missing heritability’).

The remaining unexplained heritability is believed to be due to a combination of factors:

- additional common variants of smaller effect size (which current GWAS are still underpowered for detection),
- additional common variants missed by GWAS due to incomplete coverage,
- genetic risk caused by low frequency and rare variants, which are not detectable by GWAS,
- synthetic associations attributable to rare variants and
- epistasis (known as gene-gene interaction).

Other factors such as copy number variations (CNVs) and gene-environmental interactions could also play a role in AD pathogenesis and contribute to the missing heritability (Morgan, 2011).

Additional common variants

It is plausible that additional common variants are associated with risk of LOAD, but have not yet been discovered due to insufficient power. The sample size of a GWAS determines the effect size of a common variant that can be detected. Identification of associations with common variants of smaller effect sizes may continue to provide insights into the complex biological pathways involved in AD. However, identification of these variants is unlikely to have any immediate consequences in terms of disease prediction and diagnosis (Seshadri et al., 2010).

Rare variants

Another explanation of missing heritability is due to low frequency and rare variants, which occur with a frequency < 5%. The existing commercial genotyping chips for GWAS are not designed for capturing SNPs with a MAF less than 5%. There is increasing evidence that these low frequency rare variants can make a significant

contribution to the heritability of complex traits and diseases (Rivas et al., 2011). These less frequent variants are often found to have larger effect sizes than common variants (Bodmer and Bonilla, 2008).

Synthetic associations

It has been suggested that a proportion of GWAS signals could be attributable to casual rare variants of larger effect size due to incomplete LD with these rare variants (Dickson et al., 2010).

Detection of a GWAS signal may therefore underestimate the actual effect size of the rare variants (Wang et al., 2010), although the actual number of common variants attributable to these variants is still unclear. It has also been argued that synthetic associations attributable to rare variants do not explain most of GWAS results (Wray et al., 2011).

Genetic architecture of complex traits

**Figure 1.6** illustrates the predicted genetic architecture of complex disorders. According to their allele frequency and effect size, SNPs can be separated into five different categories:

- common variants of large effect size,
- common variants of small effect size,
- low frequency variants with intermediate effect,
- rare variants causing Mendelian diseases and
- rare variants of small effect size.

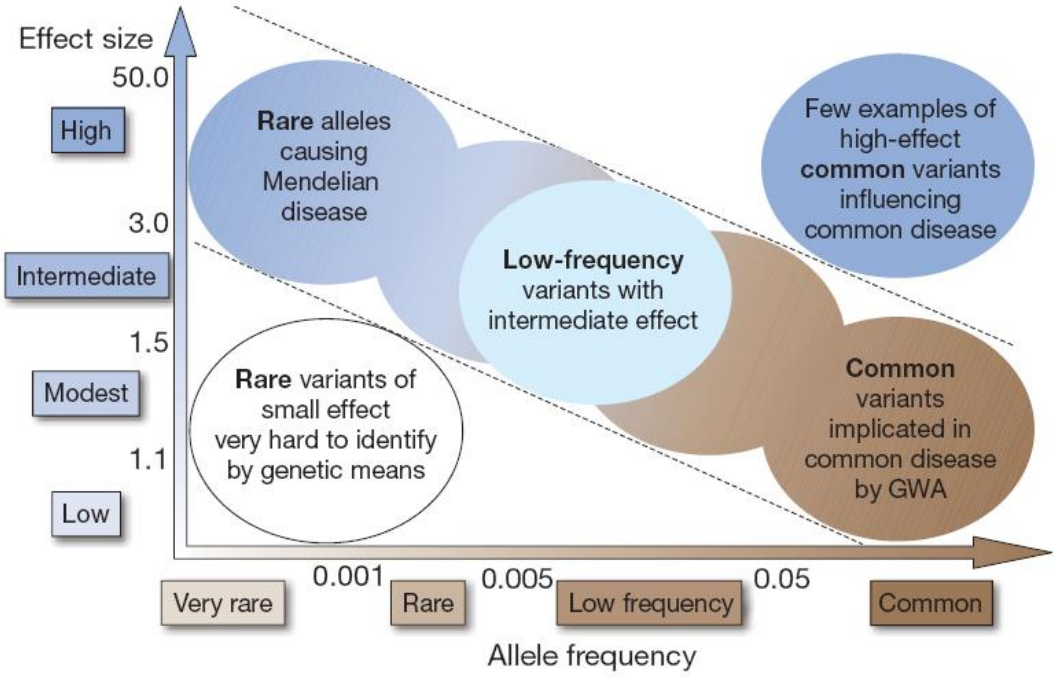



**Figure 1.6 Genetic architecture of complex traits.** Figure summarizing the role of genetic variants in complex disorders by allele frequencies and their genetic risk. Allele frequency is shown on the x-axis versus odds ratios on the y-axis. Common variants (MAF > 0.05) can be detected through genome-wide association studies, whereas low frequency and rare variants may only be ascertained through direct genotyping and sequencing projects. (Adapted from Manolio et al., 2009)

Linkage analysis (using family pedigrees) is a powerful tool to map the location of disease-causing loci in reference to the human genome. It was one of the most widely used technologies preceding the GWAS era. However, linkage analysis is not capable of identifying SNPs with small effect size. *APOE* ε2 and ε4 are examples of common variants with large effect size OR = ~4 (Bertram, 2011). In very few cases however do common genetic variants exist that have a high risk (i.e. OR > 4) associated with complex disorders.

The vast majority of common variants have shown to exert only a small effect size with OR ranges 1.1-1.5. These odds ratio have been consistently observed through studies of many complex disorders (Bodmer and Bonilla, 2008). These genetic loci are difficult to detect using pedigree information. The advent of GWAS enabled systematic detection of the associations between common variants and disease given that a large enough sample size is utilized to provide adequate power of detection (**Figure 1.6**).

There is emerging evidence that less frequent and rare variants may contribute to a significant proportion of the missing heritability of LOAD. Although these variants are only present in a small proportion of the population, the effect sizes (of these variants) are often found to be higher than the association with common variants (**Figure 1.6**).

Rare variants of small effect size may also exist (**Figure 1.6**). This type of variant is difficult to detect by any genetic means. Methods to detect these rare variants require further exploration.

Epistasis

One of the possible explanations of the missing heritability is epistasis. It is a measure of the interaction between two or more genetic loci (synergistic or antagonistic) contributing to the risk of disease. PLINK ('--epistasis' or '--fast-epistasis') (Purcell et al., 2007) and synergy factor analysis (Combarros et al., 2009)

are among the methods being widely used for assessing epistatic interactions between genes. Epistatic interactions between SNPs in the regulatory regions of *IL6* and *IL10* have been reported, and have been shown to be associated with a reduced risk of AD (Infante et al., 2004). This interaction was replicated in a follow-up study (Combarros et al., 2009).

## 1.10 Summary

Alzheimer's disease is the most common form of dementia in the elderly, accounting for approximately two thirds of all dementia cases (Blennow et al., 2006). With the average life expectancy continuing to rise (i.e. the population ageing), the number of AD cases is likely to increase in the near future. The number of individuals suffering from AD is expected to rise to ~115.4 million worldwide by 2050 (Ferri et al., 2009).

Genetic research in a small proportion (~1% to 2%) of AD patients with an autosomal dominant pattern of inheritance has contributed greatly to our understanding of AD pathogenesis by identifying causal mutations in three genes - *APP*, *PSEN1* and *PSEN2* (See **Amyloid Cascade; Introduction 1.5** for details). These AD cases are known as FAD, as they show AD symptoms early in life (before the age of 65) (Bertram and Tanzi, 2008).

LOAD represents the vast majority of AD cases, and their development is likely to be affected by both genetic and environmental factors. This non-Mendelian form of AD is still highly heritable, with an estimated heritability ranging from 60% to 80% (Gatz et al., 2006). Mutations in genes causing the early onset form of Alzheimer's disease, including *APP*, *PSEN1* and *PSEN2*, do not appear to be strongly associated with the risk of late onset form of Alzheimer's disease (Bertram, 2011).

Apart from *APOE* gene, nine additional genes (*CLU, PICALM, CR1, BIN1, ABCA7, MS4A6A, CD33, CD2AP* and *EPHA1*) have been identified by GWAS and confirmed

by several replication studies as influencing the risk for LOAD (Harold et al., 2009; Hollingworth et al., 2011; Lambert et al., 2009; Naj et al., 2011; Seshadri et al., 2010) (**Genetic risk factors in LOAD; Introduction 1.8**). Despite these successes, there is increasing evidence to suggest that a large proportion of the genetic variation contributing to AD risk remains unidentified (Sherva and Farrer, 2011) (See **The missing heritability of LOAD; Introduction 1.9** for details).

Early GWAS conducted during 2007 and 2009 generally failed to produce any convincing results due to the lack of power (Sherva and Farrer, 2011). It is believed that by combining each individually underpowered GWAS power could increase thus allowing identification of genuine associations and previous spurious associations will likely diminish. This formed the basis of the study as described in **Chapter 3 - Analysis of Genome Wide Association Study (GWAS) data looking for replicating signals in LOAD**.

# Chapter 2: Materials and methods

## 2.1 Study samples

Next generation sequencing

150 DNA samples (75 AD cases and controls) were used for next generation sequencing (**Chapter 4**). These samples came from three research centres: Nottingham, Manchester and Leeds (**Appendix 8.1**).

Samples used for studying human ageing

SNP rs4110518 was genotyped in 462 samples (335 AD cases and 127 controls) with age-at-death (AAD) information (**Chapter 5**). Sample IDs of these samples are not shown.

All DNA samples used are part of Alzheimer's Research UK (ARUK) collection. Approval was obtained from the ethics committee or institutional review board of each institution responsible for the ascertainment and collection of samples. Written informed consent was obtained for all individuals that participated in this study.

Except the autopsy sample cohort, which were examined post-mortem, all other AD cases were diagnosed according to NINCDS-ADRDA (National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association), DSM-IV (Diagnostic and Statistical Manual of Mental Disorders IV) or CERAD (Consortium to Establish a Registry for Alzheimer's Disease) criteria (McKhann et al., 1984; Mirra et al., 1991).

## 2.2 Laboratory methods

A number of laboratory methods were utilized to generate biological data for analysis in **Chapters 4 and 5**. Specifically, DNA extraction and quantitation, LR-PCR, agarose gel electrophoresis, BigDye® sequencing and gel extraction were utilized to produce DNA pools for next generation sequencing (**Chapter 4**). TaqMan® genotyping assays were used to validate and replicate SNP results as described in **Chapter 4** and **Chapter 5**.

### 2.2.1 DNA extraction from post-mortem brain tissues

DNA was extracted from brain tissues (~50mg) using QIAGEN® DNeasy Blood & Tissue Kit according to the manufacturer's protocol. Proteins and RNA were degraded by addition of 20µl protease K (20mg/ml; QIAGEN®) and 4µl RNase A (provided in the kit). DNeasy® mini spin columns (provided) were used to purify DNA by selective binding of DNA to the membrane as contaminants pass through. The DNA was eluted using 150µl elution buffer (supplied in the kit) and stored at −20°C prior to use in PCR.

### 2.2.2 DNA quantitation using NanoDrop®

DNA was quantified using NanoDrop® spectrophotometer using a standard laboratory protocol.

The inability of distinguishing UV absorption of free nucleic acid from double stranded DNA meant that this technology is not sufficiently accurate to be used for experiments that require a precise DNA concentration, e.g. next generation sequencing library preparation, real-time PCR and DNA cloning. Alternative methods such as Qubit®, Quanti-iT PicoGreen® should be used instead.

### 2.2.3 Long-range polymerase chain reaction (LR-PCR)

Primers (sense and antisense) for LR-PCR were designed using primer-BLAST program (http://www.ncbi.nlm.nih.gov/tools/primer-blast/). The quality of the primers were measured using Primer3 (v 0.4.0) (http://frodo.wi.mit.edu/primer3/) (Rozen and Skaletsky, 2000) and SNPcheck (v 2.0) (https://ngrl.manchester.ac.uk/SNPCheckV2) using default settings.

The following primers were used for *TRIM15* 'A' and 'B' amplicons:

*TRIM15* 'A'

- Sense: ATGGGTGAAGGACCGTGGCT
- Anti-sense: AGGAAAGTGCCCCAAGGCCA

*TRIM15* 'B'

- Sense: AGGGGAAGGCGCCACAGTTT
- Anti-sense: ACAGGAGAATGGGCCCCACA

PCR amplification

*TRIM15* 'A' and 'B' amplicons were LR-PCR amplified using FINNZYMES Phusion™ High-Fidelity DNA Polymerase on an Applied Biosystem Veriti® 96-well thermal cycler.

The Phusion™ DNA polymerase, according to the manufacturer, has the following advantages:

- Allows amplification of large size DNA amplicons (up to 15kb using genomic DNA)
- High Fidelity – ~50-fold more accurate than *Taq* DNA polymerase
- High speed – 15-30 seconds per 1 kb

Cycling conditions used for *TRIM15* 'A' and 'B' amplicons are shown in **Table 2.1**.

PCR products were stored at –20°C.

**Table 2.1 Amplification of *TRIM15* gene by LR-PCR.** Table showing A) cycling programs for *TRIM15* 'A' and 'B' amplicons and B) reagents used for each of the 50µl reactions. The optimal annealing temperature for each amplicon (indicated by '*') was determined by performing a temperature gradient optimisation.

| **A)** | | |
|---|---|---|
| ***TRIM15* 'A' amplicon (1,984 bp)** | | |
| **Temperature** | **Time** | **Cycles** |
| 98°C | 30 seconds | 1 |
| 98°C | 10 seconds | 35 |
| 70.3°C* | 30 seconds | |
| 72°C | 1 minute | |
| 72°C | 7 minutes | 1 |
| 10°C | | Hold |
| ***TRIM15* 'B' amplicon (4,935 bp)** | | |
| **Temperature** | **Time** | **Cycles** |
| 98°C | 30 seconds | 1 |
| 98°C | 10 seconds | 35 |
| 71.2°C* | 30 seconds | |
| 72°C | 2 minutes and 30 seconds | |
| 72°C | 7 minutes | 1 |
| 10°C | | Hold |

| **B)** | | |
|---|---|---|
| **Reagent** | **Concentration** | **Volume** |
| Phusion HF or GC buffer | 5x | 10µl |
| dNTPs | 10mM | 1µl |
| forward primer | 100µM | 0.25µl |
| reverse primer | 100µM | 0.25µl |
| Phusion® DNA Polymerase | 2U/µl | 0.5µl |
| DNA | 50ng/µl | 1µl |
| Nuclease free $H_2O$ | NA | 37µl |
| **Total:** | | 50µl |

Optimisation of conditions for LR-PCR

For each primer pair, the reaction was optimised using the strategy:

- Annealing temperature was optimised by performing temperature gradients, 57-72°C initially followed by a smaller temperature gradient according to the band intensity (e.g. 62-67°C).
- $Mg^2$+ concentration was optimised by performing a Magnesium gradient – 1.5mM, 2mM, 2.5mM, 3mM and 3.5mM.
- Addition of dimethyl sulfoxide (DMSO) was attempted to test if it improves the performance of the LR-PCR; DMSO may improve the LR-PCR performance by inhibiting formation of any secondary structures in the DNA template and facilitating complete DNA denaturation.

### 2.2.4 Agarose gel electrophoresis

PCR products were visualised by agarose gel electrophoresis. Small 0.7% agarose gels (gel volume 25ml) were prepared as follows:

- 0.17g of agarose powder (Fisher Scientific®)
- 25ml 1xTAE (40mM Tris acetate, 1mM EDTA)
- 3µl of ethidium bromide (EtBr) (10 mg/ml, Pharmacia Biotech)

Medium (50ml 1xTAE buffer, 0.35g agarose and 5µl of EtBr) and large agarose gels (80ml 1xTAE buffer, 0.56g agarose and 8ul EtBr) were also used when necessary.

PCR products mixed with DNA loading buffer (Fermentas®) were subjected to electrophoresis together with the GeneRuler$^{TM}$ 1kb ladder at ~80V. The DNA was visualised using a UV transilluminator.

### 2.2.5 BigDye® sequencing

Sanger sequencing was used to examine if the LR-PCR products were of the correct fragment (by comparing with reference human genome sequence, hg19) prior to DNA pooling for next generation sequencing (**Chapter 4**).

LR-PCR products were purified using ExoSAP-IT® (composed of Exonuclease I and Shrimp Alkaline Phosphatase) and sequenced using the BigDye® Terminator v3.1 sequencing kit according to manufacturer's recommendations. Addition of ExoSAP-IT® facilitates removal of unincorporated dNTPs and residual primers.

BigDye® terminator reaction premix contains essential reagents including AmpliTaq® DNA polymerase, deoxynucletides (dNTPs) and fluorescently labelled dideoxynucleotides (ddNTPs). Cycling programs and reagents used in each of the reactions are shown in **Table 2.2**.

After the sequencing reaction, the mixture was filtered by running through Edge Biosystem Performa® DTR Gel Filtration cartridges to remove excess ddNTPs. The reaction was dried on a thermal block at 90°C. The dried DNA pellets were stored at −20°C prior to capillary electrophoresis using an ABI® 3130 Genetic Analyser. Capillary electrophoresis was performed by the Molecular Diagnostic Lab at the University of Nottingham.

**Table 2.2 Sequencing using BigDye® (v 3.1).** Table showing A) cycling conditions and B) reagents used for the ExoSAP-IT® treatment and BigDye® sequencing reactions.

| **A) ExoSAP-IT® treatment** | | |
|---|---|---|
| **Temperature** | **Time** | **Cycles** |
| 37°C | 15 minutes | 1 |
| 80°C | 15 minutes | 1 |
| 10°C | | Hold |
| **Sequencing reaction using BigDye®** | | |
| **Temperature** | **Time** | **Cycles** |
| 96°C | 1 minute | 1 |
| 96°C | 30 seconds | 25 cycles |
| 50°C | 15 seconds | |
| 60°C | 4 minutes | |
| 10°C | | Hold |

| **B) ExoSAP-IT® treatment** | | |
|---|---|---|
| **Reagent** | **Concentration** | **Volume** |
| LR-PCR product | ~200ng/µl | 5µl |
| ExoSAP-IT premix | 100% | 2µl |
| **Total:** | | 7µl |
| **BigDye® sequencing** | | |
| **Reagent** | **Concentration** | **Volume** |
| ExoSAP-IT purified PCR product | | 5µl |
| Sequencing primer | 5µM | 1µl |
| BigDye® terminator reaction premix | 100% | 3µl |
| ABI® sequencing buffer | 5x | 1µl |
| **Total:** | | 10µl |

### 2.2.6 Gel extraction

LR-PCR product was purified by gel extraction using QIAquick® gel extraction kit according to manufacturer's recommendations.

SYBR® Green (Invitrogen SYBR Safe™ DNA gel stain) was used (instead of EtBr) for agarose gel electrophoresis to avoid introducing damage to the DNA. The agarose gel was visualised using a Dark Reader® transilluminator, where the desired DNA band was excised from the agarose gel using a clean, sharp scalpel blade before being transferred into a 1.5ml Eppendorf tube. The agarose gel was subsequently dissolved and filtered out using the kit. The DNA was collected using 30µl nuclease free water. The elution process was repeated, and the DNA was stored at −20°C.

### 2.2.7 TaqMan® SNP Genotyping assay

TaqMan® genotyping assay was supplied at 40x concentration. The assay was diluted to a 20x working concentration by adding one volume of 1xTE buffer (10mM Tris-HCL, 1mM EDTA at pH8.0).

TaqMan® genotyping was performed using Agilent® Real-Time PCR optical 8-tube strips and optically clear 8-cap strips on the STRATAGENE Mx3000P™ Real-Time PCR System. DNA templates were diluted to 10ng/µl concentration using nuclease-free water prior to genotyping.

Cycling conditions and reagents used for the assay are summarised in **Table 2.3**. The data was analysed using MxPro QPCR (v 4.01) and results were exported into Microsoft Excel format for further analysis. TaqMan® genotyping assays are claimed to be highly accurate, although false positive amplifications are possible due to the high throughput and repetitive nature of the 5' nuclease assay. Special laboratory practices are necessary to avoid false positive results (Kwok and Higuchi, 1989).

**Table 2.3 TaqMan® genotyping assay.** Table showing A) cycling program and B) reagents used for the TaqMan® genotyping assay on STRATAGENE Mx3000P™ Real-Time PCR System.

| **A)** | | |
|---|---|---|
| **TaqMan® genotyping assay** | | |
| **Temperature** | **Time** | **Cycles** |
| 50°C | 2 minutes | 1 |
| 95°C | 10 minutes | 1 |
| 95°C | 15 seconds | 55 |
| 60°C | 1 minute | |
| 10°C | | Hold |

| **B)** | | |
|---|---|---|
| **Reagent** | **Concentration** | **Volume** |
| TaqMan® genotyping assay | 20x | 1µl |
| TaqMan® Universal PCR MasterMix | 2x | 9µl |
| Nuclease-free water | NA | 8µl |
| DNA template | 10ng/µl | 2µl |
| **Total:** | | 20µl |

## 2.3 Bioinformatics tools and data analysis

Open source bioinformatic tools have been utilized to enable a wide range of data analysis presented in this thesis. Where indicated, 'In house' Perl scripts were written and have been implemented to facilitate various data analysis (See **Methods 2.3.2**).

A range of bioinformatics tools are described in this section, facilitating data analysis as described in **Chapters 3, 4 and 5**. Specifically, PLINK software (v1.06) was utilized in **Chapter 3** and **Chapter 5** to analyse GWAS data and produce association results. The Haploview program was used in **Chapter 3** to produce LD plots, and in **Chapter 5** to create the Manhattan plot. VISTA and ECR browser were utilized to analyse conservation and facilitating selection of conserved regions for next generation sequencing (**Chapter 4**). EIGENSTRAT was utilized in **Chapter 5** to analyse population stratification using LOAD GWAS data. QUANTO (v1.2.4) was utilized to perform power calculations for the association studies in **Chapter 3, 4 and 5**.

### 2.3.1 PLINK

PLINK is a powerful whole-genome association and linkage analysis toolset developed by Purcell et al., 2007. It has become one of the most reputable bioinformatic toolsets for GWAS data analysis to date. By September 2011, the corresponding manuscript (Purcell et al., 2007) has been cited by over 2,400 peer-reviewed scientific papers.

One of the advantages of using PLINK is it provides a comprehensive range of tools, including GWAS data manipulation, quality control, association studies (single SNP analysis and haplotype analysis), transmission disequilibrium testing (TDT), GWAS meta-analysis, epistasis, imputation and permutation. The results generated by

PLINK can be used directed by other bioinformatic software (e.g. Haploview) for further in-depth analysis.

PLINK is operated through the command line interface as currently there is no viable GUI available for handling large datasets.

PLINK file format

A summary of PLINK input file format is shown in **Table 2.4**. It should be noticed that data stored in PED and MAP files are interlinked, genotype data stored in the PED file correlates to the corresponding MAP file. Therefore, any manual changes on these files should be avoided, as it could render the data unusable by introducing error.

Missing data in PLINK is by default represented as either -9 or 0, with an exception of missing genotype data which is presented as 0 (zero).

**Table 2.4 Summary of PLINK input file format.** Table showing the PLINK input file format together with a brief description.

| File types | Description |
|---|---|
| **PED and MAP** | • Standard (generic) PLINK input file format<br>• PED file is comprised of six compulsory columns (Family ID, Individual ID, Paternal ID, Maternal ID, Sex and Phenotype) and variable number of genotype columns (column 7 and onwards).<br>• MAP file consists of exactly four columns (Chromosome number, SNP identifier, Genetic distance and Base-pair position)<br>• no headers |
| **BED, FAM and BIM** | • PLINK binary file format<br>• BED files store genotype information in a compressed binary format, which is unreadable using a text editor.<br>• The FAM file consists of exactly six columns as in the PED file. The BIM file includes the first four columns of the MAP file plus two additional columns showing the corresponding SNP genotypes.<br>• no headers |
| **PHENO** | • PHENO file is able to store multiple alternative phenotypes for analysis without modification of the PED file or FAM file.<br>• The PHENO file consists of two compulsory columns (family ID and individual ID) and variable number of phenotype columns.<br>• The PHENO file is included in the analysis by specifying '--pheno' in PLINK.<br>• Requires headers, two compulsory columns headers, FID and IID, representing family ID and individual ID |
| **COVAR** | • Stores covariate information to be included in PLINK analysis.<br>• The covariate (COVAR) file consists of two compulsory columns which are identical to the first two columns of the PHENO file, and variable number of covariate columns.<br>• The COVAR file is included in the analysis by specifying '--covar' in PLINK.<br>• Requires headers, two compulsory columns headers, FID and IID, representing family ID and individual ID |

It is recommended to convert the standard PLINK format (PED and MAP) to the binary format when handling large datasets. Binary files are much smaller in file size, PLINK analysis performs much quicker using the binary input format.

PLINK analysis including alternative phenotype and covariates is specified by '--pheno' and '--covar' followed by the PHENO and COVAR file names, respectively.

One of the utilities of using PHENO file is to perform expression quantitative trait loci (eQTL) analysis, where hundreds of thousands of gene expression data could be analysed all at once.

Adjusting for covariates is crucial in GWAS analysis, as it ensures that the association signal identified is not due to underlying biases such as age, gender, centre of study and other sample heterogeneity.

By default, PLINK represents data using number codings as listed:

- Phenotypes – 1 and 2 represents controls and cases
- Gender – 1 and 2 represents males and females
- Genotypes – 1 and 2 represents minor and major allele (also coded as A, T, C or G)

Case/Control association analysis

PLINK provides a number of methods for case/controls association studies. The most commonly used method '--assoc' performs allelic dosage analysis (Wald test) on query SNPs. For example:

**plink --file mydata --assoc**

The command generates an output file 'plink.assoc' which contains the following fields:

- CHR - chromosome number,

- SNP - SNP ID,
- BP - base-pair position,
- A1- the minor allele based on whole sample,
- F_A - frequency of this allele in cases,
- F_U - frequency of this allele in controls,
- A2 - the major allele,
- CHISQ - chi-squared statistics on 1 degree of freedom,
- P - asymptotic p-value and
- OR - estimated odds ratio.

Optional addition of '--ci 0.95' within the command line calculates 95% confidence interval for ORs (odds ratios).

'--assoc' examines potential association much faster than logistic regression analysis; however it does not allow inclusion of covariates.

Logistic regression analysis

Logistic regression analyses are more sophisticated and allow inclusion of covariates. The logistic regression model is more robust than linear regression as it can handle non-linear effects and it does not make assumptions on distribution of the explanatory variables (e.g. a normal distribution) (Bewick et al., 2005).

Logistic regression analysis still has a number of inbuilt assumptions, and requires much larger sample sizes than a standard linear regression analysis. In addition, the logistic regression analysis in PLINK assumes the phenotype (e.g. disease trait) is binary. Therefore, the analysis is quantitative rather than qualitative. This is considered a limitation of logistic regression analysis, as taking into account the severity of AD and disease related endophenotypes is likely to further increase power enabling identification of genuine disease associated variants (Plomin et al., 2009) that could be missed using the current approach.

The analysis can be implemented with the following command line:

**plink --file mydata --logistic --covar myfile.covar --covar-name age,sex,APOEstatus**

'--file' specifies the data files to be analysed as 'mydata.ped' and 'mydata.map'. '--logistic' indicates the logistic regression analysis is utilized in the analysis. '--covar' and '--covar-name' specifies the covariate terms.

Association analysis using different genetic models

Different genetic models can be specified and analysed using logistic regression as shown:

- Additive inheritance model: '--logistic' on its own
- Dominant inheritance model: '--logistic --dominant'
- Recessive inheritance model: '--logistic --recessive'

An alternative way to include different genetic models in the analysis is to use '--model' command, though this does not allow adjusting for covariates. The following tests are provided in PLINK '--model' command:

- Cochran-Armitage trend test
- Genotypic (2 df) test
- Dominant gene action (1 df) test
- Recessive gene action (1 df) test

GWAS Quality controls (QC)

SNPs and individuals can be filtered out from an analysis by addition of QC filters in the command line. '--geno' and '--mind' exclude SNPs and individuals according to genotyping rate. '--maf' excludes SNPs below a user-defined minor allele frequency. '--hwe' removes SNPs from the analysis according to Hardy-Weinberg Disequilibrium

p-values. Furthermore, in studies including families pedigrees, Mendelian errors can be detected by '--me' command.

Combinations of these methods provide flexibility and control in response to various GWAS data analysis types.

Quantitative Trait analysis

Two main methods are provided in PLINK to perform quantitative trait analysis, '--assoc' and '--linear'. The PLINK program automatically engages a quantitative trait analysis when it encounters integers other than 0, 1, 2 or -9 in the sixth column of the PED file.

The '--assoc' command does not take into account covariates in a quantitative trait analysis, conversely '--linear' does allow covariates ('--covar' and 'covar-name') to be included in the analysis.

The versatility of PLINK also allows for different genetic models to be explored in quantitative trait analysis, as previously described.

PLINK gene report function

The PLINK gene report function can annotate SNPs according to their base pair coordinates relative to genes. Two files are required in this analysis, a PLINK results file (e.g. '.assoc') and a file containing coordinate information of known human genes.

The gene list (glist-hg18), which consists of ~20,000 human genes was downloaded from the PLINK website at http://pngu.mgh.harvard.edu/~purcell/plink/res.shtml. This list consists of four columns - Chromosome, Start base pair position, Stop base pair position and Gene name. This facility can be implemented using a similar command to the below example:

**plink --gene-report results.assoc.linear --gene-list glist-hg18.txt --pfilter 0.05 --gene-list-border 20 --out outputfile**

'--pfilter 0.05' specifies that the association p-value threshold is equal to 0.05, indicating that the output would only present SNPs with p-value less than 0.05. '--gene-list-border 20' indicates the maximum distance between the SNP and the reported gene is equal to 20kb.

A PERL program was written (**Appendix 8.4.4**) as a plug-in to complement the PLINK gene report function. The program tabulates the PLINK output into a format enabling further manipulation and analysis (**Figure 2.1**).

This PERL program consists of two files 'gene_report_plugin.pl' and 'modules.pm'. The program can be executed via the command line interface 'perl gene_report_plugin.pl' or the appropriate file icon. Input file name can be specified by editing 'gene_report_plugin.pl' file using a standard text file editor, such as WordPad in Windows. Results are provided in a tab-delimited text file format, which can be accessed via conventional statistical tools (**Figure 2.2**).

**ARSB** -- chr5:78088792..78338113 ( 249.321kb ) including 20kb border

| DIST | CHR | SNP | BP | A1 | TEST | NMISS | OR | SE | L95 | U95 | STAT | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3.124kb | 5 | rs2173012 | 78111916 | 1 | ADD | 981 | 0.5232 | 0.2234 | 0.3377 | 0.8106 | -2.9 | 0.003734 |
| 16.4kb | 5 | rs3846677 | 78125189 | 1 | ADD | 985 | 0.7735 | 0.1217 | 0.6093 | 0.9818 | -2.111 | 0.03479 |
| 33.34kb | 5 | rs12054837 | 78142127 | 1 | ADD | 985 | 1.609 | 0.2004 | 1.086 | 2.382 | 2.373 | 0.01764 |

**ARSG** -- chr17:63746917..63948595 ( 201.678kb ) including 20kb border

| DIST | CHR | SNP | BP | A1 | TEST | NMISS | OR | SE | L95 | U95 | STAT | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3.352kb | 17 | rs7216182 | 63770269 | 1 | ADD | 972 | 0.8158 | 0.1026 | 0.6672 | 0.9977 | -1.983 | 0.04738 |
| 60.15kb | 17 | rs9890694 | 63827063 | 1 | ADD | 982 | 0.7381 | 0.1529 | 0.547 | 0.996 | -1.986 | 0.04705 |
| 81.57kb | 17 | rs8075800 | 63848486 | 1 | ADD | 971 | 0.6713 | 0.1507 | 0.4997 | 0.902 | -2.645 | 0.008173 |
| 111.4kb | 17 | rs3744301 | 63878334 | 1 | ADD | 981 | 0.6109 | 0.1994 | 0.4133 | 0.9029 | -2.472 | 0.01343 |
| 140.4kb | 17 | rs2160725 | 63907321 | 1 | ADD | 973 | 1.27 | 0.09954 | 1.045 | 1.543 | 2.401 | 0.01636 |

**ARSJ** -- chr4:115020888..115140327 ( 119.439kb ) including 20kb border

| DIST | CHR | SNP | BP | A1 | TEST | NMISS | OR | SE | L95 | U95 | STAT | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 21.3kb | 4 | rs12645879 | 115062191 | 1 | ADD | 980 | 0.8173 | 0.09698 | 0.6758 | 0.9884 | -2.081 | 0.03748 |

**ART3** -- chr4:77131360..77272979 ( 141.619kb ) including 20kb border

| DIST | CHR | SNP | BP | A1 | TEST | NMISS | OR | SE | L95 | U95 | STAT | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -18.36kb | 4 | rs6849878 | 77132997 | 1 | ADD | 983 | 0.799 | 0.1078 | 0.6468 | 0.9869 | -2.083 | 0.03729 |

**ARV1** -- chr1:229161445..229223102 ( 61.657kb ) including 20kb border

| DIST | CHR | SNP | BP | A1 | TEST | NMISS | OR | SE | L95 | U95 | STAT | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -2.028kb | 1 | rs13374343 | 229179417 | 1 | ADD | 976 | 0.6999 | 0.1691 | 0.5025 | 0.9748 | -2.111 | 0.03478 |

**ARVCF** -- chr22:18317418..18404309 ( 86.891kb ) including 20kb border

| DIST | CHR | SNP | BP | A1 | TEST | NMISS | OR | SE | L95 | U95 | STAT | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 21.14kb | 22 | rs2073746 | 18358558 | 1 | ADD | 881 | 0.7304 | 0.1235 | 0.5733 | 0.9304 | -2.544 | 0.01097 |
| 33.37kb | 22 | rs9617857 | 18370789 | 1 | ADD | 984 | 0.7137 | 0.1281 | 0.5552 | 0.9174 | -2.633 | 0.008458 |
| 33.65kb | 22 | rs9618725 | 18371067 | 1 | ADD | 965 | 0.7188 | 0.1302 | 0.5569 | 0.9278 | -2.536 | 0.01122 |

**Figure 2.1 PLINK gene-report function output.** Only SNPs that pass the user input filters are listed (i.e. '--pfilter 0.05' and '--gene-list-border 20'). Gene details are shown together with corresponding SNPs (DIST - distance between the SNP and start of the gene, CHR - chromosome number, SNP - rs identifier, BP - base pair position, A1 - minor allele code, TEST - model of test, NMISS - number of non-missing individuals, OR - odds ratio, SE - standard error, L95 and U95 - lower and upper 95% confidence interval, STAT - association test statistics, P - p-value of the association). The corresponding genes (shown in bold) are automatically sorted in alphabetical order.

| SNP | CHR | BP | P | OR | GENE | LENGTH | DIST |
|---|---|---|---|---|---|---|---|
| rs2173012 | 5 | 78111916 | 0.003734 | 0.5232 | ARSB | 209.321kb | 3.124kb |
| rs3846677 | 5 | 78125189 | 0.03479 | 0.7735 | ARSB | 209.321kb | 16.4kb |
| rs12054837 | 5 | 78142127 | 0.01764 | 1.609 | ARSB | 209.321kb | 33.34kb |
| rs7216182 | 17 | 63770269 | 0.04738 | 0.8158 | ARSG | 161.678kb | 3.352kb |
| rs9890694 | 17 | 63827063 | 0.04705 | 0.7381 | ARSG | 161.678kb | 60.15kb |
| rs8075800 | 17 | 63848486 | 0.008173 | 0.6713 | ARSG | 161.678kb | 81.57kb |
| rs3744301 | 17 | 63878334 | 0.01343 | 0.6109 | ARSG | 161.678kb | 111.4kb |
| rs2160725 | 17 | 63907321 | 0.01636 | 1.27 | ARSG | 161.678kb | 140.4kb |
| rs12645879 | 4 | 115062191 | 0.03748 | 0.8173 | ARSJ | 79.439kb | 21.3kb |
| rs6849878 | 4 | 77132997 | 0.03729 | 0.799 | ART3 | 101.619kb | -18.36kb |
| rs13374343 | 1 | 229179417 | 0.03478 | 0.6999 | ARV1 | 21.657kb | -2.028kb |
| rs2073746 | 22 | 18358558 | 0.01097 | 0.7304 | ARVCF | 46.891kb | 21.14kb |
| rs9617857 | 22 | 18370789 | 0.008458 | 0.7137 | ARVCF | 46.891kb | 33.37kb |
| rs9618725 | 22 | 18371067 | 0.01122 | 0.7188 | ARVCF | 46.891kb | 33.65kb |

**Figure 2.2 Output of the PERL program for PLINK 'gene-report' function.** The results from PLINK output. **Figure 2.1** are converted into a simplified tabulated format. SNP - SNP Identifier, CHR - chromosome number, BP - base pair position are shown together with P - p-value of the association, OR - odds ratios, GENE - gene name, LENGTH - size of the gene and DIST - distance between the SNP and the start position of the corresponding gene (negative values indicate that SNP is located before the start position of the gene).

### 2.3.2 PERL programming language

PERL is a programming language that has been widely utilized for the development of novel bioinformatic applications. Analysis of large-scale genomic data is often challenged by the lack of suitable bioinformatic programs. Writing 'in house' bioinformatic tools not only permits exploiting new ideas, but also reduces time for tasks which would otherwise be laborious.

In this thesis, four Perl programs were developed 'in house' for the following calculations:

- Determination of common SNPs between different genotyping chip platforms (e.g. Illumina HumanHap300 versus Illumina HumanHap610, which is a pre-requisite step for principal component (PC) analysis (as described in **Methods 2.3.9**). The program was documented in **Appendix 8.4.1**.
- Calculation of the number of independent tests in GWAS to enable an accurate multiple testing adjustment for GWAS analysis (as described in **Methods 2.3.2**). **Appendix 8.4.2**.
- A GWAS meta-analysis tool taking into account LD, shown in **Chapter 3**. **Appendix 8.4.3**.
- A plug-in for PLINK gene-report function (as described in **Methods 2.3.1**). **Appendix 8.4.4**.

### 2.3.3 Calculation of the number of independent tests

GWAS has given insights into the aetiology of many complex diseases including LOAD. However, due to the large number of SNPs tested all at once, apparently 'significant' findings may arise simply due to chance. However, the majority of these findings are likely to be false positives. Thus a very stringent significance threshold is necessary to provide confidence in the findings. For instance, in a GWAS using 500,000 independent markers, 25,000 would be expected to show a nominal p-value $< 5 \times 10^{-2}$ by chance alone and five out of this 25,000 could be significant with p-values $< 1 \times 10^{-5}$. The most widely used methods for solving this multiple testing issue is to use Bonferroni correction, where it suggests that if 'n' independent tests are carried out, the significance level for the entire series of tests is equal to the p-value of a single test divided by 'n'. The significance threshold of $p = 5 \times 10^{-8}$ has been widely used to infer a genuine association in GWAS (Bertram et al., 2008).

It is generally believed that Bonferroni correction is overly conservative in GWAS findings (Sherva and Farrer, 2011). A p-value of $5 \times 10^{-8}$ is equivalent to a p-value of 0.05 after a Bonferroni correction of 1,000,000 independent tests, whereas early GWAS only possessed ~500,000 SNPs (Affymetrix 500K chip) or ~610,000 SNPs (Illumina 610 chip). Second, due to the existence of LD between SNPs on these genotyping chips, a large number of SNPs are not independent. Taken together, it implies that a SNP with p-value $> 5 \times 10^{-8}$ may well harbour genuine associations.

Linkage disequilibrium (LD) measures the probability that alleles at two loci are co-inherited, the LD value is affected by genetic recombination (Wray et al., 2011).

A more accurate Bonferroni correction p-value threshold can be generated using the exact number of independent tests, where multiple SNPs are counted as a single independent test if they are in perfect LD (i.e. with $r^2 = 1$). It is conceivable that using imperfect proxies ($r^2 < 1$) is likely to further reduce the number of independent test.

The results, however, must be interpreted with caution, as lowering the LD $r^2$ value will likely introduce errors. The relationship between the LD $r^2$ value and the amount of noise contributing to the calculation of genome wide significance threshold requires further investigation.

In order to calculate the exact number of independent test, SNPs genotyped on GWAS chips were extracted from the HapMap dataset using '--extract' and '--make-bed' command. Within this file, SNPs in perfect LD are ascertained using the following PLINK commands:

**plink --bfile 'file' --r2 --ld-window-kb 1000 --ld-window 99999 --ld-window-r2 1 --out 'file.ld'**

'--r2' is the command for calculating $r^2$ value of LD. '--ld-window-kb 1000' indicates the calculation is undertaken within 1Mb distance of index SNPs saved in the input file. '--ld-window 99999' specifies the maximum number of pair-wise combinations to be calculated for each SNP is 99999. '--ld-window-r2 1' indicates the LD $r^2$ threshold is equal to 1.

The LD $r^2$ values calculated using PLINK are based on haplotype frequencies estimated via the Expectation Maximisation (EM) algorithm.

Given that any two SNPs with their base-pair positions more than 1 Mb apart highly unlikely to be in perfect LD, the calculation is conducted within a window of 1Mb either side of the index SNP.

A PERL script (**Appendix 8.4.2**) was written to calculate the number of LD clusters and the number of SNPs in perfect LD. The number of independent tests was then calculated using the formula:

$$\textbf{Number of independent tests}$$
$$= [\textbf{Number of SNPs on the chip} - \textbf{Number of SNPs in perfect linkage}]$$
$$+ \textbf{Number of LD Clusters}$$

The PLINK output file (.ld) was used for this calculation using this PERL script. The genome-wide significant thresholds were calculated based on the number of actual independent tests. This approach has been used in studies described in **Chapter 3** and **Chapter 5**.

### 2.3.4 Linkage disequilibrium analysis

LD patterns dramatically increase the coverage of SNPs chips used in GWAS. In cross-platform meta-analysis, LD further increases the number comparable SNPs between different studies.

There are two ways of measuring the strength of LD - $r^2$ and D'. $r^2$ is more frequently used in comparison of SNPs with similar allele frequencies, whereas D' is often used when assessing the relationship between common and rare variants (Wang et al., 2010; Wray et al., 2011).

The $r^2$ value is considered more stringent than D'. A LD value of $r^2$ equal to 1 produces a D' value also equal to 1, whereas if LD value D' equal to 1, $r^2$ value can range from close to 0 to 1.

The value of D' is not affected by the difference in allele frequencies between two SNPs. Measures of the linkage between two SNPs can be assigned into four categories - Perfect LD, Complete LD, Moderate LD and no evidence of LD.

CandiSNPer

CandiSNPer is a web based bioinformatic application which allows efficient search of SNP LD patterns based on user-specified parameters (input SNP rs number, LD $r^2$ value, output window sizes and population sizes), and simultaneously annotates tagged SNPs, which are in LD with the index SNP (initial input SNP), based on its functions (Schmitt et al., 2010).

CandiSNPer automatically categorize SNPs into different functional classes, and annotates them in different colours. The program can be accessed through the website http://www2.hu-berlin.de/wikizbnutztier/software /CandiSNPer.

Default functional classes in CandiSNPer are:

Class 1: Stop lost, Stop gained, Frameshift

Class 2: Nonsynonymous coding, Splice site, Essential splice site

Class 3: Synonymous coding, 5' UTR, 3'UTR, Upstream, Downstream

Class 4: Intronic, Pseudogene

Class 5: Intergenic

Class 6: Start SNP: rs number

CandiSNPer directly retrieves the latest version of SNP data (in real time) from the Ensembl database. It calculates both LD $r^2$ and D' values and provides the results in a graphical HTML format. Furthermore, CandiSNPer automatically predicts and highlights the LD block where the index SNP is located.

SNAP (SNP Annotation and Proxy Search)

SNAP is a web based bioinformatics tool for assessment of LD between SNPs (Johnson et al., 2008). The SNAP program is accessible at http://www.broadinstitute.org/mpg/snap/ldsearch.php.

It provides an efficient method to retrieve proxies for SNPs under investigation. The LD values between SNPs are calculated using the HapMap and pilot 1000 genome data. Furthermore, SNAP provides a function to graphically represent a 'regional LD plot' for use in publications.

### 2.3.5 Haploview

Haploview is a bioinformatic program designed to compute linkage disequilibrium statistics and population haplotype patterns using a wide range of genotype data input formats (Barrett et al., 2005). The software is written and operated within the Java scripting language.

Haploview has been widely used for genetic studies, including association studies, haplotype analysis, and calculation of SNP coverage in GWAS using the Haploview 'tagger' program.

Generation of a Manhattan Plot

A Manhattan plot is a useful method to visualise the results of a GWAS, facilitating the identification of associated SNPs with disease or traits of interest.

Data in PLINK format was loaded into the Haploview program using the 'Locus Information File' input box. Parameters can be adjusted using the pop-up window after clicking the 'Plot' button. The following parameters were adjusted:

- 'chromosomes' was selected in 'x-axis' dropdown list
- 'p' was selected in 'y-axis' dropdown list
- '-log10' was selected in 'scale' dropdown list
- '>' was selected in 'suggestive (blue line)' dropdown list and 4.3 was inputted
- '>' was selected in 'significant (red line)' dropdown list and 7.3 was inputted

All other parameters were in default setting.

The appearance of the Manhattan plot could be adjusted via the 'properties' option provided. An example Manhattan plot is shown in **Figure 2.3** using the Mayo GWAS data (Carrasquillo et al., 2009). The x-axis and y-axis represents chromosomal position and $-\log_{10}$GWAS (p-value), respectively.

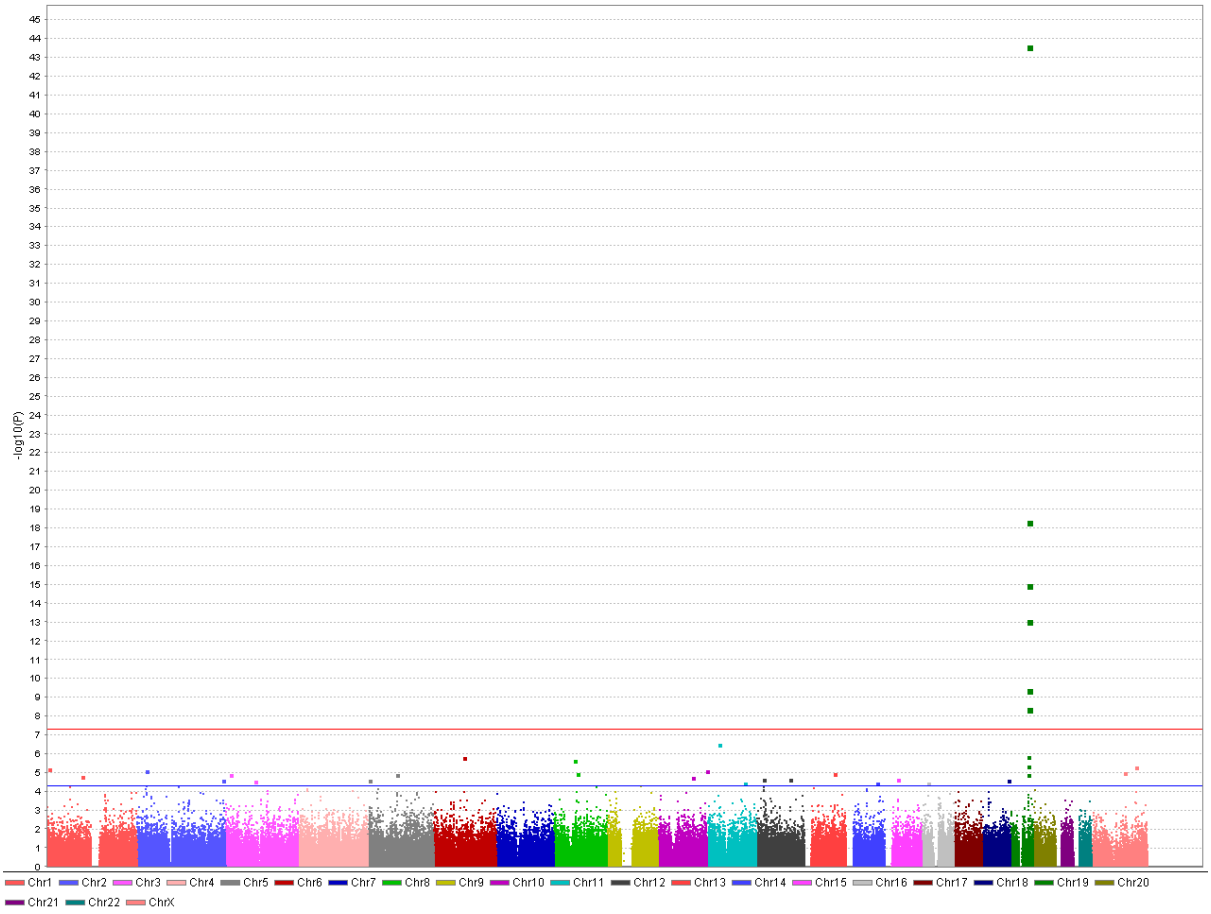



**Figure 2.3 Manhattan plot depicting GWAS output using LOAD GWAS data.** The data consists of 1,998 individuals (799 LOAD cases and 1,199 controls) and 313,330 SNPs. Chromosomal position is shown on the x-axis versus -log10 GWAS p-value on the y-axis. Red and blue horizontal lines represent p-value threshold 5 x $10^{-8}$ and 5 x $10^{-5}$, respectively. SNPs are represented by dots highlighted in different colours according to chromosomal locations. A series of green vertical dots represents SNPs in LD with *APOE* ε4 genotypes. The plot is shown for illustrative purposes only, and has not been used for actual studies.

Haploview 'tagger' program

The Haploview 'tagger' program is powerful tool to estimate the coverage of SNPs typed on a GWAS chip with respect to all known SNPs.

The HapMap genotype data release #24 was used in this calculation. Data in HapMap format was loaded into the Haploview program using the 'browse' button provided. The 'tag' and 'non-tag' SNPs were selected using tick boxes - 'force include' for 'tag' SNPs and 'force exclude' for 'non-tag' SNPs.

The coverage was calculated based on the number of SNPs (within the specified genome region) which are captured by the 'tag' SNPs. Genome regions were specified by selection of all SNPs within these regions. Selecting SNPs was undertaken using 'capture this allele?' tick boxes provided.

A SNP was defined as being captured by the 'tag' SNP if the two SNPs showed a pairwise LD $r^2 \geq 0.8$. Other parameters were in default.

An example of Haploview 'tagger' program output is demonstrated in **Figure 2.4**. In the example, two SNPs were selected as 'tag' SNPs, which captured 18% (13 out of 69) SNPs with $r^2 \geq 0.8$. 56 SNPs have not been captured.

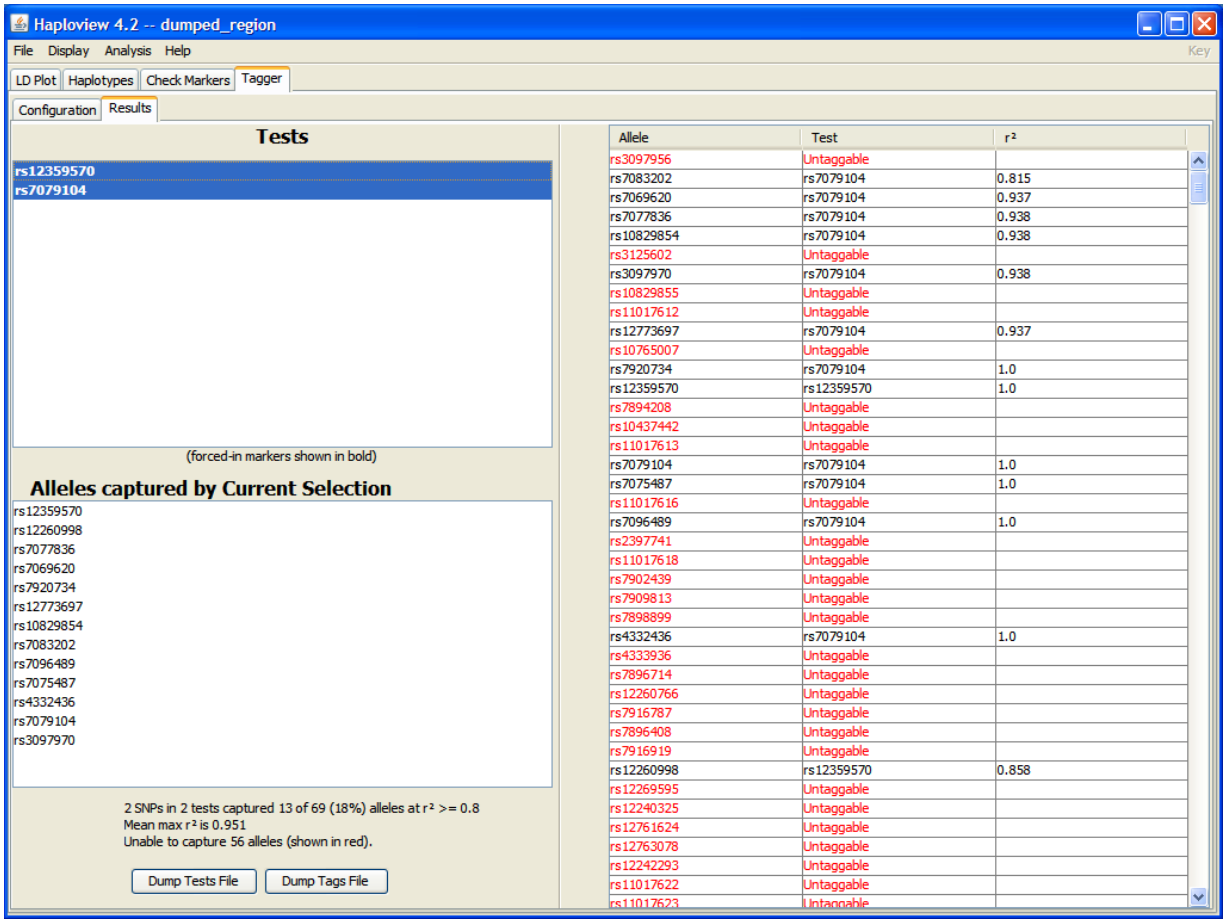



**Figure 2.4 Haploview 'Tagger' program output.** Two SNPs selected as 'Tests' (rs12359570 and rs7079104) captured SNPs listed as 'Allele captured'. The 'tagged' SNPs are listed in black font, and 'untagged' SNPs are in red font. This figure is shown for illustrative purposes only.

Generation of a Haploview LD plot

The Haploview LD plot has been widely used in genetic studies to interpret association study results, e.g. if two SNPs are in the same LD block, the association observed with one SNP could be in fact due to LD with the other functional SNP which is showing an effect.

Data were loaded into Haploview using the same method as mentioned earlier. SNPs to be included in the LD plot were selected using the tick boxes provided. The LD plot is shown by simply selecting the 'LD plot' tag.

Haploview provides three different algorithms for estimating LD blocks:

- 'confidence intervals' (the default setting)(Gabriel et al., 2002),
- 'Four Gamete Rule' (Wang et al., 2002) and
- 'Solid Spine of LD' (Barrett et al., 2005).

An example LD plot is shown in **Figure 2.5**. Colour of each rhombus represents strength of LD between SNPs:

- Red - perfect LD, reflected by both $r^2$ and D' values equal to 1. Genotypes in one SNP perfectly inform the genotype of the other.
- Blue - complete LD, where D' value equal to 1 and $r^2 < 1$. Complete LD refers to a scenario where two SNPs possess a significantly different MAF, and the alleles of the two SNPs are coupled as much as is possible given the different allele frequencies (Wray et al., 2011).
- Light red (or pink) - moderate LD, where both $r^2$ and D' value are less than 1.
- White - the two SNPs are independent.

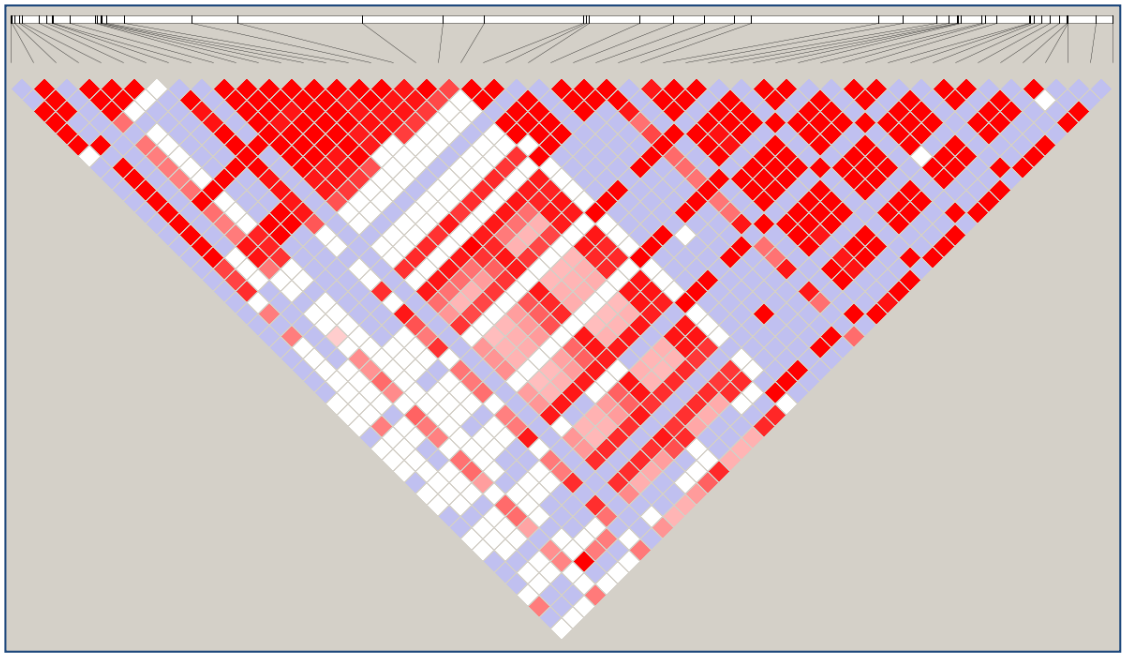


**Figure 2.5 Haploview LD plot.** The strength of LD was represented by different colours (red - perfect LD, blue - complete LD, pink - moderate LD and white - no evidence of LD). The genomic region is represented by the horizontal bar shown at the top, and the physical distance between SNPs are as indicated by connecting solid lines. This LD plot is shown for illustrative purposes only and has not been used for actual studies.

### 2.3.6 Conservation analysis

Both VISTA browser (Frazer et al., 2004) and Evolutionary Conserved Regions (ECR) browser (Ovcharenko et al., 2004) are designed to examine conservation between the human genome and the genomes of vertebrate animal species (such as mouse, rat, chimpanzee, rhesus monkey, dog, cow, opossum, chicken, frog, zebrafish).

Animals such as mouse, rat and rhesus monkey or chimpanzee are the most widely used animal species for studying conservation, likely contributed by the fact that mice and rats are the standard laboratory animals, whereas monkey and chimpanzee share high degree of homology with human.

The conservation scores are pre-computed for both VISTA and ECR browsers, allowing rapid retrieval of data from them both.

SNPs that fall in a conserved region are considered more likely to be functional than anonymous polymorphisms (Carrasquillo et al., 2009). This is also supported by the fact that regions such as exons and untranslated regions (UTRs) are more likely to be conserved than introns and intergenic regions.

Both VISTA and ECR browsers are implemented in Java programming language. They share a high degree of similarity such as both providing very similar graphical user interfaces (GUI). Sequences and annotation data utilized in the ECR browser are directly downloaded from the UCSC Genome Browser in real-time.

VISTA browser

The VISTA browser can be accessed via the website http://pipeline.lbl.gov/. For *TRIM15* gene analysis (as described in **Chapter 4**), 'Vertebrate -> Human -> Mar. 2006' was selected in the drop-down list, and base pair coordinates 'chr6:30237972-30251445' (NCBI36/HG18) was entered in the 'position' input box. The coordinates

were adjusted in order that that flanking regions of the gene are included in the analysis. The following parameters were adjusted:

- Animal species 'Rhesus, Dog, Horse, Mouse, Rat and Chicken' were selected using the 'select/add' dropdown list. The actual sequence alignments between human genome and animal genome were retrieved via the 'alignment' icon.
- The range of the conservation score was adjusted to between 50% and 100% using 'minimal y' and 'maximal y' input boxes provided.
- The significance threshold was adjusted to 70% using the 'con identity' input box.

All other parameters were in default.

With the default setting, conservation scores exhibited by the software are based on alignments of 100 nucleotide bases at a time. At each new chromosomal position, the score is recomputed by shifting one nucleotide. The final conservation plot is represented in the form of a curve. Conserved genome regions are automatically highlighted by the software in different colours.

ECR browser

The ECR browser can be accessed via the website http://ecrbrowser.dcode.org/. Genomic coordinates were input using the dropdown list and input boxes provided. Parameters were adjusted using a similar method as for the VISTA browser.

A unique feature of ECR browser is it highlights conserved regions with clickable rectangles (in pink colour). Clicking these rectangles provides access to detailed percentage identity, corresponding sequence alignment, and a hyperlink to predict transcription factor binding sites within this region.

### 2.3.7 Power Calculation for SNP discovery

The power of a statistical test is defined as the probability that it will correctly lead to the rejection of a false null hypothesis. In SNP discovery, it refers to the probability of detecting a SNP with given minor allele frequencies (MAF).

Power calculations were performed to calculate sample sizes required in order to detect a SNP. Sequencing of a single chromosome provides a probability of 0.01 to detect a SNP with MAF of 0.01. This is represented as P[detection] = 0.01. Consequently, P[non-detection] = 0.99.

If two chromosomes are sequenced, P[detection] and P[non-detection] are shown as follows:

$$\mathbf{P[detection] = 0.01x0.01 + 0.01x0.99 + 0.01x0.99 = 0.199}$$

$$\mathbf{P[nondetection] = 0.99^2 = 0.9801}$$

If 'n' chromosomes are sequenced, then

$$\mathbf{P[detection] = 1 - 0.99^n}$$

$$\mathbf{P[nondetection] = 0.99^n}$$

If 95% power is required, then

$$\mathbf{P[detection] = 1 - 0.99^n = \ 0.95}$$

Therefore,

$$\mathbf{n = \frac{[log 0.05]}{[log 0.99]} = 298}$$

Therefore, in order to detect SNPs with MAF 0.01 with 95% power, sequencing of 298 chromosomes (i.e. 149 individuals) are necessary. The following formula has been used to calculate sample size (n) with any specified power and MAFs.

$$\mathbf{n = \frac{[log(1 - power)]}{[log(1 - MAF)]}}$$

The 95% power shown here is for illustrative purpose only, rather than what has actually been used in the study.

### 2.3.8 Power calculation for detecting an association

QUANTO (v1.2.4) is a bioinformatics application designed to calculate power (or sample size required to achieve certain power) for genetic based association studies. Three statistical models are provided in QUANTO, the main effects caused by genes, gene-environment interaction and gene-gene interaction. The program provides GUIs allowing modification of input parameters.

The gene-environment interaction model was utilized, as LOAD is affected by both genetic and environmental factors (Avramopoulos, 2009). The model 'gene-environment interaction' was selected (via 'Parameters -> Outcome/Design -> Disease').

QUANTO requires mandatory configuration of four groups of parameters: 'Gene', 'Environment', 'Outcome Model' and 'Power' (**Figure 2.6**). These parameters must be adjusted in order.

'Gene G' (Genetic effect parameters used in this instance) **Figure 2.6a**

- Allele frequency: 0.01- 0.05; Increments: 0.01
- Inheritance mode: Log additive
- Susceptibility frequency: generated by the software from the allele frequency and inheritance model specified

'Environment' **Figure 2.6b**

- Population prevalence: 0.24 (representing an environmental component of LOAD equal to 0.24)

'Outcome Model' **Figure 2.6c**

- Population risk: 0.12 (an average risk of LOAD estimated between 65 to 100 years of age).
- Genetic effect size: 2 (represented by ORs)

- Environmental effect size: 1.2
- Gene-environmental interaction: no interaction was assumed
- OR summary: computed based on the inputs.

'Power' **Figure 2.6d**

- Sample size: represents the number of case-control pairs (75 was used as this represents the number of AD/Control samples analysed in Chapter 4)
- Type I error rate: 0.05 (maximum tolerated type I error)
- Perform 2 df test: yes (enables calculation of power taking into account effects from both genetic and environmental factors)

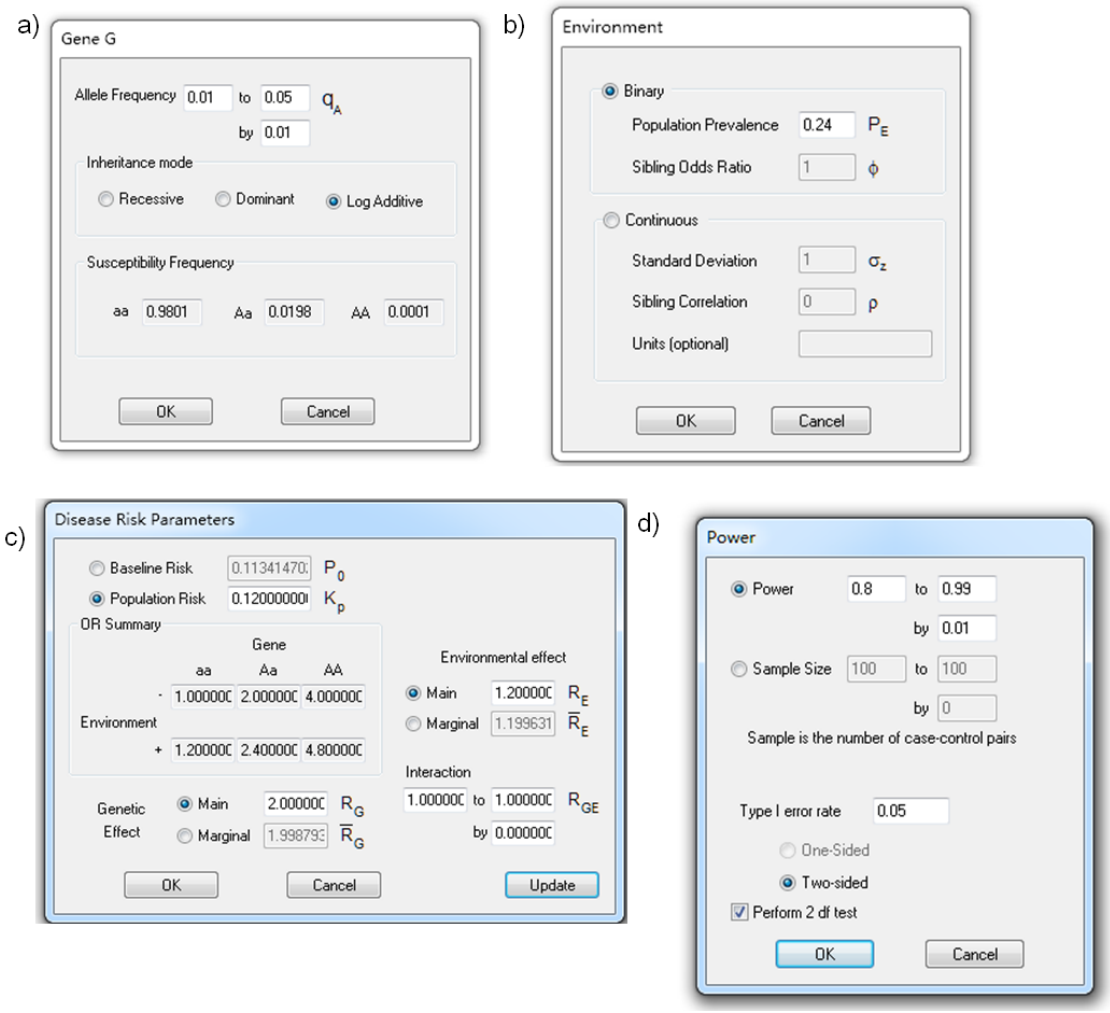



**Figure 2.6 Input dialogues for performing power calculation using QUANTO (v 1.2.4).** Four input dialogues windows: 'Gene G', 'Environment', 'Disease Risk Parameters' and 'Power' are shown with example input.

### 2.3.9 Population stratification analysis

EIGENSTRAT

The association identified in a case/control or quantitative trait GWAS analysis could be due to underlying population substructure (i.e. population stratification) and therefore must be considered in downstream analysis.

EIGENSTRAT is a bioinformatic program for calculation of heterogeneity between samples according to SNPs typed on GWAS chip platforms (Price et al., 2006). The software evaluates all possible systematic bias of allele frequencies between different GWAS datasets and presents these differences in the form of principal components (PCs). These PC values calculated can be adjusted and controlled in a standard logistic regression (or linear regression) analysis by including them as covariates. PC analysis reduces the genotype data to a number of dimensions, defined as the top eigenvectors of a covariance matrix between samples (Price et al., 2006).

EIGENSTRAT estimates genetic outliers, which are defined as any individual whose ancestry is at least 6 standard deviations (SD) from the mean on one of the top ten axes of variation. Genetic outliers are often an indication of individuals carrying suspicious genotypes possibly due to genotyping errors.

Genomic control inflation factor ($\lambda$), a representation of overall inflation of association p-values, can be calculated using EIGENSTRAT. This inflation could be due to variety of QC issues, including population stratification, centre effects and genotyping errors.

Reference datasets are required by EIGENSTRAT to provide a baseline for the analysis, HapMap data release 23 was used to fulfil this requirement. These datasets

are publicly available from HapMap (The International HapMap Consortium, 2003), three populations were downloaded are shown:

- CEU founders (release 23, 60 individuals, filtered 2.3 million SNPs) - US Utah population with Northern and Western European ancestry (samples collected in 1980 by the Centre d'Etude du Polymorphisme Humain CEPH).
- JPT + CHB founders (release 23, 90 individuals, filtered 2.2 million SNPs) - 45 unrelated Japanese in Tokyo, Japan, and 45 unrelated Han Chinese in Beijing, China.
- YRI founders (release 23, 60 individuals, filtered 2.6 million SNPs) - Yoruba people in Ibadan, Nigeria.

The filtered HapMap file included SNPs with MAF greater than 0.01 and genotyping rate greater than 0.95.

Data preparation for PC analysis

To ensure the PCs calculated (representing population stratification) were not affected by a large number of SNPs which are in LD, it is essential to create a LD pruned GWAS dataset (including both study samples and HapMap samples). It should be emphasized that it is not essential to use any specific set of SNPs since common SNPs are in LD with each other. An accurate estimation of PCs requires a minimum of 20,000 SNPs on the chip (Price et al., 2006).

The common SNPs between different GWAS datasets and HapMap data #23 were determined using an 'in house' PERL program as described in **Methods 2.3.2**. Prior to analysis using EIGENSTRAT, the HapMap CEU population dataset were pruned using the following commands in PLINK:

**plink --bfile hapmap_CEU_r23a_filtered --extract MayoSNPs.txt --make-bed --out hapmapCEUr23aM**

'--bfile' indicates the input PLINK files are in binary format. '--extract MayoSNPs.txt' extracts all common SNPs which were stored in the 'MayoSNPs.txt' file. '--make-bed' specifies the output file (in PLINK binary format), and '--out' specifies the output file name.

A LD pruned dataset was generated using the following command:

**plink --bfile hapmapCEUr23aM --indep-pairwise 1500 150 0.2 --out hapmapMP**

'--indep-pairwise 1500 150 0.2' is the main method for pruning out SNPs with pairwise LD $r^2$ value greater than 0.2 across sliding windows (window size of 1500 SNPs and 150 SNPs to shift the window). This command generated two files 'hapmapMP.prune.in' consists of LD pruned SNPs, whereas 'hapmapMP.prune.out' file is comprised of all remaining SNPs (which have been pruned out).

Each of the GWAS datasets and three HapMap datasets (as mentioned earlier) were subject to pruning using the LD pruned SNPs. Commands used to prune HapMap data CEU population are shown:

**plink --bfile hapmap_CEU_r23a_filtered --extract hapmapMP.prune.in --recode12 --out CEU_PR**

**plink --file CEU_PR --make-bed --out CEU_PRB**

'--recode12' indicates that all SNPs are converted into the same coding format ('1' and '2' coding) in PLINK format (as described in **Methods 2.3.1**). '--make-bed' indicates that standard PLINK files (PED and MAP) are converted into the PLINK binary format (BED, BIM and FAM). This process was repeated for the other populations and the GWAS dataset.

These four files were then merged into a single file using PLINK '--bmerge' command under 'Consensus call' mode.

**plink --bfile AGE_PRB --merge-list Mergelist.txt --make-bed --out merge_ACCY**

'--merge-list' is the main method for merging binary PLINK datasets. The 'Mergelist.txt' file contains three rows as shown:

**CEU_PRB.bed CEU_PRB.bim CEU_PRB.fam**

**CHBJPT_PRB.bed CHBJPT_PRB.bim CHBJPT_PRB.fam**

**YRI_PRB.bed YRI_PRB.bim YRI_PRB.fam**

SNPs with a genotyping rate less than 0.95 (--geno 0.05) were excluded (**Methods 2.3.1**). In order to calculate PC values, it is necessary that all samples are converted into 'controls' in the merged dataset.

Calculation of Eigen-values, principal components and production of an MDS plot

The following commands were used to convert the merged dataset (in PLINK binary format) into EIGENSTRAT format.

**plink --bfile merge_ACCYqc --recode --out merge_ACCYqc --noweb**

**../bin/convert -p par.PED.EIGENSTRAT**

File 'par.PED.EIGENSTRAT' contains parameters for EIGENSTRAT 'convert' command, which consists of the following lines:

**genotypename: merge_ACCYqc.ped**
**snpname: merge_ACCYqc.map**
**indivname: merge_ACCYqc.ped**
**outputformat: EIGENSTRAT**
**genotypeoutname: merge_ACCYqc.eigenstratgeno**
**snpoutname: merge_ACCYqc.snp**
**indivoutname: merge_ACCYqc.ind**
**familynames: YES**

The principal component analysis was executed using the following command,

**../bin/smartpca.perl -i merge_ACCYqc.eigenstratgeno -a merge_ACCYqc.snp -b merge_ACCYqc.ind -k 10 -o merge_ACCYqc.pca -p merge_ACCYqc.plot -e merge_ACCYqc.eval -l merge_ACCYqc.log**

'-i', '-a' and '-b' specify the input files in EIGENSTRAT format - genotype file ('.eigenstratgeno'), SNP file ('.snp') and individual file ('.ind').'-k 10' indicates the number of PCs to be shown in the output is equal to 10. '-o' specifies the output file for storing PC values. '-e' specifies the output file for storing Eigen-values.
'-l' specifies name of the log file.

Although not explicitly stated in the command line, several useful outputs were generated as listed:

- a multidimensional scaling (MDS) plot in both '.ps' and '.pdf' format.
- a '.exec' file, which was used to calculate genomic control inflation factor (λ) by EIGENSTRAT, as well as to create the covariate file for subsequent GWAS analysis using PLINK.
- genetic outliers, which were calculated automatically. The results can be found in the '.log' file.

The significance of each PC axes were calculated using the following command ('twtable' is a pre-made reference table, which was copied into the working directory).

**../bin/twstats -t twtable -i merge_ACCYqc.eval > merge_ACCYqc.Sout**

Calculation of genomic control inflation factor (λ)

Genomic control inflation factor (λ) is an important estimator of population stratification. The high λ value indicates the data is inflated, and the corresponding analysis is more likely to generate false positive outputs as a result. A λ value of equal or greater than 1.1 is often treated as unacceptable, and is an indication of the existence of significant bias in the GWAS data. EIGENSTRAT provides tools to calculate the genomic control inflation factor for data with and without correction of the PCs.

Genetic outliers were removed from the merged data using '--remove' command in PLINK,

**plink --bfile study_data --remove outlier.txt --make-bed --out new_dataset**

'outlier.txt' consists of a list of individuals identified as genetic outliers (one individual per row).

In order to proceed with the calculation, the phenotype data in PLINK PED file was replaced with the actual phenotype values (e.g. age-at-death (AAD) values as described in **Chapter 5**). The PLINK binary format was converted into EIGENSTRAT format using the 'convert' command as mentioned.

The '.pca' file was generated using command

**evec2pca.perl 10 merge_ACCYqc.pca.evec merge_ACCYqc.ind merge_ACCYqc.pca**

The command is in a format:

**evec2pca.perl $k $evec $b $o, where $k, $b and $o**

$k, $b and $o corresponds to '-k', '-b' and '-o' as previously described, and $evec specifies the '.evec' filename generated using the same methods as described.

Finally, the genomic control inflation factor was calculated using command,

**../bin/smarteigenstrat.perl -I study_data.eigenstratgeno -a study_data.snp -b study_data.ind -k 10 -p merge_ACCYqc.pca -q YES -l smarteigenstrat.log -o study_data.chisq**

**../bin/gc.perl study_data.chisq study_data.chisq.GC**

'-k' specifies the number of PCs to be adjusted for calculation of λ value. '-o' specifies the output '.GC' filename. '-p' specifies the input '.pca' filename from previous analysis. '-q YES' indicates that the analysis used are in quantitative trait phenotypes

This analysis was performed iteratively, including between 0 and 10 PCs. Each calculation generated a single λ value (11 λ values in total). The number of PC axes to be included as covariates in the GWAS analysis is ascertained when the lowest λ value was acquired after comparison of all 11 λ values (see **Chapter 5** for details).

Generate a Q-Q plot

A Q-Q plot is useful in examining the general quality of GWAS data. Two publicly available methods are available to draw a Q-Q plot - the 'estlambda' function in GenABEL (v 1.6.5) (Aulchenko et al., 2007) and 'ggd.qqplot' function (Turner et al., 2011). Both methods are written in R statistical programming language.

GenABEL 'estlambda' is more flexible than the 'ggd.qqplot' method. The former allows the user to specify the plot range using 'xlim' and 'ylim' parameters.

Furthermore, GenABEL 'estlambda' provides an approximate estimation of genomic control inflation factor, though not to the same accuracy of EIGENSTRAT calculations.

Data was loaded into R using command:

**> mydata <- read.table("filename.txt", header=T)**

'estlambda' function in GenABEL was executed using the following commands:

```
> library(GenABEL)
> estlambda(mydata$P)
```

The following function was needed to run in R 'ggd.qqplot()':

```
ggd.qqplot = function(pvector, main=NULL, ...) {
o = -log10(sort(pvector,decreasing=F))
e = -log10( 1:length(o)/length(o) )
plot(e,o,pch=1,cex=1, main=main, ...,
xlab=expression(Expected~~-log[10](italic(p))),
ylab=expression(Observed~~-log[10](italic(p))),
xlim=c(0,max(e)), ylim=c(0,max(o)))
lines(e,e,col="red")
}
```

'ggd.qqplot' was executed by typing the following command in R:

```
> ggd.qqplot(mydata$P)
```

## 2.4 Bioinformatics tools for next generation sequencing data analysis

Next generation sequencing is a high-throughput sequencing technology, which generates enormous amount of sequencing data with a much lowered financial cost than traditional Sanger sequencing. The innovation in the sequencing technology leads to profound changes in methods of analysing the sequencing data.

This methods section describes a number of bioinformatics tools for next generation sequencing data analysis as described in **Chapter 4**.

### 2.4.1 Read alignment and basic data format and manipulation

BioScope® (v 1.3)

BioScope® is a commercial software package, which is part of ABI SOLiD™ sequencing pipeline. The ABI SOLiD™ system uses a technology known as 'colour space' system or '2-barcoded encoding system', which means every single nucleotide is interrogated twice. The SOLiD™ system claims to be highly accurate with the majority of base calls achieving accuracy in excess of 99.99%.

BioScope® (v 1.3) was used to perform conversion of SOLiD colour space (CS) calls into nucleotide calls and perform alignment of short sequencing reads to the current reference genome sequences (human genome build 19 GRCh37/hg19 assembly).

The BioScope® alignment algorithm produces mapping statistics and generates a mapping quality value (range 0 - 100) for each read, which can be used to filter poorly aligned reads.

BioScope® is capable of aligning both 'mate-pair' and 'paired-end' library runs (**Chapter 4**). The output of BioScope® alignment is a mapped BAM file and a detailed report of mapping statistics.

The SNP calling algorithm and associated tools provided in BioScope® are only suitable for individual barcoded data, and not appropriate for pooled next generation sequencing data.

SAMtools

The SAM file (sequence alignment/map) and BAM file (binary format of SAM) are generic file formats for storing aligned next generation sequencing data (Li et al., 2009).

A number of tools are provided by SAMtools to manipulate the next generation sequencing data, which are compulsory for downstream analysis such as SNP calling.

After the BioScope® alignment, two BAM files were generated (separate case control pools) from the CSFASTQ and QV files. CSFASTQ and QV files are the original raw data file formats representing colour space FASTQ file and quality value file, respectively. Each of the colour space calls stored in the CSFASTQ file is provided with a quality score which is saved in the QV file.

Both BAM files are sorted and indexed using SAMtools commands:

**samtools sort F3_Morgan_control.renum.csfasta.ma.bam F3_Morgan_control.renum.csfasta.ma.sorted**

**samtools sort F3_Morgan_case.renum.csfasta.ma.bam F3_Morgan_case.renum.csfasta.ma.sorted**

**samtools index F3_Morgan_control.renum.csfasta.ma.sorted.bam**

**samtools index F3_Morgan_case.renum.csfasta.ma.sorted.bam**

Aligned reads for the two *TRIM15* fragments ('A' and 'B') were extracted from the control datasets:

**samtools view F3_Morgan_control.renum.csfasta.ma.bam 6:30130365-30143332 -bo TRIM15_AandB_BioScope1.3_Control.bam**

'-bo' specifies output filename (in binary format).

The SAMtools 'view' function was used to convert BAM to SAM, as well as to retrieve a subset of next generation sequencing data. The retrieved sequencing data was sorted and indexed (which generates a new index BAI file):

**samtools sort TRIM15_AandB_BioScope1.3_Control.bam TRIM15_AandB_Bioscope1.3_ControlS**

**samtools index TRIM15_AandB_BioScope1.3_ControlS.bam**

The genome reference FASTA file was indexed using 'faidx' function in SAMtools (which generates a new FASTA index FAI file):

**samtools faidx valid_6.fa**

A pileup file (position based output) was generated using the 'pileup' function:

**samtools pileup -vcf valid_6.fa TRIM15_AandB_BioScope1.3_ControlS.bam**

The same commands were used for extracting, indexing and sorting of sequencing reads for the case pool.

Additional functions are provided in SAMtools,

- 'merge' function, which allow user to merge multiple sorted BAM files
- 'tview' function, an alignment viewer.

SAM file format

SAM file is in a tab-delimited text file format which consists of an optional header section and a compulsory alignment section. Each header starts with the '@' symbol and is followed by a two-letter code (e.g. @RG) and a colon (':'). **Table 2.5** lists all these two-letter codes (known as 'TAGs') and their definitions.

An example header row is shown:

**@RGID:20101014202018783PL:SOLiDLB:lib1-50FPI:0DT:2010-10-14T13:20:18-0700SM:Morgan_controlCN:freetext**

This read header can be interpreted as shown:

- Read group (RG) identifier (ID): 20101014202018783
- Platform (PL): SOLiD
- Library (LB): 1 to 50F
- Predicted median insert size (PI): 0
- Date of the run was produced (DT): 2010-10-14T13:20:18-0700
- Sample (SM): Morgan_control
- Name of the sequencing centre producing the read: freetext

It is noteworthy that each read from an individual pool is labelled with the same @RG, reflecting the pooling strategy used.

In the alignment section, each read occupies one row, consists of 11 mandatory fields for storing essential alignment information, as well as variable number of optional fields for flexibility (**Figure 2.7**).

**Table 2.5 Header summary of the SAM file format.** Figure showing a list of two-letter header codes and their definitions (Adopted from Li et al., 2009).

| Tag | | Description |
|---|---|---|
| **@HD** | | The header line. The first line if present. |
| | **VN*** | Format version. *Accepted* format: /^[0-9]+\.[0-9]+$/. |
| | **SO** | Sorting order of alignments. *Valid values*: **unknown** (default), **unsorted**, **queryname** and **coordinate**. For coordinate sort, the major sort key is the **RNAME** field, with order defined by the order of **@SQ** lines in the header. The minor sort key is the **POS** field. For alignments with equal **RNAME** and **POS**, order is arbitrary. All alignments with `*' in **RNAME** field follow alignments with some other value but otherwise are in arbitrary order. |
| **@SQ** | | Reference sequence dictionary. The order of **@SQ** lines defines the alignment sorting order. |
| | **SN*** | Reference sequence name. Each **@SQ** line must have a unique **SN** tag. The value of this field is used in the alignment records in **RNAME** and **PNEXT** fields. Regular expression: [!-)+-<>-~][!-~]* |
| | **LN*** | Reference sequence length. *Range*: [1, $2^{29}$ −1] |
| | **AS** | Genome assembly identifier. |
| | **M5** | MD5 checksum of the sequence in the uppercase, with gaps and spaces removed. |
| | **SP** | Species. |
| | **UR** | URI of the sequence. This value may start with one of the standard protocols, e.g http: or ftp:. If it does not start with one of these protocols, it is assumed to be a file-system path. |
| **@RG** | | Read group. Unordered multiple @RG lines are allowed. |
| | **ID*** | Read group identifer. Each **@RG** line must have a unique **ID**. The value of **ID** is used in the **RG** tags of alignment records. Must be unique among all read groups in header section. Read group **IDs** may be modified when merging SAM files in order to handle collisions. |
| | **CN** | Name of sequencing center producing the read. |
| | **DS** | Description. |
| | **DT** | Date the run was produced (ISO8601 date or date/time). |
| | **FO** | Flow order. The array of nucleotide bases that correspond to the nucleotides used for each flow of each read. Multi-base flows are encoded in IUPAC format, and non-nucleotide flows by various other characters. *Format*: /*\|[ACMGRSVTWYHKDBN]+/ |
| | **KS** | The array of nucleotide bases that correspond to the key sequence of each read. |
| | **LB** | Library. |
| | **PG** | Programs used for processing the read group. |
| | **PI** | Predicted median insert size. |
| | **PL** | Platform/technology used to produce the reads. *Valid values*: **CAPILLARY**, **LS454**, **ILLUMINA**, **SOLID**, **HELICOS**, **IONTORRENT** and **PACBIO**. |
| | **PU** | Platform unit (e.g. flowcell-barcode.lane for Illumina or slide for SOLiD). Unique identifier. |
| | **SM** | Sample. Use pool name where a pool is being sequenced. |
| **@PG** | | Program. |
| | **ID*** | Program record identifier. Each **@PG** line must have a unique **ID**. The value of **ID** is used in the alignment **PG** tag and **PP** tags of other **@PG** lines. **PG IDs** may be modified when merging SAM files in order to handle collisions. |
| | **PN** | Program name |
| | **CL** | Command line |
| | **PP** | Previous **@PG-ID**. Must match another **@PG** header's **ID** tag. **@PG** records may be chained using **PP** tag, with the last record in the chain having no **PP** tag. This chain defines the order of programs that have been applied to the alignment. **PP** values may be modified when merging SAM files in order to handle collisions of **PG IDs**. The first **PG** record in a chain (i.e. the one referred to by the **PG** tag in a SAM record) describes the most recent program that operated on the SAM record. The next **PG** record in the chain describes the next most recent program that operated on the SAM record. The **PG ID** on a SAM record is not required to refer to the newest **PG** record in a chain. It may refer to any **PG** record in a chain, implying that the SAM record has been operated on by the program in that **PG** record, and the program(s) referred to via the **PP** tag. |
| | **VN** | Program version |
| **@CO** | | One-line text comment. Unordered multiple **@CO** lines are allowed. |

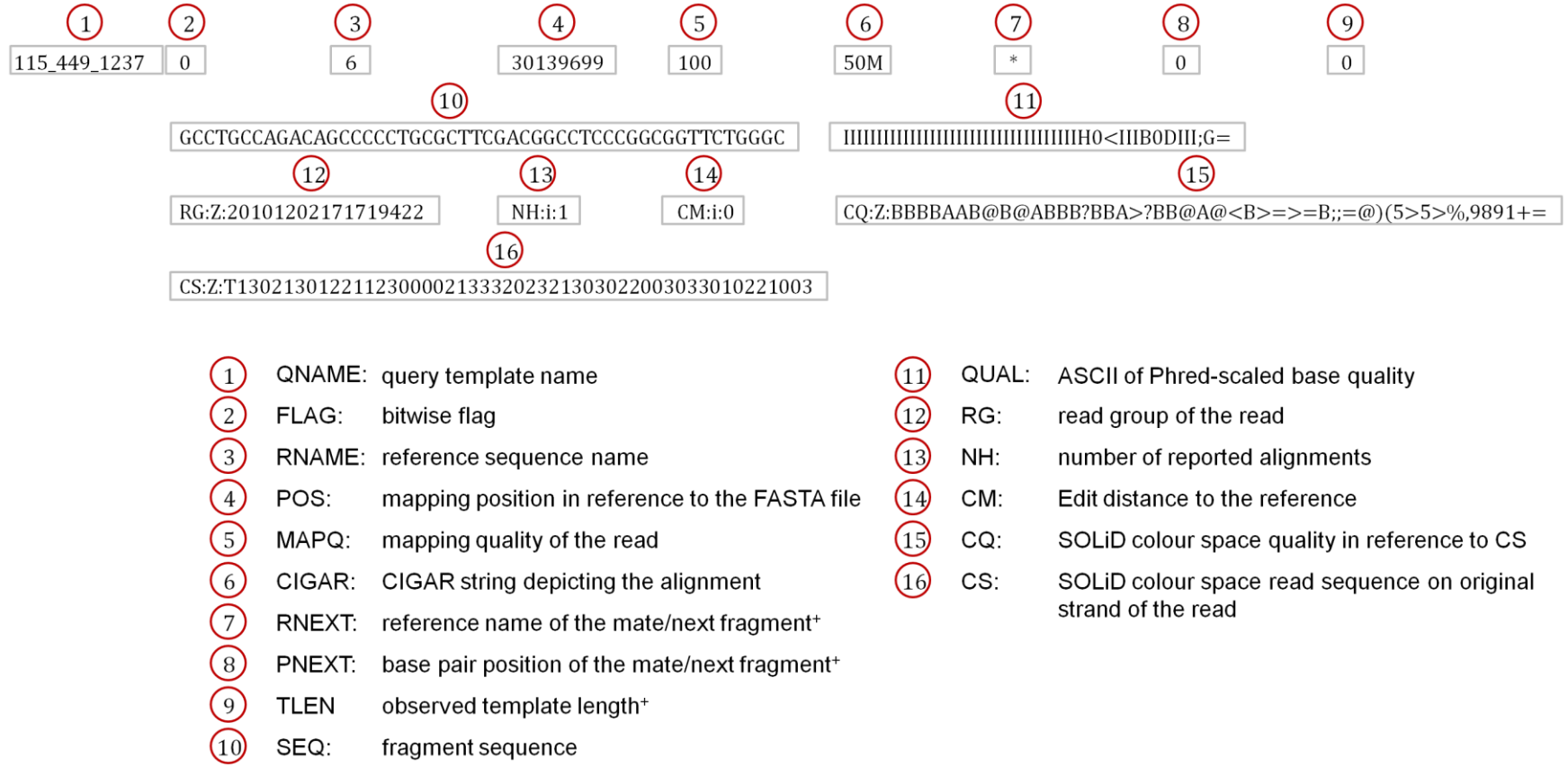



**Figure 2.7 Summary of SAM file format.** Figure showing all elements of an individual read in the SAM file format, and corresponding descriptions. '+' indicates the relevant field is inapplicable to the 'paired end' sequencing data (See **Chapter 4** for definition).

PILEUP file format

Each row of the pileup file describes alignment information of reads at each nucleotide base position. There is no header section in the pileup file.

Two pileup formats are in use. The 'standard sequence pileup' format consists of six columns: chromosome number, base pair position, reference nucleotide base, number of reads aligned to this site, read bases and base qualities.

An example row of the SAMtools standard '.pileup' file is shown:

**1 10000 T 22 ….,,.,.C.,,,.,..G. +7<;<<<<<<<&<=<<:;<<&<**

In this example, there are 22 reads mapped to this chromosomal position, on chromosome 1, at base coordinate 10000 and with the reference allele 'T'. The dot ('**.**') symbol in read bases represent a single read is mapped to the forward strand and the base matches the reference allele 'T'. The comma ('**,**') symbol indicates a read, which matched to the reference allele, however mapped to the reverse strand. Any other letters indicate a possible variant (e.g. 'C' and 'G' is shown in this example), where uppercase lettering indicates a non-reference nucleotide base mapped to the forward strand and a lowercase indicates the base is mapped to the reverse strand. The quality of each nucleotide base was represented in a single ASCII code (**+7<;<<<<<<<&<=<<:;<<&<**); interpretation of this code is discussed in more detail in **Chapter 4**, **Table 4.1**.

'Consensus sequence pileup' format differs from the 'standard sequence pileup' where it consists of four additional columns between the 'reference nucleotide base' and 'number of reads aligned' columns. These four columns are consensus base, consensus quality, SNP quality and maximum mapping quality, respectively.

### 2.4.2 FastQC (v0.9.4) quality assessment

FastQC is a bioinformatic program designed for assessment of quality of raw next generation sequencing data. Obvious sequencing errors can be effectively revealed by performing QC checks using the software.

Several input formats are supported, including SAM, BAM, FASTQ and CSFASTQ.

The program performs a series QC checks: basic sequence stats, sequence quality (per base and per read), per base sequence content, per base GC content, per sequence GC content, sequence length distribution and any evidence of sequence over-representation.

Each of the tests performed is automatically flagged as a pass (green tick), warning (an orange warning sign) or failed (a cross symbol in red) according to the QC calculations.

The software is written entirely in Java and provided with a graphical user interface (GUI). The analysis reports from FastQC can be saved in HTML format via ‘File -> Save’.

### 2.4.3 SNP calling

Syzygy (v 1.1.0)

Syzygy is a SNP calling software designed specifically for pooled next generation sequencing data analysis. Syzygy is written in python programming language, which has several pre-requisites which influence how the program is installed in Linux. One of the pre-requisites is SAMtools as mentioned earlier.

Syzygy calls SNPs and calculates error rates of each prediction based on error models. Error models are generated while performing the analysis and calling SNPs. The error rates generated determine whether the non-reference observations from sequencing are variants or errors. Syzygy takes into account allele strand biases and calculates a LOD score of strand bias (also known as 'SLOD score'), allowing a genuine SNP call to be distinguished from errors.

Syzygy input file format

Two compulsory files are required by Syzygy - 'Target Info File' ('.tgf' file) which contains information about the sequencing target (DNA amplicons) and 'Pool Info File' ('.pif' file) which contains information about the pooling strategy.

The '.tgf' file has six columns and '.pif' file has four columns, and must be generated in the format demonstrated in **Table 2.6**.

**Table 2.6 Syzygy input files ('.tgf' and '.pif' file) for *TRIM15* 'A' and 'B' amplicons.** 'tgf' and 'pif' file formats for the *TRIM15* 'A' and 'B' amplicons, including the obligatory header sections.

'.tgf' file

| FEATURE_NAME | CHR | START_POSITION | END_POSITION | LENGTH | GENOME_BUILD |
|---|---|---|---|---|---|
| TRIM15_AandB | 6 | 30130365 | 30143332 | 12968 | 19 |

'.pif' file

| PoolBAM | Phenotype | Inds | Chroms |
|---|---|---|---|
| TRIM15_AandB_BioScope1.3_ControlS.bam | 0 | 75 | 150 |
| TRIM15_AandB_BioScope1.3_CaseS.bam | 1 | 75 | 150 |

Syzygy Implementation

Syzygy was executed using the following command:

**syzygy --pif TRIM15.pif --tgf TRIM15.tgf --samtoolspath /usr/bin --outputdir /home/mrxhs/deepseq/job6 --hg 19 --ref valid_6.fa --dbsnp TRIM15.dbsnp --skipannot true --mqthr 50 --bqthr 10 --power --rarethr 0.01**

'--samtoolpath' specifies the file directory location of SAMtools program. '--outputdir' specifies the output directory. '--hg 19' indicates human genome build 19 was used in the analysis. '--ref' specifies the reference sequence FASTA file. '--dbsnp' specifies the dbSNP file (which was downloaded from UCSC website). '--bqthr 10' and '--mqthr 50' indicate the threshold for base call quality and mapping quality is equal to 10 and 50, respectively. '--power' instructs Syzygy to calculate power for detection of a singleton. '--rarethr 0.01' indicates the rare variant threshold is equal to 0.01.

FreeBayes (v 0.4.2)

FreeBayes is a bioinformatic tools for calling SNPs developed by Marth and colleagues at Boston College (http://bioinformatics.bc.edu/marthlab/FreeBayes). It is an extension of the original Bayesian SNP caller PolyBayes (Marth et al., 1999). FreeBayes supports analysis of both pooled sequencing data and individually barcoded data.

FreeBayes is flexible and fast, and provides accurate estimations of allele frequencies together with useful information including read depth of the nucleotide base, alternative allele counts and number of reads aligned to the forward and reverse strand.

FreeBayes generates results in Variant call file (VCF) format, a format which has been widely adopted for next generation sequencing data analysis.

The software can also accommodate insertions and deletions via the inclusion of '--indels' and '--left-align-indels'.

One of the limitations of FreeBayes is that it does not distinguish between high and low quality SNPs as does Syzygy. Furthermore, although the software provides an estimation of SNP quality by 'pval' (p-value) in the output, it is not sufficiently stringent, as reflected by unrealistic number of predicted SNPs acquired 'pval' value of 1 (the highest p-value allowed to be specified).

The following command was used for analysis of *TRIM15* 'A' and 'B' amplicons (as described in **Chapter 4**).

**freebayes --fasta-reference valid_6.fa TRIM15_AandB_BioScope1.3_ControlS.bam --pooled --ploidy 150 --pvar 1 --min-mapping-quality 50 --min-base-quality 10 --region 6: 30130365..30143332**

'--fasta-reference' specifies the genome reference ('.fa' or '.fasta'), and next generation data file in BAM format. '--pvar 1' indicates the p-value (confidence of calling a SNP) is equal to 1. '--min-mapping-quality 50' indicates the mapping quality threshold is equal to 50, and '--min-base-quality 10' indicates the base quality threshold is equal to 10. '--region 6:30138938..30143332' specifies the region analysed is on chromosome 6, base pair position from 30138983 to 30143332.

VCF file format

VCF (variant call file) is a generic file format designed for storing variants information (including SNPs, Indels and structural variants) together with detailed annotations. The VCF format is compact in size and both flexible and easily extensible for further development. Furthermore, VCF files can be indexed by a program known as 'tabix',

allowing fast and simple data retrieval of variants from a range of positions on the reference genome (Li, 2011). An example command is as shown:

**tabix -hf ftp://ftp.1000genome.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/ALL.chr6.phase1.projectConsensus.genotypes.vcf.gz 6:30139699-30139699**

The VCF file is comprised of three sections: meta-information, header section and data section.

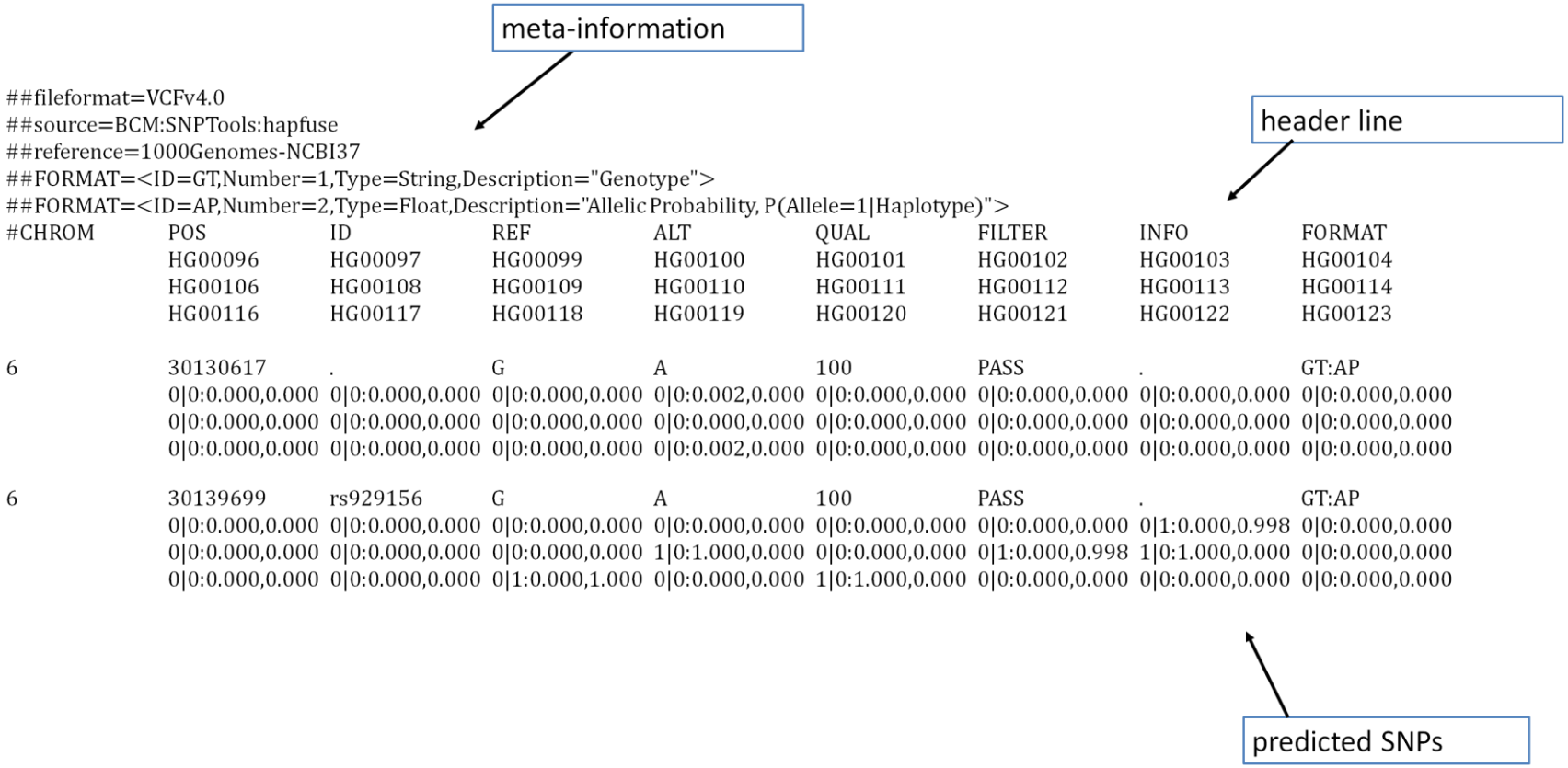
Meta-information (starts with a'##' sign) stores:

- VCF file version (depicted as 'fileformat'),
- Date,
- Source (e.g. syzygy 1.1.0),
- Filename of the reference genome, and
- Definitions of all annotations included in the VCF file.

Header line (starts with a '#' sign) consists of eight mandatory fields as listed:

- CHROM - chromosome number
- POS - base pair position
- ID - dbSNP rs number
- REF - reference bases
- ALT - alternative base,
- QUAL - Phred scaled quality score of the alternative allele
- FILTER - quality filters
- INFO - and additional information in format <key>=<data>

Predicted SNPs (third section) are presented in an ascending order of base pair positions (one SNP per row). An example VCF file downloaded using tabix from 1000 genome project is shown in **Figure 2.8**.




```
##fileformat=VCFv4.0
##source=BCM:SNPTools:hapfuse
##reference=1000Genomes-NCBI37
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AP,Number=2,Type=Float,Description="Allelic Probability, P(Allele=1|Haplotype)">
#CHROM  POS       ID        REF     ALT     QUAL    FILTER  INFO    FORMAT
                  HG00096   HG00097 HG00099 HG00100 HG00101 HG00102 HG00103 HG00104
                  HG00106   HG00108 HG00109 HG00110 HG00111 HG00112 HG00113 HG00114
                  HG00116   HG00117 HG00118 HG00119 HG00120 HG00121 HG00122 HG00123

6       30130617  .       G       A       100     PASS    .       GT:AP
        0|0:0.000,0.000 0|0:0.000,0.000 0|0:0.000,0.000 0|0:0.002,0.000 0|0:0.000,0.000 0|0:0.000,0.000 0|0:0.000,0.000 0|0:0.000,0.000
        0|0:0.000,0.000 0|0:0.000,0.000 0|0:0.000,0.000 0|0:0.000,0.000 0|0:0.000,0.000 0|0:0.000,0.000 0|0:0.000,0.000 0|0:0.000,0.000
        0|0:0.000,0.000 0|0:0.000,0.000 0|0:0.000,0.000 0|0:0.002,0.000 0|0:0.000,0.000 0|0:0.000,0.000 0|0:0.000,0.000 0|0:0.000,0.000

6       30139699  rs929156  G       A       100     PASS    .       GT:AP
        0|0:0.000,0.000 0|0:0.000,0.000 0|0:0.000,0.000 0|0:0.000,0.000 0|0:0.000,0.000 0|0:0.000,0.000 0|1:0.000,0.998 0|0:0.000,0.000
        0|0:0.000,0.000 0|0:0.000,0.000 0|0:0.000,0.000 1|0:1.000,0.000 0|0:0.000,0.000 0|1:0.000,0.998 1|0:1.000,0.000 0|0:0.000,0.000
        0|0:0.000,0.000 0|0:0.000,0.000 0|1:0.000,1.000 0|0:0.000,0.000 1|0:1.000,0.000 0|0:0.000,0.000 0|0:0.000,0.000 0|0:0.000,0.000
```

**Figure 2.8 Summary of VCF format.** Figure showing an example VCF file from 1000 genome project accessed using tabix software. The 1000 genome project data is freely available at website ftp://ftp.1000genome.ebi.ac.uk. A known SNP (rs929156) at chromosome 6 base pair position 30139699 is shown together with a novel SNP at chromosomal 6 base pair position 30130617.

### 2.4.4 SNPs annotation

Variant Classifier

Variant Classifier is a program which annotates genetic variants (SNPs, insertion and deletion) from analysis of next generation sequencing data.

A list of all current databases was displayed using script 'Show_Latest_Databases.pl' included in the program. The latest annotation data ('.coding_info' file) and the corresponding reference sequence '.fasta' file were downloaded from the most up-to-date Ensembl database using command:

**perl Extract_Cding_Info.pl -c 6 -b 30130365 -e 30143332 -O "homo sapiens" -B 60 -A 37e -f TRIM15_AandB_BioScope1.3 -x**

'-c 6' indicates chromosome number is equal to 6. '-b' and '-e' specifies the start and end base pair coordinates. '-O homo sapiens' indicates that human species is selected. '-B 60' indicates the retrieved data is in NCBI build version '60'. '-A 37e' indicates the NCBI assembly version is '37e'. '-f' specifies the output filename. '-x' instructs the software to extend the coverage to include the whole gene if necessary.

The annotation function of Variant Classifier requires an input file 'input_snps', which is comprised of 4 columns as listed:

- column 1 - SNP starting base pair position (it is compulsory that this position matches the coordinate saved in the '.coding_info' file);
- column 2 - SNP ending position (starting position plus 1);
- column 3 - strand (positive strand is noted as 1 and negative strand is denoted as 2);
- column 4 - alternative allele.

This input file does not include any header.

The annotation was executed using the commands:

**perl Classify_SNPs.pl -s input_snps -c**

**TRIM15_AandB_BioScope1.3.coding_info -n**

**TRIM15_AandB_BioScope1.3.fasta -o results.txt**

'-s' specifies the input file name as described. '-c' and '-n' specify the annotation data file ('.coding_info') and reference sequence file ('.fasta'), respectively. '-o' specifies the output file name.

Two result files were generated - 'output.normal' and 'output.denormal'. Both files contain the same information, where 'output.denormal' is designed to be read and manipulated by computer.

Polyphen-2

Polyphen-2 is a web-based bioinformatic tool for predicting SNP pathogenicity. It measures possible impact of an amino acid substitution on the structures and functions of the encoded protein (Adzhubei et al., 2010). Non-synonymous SNPs are characterized into three distinct risk groups - benign, possibly damaging and probably damaging.

*TRIM15* protein sequence was entered into the 'Amino acid sequence in FASTA format' input box. The amino acid position where the change occurred was entered into the 'position' input box. In 'Substitutions', the reference amino acid 'AA1' and substituted amino acid 'AA2' (caused by the mutation) were selected according to results obtained from Variant Classifier (See **Chapter 4** for details).

### 2.4.5 Visualisation tools

Integrative Genomic Viewer (IGV)

IGV is a visualisation tool for 'real-time' exploration of large-scale next generation sequencing data. The software is implemented in Java (Robinson et al., 2011).

IGV viewer is freely available at http://www.broadinstitute.org/software/igv/download. Two files are required by IGV: a BAM file and a reference genome FASTA file. Both files must be sorted and indexed using methods as previously described.

A '.genome' file is generated automatically after loading the FASTA file into IGV. An input box is provided to enable viewing of a user-specified genomic region.

UCSC genome browser

The UCSC genome browser provides a number of useful web-based bioinformatics tools - 'liftOver', 'In-Silico PCR' and 'UCSC custom tracks'.

liftOver

Human genome sequence is constantly under review due to technological advancement and clarification of existing data. As a result, multiple genome sequence assemblies exist, each differs in base pair coordinates. Base pair coordinates must be transformed into the same genome build before a comparison can be conducted. 'liftOver' is a tool designed to perform this conversion.

The latest genome assembly (Hg19) is currently in operation and was used by the 1000 genome project (The 1000 Genomes Project Consortium, 2010). Base pair coordinates (in BED format) are required for 'liftOver' function.

In-Silico PCR

The ‘In-Silico PCR’ program on UCSC website allows the user to search for the target DNA sequence by simply entering the PCR forward and reverse primer sequences. This tool is useful to examine whether the PCR primer designed target the correct DNA template sequence of interest.

Input boxes are provided to allow PCR primers to be uploaded. Adjustment of ‘max product size’ may be necessary depending on the size of expected PCR amplicon.

UCSC custom track

The custom track function allows users to view their own data in the UCSC genome browser and to be displayed as a ‘custom track’. Custom tracks (in BED and WIG format) can be loaded into UCSC genome browser using the ‘browse’ button provided. These input files were used to display *TRIM15* ‘A’ and ‘B’ amplicons (**Chapter 4**) as documented in **Appendix 8.5**.

# Chapter 3: Analysis of Genome Wide Association Study (GWAS) data looking for replicating signals in LOAD

## 3.1 Introduction

Alzheimer's disease is the most prevalent form of dementia. As life expectancies continue to rise, an increasing number of individuals are expected to develop AD. The number of LOAD cases worldwide was recorded as 26.6 million in 2007, this figure has been estimated to rise to over 100 million by 2050 (Brookmeyer et al., 2007). Understanding the genetic aetiology of LOAD could enable the development of effective therapeutic treatment.

Despite tremendous efforts over the last few decades, identification of genetic loci underlying LOAD has been proven difficult, with the ε4 allele of *APOE* being the only established, reproducible genetic risk factor prior to the discovery of new LOAD risk genes in 2009 (Harold et al., 2009). Genes explored in previous candidate gene studies are often based on pre-conceived functional and biological hypotheses. As a result, genes that are closely related to Aβ and tau have been extensively studied in the pathogenesis of LOAD. However, genetic defects found in genes such as *APP*, *PSEN1* and *PSEN2* do not appear to contribute to risk for LOAD, but are tightly linked to early onset Alzheimer's disease (Bertram, 2011).

GWAS in LOAD has generated significant, reproducible findings and given insight into the biological aetiology of LOAD. Nine new LOAD genes *(CLU, PICALM, CR1, BIN1, ABCA7, MS4A6A, CD33, CD2AP* and *EPHA1*) have been identified through recent large GWAS (Harold et al., 2009; Hollingworth et al., 2011; Lambert et al., 2009; Naj et al., 2011; Seshadri et al., 2010). These genes provide new impetus for drug development which could aid in slowing down disease progression and ultimately developing a cure based on the grounds of genetic associations with LOAD.

Insufficient power has meant a large number early GWAS failed to generate any significant 'hits' (McCarthy et al., 2008). These GWAS lacked the numbers of cases and controls required to detect a modest effect (ORs = ~1.25) from a common variant (Bertram et al., 2007). Despite these power issues, they are still a valuable source of data for meta-analysis purposes. Combining individually underpowered GWAS could increase power thus allowing identification of genuine associations and previous spurious associations will likely diminish.

Unfortunately, there are a number of constraints that have limited the effectiveness of whole-genome meta-analysis to date. Given that GWAS may use different genotyping platforms (such as Illumina or Affymetrix) each assaying different panels of SNPs, the number of 'matched' SNPs available for meta-analysis is limited. This is often confounded by SNP dropout during quality control procedures.

## 3.2 Aims

Genetic markers with suggestive association p-value ($5 \times 10^{-5} < p < 5 \times 10^{-8}$) may be genuine AD candidates that due to power constraints, have failed to reach genome-wide significance ($p < 5 \times 10^{-8}$).

The aim of this study was to select genes/regions that merit further investigation by identifying all SNPs with p-values within this range ($5 \times 10^{-5} < p < 5 \times 10^{-8}$) and comparing their effects across several GWAS, either directly or by using a perfect proxy ($r^2 = 1$). The approaches employed to identify replicating signals in this study can be applied to other studies to search across GWAS data from different platforms.

## 3.3 Strategy

A cross-platform comparison of four GWAS was conducted using data that are readily obtainable; subject-level genotype data from two, Reiman et al., 2007 and Carrasquillo et al., 2009, complete summary data from a third Li et al., 2008 and summary data of top SNP hits ($5 \times 10^{-5}$ to $5 \times 10^{-8}$) in the fourth Beecham et al., 2009. In each case, quality control measures had been applied by the authors prior to data release (**Table 3.1**).

Generating SNP results from subject-level genotype data (Carrasquillo et al., 2009; Reiman et al., 2007)

Datasets Reiman et al., 2007 and Carrasquillo et al., 2009 were analysed using the PLINK analysis toolset version 1.06 (**Methods 2.3.1**). GWAS data was converted into a file format appropriate for PLINK (PED and MAP) before analysis. GWAS outputs were generated from genotyping data using '--assoc' command.

To make the Reiman et al., 2007 Affymetrix data comparable with Carrasquillo et al., 2009 Illumina data, the SNP ID was translated from Affymetrix SNP ID format to dbSNP ID format (rs number). A PERL script was written to perform the translation process (**Methods 2.3.2**; **Methods 8.4.1**).

As the sex status of individuals was unspecified in the Reiman et al., 2007 dataset, the '--allow-no-sex' command was utilised to instruct PLINK to ignore unspecified sex and include all samples in the calculations.

Only limited information was obtained for the Beecham et al., 2009 and Li et al., 2008 studies. It was not possible to merge datasets, since the two studies used different chip platforms.

**Table 3.1 Summary of the four GWAS analysed in this study.** The number of SNPs following QC, the platform utilised and the percentage of SNPs excluded in each study is listed. Also shown are the number of perfect proxies ($r^2 = 1$) in the data (post QC) together with the number of clusters into which these SNPs fall. The number of independent tests for multiple testing corrections of combined p-values is shown in the last column and was calculated as described (**Methods 2.3.3**).

| Study | Number of SNPs (post QC) | CHIP platform | Excluded SNPs (%) | Number of SNPs with LD ($r^2 = 1$) | Number of LD Clusters ($r^2 = 1$) | Number of Independent Tests |
|---|---|---|---|---|---|---|
| Beecham et al., 2009 | 532,000 | Illumina 550 | 4% | - | - | - |
| Carrasquillo et al., 2009 | 313,330 | Illumina 300 | 1% | 26,284 | 11,539 | 298,585 |
| Li et al., 2008 | 469,438 | Affymetrix 500K | 5% | 128,139 | 42,634 | 383,933 |
| Reiman et al., 2007 | 312,316 | Affymetrix 500K | 38% | 83,739 | 29,678 | 258,255 |

Comparing p-values across different GWAS

For each of the GWAS, all SNPs with p-values between $5 \times 10^{-5}$ to $5 \times 10^{-8}$ were compared across the other studies (where possible) either directly or by using a perfect proxy ($r^2 = 1$). SNAP (SNP Annotation and Proxy Search) (http://www.broad.mit.edu /mpg/snap) was used to identify SNP proxies using the HapMap Resource CEU population - release 23 as the reference dataset (**Methods 2.3.4**).

Direct proxies were used in order to capture the maximum number of SNPs across the different chip platforms (each has their own SNP portfolio). Imputation attempts for SNPs in *TRIM15* using PLINK yielded limited information when merging the datasets with the reference datasets. Imputed SNPs generated PLINK INFO (information content metric) scores lower than 0.8, indicating unreliably imputed SNPs. This low score is due to poor LD architecture within this region and the limited availability of data.

The significance band $5 \times 10^{-5}$ to $5 \times 10^{-8}$ was used to search for potential new AD candidates that have failed to reach genome-wide significance due to limited power of the GWAS to date. Extending to a lower cut-off ($p > 10^{-5}$) may reveal more substantial information and this could well be a viable approach to use on larger GWAS datasets as they become available. Any SNPs with p-values below $5 \times 10^{-8}$ were not included in this analyses as they would have been identified as genome wide significant; effectively this resulted in all SNPs in the *APOE* region on chromosome 19 being removed – this region replicated across all the studies.

SNPs were selected for further analysis as described below:

- SNPs with p-values $5 \times 10^{-5}$ to $5 \times 10^{-8}$ were selected from each of the GWAS.
- SNP p-values were determined for the same SNPs (or proxies $r^2 = 1$) across the remaining studies.

- The Fisher's combined p-value test was used as a summary statistic to give an overall value of association. It has to be noted that this test does not correct for disparate effects created by alleles whose direction of association differs between studies – the so-called 'flippers'. For the resultant p-value to be meaningful all effects must be in the same direction.
- Combined p-values were corrected for the number of independent SNPs on the highest density platform utilised following QC (**Methods 2.3.3**).

It is only possible to access the 'top hits' from Beecham et al., 2009, which limited the comparison across all four studies.

Meta-analysis of odds ratios

Any SNPs that showed a corrected combined p-value ($p < 0.05$) were further analysed by comparing their corresponding odds ratios across multiple GWAS datasets. The random-effects method was implemented in the StatsDirect software package. In contrast to Fisher's combined probability test, random-effect meta-analysis accounts for the direction of effect. Significance is only obtained when the effects are all in the same direction. A SNP could therefore be significantly associated using Fisher's method but fail odds ratio meta-analysis.

Gene-centric analysis for *TRIM15*

A gene-centric approach was used to conduct an in depth SNP analysis of *TRIM15*, the only genetic locus achieved significant by both Fisher's combined probability test and random-effect meta-analysis (See results for details). The LD architecture surrounding this gene was identified using LD plots generated in Haploview (v 4.0) using HapMap CEU population data (**Methods 2.3.5**). SNPs flanking the gene (20 kb either side) were also analysed. The base pair coordinates were obtained from HapMap database. The study-specific p-values for allelic association for each of the *TRIM15* SNPs were generated in PLINK using the data from the Reiman et al., 2007

and Carrasquillo et al., 2009; the values from the summary data were used for Li et al., 2008.

## 3.4 Results

Analysis of GWAS

The SNP's with p-values 5 x $10^{-5}$ to 5 x $10^{-8}$ were identified for each study and then compared across all datasets. Four tables were created, one table for each of the GWAS listing the SNPs that were in this significance band together with the corresponding SNP p-values in the three other GWAS (irrespective of their significance values). **Table 3.2** compares GWAS output for all four studies, whereas tables **Table 3.3**, **Table 3.4** and **Table 3.5** compare data from the remaining three GWAS.

Combined p-values were determined for SNPs that occurred in at least two studies. SNPs with combined p-values of $10^{-8}$ were corrected for multiple testing. Using this approach, three SNPs were identified. SNP rs929156 (**Table 3.2** – Beecham et al., 2009 as primary comparator) had a combined p-value of 8.77 x $10^{-8}$, corrected p-value ($p$ = 0.0467); this occurs in an exonic sequence of the *TRIM15* gene on chromosome 6. Using Li et al., 2008 as the primary dataset to compare with, SNP rs11682545 (**Table 3.3**) gave a combined p-value of 7.98 x $10^{-8}$, corrected p-value ($p$ = 0.0306). This SNP occurs downstream of the *TFCP2L1* gene on chromosome 2. The third SNP (rs7077757) was identified in **Table 3.4** (Reiman et al., 2007 as the primary dataset) with a combined p-value of 6.35 x $10^{-8}$, corrected p-value ($p$ = 0.0244). This occurs in intronic sequence of the *RBM20* gene on chromosome 10. No combined p-values of less than $10^{-8}$ were evident using the Carrasquillo et al., 2009 study as the primary comparator (**Table 3.5**).

**Table 3.2 Beecham et al., 2009 GWAS SNPs ($5 \times 10^{-5} < p < 5 \times 10^{-8}$) compared with Reiman et al., 2007, Li et al., 2008 and Carrasquillo et al., 2009.**

| Beecham et al., 2009 | | | | | | Reiman et al., 2007 | | Li et al., 2008 | | Carrasquillo et al., 2009 | | Combined p-value | Corrected p-value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SNP** | **CHR** | **BP** | **Gene** | **Position** | **p-value** | **SNP** | **p-value** | **SNP** | **p-value** | **SNP** | **p-value** | | |
| **rs9659092** | 1 | 50216176 | | | 4.54E-06 | rs12022125 | 4.04E-01 | rs12022125 | 1.48E-01 | - | | 2.71E-07 | |
| **rs3807031** | 6 | 30141863 | *PPP1R11* | Promoter | 1.16E-05 | - | | - | | rs3807031 | 4.94E-01 | 5.73E-06 | |
| **rs1415985** | 1 | 49703336 | | | 1.23E-05 | rs12022125 | 4.04E-01 | rs12022125 | 1.48E-01 | - | | 7.35E-07 | |
| **rs4926831** | 1 | 50062688 | | | 1.23E-05 | rs4926831 | 6.32E-01 | rs4926831 | 5.17E-01 | - | | 4.02E-06 | |
| **rs929156** | 6 | 30247678 | *TRIM15* | Exon 7 | 1.69E-05 | rs2844775 | 2.50E-01 | rs2844775 | 2.34E-01 | rs929156 | 8.87E-02 | 8.77E-08 | 4.67E-02 |
| **rs11583200** | 1 | 50332407 | | | 1.83E-05 | - | | - | | rs11583200 | 5.75E-01 | 1.05E-05 | |
| **rs11754661** | 6 | 151248771 | *MTHFD1L* | Intron | 2.01E-05 | - | | - | | rs11754661 | 6.27E-01 | 1.26E-05 | |
| **rs3746319** | 19 | 49304071 | *ZNF224* | Exon 6 | 2.96E-05 | - | | - | | rs3746319 | 9.85E-01 | 2.92E-05 | |
| **rs2180566** | 20 | 29482515 | *DEFB123* | Promoter | 3.80E-05 | - | | - | | rs2180566 | 4.75E-01 | 1.80E-05 | |
| **rs2061332** | 19 | 49305501 | *ZNF224* | D'stream | 3.93E-05 | rs2061332 | 1.49E-02 | rs2061332 | 7.22E-01 | rs2061332 | 8.70E-01 | 3.68E-07 | |
| **rs2681411** | 3 | 123268321 | *CD86* | Intron | 4.21E-05 | - | | - | | rs2681411 | 3.09E-01 | 1.30E-05 | |
| **rs2119067** | 2 | 165835529 | | | 4.38E-05 | - | | - | | rs2119067 | 1.58E-01 | 6.92E-06 | |
| **rs1402627** | 18 | 4123739 | | | 4.42E-05 | - | | - | | rs1402627 | 8.01E-01 | 3.54E-05 | |
| **rs659628** | 13 | 76361237 | *KCTD12* | Promoter | 4.46E-05 | rs659628 | 4.49E-01 | rs659628 | 1.00E+00 | - | | 2.00E-05 | |
| **rs9455973** | 6 | 168325855 | | | 4.47E-05 | rs9455973 | 9.79E-01 | rs9455973 | 5.99E-01 | rs9455973 | 6.27E-01 | 1.64E-05 | |
| **rs6059244** | 20 | 29474144 | | | 4.76E-05 | - | | - | | rs6059244 | 5.43E-01 | 2.59E-05 | |
| **rs11205641** | 1 | 49957662 | | | 8.41E-05 | rs11205641 | 3.40E-01 | rs11205641 | 4.79E-01 | rs11205641 | 3.85E-01 | 5.27E-06 | |

Each row represents a SNP with a p-value between $5 \times 10^{-5}$ and $5 \times 10^{-8}$. **The p-values are highlighted yellow if *corr-p* < 0.05 and they replicated across two or more studies.** Data from a perfect proxy SNP was used if data for the initial SNP was unavailable. If a perfect proxy was used the corresponding rs number is listed. The combined p-values across studies are as shown. The final column shows the corrected p-value adjusted as described (**Chapter 3 - section 3.3**).

**Table 3.3 Li et al., 2008 GWAS SNPs ($5 \times 10^{-5} < p < 5 \times 10^{-8}$) compared with Carrasquillo et al., 2009 and Reiman et al., 2007.**

| Li et al., 2008 | | | | | | Carrasquillo et al., 2009 | | Reiman et al., 2007 | | Combined p-value | Corrected p-value |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **SNP** | **CHR** | **BP** | **Gene** | **Position** | **p-value** | **SNP** | **p-value** | **SNP** | **p-value** | | |
| **rs4735627** | 8 | 100705091 | *VPS13B* | Intron | 3.51E-06 | rs4735627 | 8.73E-01 | rs4735627 | 7.66E-01 | 2.35E-06 | |
| **rs7336489** | 13 | 59171299 | *BC041395* | Intron | 5.38E-06 | - | | rs7336489 | 8.78E-01 | 4.72E-06 | |
| **rs370672** | 5 | 2501146 | | | 9.37E-06 | - | | rs370672 | 1.62E-01 | 1.52E-06 | |
| **rs4684083** | 3 | 163865 | | | 9.73E-06 | - | | rs4684083 | 6.72E-01 | 6.54E-06 | |
| **rs11682545** | 2 | 121662295 | *TFCP2L1* | Downstream | 1.29E-05 | - | | rs11682545 | 6.18E-03 | 7.98E-08 | 3.06E-02 |
| **rs6805482** | 3 | 25435600 | | | 1.78E-05 | - | | rs6805482 | 9.27E-01 | 1.65E-05 | |
| **rs11166407** | 1 | 100410296 | *LRRC39* | Intron | 2.00E-05 | - | | rs11166407 | 8.62E-02 | 1.72E-06 | |
| **rs8014810** | 14 | 35394781 | *BRMS1L* | Intron | 2.00E-05 | rs2274068 | 2.33E-01 | rs8014810 | 3.84E-01 | 1.79E-06 | |
| **rs541392** | 10 | 130941167 | | | 2.76E-05 | rs476628 | 3.66E-01 | rs541392 | 4.19E-01 | 4.23E-06 | |
| **rs13180602** | 5 | 160213616 | *ATP10B* | Upstream | 2.79E-05 | rs4559036 | 7.00E-02 | rs13180602 | 4.03E-01 | 7.87E-07 | |
| **rs11751998** | 6 | 11297073 | *NEDD9* | Intron | 3.42E-05 | rs10484448 | 4.86E-01 | - | | 1.66E-05 | |
| **rs6571727** | 14 | 35210859 | *GARNL1* | Intron | 3.49E-05 | rs6571727 | 2.14E-01 | rs10132580 | 7.61E-01 | 5.67E-06 | |
| **rs4483549** | 11 | 90595620 | | | 3.58E-05 | rs4483549 | 3.10E-01 | rs4483549 | 2.12E-01 | 2.35E-06 | |
| **rs1914516** | 2 | 215270178 | | | 3.61E-05 | - | | rs1914516 | 2.21E-01 | 7.98E-06 | |
| **rs4905898** | 14 | 99345451 | *EML1* | Intron | 3.61E-05 | rs10141863 | 7.74E-01 | rs4905897 | 5.44E-01 | 1.52E-05 | |
| **rs4687319** | 3 | 193526543 | *FGF12* | Intron | 4.60E-05 | - | | rs4687319 | 6.18E-01 | 2.84E-05 | |
| **rs16897530** | 8 | 100725659 | *VPS13B* | Intron | 4.74E-05 | - | | rs16897530 | 9.66E-01 | 4.58E-05 | |
| **rs4438299** | 16 | 60259838 | *CDH8* | Intron | 4.90E-05 | rs4438299 | 9.09E-01 | rs4438299 | 8.81E-01 | 3.93E-05 | |

Each row represents a SNP with a p-value between $5 \times 10^{-5}$ and $5 \times 10^{-8}$. **The p-values are highlighted yellow if *corr-p* < 0.05 and they replicated across two or more studies.** Data from a perfect proxy SNP was used if data for the initial SNP was unavailable. If a perfect proxy was used the corresponding rs number is listed. The same platform was used in the Reiman et al., 2007 and Li et al., 2008 studies. The combined p-values across studies are as shown. The final column shows the corrected p-value adjusted as described (**Chapter 3 - section 3.3**).

**Table 3.4 Reiman et al., 2007 GWAS SNPs ($5 \times 10^{-5} < p < 5 \times 10^{-8}$) compared with Li et al., 2008 and Carrasquillo et al., 2009.**

| Reiman et al., 2007 | | | | | | Li et al., 2008 | | Carrsquillo et al., 2009 | | Combined p-value | Corrected p-value |
|---|---|---|---|---|---|---|---|---|---|---|---|
| SNP | CHR | BP | Gene | Position | p-value | SNP | p-value | SNP | p-value | | |
| **rs10824310** | 10 | 53680643 | *PRKG1* | Intron | 6.03E-07 | rs10824310 | 3.06E-01 | - | | 1.84E-07 | |
| **rs17330779** | 7 | 107663071 | *NRCAM* | Intron | 8.80E-07 | rs17330779 | 5.31E-01 | - | | 4.67E-07 | |
| **rs6784615** | 3 | 52468315 | *NISCH* | Intron | 9.89E-07 | rs6784615 | 6.14E-01 | - | | 6.07E-07 | |
| **rs12162084** | 16 | 26553533 | | | 1.30E-06 | rs12162084 | 7.61E-01 | - | | 9.88E-07 | |
| **rs2517509** | 6 | 31138101 | | | 1.35E-06 | rs2517509 | 3.83E-01 | - | | 5.16E-07 | |
| **rs7077757** | 10 | 112527724 | *RBM20* | Intron | 1.52E-06 | rs7077757 | 4.18E-02 | - | | 6.35E-08 | 2.44E-02 |
| **rs249153** | 12 | 93837244 | | | 2.66E-06 | rs249153 | 8.25E-02 | rs249153 | 7.17E-01 | 1.58E-07 | |
| **rs10747758** | 12 | 54287453 | | | 3.03E-06 | rs10747758 | 2.45E-01 | - | | 7.42E-07 | |
| **rs11958566** | 5 | 117719226 | | | 4.16E-06 | rs11958566 | 6.16E-01 | - | | 2.56E-06 | |
| **rs17505622** | 13 | 101759124 | *FGF14,LOC283480* | Intron | 5.47E-06 | rs17505622 | 2.55E-01 | - | | 1.39E-06 | |
| **rs7079348** | 10 | 77742377 | *C10ORF11* | Intron | 8.70E-06 | rs7079348 | 3.85E-01 | - | | 3.35E-06 | |
| **rs475093** | 1 | 43383592 | *LOC440585* | Intron | 8.86E-06 | rs475093 | 7.10E-01 | - | | 6.29E-06 | |
| **rs11748700** | 5 | 15773106 | *FBXL7* | Intron | 1.09E-05 | rs11748700 | 2.40E-01 | - | | 2.62E-06 | |
| **rs7817227** | 8 | 27951747 | | | 1.47E-05 | rs7817227 | 4.99E-01 | - | | 7.35E-06 | |
| **rs17126808** | 8 | 18457737 | *PSD3* | Intron | 1.89E-05 | rs17126808 | 7.88E-01 | - | | 1.49E-05 | |
| **rs950922** | 1 | 21747977 | *ALPL* | Intron | 1.96E-05 | rs950922 | 3.45E-01 | - | | 6.74E-06 | |
| **rs16842422** | 1 | 196346167 | | | 1.99E-05 | rs16842422 | 7.48E-01 | - | | 1.49E-05 | |
| **rs4759173** | 12 | 54262230 | | | 1.99E-05 | rs4759173 | 4.52E-01 | rs10876820 | 4.45E-01 | 4.00E-06 | |
| **rs2122339** | 4 | 27290902 | | | 2.12E-05 | rs2122339 | 5.96E-01 | - | | 1.27E-05 | |
| **rs4394475** | 9 | 90496717 | | | 2.18E-05 | rs4394475 | 5.23E-01 | - | | 1.14E-05 | |
| **rs10783760** | 12 | 54260896 | | | 2.22E-05 | rs10783760 | 3.65E-01 | rs10876820 | 4.45E-01 | 3.62E-06 | |
| **rs13213247** | 6 | 81560955 | | | 2.29E-05 | rs13213247 | 5.73E-01 | rs16892136 | 4.17E-01 | 5.46E-06 | |
| **rs7097398** | 10 | 91782821 | | | 2.60E-05 | rs7097398 | 8.02E-01 | - | | 2.08E-05 | |
| **rs9982394** | 21 | 41191871 | | | 2.68E-05 | rs9982394 | 3.06E-01 | - | | 8.19E-06 | |
| **rs9934599** | 16 | 69220773 | *IL34* | Upstream | 2.68E-05 | - | | rs9934599 | 4.46E-01 | 1.20E-05 | |
| **rs7031458** | 9 | 84704086 | | | 2.74E-05 | rs7031458 | 2.16E-02 | - | | 5.91E-07 | |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **rs1923924** | 9 | 1581055 | | | 2.98E-05 | rs1923924 | 5.00E-01 | - | | 1.49E-05 |
| **rs249154** | 12 | 93848520 | | | 3.12E-05 | rs249154 | 1.14E-01 | rs249153 | 7.17E-01 | 2.55E-06 |
| **rs17151710** | 5 | 123739233 | | | 3.13E-05 | rs17151710 | 7.59E-01 | - | | 2.38E-05 |
| **rs17048904** | 4 | 118081372 | | | 3.50E-05 | rs17048904 | 1.00E+00 | - | | 3.50E-05 |
| **rs7134292** | 12 | 54260239 | | | 3.68E-05 | rs7134292 | 3.23E-01 | rs10876820 | 4.45E-01 | 5.30E-06 |
| **rs7585710** | 2 | 10819621 | *ATP6V1C2* | Intron | 3.76E-05 | rs7585710 | 1.00E+00 | - | | 3.76E-05 |
| **rs12044355** | 1 | 229901524 | *DISC1** | Intron | 3.93E-05 | rs12044355 | 9.07E-01 | - | | 2.92E-07***** |
| **rs6888935** | 5 | 117745419 | | | 3.93E-05 | rs6888935 | 9.96E-01 | - | | 3.92E-05 |
| **rs17586545** | 14 | 51101242 | *LOC645380,LOC651876* | Intron | 4.11E-05 | rs17586545 | 8.87E-01 | - | | 3.65E-05 |
| **rs1038891** | 11 | 40877959 | | | 4.48E-05 | rs1038891 | 4.84E-01 | - | | 2.17E-05 |
| **rs6094514** | 20 | 44993488 | *EYA2* | Intron | 4.49E-05 | rs6094514 | 3.40E-01 | rs11700355 | 5.60E-01 | 8.54E-06 |
| **rs10248657** | 7 | 112741449 | | | 4.56E-05 | rs10248657 | 8.88E-01 | - | | 4.05E-05 |

Each row represents a SNP with a p-value between 5 x $10^{-5}$ and 5 x $10^{-8}$. **The p-values are highlighted yellow if *corr-p* < 0.05 and they replicated across two or more studies.** Data from a perfect proxy SNP was used if data for the initial SNP was unavailable. If a perfect proxy was used the corresponding rs number is listed. The combined p-values across studies are as shown. The final column shows the corrected p-value adjusted as described. *DISC1* is starred to indicate that the combined p-value listed has included the data ($p$ = 8.20 x $10^{-3}$) from the Beecham et al., 2009 study (Chapter 3 - section 3.3).

**Table 3.5 Carrasquillo et al., 2009 GWAS SNPs ($5 \times 10^{-5} < p < 5 \times 10^{-8}$) compared with Li et al., 2008 and Reiman et al., 2007.**

| Carrasquillo et al., 2009 | | | | | | Li et al., 2008 | | Reiman et al., 2007 | | Combined p-value | Corrected p-value |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **SNP** | **CHR** | **BP** | **Gene** | **Position** | **p-value** | **SNP** | **p-value** | **SNP** | **p-value** | | |
| **rs2318144** | 8 | 58277297 | *ncRNA* | | 2.22E-06 | rs17194995 | 2.04E-01 | rs17194995 | 3.13E-01 | 1.42E-07 | |
| **rs1279795** | 23 | 123152101 | | | 5.02E-06 | rs1279795 | 8.42E-01 | - | | 4.22E-06 | |
| **rs3007421** | 1 | 6452776 | *PLEKHG5* | Intron | 6.54E-06 | rs3007421 | 6.51E-01 | rs3007421 | 4.68E-01 | 1.99E-06 | |
| **rs6546452** | 2 | 25834776 | | | 8.55E-06 | rs17680828 | 9.00E-01 | rs17680828 | 9.68E-01 | 7.45E-06 | |
| **rs7318037** | 13 | 81367146 | | | 1.15E-05 | rs4456389 | 9.82E-01 | rs4456389 | 2.39E-01 | 2.70E-06 | |
| **rs2118732** | 5 | 79419032 | | | 1.32E-05 | rs7736549 | 5.49E-01 | - | | 7.25E-06 | |
| **rs8039031** | 15 | 34954382 | *MEIS2* | Downstream | 2.26E-05 | rs8039031 | 5.04E-01 | rs8039031 | 9.92E-02 | 1.13E-06 | |
| **rs7245160** | 18 | 70417826 | *AK056288/LOC400657* | Upstream | 2.66E-05 | rs7245160 | 4.60E-01 | rs7245160 | 4.15E-01 | 5.08E-06 | |
| **rs856675** | 14 | 84405968 | | | 3.83E-05 | rs17737309 | 7.10E-01 | rs17737309 | 2.87E-01 | 7.81E-06 | |

Each row represents a SNP with a p-value between $5 \times 10^{-5}$ and $5 \times 10^{-8}$. **Data from a perfect proxy SNP was used if data for the initial SNP was unavailable.** If a perfect proxy was used the corresponding rs number is listed. The combined p-values across studies are as shown. The final column shows the corrected p-value adjusted as described. No SNPs replicated across studies using the Carrasquillo et al., 2009 GWAS as the primary dataset (**Chapter 3 - section 3.3**).

Meta-analysis of odds ratio for candidate SNPs

A random-effects meta-analysis (also known as DerSimonian-Laird test) of the allelic odds ratios was performed for the three SNPs identified as mentioned (DerSimonian and Laird, 1986). The *TRIM15* SNP (rs929156) gave odds ratios in the same direction (causative, **Table 3.6**) across three studies and random effect meta-analysis gave an odds ratio of 1.1 (95% CI 1.0-1.2; $p = 0.03$). *RBM20* ($p = 0.95$) and *TFCP2L1* ($p = 0.74$) SNPs were not significant following meta-analysis.

Gene-centric analysis of *TRIM15*

A gene-centric analysis of *TRIM15* was undertaken (**Figure 3.1**) to explore the genetic architecture in more detail. The histogram shows the SNPs present in three different GWAS (Carrasquillo et al., 2009; Li et al., 2008; Reiman et al., 2007), their associated p-values together with their degree of linkage.

**Table 3.6 Comparison of odds ratios across GWAS for selected SNPs.** If the SNP was not present in a GWAS, odds ratio of a perfect proxy ($r^2 = 1$) was used. The proxy SNP ID is shown underneath the corresponding odds ratios. The data shown is for the allelic association model. The 95% confidence interval (CI) for odds ratios are shown in brackets. The results from random effects meta-analysis of these odds ratios are given in the final column.

| Gene | OR (95% CI) | | | |
|---|---|---|---|---|
| | **Reiman et al., 2007** | **Carrasquillo et al., 2009** | **Li et al., 2008** | **Random effects Meta-analysis of OR's** |
| ***TRIM15*** **(rs929156)** | 1.1 (0.9-1.3) (rs2844775) | 1.1 (1.0-1.3) | 1.1 (0.9-1.3) (rs2844775) | 1.1 (1.0-1.2) *p* = 0.03 |
| ***TFCP2L1*** **(rs11682545)** | 0.8 (0.7-0.9) | - | 1.3 (1.1-1.5) | 1.0 (0.7-1.6) *p* = 0.95 |
| ***RBM20*** **(rs7077757)** | 0.6 (0.5-0.8) | - | 1.3 (1.0-1.5) | 0.9 (0.5-1.7) *p* = 0.74 |

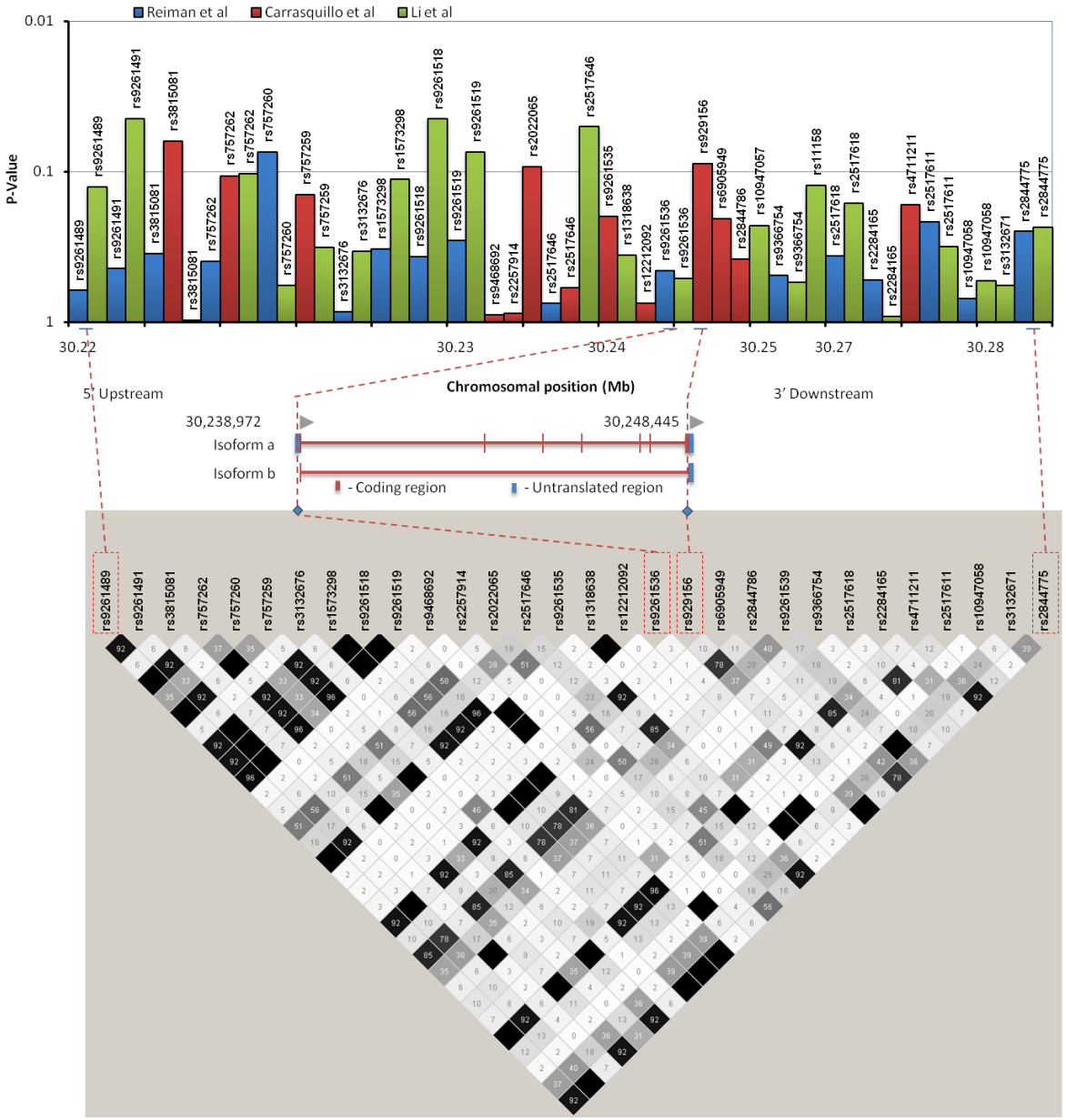



**Figure 3.1 Schematic overview of the *TRIM15* gene and the LD plot for this region.** The histogram depicts all GWAS SNPs in *TRIM15*, their p-values and IDs are shown at the top of the figure. These studies are colour-coded as indicated at the top of the figure. The two *TRIM15* isoforms and their chromosomal positions are as depicted in HapMap (release 23). The LD plot is for the GWAS variants (Haploview 4.0, $r^2$ values with $r^2$ colour scheme). The positions of SNPs with respect to the gene are indicated on the LD plot. The SNPs at the boundaries of this LD block are also shown. LD values are represented by different colours (black - strong LD, grey - moderate LD, and white - no evidence of LD).

## 3.5 Discussion

The *APOE* region on chromosome 19 was confirmed as a genetic-risk factor in LOAD by all four GWAS with SNP p-values ranging from $10^{-36}$ to $10^{-44}$. Apart from those in LD with the *APOE* locus, there were no other SNPs across the four GWAS with p-values less than $10^{-8}$. Genes with suggestive significance ($10^{-5} < p < 10^{-8}$) across different GWAS may infer a genuine LOAD candidate.

Approximately 700 genes and 3000 polymorphisms have been assessed as genetic risk factors in association with AD (http://www.alzgene.org/) as of October 2011 (Bertram et al., 2007). Except for the *APOE* gene, most of the genes have conflicting reports with regard to their associations. However, each of the studies often uses different populations with varying male and female percentages, as well as differing age ranges and sample sizes. Results are therefore not always directly comparable between different studies (Bertram et al., 2007). The study approach used here may help identify potential LOAD candidate genes whose signals replicate across studies.

GWAS association analysis uses very stringent significance levels to avoid the large number of false positives potentially arising from the confounding effects of population substructure and testing of a very large number of SNPs simultaneously (Bodmer and Bonilla, 2008). For example, in a GWAS using 500,000 independent markers, 25,000 can be expected to show a nominal p-value ($p < 5 \times 10^{-2}$) by chance alone and five out of this 25,000 may be significant with p-values ($p < 1 \times 10^{-5}$). A widely accepted p-value ($p < 5 \times 10^{-8}$) is used to indicate a genuine disease association in GWAS (Bertram and Tanzi, 2008). However, the SNPs on different chip platforms are often not independent. Many SNPs are in LD with other SNPs, potentially reducing the number of independent markers available for analysis. Secondly, the genotyping rate never reaches 100%, and after quality control, significant numbers of SNPs are excluded from study (**Table 3.1**). This suggests that

a p-value of $< 5 \times 10^{-8}$ may in some instances be too stringent and SNPs with p-values between $10^{-5}$ and $10^{-8}$ might well harbour genuine associations.

The potential role of *TRIM15*

TRIM15 is a member of the tripartite motif (TRIM) family. The TRIM motif includes three zinc-binding domains, a RING, a B-box type 1 and B-box type 2, and a coiled-coil region. The protein is localized to the cytoplasm. Two isoforms have been identified and described, however their biological functions have not as yet been identified. *TRIM15* is ubiquitously expressed in various tissues. However, the biological role of TRIM15 has not yet been determined (Shiina et al., 2006).

SNP rs929156 in *TRIM15* is located in an exon in one of the two *TRIM15* transcripts. It changes the amino acid from a small, polar Serine to a medium-sized, polar Asparagine. It is located in a B30.2 SPRY like domain (position: 276-465 amino acids). The B30.2-like domain is a conserved domain found in nuclear and cytoplasmic proteins, as well as transmembrane and secreted proteins. The B30.2-like domain may also be associated with a zinc-binding B-box domain in the N-terminal (Henry et al., 1998). The SPRY domain is proposed to be a protein interacting module, which recognizes and interacts with specific individual partner proteins (Woo et al., 2006). The potential effects of this SNP on protein structure require further investigation.

The only other *TRIM15* SNP in these GWAS rs9261536 is located in the 5' untranslated region (UTR), which may harbour potential regulatory elements (i.e. a promoter region or a binding site for an associated transcription factor - **Figure 3.1**). Possible linkage has been observed between this *TRIM15* SNP and SNPs in Human leukocyte antigen A (*HLA-A*) with $r^2$ value 0.77 (**Figure 3.2**).

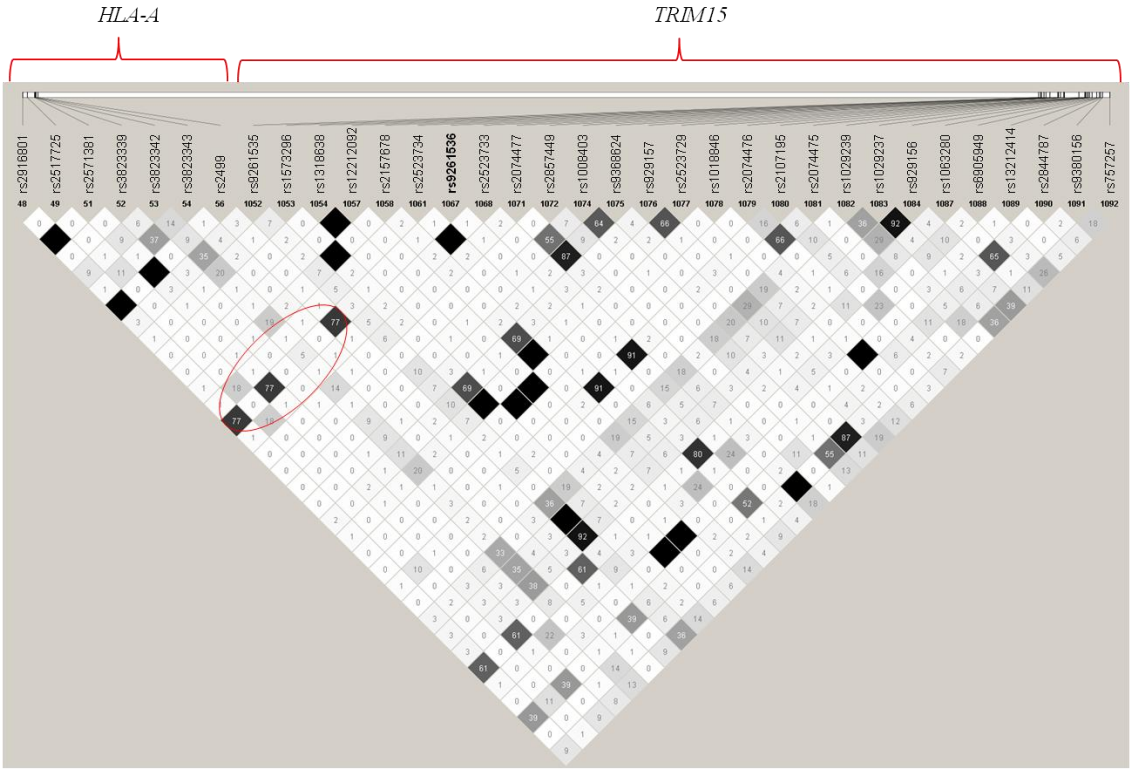



**Figure 3.2 Illustration of LD between *TRIM15* and *HLA-A* genes.** The red ellipse highlights the linkage between *TRIM15* SNP rs9261536 (shown in bold) and three known SNPs (rs2916801, rs2571381 and rs2499) in *HLA-A* with $r^2$ value = 0.77. The LD plot was generated using HapMap data (CEU population release 23) and the program Haploview version 4.0 (**Methods 2.3.5**).

*TRIM15* is surrounded by a number of *HLA* genes which are associated with the human immune system. This group of *HLA* genes encode cell-surface antigen-presenting proteins, which are essential elements in human immune responses. *HLA-A* is necessary for immune recognition and apoptosis, and mutations in *HLA-A* have been reported as risk factors for various cancers (Hu et al., 2009). Ma et al., 2008 showed that mutations in *HLA-A* are associated with earlier age at onset of AD (2.4 years, $p = 0.03$) for non-carriers of *APOE* ε4 (Ma et al., 2008).

The significance of identified SNPs

In this study, an approach was described to detect replicating signals across different GWAS platforms in an effort to identify LOAD candidate genes that have failed to reach genome-wide significance previously. Using the data from the four studies listed has generally failed to produce any convincing replicating signals with the possible exception of the *TRIM15* gene which contains the only SNP (rs929156) whose combined p-value ($p = 8.77 \times 10^{-8}$) survives multiple testing correction (*corr-p* = 0.0467) and where the meta-analysis of odds ratios is also tentatively significant (OR = 1.1, 95% CI 1.0-1.2, $p = 0.03$) with no evidence of between-study heterogeneity (Breslow-Day $p = 0.90$) (**Figure 3.3**).

The remaining two SNPs that had p-values of $10^{-8}$ failed the meta-analysis of odds ratios because their effects were discordant between studies (**Table 3.6**).

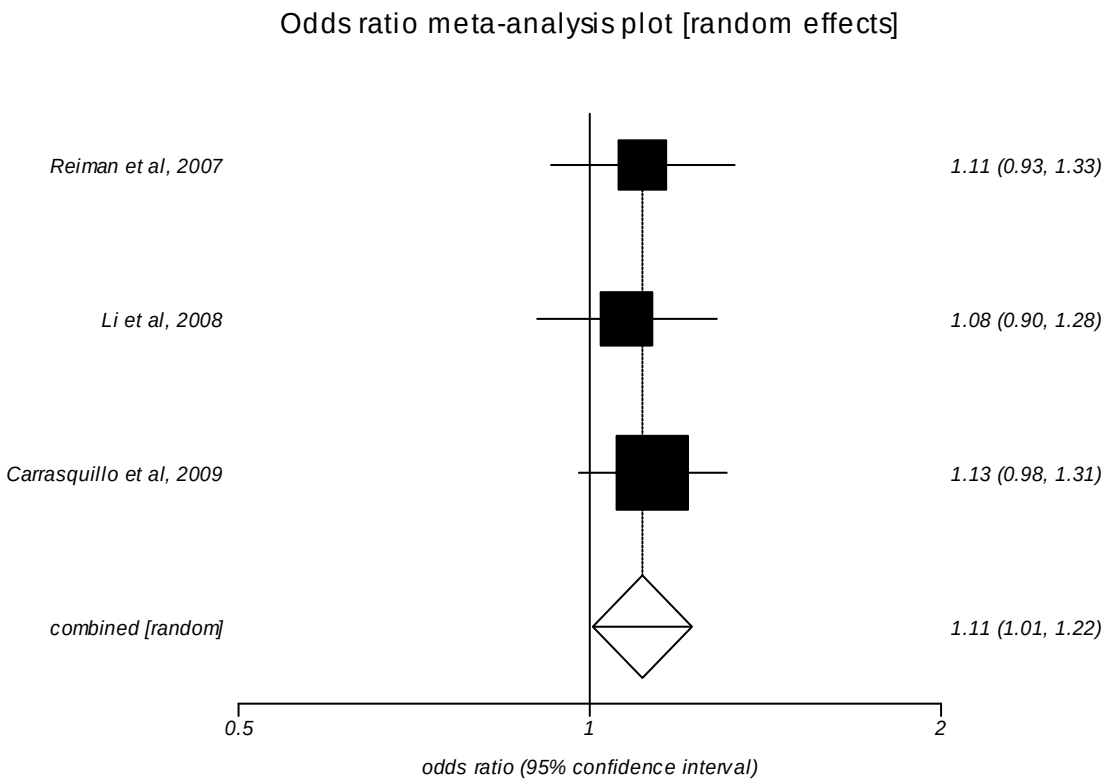



**Figure 3.3 Forest plot depicting effects of SNP rs929156 from different GWAS.** The effect of SNP rs929156 for each individual GWAS together with the meta-analysis results (combined [random]) are as indicated. The size of the squares represents the weight (Mantel-Haenzel weight) of the corresponding study in the meta-analysis. The line either side of the square represents 95% confidence intervals for the odds ratio. Confidence intervals of pooled estimates are displayed as a horizontal line through the diamond. The dashed vertical line (linking squares and the diamond) represents the odds ratio of the meta-analysis. Odds ratio and 95% confidence interval of each study and combined are documented to the right of the forest plot.

In the UK LOAD GWAS paper (Harold et al., 2009) the *TRIM15* SNP, rs929156, was shown to be modestly associated with AD ($p$ = 0.049). Adding this data results in a Fisher's combined p-value of $4.30 \times 10^{-9}$ strengthening the evidence of association for this SNP. The odds ratio from the UK GWAS (OR = 1.07) was also compatible with the odds ratio observed in the random effect meta-analysis in this study (OR = 1.11).

An issue which is evident in this study is the difficulty that exists when trying to compare data across different chip platforms where the SNP complement differs. Surprisingly few perfect proxies available resulted in a significant loss of data and a reduction of power to detect new signals.

Technology continues to advance; the latest Illumina GWAS chip is capable of genotyping ~5 million SNPs from the international HapMap project, as well as SNPs identified by the 1000 genomes project with MAF above 1%. The approach described may prove to be useful when larger datasets (generated using these chips) are analysed.

## 3.6 Bioinformatics Application Note

As the number of publically available GWAS datasets continues to grow, bioinformatic tools which enable routine manipulation of data are becoming increasingly useful. In addition, whole genome meta-analysis is labour intensive without suitable bioinformatics software.

An 'LD aware' bioinformatics application was developed enabling efficient comparison of SNPs effects across multiple GWAS datasets using Fisher's combined probability test from PLINK (v1.06) 'LD clumped' output (**Appendix 8.4.3**).

Implementation

PLINK (v1.06) provides an 'ld-clump' analysis which allows automatic calculation of 'clumps' (blocks of SNPs in LD) across genotyping chip platforms. The software developed uses the output file from this 'LD clump' analysis and performs the downstream meta-analysis of SNPs in each clump (taking one SNP/proxy ($r^2$=1) from each study and combining their p-values).

The application consists of two files 'meta_analysis.pl', 'modules.pm', where 'meta_analysis.pl' is an executable file when PERL language is installed.

'meta_analysis.pl' file can be edited using a conventional text file editor, and allows users to define two parameters; i) the location and the filename of the input ('ld_clump') file generated in PLINK, and ii) the type of the study - case/control (CC) analysis or quantitative trait (QT) analysis. Fields requiring modification were annotated in the 'meta_analysis.pl' file. The default input is 'case/control' analysis. Failure to adjust the parameter for the correct type of analysis will generate false output in the results file.

The program was designed to handle an unlimited number of GWAS datasets and unlimited SNPs. However, currently the application has only been validated using up to ten modelled datasets.

Issues and problem-solving

It is worth noticing that, when generating the 'clump' files, PLINK requires a reference GWAS dataset (e.g. HapMap data) to calculate LD values between two SNPs. Given that LD values vary between ethnic populations, it is imperative that the reference dataset and the GWAS datasets are from the same population thus avoiding stratification issues.

It is common that a SNP in one GWAS has multiple perfect SNP proxies ($r^2$= 1) in another independent GWAS. In this situation, although it is considered appropriate to use any pair of SNPs to perform meta-analysis, the application decides which proxy to use based on which proxy is closest to the index SNP (as in physical distance). The distance between the SNP/proxy and the index SNP is annotated in the results file. It should be noted that if an index SNP is unique to one study and does not have a perfect proxy in any other studies, no meta-analysis results will be displayed for this SNP.

This application only uses perfect proxies ($r^2$ = 1). This is a limitation of the software as using imperfect proxies ($r^2$ < 1) will increase the number of comparable SNPs between studies. Currently, there is no weighting algorithm implemented in the program therefore any SNP p-value inferred from a proxy with $r^2$ < 1 will be inaccurately treated as a perfect match.

To use imperfect proxies, simply alter the (--clump-r2) parameter in PLINK, and run the application as usual. This may indeed be a valuable approach to increase coverage, analogous to imputation, but until the output is weighted accordingly, the results will have to be interpreted with caution.

A ‘flipper’ refers to a SNP or SNP proxy in one dataset which has the opposite effect to that observed in the original study. This application compares the OR (case/control analysis) or regression coefficient (from quantitative trait analysis) of all SNP pairs from different studies, and annotates according to the following rules;

i) ‘YES – flipper’ – ORs are in opposite directions in two (or more) GWAS studies, irrespective of missing OR data in additional studies, ii) ‘NO - non-flipper’ - all SNPs OR in the same direction and OR data is present for all studies, and iii) ‘NA - not applicable’ - there is either no OR data, or datasets are missing OR data making it inappropriate to call a ‘non-flipper’. In studies with missing OR or BETA, the field has to be encoded as ‘-9’ in the ‘.assoc’ file for subsequent ‘ld-clumping’ analysis in PLINK.

The application automatically recognizes the number of GWAS from PLINK ‘ld-clump’ output files and tabulates the results accordingly. Although the application was designed for GWAS meta-analysis, the user can perform analysis on much smaller datasets.

This approach is advised to be used prior to more formal meta-analysis. It is essential that any potential finding that emerges using the application is verified by further investigation in a rigorous manner. Adjusting the genotypic data for covariates and taking into account heterogeneity between studies/samples will verify if observations involving both ‘flipping’ and ‘non-flipping’ alleles are likely to be genuine and worthy of downstream study.

An example

Two late-onset Alzheimer's disease (LOAD) GWAS datasets - Carrasquillo et al., 2009 and Reiman et al., 2007 and the 'top hits' tabulated in Beecham et al., 2009 were used to test the performance of the application. The sample sizes and genotyping platforms for the three studies are - Carrasquillo et al., 2009, 799 LOAD cases and 1199 controls on Illumina 300 chip, Reiman et al., 2007, 859 LOAD cases and 552 controls on Affymetrix 500K chip and Beecham et al., 2009, 492 LOAD cases and 496 controls on Illumina 550 chip. The total sample size of all three GWAS is 4,397 (2,150 LOAD cases and 2,247 controls). No apolipoprotein E (*APOE*) related SNPs were listed in Beecham et al., 2009 'top hits'. These samples were estimated to provide over 93% power to detect an association with a common SNP (MAF > 10% and OR > 1.3). This estimation has to be treated with caution as it is based on a number of assumptions (such as effect size and mode of inheritance), and gene-environment interaction (GxE) has not been taken into account.

Before using the software, a number of PLINK analyses were undertaken.

1) Subject-level genotype data was obtained from Carrasquillo et al., 2009 and Reiman et al., 2007. The files were converted into PLINK format where necessary, and the SNP identifiers were converted into dbSNP rs number using an 'in house' program written in PERL (**Methods 2.3.2**). The GWAS output was generated using the PLINK '--assoc' command.

2) As Beecham et al., 2009 GWAS data was not available, a file called 'Beecham.assoc' was manually generated conforming to the format of a PLINK '.assoc' file (**Methods 2.3.1**). Three compulsory columns are required in the '.assoc' file with the headers 'SNP', 'OR' and 'P'. All other information such as 'BP', 'CHR', 'A1' in the standard '.assoc' file are not required, and can simply be ignored. As OR

data was not included in the Beecham et al., 2009 data, these values were set to '-9' in the 'OR' column.

3) The 'ld-clump' analysis was performed using PLINK (v1.06) using the three files generated 'Carrasquillo.assoc', 'Reiman.assoc' and 'Beecham.assoc'. The filtered version of HapMap data (CEU population, release 23) in PLINK binary format (BED, BIM and FAM) was downloaded from the PLINK website 'http://pngu.mgh.harvard.edu/~purcell/plink/res.shtml' (Purcell et al., 2007). The HapMap data contains 2.3 million SNPs.

The 'ld-clump' analysis was performed using the following PLINK command:

**plink --bfile HapMapCEU23**
**--clump Carrasquillo.assoc,Reiman.assoc,Beecham.assoc**
**--clump-verbose**
**--clump-annotate OR**
**--clump-p1 1**
**--clump-p2 1**
**--clump-r2 0.99**
**--out ld_clump**
**--noweb**

The PLINK method reads '--bfile' (the HapMap data in PLINK format) and 'clumps' the three datasets based on HapMap LD $r^2$ values. '--clump-verbose' instructs PLINK to generate a detailed report of SNPs in each clump. The output of ORs was specified using '--clump-annotate OR' ('--clump-annotate BETA' was used for quantitative trait analysis). All SNPs were used to perform the 'ld-clump' analysis irrespective of p-values ('--clump-p1 1' and '--clump-p2 1'). '--clump-r2 0.99' ensures that only SNPs which are perfect proxies ($r^2 > 0.99$) are clumped ('--clump-r2 1' does not work).

A single output file 'ld_clump.clumped' was generated using '--out ld_clump'. All of these commands listed are compulsory to the subsequent analysis except '--clump-p1' and '--clump-p2' which allow the user to adjust p-value threshold.

4) The 'meta_analysis.pl' file was edited using a text file editor to ensure it contains the correct PLINK 'ld-clump' output filename and correct type of analysis as described earlier.

5) The application was executed and a results file named 'results.txt' was generated automatically.

The top 6 results in **Table 3.7** illustrate the utility of this application. The top 2 SNPs are in LD with the *APOE* locus and demonstrate highly significant p-values as expected (rs1114832 Fisher's method p-value $1.09 \times 10^{-9}$ and rs10402271 Fisher's method p-value $1.13 \times 10^{-7}$); the first SNP exceeded genome-wide significance ($p = 1.67 \times 10^{-7}$) after correcting for the number of independent tests (**Methods 2.3.2**) and the second SNP approached this value.

Sub-significant hits may prove to be genuine when more datasets are included. SNP rs2318144 is located 200kb upstream of the inositol monophosphatase domain containing 1 gene, *IMPAD1*; as of October 2011 this gene has yet to figure as an AD candidate in the AlzGene forum (Bertram et al., 2007). rs3746319 ($p = 4.34 \times 10^{-7}$) was found to be located in an exon of the zinc finger protein (*ZNF224*). Although this SNP is in vicinity to the *APOE* region, the effect has been suggested to be independent of *APOE* status (Beecham et al., 2009). rs11205641 demonstrates a dichotomy of effect i.e. it's OR is not comparable between datasets (shows opposing effects) and is thus indicated as a 'flipper'. rs468345 ($p = 1.15 \times 10^{-6}$) is located ~120kb upstream of Amyloid-beta precursor protein (*APP*) gene.

**Table 3.7 Results from Meta-analysis of Carrasquillo et al., 2009, Reiman et al., 2007 and Beecham et al., 2009.** The table shows SNPs with Fisher's combined probability test p-value less than 1 x $10^{-5}$. F1, F2 and F3 refers to the studies which have been inputted to perform the meta-analysis (1, 2 and 3 refers to Carrasquillo et al., 2009, Reiman et al., 2007 and Beecham et al., 2009, respectively). KB1 and RSQ1 refer to the distance and LD between the index SNP and PROXY1. The same rule applies to KB2 and RSQ2. The suffix (1, 2 or 3) on column headers Pvalue indicated p-value in each study individually as listed. ClumpNo - index number ranked based on descending p-value in each study, CHR - chromosome number, FISHER - Fisher's combined probability test p-value and FLIPPER - indicates whether the SNP is a 'flipper'.

| ClumpNo | SNP | F1 | CHR | PROXY1 | F2 | KB1 | RSQ1 | PROXY2 | F3 | KB2 | RSQ2 | FISHER | FLIPPER | Pvalue1 | Pvalue2 | Pvalue3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | rs1114832 | 1 | 19 | rs1114832 | 2 | 0 | 1 | - | - | - | - | 1.09E-09 | NO | 1.37E-06 | 0.000799 | - |
| 4 | rs10402271 | 1 | 19 | rs10402271 | 2 | 0 | 1 | - | - | - | - | 1.13E-07 | NO | 4.54E-06 | 0.0248 | - |
| 2 | rs2318144 | 1 | 8 | rs6982990 | 2 | 0.532 | 1 | - | - | - | - | 3.57E-07 | NO | 2.22E-06 | 0.161 | - |
| 20 | rs3746319 | 3 | 19 | rs3746319 | 1 | 0 | 1 | rs2061332 | 2 | 1.43 | 1 | 4.34E-07 | NA | 0.985 | 0.0149 | 2.96E-05 |
| 7 | rs11205641 | 3 | 1 | rs11205641 | 1 | 0 | 1 | rs11205641 | 2 | 0 | 1 | 1.10E-06 | YES | 0.385 | 0.34 | 8.41E-06 |
| 107 | rs468345 | 2 | 21 | rs468345 | 1 | 0 | 1 | - | - | - | - | 1.15E-06 | NO | 0.00353 | 0.000326 | - |
| 42 | rs7679738 | 1 | 4 | rs510115 | 2 | 2.67 | 1 | - | - | - | - | 1.63E-06 | NO | 9.72E-05 | 0.0168 | - |
| 5 | rs9659092 | 3 | 1 | rs12022125 | 2 | -83.5 | 1 | - | - | - | - | 1.83E-06 | NA | - | 0.404 | 4.54E-06 |
| 3 | rs249153 | 2 | 12 | rs249153 | 1 | 0 | 1 | - | - | - | - | 1.91E-06 | NO | 0.717 | 2.66E-06 | - |
| 68 | rs9474661 | 2 | 6 | rs4486000 | 1 | -2.44 | 1 | - | - | - | - | 2.19E-06 | NO | 0.0142 | 0.000154 | - |
| 16 | rs8039031 | 1 | 15 | rs8039031 | 2 | 0 | 1 | - | - | - | - | 2.24E-06 | NO | 2.26E-05 | 0.0992 | - |
| 86 | rs4693305 | 2 | 4 | rs4693305 | 1 | 0 | 1 | - | - | - | - | 2.64E-06 | NO | 0.0119 | 0.000222 | - |
| 10 | rs7318037 | 1 | 13 | rs4456389 | 2 | 11.5 | 1 | - | - | - | - | 2.75E-06 | YES | 1.15E-05 | 0.239 | - |
| 6 | rs3007421 | 1 | 1 | rs3007421 | 2 | 0 | 1 | - | - | - | - | 3.06E-06 | NO | 6.54E-06 | 0.468 | - |
| 74 | rs385771 | 1 | 5 | rs385771 | 2 | 0 | 1 | - | - | - | - | 3.78E-06 | YES | 0.000163 | 0.0232 | - |
| 76 | rs10501120 | 1 | 11 | rs10501120 | 2 | 0 | 1 | - | - | - | - | 4.00E-06 | NO | 0.000171 | 0.0234 | - |
| 9 | rs4313171 | 2 | 8 | rs359819 | 1 | -169 | 1 | - | - | - | - | 5.01E-06 | NO | 0.501 | 1.00E-05 | - |
| 144 | rs6695249 | 1 | 1 | rs17113051 | 2 | 4.38 | 1 | - | - | - | - | 5.40E-06 | NO | 0.00045 | 0.012 | - |
| 11 | rs3807031 | 3 | 6 | rs3807031 | 1 | 0 | 1 | - | - | - | - | 5.73E-06 | NA | 0.494 | - | 1.16E-05 |
| 25 | rs2119067 | 3 | 2 | rs2119067 | 1 | 0 | 1 | - | - | - | - | 6.92E-06 | NA | 0.158 | - | 4.38E-05 |
| 35 | rs11033712 | 2 | 11 | rs12271660 | 1 | 26.9 | 1 | - | - | - | - | 7.85E-06 | YES | 0.127 | 6.18E-05 | - |
| 8 | rs6546452 | 1 | 2 | rs17680828 | 2 | 9.37 | 1 | - | - | - | - | 8.28E-06 | NO | 8.55E-06 | 0.968 | - |
| 14 | rs4759173 | 2 | 12 | rs10876820 | 1 | -22.7 | 1 | - | - | - | - | 8.86E-06 | NO | 0.445 | 1.99E-05 | - |
| 22 | rs2387100 | 3 | 13 | rs2387100 | 1 | 0 | 1 | rs9551404 | 2 | -12.5 | 1 | 9.40E-06 | NA | 0.644 | 0.382 | 3.82E-05 |
| 31 | rs7537266 | 2 | 1 | rs7537266 | 1 | 0 | 1 | - | - | - | - | 9.52E-06 | NO | 0.186 | 5.12E-05 | - |
| 17 | rs13213247 | 2 | 6 | rs16892136 | 1 | -115 | 1 | - | - | - | - | 9.55E-06 | NO | 0.417 | 2.29E-05 | - |
| 79 | rs4904864 | 1 | 14 | rs10484035 | 2 | 14 | 1 | - | - | - | - | 9.58E-06 | YES | 0.000186 | 0.0515 | - |

## 3.7 Conclusion

An approach has been described in this chapter to detect replicating signals across different GWAS in an effort to identify LOAD candidate genes that have failed to reach genome-wide significance previously.

Using data from four studies (**Table 3.1**) revealed a single SNP rs929156 (located in exon 7 of *TRIM15* gene) whose combined p-value ($p = 8.77 \times 10^{-8}$) withstands multiple testing correction ($p = 0.0467$) using perfect proxies and where the meta-analysis of odds ratios is also significant (OR 1.1, 95% CI 1.0-1.2, $p = 0.03$). Using imperfect proxies ($r^2 < 1$) (i.e. relaxing the condition of perfect LD) in this approach would likely further reduce the number of independent test, thereby lowering the genome-wide significant threshold, and more SNPs with suggestive p-values may reach genome wide significance. However, such results would need to be interpreted with caution, as lowering the LD $r^2$ value is likely to introduce errors. The relationship between the LD $r^2$ value and the amount of noise introduced by using imperfect proxies requires further investigation.

The next chapter (**Chapter 4**) describes a study investigating if the gene encompassing this SNP harbours multiple rare variants that may be associated with the disease using ABI SOLiD® next generation sequencing.

An important argument for GWAS is that the genes in which common variants are found, or genes nearby, may well contain functional rare variants; these may have high enough penetrance to be considered as candidates for possible preventive screening strategies in the future (Bodmer and Bonilla, 2008).

# Chapter 4: Next generation sequencing (NGS) of tripartite motif-containing 15 (*TRIM15*) gene using pooled DNA samples

## 4.1 Introduction

Comparing DNA sequence of LOAD patients against those free of disease symptoms allows identification of underlying genetic loci, which has been estimated to account for up to 76% of the disease risk (Gatz et al., 2006). However, in order to differentiate true signals from background noise and achieve statistical significance, sequencing of a large number of individuals is essential.

DNA sequencing technology has evolved rapidly over the last few years, with the advent of NGS enabling both reliable and economically affordable sequencing of large-scale DNA sequence (such as whole exome and whole genome sequencing) in a large number of individuals (Metzker, 2010). This enhanced capability of sequencing provides unprecedented opportunity to address major biological questions, such as the search for genetic heritability of LOAD attributable to rare variants.

Existing GWAS is not designed for capturing rare variants with allele frequency less than 5%, and insufficient coverage meant that some of the common variants are also not accounted for (Cirulli and Goldstein, 2010). NGS of targeted genomic regions using pooled DNA samples is capable of testing genetic associations of all variants within target regions provided there is sufficient power.

Sanger sequencing

Sequencing of DNA and RNA has solely relied on Sanger sequencing technology for almost 30 years since it was first developed by Frederick Sanger in 1977 (Sanger et al., 1977). It was the key technology used in identification of SNPs, copy number

variations (CNVs) and structural variants (such as insertions and deletions) prior to the advent of the NGS technology.

Automated Sanger sequencing was used for the Human Genome Project, which produced the first sequence data of the complete human genome encompassing ~3 billion nucleotide bases (International Human Genome Sequencing Consortium, 2004). The project was accomplished in 2003 through intensive worldwide collaboration and cost ~$3 billion.

Sanger sequencing is based on using fluorescently labelled dideoxynucleotide triphosphate (ddNTPs) as DNA chain terminators. As ddNTPs lack the 3'-OH group required to form the next phosphor-diester bond, addition of ddNTPs terminates the chain elongation reaction (facilitated by *Taq* polymerase and dNTPs). The concentrations of ddNTPs are much lower than dNTPs, thus they would occasionally incorporate into a growing DNA chain at random, and stop further synthesis.

The final product of Sanger sequencing is a mixture of various sizes of nucleotide fragments. Capillary electrophoresis of these short nucleotide fragments enables the DNA sequence to be recorded, facilitated by a laser and a detector (**Figure 4.1**). The Sanger sequencing method generates the DNA sequence in a format known as an electropherogram.

Sanger's method is limited by its throughput of sequencing of only ~1kb DNA template per experiment run. Additionally, the first ~50 bases of reads are often found to be of poor quality, a result likely to be due to the presence of residual unincorporated fluorescently labelled ddNTPs (Wallis and Morrell, 2010).

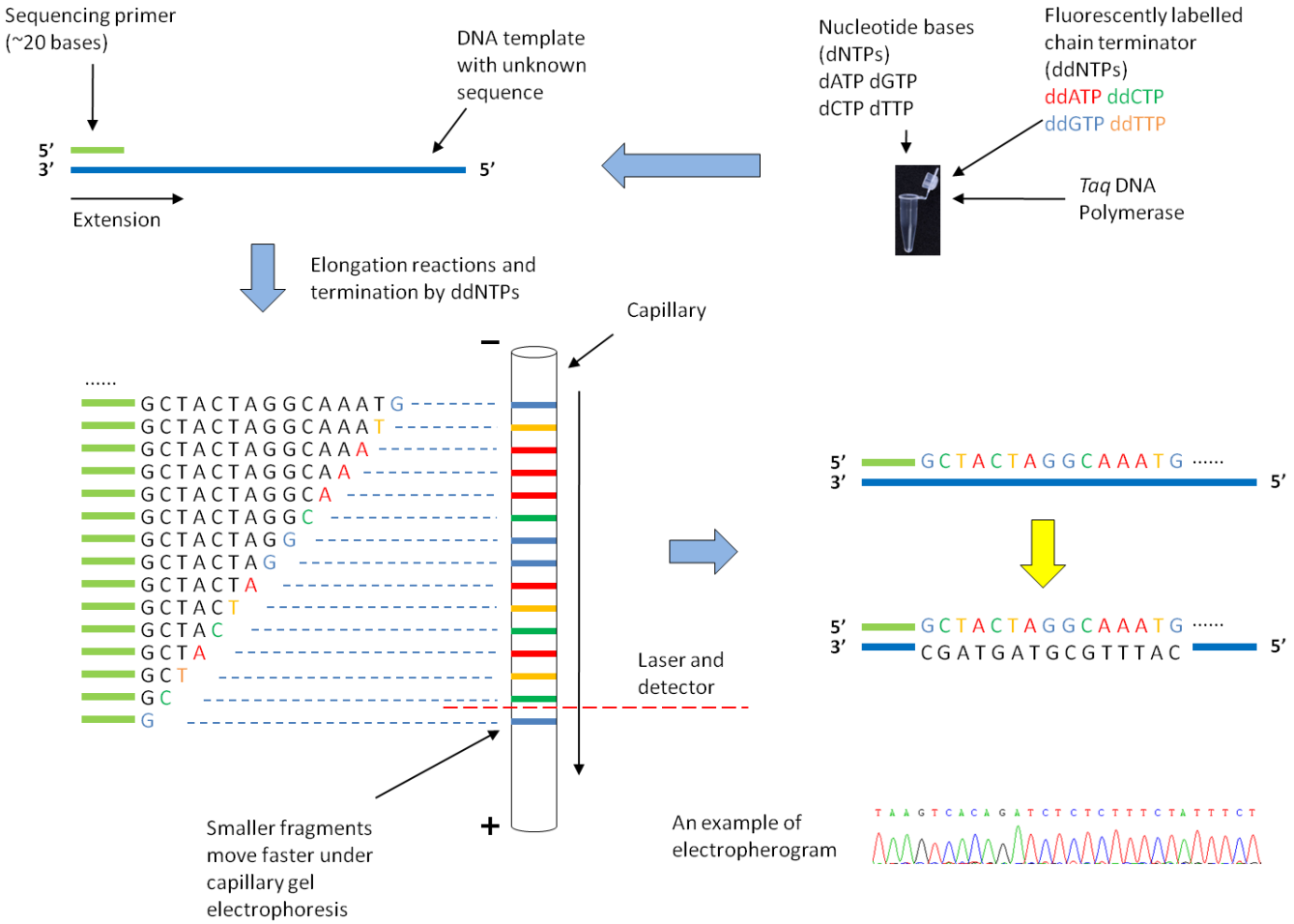



**Figure 4.1 Schematic diagram of Sanger sequencing.** In Sanger sequencing, *Taq* polymerase, dNTPs and fluorescently labelled ddNTPs, sequencing primer and DNA template are added together. A range of different lengths of nucleotide fragments are generated, which are subjected to capillary gel electrophoresis. The fluorescent signals emitted by the labelled the ddNTPs, each corresponding to the point at which the chain growth is terminated, are detected using a laser and a detector. The sequence results are in the form of an electropherogram.

Next generation sequencing

The major advantage of next generation sequencing technology is the ability of generating enormous amounts of sequencing data quickly and at substantially lowered expenditure. Millions of short reads, each of 35-150 base pair in length, can be generated in a single experiment run using the technology (Mardis, 2011; Metzker, 2010).

There has been fierce competitions in the field of next generation sequencing technology, leading to rapid evolution of technology with respect to its accuracy, throughput and speed (Metzker, 2010).

Next generation sequencing platforms

The currently available next generation sequencing platforms are Roche 454, ABI SOLiD and Illumina platforms (Genome Analyzer IIx, HiSeq 1000/2000 and HiScanSQ). Each of these platforms uses distinct chemistry, both Roche 454 and Illumina platforms use a method known as ‘sequencing by synthesis’, compared to ‘sequencing by ligation’ used by ABI SOLiD®.

Roche 454, also known as pyrosequencing, detects each polymerase catalysed nucleotide incorporation event marked by the release of an inorganic pyrophosphate. Unlike Illumina platforms, which use modified nucleotides, the pyrosequencing method adds only a single type of dNTP (i.e. dATP, dCTP, dTTP or dGTP) at a time. This extension step is immediately followed by a temporary pause, allowing the signal (release of inorganic pyrophosphate) to be detected by non-electrophoretic bioluminescence (Ronaghi et al., 1998). The signal intensity is directly proportional to the number of incorporated nucleotides, where incorporation of three dATPs would result in three times the intensity of a signal observed from incorporation of a single dATP. However, incorporation of eight or more of the same nucleotides can cause

the signal to become saturated, prohibiting detection of longer repetitive nucleotide sequences (Metzker, 2010).

Illumina platforms use a method known as cyclic reversible termination, where each cycle comprises nucleotide incorporation, fluorescence imaging and cleavage (Metzker, 2005). Like Sanger sequencing, the Illumina platform utilizes chain terminators, with the exception that chain termination is reversible and the reaction restarts after imaging has taken place.

ABI SOLiD uses a technique known as the ‘colour space’ system or ‘2-barcoded encoding system’. Processes involved in ABI SOLiD next generation sequencing are summarized in **Figure 4.2**.

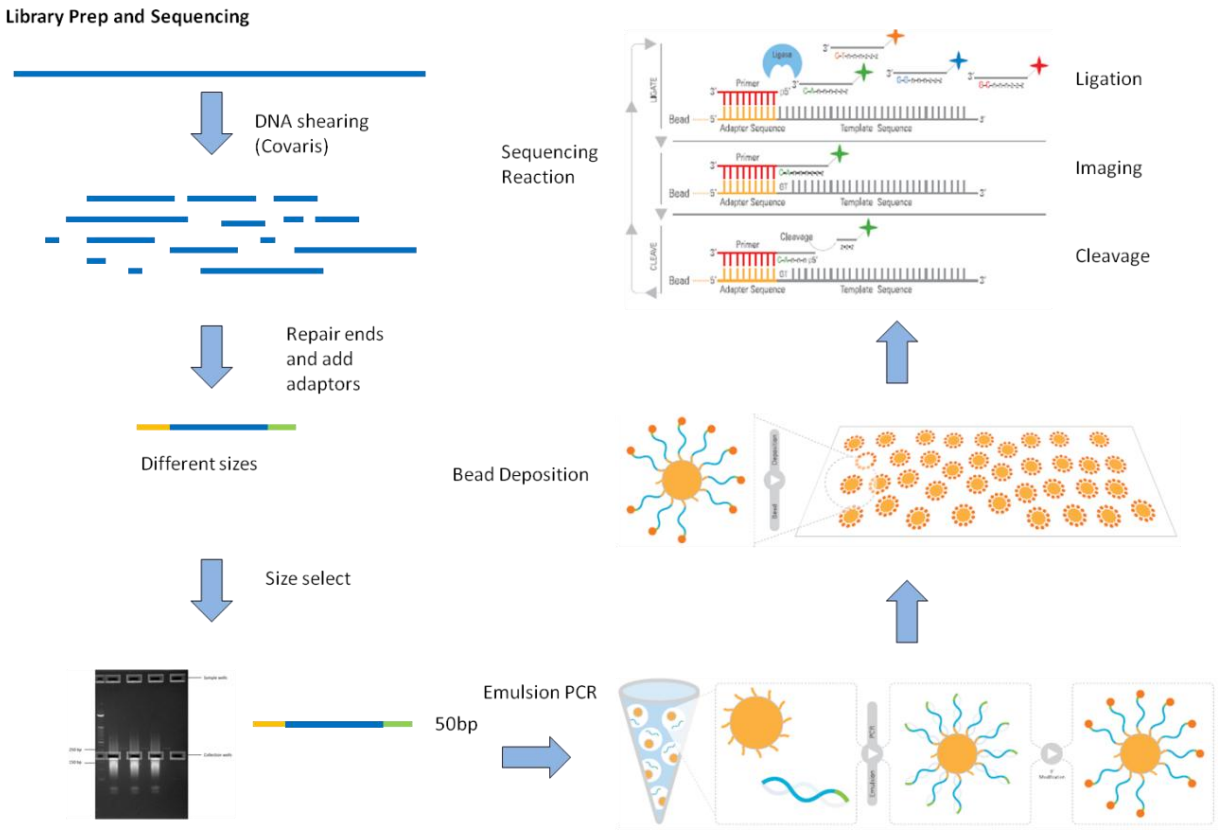



**Figure 4.2 Overview of library preparation of ABI SOLiD®.** The flow diagram summarizes the processes involved in next generation sequencing using the ABI SOLiD® sequencer. The LR-PCR amplified DNA was sheared into random sizes using the Covaris AFA$^{TM}$ – similar technology to a sonicator with improved control of wavelengths and isothermal advantage. The sheared DNA's are end-repaired, and two adaptors (known as 'P1' and 'P2') are ligated to both ends of the DNA fragments. A specific length of DNA fragments (e.g. 50bp as used in this study) are extracted using a size selection gel. This is followed by amplification via emulsion-PCR using two primers: 'A1' and 'A2', which are complimentary to the 'P1' and 'P2' adaptors. 'A1' primers are coated on polystyrene beads, which enable enrichment of DNA fragments. These polystyrene beads, with DNA attached, are deposited onto a glass slide, where the sequencing reactions take place. The ABI SOLiD® sequencer supports flexible slide segmentation (also known as flow cells), which enables several independent samples to be run simultaneously. Millions of random short DNA fragments are sequenced in parallel using the 'colour space' system, where each nucleotide is interrogated twice.

Although the NGS platforms described are distinct from each other in many aspects, they share substantial similarities:

- library preparation – all platforms involve DNA shearing into smaller fragments followed by end-repair and addition of adaptors (short DNA fragments) onto both ends of the template DNA.
- amplification of DNA template – all NGS platforms require amplification of DNA template, so that the sequencing reaction produces sufficient signals which can be detected by the instruments' optical systems. Given that no DNA polymerase is 100% accurate, this has been considered to be a limitation of NGS technology (Mardis, 2011).
- repeating steps – all platforms perform sequencing reactions using a series of repeating steps, which are performed automatically.

Sequencing library preparation

Two major methods exist to create a next generation sequencing library: pair-end and mate-paired sequencing libraries, where reads generated from sequencing of these libraries are known as paired-end and mate-paired reads, respectively.

Single-end sequencing, with each DNA fragment only sequenced from one end, has been largely superseded by paired-end sequencing as a result of the lack of accuracy. Single-end sequencing results in a higher proportion of reads incapable of being aligned uniquely, resulting in these reads being unsuitable for variant discovery (Mardis, 2011).

Paired-end sequencing allows a DNA fragment to be sequenced from both ends, thus improving the confidence when it comes to calling SNPs (Mardis, 2011).

It should be emphasized that paired-end reads are from a single location of a genome region in comparison to the mate-pair sequencing. The difficulty of mapping

reads back to a locus of origin which is repetitive is considered a limitation for paired-end sequencing.

This led to the development mate-pair sequencing. Reads in mate-pair sequencing are made up of two short DNA segments derived from two genomic locations with the distance between the two known (**Figure 4.3**). Given one end of the read of a mate pair uniquely aligns to the reference sequence and the distance between the two ends is known, the location of the other end of the read should be obvious. Consequently, the use of a mate-pair sequencing library can greatly improve the coverage of next generation sequencing across the target regions.

However, as this technology is based on circularization of large DNA molecules, the low yield of circularization (directly proportional to the DNA molecules used) means the technology is DNA expensive (Mardis, 2011). Furthermore, mate-pair sequencing requires extra experimental steps and raises more challenges for mapping and alignment, which in turn could result in more reagent cost and longer time to process the data.

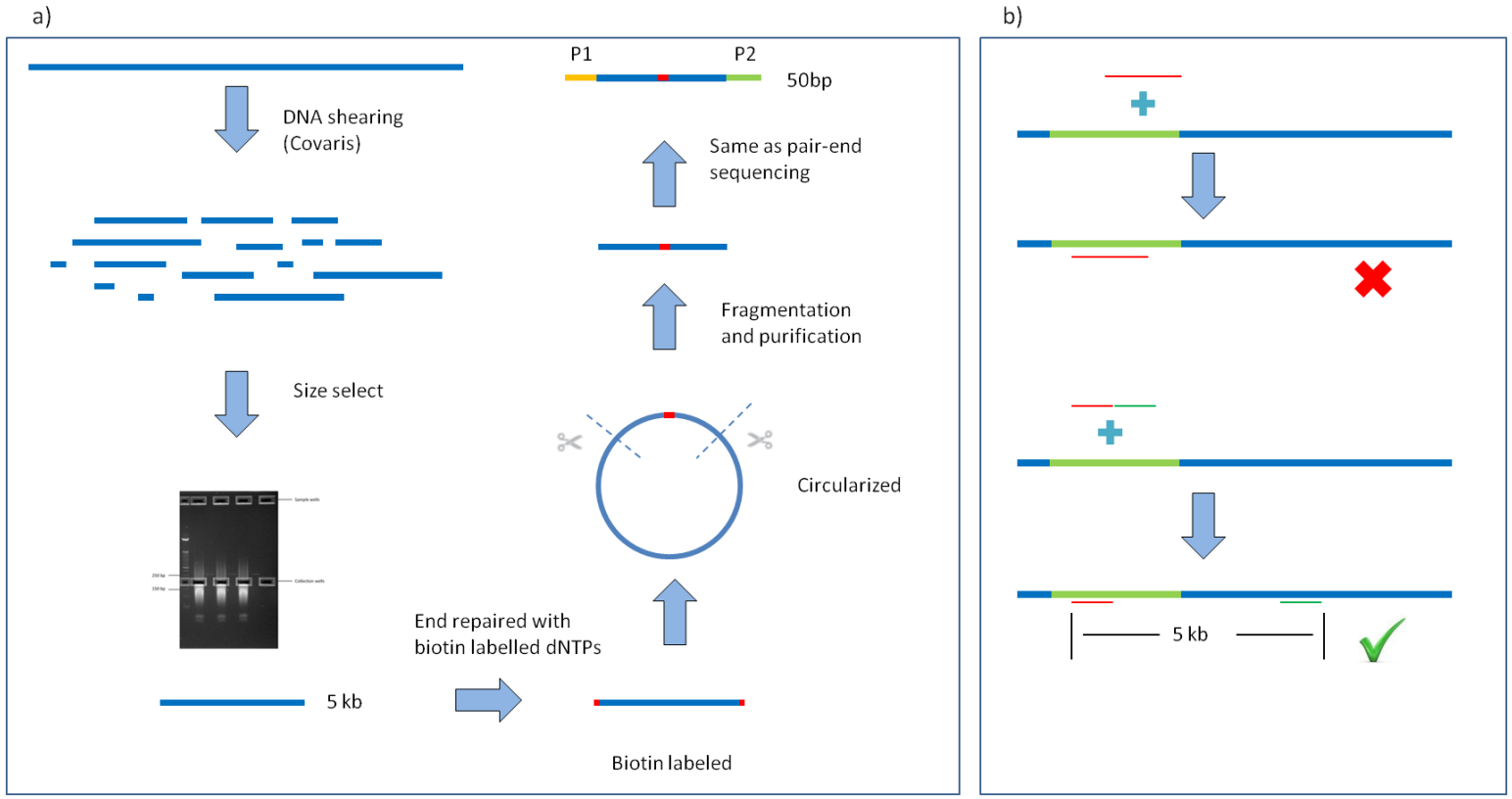



**Figure 4.3 Overview of 'mate-pair sequencing' library preparation.** a) summarizes steps involved in library preparation for mate-pair sequencing run. DNA template is represented by blue bars. Larger sizes of DNA fragments (e.g. 5 kb) are selected in contrast to the library preparation for paired-end sequencing. These DNA fragments are end-repaired with biotin labelled dNTPs, which is followed by circularization. Non-circularized DNA fragments are removed by digestion. After further fragmentation, biotin labelled DNA fragments are purified. These fragments are then end-repaired, and two adaptors ('P1' and 'P2') are added, attaching to both ends of the amplicon. The rest of the sequencing reaction is identical to paired-end sequencing. b) illustrates two scenarios of alignment to the reference genome sequence, top – paired-end read is unable to map to repetitive genomic region (highlighted in green) and bottom – as one end of the mate-pair read uniquely aligns to a non-repetitive genomic sequence, the locus of origin for the other end is obvious. Mate-pair reads outperform paired-end reads on mapping to repetitive genomic regions and thus improve sequencing coverage.

ABI SOLiD® Colour space system

SOLiD stands for Sequencing by Oligonucleotide Ligation and Detection. A unique feature of SOLiD technology is it uses the colour space system with each nucleotide base being interrogated twice. This double interrogation greatly increases the accuracy of each nucleotide call.

SOLiD uses 16 two-base-encoded probes as illustrated in **Figure 4.4a**. Four specific dinucleotides are labelled with a single fluorescent dye, and a total of four fluorescent dyes are used in ABI SOLiD® next generation sequencing.

Colour space calls are converted into nucleotide calls based on the chart shown in **Figure 4.4a** using BioScope®.

Furthermore, colour space system can easily distinguish a SNP from a reading error (**Figure 4.4b**). As each nucleotide base is determined by two adjacent colours, a single colour change is an indication of a reading error.

**Figure 4.5** summarized the processes involved in ABI SOLiD® next generation sequencing. Five rounds of ligation reaction (each of ten cycles) are necessary to sequence all 50bp reads generated.

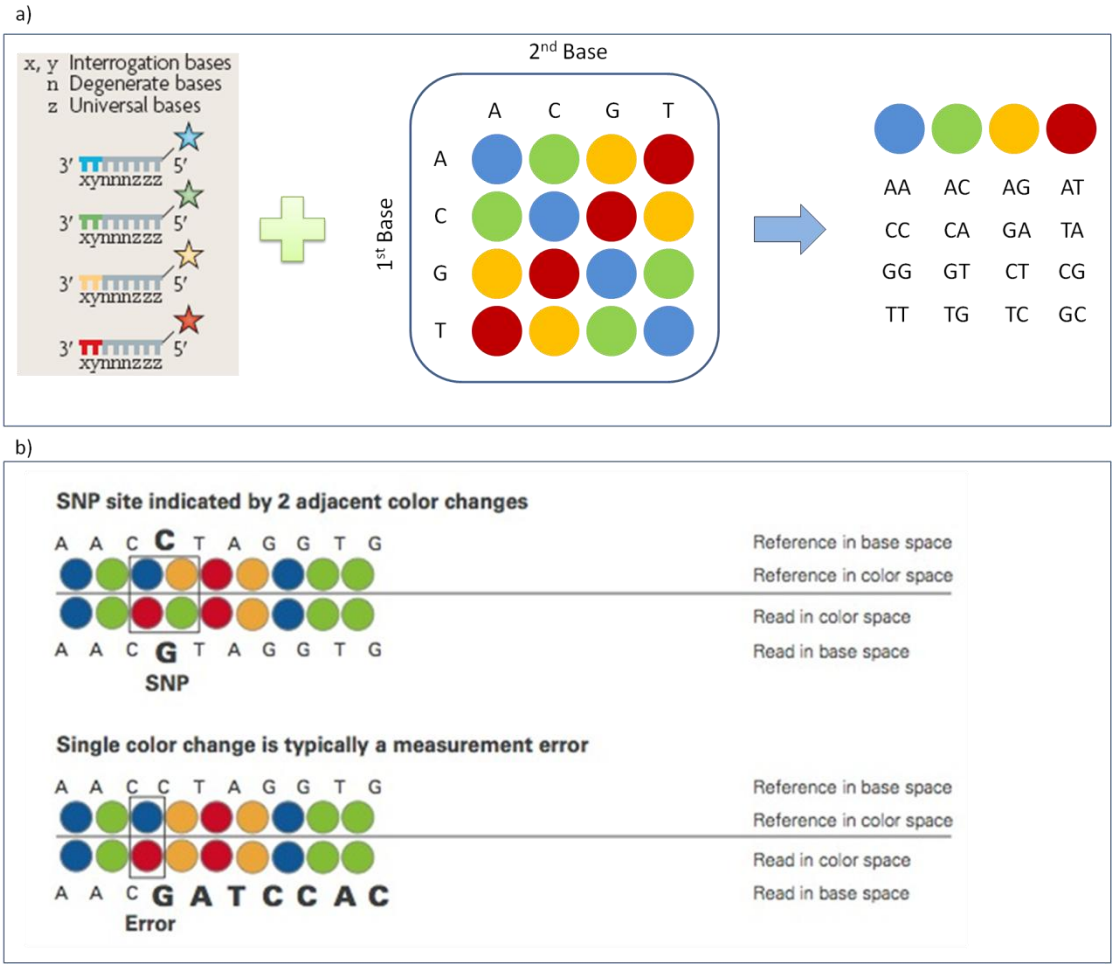



**Figure 4.4 ABI SOLiD® colour space system.** a) Each probe consists of i) specific dinucleotides at nucleotide bases 1-2, ii) degenerate nucleotides bases (denoted as 'n' (often RNA nucleotides) and 'z' (inosine)) at nucleotide bases 3-8, iii) restriction site between base pair 5 and 6 and iii) fluorescent dye labelled at the 5' end. The specific dinucleotides of each probe, according to the chart, are as shown. b) illustrates a scenario where a sequencing error can easily be distinguished from a real SNP call in ABI SOLiD® NGS: a SNP is represented by two adjacent colour changes, whereas a single colour change is an indication of a sequencing error.

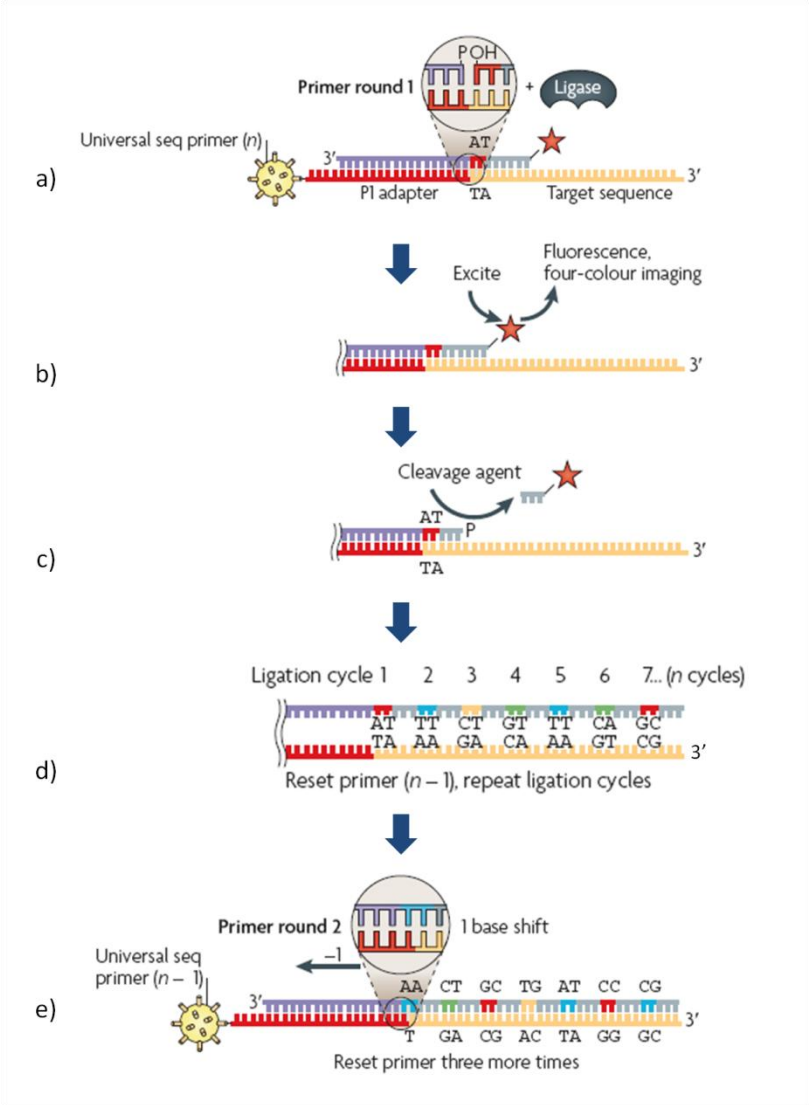



**Figure 4.5 Flow diagram of the sequencing reaction of ABI SOLiD® NGS.** The flow diagram summarizes steps involved in the sequencing reaction using by ABI SOLiD®: a) annealing of a universal sequence primer enables one of the 16 fluorescently labelled probes (complementary to the target sequence) to be ligated with the universal primer facilitated by ligase, b) four-colour imaging c) nucleotide bases 6-8 (denoted as 'z' (inosine) in **Figure 4.4a**) are subsequently cleaved off, leaving a 5' phosphate group for further ligation reactions, d) nine more cycles of the sequencing step is required to interrogate all nucleotide bases, e) the DNA is denatured and the complementary strand is discarded. A new universal sequence primer which binds to 'n-1' position is added, and is subjected to another ten cycles of ligation reactions. A total of five ligation rounds are performed in ABI SOLiD® NGS. (Adapted from Metzker, 2010)

Challenge of analysing NGS data

The distinct chemistry of NGS results in fundamental changes to the way that the data are analysed in comparison with analysing the capillary data from Sanger sequencing. An analysis pipeline of NGS data from the pooled DNA samples using ABI SOLiD® is described later in this chapter.

Furthermore, it is conceivable that the production of millions of NGS reads causes challenges to the management of information technologies such as data transfer, storage and quality control. Some NGS systems are able to generate over one billion short reads per instrument run. Therefore, sufficient large data storage and transfer devices are essential.

The delivery of high quality genome sequence and SNP calling is challenging, and is dependent upon development of suitable bioinformatics software and sufficient computing power in order to analyse the large-scale data. This includes accurate alignment/assembly of read data and production of error models to permit confident calling of novel rare SNPs (Metzker, 2010).

The accuracy of the alignment has a crucial role in variant detection. Incorrectly aligned reads may lead to errors in SNP and genotype calling. Therefore, it is important for alignment algorithms to be able to cope with sequencing errors, as well as potential real differences (e.g. SNPs and Indels) between the reference genome and the sequenced genome. In addition, the aligner must also be able to produce well-calibrated alignment quality values, as variants calls are dependent on those scores (Nielsen et al., 2011).

## 4.2 Aims

An important argument for GWAS is that the genes in which common variants are found, or genes nearby, may well contain functional rare variants; these may have high enough penetrance to be considered as candidates for possible preventive screening strategies (Bodmer and Bonilla, 2008).

The aims of this study were to i) investigate if the *TRIM15* gene amplicons 'A' and 'B' (**Figure 4.6**) harbour multiple rare variants (with allele frequency between 1% to 5%) by analysing the next generation sequencing data, and ii) prioritizing these SNPs according to their potential biological functions and associations with the risk of LOAD.

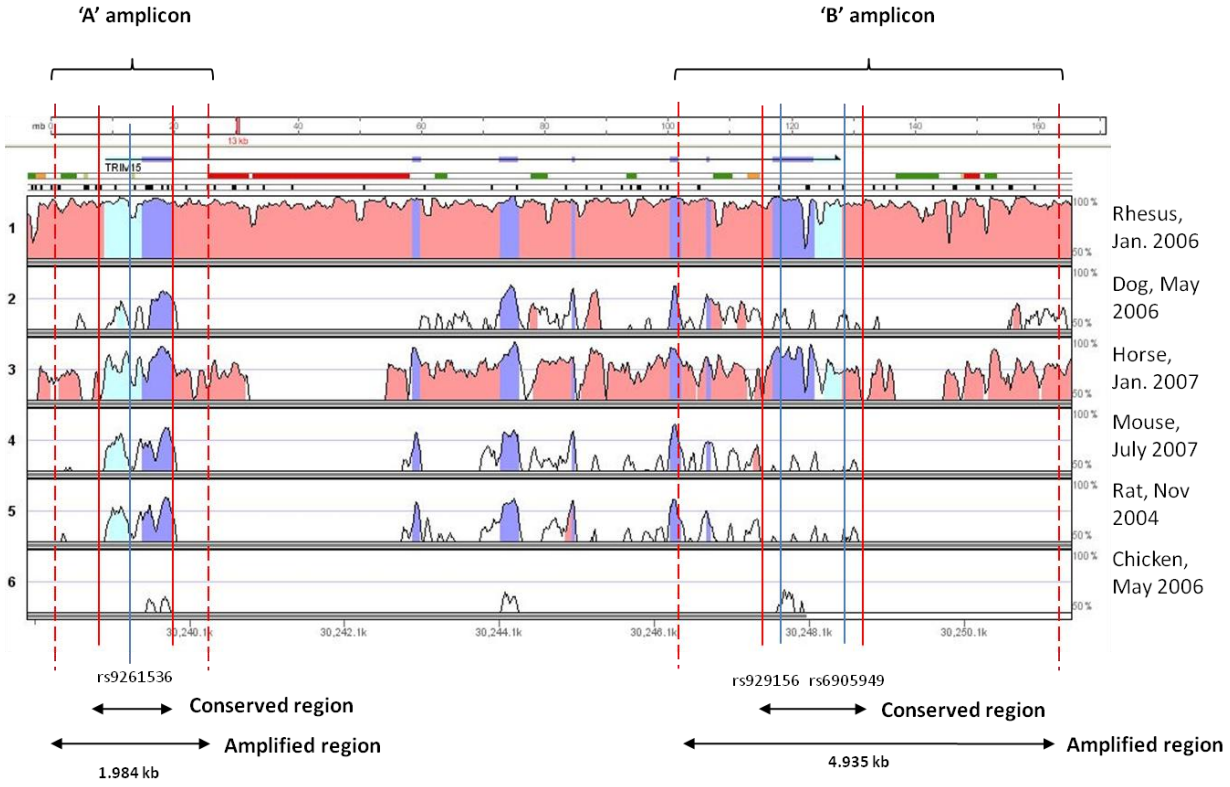



**Figure 4.6 Conservation plot using VISTA browser.** Figure showing genomic position of *TRIM15* 'A' and 'B' amplicons which were sequenced using ABI SOLiD® NGS. The blue vertical lines represent the locations of SNPs on the genotyping chips (**Chapter 3**). Conserved regions are indicated between the solid red lines, and the actual *TRIM15* 'A' and 'B' amplicons are indicated between the dotted red lines. The corresponding vertebrate species are documented on the right. Conserved regions are also highlighted in colours according to their function (red - introns and intergenic regions, blue - exons, cyan - UTRs).

## 4.3 Strategy

The next generation sequencing pipeline performed in this study are summarized in **Figure 4.7**. The project consists of the following distinct steps:

- Preparation of DNA samples
- Ascertainment of DNA region for sequencing
- Target enrichment using LR-PCR
- Equi-molar pooling of LR-PCR products
- Library preparation and the sequencing reaction
- Quality assessment
- SNP discovery and identification
- Validation and replication

Preparation of DNA samples

Genomic DNA was extracted from brain tissue using QIAGEN® DNeasy Blood and Tissue Kit (**Methods 2.2.1**). Concentrations of DNA samples were measured using NanoDrop® (**Methods 2.2.2**).

Ascertainment of DNA region for sequencing

Two conserved regions of *TRIM15* gene located at the 5' and 3' ends were ascertained using VISTA browser (**Figure 4.6**) (**Method 2.3.6**). The middle part of *TRIM15* gene (region between the 'A' and 'B' amplicons) was sequenced using Illumina HiSeq™ (ongoing project). This region was sequenced in 96 LOAD cases which were separated into 8 DNA pools, where each pool comprises 12 samples. No control subjects were sequenced for this part of the project.

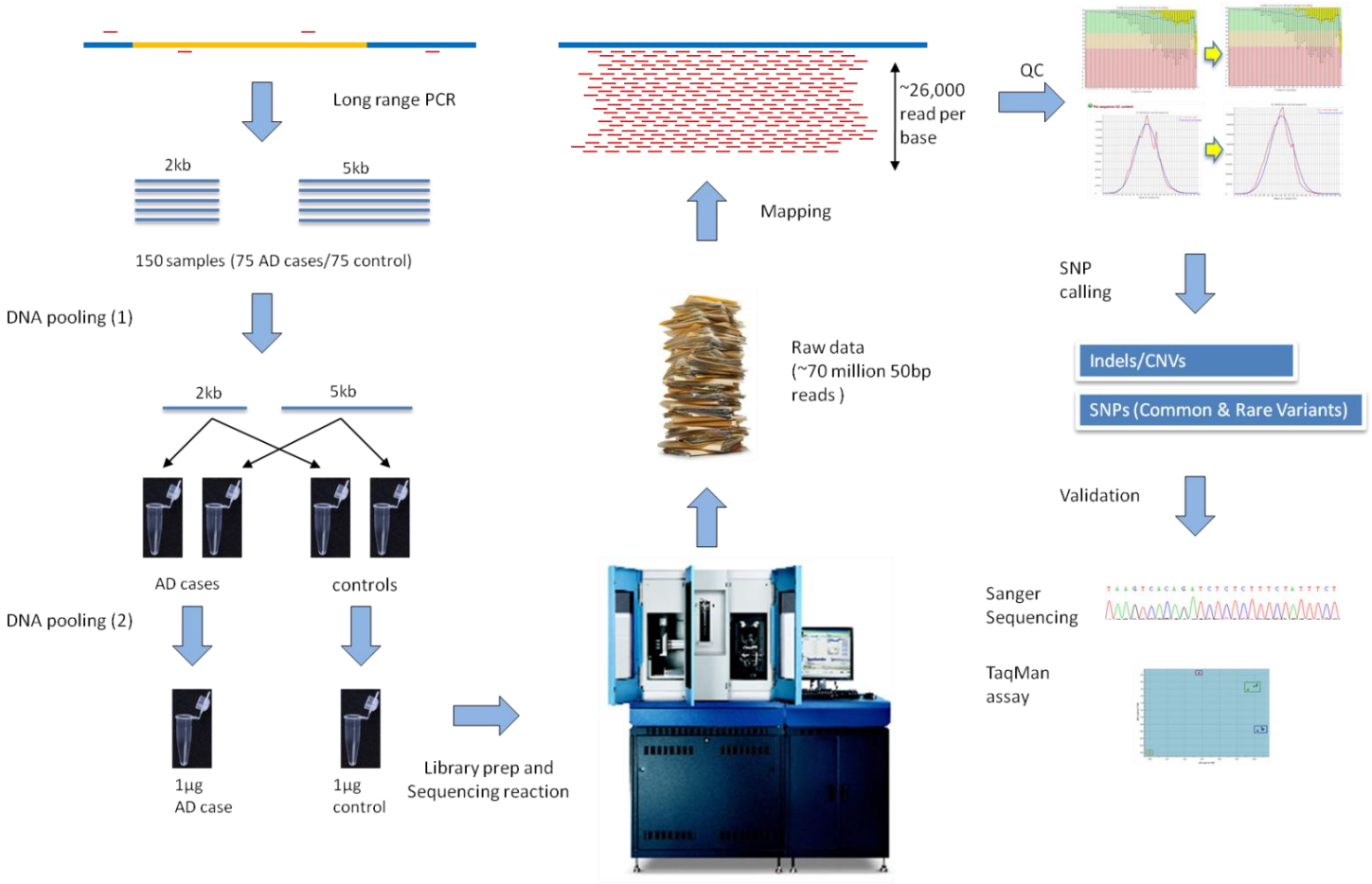



**Figure 4.7 Overview of the next generation sequencing pipeline in this study.** The flow diagram illustrates seven distinct processes conducted in this study: target enrichment using LR-PCR, equi-molar pooling of DNA samples, library preparation and sequencing using ABI SOLiD® NGS sequencer, translation of colour space calls to nucleotide calls and mapping to the reference genome sequence, quality assessment of the raw data, variant discovery and identification, validation using TaqMan® genotyping assays or Sanger sequencing.

Target enrichment

Two conserved regions of the *TRIM15* gene were enriched using LR-PCR (**Methods 2.2.3**) on 150 samples (75 LOAD cases and controls) (**Methods 2.1**). These LR-PCR products were visualized on EtBr stained 0.7% agarose gels under UV light together with blank (negative controls) and the GeneRuler$^{TM}$ 1kb Plus DNA ladder (**Methods 2.2.4**).

DNA samples were kept for NGS only if they displayed a strong intensity band on the agarose gel of the correct size and no evidence of DNA contamination noted by examining the negative control samples. A small number (2-5) DNA samples were selected and Sanger sequenced (**Methods 2.2.5**) to ensure the DNA amplified was of the correct sequence by comparison to the reference sequence (hg19) downloaded from the NCBI database.

Equi-molar pooling of LR-PCR products

Four DNA pools (AD cases and controls) for *TRIM15* 'A' and 'B' amplicons were created by pooling 5µl of each PCR amplified DNA products ('DNA pooling 1' in **Figure 4.7**).

Gel extractions were undertaken for each of the DNA pools to remove non-specific PCR products and primer-dimers (**Methods 2.2.6**). Concentrations of the cleaned LR-PCR products were measured using Qubit® (**Methods 2.2.2**).

Two DNA pools (case and control) each containing 1µg of DNA were subsequently created ('DNA pooling 2' in **Figure 4.7**) by adding all LR-PCR amplicons into the pool. Volumes were adjusted according to the concentrations measured and the sizes of the amplicons.

The amount of DNA for each amplicon contributed to the final pool was calculated using the equation:

$$\textbf{Amount of DNA} = \frac{\textbf{length of amplicon (bp)} \times \textbf{1000ng}}{\textbf{length of target (bp)}}$$

Library preparation and the sequencing reaction

Library preparation and the sequencing reaction of ABI SOLiD® were performed by the next generation sequencing unit at the University of Nottingham. Sequencing reads were aligned using BioScope® (**Methods 2.4.1**).

Quality assessment

Raw data generated from the next generation sequencing were assessed using FastQC (**Methods 2.4.2**).

Read depth and coverage

Read depth was calculated using the formula:

$$\textbf{Read Depth} = \frac{\textbf{Number of reads} \times \textbf{Length of the reads}}{\textbf{Length of the target region}}$$

Fold coverage (per individual and chromosome) was calculated using the formula:

$$\textbf{Coverage (per individual)} = \frac{\textbf{Read Depth}}{\textbf{75}}$$

$$\textbf{Coverage (per chromosome)} = \frac{\textbf{Read Depth}}{\textbf{150}}$$

An R script was written to draw the histogram depicting the fold coverage:

```
rm(mydata)
rm(count)
rm(depth_cont)
mydata <- read.table("mydata.txt", header=T)
count <- 0
for (i in mydata$cont_cov) {
  if (count == 0) {
    depth_cont <- i
    count <- 1
  }
  if (count > 0) {
    if (i > 10) {
      depth_cont <- c(depth_cont,i)
    }
  }
}
hist(depth_cont, breaks=300, ylim=c(0,100))
```

SNPs discovery and identification

SNPs were called using Syzygy and Freebayes (**Methods 2.4.3**). QC thresholds were applied to SNP calls:

- Base quality threshold (--bqthr): 10
- Mapping quality threshold (--mqthr): 50

SNPs were separated into high and low quality according to QC criteria: strand bias, read depth and error models as created by the software.

SNP annotation

SNPs were annotated using the Variant Classifier program (**Methods 2.4.4**). The Variant Classifier input file is shown in **Appendix 8.2**. The full transcript 'ENST00000376694' from the *TRIM15* gene 'ENSG00000204610' was used. The annotation data was downloaded from the most up-to-date Ensembl database. Fisher's exact test (2-sided) was utilized to examine the association of SNPs identified with the risk of LOAD using the estimated allele counts generated from both pools (case and control).

SNPs identified by NGS were visualized using UCSC custom track (**Methods 2.4.5**). Four custom tracks were entered into the genome browser: power to detect a singleton, high quality SNPs (known and novel) and association p-values.

SNP validation and replication

SNPs identified by NGS were compared with data documented in the latest SNP databases: HapMap (CEU population release 28), dbSNP (release 132) and the 1000 genome project VCF (variant call file) accessed using the tabix program (**Methods 2.4.3**).

SNPs showing significant evidence of association (with LOAD), as suggested by NGS data, were validated by direct genotyping of samples that were used in creating the sequencing library for NGS using Sanger sequencing (**Methods 2.2.4**) and TaqMan® genotyping assay (**Methods 2.2.7**).

Replication studies were performed for SNPs that validated. SNPs showing consistent MAF were further genotyped in an independent sample cohort (93 AD cases and controls) using the TaqMan® genotyping assay (**Methods 2.2.7**). SNP validation and replication experiments were performed by Narat Pititaweewat and Rebecca Gibbons (MSc students in our laboratory).

## 4.4 Results

LR-PCR of *TRIM15* 'A' and 'B'

Two conserved regions of *TRIM15* gene ('A' and 'B') were successfully amplified using LR-PCR. The sizes of these two amplicons are 1,984 bp and 4,935 bp, respectively (**Figure 4.6**).

Coverage of *TRIM15* gene

The latest human genome build hg19/GRCH37 was used as the reference genome sequence. Both amplicons ('A' and 'B') encompasses the following genomic regions of *TRIM15* gene:

**Promoter region of *TRIM15*:**

- 618 bp (chr6: 30130365-30130982) upstream of *TRIM15* gene

**Untranslated regions (UTR):**

- complete 5' UTR (length = 479 bp; chr6: 30130983-30131461)
- complete 3' UTR (length = 340 bp; chr6: 30140127-30140466)

**Exons:**

- complete exon 1 (length = 381 bp; chr6: 30131462-30131842)
- complete exon 6 (length = 33 bp; chr6: 30138753-30138785)
- complete exon 7 (length = 518 bp; chr6: 30139609-30140126)

**Introns:**

- complete intron 6 (length = 823 bp; chr6: 30138786-30139608)
- 506 bp of intron 1 (chr6: 30131843-30132348) (full size of intron 1 = 3110 bp)
- 355 bp of intron 5 (chr6: 30138398-30138752) (full size of intron 5 = 359 bp)

**3' downstream of *TRIM15* gene:**

- 2866 bp (chr6: 30140467-30143332)

Altogether these genomic regions range from 30130365 to 30143332 on chromosome 6.

LR-PCR optimisation

Optimisation of conditions for LR-PCR is key to successful amplification of large size amplicons (> 1 kb), which are more sensitive to experimental conditions compared with amplification of small amplicons (< 1 kb). Furthermore, LR-PCRs require longer time to run in comparison to standard PCR using *Taq* polymerase.

Experiment conditions were successfully optimised for LR-PCR amplification of *TRIM15* 'A' and 'B' fragments, and the optimisation results are shown in **Figure 4.8** (**Methods 2.2.3**).

LR-PCR and equimolar pooling

*TRIM15* 'A' and 'B' amplicons were successfully amplified by LR-PCR, pooled in equi-molar amounts and cleaned via gel extraction (**Figure 4.9**).

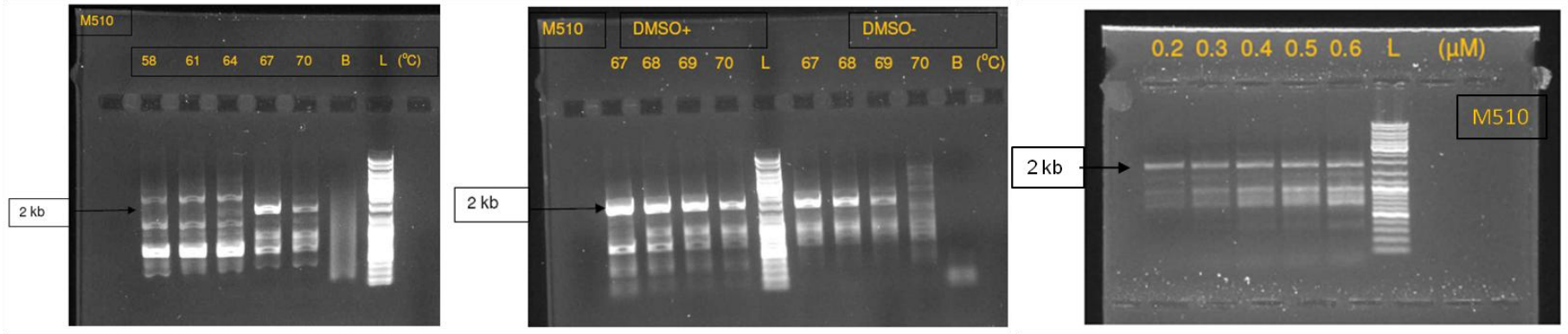



**Figure 4.8 Optimisation of LR-PCR.** EtBr stained 0.7% agarose gel depicting PCR optimisation result using temperature gradient 58°C to 70°C (left), 67°C to 70°C (middle) both in presence and absence of DMSO (middle) and a gradient with varying primer concentrations (right). The corresponding temperature and primer concentrations are shown together with sample ID on the top of the gel. B – blank (negative control), L – DNA ladder (GeneRuler$^{TM}$ 1kb ladder)

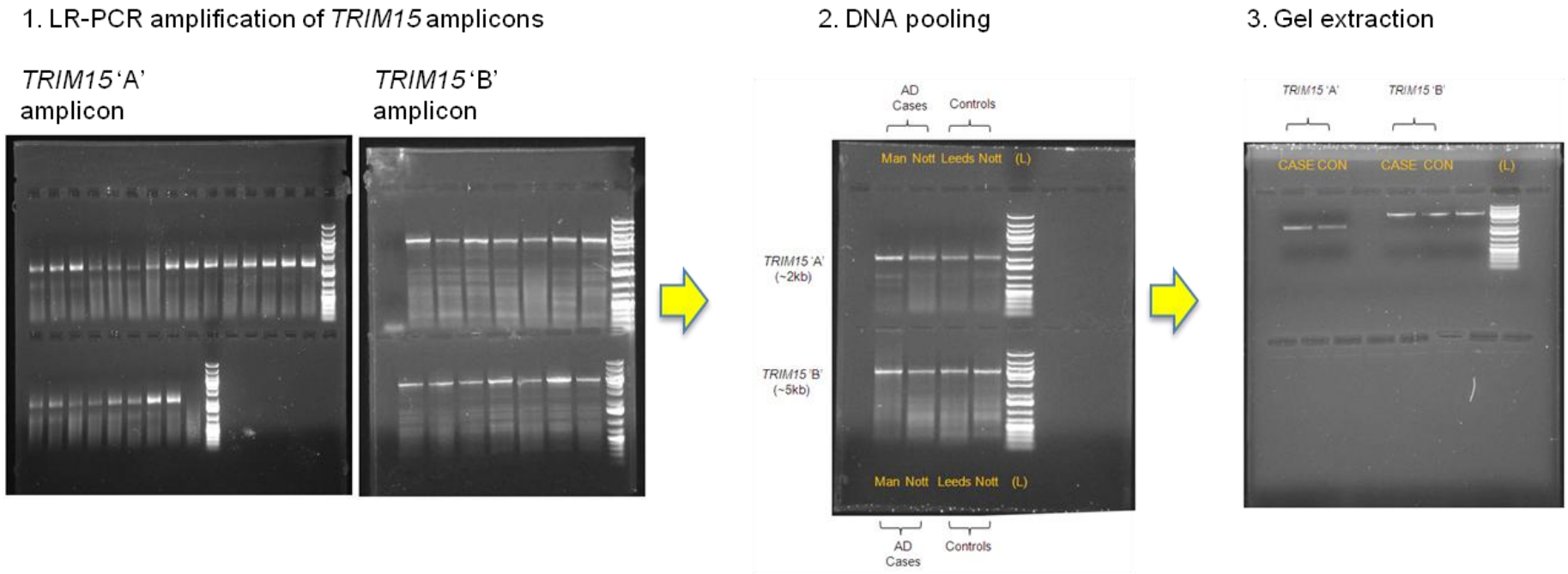



**Figure 4.9 LR-PCR and equimolar pooling.** *TRIM15* amplicons ('A' and 'B') were successfully amplified in 75 AD cases and 75 controls (left). DNA samples of the same amplicon (AD cases and controls) were pooled (middle). Gel extractions were undertaken for all DNA pools created, and visualized using EtBr stained 0.7% agarose gel (right).

General statistics and quality assessment of raw NGS data

The raw next generation sequencing data is represented by a total of 78,133,090 short oligonucleotide reads (case pool: 34,043,745; control pool: 44,089,345).

82.76% of reads for the control pool (36,489,719 reads) and 87.07% of reads for the case pool (29,641,561 reads) were successfully mapped to the reference genome sequence (hg19) using BioScope® (v 1.3) (performed by the next generation sequencing unit at University of Nottingham). ~15% of reads produced by NGS were unable to be mapped to the reference genome sequence, likely to be due to the fact that these reads belonged to genome sequences which are repetitive.

Of those reads that have been mapped to the reference genome sequence, 89.76% (26,607,720) and 91.46% (33,372,196) reads were mapped to targeted genome regions (i.e. regions which were enriched by LR-PCR).

Furthermore, almost all mapped reads are of 50bp in length as expected (**Figure 4.10**).

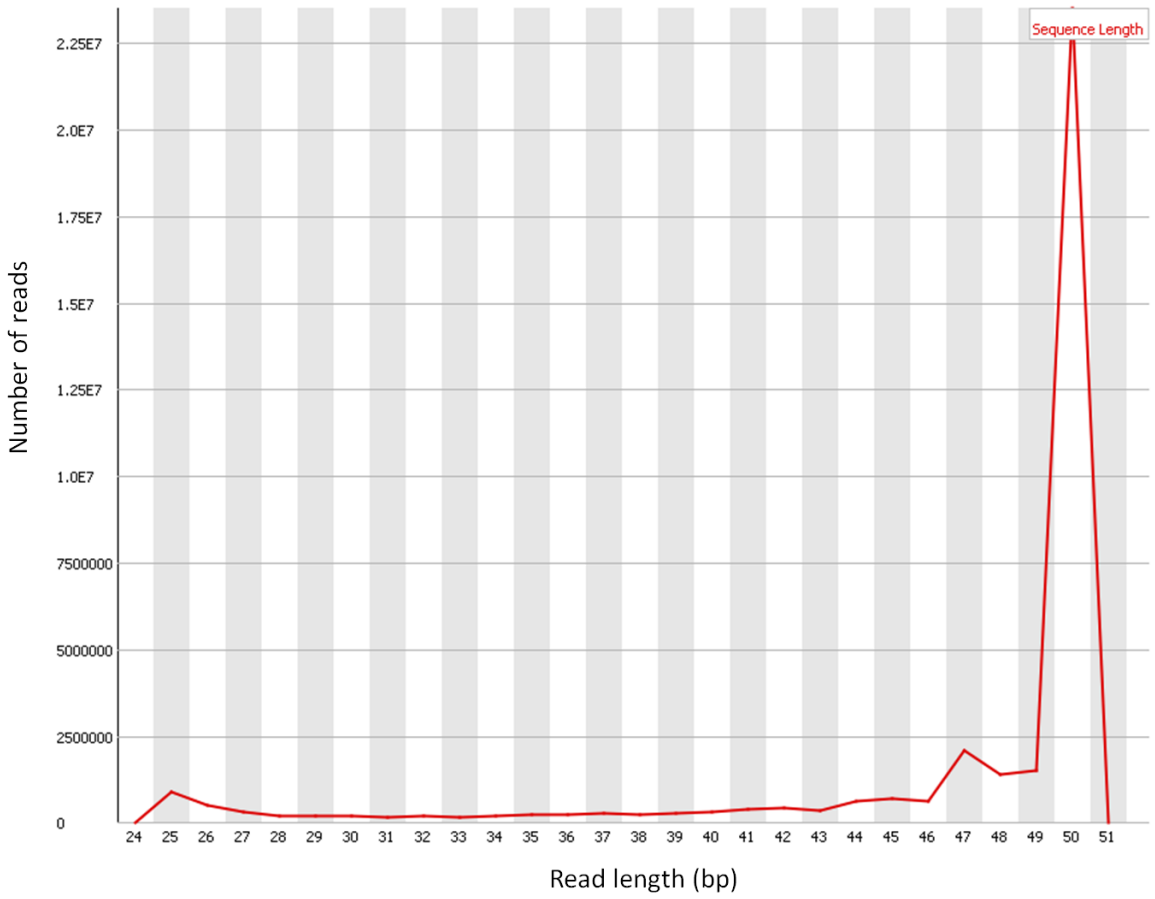



**Figure 4.10 Read length of the control pool.** The x and y axis represent read length (bp) and corresponding read counts, respectively. Almost all reads are 50bp in length as expected. The diagram was generated using FastQC (**Methods 2.4.2**).

Quality scores

The quality of each sequencing read was measured using two scores: base quality and mapping quality. Base quality values are indicated as ASCII codes (range 33-73) (**Methods 2.4.1**), with each ASCII code representing a single PHRED score (range 0-40). The relationship between a PHRED score and probability of being a wrong call is shown (Nielsen et al., 2011):

$$\mathbf{PHRED\ Score} = -\mathbf{10} \times \mathbf{log}\,\boldsymbol{P}(\mathbf{error})$$

A base quality of 40 is currently the highest quality that the NGS instrument (including ABI SOLiD®) can generate, which indicates a p-value of 0.0001 or the nucleotide call is 99.99% accurate (**Table 4.1**).

Similarly, mapping quality is also represented by PHRED scores with wider quality score range (0-100).

Due to the cyclic nature of the next generation sequencing, base qualities are higher in early cycles of the sequencing reaction, and lower in the later cycles (**Figure 4.11**). It is likely that reagents required for the experiment are diminished in the later cycles.

Number of reads mapped to *TRIM15* 'A' and 'B' amplicons

3,134,504 reads and 3,492,467 reads were mapped to *TRIM15* gene 'A' and 'B' amplicons, respectively.

**Table 4.1 ASCII codes depicting base qualities in next generation sequencing data.** The table shows nucleotide base qualities between 0 and 40. ASCII code is shown together with corresponding base quality scores (Phred scaled) and *p* (error) – likelihood of being a wrong call.

| ASCII code | Corresponding base quality | *p* (error) | ASCII code | Corresponding base quality | *p* (error) |
|---|---|---|---|---|---|
| ! | 0 | 1 | 6 | 21 | 0.007943 |
| “ | 1 | 0.794328 | 7 | 22 | 0.00631 |
| # | 2 | 0.630957 | 8 | 23 | 0.005012 |
| $ | 3 | 0.501187 | 9 | 24 | 0.003981 |
| % | 4 | 0.398107 | : | 25 | 0.003162 |
| & | 5 | 0.316228 | ; | 26 | 0.002512 |
| ‘ | 6 | 0.251189 | < | 27 | 0.001995 |
| ( | 7 | 0.199526 | = | 28 | 0.001585 |
| ) | 8 | 0.158489 | > | 29 | 0.001259 |
| * | 9 | 0.125893 | ? | 30 | 0.001 |
| + | 10 | 0.1 | @ | 31 | 0.000794 |
| , | 11 | 0.079433 | A | 32 | 0.000631 |
| - | 12 | 0.063096 | B | 33 | 0.000501 |
| . | 13 | 0.050119 | C | 34 | 0.000398 |
| / | 14 | 0.039811 | D | 35 | 0.000316 |
| 0 | 15 | 0.031623 | E | 36 | 0.000251 |
| 1 | 16 | 0.025119 | F | 37 | 0.0002 |
| 2 | 17 | 0.019953 | G | 38 | 0.000158 |
| 3 | 18 | 0.015849 | H | 39 | 0.000126 |
| 4 | 19 | 0.012589 | I | 40 | 0.0001 |
| 5 | 20 | 0.01 | | | |

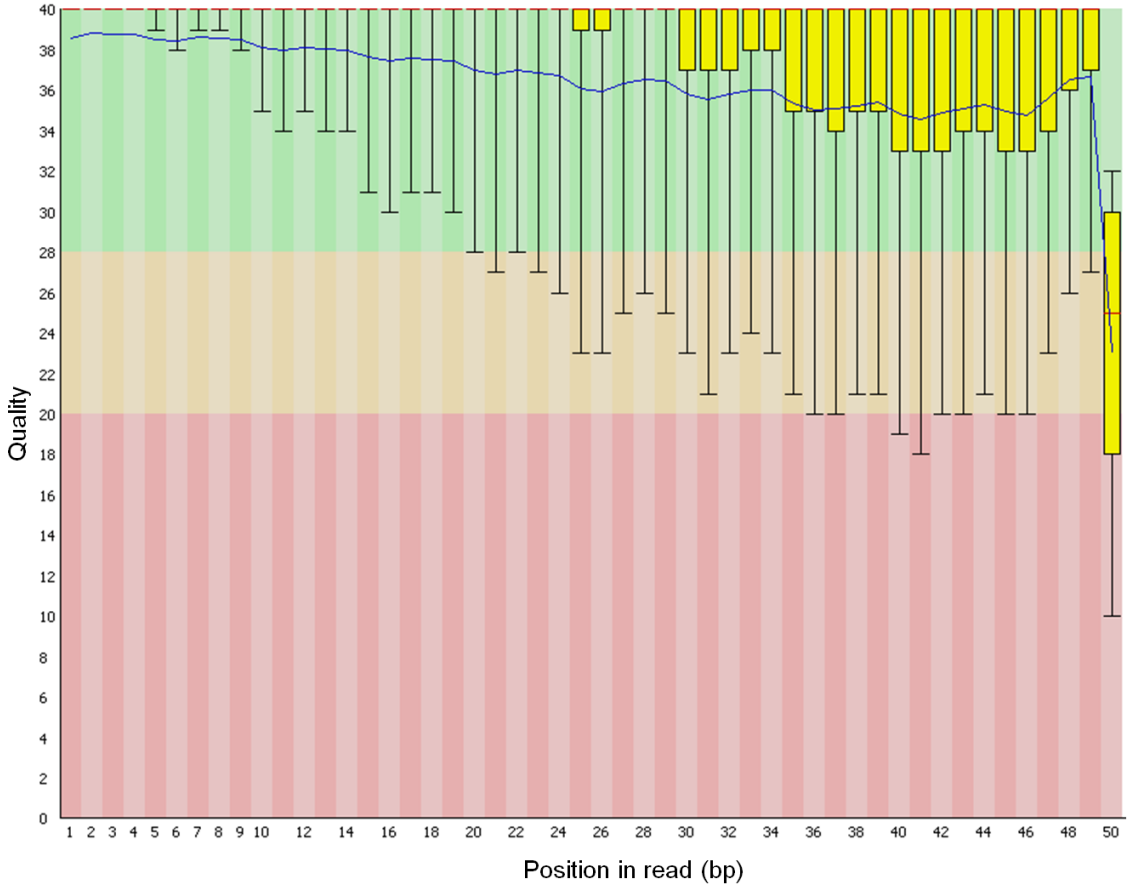



**Figure 4.11 Quality scores across all nucleotide bases in the control pool.** The x and y axis represent base position in reads and base quality (PHRED scaled), respectively. Nucleotide bases of high (28-40), moderate (20-28) and low (0-20) quality are shown in green, orange and red, respectively. The central box (in yellow) represents the distance between the first and third quartile with the median marked with a red line. The upper and lower whiskers represent the 10%-90% quartiles. The mean base quality scores are represented by a blue curve.

Power calculation

Three power calculations were performed:

- Power to discover SNPs given the number of samples sequenced
- Power to detect a singleton (a single alternative allele) according to the read depth (number of reads aligned to target regions).
- Power to detect an association between SNP discovered and risk of LOAD given the effect size (i.e. odds ratios) and allele frequencies.

Power for SNP discovery

Power to detect a SNP given the number of DNA samples sequenced was calculated and is shown in **Table 4.2** using **Methods 2.3.7**. The power to detect a SNP with MAF of 0.01 using 75 samples was estimated to be 78%. As a result, this study does not have sufficient power to detect a SNP with allele frequency less than 0.01.

Therefore, SNPs with MAF less than 0.01 in both AD cases and controls were removed from further analysis as a result of the lack of power.

**Table 4.2 Relationship between MAF, power and sample size required.** Minor allele frequency, power to detect SNPs and sample size required are as indicated. Sample size required to obtain 95% power to detect a SNP with MAF between 0.001 and 0.5 are shown on the left. Power to detect SNPs with sequencing of 75 samples (with MAF between 0.001 and 0.5) are shown on the right. Power to detect SNPs with MAF 0.01 using 75 samples is highlighted in yellow.

| Minor allele frequency | Power | Sample size required | | Minor allele frequency | Power | Sample size |
|---|---|---|---|---|---|---|
| 0.001 | 0.95 | 1497 | | 0.001 | 14% | 75 |
| 0.005 | 0.95 | 299 | | 0.005 | 53% | 75 |
| 0.01 | 0.95 | 149 | | 0.01 | 78% | 75 |
| 0.02 | 0.95 | 74 | | 0.02 | 95% | 75 |
| 0.03 | 0.95 | 49 | | 0.03 | 99% | 75 |
| 0.04 | 0.95 | 37 | | 0.04 | >99% | 75 |
| 0.05 | 0.95 | 29 | | 0.05 | >99% | 75 |
| 0.1 | 0.95 | 14 | | 0.1 | ~100% | 75 |
| 0.2 | 0.95 | 7 | | 0.2 | ~100% | 75 |
| 0.3 | 0.95 | 4 | | 0.3 | ~100% | 75 |
| 0.4 | 0.95 | 3 | | 0.4 | ~100% | 75 |
| 0.5 | 0.95 | 2 | | 0.5 | ~100% | 75 |

Power to detect a singleton

Syzygy provides '--power' function to estimate power to detect a singleton (a single alternative allele) at each nucleotide base according to the read depth – number of reads that mapped to the target regions.

The majority of the base pair positions in the *TRIM15* 'A' and 'B' amplicons showed a power value of 100, suggesting ~100% power to detect a singleton.

Power to detect a singleton can be calculated using the method described in **Methods 2.3.7**. For example, using the average read depth of ~20,500 in the control pool, the power to detect a singleton can be calculated:

$$\mathbf{P[detection]} = \mathbf{1} - \left(\mathbf{1} - \left(\frac{\mathbf{1}}{\mathbf{150}}\right)\right)^{\mathbf{20500}} = \mathbf{1} - \mathbf{2.8} \times \mathbf{10^{-60}}$$

$$= \sim \mathbf{100}\%$$

SNP with the 'power' value given by Syzygy less than 80 have been excluded from further analysis.

Power to detect an association

Power to detect an association was calculated using QUANTO v1.2.4 and the results are shown in **Table 4.3**. As indicated, the study has 80% power to detect an association for common variants with OR > 3 or a rare variants with OR > 4. This estimation, however, should be interpreted with caution as it is based on a number of assumptions (such as effect size and mode of inheritance), and gene-environment interaction (GxE) has not been taken into account.

**Table 4.3 Sample size required to detect an association calculated using QUANTO (v 1.2.4).** The required sample sizes (to have 80% and 95% power as indicated) were calculated for common variants (0.5 > MAF > 0.05) with odds ratio between 1.1 and 4.0, and rare variants (0.05 > MAF > 0.01) with odds ratio between 1.5 and 12.0. The '*' indicates sample size was calculated based on allele frequencies 0.10 to 0.20 instead of 0.05 to 0.5. All sample sizes shown are the the number of case-control pairs.

| **Common Variants (0.5 > MAF > 0.05)** | | | **Rare Variants (0.05 > MAF > 0.01)** | | |
|---|---|---|---|---|---|
| **OR** | **Sample required to have 80% power** | **Sample required to have 95% power** | **OR** | **Sample required to have 80% power** | **Sample required to have 95% power** |
| **4.0*** | 32-47 | 50-76 | **12.0** | 27-107 | 42-172 |
| **3.0*** | 48-75 | 76-119 | **11.0** | 28-114 | 45-183 |
| **2.0** | 88-350 | 141-561 | **10.0** | 30-123 | 48-197 |
| **1.9** | 102-412 | 162-661 | **9.0** | 33-135 | 53-216 |
| **1.8** | 120-498 | 191-797 | **8.0** | 37-150 | 58-241 |
| **1.7** | 145-619 | 232-991 | **7.0** | 41-172 | 66-275 |
| **1.6** | 183-800 | 293-1283 | **6.0** | 48-204 | 77-326 |
| **1.5** | 243-1094 | 389-1753 | **5.0** | 60-256 | 95-409 |
| **1.4** | 349-1618 | 559-2593 | **4.0** | 81-353 | 129-565 |
| **1.3** | 569-2719 | 912-4357 | **3.0** | 131-589 | 210-943 |
| **1.2** | 1171-5768 | 1876-9244 | **2.0** | 350-1623 | 561-2601 |
| **1.1** | 4264-21697 | 6834-34778 | **1.5** | 1094-5161 | 1753-8273 |

High quality SNPs

77 high quality SNPs were called using BioScope (v1.3) (**Methods 2.4.1**) and Syzygy (v1.1.0) (**Methods 2.4.3**).

21 SNPs with MAF less than 0.01 in both case and control pool were excluded from further analysis; 56 SNPs remained for further analysis.

Of the remaining high quality SNPs, 31 were known and 25 were novel according to the dbSNP database (release#132)

Further QC, taking into account strand bias, clustering and low read depth, were applied to these 25 novel variants, resulting in 15 being removed. This QC did not apply to known SNPs.

Known SNPs

All 31 known SNPs show identical alternative alleles as documented in the dbSNP database (**Table 4.4**).

Approximately half of these SNPs (15 out of the 31) were found to be documented in the latest version of HapMap database (release 28 - August 2010).

With exception of a single SNP (rs1063280), which showed a significantly different allele frequency in comparison to the allele frequency quoted in HapMap, comparison of the remaining 14 SNPs showed compelling consistency with a correlation coefficient ($R^2$) of 0.966 (**Figure 4.12**).

Allele frequency of the 'C' allele (alternative allele) of SNP rs1063280 was estimated to be 0.21. The frequency of the same SNP is however shown to be 0.89 in HapMap (CEU population release 28). The discrepancy in allele frequency suggests either the presence of possible population heterogeneity, an error caused by read misalignment or more likely a mis-call due to genotyping error.

**Table 4.4 Comparison of alternative alleles of known SNPs.** Table shows all known SNPs with allele frequency greater than 0.01 in the case or control pools identified by next generation sequencing. All alternative alleles are shown in the forward strand. Chromosome number (CHR), SNP identifier (SNP), base pair positions (BP) are as indicated together with reference allele (RA), observed alternative allele (Observed AA) and alternative allele documented in dbSNP database (release 132).

| CHR | SNP | BP | RA | Observed AA | AA found in dbSNP #132 |
|---|---|---|---|---|---|
| 6 | rs60650863 | 30130456 | A | G | G |
| 6 | rs17188113 | 30131123 | T | G | G |
| 6 | rs62407492 | 30131331 | T | C | C |
| 6 | rs9261536 | 30131349 | T | C | C |
| 6 | rs35278640 | 30131503 | T | C | C |
| 6 | rs2523733 | 30131515 | C | A | A |
| 6 | rs11961941 | 30131527 | G | T | T |
| 6 | rs17194460 | 30131546 | A | G | G |
| 6 | rs17194467 | 30131585 | G | A | A |
| 6 | rs17194474 | 30131711 | G | C | C |
| 6 | rs2074477 | 30132035 | G | A | A |
| 6 | rs41272587 | 30132100 | G | C | C |
| 6 | rs114344980 | 30138489 | G | A | A |
| 6 | rs1029238 | 30138645 | G | A | A |
| 6 | rs1029237 | 30138662 | C | T | T |
| 6 | rs41272591 | 30138853 | C | A | A |
| 6 | rs41272595 | 30138865 | T | C | C |
| 6 | rs115440118 | 30138895 | C | T | T |
| 6 | rs41272599 | 30139021 | A | T | T |
| 6 | rs929156 | 30139699 | G | A | A |
| 6 | rs1063280 | 30140342 | T | C | C |
| 6 | rs6905949 | 30140525 | T | C | C |
| 6 | rs13212414 | 30140540 | A | T | T |
| 6 | rs2844787 | 30140913 | G | T | T |
| 6 | rs9380156 | 30141042 | A | G | G |
| 6 | rs13213365 | 30141204 | C | T | T |
| 6 | rs757258 | 30142253 | T | C | C |
| 6 | rs757257 | 30142458 | G | A | A |
| 6 | rs961039 | 30142674 | G | A | A |
| 6 | rs957765 | 30142690 | G | A | A |
| 6 | rs2394737 | 30142999 | A | G | G |

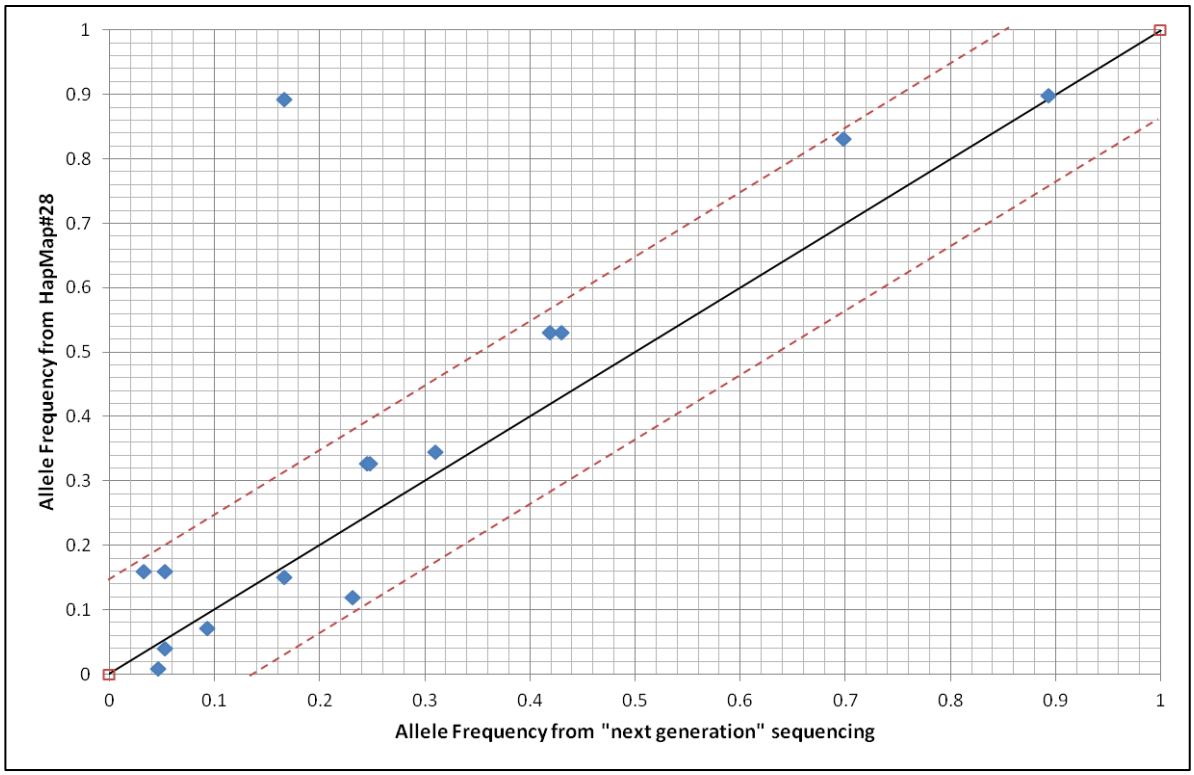



**Figure 4.12 Comparison of allele frequencies with HapMap data.** Scatter plot showing correlation between allele frequencies estimated from NGS (x-axis) and from HapMap (CEU population - release 28) (y-axis) in the control pool. Each SNP is represented by a blue diamond. Black line: expected line if allele frequencies from both sets are identical. The dashed line highlights that the majority of the SNPs exhibited similar allele frequency when compared with HapMap data (CEU population release 28).

Three out of the 31 high quality SNPs with dbSNP rs numbers showed significant evidence of association with risk of LOAD (*uncorr-p* < 0.05) – rs41272591 (*p* = 0.0006), rs9380156 (*p* = 0.0044) and rs6905949 (*p* = 0.034). All three SNPs are common (MAF > 5%) (**Table 4.5**). SNPs rs9380156 and rs6905949 exist in HapMap (CEU population release 28) with MAF 0.16 and 0.15, respectively. SNP rs41272591 does not exist in HapMap (CEU population release 28).

Direct genotyping of the two most significant SNPs (rs41272591 and rs9380156) using TaqMan® genotyping assays using the original 75 case and control samples showed similar allele frequency in the case pool as estimated by NGS.

Minor allele frequency (MAF) of SNP rs41272591:

- Case pool: 0.20 (TaqMan) and 0.19 (NGS)
- Control pool: 0.11 (TaqMan) and 0.05 (NGS)

Minor allele frequency (MAF) of SNP rs9380156:

- Case pool: 0.20 (TaqMan) and 0.17 (NGS)
- Control pool: 0.11 (TaqMan) and 0.05 (NGS)

Calculation of LD between the two SNPs showed strong evidence of linkage ($r^2$ = 0.864 and D' = 0.93 in the control pool and $r^2$ = 0.918 and D' = 0.958 in the case pool). As a result, both SNPs showed identical and significant association with LOAD (*p* = 0.036, OR = 2.09) using the same statistical test (Fisher's exact test (2-sided)).

Genotyping using an independent sample cohort (90 LOAD cases and 91 controls) showed no significant difference in allele frequencies (cases compared with controls) (*p* = 0.89, OR = 1.05), although the odds ratio appears to be in the same direction as suggested by the NGS.

**Table 4.5 Summary of all high quality SNPs identified in *TRIM15* 'A' and 'B' amplicons with dbSNP rs numbers.** Chromosome number (CHR), base pair position (BP) and SNP identifier (SNP) are indicated together with fold coverage, power to detect a singleton (Power), observed alleles (reference and alternative), annotation, MAF (case and control pool) and Fisher's exact test (2-sided) p-values. SNPs with association p-value ($p < 0.05$) are highlighted in yellow.

| CHR | BP | SNP | Fold coverage (case/control) | Power (case/control) | Allele (ref/alt) | Annotation | MAF in case pool | MAF in control pool | Fisher's exact test (2-sided) p-value |
|---|---|---|---|---|---|---|---|---|---|
| 6 | 30138853 | rs41272591 | 330/402 | 100/100 | C/A | INTRON.6 +67 | 0.1863 | 0.0508 | 0.0006 |
| 6 | 30141042 | rs9380156 | 238/404 | 100/100 | A/G | 3' DOWNSTREAM | 0.163 | 0.0533 | 0.0044 |
| 6 | 30140525 | rs6905949 | 364/372 | 100/100 | T/C | 3' DOWNSTREAM | 0.0803 | 0.1665 | 0.0340 |
| 6 | 30138645 | rs1029238 | 270/266 | 100/100 | G/A | INTRON.5 +251 | 0.033 | 0.0932 | 0.0554 |
| 6 | 30131515 | rs2523733 | 160/152 | 100/100 | C/A | EXON.1 | 0.3344 | 0.2314 | 0.0725 |
| 6 | 30130456 | rs60650863 | 340/276 | 100/100 | A/G | PROMOTER -536 | 0.0199 | 0.0666 | 0.0853 |
| 6 | 30140342 | rs1063280 | 54/32 | 96/74 | T/C | 3' UTR | 0.2476 | 0.1662 | 0.1163 |
| 6 | 30131546 | rs17194460 | 178/166 | 100/100 | A/G | EXON.1 | 0.0333 | 0.0781 | 0.1320 |
| 6 | 30139699 | rs929156 | 186/160 | 100/100 | G/A | EXON.7 | 0.2355 | 0.3099 | 0.1933 |
| 6 | 30138865 | rs41272595 | 320/392 | 100/100 | T/C | INTRON.6 +79 | 0 | 0.0199 | 0.2475 |
| 6 | 30138895 | rs115440118 | 386/474 | 100/100 | C/T | INTRON.6 +109 | 0 | 0.0199 | 0.2475 |
| 6 | 30140913 | rs2844787 | 392/420 | 100/100 | G/T | 3' DOWNSTREAM | 0.7586 | 0.6988 | 0.2982 |
| 6 | 30131349 | rs9261536 | 138/186 | 100/100 | T/C | 5'UTR | 0.8457 | 0.8938 | 0.3030 |
| 6 | 30138662 | rs1029237 | 262/278 | 100/100 | C/T | INTRON.5 +268 | 0.1969 | 0.2449 | 0.4057 |
| 6 | 30142253 | rs757258 | 152/202 | 100/100 | T/C | 3' DOWNSTREAM | 0.0363 | 0.0603 | 0.4127 |
| 6 | 30142690 | rs957765 | 186/234 | 100/100 | G/A | 3' DOWNSTREAM | 0.2062 | 0.2484 | 0.4907 |
| 6 | 30138489 | rs114344980 | 214/160 | 100/100 | G/A | INTRON.5 +95 | 0 | 0.0134 | 0.4983 |
| 6 | 30139021 | rs41272599 | 302/370 | 100/100 | A/T | INTRON.6 +235 | 0 | 0.0133 | 0.4983 |
| 6 | 30131711 | rs17194474 | 250/290 | 100/100 | G/C | EXON.1 | 0.0333 | 0.0533 | 0.5724 |
| 6 | 30140540 | rs13212414 | 374/386 | 100/100 | A/T | 3' DOWNSTREAM | 0.0333 | 0.0533 | 0.5724 |
| 6 | 30142674 | rs961039 | 26/30 | 20/28 | G/A | 3' DOWNSTREAM | 0.2112 | 0.2405 | 0.6793 |
| 6 | 30131503 | rs35278640 | 188/194 | 100/100 | T/C | EXON.1 | 0.0133 | 0.0278 | 0.6843 |
| 6 | 30141204 | rs13213365 | 258/268 | 100/100 | C/T | 3' DOWNSTREAM | 0.0266 | 0.0401 | 0.7497 |
| 6 | 30142999 | rs2394737 | 316/324 | 100/100 | A/G | 3' DOWNSTREAM | 0.408 | 0.4195 | 0.9067 |
| 6 | 30142458 | rs757257 | 336/378 | 100/100 | G/A | 3' DOWNSTREAM | 0.4202 | 0.4305 | 0.9071 |
| 6 | 30131123 | rs17188113 | 360/420 | 100/100 | T/G | 5'UTR | 0.0266 | 0.0266 | 1.0000 |
| 6 | 30131331 | rs62407492 | 156/200 | 100/100 | T/C | 5'UTR | 0.0421 | 0.0466 | 1.0000 |
| 6 | 30131527 | rs11961941 | 152/144 | 100/100 | G/T | EXON.1 | 0.0541 | 0.0465 | 1.0000 |
| 6 | 30131585 | rs17194467 | 174/182 | 100/100 | G/A | EXON.1 | 0.0133 | 0.0128 | 1.0000 |
| 6 | 30132035 | rs2074477 | 358/428 | 100/100 | G/A | INTRON.1 +192 | 0.0255 | 0.0333 | 1.0000 |
| 6 | 30132100 | rs41272587 | 352/428 | 100/100 | G/C | INTRON.1 +257 | 0.0266 | 0.0333 | 1.0000 |

Novel rare variants

Ten high quality novel variants were identified in *TRIM15* 'A' and 'B' amplicons (**Table 4.6**). All of which were estimated to have an allele frequency between 0.01 and 0.05. No novel common variants were identified. Using the tabix program (as described in **Methods 2.4.3**), four SNPs (out of the ten) have also been identified by the 1000 genome project; suggesting these SNPs are likely to be genuine. dbSNP rs numbers have yet to be assigned to these novel SNPs.

The average fold coverage for these high quality novel rare variants was calculated as 230 and 242 in the case and control pools, respectively. As a result, all ten novel rare variants acquired maximum power to detect a singleton, which in turn provides confidence in them being genuine SNPs.

None of these novel rare variants however showed significant evidence of association with LOAD ($p < 0.05$). This is perhaps unsurprising, as an association study requires much larger sample size than SNP discovery.

Interestingly, one of the rare SNPs (located at chr6: 30131558) showed an allele frequency of 0.00 in controls and 0.0133 in cases, was found to cause a non-synonymous change (H33Y, histidine->tyrosine at amino acid position 33) and was predicted to be 'probably damaging' by Polyphen-2 (**Methods 2.4.4**).

Furthermore, a single high quality novel SNP (chr6: 30142265) was identified as a deletion of a single 'T' allele out of eight consecutive 'T' repeats. The frequencies of this deletion were estimated to be identical in both pools (**Figure 4.13**).

**Table 4.6 Summary of all high quality novel rare variants identified in *TRIM15* 'A' and 'B' amplicons.** Chromosome number (CHR), base pair position (BP), read coverage (per chromosome) are as indicated as well as power to detect a singleton (Power), observed alleles (reference and alternative), annotation, MAF in case and control pools and Fisher's exact test (2-sided) p-values. The novel rare variant that causes the non-synonymous change and predicted to be 'probably damaging' by Polyphen-2 is highlighted in yellow. SNPs which have been found by the 1000 genome project are highlighted in green. As no SNPs showed significant ($p < 0.05$) evidence of association, the table is presented in ascending order of base pair coordinates.

| CHR | BP | Read coverage (case/control) | Power (case/control) | Allele (ref/alt) | Annotation | MAF in case pool | MAF in control pool | Fisher's exact test (2-sided) p-value |
|---|---|---|---|---|---|---|---|---|
| 6 | 30130617 | 152/212 | 100/100 | G/A | PROMOTER -375 | 0 | 0.0133 | 0.4983 |
| 6 | 30131558 | 174/172 | 100/100 | C/T | EXON.1 | 0.0133 | 0 | 0.4983 |
| 6 | 30131764 | 254/138 | 100/100 | C/G | EXON.1 | 0.0133 | 0.0263 | 0.6843 |
| 6 | 30132314 | 216/298 | 100/100 | T/A | INTRON.1 +471 | 0.0066 | 0.0133 | 1.0000 |
| 6 | 30138476 | 182/134 | 100/100 | T/C | INTRON.5 +82 | 0 | 0.0198 | 0.2475 |
| 6 | 30138597 | 302/316 | 100/100 | T/A | INTRON.5 +203 | 0 | 0.0133 | 0.4983 |
| 6 | 30139155 | 344/426 | 100/100 | C/G | INTRON.6 +369 | 0 | 0.0133 | 0.4983 |
| 6 | 30139396 | 242/264 | 100/100 | T/G | INTRON.6 +610 | 0.0133 | 0.0136 | 1.0000 |
| 6 | 30139477 | 264/238 | 100/100 | G/T | INTRON.6 +691 | 0 | 0.0196 | 0.2475 |
| 6 | 30142265 | 162/220 | 100/100 | T/- | 3' DOWNSTREAM | 0.0133 | 0.0133 | 1.0000 |

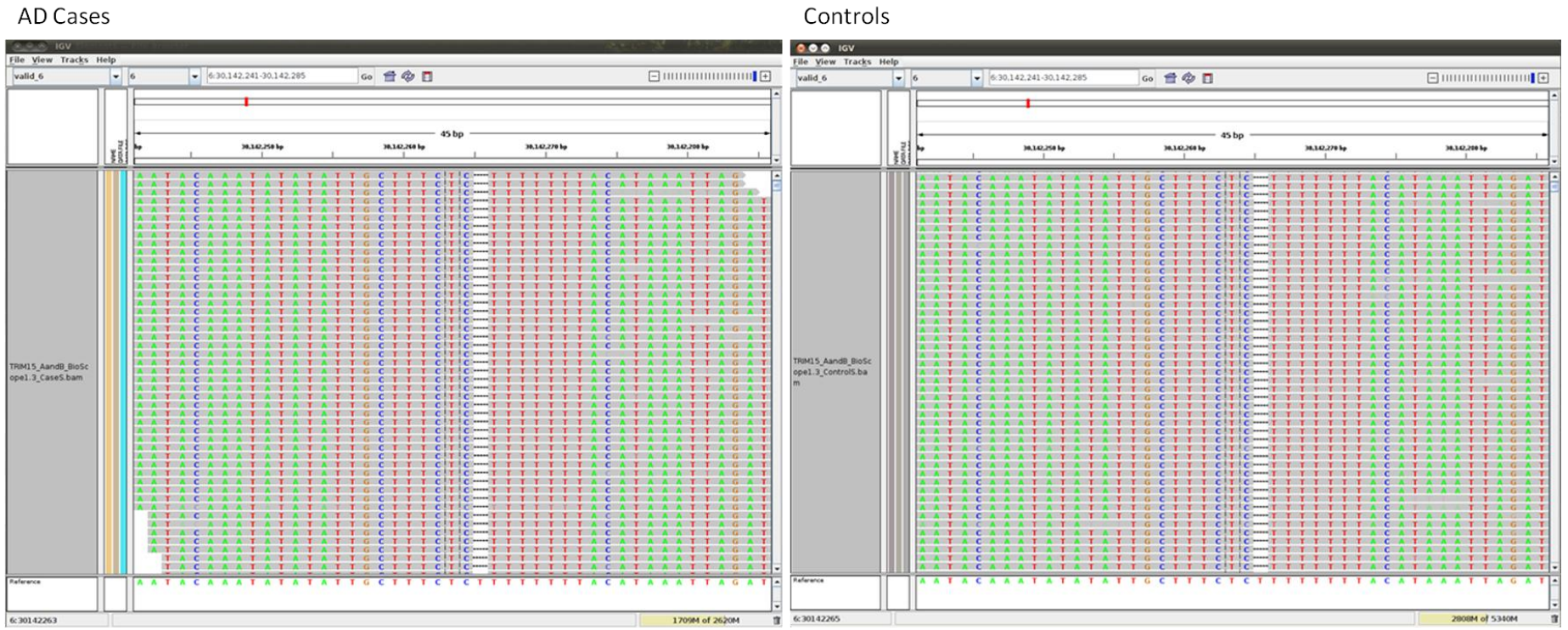



**Figure 4.13 Output from IGV viewer depicting deletion of a 'T' allele from eight consecutive 'T' repeats.** The deletion is as indicated (marked with '---') – case pool on the left and control pool on the right. Nucleotide bases 'A', 'T', 'C' and 'G' are coloured in green, red, blue and brown, respectively. Chromosomal locations (on the top) are shown together with reference human genome sequence (hg19) (at the bottom).

UCSC custom tracks

All high quality SNPs (both common and rare) identified by NGS are displayed in UCSC custom tracks along with comprehensive UCSC annotations (**Figure 4.14** and **Figure 4.15**) (**Methods 2.4.5**).

Repetitive genomic regions are more likely to suffer from power issues due to insufficient coverage and read depth (Treangen and Salzberg, 2011). Interestingly, repeat masked region (using RepeatMasker) does not necessarily imply a drop of power. Only a few genomic regions masked by RepeatMasker demonstrated this loss of read depth and power, whereas the majority acquired full power of detection of a singleton in comparison with non-repeat masked regions.

Furthermore, it is noteworthy that regions that have not been masked by RepeatMaster occasionally show similar loss of power, however, with 'sharp' appearance (instead of 'broad' for repeat-masked regions).

A single region of *TRIM15* 'A' amplicon showed a drop of power to detect a singleton (**Figure 4.14**). On closer inspection (using IGV viewer **(Methods 2.4.5**)), this region was found to be highly repetitive with 'TAAA' repeats (**Figure 4.16**).

Multiple repetitive regions exist for *TRIM15* 'B' amplicon. A number of these regions showed significantly reduction of power and ability to call SNPs (**Figure 4.15**).

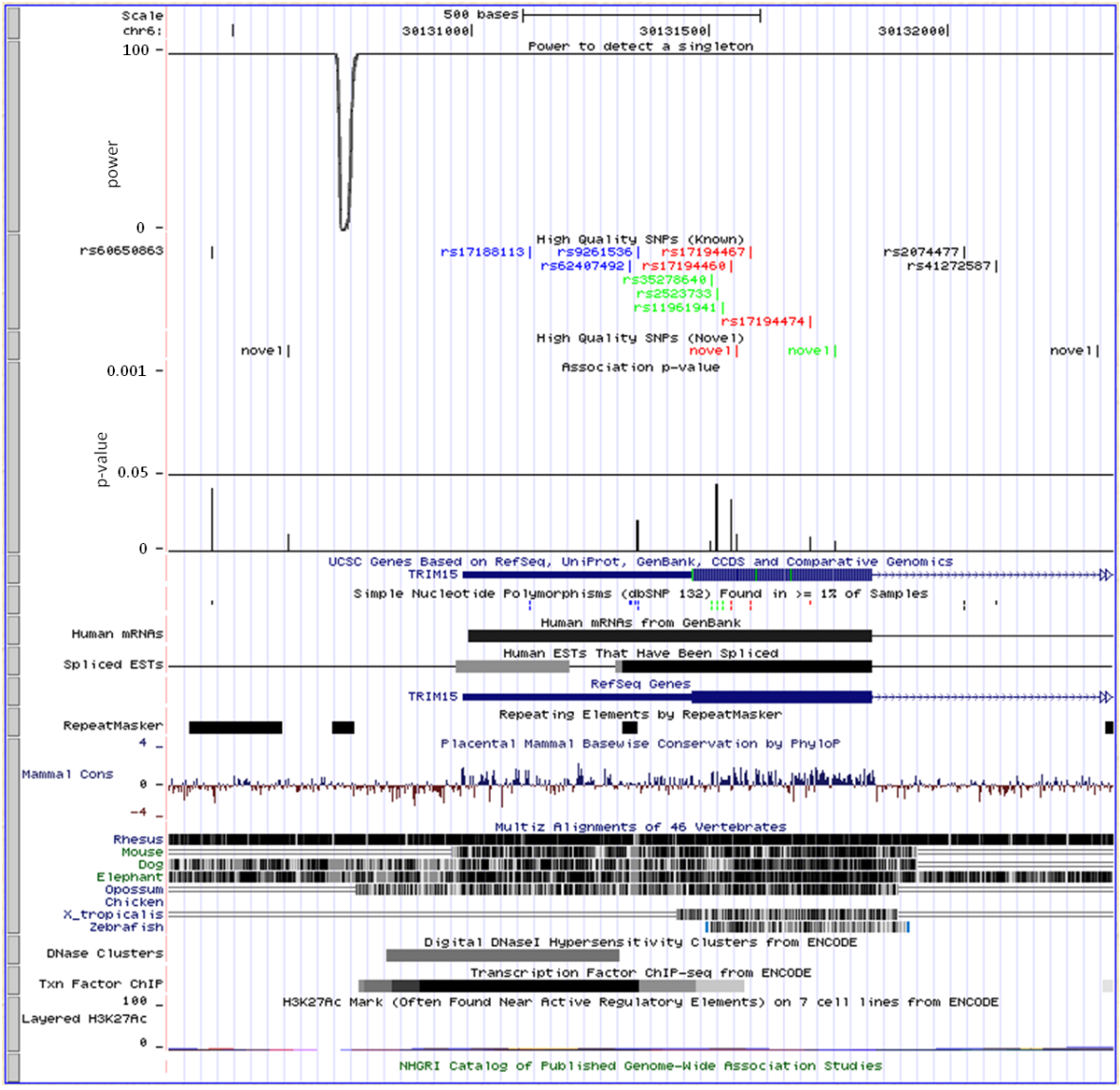



**Figure 4.14 Summary of high quality SNPs identified in *TRIM15* 'A' amplicon.** Power to detect a singleton is shown (at the top) together with high quality SNPs (both known and novel) (in the middle) and association p-values (underneath SNP representation). SNPs are highlighted in colours according to their chromosomal position and their function - red (exon, non-synonymous), green (exon, synonymous), blue (UTR), and black (introns, 5' upstream of *TRIM15* gene). The p-value threshold ($p$ = 0.05) is represented by a horizontal line.

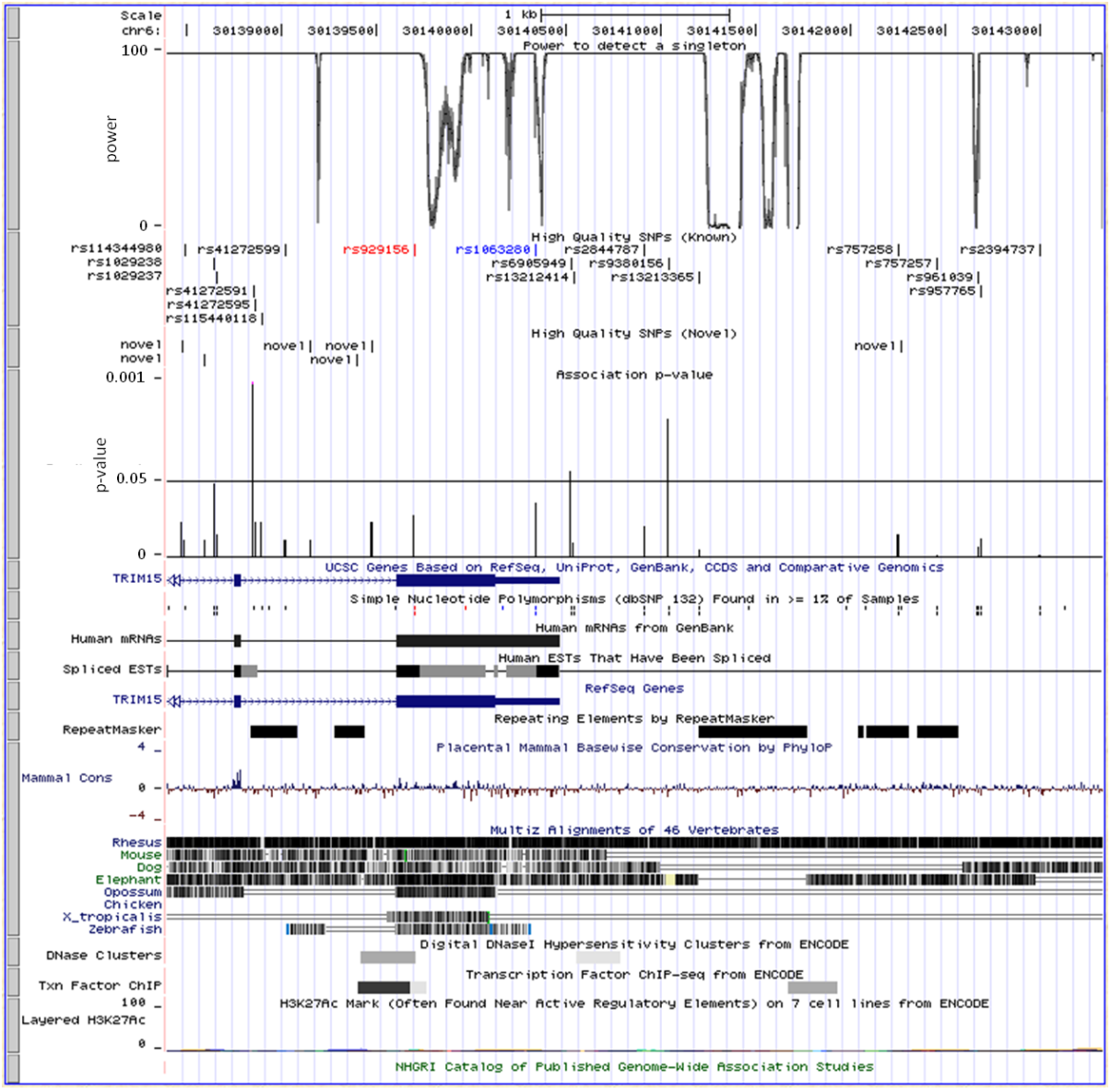



**Figure 4.15 Summary of high quality SNPs identified in *TRIM15* 'B' amplicon.** Power to detect a singleton is shown (at the top) together with high quality SNPs (both known and novel) (in the middle) and association p-values (underneath SNPs representation). SNPs are highlighted in colours according to their chromosomal positions and functions - red (exon, non-synonymous), green (exon, synonymous), blue (UTR), and black (introns, 3' downstream of *TRIM15* gene). The p-value threshold ($p$ = 0.05) is represented by a horizontal line.

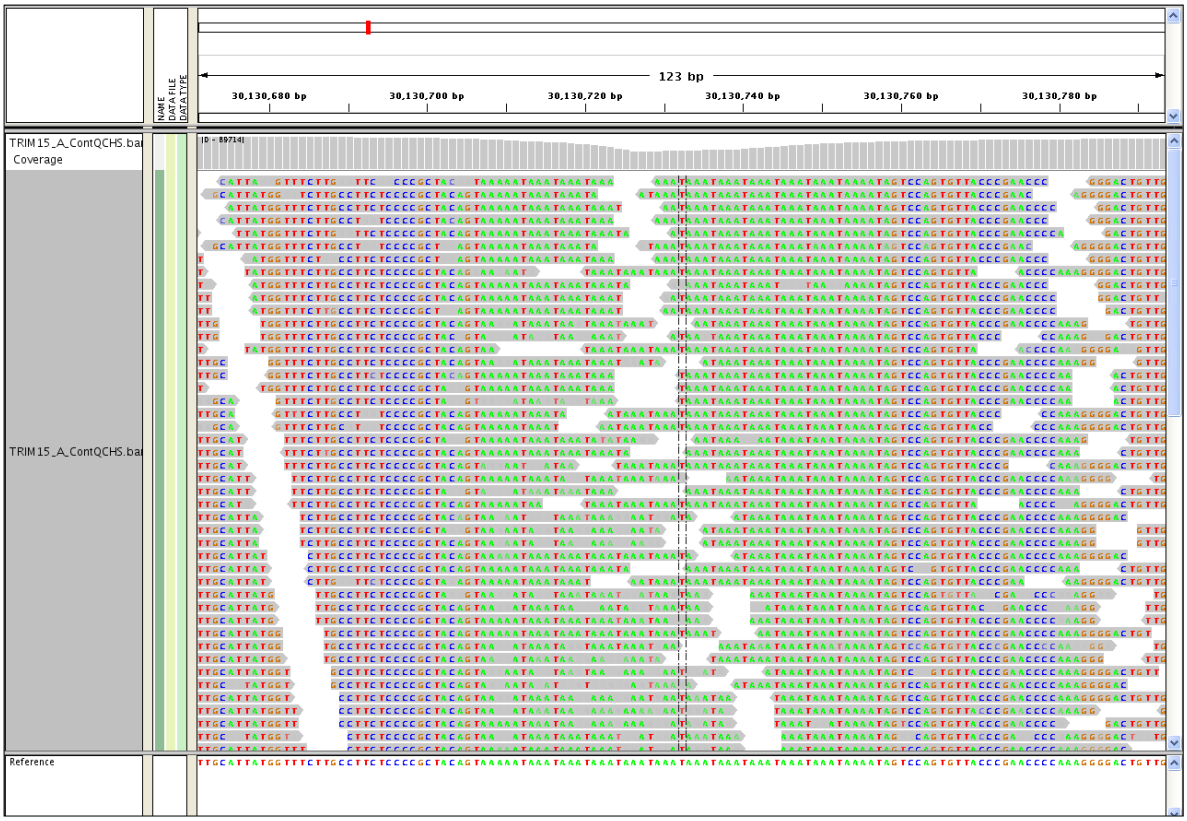



**Figure 4.16 Output from IGV viewer showing 'TAAA' repeats found in the *TRIM15* 'A' amplicon.** Nucleotide bases 'A', 'T', 'C' and 'G' are coloured in green, red, blue and brown, respectively. Chromosomal locations (on the top) are shown together with reference human genome sequence (hg19) (at the bottom).

Overrepresentation of reads

A large number of overrepresented reads (50bp either side of the original amplicons) were observed. This overrepresentation of reads unnecessarily wasted a large number of reads, and thus reduced the throughput and capacity of the next generation sequencing.

This overrepresentation is caused by bias where short nucleotide fragments at both ends of the amplicons are more likely to be presented in the library in comparison to nucleotide fragments located elsewhere (**Figure 4.2**).

Reducing this overrepresentation of reads is thus capable of increasing the average read depth across the targeted interval, and provides more power to call SNPs. A previous study suggested that using 5'-blocked primers in LR-PCR can significantly reduces this overrepresentation (Harismendy and Frazer, 2009).

It should be emphasized that this overrepresentation of reads is only applicable for NGS data using PCR based enrichment, and enrichment using commercial kits such as Agilent SureSelect® are not affected in this manner.

As only a small number of genes were inputted into this NGS project, despite the reads overrepresentation issue, the average read depth and fold coverage is still far greater than what is generally accepted as deep/high coverage; > 20x as suggested by Nielsen et al., 2011 and 42x as depicted by the 1000 Genome Project (The 1000 Genomes Project Consortium, 2010). The average coverage for all *TRIM15* high quality SNPs was ~240-fold in the case pool and ~260-fold in the control pool (**Figure 4.17**).

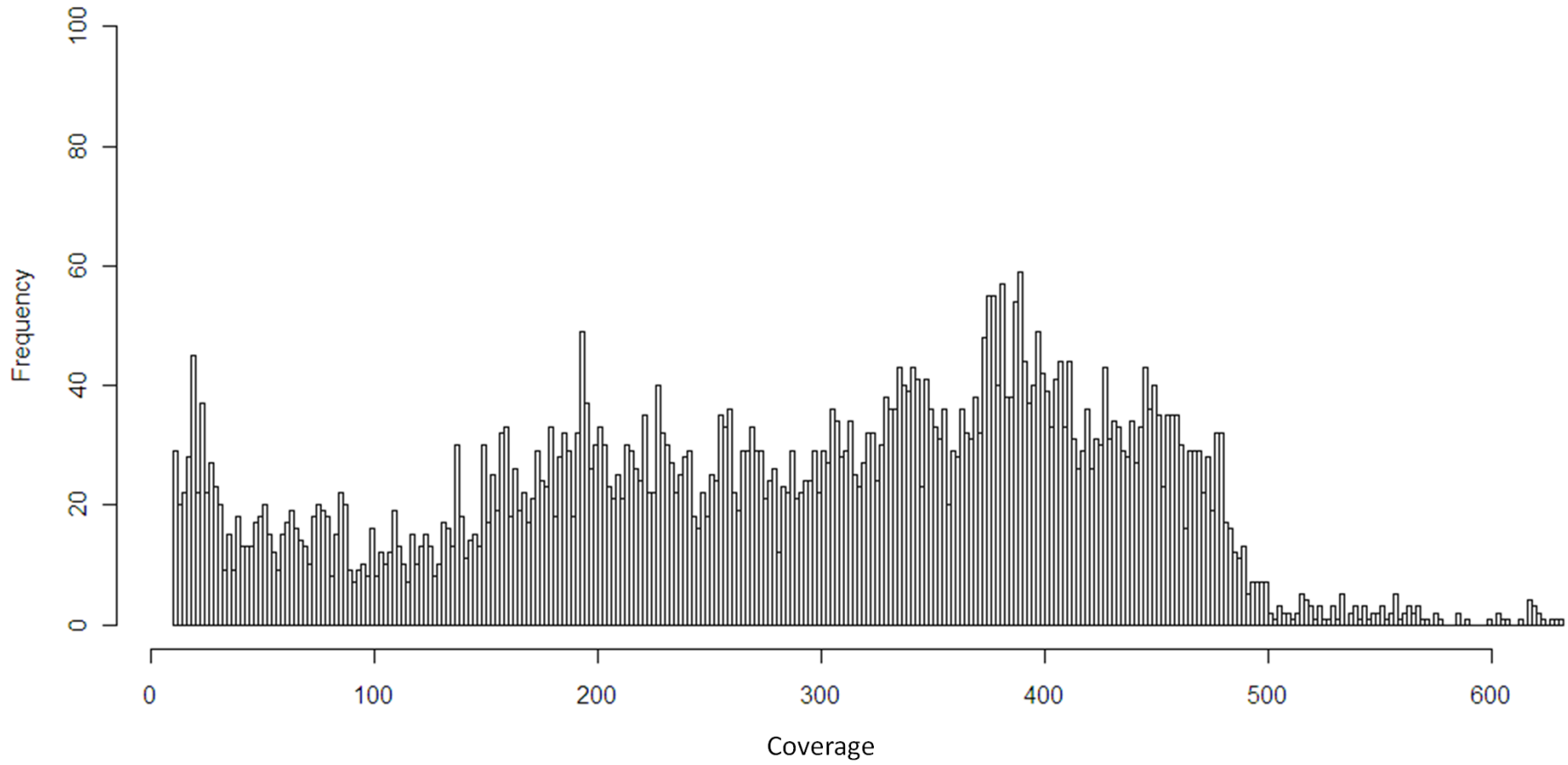



**Figure 4.17 Histogram depicting read coverage in the control pool.** The x and y axis represent the fold coverage and frequency of reads, respectively. The overrepresented regions (i.e. 50bp each side of the amplicon) and base pair positions with less than 10-fold coverage are not shown. As indicated, the majority of *TRIM15* sequence acquired an average read coverage between 200 and 400 fold.

## 4.5 Discussion

41 high quality SNPs were identified within the *TRIM15* 'A' and 'B' amplicons from analysing the next generation sequencing data generated by ABI SOLiD® sequencing.

The sequencing has been successful, as reflected by a number of observations:

- High on-target rate – ~90% of mapped reads were found to be on target (LR-PCR enriched regions).
- Consistent alternative alleles – 100% of known SNPs identified exhibited identical alternative alleles as documented in the dbSNP database.
- Consistent allele frequencies – 96.6% correlation coefficient of allele frequency as compared with HapMap data (14 SNPs)
- High discovery rate – all documented SNPs in HapMap and dbSNP with MAF greater than 0.01 were detected.

It is noteworthy that all allele frequencies estimated in the case and control pool are multiples of 0.0066, which is equivalent to one minor allele out of 150 (i.e. 1/150 = 0.0066). For example, a SNP with three minor alleles would therefore result in a minor allele frequency estimation of 0.0199.

Calculation of Power

This study was estimated to have ~78% power to detect a SNP with minor allele frequency (MAF) 0.01, and ~95% power to detect a SNP with MAF 0.02 (**Table 4.2**).

As only a small number of genomic regions were inputted for NGS, the majority of sequence regions within *TRIM15* 'A' and 'B' amplicons acquired maximum power to detect a singleton. The average read-depth and fold-coverage for *TRIM15* 'A' and 'B'

amplicons were estimated to be 18,382 and 245-fold in cases and 20,543 and 274-fold in controls (**Figure 4.17**).

Discovery of a SNP is dependent upon: i) SNP present in the number of samples chosen to be sequenced and ii) sufficient number of high quality and well mapped reads overlapping the SNP site (i.e. read-depth) (The 1000 Genomes Project Consortium, 2010). As the throughput of NGS per run is fixed, sequencing of more DNA samples would avoid missing variants not represented by these samples, but decreases the read-depth for each individual DNA sample and therefore leads to loss of sensitivity and accuracy.

The power to detect an association with the risk of LOAD is dependent upon: i) effect size carried by the SNP (i.e. odds ratios) and the number of samples sequenced. With 75 matched case and control pairs, the power to detect an association was estimated to be ~80% for common variants with OR ≥ 3 and rare variants with OR ≥ 4 (**Table 4.3**). An OR > 4 for common variants or an OR > 6 for rare variants is required to have 95% power.

As rare variants are more likely to have greater odds ratios in comparison with common variants according to studies of multiple complex disorders (Bodmer and Bonilla, 2008), these variants are therefore more likely to become statistically significant with smaller sample sizes. None of the high quality variants identified within *TRIM15* 'A' and 'B' amplicons exhibited an odds ratio above 6.

Ascertainment of base quality and mapping quality thresholds

The base quality threshold of 10 and mapping quality threshold of 50 were used in this study to call SNPs using Syzygy. FreeBayes was used to validate and confirm the Syzygy outputs (**Methods 2.4.3**).

Base quality and mapping quality thresholds are tightly correlated; a read with low base quality at multiple sites of a read, results in overall low mapping quality for that read.

The same base quality threshold was used in this project as used by the 1000 genome project, and a more stringent mapping quality threshold of 50 (instead of 20) has been used due to the high read-depth of the NGS data (The 1000 Genomes Project Consortium, 2010).

High quality rare variants

Ten novel rare variants with MAFs between 0.01 and 0.05 were identified within the *TRIM15* 'A' and 'B' amplicons, none of which were shown to be associated with LOAD with statistical significance $p < 0.05$ (**Table 4.3**).

Nine out of the ten were found to be single nucleotide polymorphisms (SNPs); six SNPs were found to be located in introns, two in exon 1, and one was shown to be located in a predicted promoter region of the *TRIM15* gene. Of the 2 exonic SNPs, one (located at chr6: 30131558) was identified as a non-synonymous SNP, the other (located at chr6: 30131764) was found to be synonymous.

The non-synonymous SNP appears to be particularly interesting; not only is the SNP predicted to cause a coding change from a positive charged histidine to a neutral tyrosine, but is also present in cases only (with minor allele frequency 0.0133) and absent from the controls. Additionally, this coding change was predicted to be 'probably damaging' with respect to the encoded protein structure by Polyphen-2 (**Methods 2.4.4**).

A deletion variant was also detected; a single 'T' allele out of eight consecutive 'T' repeats with a frequency of 0.0133 in both case and control pools. Interestingly, a known deletion variant rs5875237 has been documented in dbSNP database, at a

neighbouring site (chr6: 30142266) as this variant found by NGS (chr6: 30142265). It is likely that the novel deletion variant found in this study is the same SNP as documented in the database.

Further experimental validation (using alternative genotyping methods) of these SNPs are required to confirm if they are genuine and worthy of further investigation.

High quality common variants

Three common variants (rs41272591, rs9380156 and rs6905949) exhibited significant evidence of associations with LOAD.

Two of the most significant SNPs (rs41272591 and ra9380516 which were found to be in tight LD), have been validated by TaqMan® genotyping assays with allele frequencies similar to those estimated by NGS, and did not show significant evidence of association using an independent sample cohort (90 cases and 91 controls) ($p$ = 0.89, OR = 1.05). It is perhaps unsurprising as the initial p-values do not withstand multiple testing after Bonferroni correction. In a study of 41 independent observations, assuming that 5% would be expected to appear due to chance, two SNPs would be expected to show a significant p-value ($p < 0.05$) and two have been detected.

No further efforts were made to replicate the nominal association seen with SNP rs9380516.

Low quality variants

281 low quality variants were identified using Syzygy (**Methods 2.4.3**). The majority of these low quality variants are likely to be due to errors, as discussed below, and therefore no further efforts were made to validate any of them.

Discussion over errors of NGS using pooled DNA samples

An accurate estimation of allele frequencies using pooled DNA samples is dependent upon equal representation of each individual DNA sample in the library (DNA pool) for sequencing.

Although gel extractions were performed to remove undesired non-specific PCR products and primer dimers, this however was only undertaken after the initial DNA pooling (**Figure 4.7**).

As the exact amount of the DNA amplicon out of the 5µl (which had been pooled) may vary, it is likely that some of the DNA samples are more concentrated than others. Ideally the library is better created by gel extractions of all LR-PCR samples individually, this however means more cost, longer time to perform, and potential to introduce errors.

Barcoding of each individual sample prior to DNA pooling and target enrichment (often with short non-palindromic oligos) enables downstream identification of where the read has originated.

Employing individually bar-coded DNA library samples is an alternative approach to solve equi-molar pooling issues, where estimation of allele frequencies is less sensitive to differential representation of different DNA samples in the library.

NGS using bar-coded DNA libraries results in more accuracy in calling SNPs and prediction of allele frequencies. An alternative allele from individually bar-coded samples would be represented by 50% of reads with the same barcode, whereas by only 0.66% in this study. As a result, differentiating errors (caused by sequencing and read misalignment) from a real SNP would be much easier in bar-coded NGS data.

Furthermore, NGS with bar-codes enables identification of each individual genotype, which is useful in downstream analyses, e.g. imputation, studies of LD between

common and rare variants and meta-analysis taking into account heterogeneity between samples through including covariates. However, applying an individual barcode to each DNA sample is significantly more expensive than using a pooled DNA strategy (which is considered a cost effective approach).

Furthermore, using multiple smaller DNA pools would be better in comparison with using a single large DNA pool. First, each alternative allele would be representing a higher allele frequency, making it easier to be differentiated from errors. Second, validation of genotype data would be easier as sequencing of a smaller number of DNA samples will be required (instead of sequencing all DNA samples in the pool). The above therefore are all considered limitations of using a DNA pooling strategy in next generation sequencing.

NGS is still at its early stage. The highest capacity sequencer currently available requires 8-14 days to produce data, which has limited the application of this technology in clinical diagnosis.

The most anticipated ‘real-time next generation’ sequencer is in trial and under development. One of the examples is nanopore technology, which claims to be able to sequence DNA fragments while they pass through the nanopore. This technology, once developed, is thought to be able to sequence the entire human genome in less than 24 hours for $1000 (Branton et al., 2008).

It is foreseeable that in the not too distant future, NGS technologies will be able to obtain sequencing data from a single cell, allowing investigation of somatic mutations responsible for disease, particularly in cancer genomics (Metzker, 2010).

## 4.6 Conclusion

A next generation sequencing pipeline was described in an effort to investigate if the *TRIM15* gene 'A' and 'B' amplicons harboured multiple rare variants that may be associated with disease. The pipeline described can be applied to other studies that use the ABI SOLiD® next generation sequencing platform.

A total of ten high quality rare variants were successfully identified. Four of which are likely to be genuine as they have also been found by the 1000 genome project (The 1000 Genomes Project Consortium, 2010). The remaining six rare SNPs are novel, and have not been reported before. Direct genotyping using an alternative approach (e.g. Sanger sequencing or TaqMan® genotyping assay) will be necessary to confirm if these SNPs are genuine, and worthy of further investigation.

Furthermore, none of the high quality rare variants identified were found to be significantly associated with LOAD ($p > 0.05$). This is perhaps unsurprising, as an association study would require much larger sample size than SNP discovery.

Although not statistically significant, the rare SNP (located at chr6: 30131558) appears to be interesting; not only as it appear to cause a non-synonymous change in exon 1 of the *TRIM15* gene, but also as it is suggestively associated with LOAD (found only in cases with MAF 1.33% and absent in controls), and predicted to be 'probably damaging' by Polyphen-2. If validated, this SNP may prove to be genuinely associated with AD when larger samples are analysed.

# Chapter 5: Genetic variants influencing human ageing from LOAD Genome-Wide Association Studies (GWAS)

## 5.1 Introduction

Human ageing has long been considered a natural process of deterioration, where an accumulation of damage occurs to DNA molecules, cells and tissues over the life-time. As a result, it causes frailty and malfunctions of various parts of the body, and eventually leads to death. It has become increasingly evident that the human ageing process, like all other biological processes, is subject to regulation by signalling pathways and transcription factors (Kenyon, 2010).

Human ageing and longevity are closely related. There are thoughts that achieving longevity might mean merely adding a few years at the late stage of a life-span. Others thought that increasing the life-span could slow down the ageing process (i.e. people stay young and healthy for longer). The latter has become increasingly accepted. In *C.elegans*, long-lived mutants appear to remain young after normal worms 'look' old by assessing the rate of tissue decline (Garigan et al., 2002; Herndon et al., 2002).

Human ageing is affected by both genetic and environment factors (Cutler and Mattson, 2006). The heritability of ageing is estimated to be 20-30% to reach mid-eighties estimated from twin studies (Herskind et al., 1996). Furthermore, previous studies have shown that siblings of centenarians have an approximately 4-fold higher chance of survival to their early 90s compared with siblings of individuals who die at 73 years of age (Perls et al., 1998). A larger study conducted by the same research group has shown an even higher fold increase (8-18 fold) in the 'risk' of longevity for siblings of centenarians compared with random controls (US 1900 birth cohort) (Perls et al., 1998). Evidence indicates strong familial aggregation towards human ageing.

Given the fact of the likely existence of ageing genes, a large number of studies have been undertaken using different approaches to elucidate the genetic variants contributing to successful ageing and longevity. As a result, various genes have been reported as susceptibility loci for human ageing, such as *PON1, APOE, IL6, IL10, GSTT1*, and *SIRT3* (Glatt et al., 2007; Martin et al., 2007). These candidate genes highlight biological pathways that maybe important in human ageing, such as lipid/cholesterol metabolism [GO:0006629] [GO:0008203] (*APOE* and *PON1*) (de Chaves and Narayanaswami, 2008; Efrat and Aviram, 2010), immune system processes [GO:0002376] (*IL6* and *IL10*) (Jylhava and Hurme, 2009), drug metabolism [KEGG:hsa00982] (*GSTT1*) (Glatt et al., 2007) and energy metabolism in mitochondria (*SIRT3*) (Polito et al., 2010).

GenAge is a database of genes related to ageing (http://genomics.senescence.info/). To date, over 250 genes have been recorded by the GenAge database based on extensive literature reviews. All of these genes have shown possible association with human ageing (de Magalhaes et al., 2009). Most of these genes are playing critical parts in a variety of biological pathways, and a significant number of these genes (>100) are also related to severe human diseases. It is generally believed that genes and bio-markers implicated in age-related diseases such as cancer, coronary artery disease (CAD), cerebra-vascular disease (CVD) and Alzheimer's disease (AD) have a role in successful ageing (Panza et al., 2009; Wang et al., 2009). Identification of genuine ageing genes may uncover 'master genes' that increase our understanding of many age-related diseases.

The molecular genetics underlying the human ageing process is complex and it is suggested that successfully ageing is likely due to numerous genes and environmental factors, each exerting a small effect (Lescai et al., 2009; Plomin et al., 2009).

Insights into human ageing have been gained from studying model organisms. Extension of lifespan can be achieved by manipulating a few genes in laboratory animals, such as flies, worms and mice (Kenyon, 2010). The insulin/IGF-1 (insulin-like growth factor 1) signalling pathway has a well-established role in influencing lifespan within model organisms with large effect (Clancy et al., 2001; Holzenberger et al., 2003; Kenyon et al., 1993). Genetic inactivation of the *daf-2* gene (encoding the IGF-1 receptor homolog in *C.elegans*) increases the lifespan of *C.elegans* by approximately 100% (Sebastiani et al., 2009). Interestingly, there is emerging evidence that genes such as *IGF1/IGF1R* (the orthologues of which play a major part in ageing in animals) can play a role in human lifespan. Loss of function mutations in *IGF1R* have been found to be overrepresented in Ashkenazi Jewish centenarians compared with controls (Suh et al., 2008).

Ageing genes in human may not only increase the life-span but also postpone age-related diseases. A previous study has indicated a significantly decreased prevalence of age-related diseases in offspring of long-lived parents (hypertension by 23%, diabetes mellitus by 50%, heart attacks by 60%, and no incidences of strokes) compared with several age-matched control groups (Atzmon et al., 2004).

Characterising various genetic and environmental factors influencing human life-span is one of the world's major scientific challenges (Jylhava and Hurme, 2009). To date, GWAS are one of the most widely adopted approaches for identifying common genetic variations associated with human diseases. It has been suggested that with increasing sample size, promising signals of association between human traits and genetic variants can be revealed (McCarthy et al., 2008).

## 5.2 Aims

Age is one of the biggest risk factor for many age-related diseases including Alzheimer's disease. The prevalence of Alzheimer's disease rises from less than 0.6% in persons aged between 65 and 69 years to over 22% in persons aged 90 years and older (average age-at-onset: 84) (Corder et al., 1993; Lobo et al., 2000). The risk of developing AD approximately doubles every five years after the age of 65 (Feulner et al., 2009).

The aim of this study was to i) investigate whether the 'known' LOAD genes play a role in human ageing, and ii) search for candidate genetic risk factors associated with human survival and ageing, which may merit further study.

## 5.3 Strategy

Through collaborative efforts, a combined GWAS dataset (subject-level genotype data) was generated from 1,385 subjects (1,047 LOAD cases and 338 controls) with documented age-at-death (AAD). All of these data were subject to subsequent QC procedures and analysis. The data analysis was performed using PLINK (**Methods 2.3.1**).

Datasets merging and QC

The data was obtained from nine research centres, three from the USA and six from the UK (ARUK consortium). All studies used the Illumina 610 QuadChip, except the Mayo data which used the Illumina HumanHap300 chip (Carrasquillo et al., 2009). The Illumina 610 QuadChip includes all SNPs presented in Illumina 300 chip, which enabled merging of all the datasets (**Table 5.1**).

Individual data characterized as 'AUT - autopsy' or having AAD information were extracted from the Mayo dataset (Carrasquillo et al., 2009) using the '--keep' and '--make-bed' command in PLINK. This was repeated for samples from ARUK GWAS data (Nottingham, Bristol, Manchester, Belfast, Oxford and London), National Institute of Mental Health (NIMH) and Washington University (WashU) where possible. All GWAS datasets were transformed into the same PLINK format (1 and 2 coding in PLINK binary format). Any samples which overlapped between GWAS datasets were removed. Each sample was checked individually for discrepancies between AAD and age at sampling (AAS). Samples with AAS greater than AAD were removed from further analysis. Data merging was performed using (--bmerge) in PLINK under "Consensus call" mode (shown as merged data 1 in **Figure 5.1**).

**Table 5.1 Summary of sample information.** GWAS data obtained from a total of nine centres, three from the USA (NIMH, WashU and Mayo) and six from the UK (Nottingham, Bristol, Manchester, Belfast, Oxford, London). Sample size, number of cases and controls, males and females for each cohort are as indicated together with details of the mean age at death in years for cases and controls and genotyping chip used in each study.

| Dataset | Sample Size | AD Status (AD/Controls) | Gender (Male/Female) | Genotyping Chip | Mean AAD (AD/Controls) | Origin |
|---|---|---|---|---|---|---|
| Mayo | 434 | 220/214 | 246/188 | Illumina 300 | 73.5/71.7 | USA |
| NIMH | 46 | 46/0 | 12/34 | Illumina 610 | 78.1/- | USA |
| WashU | 332 | 294/38 | 140/192 | Illumina 610 | 84.1/86.1 | USA |
| Belfast | 235 | 213/22 | 99/136 | Illumina 610 | 82.2/83.1 | UK |
| Bristol | 59 | 43/16 | 21/38 | Illumina 610 | 82.5/81.6 | UK |
| London | 238 | 194/44 | 83/155 | Illumina 610 | 86.1/83.0 | UK |
| Manchester | 1 | 1/0 | 0/1 | Illumina 610 | 79.0/- | UK |
| Nottingham | 39 | 35/4 | 18/21 | Illumina 610 | 83.6/79.5 | UK |
| Oxford | 1 | 1/0 | 0/1 | Illumina 610 | 79.5/- | UK |
| **Pooled** | 1385 | 1047/338 | 619/766 | - | 81.5/76.1 | - |

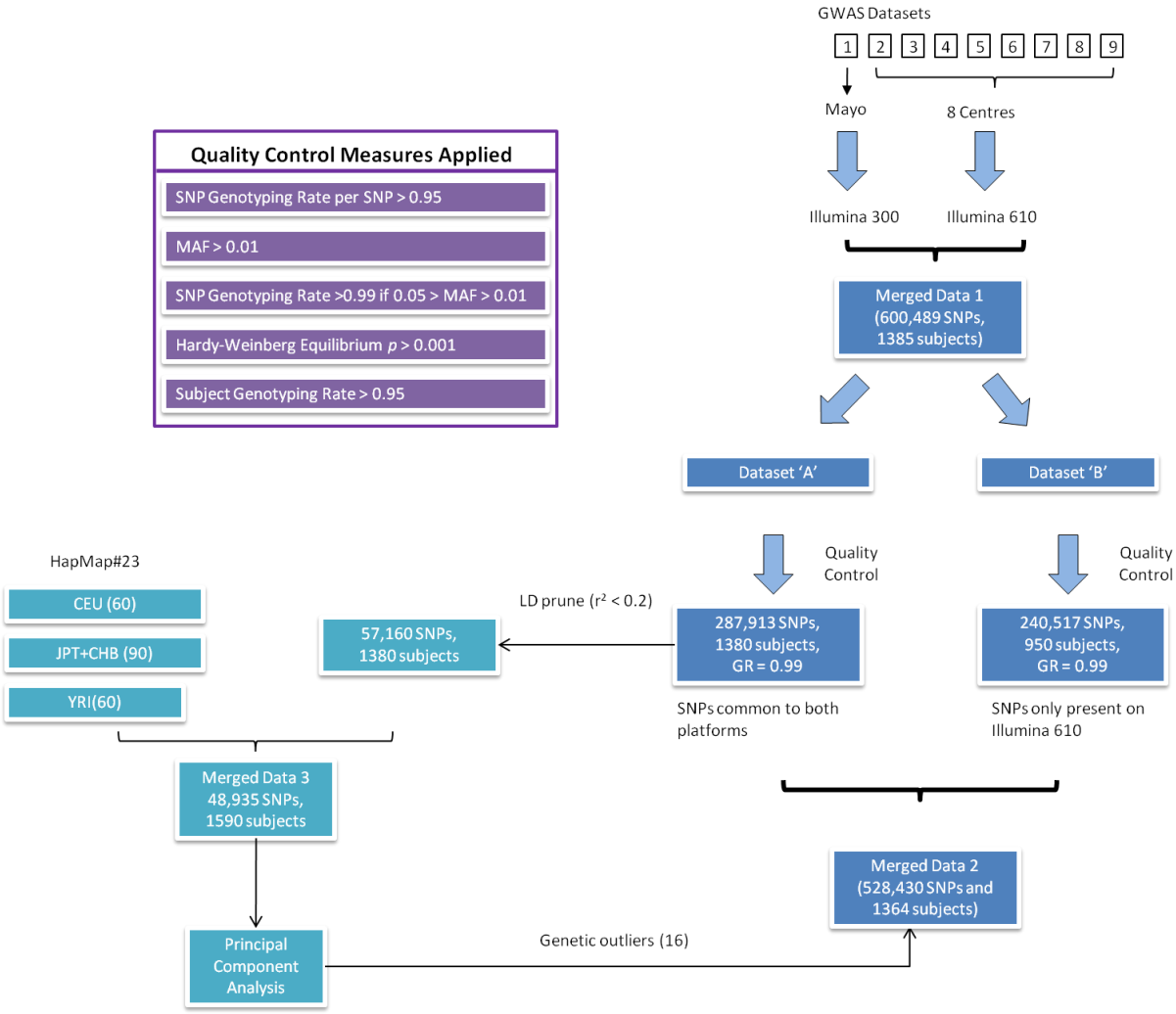



**Figure 5.1 GWAS data QC and data merging strategy.** Flow diagram summarizing the processes undertaken for data preparation and QC prior to subsequent analyses. The data was merged together under PLINK "Consensus call" mode (Merged Data 1). This data was split into two groups (Dataset 'A' and 'B') according to genotyping rate (SNPs which had > 95% genotyping rate for all samples (both chips) and the rest of the SNPs with genotyping rate > 95% for samples typed on the Illumina 610 chip). Both of these groups were subject to QC separately. The two datasets were then merged (Merged Data 2). Dataset 'A' (which contained SNPs common to both platforms) was LD pruned and merged with HapMap data (CEU, CHB/JPT and YRI) to form 'Merged Data 3'. This was then used in a Principal Components Analysis which revealed 16 individuals as genetic outliers. These were removed from 'Merged Data 2'. GR – genotyping rate

QC procedures were undertaken for the merged data to account for population stratification and differences in Illumina chip versions. Data merging and QC procedures are illustrated in **Figure 5.1**. The merged data was separated into two GWAS datasets ('A' and 'B') using the '--geno 0.05' command in PLINK.

For both of the GWAS datasets, the following QC procedures were carried out in order.

1) SNPs with a genotyping rate less than 0.95 (--geno 0.05) were excluded from further analysis.

2) SNPs with a minor allele frequency (MAF) less than 0.01 (--maf 0.01) were excluded from further analysis.

3) A list of SNPs with MAF's between 0.01 and 0.05 was generated (--freq). Within this shortlist, SNPs with genotyping rate less than 0.99 (--geno 0.01) were excluded (--exclude) from further analysis.

4) SNPs with a Hardy-Weinberg Equilibrium p-value less than 0.001 (--hwe 0.001 --hwe-all) were excluded from further analysis, irrespective of status (AD cases or controls).

5) Individuals with a genotyping rate less than 0.95 (--mind 0.05) were excluded from further analysis.

6) Using the GWAS dataset 'A', a LD pruned subset of 57,160 SNPs common to all arrays and HapMap data (--indep-pairwise) was generated using a PERL script written 'in-house' (**Appendix 8.4.1**). No two SNPs within this list had a LD $r^2$ value greater than 0.2 across sliding windows (window size of 1,500 SNPs and 150 SNPs to shift the window). The subset of SNPs was used by EIGENSTRAT (Price et al., 2006) for following calculations:

- To detect genetic outliers (See **Methods 2.3.9** for definition),
- To calculate principal components (PCs),
- To generate a population stratification plot (as described in **Methods 2.3.9**)

7) Genomic control inflation factor ($\lambda$) was calculated using the GWAS dataset 'A' with AAD using EIGENSTRAT (**Methods 2.3.9**).

8) Extraction of PCs from EIGENSTRAT results.

Principal component analysis (PCA) using EIGENSTRAT reduces the genotype data to a number of dimensions, which are defined as the top eigenvectors of a covariance matrix between samples (Price et al., 2006). All significant PC axes ($p < 0.05$) were taken into account in the analysis. The number of PC axes to be included as covariates was ascertained using the method as described (**Methods 2.3.9**).

After the above QC, dataset 'A' consisted of SNPs common to both Illumina 300 and Illumina 610 chips, whereas dataset 'B' consisted of SNPs only common to Illumina 610 chips. The two GWAS datasets were merged using the same methods (--bmerge) as described (shown as merged data 2 in **Figure 5.1**).

SNPs with allele frequency bias due to inter-chip and inter-cohort differences can cause inflation of type I error rate. A box & whisker plot was drawn using StatsDirect (v 2.7.8). Only two centres (Mayo and NIMH) showed significant differences in AAD range compared with the rest of the data. Therefore, two logistic regression tests were undertaken using WashU data as a control and the Mayo and NIMH data as cases. The test incorporated the top six PCs and AAD as covariates. For each comparison, a Q-Q plot of $\chi 2$ of observed versus expected p-values was generated using GenABEL (v1.6.4) (Aulchenko et al., 2007) (**Methods 2.3.9**).

Investigation of effect of LOAD genes in ageing

Ten known LOAD susceptibility genes (*APOE, CLU, PICALM, CR1, BIN1, ABCA7, MS4A6A, CD33, CD2AP* and *EPHA1*) were tested for association with ageing using the most significant SNPs found in the previous GWAS (Harold et al., 2009; Hollingworth et al., 2011; Lambert et al., 2009; Naj et al., 2011; Seshadri et al., 2010). The best proxy was used to inform the effect of SNPs if they were not present in the merged dataset (**Table 5.2**).

**Table 5.2 SNP selection.** The most significant SNPs found in previous LOAD GWAS were selected for testing their effects with human lifespan. The SNPs, associated genes and the GWAS study are as indicated. LD - linkage disequilibrium shown in $r^2$. '-' indicates the original index SNP was used rather than a proxy.

| SNP used in this study | Gene | SNP cited in literature | LD ($r^2$) | Literature |
|---|---|---|---|---|
| rs2075650 | APOE | rs2075650 | - | Harold et al., 2009 |
| rs11136000 | CLU | rs11136000 | - | Harold et al., 2009 |
| rs3851179 | PICALM | rs3851179 | - | Harold et al., 2009 |
| rs3818361 | CR1 | rs3818361 | - | Hollingworth et al., 2011 |
| rs744373 | BIN1 | rs744373 | - | Hollingworth et al., 2011 |
| rs3764650 | ABCA7 | rs3764650 | - | Hollingworth et al., 2011 |
| rs610932 | MS4A6A | rs610932 | - | Hollingworth et al., 2011 |
| rs3865444 | CD33 | rs3865444 | - | Hollingworth et al., 2011 |
| rs1485780 | CD2AP | rs9349407 | 0.913 | Hollingworth et al., 2011 |
| rs11767557 | EPHA1 | rs11767557 | - | Hollingworth et al., 2011 |

Genome-wide association studies

Quantitative Trait (QT) analysis of all SNPs was performed using multivariable linear regression (--linear) adjusted for AD status, gender and the top six PCs (**Methods 2.3.1**). AAD was used as a continuous trait in this analysis thus giving maximum statistical power. Manhattan plots were drawn using Haploview (v 4.1) to visualize GWAS results (**Methods 2.3.5**). A histogram of AAD of all individuals that passed QC was drawn using StatsDirect software (v 2.7.8).

Each SNP was annotated with its gene name using the PLINK gene report function (**Methods 2.3.1**) and a PERL script developed 'in-house' (**Appendix 8.4.4**).

Genotyping of SNP rs4110518

SNP rs4110518 was genotyped using an independent sample cohort (n = 487) using TaqMan® genotyping assay (**Methods 2.2.7**).

Power calculation

Power calculations were undertaken using QUANTO v1.2.4 (**Methods 2.3.8**). The required sample size was estimated using an additive model created by the software.

MAF analysis of SNPs responsible for LOAD

The full range of AAD (58-108 years) was separated into five age-at-death categories, the boundaries of which were selected to ensure each group contains an approximately equal number of samples. This was carried out using "Grouping => Categorise" function in StatsDirect. For each of the LOAD gene loci, the allele frequency was calculated and stratified by AAD category.

The separation into five age-at-death categories were used only to facilitate visualisation of MAF of candidate SNPs in the different age ranges, and was not used to generate p-values.

## 5.4 Results

Dataset composition and QC

The combined GWAS dataset had a sample size 1,385 before QC. After QC, four participants were removed for low genotyping rate (--mind 0.05) from Mayo GWAS (Carrasquillo et al., 2009). A single sample from Bristol was removed due to discrepancies between AAD and age-at-onset (AAO). An additional 16 samples were removed as genetic outliers by PCs analysis using EIGENSTRAT. This included 14 samples from the Mayo data, one from NIMH and one from Belfast. The mean AAD in the pooled dataset (post QC) was greater than 80 years of age (**Table 5.1**).

The AAD histogram follows a normal distribution (**Figure 5.2**), with mean AAD of 80.2 years of age (SD = 8.9 years). The box and whisker plot depicting AAD of each centre was as illustrated (**Figure 5.3**).

The multidimensional scaling plot (MDS) demonstrated three distinct clusters. As expected, each cluster represents different population ancestry - European (CEU), Asian (CHB and JPT), and Yoruban (YRI) (**Figure 5.4**). UK, USA and HapMap_CEU samples formed a single cluster. On closer inspection, slight deviation between UK and USA samples exists and this was accounted for by including PCs as covariates.

Genomic control inflation factor ($\lambda$) was calculated using EIGENSTRAT by iteratively including zero to ten PCs (Price et al., 2006). Including six PCs generated the lowest genomic control inflation factor ($\lambda$ = 1.003) (**Table 5.3**).

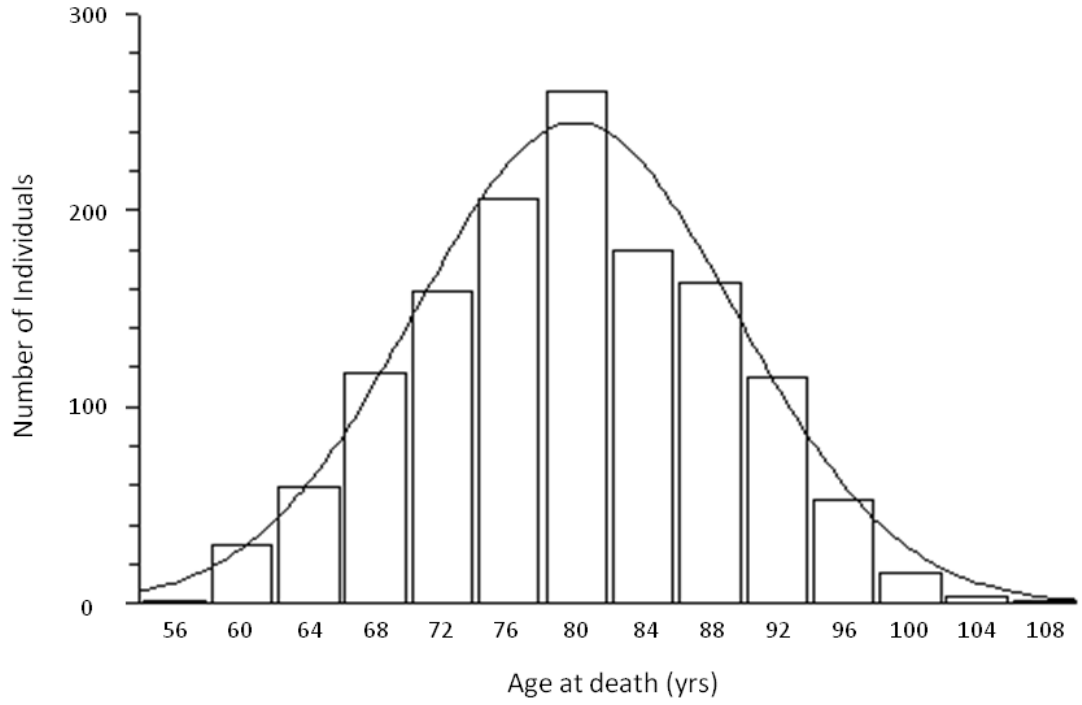



**Figure 5.2 Histogram plot representing the spread of AAD of samples included in this study.** The x and y axis represents age-at-death (AAD) in years and number of individuals, respectively. This graph follows a normal distribution, with mean AAD 80.2 years (n = 1,385).

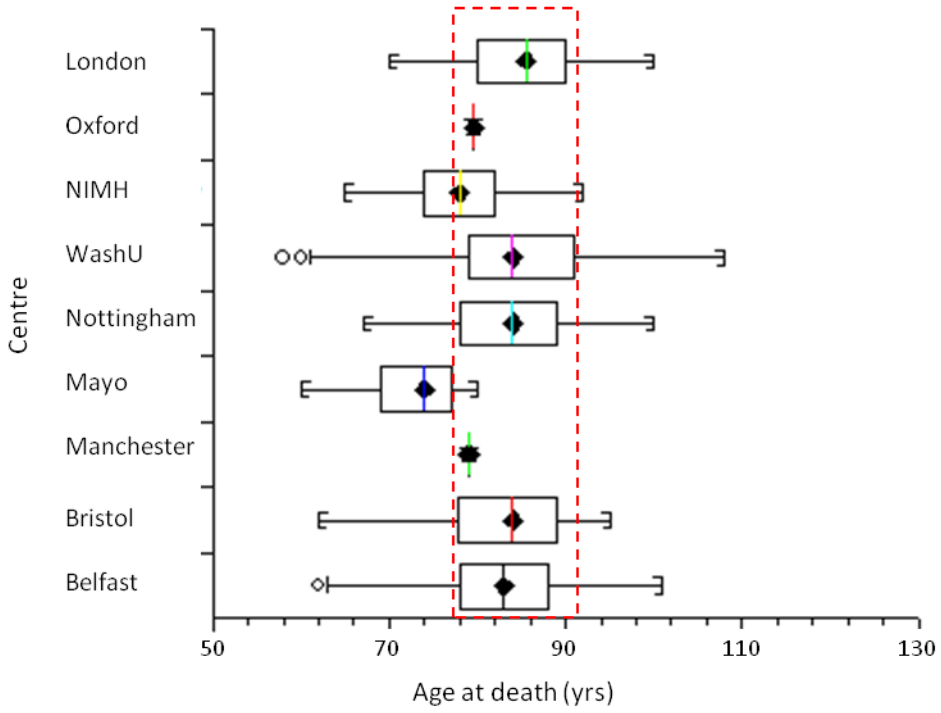



**Figure 5.3 Box and whisker plot, showing the Age at Death (AAD) distribution for each centre.** The central box represents the distance between the first and third quartiles with median marked with a diamond. The circles indicate that an individual's AAD is outside 2 times the interquartile range. The dashed rectangle highlights that the majority of the data have a similar range of AAD with the exception of the NIMH and Mayo data.

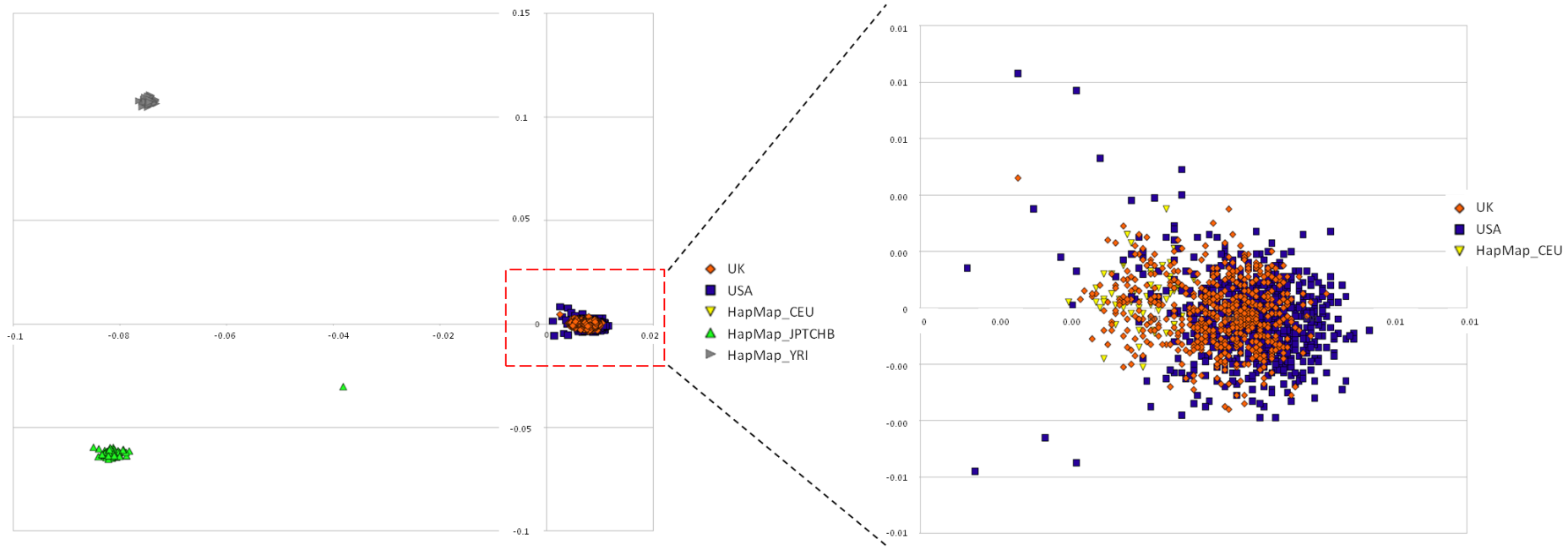



**Figure 5.4 Multi-dimensional scaling (MDS) plot depicting the Principal Component Analysis of Merged Data 3.** Population stratification was tested using HapMap data #23 as reference. UK and USA and HapMap CEU samples formed a single cluster (shown inside the dashed rectangle). One HapMap individual from the Asian samples appears to have dual ethnicity. The diagram inset shows a magnified section of UK, USA and HapMap CEU samples.

**Table 5.3 Calculation of genomic control inflation factor (λ).** Table showing genomic control inflation factor (λ) calculated before and after taking into account population stratification. The top six PCs were used in subsequently analyses as covariates (highlighted yellow).

| Number of top PCs adjusted | λ (before) | λ (after) |
|---|---|---|
| 0 | 1.020 | 1.020 |
| 1 | 1.020 | 1.018 |
| 2 | 1.020 | 1.018 |
| 3 | 1.020 | 1.007 |
| 4 | 1.020 | 1.008 |
| 5 | 1.020 | 1.006 |
| 6 | 1.020 | 1.003 |
| 7 | 1.020 | 1.003 |
| 8 | 1.020 | 1.003 |
| 9 | 1.020 | 1.004 |
| 10 | 1.020 | 1.005 |

It was noted that there is a difference in AAD between LOAD cases (mean AAD = 81.63 years) and controls (mean AAD = 76.09 years), and similarly between male (mean AAD = 77.93 years) and female (mean AAD = 82.17 years). ANOVA tests of the variance of AAD between these groups were found to be significant ($p < 0.001$). This confirmed that AD status and gender were appropriate covariates.

After stringent QC, there were 1,364 samples (1,031 LOAD cases and 333 controls, 608 male and 756 female) and 528,430 SNPs remaining for further analysis.

Analysis and results

Assessment of the ten LOAD susceptibility genes yielded compelling evidence of association between *APOE* locus (rs2075650) and human ageing (*uncorr-p* = $5.27 \times 10^{-4}$), which withstood multiple testing after Bonferroni correction for ten independent tests (**Table 5.4A**).

In addition to examining the association of these ten LOAD susceptibility genes with ageing, analyses including all SNPs on the Illumina 610 chip (post QC) were undertaken. The genome-wide significance threshold was calculated ($p = 1.04 \times 10^{-7}$) using Bonferroni correction for the number of independent tests (N = 483,066), which was estimated using **Methods 2.3.3**.

No variants appear to be associated with ageing with a genome-wide level of significance ($p < 1.04 \times 10^{-7}$). There were 41 SNPs with p-value ($p \leq 5 \times 10^{-5}$). These SNPs span the genome, representing 35 distinct signals (pairwise $r^2 \leq 0.8$) across 13 chromosomes. 24 of them are located within 20kb of known human genes with a wide range of functions. SNPs with p-value ($p < 5 \times 10^{-5}$) are shown in **Table 5.4B**. These signals are at best tentative but may merit study in larger sample sets.

**Table 5.4 Summary of results.** Table showing the results of the analysis of A) the ten documented LOAD susceptibility loci (*APOE, CLU, PICALM, CR1, BIN1, ABCA7, MS4A6A, CD33, CD2AP* and *EPHA1*) with human ageing. B) SNPs (rs number [major/minor allele]) with association p-value < 5 x $10^{-5}$. Chromosome number (CHR), base pair position (BP) and gene name (Gene) is shown together with p-value (P) and direction of effect. '+' indicates the minor allele of any given SNP is protective, whereas '-' means the minor allele of the SNP has a detrimental effect on ageing. In table B), the gene name is shown if the SNP is within 20kb of a known gene.

**A)**

| SNP [major/minor] | CHR | BP | Gene | P | Direction of Effect |
|---|---|---|---|---|---|
| rs2075650[T/C] | 19 | 50087459 | APOE | 5.27E-04 | - |
| rs3764650[A/C] | 19 | 997520 | ABCA7 | 1.35E-01 | - |
| rs610932[C/A] | 11 | 59695883 | MS4A6A | 1.76E-01 | + |
| rs3851179[G/A] | 11 | 85546288 | PICALM | 2.27E-01 | + |
| rs11767557[A/G] | 7 | 142819261 | EPHA1 | 2.67E-01 | - |
| rs3865444[C/A] | 19 | 56419774 | CD33 | 4.42E-01 | + |
| rs3818361[G/A] | 1 | 205851591 | CR1 | 5.63E-01 | + |
| rs1485780[A/C] | 6 | 47664589 | CD2AP | 6.04E-01 | - |
| rs744373[A/G] | 2 | 127611085 | BIN1 | 6.89E-01 | + |
| rs11136000[G/A] | 8 | 27520436 | CLU | 9.75E-01 | - |

**B)**

| SNP [major/minor] | CHR | BP | Gene | P | Direction of Effect |
|---|---|---|---|---|---|
| rs987839[T/C] | 12 | 21266105 | SLCO1B1 | 3.19E-06 | + |
| rs17205854[G/A] | 5 | 64458658 | | 3.74E-06 | + |
| rs17811551[T/C] | 5 | 64462993 | ADAMTS6 | 3.74E-06 | + |
| rs1857821[A/G] | 4 | 77101003 | NAAA | 5.12E-06 | + |
| rs7525717[G/A] | 1 | 56700226 | | 5.76E-06 | - |
| rs4673651[A/G] | 2 | 212712848 | ERBB4 | 6.08E-06 | - |
| rs17049647[G/T] | 2 | 130093855 | | 7.12E-06 | + |
| rs10518142[G/T] | 4 | 77061898 | NAAA | 8.89E-06 | + |
| rs2444861[A/G] | 8 | 99170108 | C8orf47 | 8.99E-06 | + |
| rs1418425[G/A] | 1 | 111270409 | | 9.64E-06 | - |
| rs2271528[C/T] | 4 | 77107860 | SDAD1 | 1.02E-05 | + |
| rs1555453[A/C] | 9 | 27316780 | MOBKL2B | 1.07E-05 | + |
| rs13111494[A/G] | 4 | 77204512 | ART3 | 1.28E-05 | + |
| rs12740413[C/T] | 1 | 16388466 | ARHGEF19 | 1.30E-05 | - |
| rs3210458[C/T] | 3 | 142494320 | ACPL2 | 1.52E-05 | + |

| | | | | | |
|---|---|---|---|---|---|
| rs4859571[G/A] | 4 | 77076333 | NAAA | 1.56E-05 | + |
| rs7803143[T/C] | 7 | 6211011 | PSCD3 | 1.57E-05 | - |
| rs7622678[T/C] | 3 | 198798845 | BDH1 | 1.93E-05 | - |
| rs4720752[G/A] | 7 | 7735965 | RPA3 | 1.93E-05 | - |
| rs6962026[C/T] | 7 | 6213048 | PSCD3 | 2.07E-05 | - |
| rs10901296[C/T] | 9 | 132755477 | ABL1 | 2.37E-05 | + |
| rs680109[C/A] | 11 | 105255919 | GRIA4 | 2.52E-05 | + |
| rs1537438[G/T] | 13 | 26806973 | | 2.53E-05 | + |
| rs11206814[C/T] | 1 | 56690636 | | 2.67E-05 | - |
| rs12562047[A/C] | 1 | 164095780 | UCK2 | 2.70E-05 | + |
| rs2710548[A/G] | 4 | 126492980 | FAT4 | 2.79E-05 | - |
| rs6454676[C/T] | 6 | 88934174 | | 2.90E-05 | + |
| rs17047650[C/T] | 3 | 68547103 | FAM19A1 | 3.19E-05 | + |
| rs10433502[C/T] | 3 | 68569792 | FAM19A1 | 3.49E-05 | + |
| rs10485170[T/C] | 6 | 88939371 | | 3.67E-05 | + |
| rs12257410[A/C] | 10 | 13832528 | FRMD4A | 3.77E-05 | + |
| rs3125524[C/T] | 10 | 133104931 | | 3.83E-05 | + |
| rs4280854[A/G] | 5 | 105923201 | | 3.85E-05 | + |
| rs1037381[A/G] | 2 | 105669675 | | 3.99E-05 | - |
| rs6532496[A/G] | 4 | 95799427 | PDLIM5 | 4.00E-05 | - |
| rs17618813[G/A] | 4 | 114153483 | ANK2 | 4.16E-05 | + |
| rs7103504[G/A] | 11 | 99006474 | CNTN5 | 4.39E-05 | - |
| rs10085518[C/T] | 7 | 6252040 | PSCD3 | 4.59E-05 | - |
| rs4686837[G/A] | 3 | 188222371 | ST6GAL1 | 4.60E-05 | + |
| rs6491207[T/C] | 13 | 26828900 | | 4.63E-05 | + |
| rs7952321[G/T] | 11 | 55539349 | OR5AS1 | 4.68E-05 | - |

Without conducting logistic regression comparison, initial analysis of the association study suggested two genome-wide significant SNPs - rs4110518 ($p = 5.96 \times 10^{-9}$) and rs2944476 ($p = 2.19 \times 10^{-8}$). Comparing SNPs between NIMH and WashU data showed no significant difference in allele frequency, whereas five SNPs showed significant difference in allele frequency comparing Mayo data with WashU data after taking into account population stratification (i.e. PCs) and AAD. These five SNPs are rs4110518, rs2944476, rs10460926, rs10953303, rs7172278 (**Figure 5.5** and **Figure 5.6**).

It is perhaps unsurprising that the two SNPs which showed genome-wide level of significance overlap with the five SNPs that showed significant bias between Mayo and WashU data, as the AAD of the Mayo data is significantly younger (as previously described). It is not possible to correct for centre, as the spread of the AAD is considered crucial in detecting genuine ageing associated variants. The difference in allele frequency due to samples with young AAD in Mayo and old AAD in WashU may well represent genuine associations. Including centre as a covariate would abolish the ability to detect this effect.

The Manhattan plot shown in **Figure 5.6** represents a scenario before removal of these five false positive SNPs.

TaqMan® genotyping of SNP rs4110518

SNP rs4110518 was genotyped using TaqMan® genotyping assay (**Methods 2.2.7**). A total of nine samples (four major homozygotes, four heterozygotes and one minor homozygote with GWAS data) were genotyped using the assay as positive controls. The positive control results showed identical genotypes as derived from the GWAS chips. The replication cohort consisted of 462 samples (**Figure 5.7**).

As expected, SNP genotype data (rs4110518) from the replication cohort showed no significant evidence of association with human ageing ($p = 0.5064$).

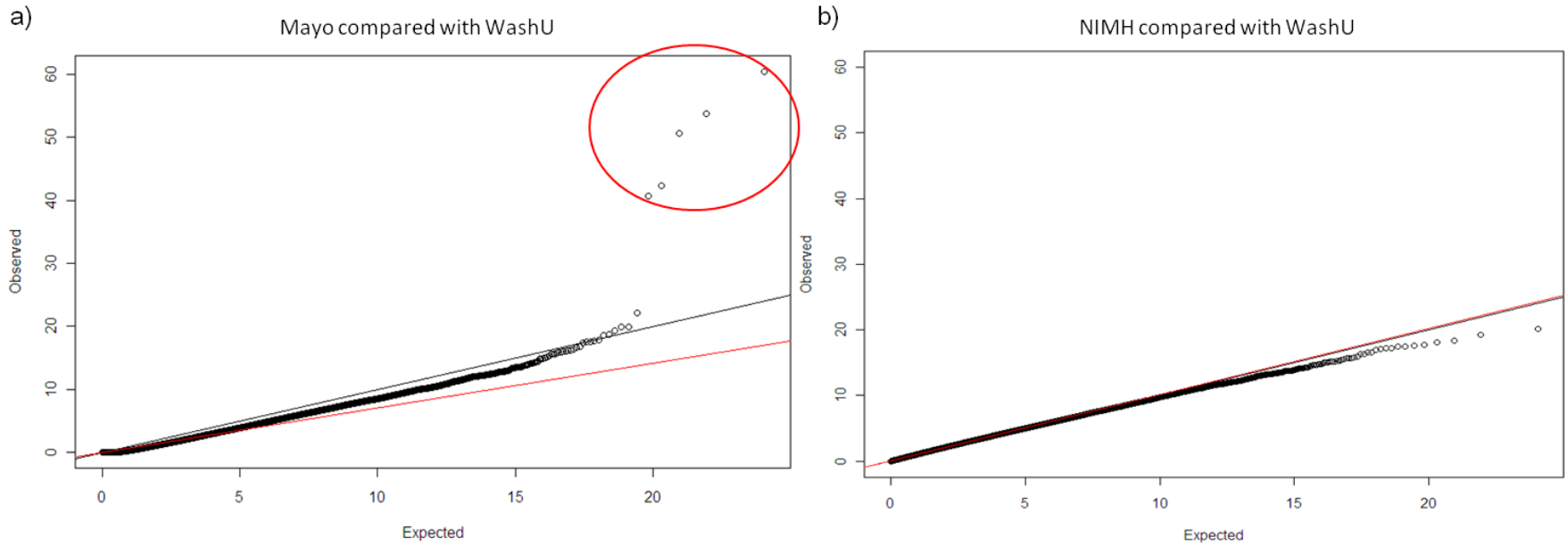



**Figure 5.5 QQ plot of χ2 - χ2 p-values to determine bias in SNP frequencies observed in Mayo (a), NIMH (b) versus WashU data.** a) Logistic regression (Mayo versus WashU samples) adjusted for the top six PCs and AAD. Five SNPs (circled) showed significant bias in the Mayo compared with WashU data taking into account population stratification and AAD. b) Logistic regression comparing NIMH data versus WashU data taking into account population stratification and AAD. No bias was observed in NIMH compared with WashU. Solid line: expected under null hypothesis i.e. no significant difference (or no significant association); Open circles: data points. Red line: fitted slope of all data points. The diagram was drawn using GenABEL in R (v2.12.1).

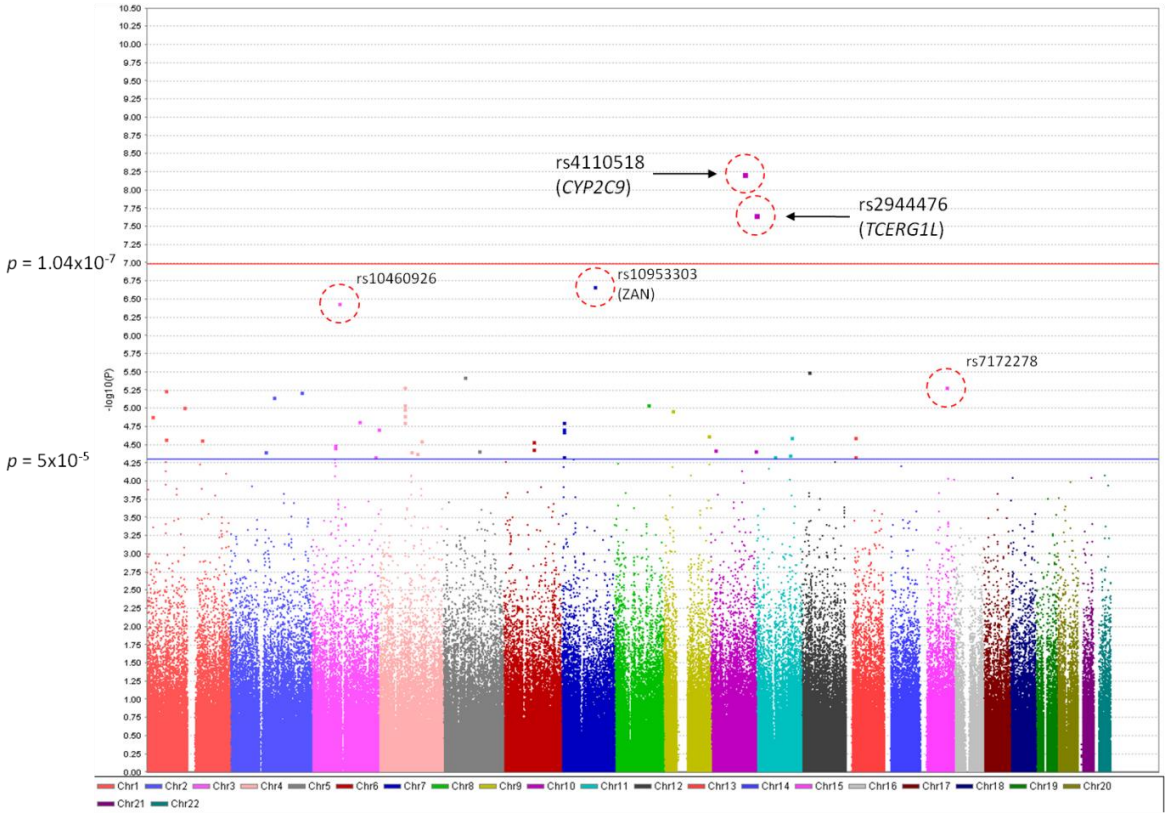



**Figure 5.6 Manhattan plot of GWAS in human ageing.** Chromosomal position is shown on the x-axis versus -log10 GWAS p-value on the y-axis. The threshold for genome-wide significance ($p$ = 1.04 x $10^{-7}$) and p-value threshold ($p$ = 5 x $10^{-5}$) are indicated by the horizontal lines. SNPs between these thresholds show "suggestive" associations. The five SNPs (highlighted by circles) exhibit significant differences in allele frequencies between samples from Mayo and WashU (see **Figure 5.5**). Two of the five SNPs (rs4110518 and rs2944476) showed spurious genome-wide significant signals as a result of this bias.

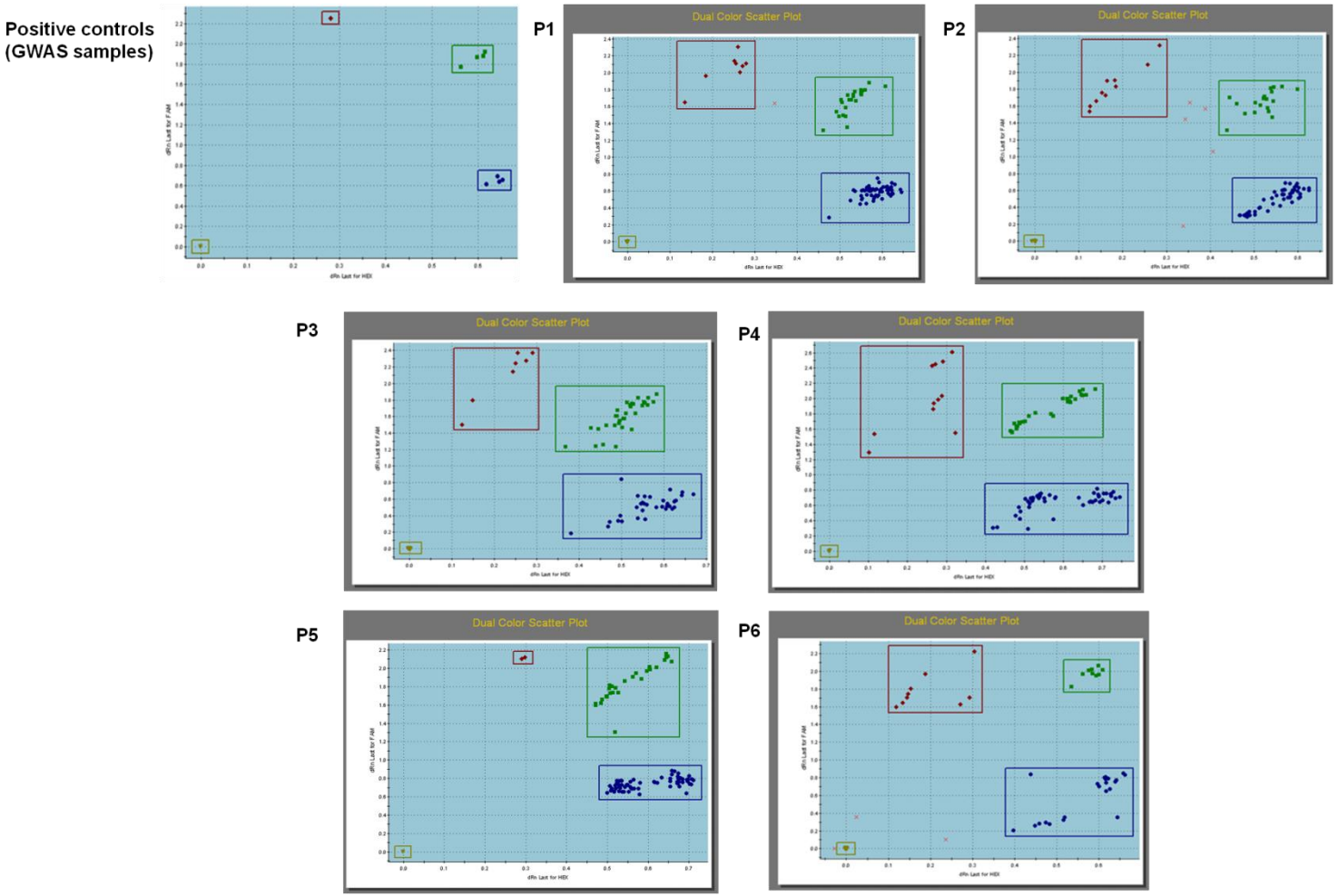



**Figure 5.7 TaqMan® genotyping assays for SNP rs4110518.** TaqMan® results, dual scatter plots, are shown for positive control samples and the six 96-well plates (P1 to P6). The HEX and FAM signal thresholds are indicated by horizontal lines (green – HEX and blue – FAM). Individual genotypes are determined according to the signal intensities of both TaqMan® probes (HEX/FAM): depicted using different coloured points – blue (individual homozygous for the major allele 'C'), red (individual homozygous for the minor allele 'T'), green (heterozygote) and yellow (no template controls). The corresponding amplification curves are shown in **Appendix 8.3**.

Power Calculation

Power calculations indicate that a sample size between ~3,000 and ~15,000 is required in order to have 80% power to detect an association with a MAF ranging above 0.05. Approximately ~4,000 to ~19,000 samples will be needed if 95% power is required.

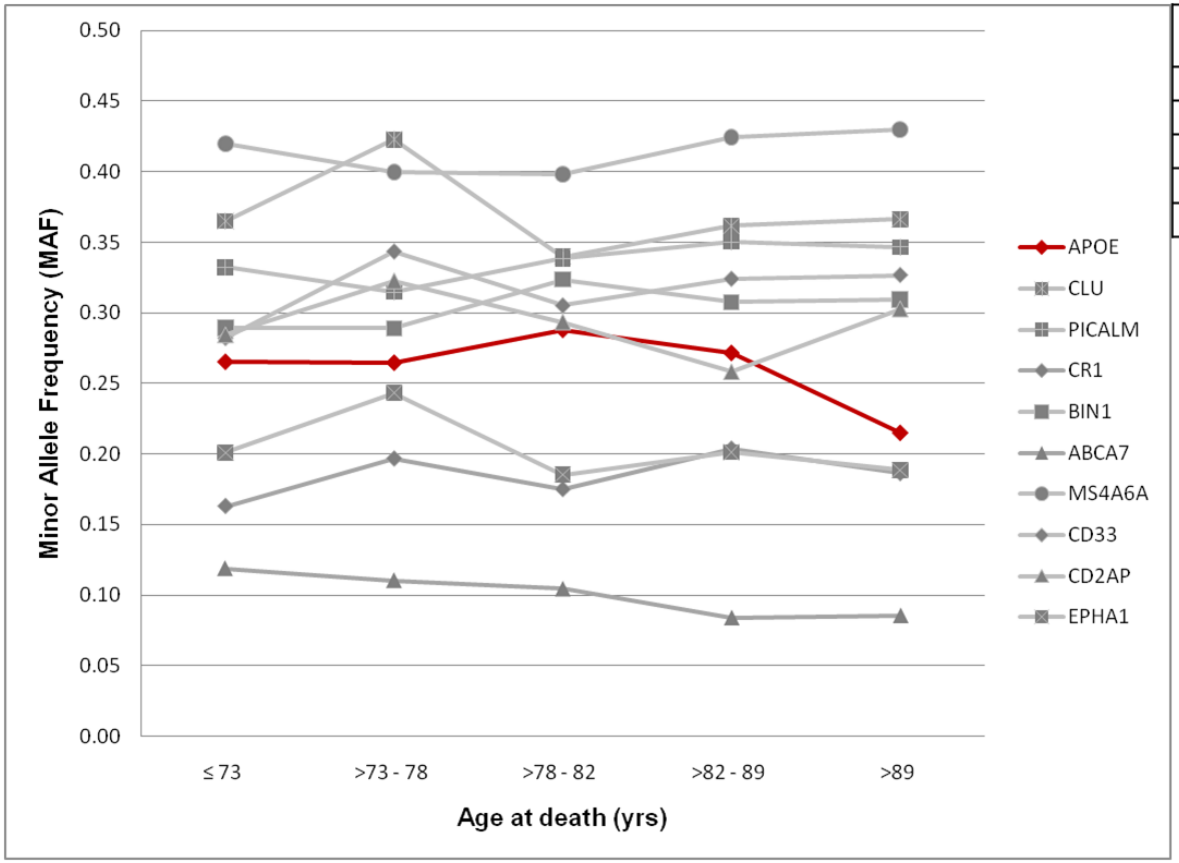
This estimation should be interpreted with caution as it is based on a number of assumptions (such as effect size and mode of inheritance), gene-environment interaction (GxE) has also not been taken into account.

## 5.5 Discussion

It is known that age is one of the biggest risk factors for LOAD. The prevalence of LOAD was estimated ranging from 0.6% in persons aged 65 to 69 years to 22.2% in persons aged 90 and older (Lobo et al., 2000). Since age is one of the biggest risk factors for LOAD, it is important to understand whether genes involved in LOAD play a role in successful ageing and longevity.

In this study, an association test of the top GWAS LOAD genes (*APOE, CLU, PICALM, CR1, BIN1, ABCA7, MS4A6A, CD33, CD2AP* and *EPHA1*) with human ageing was performed using the most significant SNPs found in previous studies. Apart from the well documented association between *APOE* and LOAD, the association with the other nine genes was identified recently through large GWAS, each with a sample size of over 10,000 (Harold et al., 2009; Hollingworth et al., 2011; Lambert et al., 2009; Naj et al., 2011; Seshadri et al., 2010).

The results of this study provided compelling evidence of association between *APOE* locus (rs2075650) and human ageing ($p = 5.27 \times 10^{-4}$) (**Table 5.4A**) with risk effect based on the analysis of 1,364 samples using AAD as a continuous trait. The minor allele frequency plot (**Figure 5.8**) shows that the MAF of this SNP significantly decreases in the old AAD category (MAF = 0.21, AAD > 89 years of age, n = 228) compared with the other four younger AAD categories (MAF = 0.27, AAD ≤ 89, n =1,136). Interestingly, individuals homozygous for the minor allele ‘G’ showed significantly lowered AAD ($p = 0.002$) compared with individuals homozygous for the major allele ‘A’. No effect was seen for the individuals carrying ‘AG’ genotype ($p = 0.891$).




| AAD Category | Count |
|---|---|
| ≤ 73 | 313 |
| >73 - 78 | 259 |
| >78 - 82 | 260 |
| >82 - 89 | 304 |
| >89 | 228 |

**Figure 5.8 Minor Allele Frequency (MAF) analysis for ten LOAD genes with respect to ageing.** The figure shows the relationship between SNP MAFs and human ageing, where AAD is separated into five categories. Each AAD category contains roughly equal amounts of samples to avoid bias in sample sizes. All ten documented LOAD genes are shown together with the *APOE* locus highlighted in red (all other loci in grey). The *APOE* locus (rs2075650) showed significant association with ageing, with MAF = 0.27, AAD ≤ 89 years of age (n = 1,136) and MAF = 0.21, AAD > 89 years of age (n = 228). None of the other gene loci were significantly associated with ageing. The analysis of quantitative trait was conducted on the whole dataset. The stratification of age at death into five age-at-death categories was only used to facilitate visualisation of the minor allele frequency of the candidate genes in the different age ranges, and was not used to generate p-values.

To date, *APOE* has been extensively examined with respect to human ageing due to its role in AD and vascular disease. A longitudinal study following subjects for 18 years using 1,094 individuals aged 75 and older showed that the risk of mortality was affected by the *APOE* gene. Risk was increased by 22% in those carrying the *APOE* ε4 allele, decreased by 28% in those carrying the *APOE* ε2 allele, and individuals carrying the *APOE* ε3 allele showed no significant difference in risk (Lewis and Brunner, 2004; Rosvall et al., 2009). The association between *APOE* ε2 variant and ageing has been investigated in Finish centenarians, where a trend of association was observed – 9%, 21%, and 25% in people aged 100 to 101, 102 to 103 and 104 years and older, respectively (Frisoni et al., 2001). SNP rs2075650 is known to be in tight LD with the *APOE* ε4 allele (Yu et al., 2007). The direction of the effect of rs2075650 in this study is compatible with previous findings for the *APOE* ε4 allele (Christensen et al., 2006). APOE is a major transporter of cholesterol and has been implicated in multiple age-related diseases including LOAD and vascular diseases (Panza et al., 2007).

No evidence of association was observed with the remaining LOAD genes implying that these genes are genuine LOAD genes with no detectable effects on human ageing. However, the possibility of these genes having weak effects on ageing that the dataset analysed was not sufficiently large enough to detect cannot be ruled out.

All SNPs on the Illumina chips were subsequently analysed after stringent QC procedures. The mean AAD for the samples analysed was over 80 years (**Table 5.1**). This minimized the possibility of early death (prior to age 40 years) as a result of underlying non-genetic factors or highly penetrant genetic factors affecting the analysis (McGue et al., 1993).

In an assessment of all SNPs on the chip, none were found to approach genome-wide significance as calculated for this study ($p < 1.04 \times 10^{-7}$). The inability to detect

any novel ageing associated variants is likely the result of a lack of power. The calculation of power using QUANTO (v 1.2.4) has suggested a much larger sample size is required in order to detect an association with common variants. With rare exceptions, common variants are known to exert only small to moderate effects, according to previous studies of many complex disorders and traits (Bodmer and Bonilla, 2008).

GWAS provide a method of identifying common genetic variations associated with disease or phenotype in an unbiased manner. However, it comes with a price of correction for multiple testing given that hundreds and thousands of SNPs are tested simultaneously. A very stringent significance threshold ($p < 5 \times 10^{-8}$) is often used to infer a genome-wide significant association and to avoid large number of false positives (Bertram et al., 2008).

The analysis was conducted using AAD as a quantitative trait; this is believed to provide more power compared with a traditional case/control approach. The advantage of statistical power gained compared with the case/control analysis is dependent upon the design of the study. For example, dichotomizing the AAD distribution into cases and controls would give less power than comparing the low and high extremes of the quantitative trait (Plomin et al., 2009). Increasing statistical power by including more samples is imperative to elucidate genuine genetic associations in this study. Including more samples with the extreme phenotypes (e.g. exceptional longevity - nonagenarians and centenarians) would give more power than addition of samples of average AAD (Plomin et al., 2009; Tan et al., 2010).

Domestic and international collaborations are often required to raise sufficiently large sample sizes in order to have adequate power to detect genuine disease associations. This is especially true for SNPs with a small effect size. However, such combined analysis can in some instances generate new problems. For instance,

inter-chip and inter-cohort differences could create spurious genome-wide significant associations. More importantly, these SNPs may pass all conventional QC filtering (e.g. Hardy Weinberg Equilibrium p-value, minor allele frequency, genotyping rate threshold) increasing the likelihood of generating false positive results, which are not corrected for by principal component analysis.

As shown in **Figure 5.5** and **Figure 5.6,** ignoring comparison of data from centres (Mayo and WashU) gave spurious genome-wide significant associations (rs4110518, $p = 5.96 \times 10^{-9}$ and rs2944476, $p = 2.19 \times 10^{-8}$). Therefore, extra caution should be made when performing GWAS analysis which utilises data from multiple centres.

Including 'Centre' as a covariate has been widely used to solve such problems raised by centres and this is largely effective. However this is not always possible, especially in circumstances where the number of cases and controls are significantly different between centres. In this study, correcting for centre was not possible due to the AAD bias in the centres sampled. The overall spread of AAD is crucial to this analysis and the difference in allele frequency between individuals with relatively young AAD (Mayo) and relatively old (WashU) may well represent genuine associations.

Furthermore, samples were included from both LOAD cases and controls in this study which is intended to achieve maximum power to detect novel ageing associated variants. Ideally this test is better performed using only control samples. Considering that 'pure controls' where individuals die without experiencing any age-related diseases probably do not exist, it was considered valid to undertake an analysis using both sets. However, due to the large number of AD cases that have been used relative to the number of controls (about three quarter of the total), any association with human ageing implicated in the study may be biased and specific to

the AD population. Follow-up studies using only control samples will be required to confirm these associations.

## 5.6 Conclusion

In this chapter, a study was conducted to investigate genetic factors influencing human ageing using LOAD GWAS data from the GERAD1 (Genetic and Environmental Risk for Alzheimer's Disease) and Mayo GWAS datasets that had documented age-at-death (**Table 5.1**).

The study, which consists of both a candidate gene study and genome-wide analysis, were conducted using age-at-death as a quantitative trait, a method likely to provide more power than a traditional case/control analysis (Plomin et al., 2009).

Testing of the 'known' LOAD genes for association with human ageing may provide insights on whether these genes are directly associated with the disease or indirectly by allowing successful ageing. The *APOE* locus (rs2075650) showed compelling evidence of association with human ageing with a p-value which withstood multiple test correction for ten independent tests (*uncorr-p* = $5.27 \times 10^{-4}$); the number of genes tested. This effect is consistent with previous reports of an association of the *APOE* locus with human ageing (Deelen et al., 2011; Panza et al., 2009).

None of the other nine genetic loci (*CLU*, *PICALM*, *CR1*, *BIN1*, *ABCA7*, *MS4A6A*, *CD33*, *CD2AP* and *EPHA1*) showed significant evidence of association (*uncorr-p* > 0.05), suggesting these genes are genuine LOAD genes with no detectable effect on ageing in this study.

No SNPs were found to approach genome-wide significance in an assessment using all SNPs available on the genotyping chips after stringent QC. This is likely to be due to the lack of power. Increasing statistical power by including more samples, especially individuals with exceptional longevity (e.g. nonagenarians and

centenarians), will be imperative in identification of genuine genetic association with human ageing.

Presenting data on genes that lie between $10^{-5}$ and $10^{-8}$ (**Table 5.4B**) may enable groups to identify genes for future study especially if there is overlap with other studies. Additionally this data could be used as part of a larger meta-analysis.

# Chapter 6: General discussion

Alzheimer's disease is the most prevalent form of dementia, representing the majority (~60%) of dementia cases. Approximately 35 million people are affected by Alzheimer's disease worldwide as of 2009, and it has been estimated that this figure will increase to ~65 million and ~115 million by 2030 and 2050, respectively (Ferri et al., 2009).

In Mendelian disorders, presence or absence of a disease is often caused by mutations in a single gene. These mutations are found to be 100% penetrant (i.e. all individuals would get the disease if they carry the mutation). As opposed to Mendelian type of disorders, the penetrances of genetic factors found responsible for LOAD are low. As a result, it is intractable to map these genes through linkage analysis using family pedigrees with the exception of *APOE* ε4 (Brookfield, 2010).

GWAS is a powerful approach in identifying susceptibility genes responsible for common diseases. In comparison with candidate gene studies, GWAS is known to have several major advantages:

- not limited to pre-defined set of candidate genes,
- the ability to adjust and account for complex population substructures (e.g. principal component analysis using EIGENSTRAT),
- fine mapping via imputation analysis,
- serve as a replication dataset for proposed associations without having to perform additional genotyping experiments (Bertram, 2011).

Since 2009, nine novel genetic loci (in addition to *APOE* ε4) have been unequivocally identified and confirmed by several large GWAS as associated with the risk of LOAD though large consortium efforts (Harold et al., 2009; Hollingworth et al., 2011; Lambert et al., 2009; Naj et al., 2011; Seshadri et al., 2010).

These genes highlighted biological pathways: immune system function, cholesterol metabolism, synaptic cell membrane processes and endocytosis, which provide novel insights into the development of effective therapeutic treatments and more accurate diagnosis (Morgan, 2011).

Despite recent success in identification of common variants associated with risk of LOAD, there is increasing evidence that a substantial fraction of the genetic variation contributing to AD risk remained unexplained (See **Introduction 1.9**).

Insufficient power has meant a large number of early GWAS failed to generate any convincing results. From 2007 to 2009, there were ten published GWAS of AD performed by individual research teams reporting on nine distinct datasets. Although nearly all these studies confirmed the association with the APOE locus, collectively very few novel AD risk genes were identified, and nearly all these findings were not confirmed in independent samples (Sherva and Farrer, 2011).

A cross-platform meta-analysis of four GWAS was performed in an effort to search for novel genetic associations (**Chapter 3**), as combining individually underpowered GWAS would increase power thus allowing identification of novel genetic risk factors associated with LOAD. A single SNP rs929156 located in exon 7 of *TRIM15* gene showed a significant evidence of association with risk of LOAD (*uncorr-p* = $8.77 \times 10^{-8}$), and the random effect meta-analysis of odds ratios was also found significant (*p* = 0.03). The minor allele of this SNP was found to be risky with odds ratio 1.11 (95% CI 1.01-1.22).

An approach was described in this study which may prove to be useful when larger datasets are utilized. In addition, a PERL script was written to automate the cross-platform GWAS meta-analysis using Fisher's combined probability test (**Chapter 3**).

Given the hypothesis that GWAS signals may be attributable to multiple rare variants nearby (Dickson et al., 2010), it is imperative to determine whether the region

surrounding the GWAS SNPs harbours multiple rare variants which are associated with the disease.

A next generation sequencing pipeline was described which can be applied to other studies which use the ABI SOLiD® next generation sequencing platform.

Evolutionary conserved regions encompassing the SNP (rs929156) were enriched by LR-PCR and sequenced using ABI SOLiD® next generation sequencing (**Chapter 4**). A pooled DNA strategy using 150 samples (75 LOAD cases and controls) was employed, where each pool was estimated to provide ~80% power to detect a SNP with an allele frequency of ~1%, and >95% power to detect a SNP with allele frequency above 2%.

SNP rs929156, which was found significantly associated with LOAD from the cross-platform meta-analysis, did not however show significant evidence of association by analysing the NGS data (*uncorr-p* = 0.193). This is perhaps unsurprising as the number of samples sequenced by NGS is likely to be underpowered for a case/control study.

With respect to SNP discovery, all 31 SNPs with documented allele frequency greater than 1% in the latest SNP databases (dbSNP#132 and HapMap#28 CEU population) were identified. All of which exhibited identical alternative alleles as documented in the database. With the exception of a single SNP, the remaining 14 SNPs showed a compelling correlation coefficient (96.6%) of allele frequency in comparison to the allele frequency quoted in the HapMap database (CEU population, release 28).

Ten high quality novel rare variants were identified after in-depth quality control measures. Four of which have also been discovered by the 1000 genome project, suggesting these SNPs are likely to be genuine (The 1000 Genomes Project Consortium, 2010).

Genotyping of the remaining novel rare variants using alternative genotyping methods (e.g. TaqMan® genotyping assay, Sanger sequencing) are required to confirm if they are genuine SNPs, and worthy of further investigations.

Three common SNPs (rs41272591, rs9380156 and rs6905949) with MAF greater than 5% showed significant ($p < 0.05$) evidence of association with LOAD. The top two SNPs were validated using TaqMan® genotyping assays with allele frequency similar to those estimated by NGS. Genotyping of an independent sample cohort (90 LOAD cases and 91 controls) however failed to achieve significance ($p = 0.89$).

As a pilot study, the NGS study has only a limited power to detect an association: ~ 80% power to detect common variants if OR > 3 and rare variants if OR > 4 estimated using QUANTO (v 1.2.4).

None of the rare variants were found to be significantly associated with LOAD, which is likely the result of lack of power. The odds ratios for these high quality rare variants were found to be less than 4, consistent with the power calculation.

Interestingly, a coding change (H33Y) (located at chr6: 30131558) was found only in cases (with MAF = 1.33%) and absent in controls, and predicted to be 'probably damaging' by Polyphen-2. Provided with sufficient power, this association may prove to be genuine.

In **Chapter 5**, bioinformatic analyses were undertaken to investigate genetic variants influencing human life-span using late-onset Alzheimer's disease GWAS data with documented AAD.

As age is one of the biggest risk factor in Alzheimer's disease, it is important to understand whether LOAD genes are directly associated with the disease or indirectly by allowing successful ageing.

The ten most promising LOAD genes (*APOE*, *CLU*, *PICALM*, *CR1*, *BIN1, ABCA7*, *MS4A6A*, *CD33*, *CD2AP* and *EPHA1*) identified through recent large GWAS (Harold et al., 2009; Hollingworth et al., 2011; Lambert et al., 2009; Naj et al., 2011; Seshadri et al., 2010) were tested for association with human ageing in a merged dataset (from ARUK and GERAD consortia) using the most significant SNPs found in previous studies.

The results of the analysis provided compelling evidence of association between the *APOE* locus (rs2075650) and human ageing ($p = 5.27 \times 10^{-4}$) as expected. The minor allele of this SNP was found to be overrepresented in individuals with a young AAD (MAF = 0.27; ≤ 89 years of age) in comparison to individuals with an old AAD (MAF = 0.21; > 89 years of age). None of the other LOAD gene loci showed significant evidence of association ($p < 0.05$) with ageing, suggesting that these genes are likely to be genuine LOAD genes with no detectable effect on human ageing.

A genome wide analysis was performed in an effort to search for novel genetic risk factors associated with human lifespan. No SNPs were found to be associated with human ageing with genome-wide level of significance after assessing all SNPs on the chip. Increasing statistical power by including more samples is paramount to enable detection of novel genetic associations with human ageing. Tan et al., 2010 have shown that increasing sample age from nonagenarians to centenarians further increases the power to discover variants associated with human ageing.

Furthermore, the chapter highlighted the importance of quality control procedures taking into account inter-chip and inter-cohort differences in an analysis; these differences may lead to spurious genome wide significant associations, which may pass all conventional QC filters (e.g. Hardy Weinberg Equilibrium p-value, minor allele frequency, genotyping rate threshold) and not be corrected for by principal component analysis.

Future perspective

Late-onset Alzheimer's disease is a multifactorial and complex disease affected by both environmental and genetic factors, where genetic factors are estimated to contribute as much as 76% of LOAD cases (Harold et al., 2009; van Es and van den Berg, 2009).

The *APOE* gene found almost 20 years ago remains the single most outstanding risk factor (Corder et al., 1993; Hardy and Higgins, 1992), which has been estimated to account for ~25% of the risk of LOAD. The newly found LOAD genes (*CLU*, *PICALM*, *CR1*, *BIN1, ABCA7*, *MS4A6A*, *CD33*, *CD2AP* and *EPHA1*) were estimated to contribute to another ~30% of the population risk for the disease (Naj et al., 2011). However, these estimations are likely to be inflated, and the true proportions of the genetic predisposition accounted for by these genes are likely to be much lower (Sherva and Farrer, 2011). Additional genetic risk factors are likely to be found and replicated in the near future, where it has become increasingly evident that less frequent and rare variants are also playing a role (Manolio et al., 2009).

*TRIM15* acts as a potential LOAD candidate gene, and the association with LOAD remains to be confirmed through larger studies. *TRIM15* encodes for a protein likely to be involved in the innate immune system, one of the pathways known to be involved in LOAD pathogenesis (Jones et al., 2010; McNab et al., 2010; Morgan, 2011).

Furthermore, genetic loci identified by GWAS are unlikely to be functional, and merely act as proxies in LD with functional variants. The advent of next generation sequencing technology, for the first time in human genetic research, enables identification and testing of novel functional variants at base-pair resolution with affordable costs (Bertram, 2011). With the improved understanding of pathways uncovered by genetic research in LOAD, it is hoped that in the not too distant future,

earlier diagnosis and better therapy targeting the root cause of the devastating disease will become available.

# 7. References

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat Methods *7*, 248-249.

Arendt, J., Van Someren, E.J., Appleton, R., Skene, D.J., and Akerstedt, T. (2008). Clinical update: melatonin and sleep disorders. Clin Med *8*, 381-383.

Areosa, S.A., Sherriff, F., and McShane, R. (2005). Memantine for dementia. Cochrane Database Syst Rev, CD003154.

Atzmon, G., Schechter, C., Greiner, W., Davidson, D., Rennert, G., and Barzilai, N. (2004). Clinical phenotype of families with longevity. J Am Geriatr Soc *52*, 274-277.

Aulchenko, Y.S., Ripke, S., Isaacs, A., and van Duijn, C.M. (2007). GenABEL: an R library for genome-wide association analysis. Bioinformatics *23*, 1294-1296.

Avila, J., Lucas, J.J., Perez, M., and Hernandez, F. (2004). Role of tau protein in both physiological and pathological conditions. Physiol Rev *84*, 361-384.

Avramopoulos, D. (2009). Genetics of Alzheimer's disease: recent advances. Genome Med *1*, 34.

Baig, S., Joseph, S.A., Tayler, H., Abraham, R., Owen, M.J., Williams, J., Kehoe, P.G., and Love, S. (2010). Distribution and expression of picalm in Alzheimer disease. J Neuropathol Exp Neurol *69*, 1071-1077.

Ballard, C., Gauthier, S., Corbett, A., Brayne, C., Aarsland, D., and Jones, E. (2011). Alzheimer's disease. Lancet *377*, 1019-1031.

Bareggi, S.R., and Cornelli, U. (2011). Clioquinol: Review of its Mechanisms of Action and Clinical Uses in Neurodegenerative Disorders. CNS Neurosci Ther.

Barrett, J.C., Fry, B., Maller, J., and Daly, M.J. (2005). Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics *21*, 263-265.

Beecham, G.W., Martin, E.R., Li, Y.J., Slifer, M.A., Gilbert, J.R., Haines, J.L., and Pericak-Vance, M.A. (2009). Genome-wide association study implicates a chromosome 12 risk locus for late-onset Alzheimer disease. Am J Hum Genet *84*, 35-43.

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

Bertram, L., Lange, C., Mullin, K., Parkinson, M., Hsiao, M., Hogan, M.F., Schjeide, B.M., Hooli, B., Divito, J., Ionita, I.*, et al.* (2008). Genome-wide association analysis reveals putative Alzheimer's disease susceptibility loci in addition to APOE. Am J Hum Genet *83*, 623-632.

Bertram, L., and Tanzi, R.E. (2008). Thirty years of Alzheimer's disease genetics: the implications of systematic meta-analyses. Nat Rev Neurosci *9*, 768-778.

Bertram, L., Lill, C.M., and Tanzi, R.E. (2010). The genetics of Alzheimer disease: back to the future. Neuron *68*, 270-281.

Bertram, L. (2011). Alzheimer's Genetics in the GWAS Era: A Continuing Story of 'Replications and Refutations'. Curr Neurol Neurosci Rep *11*, 246-253.

Bewick, V., Cheek, L., and Ball, J. (2005). Statistics review 14: Logistic regression. Crit Care *9*, 112-118.

Blennow, K., de Leon, M.J., and Zetterberg, H. (2006). Alzheimer's disease. Lancet *368*, 387-403.

Bodmer, W., and Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet *40*, 695-701.

Branton, D., Deamer, D.W., Marziali, A., Bayley, H., Benner, S.A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X.*, et al.* (2008). The potential and challenges of nanopore sequencing. Nat Biotechnol *26*, 1146-1153.

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

Brookmeyer, R., Johnson, E., Ziegler-Graham, K., and Arrighi, H.M. (2007). Forecasting the global burden of Alzheimer's disease. Alzheimers Dement *3*, 186-191.

Brzezinski, A., Vangel, M.G., Wurtman, R.J., Norrie, G., Zhdanova, I., Ben-Shushan, A., and Ford, I. (2005). Effects of exogenous melatonin on sleep: a meta-analysis. Sleep Med Rev *9*, 41-50.

Bu, G. (2009). Apolipoprotein E and its receptors in Alzheimer's disease: pathways, pathogenesis and therapy. Nat Rev Neurosci *10*, 333-344.

Buchhave, P., Blennow, K., Zetterberg, H., Stomrud, E., Londos, E., Andreasen, N., Minthon, L., and Hansson, O. (2009). Longitudinal study of CSF biomarkers in patients with Alzheimer's disease. PLoS One *4*, e6294.

Calero, M., Rostagno, A., Matsubara, E., Zlokovic, B., Frangione, B., and Ghiso, J. (2000). Apolipoprotein J (clusterin) and Alzheimer's disease. Microsc Res Tech *50*, 305-315.

Campion, D., Dumanchin, C., Hannequin, D., Dubois, B., Belliard, S., Puel, M., Thomas-Anterion, C., Michon, A., Martin, C., Charbonnier, F.*, et al.* (1999). Early-onset autosomal dominant Alzheimer disease: prevalence, genetic heterogeneity, and mutation spectrum. Am J Hum Genet *65*, 664-670.

Cardinali, D.P., Brusco, L.I., Liberczuk, C., and Furio, A.M. (2002). The use of melatonin in Alzheimer's disease. Neuro Endocrinol Lett *23 Suppl 1*, 20-23.

Carrasquillo, M.M., Zou, F., Pankratz, V.S., Wilcox, S.L., Ma, L., Walker, L.P., Younkin, S.G., Younkin, C.S., Younkin, L.H., Bisceglio, G.D.*, et al.* (2009). Genetic variation in PCDH11X is associated with susceptibility to late-onset Alzheimer's disease. Nat Genet *41*, 192-198.

Carrasquillo, M.M., Belbin, O., Hunter, T.A., Ma, L., Bisceglio, G.D., Zou, F., Crook, J.E., Pankratz, V.S., Dickson, D.W., Graff-Radford, N.R.*, et al.* (2010). Replication of CLU, CR1, and PICALM Associations With Alzheimer Disease. Arch Neurol.

Carrette, O., Demalte, I., Scherl, A., Yalkinoglu, O., Corthals, G., Burkhard, P., Hochstrasser, D.F., and Sanchez, J.C. (2003). A panel of cerebrospinal fluid potential biomarkers for the diagnosis of Alzheimer's disease. Proteomics *3*, 1486-1494.

Cash, A.D., Aliev, G., Siedlak, S.L., Nunomura, A., Fujioka, H., Zhu, X., Raina, A.K., Vinters, H.V., Tabaton, M., Johnson, A.B.*, et al.* (2003). Microtubule reduction in Alzheimer's disease and aging is independent of tau filament formation. Am J Pathol *162*, 1623-1627.

Chibnik, L.B., Shulman, J.M., Leurgans, S.E., Schneider, J.A., Wilson, R.S., Tran, D., Aubin, C., Buchman, A.S., Heward, C.B., Myers, A.J.*, et al.* (2011). CR1 is associated with amyloid plaque burden and age-related cognitive decline. Ann Neurol *69*, 560-569.

Christensen, K., Johnson, T.E., and Vaupel, J.W. (2006). The quest for genetic determinants of human longevity: challenges and insights. Nat Rev Genet *7*, 436-448.

Churcher, I. (2006). Tau therapeutic strategies for the treatment of Alzheimer's disease. Curr Top Med Chem *6*, 579-595.

Cirrito, J.R., Deane, R., Fagan, A.M., Spinner, M.L., Parsadanian, M., Finn, M.B., Jiang, H., Prior, J.L., Sagare, A., Bales, K.R.*, et al.* (2005). P-glycoprotein deficiency at the blood-brain barrier increases amyloid-beta deposition in an Alzheimer disease mouse model. J Clin Invest *115*, 3285-3290.

Cirulli, E.T., and Goldstein, D.B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet *11*, 415-425.

Clancy, D.J., Gems, D., Harshman, L.G., Oldham, S., Stocker, H., Hafen, E., Leevers, S.J., and Partridge, L. (2001). Extension of life-span by loss of CHICO, a Drosophila insulin receptor substrate protein. Science *292*, 104-106.

Combarros, O., Cortina-Borja, M., Smith, A.D., and Lehmann, D.J. (2009). Epistasis in sporadic Alzheimer's disease. Neurobiol Aging *30*, 1333-1349.

Corder, E.H., Saunders, A.M., Strittmatter, W.J., Schmechel, D.E., Gaskell, P.C., Small, G.W., Roses, A.D., Haines, J.L., and Pericak-Vance, M.A. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science *261*, 921-923.

Cousin, M.A., and Robinson, P.J. (2001). The dephosphins: dephosphorylation by calcineurin triggers synaptic vesicle endocytosis. Trends Neurosci *24*, 659-665.

Cribbs, D.H. (2010). Abeta DNA vaccination for Alzheimer's disease: focus on disease prevention. CNS Neurol Disord Drug Targets *9*, 207-216.

Crocker, P.R., Paulson, J.C., and Varki, A. (2007). Siglecs and their roles in the immune system. Nat Rev Immunol *7*, 255-266.

Cunnane, S.C., Plourde, M., Pifferi, F., Begin, M., Feart, C., and Barberger-Gateau, P. (2009). Fish, docosahexaenoic acid and Alzheimer's disease. Prog Lipid Res *48*, 239-256.

Cupers, P., Orlans, I., Craessaerts, K., Annaert, W., and De Strooper, B. (2001). The amyloid precursor protein (APP)-cytoplasmic fragment generated by gamma-

secretase is rapidly degraded but distributes partially in a nuclear fraction of neurones in culture. J Neurochem *78*, 1168-1178.

Cutler, R.G., and Mattson, M.P. (2006). The adversities of aging. Ageing Res Rev *5*, 221-238.

Davis, D.G., Schmitt, F.A., Wekstein, D.R., and Markesbery, W.R. (1999). Alzheimer neuropathologic alterations in aged cognitively normal subjects. J Neuropathol Exp Neurol *58*, 376-388.

de Chaves, E.P., and Narayanaswami, V. (2008). Apolipoprotein E and cholesterol in aging and disease in the brain. Future Lipidol *3*, 505-530.

de Magalhaes, J.P., Budovsky, A., Lehmann, G., Costa, J., Li, Y., Fraifeld, V., and Church, G.M. (2009). The Human Ageing Genomic Resources: online databases and tools for biogerontologists. Aging Cell *8*, 65-72.

Deane, R., Wu, Z., Sagare, A., Davis, J., Du Yan, S., Hamm, K., Xu, F., Parisi, M., LaRue, B., Hu, H.W.*, et al.* (2004). LRP/amyloid beta-peptide interaction mediates differential brain efflux of Abeta isoforms. Neuron *43*, 333-344.

Deane, R., Sagare, A., Hamm, K., Parisi, M., Lane, S., Finn, M.B., Holtzman, D.M., and Zlokovic, B.V. (2008). apoE isoform-specific disruption of amyloid beta peptide clearance from mouse brain. J Clin Invest *118*, 4002-4013.

Deane, R., Sagare, A., and Zlokovic, B.V. (2008). The role of the cell surface LRP and soluble LRP in blood-brain barrier Abeta clearance in Alzheimer's disease. Curr Pharm Des *14*, 1601-1605.

Deane, R., Bell, R.D., Sagare, A., and Zlokovic, B.V. (2009). Clearance of amyloid-beta peptide across the blood-brain barrier: implication for therapies in Alzheimer's disease. CNS Neurol Disord Drug Targets *8*, 16-30.

Deelen, J., Beekman, M., Uh, H.W., Helmer, Q., Kuningas, M., Christiansen, L., Kremer, D., van der Breggen, R., Suchiman, H.E., Lakenberg, N.*, et al.* (2011). Genome-wide association study identifies a single major locus contributing to survival into old age; the APOE locus revisited. Aging Cell *10*, 686-698.

DeMattos, R.B., Cirrito, J.R., Parsadanian, M., May, P.C., O'Dell, M.A., Taylor, J.W., Harmony, J.A., Aronow, B.J., Bales, K.R., Paul, S.M.*, et al.* (2004). ApoE and

clusterin cooperatively suppress Abeta levels and deposition: evidence that ApoE regulates extracellular Abeta metabolism in vivo. Neuron *41*, 193-202.

DerSimonian, R., and Laird, N. (1986). Meta-analysis in clinical trials. Control Clin Trials *7*, 177-188.

Di Paolo, G., Sankaranarayanan, S., Wenk, M.R., Daniell, L., Perucco, E., Caldarone, B.J., Flavell, R., Picciotto, M.R., Ryan, T.A., Cremona, O.*, et al.* (2002). Decreased synaptic vesicle recycling efficiency and cognitive deficits in amphiphysin 1 knockout mice. Neuron *33*, 789-804.

Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H., and Goldstein, D.B. (2010). Rare variants create synthetic genome-wide associations. PLoS Biol *8*, e1000294.

Dolev, I., and Michaelson, D.M. (2004). A nontransgenic mouse model shows inducible amyloid-beta (Abeta) peptide deposition and elucidates the role of apolipoprotein E in the amyloid cascade. Proc Natl Acad Sci U S A *101*, 13909-13914.

Dubois, B., Feldman, H.H., Jacova, C., Dekosky, S.T., Barberger-Gateau, P., Cummings, J., Delacourte, A., Galasko, D., Gauthier, S., Jicha, G.*, et al.* (2007). Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. Lancet Neurol *6*, 734-746.

Edbauer, D., Willem, M., Lammich, S., Steiner, H., and Haass, C. (2002). Insulin-degrading enzyme rapidly removes the beta-amyloid precursor protein intracellular domain (AICD). J Biol Chem *277*, 13389-13393.

Edbauer, D., Winkler, E., Regula, J.T., Pesold, B., Steiner, H., and Haass, C. (2003). Reconstitution of gamma-secretase activity. Nat Cell Biol *5*, 486-488.

Efrat, M., and Aviram, M. (2010). Paraoxonase 1 Interactions with HDL, Antioxidants and Macrophages Regulate Atherogenesis - A Protective Role for HDL Phospholipids. Adv Exp Med Biol *660*, 153-166.

Ertekin-Taner, N. (2010). Genetics of Alzheimer disease in the pre- and post-GWAS era. Alzheimers Res Ther *2*, 3.

Falgarone, G., and Chiocchia, G. (2009). Chapter 8: Clusterin: A multifacet protein at the crossroad of inflammation and autoimmunity. Adv Cancer Res *104*, 139-170.

Ferri, C.P., Sousa, R., Albanese, E., Ribeiro, W.S., and Honyashiki, M. (2009). World Alzheimer Report 2009 - Executive Summary. Edited by Prince M. Alzheimer's Disease International, 1-22.

Feulner, T.M., Laws, S.M., Friedrich, P., Wagenpfeil, S., Wurst, S.H., Riehle, C., Kuhn, K.A., Krawczak, M., Schreiber, S., Nikolaus, S.*, et al.* (2009). Examination of the current top candidate genes for AD in a genome-wide association study. Mol Psychiatry.

Folstein, M.F., Folstein, S.E., and McHugh, P.R. (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res *12*, 189-198.

Frank, S., Copanaki, E., Burbach, G.J., Muller, U.C., and Deller, T. (2009). Differential regulation of toll-like receptor mRNAs in amyloid plaque-associated brain tissue of aged APP23 transgenic mice. Neurosci Lett *453*, 41-44.

Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M., and Dubchak, I. (2004). VISTA: computational tools for comparative genomics. Nucleic Acids Res *32*, W273-279.

Frisoni, G.B., Louhija, J., Geroldi, C., and Trabucchi, M. (2001). Longevity and the epsilon2 allele of apolipoprotein E: the Finnish Centenarians Study. J Gerontol A Biol Sci Med Sci *56*, M75-78.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M.*, et al.* (2002). The structure of haplotype blocks in the human genome. Science *296*, 2225-2229.

Garigan, D., Hsu, A.L., Fraser, A.G., Kamath, R.S., Ahringer, J., and Kenyon, C. (2002). Genetic analysis of tissue aging in Caenorhabditis elegans: a role for heat-shock factor and bacterial proliferation. Genetics *161*, 1101-1112.

Gatz, M., Reynolds, C.A., Fratiglioni, L., Johansson, B., Mortimer, J.A., Berg, S., Fiske, A., and Pedersen, N.L. (2006). Role of genes and environments for explaining Alzheimer disease. Arch Gen Psychiatry *63*, 168-174.

Gehrman, P.R., Connor, D.J., Martin, J.L., Shochat, T., Corey-Bloom, J., and Ancoli-Israel, S. (2009). Melatonin fails to improve sleep or agitation in double-blind randomized placebo-controlled trial of institutionalized patients with Alzheimer disease. Am J Geriatr Psychiatry *17*, 166-169.

Glatt, S.J., Chayavichitsilp, P., Depp, C., Schork, N.J., and Jeste, D.V. (2007). Successful aging: from phenotype to genotype. Biol Psychiatry *62*, 282-293.

Goodger, Z.V., Rajendran, L., Trutzel, A., Kohli, B.M., Nitsch, R.M., and Konietzko, U. (2009). Nuclear signaling by the APP intracellular domain occurs predominantly through the amyloidogenic processing pathway. J Cell Sci *122*, 3703-3714.

Guerreiro, R.J., and Hardy, J. (2011). Alzheimer's disease genetics: lessons to improve disease modelling. Biochem Soc Trans *39*, 910-916.

Haass, C. (2004). Take five--BACE and the gamma-secretase quartet conduct Alzheimer's amyloid beta-peptide generation. EMBO J *23*, 483-488.

Haass, C., and Selkoe, D.J. (2007). Soluble protein oligomers in neurodegeneration: lessons from the Alzheimer's amyloid beta-peptide. Nat Rev Mol Cell Biol *8*, 101-112.

Haberland, C. (2010). Frontotemporal dementia or frontotemporal lobar degeneration--overview of a group of proteinopathies. Ideggyogy Sz *63*, 87-93.

Hardy, J.A., and Higgins, G.A. (1992). Alzheimer's disease: the amyloid cascade hypothesis. Science *256*, 184-185.

Hardy, J., and Selkoe, D.J. (2002). The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. Science *297*, 353-356.

Harismendy, O., and Frazer, K. (2009). Method for improving sequence coverage uniformity of targeted genomic intervals amplified by LR-PCR using Illumina GA sequencing-by-synthesis technology. Biotechniques *46*, 229-231.

Harold, D., Abraham, R., Hollingworth, P., Sims, R., Gerrish, A., Hamshere, M.L., Pahwa, J.S., Moskvina, V., Dowzell, K., Williams, A.*, et al.* (2009). Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nat Genet *41*, 1088-1093.

Hemming, M.L., Patterson, M., Reske-Nielsen, C., Lin, L., Isacson, O., and Selkoe, D.J. (2007). Reducing amyloid plaque burden via ex vivo gene delivery of an Abeta-degrading protease: a novel therapeutic approach to Alzheimer disease. PLoS Med *4*, e262.

Henry, J., Mather, I.H., McDermott, M.F., and Pontarotti, P. (1998). B30.2-like domain proteins: update and new insights into a rapidly expanding family of proteins. Mol Biol Evol *15*, 1696-1705.

Herndon, L.A., Schmeissner, P.J., Dudaronek, J.M., Brown, P.A., Listner, K.M., Sakano, Y., Paupard, M.C., Hall, D.H., and Driscoll, M. (2002). Stochastic and genetic factors influence tissue-specific decline in ageing C. elegans. Nature *419*, 808-814.

Herskind, A.M., McGue, M., Holm, N.V., Sorensen, T.I., Harvald, B., and Vaupel, J.W. (1996). The heritability of human longevity: a population-based study of 2872 Danish twin pairs born 1870-1900. Hum Genet *97*, 319-323.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A *106*, 9362-9367.

Hollingworth, P., Harold, D., Sims, R., Gerrish, A., Lambert, J.C., Carrasquillo, M.M., Abraham, R., Hamshere, M.L., Pahwa, J.S., Moskvina, V.*, et al.* (2011). Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. Nat Genet *43*, 429-435.

Holmes, C., Boche, D., Wilkinson, D., Yadegarfar, G., Hopkins, V., Bayer, A., Jones, R.W., Bullock, R., Love, S., Neal, J.W.*, et al.* (2008). Long-term effects of Abeta42 immunisation in Alzheimer's disease: follow-up of a randomised, placebo-controlled phase I trial. Lancet *372*, 216-223.

Holzenberger, M., Dupont, J., Ducos, B., Leneuve, P., Geloen, A., Even, P.C., Cervera, P., and Le Bouc, Y. (2003). IGF-1 receptor regulates lifespan and resistance to oxidative stress in mice. Nature *421*, 182-187.

Hu, S.P., Zhou, G.B., Luan, J.A., Chen, Y.P., Xiao, D.W., Deng, Y.J., Huang, L.Q., and Cai, K.L. (2009). Polymorphisms of HLA-A and HLA-B genes in genetic susceptibility to esophageal carcinoma in Chaoshan Han Chinese. Dis Esophagus.

Ihalainen, J., Sarajarvi, T., Rasmusson, D., Kemppainen, S., Keski-Rahkonen, P., Lehtonen, M., Banerjee, P.K., Semba, K., and Tanila, H. (2011). Effects of memantine and donepezil on cortical and hippocampal acetylcholine levels and object recognition memory in rats. Neuropharmacology.

Infante, J., Sanz, C., Fernandez-Luna, J.L., Llorca, J., Berciano, J., and Combarros, O. (2004). Gene-gene interaction between interleukin-6 and interleukin-10 reduces AD risk. Neurology *63*, 1135-1136.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931-945.

Jenne, D.E., and Tschopp, J. (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*, 7123-7127.

Jiang, Q., Lee, C.Y., Mandrekar, S., Wilkinson, B., Cramer, P., Zelcer, N., Mann, K., Lamb, B., Willson, T.M., Collins, J.L.*, et al.* (2008). ApoE promotes the proteolytic degradation of Abeta. Neuron *58*, 681-693.

Johnson, A.D., Handsaker, R.E., Pulit, S.L., Nizzari, M.M., O'Donnell, C.J., and de Bakker, P.I. (2008). SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics *24*, 2938-2939.

Jones, L., Holmans, P.A., Hamshere, M.L., Harold, D., Moskvina, V., Ivanov, D., Pocklington, A., Abraham, R., Hollingworth, P., Sims, R.*, et al.* (2010). Genetic evidence implicates the immune system and cholesterol metabolism in the aetiology of Alzheimer's disease. PLoS One *5*, e13950.

Jones, S.E., and Jomary, C. (2002). Clusterin. Int J Biochem Cell Biol *34*, 427-431.

Jun, G., Naj, A.C., Beecham, G.W., Wang, L.S., Buros, J., Gallins, P.J., Buxbaum, J.D., Ertekin-Taner, N., Fallin, M.D., Friedland, R.*, et al.* (2010). Meta-analysis confirms CR1, CLU, and PICALM as alzheimer disease risk loci and reveals interactions with APOE genotypes. Arch Neurol *67*, 1473-1484.

Jylhava, J., and Hurme, M. (2009). Gene variants as determinants of longevity: focus on the inflammatory factors. Pflugers Arch *459*, 239-246.

Kauwe, J.S., Wang, J., Mayo, K., Morris, J.C., Fagan, A.M., Holtzman, D.M., and Goate, A.M. (2009). Alzheimer's disease risk variants show association with cerebrospinal fluid amyloid beta. Neurogenetics *10*, 13-17.

Kenyon, C., Chang, J., Gensch, E., Rudner, A., and Tabtiang, R. (1993). A C. elegans mutant that lives twice as long as wild type. Nature *366*, 461-464.

Kenyon, C.J. (2010). The genetics of ageing. Nature *464*, 504-512.

Khera, R., and Das, N. (2009). Complement Receptor 1: disease associations and therapeutic implications. Mol Immunol *46*, 761-772.

Kim, W.S., Guillemin, G.J., Glaros, E.N., Lim, C.K., and Garner, B. (2006). Quantitation of ATP-binding cassette subfamily-A transporter gene expression in primary human brain cells. Neuroreport *17*, 891-896.

Kim, J., Basak, J.M., and Holtzman, D.M. (2009). The role of apolipoprotein E in Alzheimer's disease. Neuron *63*, 287-303.

Konietzko, U., Goodger, Z.V., Meyer, M., Kohli, B.M., Bosset, J., Lahiri, D.K., and Nitsch, R.M. (2008). Co-localization of the amyloid precursor protein and Notch intracellular domains in nuclear transcription factories. Neurobiol Aging *31*, 58-73.

Koudinov, A.R., Berezov, T.T., Kumar, A., and Koudinova, N.V. (1998). Alzheimer's amyloid beta interaction with normal human plasma high density lipoprotein: association with apolipoprotein and lipids. Clin Chim Acta *270*, 75-84.

Kuperstein, I., Broersen, K., Benilova, I., Rozenski, J., Jonckheere, W., Debulpaep, M., Vandersteen, A., Segers-Nolten, I., Van Der Werf, K., Subramaniam, V.*, et al.* (2010). Neurotoxicity of Alzheimer's disease Abeta peptides is induced by small changes in the Abeta42 to Abeta40 ratio. EMBO J *29*, 3408-3420.

Kwok, S., and Higuchi, R. (1989). Avoiding false positives with PCR. Nature *339*, 237-238.

Lai, K.O., and Ip, N.Y. (2009). Synapse development and plasticity: roles of ephrin/Eph receptor signaling. Curr Opin Neurobiol *19*, 275-283.

Lambert, J.C., and Amouyel, P. (2011). Genetics of Alzheimer's disease: new evidences for an old hypothesis? Curr Opin Genet Dev *21*, 295-301.

Lambert, J.C., Heath, S., Even, G., Campion, D., Sleegers, K., Hiltunen, M., Combarros, O., Zelenika, D., Bullido, M.J., Tavernier, B.*, et al.* (2009). Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nat Genet *41*, 1094-1099.

Lee, H.G., Perry, G., Moreira, P.I., Garrett, M.R., Liu, Q., Zhu, X., Takeda, A., Nunomura, A., and Smith, M.A. (2005). Tau phosphorylation in Alzheimer's disease: pathogen or protector? Trends Mol Med *11*, 164-169.

Lescai, F., Marchegiani, F., and Franceschi, C. (2009). PON1 is a longevity gene: results of a meta-analysis. Ageing Res Rev *8*, 277-284.

Lewis, S.J., and Brunner, E.J. (2004). Methodological problems in genetic association studies of longevity--the apolipoprotein E gene as an example. Int J Epidemiol *33*, 962-970.

Li, H., Wetten, S., Li, L., St Jean, P.L., Upmanyu, R., Surh, L., Hosford, D., Barnes, M.R., Briley, J.D., Borrie, M.*, et al.* (2008). Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. Arch Neurol *65*, 45-53.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Li, H. (2011). Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics *27*, 718-719.

Liang, Y., Buckley, T.R., Tu, L., Langdon, S.D., and Tedder, T.F. (2001). Structural organization of the human MS4A gene cluster on Chromosome 11q12. Immunogenetics *53*, 357-368.

Liolitsa, D., Powell, J., and Lovestone, S. (2002). Genetic variability in the insulin signalling pathway may contribute to the risk of late onset Alzheimer's disease. J Neurol Neurosurg Psychiatry *73*, 261-266.

Lobo, A., Launer, L.J., Fratiglioni, L., Andersen, K., Di Carlo, A., Breteler, M.M., Copeland, J.R., Dartigues, J.F., Jagger, C., Martinez-Lage, J.*, et al.* (2000). Prevalence of dementia and major subtypes in Europe: A collaborative study of population-based cohorts. Neurologic Diseases in the Elderly Research Group. Neurology *54*, S4-9.

Lovell, M.A., and Markesbery, W.R. (2007). Oxidative DNA damage in mild cognitive impairment and late-stage Alzheimer's disease. Nucleic Acids Res *35*, 7497-7504.

Lue, L.F., Kuo, Y.M., Roher, A.E., Brachova, L., Shen, Y., Sue, L., Beach, T., Kurth, J.H., Rydel, R.E., and Rogers, J. (1999). Soluble amyloid beta peptide concentration as a predictor of synaptic change in Alzheimer's disease. Am J Pathol *155*, 853-862.

Lynch, D.K., Winata, S.C., Lyons, R.J., Hughes, W.E., Lehrbach, G.M., Wasinger, V., Corthals, G., Cordwell, S., and Daly, R.J. (2003). A Cortactin-CD2-associated protein (CD2AP) complex provides a novel link between epidermal growth factor receptor endocytosis and the actin cytoskeleton. J Biol Chem *278*, 21805-21813.

Ma, S.L., Tang, N.L., Tam, C.W., Lui, V.W., Suen, E.W., Chiu, H.F., and Lam, L.C. (2008). Association between HLA-A alleles and Alzheimer's disease in a southern Chinese community. Dement Geriatr Cogn Disord *26*, 391-397.

Mahley, R.W. (1988). Apolipoprotein E: cholesterol transport protein with expanding role in cell biology. Science *240*, 622-630.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A.*, et al.* (2009). Finding the missing heritability of complex diseases. Nature *461*, 747-753.

Mardis, E.R. (2011). A decade's perspective on DNA sequencing technology. Nature *470*, 198-203.

Marth, G.T., Korf, I., Yandell, M.D., Yeh, R.T., Gu, Z., Zakeri, H., Stitziel, N.O., Hillier, L., Kwok, P.Y., and Gish, W.R. (1999). A general approach to single-nucleotide polymorphism discovery. Nat Genet *23*, 452-456.

Martin, G.M., Bergman, A., and Barzilai, N. (2007). Genetic determinants of human health span and life span: progress and new opportunities. PLoS Genet *3*, e125.

Martins, I.J., Hone, E., Foster, J.K., Sunram-Lea, S.I., Gnjec, A., Fuller, S.J., Nolan, D., Gandy, S.E., and Martins, R.N. (2006). Apolipoprotein E, cholesterol metabolism, diabetes, and the convergence of risk factors for Alzheimer's disease and cardiovascular disease. Mol Psychiatry *11*, 721-736.

Marzolo, M.P., and Bu, G. (2009). Lipoprotein receptors and cholesterol in APP trafficking and proteolytic processing, implications for Alzheimer's disease. Semin Cell Dev Biol *20*, 191-200.

May, P.C., Lampert-Etchells, M., Johnson, S.A., Poirier, J., Masters, J.N., and Finch, C.E. (1990). Dynamics of gene expression for a hippocampal glycoprotein elevated in Alzheimer's disease and in response to experimental lesions in rat. Neuron *5*, 831-839.

Maynard, C.J., Bush, A.I., Masters, C.L., Cappai, R., and Li, Q.X. (2005). Metals and amyloid-beta in Alzheimer's disease. Int J Exp Pathol *86*, 147-159.

McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P., and Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet *9*, 356-369.

McGue, M., Vaupel, J.W., Holm, N., and Harvald, B. (1993). Longevity is moderately heritable in a sample of Danish twins born 1870-1880. J Gerontol 48, B237-244.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., and Stadlan, E.M. (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*, 939-944.

McNab, F.W., Rajsbaum, R., Stoye, J.P., and O'Garra, A. (2010). Tripartite-motif proteins and innate immune regulation. Curr Opin Immunol *23*, 46-56.

Metzker, M.L. (2005). Emerging technologies in DNA sequencing. Genome Res *15*, 1767-1776.

Metzker, M.L. (2010). Sequencing technologies - the next generation. Nat Rev Genet *11*, 31-46.

Miller, E.R., 3rd, Pastor-Barriuso, R., Dalal, D., Riemersma, R.A., Appel, L.J., and Guallar, E. (2005). Meta-analysis: high-dosage vitamin E supplementation may increase all-cause mortality. Ann Intern Med *142*, 37-46.

Mirra, S.S., Heyman, A., McKeel, D., Sumi, S.M., Crain, B.J., Brownlee, L.M., Vogel, F.S., Hughes, J.P., van Belle, G., and Berg, L. (1991). The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part II. Standardization of the neuropathologic assessment of Alzheimer's disease. Neurology *41*, 479-486.

Morgan, K. (2011). The three new pathways leading to Alzheimer's disease. Neuropathol Appl Neurobiol *37*, 353-357.

Mucke, L. (2009). Neuroscience: Alzheimer's disease. Nature *461*, 895-897.

Naj, A.C., Jun, G., Beecham, G.W., Wang, L.S., Vardarajan, B.N., Buros, J., Gallins, P.J., Buxbaum, J.D., Jarvik, G.P., Crane, P.K.*, et al.* (2011). Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. Nat Genet *43*, 436-441.

Neuropathology Group of the Medical Research Council Cognitive Function and Ageing Study (2001). Pathological correlates of late-onset dementia in a multicentre, community-based population in England and Wales. Neuropathology Group of the Medical Research Council Cognitive Function and Ageing Study (MRC CFAS). Lancet *357*, 169-175.

Nielsen, R., Paul, J.S., Albrechtsen, A., and Song, Y.S. (2011). Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet *12*, 443-451.

Orgogozo, J.M., Gilman, S., Dartigues, J.F., Laurent, B., Puel, M., Kirby, L.C., Jouanny, P., Dubois, B., Eisner, L., Flitman, S*., et al.* (2003). Subacute meningoencephalitis in a subset of patients with AD after Abeta42 immunization. Neurology *61*, 46-54.

Ovcharenko, I., Nobrega, M.A., Loots, G.G., and Stubbs, L. (2004). ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. Nucleic Acids Res *32*, W280-286.

Panza, F., D'Introno, A., Capurso, C., Colacicco, A.M., Seripa, D., Pilotto, A., Santamato, A., Capurso, A., and Solfrizzi, V. (2007). Lipoproteins, vascular-related genetic factors, and human longevity. Rejuvenation Res *10*, 441-458.

Panza, F., Frisardi, V., Capurso, C., D'Introno, A., Colacicco, A.M., Seripa, D., Pilotto, A., Vendemiale, G., Capurso, A., and Solfrizzi, V. (2009). Apolipoprotein E, dementia, and human longevity. J Am Geriatr Soc *57*, 740-742.

Pardossi-Piquard, R., Petit, A., Kawarai, T., Sunyach, C., Alves da Costa, C., Vincent, B., Ring, S., D'Adamio, L., Shen, J., Muller, U*., et al.* (2005). Presenilin-dependent transcriptional control of the Abeta-degrading enzyme neprilysin by intracellular domains of betaAPP and APLP. Neuron *46*, 541-554.

Perls, T.T., Bubrick, E., Wager, C.G., Vijg, J., and Kruglyak, L. (1998). Siblings of centenarians live longer. Lancet *351*, 1560.

Plomin, R., Haworth, C.M., and Davis, O.S. (2009). Common disorders are quantitative traits. Nat Rev Genet *10*, 872-878.

Polito, L., Kehoe, P.G., Forloni, G., and Albani, D. (2010). The molecular genetics of sirtuins: association with human longevity and age-related diseases. Int J Mol Epidemiol Genet *1*, 214-225.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet *38*, 904-909.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J*., et al.* (2007). PLINK: a tool set for whole-

genome association and population-based linkage analyses. Am J Hum Genet *81*, 559-575.

Qiu, W.Q., and Folstein, M.F. (2006). Insulin, insulin-degrading enzyme and amyloid-beta peptide in Alzheimer's disease: review and hypothesis. Neurobiol Aging *27*, 190-198.

Qiu, W.Q., Walsh, D.M., Ye, Z., Vekrellis, K., Zhang, J., Podlisny, M.B., Rosner, M.R., Safavi, A., Hersh, L.B., and Selkoe, D.J. (1998). Insulin-degrading enzyme regulates extracellular levels of amyloid beta-protein by degradation. J Biol Chem *273*, 32730-32738.

Quinn, J.F., Raman, R., Thomas, R.G., Yurko-Mauro, K., Nelson, E.B., Van Dyck, C., Galvin, J.E., Emond, J., Jack, C.R., Jr., Weiner, M.*, et al.* (2010). Docosahexaenoic acid supplementation and cognitive decline in Alzheimer disease: a randomized trial. JAMA *304*, 1903-1911.

Reiman, E.M., Webster, J.A., Myers, A.J., Hardy, J., Dunckley, T., Zismann, V.L., Joshipura, K.D., Pearson, J.V., Hu-Lince, D., Huentelman, M.J.*, et al.* (2007). GAB2 alleles modify Alzheimer's risk in APOE epsilon4 carriers. Neuron *54*, 713-720.

Reiman, E.M., Chen, K., Liu, X., Bandy, D., Yu, M., Lee, W., Ayutyanont, N., Keppler, J., Reeder, S.A., Langbaum, J.B.*, 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*, 6820-6825.

Ritchie, C.W., Bush, A.I., Mackinnon, A., Macfarlane, S., Mastwyk, M., MacGregor, L., Kiers, L., Cherny, R., Li, Q.X., Tammer, A.*, et al.* (2003). Metal-protein attenuation with iodochlorhydroxyquin (clioquinol) targeting Abeta amyloid deposition and toxicity in Alzheimer disease: a pilot phase 2 clinical trial. Arch Neurol *60*, 1685-1691.

Rivas, M.A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C.K., Boucher, G., Ripke, S., Ellinghaus, D., Burtt, N.*, et al.* (2011). Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet *43*, 1066-1073.

Robinson, D.M., and Keating, G.M. (2006). Memantine: a review of its use in Alzheimer's disease. Drugs *66*, 1515-1534.

Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat Biotechnol *29*, 24-26.

Ronaghi, M., Uhlen, M., and Nyren, P. (1998). A sequencing method based on real-time pyrophosphate. Science *281*, 363, 365.

Rosvall, L., Rizzuto, D., Wang, H.X., Winblad, B., Graff, C., and Fratiglioni, L. (2009). APOE-related mortality: effect of dementia, cardiovascular disease and gender. Neurobiol Aging *30*, 1545-1551.

Rozen, S., and Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol *132*, 365-386.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A *74*, 5463-5467.

Sato-Harada, R., Okabe, S., Umeyama, T., Kanai, Y., and Hirokawa, N. (1996). Microtubule-associated proteins regulate microtubule function as the track for intracellular membrane organelle transports. Cell Struct Funct *21*, 283-295.

Saunders, A.M., Strittmatter, W.J., Schmechel, D., George-Hyslop, P.H., Pericak-Vance, M.A., Joo, S.H., Rosi, B.L., Gusella, J.F., Crapper-MacLachlan, D.R., Alberts, M.J.*, et al.* (1993). Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. Neurology *43*, 1467-1472.

Schmitt, A.O., Assmus, J., Bortfeldt, R.H., and Brockmann, G.A. (2010). CandiSNPer: a web tool for the identification of candidate SNPs for causal variants. Bioinformatics *26*, 969-970.

Sebastiani, P., Montano, M., Puca, A., Solovieff, N., Kojima, T., Wang, M.C., Melista, E., Meltzer, M., Fischer, S.E., Andersen, S.*, et al.* (2009). RNA editing genes associated with extreme old age in humans and with lifespan in C. elegans. PLoS One *4*, e8210.

Selkoe, D.J. (1991). The molecular pathology of Alzheimer's disease. Neuron *6*, 487-498.

Seshadri, S., Fitzpatrick, A.L., Ikram, M.A., DeStefano, A.L., Gudnason, V., Boada, M., Bis, J.C., Smith, A.V., Carassquillo, M.M., Lambert, J.C.*, et al.* (2010). Genome-wide analysis of genetic loci associated with Alzheimer disease. JAMA *303*, 1832-1840.

Shepherd, C., McCann, H., and Halliday, G.M. (2009). Variations in the neuropathology of familial Alzheimer's disease. Acta Neuropathol.

Sherva, R., and Farrer, L.A. (2011). Power and pitfalls of the genome-wide association study approach to identify genes for Alzheimer's disease. Curr Psychiatry Rep *13*, 138-146.

Shibata, M., Yamada, S., Kumar, S.R., Calero, M., Bading, J., Frangione, B., Holtzman, D.M., Miller, C.A., Strickland, D.K., Ghiso, J.*, et al.* (2000). Clearance of Alzheimer's amyloid-ss(1-40) peptide from brain by LDL receptor-related protein-1 at the blood-brain barrier. J Clin Invest *106*, 1489-1499.

Shiina, T., Ota, M., Shimizu, S., Katsuyama, Y., Hashimoto, N., Takasu, M., Anzai, T., Kulski, J.K., Kikkawa, E., Naruse, T.*, et al.* (2006). Rapid evolution of major histocompatibility complex class I genes in primates generates new disease alleles in humans via hitchhiking diversity. Genetics *173*, 1555-1570.

Sleegers, K., Lambert, J.C., Bertram, L., Cruts, M., Amouyel, P., and Van Broeckhoven, C. (2010). The pursuit of susceptibility genes for Alzheimer's disease: progress and prospects. Trends Genet.

Small, S.A., and Duff, K. (2008). Linking Abeta and tau in late-onset Alzheimer's disease: a dual pathway hypothesis. Neuron *60*, 534-542.

Stewart, R., and Liolitsa, D. (1999). Type 2 diabetes mellitus, cognitive impairment and dementia. Diabet Med *16*, 93-112.

Sucher, N.J., Awobuluyi, M., Choi, Y.B., and Lipton, S.A. (1996). NMDA receptors: from genes to channels. Trends Pharmacol Sci *17*, 348-355.

Suh, Y., Atzmon, G., Cho, M.O., Hwang, D., Liu, B., Leahy, D.J., Barzilai, N., and Cohen, P. (2008). Functionally significant insulin-like growth factor I receptor mutations in centenarians. Proc Natl Acad Sci U S A *105*, 3438-3442.

Swerdlow, R.H. (2011). Brain aging, Alzheimer's disease, and mitochondria. Biochim Biophys Acta *1812*, 1630-1639.

Tan, Q., Zhao, J.H., Li, S., Kruse, T.A., and Christensen, K. (2010). Power assessment for genetic association study of human longevity using offspring of long-lived subjects. Eur J Epidemiol *25*, 501-506.

Tanzi, R.E., Moir, R.D., and Wagner, S.L. (2004). Clearance of Alzheimer's Abeta peptide: the many roads to perdition. Neuron *43*, 605-608.

Tanzi, R.E., and Bertram, L. (2005). Twenty years of the Alzheimer's disease amyloid hypothesis: a genetic perspective. Cell *120*, 545-555.

Tapiola, T., Alafuzoff, I., Herukka, S.K., Parkkinen, L., Hartikainen, P., Soininen, H., and Pirttila, T. (2009). Cerebrospinal fluid {beta}-amyloid 42 and tau proteins as biomarkers of Alzheimer-type pathologic changes in the brain. Arch Neurol *66*, 382-389.

Tariot, P.N., Farlow, M.R., Grossberg, G.T., Graham, S.M., McDonald, S., and Gergel, I. (2004). Memantine treatment in patients with moderate to severe Alzheimer disease already receiving donepezil: a randomized controlled trial. JAMA *291*, 317-324.

Tebar, F., Bohlander, S.K., and Sorkin, A. (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*, 2687-2702.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061-1073.

The International HapMap Consortium (2003). The International HapMap Project. Nature *426*, 789-796.

Tiraboschi, P., Hansen, L.A., Masliah, E., Alford, M., Thal, L.J., and Corey-Bloom, J. (2004). Impact of APOE genotype on neuropathologic and neurochemical markers of Alzheimer disease. Neurology *62*, 1977-1983.

Treangen, T.J., and Salzberg, S.L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet *13*, 36-46.

Turner, S., Armstrong, L.L., Bradford, Y., Carlson, C.S., Crawford, D.C., Crenshaw, A.T., de Andrade, M., Doheny, K.F., Haines, J.L., Hayes, G.*, et al.* (2011). Quality control procedures for genome-wide association studies. Curr Protoc Hum Genet *Chapter 1*, Unit1 19.

van Es, M.A., and van den Berg, L.H. (2009). Alzheimer's disease beyond APOE. Nat Genet *41*, 1047-1048.

Vijg, J. (2009). SNP'ing for longevity. Aging (Albany NY) *1*, 442-443.

Wallis, Y., and Morrell, N. (2010). Automated DNA sequencing. Methods Mol Biol *688*, 173-185.

Wang, J., Zhang, S., Wang, Y., Chen, L., and Zhang, X.S. (2009). Disease-aging network reveals significant roles of aging genes in connecting genetic diseases. PLoS Comput Biol *5*, e1000521.

Wang, K., Dickson, S.P., Stolle, C.A., Krantz, I.D., Goldstein, D.B., and Hakonarson, H. (2010). Interpretation of association signals and identification of causal variants from genome-wide association studies. Am J Hum Genet *86*, 730-742.

Wang, N., Akey, J.M., Zhang, K., Chakraborty, R., and Jin, L. (2002). Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. Am J Hum Genet *71*, 1227-1234.

Winblad, B., Engedal, K., Soininen, H., Verhey, F., Waldemar, G., Wimo, A., Wetterholm, A.L., Zhang, R., Haglund, A., and Subbiah, P. (2001). A 1-year, randomized, placebo-controlled study of donepezil in patients with mild to moderate AD. Neurology *57*, 489-495.

Woo, J.S., Imm, J.H., Min, C.K., Kim, K.J., Cha, S.S., and Oh, B.H. (2006). Structural and functional insights into the B30.2/SPRY domain. EMBO J *25*, 1353-1363.

Wray, N.R., Purcell, S.M., and Visscher, P.M. (2011). Synthetic associations created by rare variants do not explain most GWAS results. PLoS Biol *9*, e1000579.

Wyss-Coray, T., Yan, F., Lin, A.H., Lambris, J.D., Alexander, J.J., Quigg, R.J., and Masliah, E. (2002). Prominent neurodegeneration and increased plaque formation in complement-inhibited Alzheimer's mice. Proc Natl Acad Sci U S A *99*, 10837-10842.

Yaari, R., and Corey-Bloom, J. (2007). Alzheimer's disease. Semin Neurol *27*, 32-41.

Yang, Q., Khoury, M.J., Friedman, J.M., and Flanders, W.D. (2003). On the use of population attributable fraction to determine sample size for case-control studies of gene-environment interaction. Epidemiology *14*, 161-167.

Yin, G.N., Lee, H.W., Cho, J.Y., and Suk, K. (2009). Neuronal pentraxin receptor in cerebrospinal fluid as a potential biomarker for neurodegenerative diseases. Brain Res *1265*, 158-170.

Yu, C.E., Seltman, H., Peskind, E.R., Galloway, N., Zhou, P.X., Rosenthal, E., Wijsman, E.M., Tsuang, D.W., Devlin, B., and Schellenberg, G.D. (2007).

Comprehensive analysis of APOE and selected proximate markers for late-onset Alzheimer's disease: patterns of linkage disequilibrium and disease/marker association. Genomics *89*, 655-665.

Zhang, Y., Tian, Q., Zhang, Q., Zhou, X., Liu, S., and Wang, J.Z. (2009). Hyperphosphorylation of microtubule-associated tau protein plays dual role in neurodegeneration and neuroprotection. Pathophysiology.

# 8. Appendices

## 8.1 DNA samples sequenced by next generation sequencing.

*TRIM15* gene was sequenced in 150 samples (75 AD cases and controls) by ABI SOLiD® next generation sequencing. The study is described in **Chapter 4**.

| TRIM15 A | | | | TRIM15 B | | | |
|---|---|---|---|---|---|---|---|
| AD cases | | control | | AD cases | | control | |
| AD203 | M390 | L1 | L67 | AD203 | M41 | L11 | L69 |
| AD206 | M392 | L12 | L68 | AD208 | M411 | L12 | L7 |
| AD207 | M393 | L13 | L69 | AD209 | M419 | L13 | L71 |
| AD208 | M394 | L14 | L7 | AD210 | M424 | L15 | L72 |
| AD209 | M395 | L16 | L71 | AD211 | M426 | L16 | L73 |
| AD210 | M396 | L17 | L72 | AD212 | M429 | L17 | L74 |
| AD212 | M397 | L18 | L74 | AD213 | M438 | L18 | L75 |
| AD215 | M398 | L19 | L75 | AD216 | M447 | L19 | L76 |
| AD216 | M400 | L2 | L77 | AD218 | M455 | L2 | L77 |
| AD217 | M401 | L20 | L80 | AD219 | M46 | L21 | L79 |
| AD218 | M41 | L21 | L83 | AD221 | M503 | L22 | L8 |
| AD223 | M410 | L23 | L85 | AD222 | M504 | L23 | L80 |
| AD224 | M411 | L24 | L86 | AD223 | M505 | L25 | L81 |
| AD226 | M419 | L29 | L87 | AD224 | M506 | L28 | L82 |
| AD227 | M424 | L3 | L88 | AD225 | M507 | L29 | L83 |
| AD229 | M426 | L31 | L9 | AD226 | M508 | L3 | L84 |
| AD232 | M429 | L32 | L91 | AD232 | M509 | L31 | L85 |
| AD233 | M438 | L34 | L92 | AD233 | M510 | L32 | L86 |
| AD235 | M441 | L35 | L95 | M032 | M511 | L33 | L87 |
| AD236 | M447 | L38 | L96 | M050 | M512 | L35 | L88 |
| M004 | M455 | L4 | N135 | M111 | M513 | L38 | L89 |
| M005 | M503 | L41 | N138 | M123 | M514 | L4 | L91 |
| M021 | M504 | L46 | N139 | M125 | M515 | L41 | L92 |
| M022 | M505 | L47 | N140 | M325 | M517 | L44 | L93 |
| M026 | M506 | L49 | N141 | M38 | M518 | L47 | L95 |
| M050 | M507 | L50 | N142 | M385 | M519 | L49 | L96 |
| M053 | M508 | L51 | N143 | M388 | M520 | L50 | N133 |
| M070 | M509 | L53 | N145 | M389 | M521 | L51 | N134 |
| M102 | M512 | L54 | N146 | M390 | M59 | L53 | N138 |
| M106 | M513 | L55 | N148 | M392 | M616 | L54 | N139 |
| M111 | M515 | L56 | N149 | M393 | M618 | L56 | N141 |
| M125 | M517 | L59 | N150 | M394 | M619 | L6 | N142 |
| M32 | M519 | L6 | N151 | M395 | M620 | L62 | N143 |
| M325 | M520 | L60 | N152 | M396 | M624 | L63 | N145 |
| M38 | M521 | L61 | N153 | M397 | M626 | L64 | N148 |
| M385 | M619 | L62 | N155 | M398 | M632 | L65 | N154 |
| M388 | M633 | L64 | N156 | M400 | M77 | L67 | N156 |
| M389 | | L65 | | M401 | | L68 | |

## 8.2 Variant Classifier Input File

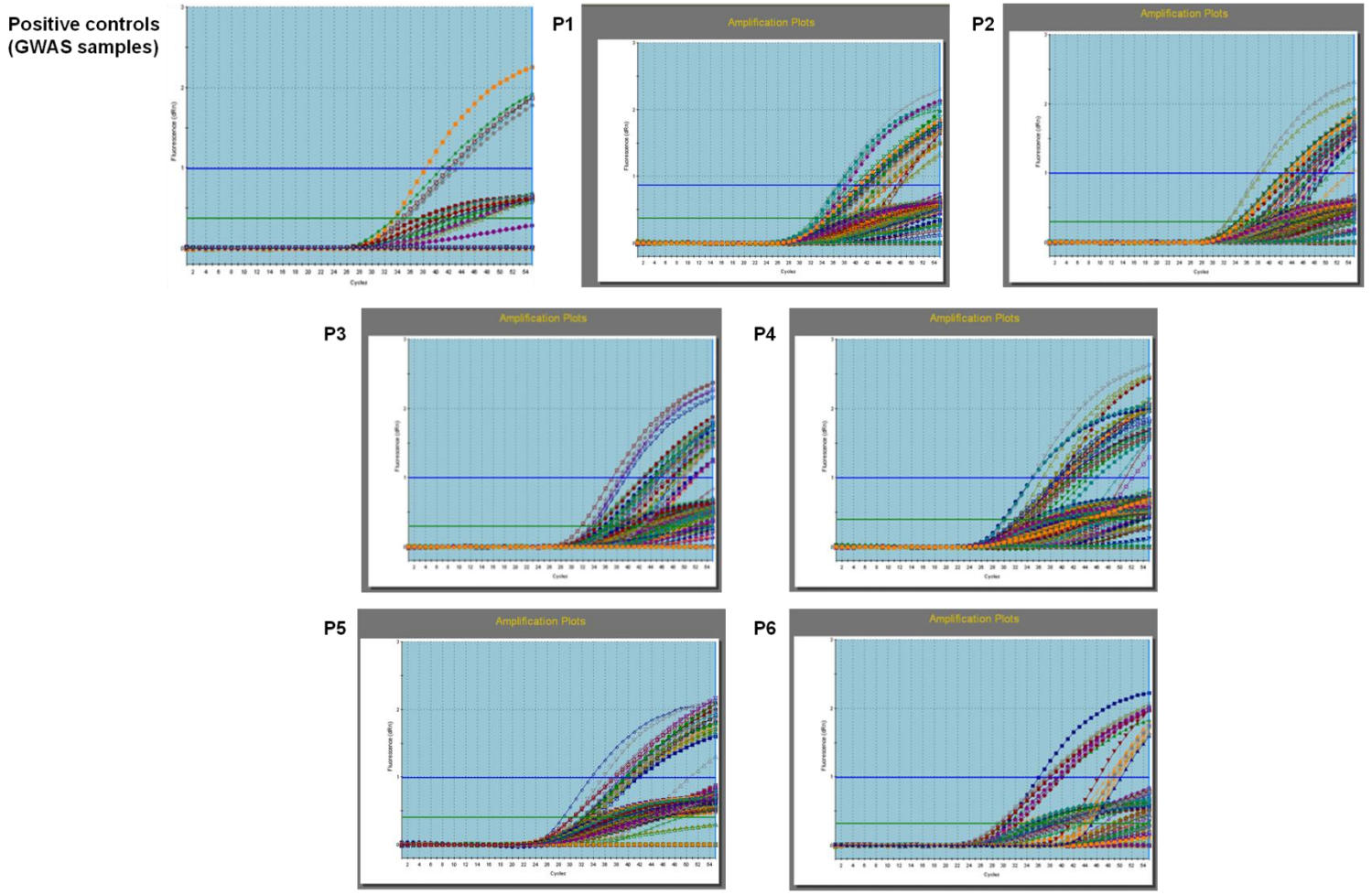
High Quality SNPs identified in *TRIM15* 'A' and 'B' amplicons using next generation sequencing were annotated using Variant Classifier (**Chapter 4**). The input file for the bioinformatic tool is shown below.

---------- VC_input.txt ----------

| | | | |
|---|---|---|---|
| 91 | 92 | 1 | G |
| 252 | 253 | 1 | A |
| 758 | 759 | 1 | G |
| 966 | 967 | 1 | C |
| 984 | 985 | 1 | C |
| 1138 | 1139 | 1 | C |
| 1150 | 1151 | 1 | A |
| 1162 | 1163 | 1 | T |
| 1181 | 1182 | 1 | G |
| 1193 | 1194 | 1 | T |
| 1220 | 1221 | 1 | A |
| 1346 | 1347 | 1 | C |
| 1399 | 1400 | 1 | G |
| 1670 | 1671 | 1 | A |
| 1735 | 1736 | 1 | C |
| 1949 | 1950 | 1 | A |
| 8111 | 8112 | 1 | C |
| 8124 | 8125 | 1 | A |
| 8232 | 8233 | 1 | A |
| 8280 | 8281 | 1 | A |
| 8297 | 8298 | 1 | T |
| 8488 | 8489 | 1 | A |
| 8500 | 8501 | 1 | C |
| 8530 | 8531 | 1 | T |
| 8656 | 8657 | 1 | T |
| 8790 | 8791 | 1 | G |
| 9031 | 9032 | 1 | G |
| 9112 | 9113 | 1 | T |
| 9334 | 9335 | 1 | A |
| 9977 | 9978 | 1 | C |
| 10175 | 10176 | 1 | T |
| 10548 | 10549 | 1 | T |
| 10677 | 10678 | 1 | G |
| 10839 | 10840 | 1 | T |
| 11888 | 11889 | 1 | C |
| 11900 | 11901 | 1 | - |

| | | | |
|---|---|---|---|
| 12093 | 12094 | 1 | A |
| 12309 | 12310 | 1 | A |
| 12325 | 12326 | 1 | A |
| 12634 | 12635 | 1 | G |

## 8.3 Amplification curve for TaqMan® genotyping of SNP rs4110518




The diagram depicts fluorescence intensities for TaqMan® genotyping of SNP rs4110518. Each amplification curve corresponds to a dual scatter plot as shown in **Figure 5.7** in **Chapter 5**.

## 8.4 PERL programs

Respective specific utilities of each PERL script listed are described in **Methods**

**2.3.2**.

### 8.4.1 Determination of common SNPs between different chip platforms

---------- perl.pl ----------

```
use strict;
use warnings;
use modules;

my $filename1 = 'mydata.bim'; # input SNP file in PLINK format
my $filename2 = 'SNPs.txt';   # a list of SNP from GWAS
my $str = modules::readData($filename1, $filename2);
```

---------- modules.pm ----------

```
package modules;

use strict;
use warnings;

# readData subroutine
sub readData {
my $filename1 = $_[0];
my $filename2 = $_[1];
my @array;
my %hash;
my %hashtwo;
open (OUT, ">results.txt") or die "Unable to open results.txt: $!\n";
open (FILE, "<$filename2") or die "Unable to open $filename2: $!\n";
my $count=0;
while (<FILE>){ chomp;
if (/^(\w+)/) {$hash{$1} ="unmatched";$hashtwo{$count}=$1;$count++;}}
open (MYFILE, "<$filename1") or die "Unable to open $filename1: $!\n";
while (<MYFILE>) { chomp;
@array = split(' ',$_);
my $SNP = $array[1]; # it specifies the column where the SNP is
located
if ($hash{$SNP}) {
# QT indexed SNPs looking for same SNPs in CC results
$hash{$SNP} = "$_";
foreach my $keys (sort {$a <=> $b} keys %hashtwo) {
print "$hashtwo{$keys}\t$hash{$hashtwo{$keys}}\n";
print OUT "$hashtwo{$keys}\t$hash{$hashtwo{$keys}}\n";}
close MYFILE;
close FILE;
close OUT;}

1;
```

### 8.4.2 Calculation of number of independent tests

---------- number_independent_test.pl ----------

```
use strict;
use warnings;
use modules;

my $filename1 = 'mydata.ld'; # .ld file from PLINK analysis
my $str = modules::readData($filename1);
```

---------- modules.pm ----------

```
package modules;
use strict;
use warnings;

# readData subroutine
sub readData {
my $input1 = $_[0]; my @array; my $col1; my $col2; my %hash;
open (MYFILE, "<$input1") or die "Unable to open $input1: $!\n";
my $count=0; my $count2=1;
while (<MYFILE>) { chomp; @array = split("\t",$_);
$col1=$array[0];$col2=$array[1];
if ($count==0) {$hash{$col1} = 1; $hash{$col2} = 1;
} else {
if (exists $hash{$col1} && exists $hash{$col2}){
if ($hash{$col1} == $hash{$col2}) { next;
} else {
if ($hash{$col1} > $hash{$col2}) { my $stringTemp = $hash{$col1};
foreach my $keys (keys %hash) {
if ($hash{$keys}==$stringTemp) {
$hash{$keys} = $hash{$col2};
}}} else { my $stringTemp = $hash{$col2};
foreach my $keys (keys %hash) {
if ($hash{$keys}==$stringTemp) { $hash{$keys} = $hash{$col1};
}}} $count2 = $count2-1;}}
elsif (exists $hash{$col1} || exists $hash{$col2}){
if (exists $hash{$col1}) {
$hash{$col2} = $hash{$col1};}
if (exists $hash{$col2}) {
$hash{$col1} = $hash{$col2};}} else {
$hash{$col1} = $count2+1; $hash{$col2} = $count2+1; $count2++;}}
$count++;}
close MYFILE;
open (OUT, ">results.txt") or die "Unable to open results.txt: $!\n";
my %hashtwo; my $count3=0;
foreach my $keys (sort {$hash{$b} <=> $hash{$a}} keys %hash) {
$count3++;
print OUT "$keys => $hash{$keys} => $count3\n";
$hashtwo{$hash{$keys}}=0; }
# $hashtwo counts the number unique values from using $hashone
my $counttwo=0;
foreach my $keystwo (keys %hashtwo) { $counttwo++;}
print "The number of LD clusters is: $counttwo\n";
print "Check [ results.txt ] for details of LD clusters";
close OUT;}

1;
```

### 8.4.3 GWAS meta-analysis

---------- meta_analysis.pl ----------

```
use strict;
use warnings;
use modules;
use Time::Local;

# type in file name
my $filename1 = 'test.txt';

# type in 'CC' for Case/Control analysis or 'QT' or Quantitative
Trait analysis, the default is for Case/Control
# analysis. NB: This parameter is for checking SNP 'flippers',
failure to adjust the parameter for correct type of
# analysis will result in incorrect output in the results file.
my $analysis = 'CC';

# Do NOT edit after this line
my $time = localtime;
print "Analysis started: $time\n";
my $arrayRef = modules::readData($filename1);
modules::printHeader($$arrayRef[1]);
print "Writting results to file [results.txt]\n";
my $SNP = modules::analyzeData($$arrayRef[0],$$arrayRef[1],$analysis);
$time = localtime;
print "Analysis finished: $time\n";
```

---------- modules.pm ----------

```
package modules;
use strict;
use warnings;

# readData subroutine
sub readData {
my $filename1 = $_[0]; my @arrayOne; my %hash; my $string;
open (OUT, ">results.txt") or die "Unable to open results.txt: $!\n";
open (FILE, "<$filename1") or die "Unable to open $filename1: $!\n";
my $countAll=0; my $count=0; my $countProxy=0; my $totalProxy;
my $fileTag; my $SNPID; my $string1; my $string2; my $string3;
my $fileNo=0;
while (<FILE>){
chomp;
if (/^(\s+)(\S+)/) {
if ($count==1) { $countAll++; @arrayOne = split(" ", $_);
$SNPID = $arrayOne[2];
$totalProxy=$arrayOne[5];#TOTAL proxy
$fileTag = $arrayOne[1];
$hash{$SNPID}The 1000 Genomes Project Consortium,
=$countAll;#sequence
$hash{$SNPID}{2}=$arrayOne[1];#file number
$hash{$SNPID}{3}=$arrayOne[0];#chromosome number
$hash{$SNPID}{4}=$arrayOne[3];#bp number
$hash{$SNPID}{5}=$arrayOne[4];#p-value
if ($arrayOne[1] > $fileNo) {
$fileNo = $arrayOne[1]; }}
if ($count==3 && $totalProxy ==0) {print "[Error1]\n";die;} #check
file integrity 1
if ($count==3) {
@arrayOne = split(" ", $_);
```

```
if ($arrayOne[1] eq $SNPID) {} else {print "[Error2]\n";die;} #check
file integrity 2
$hash{$SNPID}{6}=$arrayOne[7];#odds ratio
} if ($count>3) {
my $proxyID=0;
@arrayOne = split(" ", $_);
if (scalar(@arrayOne)==7) {
my $checkFile = checkFile($fileTag,$arrayOne[4]);
$string1 = $arrayOne[0];
$string2 = $arrayOne[1];
$string3 = $arrayOne[2];
if ($checkFile==1) {
$countProxy++;
$proxyID = $arrayOne[0];
my $fileID = $arrayOne[4];
$hash{$SNPID}{7}{$fileID}{$proxyID}The 1000 Genomes Project
Consortium, =$countProxy; #proxy count
$hash{$SNPID}{7}{$fileID}{$proxyID}{2}=$arrayOne[1]; #proxy kb
$hash{$SNPID}{7}{$fileID}{$proxyID}{3}=$arrayOne[2]; #proxy RSQ
$hash{$SNPID}{7}{$fileID}{$proxyID}{4}=$arrayOne[5]; #proxy P-value
$hash{$SNPID}{7}{$fileID}{$proxyID}{5}=$arrayOne[6]; #proxy OR
$hash{$SNPID}{8}=1; #total proxy
if ($fileID > $fileNo) {
$fileNo = $fileID;}}}
if (scalar(@arrayOne)==3) {
@arrayOne = split(" ", $_);
my $checkFile = checkFile($fileTag,$arrayOne[0]);
if ($checkFile==1) {
$countProxy++;
$proxyID =$string1; #proxy ID
my $fileID = $arrayOne[0];
$hash{$SNPID}{7}{$fileID}{$proxyID}The 1000 Genomes Project
Consortium,  = $countProxy; #proxy count
$hash{$SNPID}{7}{$fileID}{$proxyID}{2} = $string2; #proxy KB
$hash{$SNPID}{7}{$fileID}{$proxyID}{3} = $string3; #proxy RSQ
$hash{$SNPID}{7}{$fileID}{$proxyID}{4} = $arrayOne[1]; #proxy P-value
$hash{$SNPID}{7}{$fileID}{$proxyID}{5} = $arrayOne[2]; #proxy OR
$hash{$SNPID}{8}=1; #total proxy
if ($fileID > $fileNo) {
$fileNo = $fileID;
}}}} $count++; }
if (/^(\-)/) { $count=0; $totalProxy=0; $countProxy=0;}}
my @array = (\%hash,\$fileNo);
return \@array; close FILE; close OUT;}

# checkFile subroutine
sub checkFile {
my $input1 = $_[0];
my $input2 = $_[1];
if ($input1==$input2) {
return 0;
} else { return 1; }}

# analyzeData subroutine
sub analyzeData {
open (OUT, ">>results.txt") or die "Unable to open results.txt: $!\n";
my $hash = $_[0]; my $input = $_[1]; my $test = $_[2];
my $count1=1; my $maxFile = $$input-1;
foreach my $key (sort {${$hash}{$a}The 1000 Genomes Project
Consortium,  <=> ${$hash}{$b}The 1000 Genomes Project Consortium, }
keys %{$hash}) { #key is the indexID
```

```
my $totalProxy = ${$hash}{$key}{8};
if (defined $totalProxy) {
my $fileNo = ${$hash}{$key}{2};
my $indexPvalue = ${$hash}{$key}{5};
my $indexOR = ${$hash}{$key}{6};
my $indexCHR = ${$hash}{$key}{3};
my $proxyNo = ${$hash}{$key}The 1000 Genomes Project Consortium, ;
print OUT "$count1\t$key\t$fileNo\t$indexCHR\t";
my %hash2; my %hash3; my $count=0;
$hash3{$fileNo}=$indexPvalue;
foreach my $fileID (keys %{${$hash}{$key}{7}}) {
$count++;
my $arrayRef = bestSNP(\%{${$hash}{$key}{7}{$fileID}});
my $proxyID = $$arrayRef[0];
my $proxyKB = ${$hash}{$key}{7}{$fileID}{$proxyID}{2};
my $proxyRSQ = ${$hash}{$key}{7}{$fileID}{$proxyID}{3};
my $proxyPvalue = ${$hash}{$key}{7}{$fileID}{$proxyID}{4};
my $proxyOR = ${$hash}{$key}{7}{$fileID}{$proxyID}{5};
$hash2{$fileID}The 1000 Genomes Project Consortium,  = $count;
$hash2{$fileID}{2} = $proxyPvalue;
$hash2{$fileID}{3} = $proxyOR;
$hash2{$fileID}{4} = $proxyID;
$hash2{$fileID}{5} = $fileID;
$hash2{$fileID}{6} = $proxyKB;
$hash2{$fileID}{7} = $proxyRSQ;
$hash3{$fileID}=$proxyPvalue; }
my $fisherPvalue = getFisher(\%hash2,$indexPvalue);
my $proxyCount = $$fisherPvalue[1];
foreach my $key2 (sort {$hash2{$a}The 1000 Genomes Project Consortium,
<=> $hash2{$b}The 1000 Genomes Project Consortium, } keys %hash2)
{ print OUT
"$hash2{$key2}{4}\t$hash2{$key2}{5}\t$hash2{$key2}{6}\t$hash2{$key2}{
7}\t";
printSpace($maxFile,$proxyCount,1); # space 1
}
my $flipper = getFlipper(\%hash2,$indexOR,$test); # $test = "C/C" or
"QT"
$fisherPvalue = sprintf "%.2e" ,$$fisherPvalue[0];
print OUT "$fisherPvalue\t";
if ($flipper==0) {print OUT "YES\t"; } elsif ($flipper==1) { print
OUT "NO\t";} else { print OUT "NA\t";}
foreach my $key2 (sort keys %hash3) {
for (my $emptySpace=1;$emptySpace<=$$input;$emptySpace++) {
if (defined $hash3{$emptySpace}) {
} else {
$hash3{$emptySpace}="-";}}}
foreach my $key3 (sort keys %hash3) {
print OUT "$hash3{$key3}\t";# all proxy pvalue
} print OUT "\n"; $count1++;}} close OUT;}

# printSpace subroutine
sub printSpace {
my $input1 = $_[0];
my $input2 = $_[1];
my $input3 = $_[2];
if ($input1 != $input2) {
my $DETA = $input1 - $input2;
if ($input3 ==1) {
for (my $count=0;$count<$DETA;$count++) {
print OUT "-\t-\t-\t-\t";}
} elsif ($input3 ==2) {
```

```
for (my $count=0;$count<$DETA;$count++) {
print OUT "-\t";}}}}

# printHeader subroutine
sub printHeader {
my $input1= $_[0];
open (OUT, ">results.txt") or die "Unable to open results.txt: $!\n";
print OUT "ClumpNo\tSNP\tF1\tCHR\t";
my $fileNo = $$input1;
for (my $count=1;$count<$fileNo;$count++) {
my $count2 = $count+1;
print OUT "PROXY$count\tF$count2\tKB$count\tRSQ$count\t"; }
print OUT "FISHER\tFLIPPER\t";
for (my $count=1;$count<=$fileNo;$count++) {
print OUT "Pvalue$count\t"; }
print OUT "\n"; close OUT;}

# bestSNP subroutine
sub bestSNP {
my $hash =$_[0]; my $count =0; my $count2=0; my $Num;
my $firstNum; my %hash2;
foreach my $key (keys %{$hash}) {
$count2++; $hash2{${$hash}{$key}{2}}=$key; }
foreach my $key (sort {${$hash}{$a}The 1000 Genomes Project
Consortium,  <=> ${$hash}{$b}The 1000 Genomes Project Consortium, }
keys %{$hash}) {
if ($count==0) { $firstNum = ${$hash}{$key}{2}; }
if ($count>0) {
$Num = ${$hash}{$key}{2}; $firstNum = compareKB($firstNum,$Num); }
$count++;
if ($count2 == $count) {
my @array = ($hash2{$firstNum},$count); return \@array; }
}}

#getFisher subroutine
sub getFisher {
my $hash = $_[0]; my $indexPvalue = $_[1];
my $firstNum; my $Num; my $count=0; my $count2=0;
foreach my $key (keys %{$hash}) { $count2++;}
foreach my $key (sort {${$hash}{$a}The 1000 Genomes Project
Consortium,  <=> ${$hash}{$b}The 1000 Genomes Project Consortium, }
keys %{$hash}) {
if ($count==0) { $firstNum = ${$hash}{$key}{2}; }
if ($count>0) {
$Num=${$hash}{$key}{2}; $firstNum = combineP($firstNum,$Num); }
$count++;
if ($count2 ==$count){
$firstNum = combineP($firstNum,$indexPvalue);
my @array = ($firstNum,$count); return \@array; }
}}

# getFlipper subroutine
sub getFlipper {
my $hash = $_[0]; my $indexOR = $_[1]; my $test = $_[2]; my $Num;
my $firstNum; my $count=0; my $count2=0; my $NAcheck=0;
my $testCheck=0;
foreach my $key (keys %{$hash}) { $count2++; }
if ($indexOR == -9) { $NAcheck=1; }
foreach my $key (keys %{$hash}) {
if (${$hash}{$key}{3} == -9) { $NAcheck=1; }}
if ($test eq 'CC') { # analysis check "CC" or "QT"
```

```
$testCheck=1;
} elsif ($test eq 'QT') {
$testCheck=2;
} else {
print "Unrecoginized analysis!";die;}
foreach my $key (sort {${$hash}{$a}The 1000 Genomes Project
Consortium,  <=> ${$hash}{$b}The 1000 Genomes Project Consortium, }
keys %{$hash}) {
if ($count==0) { $firstNum = ${$hash}{$key}{3}; }
if ($count>0) { $Num=${$hash}{$key}{3};
if ($firstNum==-9 && $Num != -9) { $firstNum=$Num;
} elsif ($firstNum == -9 && $Num == -9) { $firstNum = -9;
} elsif ($firstNum != -9 && $Num == -9) { $firstNum = $firstNum;
} else { if ($testCheck ==1) { $firstNum = compareOR($firstNum,$Num);
} elsif ($testCheck ==2) { $firstNum = compareBETA($firstNum,$Num); }
if ($firstNum ==0) { # '0' represent flipping effect
return 0; }}} $count++;
if ($count2 ==$count){
if ($firstNum==-9 || $indexOR == -9) { return 2;
} else { if ($testCheck ==1) {
$firstNum = compareOR($firstNum,$indexOR);
} elsif ($testCheck ==2) {
$firstNum = compareBETA($firstNum,$indexOR); }
if ($firstNum ==0) { return 0;
} elsif ($NAcheck ==0) { return 1;
} else { return 2;
}}}}}

# compareBETA subroutine
sub compareBETA {
my $input1 = $_[0]; my $input2 = $_[1];
if ($input1 > 0 && $input2 >0) { return $input1;
} elsif ($input1 <0 && $input2 <0) { return $input1;
} else { return 0; }}

# compareOR subroutine
sub compareOR {
my $input1 = $_[0]; my $input2 = $_[1];
if ($input1 > 1 && $input2 >1) { return $input1;
} elsif ($input1 <1 && $input2 <1) { return $input1;
} else { return 0;}}

# combineP subroutine
sub combineP {
my $input1 = $_[0]; my $input2 = $_[1]; my $sum; my $result;
$input1 = log(1/$input1); $input2 = log(1/$input2);
$sum = $input1 + $input2; $result = 1/exp($sum); return $result;}

# compareKB subroutine
sub compareKB {
my $input1 = $_[0]; my $input2 = $_[1];
my $a = abs($input1); my $b = abs($input2);
if ($a > $b) { return $input2;
} else { return $input1; }}

1;
```

### 8.4.4 Plug-in for PLINK ‘gene report’ function

---------- gene_report_plugin.pl ----------

```
use strict;
use warnings;
use modules;

my $filename1 = 'myfile.range.report'; # input PLINK gene report file
‘.range.report’ is entered here
my $str = modules::readData($filename1);
```

---------- modules.pm ----------

```
package modules;
use strict;
use warnings;

sub readData {
my $input1 = $_[0]; my @array; my $col1;
my $col2; my %hash;
open (MYFILE, "<$input1") or die "Unable to open $input1: $!\n";
open (OUT, ">results.txt") or die "Unable to open results.txt: $!\n";
print OUT "SNP\tCHR\tBP\tP\tBETA\tGENE\tLENGTH\tDIST\n";
while (<MYFILE>) {
chomp;
if (/^\s+.+rs/) {
@array = split(' ',$_);
$hash{3} = $array[2];
$hash{4} = $array[0];
$hash{5} = $array[12];
$hash{6} = $array[7];
$hash{7} = $array[3];
$hash{8} = $array[1];
print OUT "$hash{3}\t$hash{8}\t$hash{7}\t$hash{5}\t$hash{6}\t$hashThe
1000 Genomes Project Consortium, \t$hash{2}\t$hash{4}\n"; }
if (/^(\w+)/) {
@array = split(' ',$_);
$hash{1} = $array[0]; $hash{2} = modules::math($array[4]);
}}
close MYFILE; close OUT; }

sub math {
my $input1 = $_[0]; my $first_number; my $full_number;
if ($input1 =~ m/^(\d+)(\.\d+)\w+/) {
$first_number = $1-40; $full_number = $first_number.$2."kb";
} elsif ($input1 =~ m/^(\d+)\w+$/) {
$first_number = $1-40; $full_number = $first_number."kb";
} else { die "Error in the length of the gene!";
} return $full_number;}

1;
```

## 8.5 UCSC custom tracks input files

Custom tracks input file for the *TRIM15* 'A' amplicon (used in **Chapter 4**):

```
browser position chr6:30130365-30132348
track type=wiggle_0 name="power" description="Power to detect a singleton" visibility=full graphType=points color=0,0,0 priority=10
variableStep chrom=chr6 span=8
30130365    100
30130366    100
30130367    100
...... skipped
30132346    100
30132347    100
30132348    100

track name="TRIM15_A1" description="High Quality SNPs (Known)" itemRgb=On visibility=pack priority=20
chr6  30130455    30130456    rs60650863  800   +     30130455    30130456    0,0,0
chr6  30131122    30131123    rs17188113  800   +     30131122    30131123    0,0,255
chr6  30131330    30131331    rs62407492  800   +     30131330    30131331    0,0,255
chr6  30131348    30131349    rs9261536   800   +     30131348    30131349    0,0,255
chr6  30131502    30131503    rs35278640  800   +     30131502    30131503    0,255,0
chr6  30131514    30131515    rs2523733   800   +     30131514    30131515    0,255,0
chr6  30131526    30131527    rs11961941  800   +     30131526    30131527    0,255,0
chr6  30131545    30131546    rs17194460  800   +     30131545    30131546    255,0,0
chr6  30131584    30131585    rs17194467  800   +     30131584    30131585    255,0,0
chr6  30131710    30131711    rs17194474  800   +     30131710    30131711    255,0,0
chr6  30132034    30132035    rs2074477   800   +     30132034    30132035    0,0,0
chr6  30132099    30132100    rs41272587  800   +     30132099    30132100    0,0,0

track name="TRIM15_A2" description="High Quality SNPs (Novel)" itemRgb=On visibility=pack priority=30
chr6  30130616    30130617    novel 800   +     30130616    30130617    0,0,0
chr6  30131557    30131558    novel 800   +     30131557    30131558    255,0,0
chr6  30131763    30131764    novel 800   +     30131763    30131764    0,255,0
chr6  30132313    30132314    novel 800   +     30132313    30132314    0,0,0

track type=bedGraph name="-log10(p-value)" description="Association p-value" visibility=full altColor=0,0,255 priority=40 autoScale=off viewLimits=0:3
yLineOnOff=on yLineMark=1.301 yLineOnOff=on yLineMark=0
chr6  30130455    30130456    1.0691
chr6  30130616    30130617    0.3025
chr6  30131122    30131123    0
chr6  30131330    30131331    0
chr6  30131348    30131349    0.5186
chr6  30131502    30131503    0.1647
chr6  30131514    30131515    1.1396
chr6  30131526    30131527    0
chr6  30131545    30131546    0.8794
chr6  30131557    30131558    0.3025
chr6  30131584    30131585    0
chr6  30131710    30131711    0.2423
chr6  30131763    30131764    0.1647
chr6  30132034    30132035    0
chr6  30132099    30132100    0
chr6  30132313    30132314    0
```

Custom tracks input file for the *TRIM15* 'B' amplicon (used in **Chapter 4**):

```
browser position chr6:30138398-30143332
track type=wiggle_0 name="power" description="Power to detect a singleton" visibility=full graphType=points color=0,0,0 priority=10
variableStep chrom=chr6 span=8
30138398    100
30138399    100
30138400    100
......skipped
30143329    100
30143330    100
30143331    100

track name="TRIM15_B1" description="High Quality SNPs (Known)" itemRgb=On visibility=pack priority=20
chr6  30138488    30138489    rs114344980 800   +     30138488    30138489    0,0,0
chr6  30138644    30138645    rs1029238   800   +     30138644    30138645    0,0,0
chr6  30138661    30138662    rs1029237   800   +     30138661    30138662    0,0,0
chr6  30138852    30138853    rs41272591  800   +     30138852    30138853    0,0,0
chr6  30138864    30138865    rs41272595  800   +     30138864    30138865    0,0,0
chr6  30138894    30138895    rs115440118 800   +     30138894    30138895    0,0,0
chr6  30139020    30139021    rs41272599  800   +     30139020    30139021    0,0,0
chr6  30139698    30139699    rs929156    800   +     30139698    30139699    255,0,0
chr6  30140341    30140342    rs1063280   800   +     30140341    30140342    0,0,255
chr6  30140524    30140525    rs6905949   800   +     30140524    30140525    0,0,0
chr6  30140539    30140540    rs13212414  800   +     30140539    30140540    0,0,0
chr6  30140912    30140913    rs2844787   800   +     30140912    30140913    0,0,0
chr6  30141041    30141042    rs9380156   800   +     30141041    30141042    0,0,0
chr6  30141203    30141204    rs13213365  800   +     30141203    30141204    0,0,0
chr6  30142252    30142253    rs757258    800   +     30142252    30142253    0,0,0
chr6  30142457    30142458    rs757257    800   +     30142457    30142458    0,0,0
chr6  30142673    30142674    rs961039    800   +     30142673    30142674    0,0,0
chr6  30142689    30142690    rs957765    800   +     30142689    30142690    0,0,0
chr6  30142998    30142999    rs2394737   800   +     30142998    30142999    0,0,0

track name="TRIM15_B2" description="High Quality SNPs (Novel)" itemRgb=On visibility=pack priority=30
chr6  30138475    30138476    novel 800   +     30138475    30138476    0,0,0
chr6  30138596    30138597    novel 800   +     30138596    30138597    0,0,0
chr6  30139154    30139155    novel 800   +     30139154    30139155    0,0,0
chr6  30139395    30139396    novel 800   +     30139395    30139396    0,0,0
chr6  30139476    30139477    novel 800   +     30139476    30139477    0,0,0
chr6  30142264    30142265    novel 800   +     30142264    30142265    0,0,0

track type=bedGraph name="-log10(p-value)" description="Association p-value" visibility=full altColor=0,0,255 priority=40 autoScale=off viewLimits=0:3
yLineOnOff=on yLineMark=1.301 yLineOnOff=on yLineMark=0
chr6  30138475    30138476    0.6064
chr6  30138488    30138489    0.3025
chr6  30138596    30138597    0.3025
......skipped
chr6  30142673    30142674    0.1679
chr6  30142689    30142690    0.3092
chr6  30142998    30142999    0.0425
```