From the Department of Medical Epidemiology and Biostatistics
Karolinska Institutet, Stockholm, Sweden

GENOMICS AND BIOINFORMATICS STRATEGIES IN THE STUDY OF AGING AND ALZHEIMER DISEASE.

Mun-Gwan Hong

Stockholm 2011

Published by Karolinska Institutet. Printed by Larserics Digital Print AB.

ISBN 978-91-7457-251-3

Your talents should not exceed your virtue

- Confucius -

ABSTRACT

To understand complex phenotypes, medical research has evolved from the study of single genes and proteins to approaches that encompass more comprehensive catalogues of molecules. Among the more widely used are genome-wide expression and high-throughput genotyping, the latter primarily making use of single nucleotide polymorphisms (SNPs) in what has been termed genome-wide association studies (GWAS). Because of the scale of the data sets that are being produced, unique problems have emerged that necessitate the extensive use of bioinformatics tools. This thesis has entailed the analysis of several such large data sets in the context of biological pathways and introduces several bioinformatics solutions. Paper III, IV, and V deal with this topic. This thesis is primarily oriented around the study of Alzheimer disease (AD) and aging. The questions about the etiology of AD are often concurrent with questions about the biology of aging. This thesis pursues insight on genomic factors pertaining to both inquiries, acknowledging that both the AD state and aging itself are complex and multi-factorial. Two constituent papers (I and III) address aging and two papers (II and V) deal with genetic models in the study of AD.

In paper I, we examined the association of age with several genetic markers in the insulin degrading enzyme (*IDE*) and explored possible molecular mechanisms. In contrast to women, both age-at-sampling and age-at-death of the males were significantly lower in individuals that were heterozygous at genetic loci spanning the *IDE* locus. Plasma insulin levels and the expression levels of the gene were found to be higher in those same heterozygous males.

In paper II, SNPs in 25 genes involved in cholesterol metabolism were tested for association with AD and dementia. Genetic markers in a large linkage disequilibrium block spanning *SREBF1*, *TOM1L2*, and *ATPAF2* were significantly associated with disease. Gene expression and gene network analyses supported the findings.

In paper III, we investigated the biological pathway basis of age in human brain and lymphocytes. Mitochondrial genes were negatively regulated in both tissue samples, while the protein translation genes appeared to decrease in lymphocytes but increase in brain. Those observations indicated that there are common themes across tissues, but also tissue specific changes in gene regulation. We also examined the genomic architecture of the age-regulated genes, and found that the expression of non-compact genes tend to decrease with advancing age.

A large number of genome-wide association studies (GWAS) have now been performed over the past few years. In paper IV, we developed a program that automates the conversion of SNPs to representative gene lists in order to facilitate the exploration of biological pathway in the context of GWAS.

In paper V, we employed the software developed in study IV to identify biological pathways enriched among the genes that were significantly associated from a GWAS of AD. Genes involved in intracellular protein transmembrane transport were found to be significantly overrepresented. These results highlighted the possibility that *TOMM40* contributes to AD pathology together with other translocases.

Through this thesis, several biological relationships have been identified linking AD and aging. Genetic markers in *IDE*, a gene previously claimed to be associated with AD, also associate with age. With advancing age, mitochondrial gene expression deteriorates significantly. *TOMM40* may contribute the AD pathology, together with other genes that encode proteins of the intracellular transmembrane protein transport pathway. Methodologically, pathway analyses were conducted successfully with the program, ProxyGeneLD. This enabled discoveries and discussion of the challenges that face the exploration of GWAS data sets in a pathway context. In the future, more sophisticated bioinformatics tools and enhanced gene annotation may lead to the discovery of the molecular mechanisms that dominate complex diseases and traits.

LIST OF PUBLICATIONS

This thesis is based on the following papers.

I. **Mun-Gwan Hong**, Chandra Reynolds, Margaret Gatz, Boo Johansson, Jennifer C. Palmer, Harvest F. Gu, Kaj Blennow, Patrick G. Kehoe, Ulf de Faire, Nancy L. Pedersen and Jonathan A. Prince

Evidence that the gene encoding insulin degrading enzyme influences human lifespan

Human Molecular Genetics (2008) 17:2370-2378

II. Chandra A. Reynolds, **Mun-Gwan Hong**, Ulrika K. Eriksson, Kaj Blennow, Fredrik Wiklund, Boo Johansson, Bo Malmberg, Stig Berg, Andrey Alexeyenko, Henrik Grönberg, Margaret Gatz, Nancy L. Pedersen and Jonathan A. Prince

Analysis of lipid pathway genes indicates association of sequence variation near SREBF1/TOM1L2/ATPAF2 with dementia risk.

Human Molecular Genetics (2010) 19:2068-2078

III. **Mun-Gwan Hong**, Amanda J. Myers, Patrik K. E. Magnusson and Jonathan A. Prince

Transcriptome-wide assessment of human brain and lymphocyte senescence

PLoS ONE (2008) 3:e3024

IV. **Mun-Gwan Hong**, Yudi Pawitan, Patrik K. E. Magnusson and Jonathan A. Prince

Strategies and issues in the detection of pathway enrichment in genome-wide association studies

Human Genetics (2009) 126:289-301

V. **Mun-Gwan Hong**, Andrey Alexeyenko, Jean-Charles Lambert, Philippe Amouyel and Jonathan A Prince

Genome-wide pathway analysis implicates intracellular transmembrane protein transport in Alzheimer disease

Journal of Human Genetics (2010) 55:707–709

CONTENTS

LIST OF ABBREVIATIONS

ACE	angiotensin-converting enzyme
AD	Alzheimer disease
ANOVA	analysis of variance
APOE	Apolipoprotein E
Aβ	amyloid-β
BACE	β-secretase APP-cleaving enzyme
cDNA	complementary deoxyribonucleic acid
CEU	Utah residences with Northern and Western European ancestry from the Centre d'Etude du Polymorphisme Humain collection
CLU	clusterin; a.k.a. apolipoprotein J
CR1	complement component (3b/4b) receptor 1
CSF	cerebrospinal fluid
DAVID	the database for annotation, visualization and integrated discovery
DDR	deoxyribonucleic acid damage response
DNA	deoxyribonucleic acid
ECE	endothelin-converting enzyme
EOFAD	Early-onset familial Alzheimer disease
GO	Gene Ontology
GSEA	gene set enrichment analysis
GWAS	genome-wide association study
HWE	Hardy-Weinberg equilibrium
IDE	Insulin degrading enzyme
IPA	Ingenuity pathway analysis
LD	linkage disequilibrium
LOAD	Late-onset Alzheimer disease
MAF	minor allele frequency
MI	myocardial infarction
mRNA	messenger ribonucleic acid
NCBI	National Center for Biotechnology Information
NIH	National Institutes of Health
OR	odds ratio
PCR	polymerase chain reaction
PD	Parkinsons disease
PICALM	phosphatidylinositol binding clathrin assembly protein
pmi	post mortem interval
qPCR	quantitative polymerase chain reaction
RNA	ribonucleic acid
ROS	reactive oxygen species
SNP	single nucleotide polymorphism
UTR	untranslated region

1 INTRODUCTION

Alzheimer disease (AD) is the most common cause of the dementia of the elderly, and is believed to be induced by large number of factors. Disorders such as Alzheimer disease are often termed “complex”, reflecting the challenge of identifying these underlying factors. Aging is the single factor that has the largest effect on the onset of the disease, and is itself controlled by complex biological mechanisms. Over the past decades, a considerable amount of effort has been exerted to understand the link between aging and AD.

In epidemiological studies, the various factors that contribute to a disease are generally divided into two groups, genetic and environmental. The genetic factors comprise all inherited components that can affect the predisposition of individuals to a disease of interest. The environmental factors, complementary and in some respects interacting with the genetic component, represent the events an individual is exposed to during their lifetime. Examples include both conscious choices like smoking, food preference, and degree of physical exercise, but also include chance accidents beyond the control of an individual. In this thesis, the various studies have focused primarily only on the genetic factors, more specifically single nucleotide polymorphisms (SNPs) that make up the majority of the genetic differences that can be found between individuals and between human populations.

These SNPs, also known as genetic variants, in human populations are abundant. They occur on average at about 1 site per 1000 DNA. Their relatively high frequency confers statistical power for detecting regions of interest in genetic association studies, which have been used in several of the papers presented in this thesis. Genetic association studies typically entail the examination of SNP (or other kinds of variation) frequency in relation to a disease or other phenotype like height or weight. As the technology for genotyping (i.e. reading the SNPs in an individual’s genome) has developed dramatically over the past few years, it has become a reality to examine essentially all common SNPs in an individual’s genome using arrays on a single chip. These genome-wide experiments (also called genome-wide association studies; GWAS) have produced enormous quantities of data. Various approaches to understanding this genetic data have emerged, and include, apart from the strict assessment of SNP allele frequencies for association, attempts at viewing the data in a broader biological context. One fundamental approach to investigating the data is to explore for enrichment of specific biological pathways among the genes associated with target phenotype. This thesis introduces bioinformatics strategies to deal with this research question and discusses the issues of such a strategy.

2 THE STRUCTURE OF THE HUMAN GENOME

2.1 SINGLE NUCLEOTIDE POLYMORPHISM (SNP)

The human genome is the hereditary information of our species stored in long stretches of deoxyribonucleic acid (DNA) molecules. The genome is believed to contain all information not only for the single cell of a fertilized egg to develop to all of the organs that make up the human body, but also for the body to survive in its crude environment. These potent molecules are bound into 23 pairs of bundles called chromosomes. Each chromosome has a sister chromosome which has ~99.9% identical sequences [7], and the pair are called homologous chromosomes.

Each individual in a population has a unique genomic sequence that differs from every other individual [7]. Some of these genetic variants contribute to our physical appearance and others influence the onset of the various diseases that afflict us. Variations in the genome are classified according to their size and characteristics, and include mutation of single bases to large stretches of DNA, to duplications, repeats and inversions. By far the most common however is the single nucleotide polymorphism (SNP) for which in excess of 10 million are now known to exist across world populations [8,9]. SNPs represent the consequence of mutation events that have been retained in a population, either by chance or by a selective advantage they confer, that has allowed them to attain high frequency.

The SNP is a single DNA base difference in the genome that consists of two "alleles" on the two sister chromosomes. Some individuals will be homozygous whichever the common allele at the SNP site, some individuals will be homozygous for the "rarer" allele, and some individuals will be heterozygous, having each of the two alleles. From the various large-scale genotyping projects conducted around the world in different populations, SNPs can be found at a rate of about once every ~1000 bases on average when two chromosomes are compared [7-10].

2.2 LINKAGE DISEQUILIBRIUM (LD)

The information in the human genome is transmitted to each next generation beginning with an elaborate molecular system that copies and delivers the DNA in the process of meiosis. During this process, the two copies of each chromosome undergo recombination, in which long segments of DNA cross over from one chromosome to the corresponding position on its sister chromosome. On an extended time-scale, the recombination event is rather infrequent, arising on average only once for every stretch of approximately 1 million base pairs of DNA per 100 generations [11,12]. Thus, a long DNA sequence including all of the mutations that have occurred in that specific genomic region (many of them possibly increasing disease risk) tends to be intact while

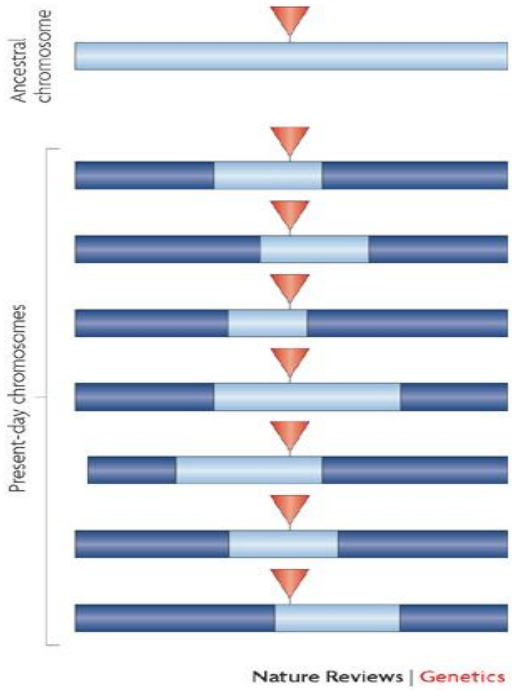

Figure 1. Preserved linkage disequilibrium surrounding ancestral mutation. A mutation occurred on the chromosome of common ancestor is indicated by red triangle. Light blue bars represent intact fragments transmitted from the ancestral chromosome, while dark blues are the parts from foreign individuals. *Reprinted by permission from Macmillan Publishers Ltd: Nature Review Genetics (L. Kruglyak[3]), copyright (2008)*

it descends through many generations (an example of a single mutation and its fate through multiple generations is depicted in (Figure 1). Figure 1 represents the cornerstone of all genetic association studies, which rely on LD to be able to observe latent effects of un-genotyped genetic sequence variants. Thus, through recombination, the "alleles" located in the vicinity of a spontaneous mutation event are physically paired with the allele generated by the mutation, and the correlation is detectable in the extant chromosomes that are present today [3]. This is one of the possible mechanisms that creates linkage disequilibrium (LD), which can more formally be defined as non-random association between two or more proximal genetic variants [13]. There are diverse other mechanisms that can generate LD, including subpopulation mixing [14] and inbreeding [15,16]. Another important aspect of LD is that it tends to decay over time. Thus, once established, LD begins to decay by recombination toward an equilibrium state, where each subsequent generation has a chance of proximal alleles becoming physically detached. This decay rate has been estimated to be about 10-8 per base pair per generation [17,18]. In other words, diminishing LD is a relatively slow process (proportional to recombination) and so young populations derived from small founder groups of people, will tend to have more LD than old populations that have been mating amongst themselves for thousands of years.

One of the commonly used statistics to describe LD between two SNPs is r^2, which is the correlation coefficient of allele frequencies of the two variants. When there are two bi-allelic (two alleles) SNPs with alleles "A", "a" at the first position and "B", "b" at the second position, the value can be calculated as depicted below [13,19]:

$$r^2 = \frac{(p_{AB} - p_A p_B)^2}{p_A(1 - p_A)p_B(1 - p_B)}$$

Where p_{AB} is the frequency of observation of both AB, and p_A and p_B are the frequencies of A and B allele respectively.

A large international effort with the goal of pursuing a LD map of the human genome was initiated in October 2002. The effort was termed the International HapMap project, and in its first phase successfully genotyped 1.1 million SNPs in 90 human individuals from 30 families who lived in Utah with Northern and Western European ancestry from the Centre d'Etude du Polymorphisme Humain collection (abbreviated to CEU) together with genotypes of Nigerian, Chinese, and Japanese samples during this first phase of the project [17]. During the second phase, an attempt was made to expand the number of tested SNPs, this time to include 3.9 million putative polymorphic variants for which genotyping assays could be developed [12]. Approximately one third of all these tested markers turned out to be non-polymorphic among the CEU samples. The remaining confirmed 2.6 million SNPs reflect a distribution across the human genome and they are estimated to cover 92% of hidden common variants (minor allele frequency $\geq 5\%$) at the threshold of $r^2 \geq 0.8$ by pairwise LD.

3 GENETIC ASSOCIATION STUDY

3.1 VOCABULARY IN GENETIC ASSOCIATION STUDIES

A locus (plural form is loci) is the position of a gene or a genetic variant on a specific chromosome. An allele is defined as each different version on a single chromosome of a genetic marker of which the variation exists in the population [20]. Since humans have two sets of chromosomes (homologous or "sister" chromosomes), each individual has two alleles at a single locus. The two collectively form the genotype of an individual for that locus. Between the two, the allele for which the frequency in the population of interest is larger is commonly known as the major allele and the other is minor allele. The frequency of the minor allele is often abbreviated to MAF (minor allele frequency) indicating how rare the allele or the genetic variant is. An individual having the same alleles at single locus is termed homozygous and an individual having different alleles is termed heterozygous.

When a genetic marker, usually SNP, is in high LD (often $r^2 \geq 0.8$) with another marker, each is called a "proxy" of the other. Especially, if the correlation coefficient is 1, they are termed as "perfect proxies"[21]. A Phenotype is defined as specific detectable characteristics of an organism [22]. Penetrance is defined as the probability that a particular allele or genotype induces a particular phenotype [20].

Under the assumptions of random mating with evenly distributed fertilities, no selective influx or efflux such as migration or natural selection, and no mutation in a large population, the allele and genotype frequencies at a locus will be unchanged through generations [23,24]. This equilibrium status is called Hardy-Weinberg equilibrium (HWE) [25], in which the genotype frequency of homozygotes is expected to be the square of the allele frequency, while the frequency of heterozygotes is simply the product of both allele frequencies. The test for HWE of a SNP is commonly performed with an asymptotic χ^2 test or exact texts [26,27].

3.2 GENETIC ASSOCIATION STUDY

Genetic association studies involve the investigation of the potential of a correlation between allelic or genotypic variation and phenotypic differences [28,29]. When the frequency of a certain variant is observed, for example, in cases at a significantly higher level than controls, it can be concluded that there is association between the genetic marker and the disease status, and that the specific allele at higher frequency may have a contributory role in the disorder. This significance is most appropriately assessed by statistical methods, some of which are described in section 7.

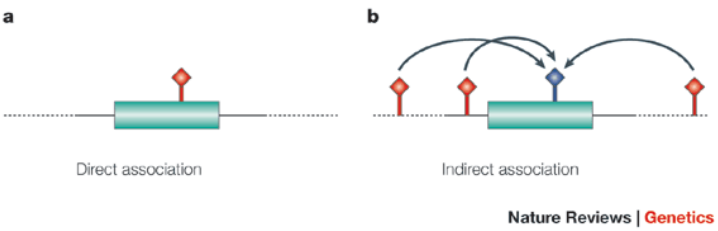

Figure 2. Testing genetic variant for association directly or indirectly. The reds indicate genotyped markers and the blue is the marker in high LD with the reds. **a**. A candidate genotype is directly tested for association. **b**. Utilizing LD structure, the proxies of the candidate are genotyped and the association is imputed.
Reprinted by permission from Macmillan Publishers Ltd: Nature Review Genetics (J. Hirschorn [6]), copyright (2005)

In contrast with other association studies in general, genetic studies carry the inherent difficulty of needing to take into account the extensive correlation that can exist between nearby markers due to LD structure. LD is a two-edged sword in that it aids in the search for association between genetic marker and phenotype, since genotyping proxies can be sufficient to observe association instead of specifically examining the precise functional SNP(s) (Figure 2) [6,30,31]. However, on the other hand, the association found between a genotype at one locus and phenotype does not denote that the tested marker is itself functional. Thus, the functional SNP or other form of variation that contributes to the difference in phenotype variance can be any genetic variant for which the frequency is highly correlated with the tested marker. This contributes to one of the greatest difficulties in attributing "marker associations" to "gene associations" since the region of study can extend across numerous gene targets, each of which can be a valid biological target in the disease under study. This fact should be taken into consideration in any genetic association study.

4 GENE EXPRESSION MEASUREMENT

4.1 BIOLOGICAL MATERIAL AND ITS PREPARATION

The molecules that carry out most of cell functions are proteins, which are encoded in the genome and are under sophisticated control. When a cell requires the creation of a particular protein, the production is initiated by a process called “transcription”, in which an enzyme complex reads the DNA code for a gene of the protein and produces messenger RNA (mRNA) which will eventually be converted to a functional protein. Since a single DNA sequence can be used as a template for thousands of protein molecules, the number of mRNAs is under careful regulatory control, where numerous factors act both by binding to genomic DNA and to the produced mRNA. In this thesis, “gene expression” refers to the mRNA expression as widely accepted to reflect the fundamental mechanism of moving from DNA sequence to the quantity of mRNA. Another key concept in the regulation of mRNA is splicing, which refers to the way in which the mRNA is processed after being transcribed from DNA. This involves the change in the exon structure of an mRNA, where complete exons may be skipped, new exons included, or changes induced between exons. This is regarded as a key way in which nature and evolution have led to a vast increase in molecular diversity, since the number of splice-forms of an mRNA can be on the order of hundreds.

mRNA as extracted from human tissue samples is chemically too unstable to be used directly in the experiments to measure its quantity. To avoid degradation, the material containing these fragile molecules should be frozen immediately after acquisition and stored in deep freeze (-80°C) [32]. If it is obtained from the tissue that may include living cells, this is even more important since cells may react to environment changes, possibly activating mRNA degradation machinery, or even promoting the rapid production of mRNA as a defense mechanism [33]. Thus as for samples obtained after death, for example, brain autopsy samples, the time that has elapsed since death (post mortem interval; pmi) to sample treatment should be minimized. Often this kind of variation among samples is interrogated in the analysis and documented and thus can be used as a covariate in statistics. However, if pmi is less than 48 hours, it affects RNA less than agonal status (the period with serious illness before death) [34,35]. To be quantified, RNA is usually reverse-transcribed (opposite process of transcription) to complementary DNA (cDNA). The most common external cause of mRNA degradation is RNAse protein molecules that exist in the environment both from humans and other organisms. RNAse proteins are abundant and a single drop of human sweat can contain millions that are all capable of rapidly degrading mRNA. This is one of the key reasons for converting mRNA rapidly to cDNA for further analysis.

4.2 QUANTITATIVE PCR

The Polymerase Chain Reaction (PCR) is arguably one of the most well known molecular biology tools, and was invented to facilitate the amplification of a particular DNA sequence by doubling the number of fragments of the sequence in each cycle. It has high fidelity and specificity in this amplification reaction, given known flanking genetic sequences around a target sequence of interest. A particular sequence can be selectively amplified in the mixture of a tremendous number of heterogeneous sequences, For example, this might involve all human cDNA sequences in a solution perhaps including twenty thousand different molecules, each with millions of copies. Because of these features, PCR frequently serves as a tool to obtain both strong evidence of the existence of a certain sequence in a DNA sample, as well as to amplify existing sequences to achieve large enough quantities for subsequent analyses.

Additionally, PCR can be also serve to measure the quantity of selected fragments in the DNA samples with high specificity by adding special fluorescent dye-attached probe and detecting emitted signal from it [36,37]. The method is called quantitative PCR (qPCR) or real-time PCR since it can track how much DNA was amplified in each cycle. The light intensity from the probe is proportional to the number of amplified fragments by PCR, and the amplification speed at same cycle is proportional to the initial quantity of the target DNA fragment. So, the quantity in the sample is postulated from a surrogate measure, Ct value. The Ct (threshold cycle) is the cycle number at the point where the signal intensity exceeds a selected threshold. The threshold is often chosen as the intensity value at the centre in the span of noise and highest level on a log scale. As an illustration, a signal change graph from a real experiment is shown in

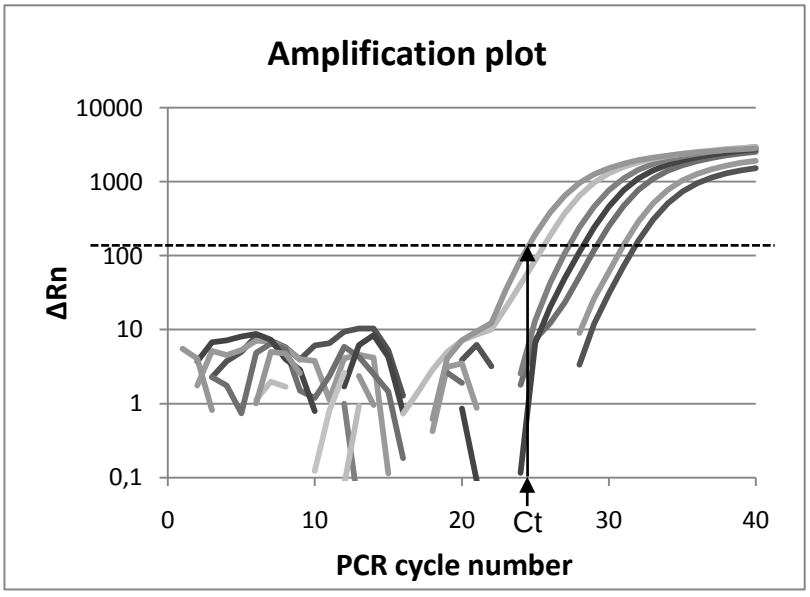

Figure 3. qPCR amplification plot. This shows the pattern of the signals from probes. Detected fluorescent intensities were adjusted by substraction to noise level intensities (ΔRn), which is shown in y axis in log scale. The dotted line indicates a selected threshold and the arrow points Ct value for the sample that had the largest number of target DNA fragments..

Figure 3.

Two kinds of quantification are usually performed with qPCR, absolute and relative quatifications. The intermediate steps in both methods are identical. The difference is that the absolute measure requires samples for which the quantity is known. For the relative quantification, the samples for which the concentration is high enough for the use of sequencial dilution is necessary. One of the common methods used to convert measured Ct to the quantity of cDNA molecules is by comparison with a standard curve, which is created by conducting identical experiments with known quantities in absolute or relative values [38]. Since the linear relationship between the log of the quantity of cDNA and Ct is expected, the observations are plotted in a standard curve. The linear equation of a trend line is then applied to estimate the quantities of unknown samples [39].

In the extraction step of mRNA from biological sample, the variation of the real quantity in various concentrations cannot be avoided. In order to adjust the difference, mRNA expression measurement of target genes are commonly accompanied with a reference gene, which is assumed to be expressed constantly in a cell. The candidates are usually housekeeping genes such as β-actin, 18S rRNA, *GAPDH*. Among them, the expression *GAPDH* showed less deviation across the combined samples of AD patients and controls [40].

4.3 MICROARRAY FOR GENOME-WIDE EXPRESSION PROFILING

A microarray-based technique to simultaneously measure the expression of the transcripts of multiple genes was developed about 15 years ago, and was created with the ultimate goal of being able to measure the complete transcriptome of biological samples [41-43]. The measurement is performed by capturing labeled cDNA sequences in the sample by hybridization to complementary oligonucleotide probes aligned in an array and attached to the solid surface of the microarray (or chip as it is sometimes called). Signal detection is then carried out usually with fluorescent labels with a scanning laser. The signal intensity represents the abundance of the corresponding mRNA in the sample. The technology involves numerous processes from the fabrication of the chips themselves to reading and interpreting the signal that induces errors to the final output, and high reproducibility is still a major goal of those involved in further developing the technology. Thus, appropriate normalization of the data is a key requirement [44].

5 CANDIDATE GENE OR PATHWAY APPROACH

5.1 CANDIDATE GENE OR PATHWAY ASSOCIATION STUDIES

Candidate gene association studies were very popular before the advent of genome-wide association studies, essentially representing the only means given technological restraints of genotyping biological samples. Candidates were typically (and still are) selected based on the previous findings of other biological studies that implicate a particular gene in a disease under study. The genetic variant(s) to be genotyped are usually chosen according to a prioritizing scheme of the individual study. Candidate gene studies often use the validated assay that were specially designed for the target marker and could be performed in the researchers own lab or in the laboratories of close collaborators. For the candidate pathway approach, which represents a middle ground between single gene association studies and the GWAS and may involve thousands of SNPs, users tend to turn to genotyping platforms for user selected SNPs that are commercially available.

5.2 INSULIN DEGRADING ENZYME (IDE)

Insulin degrading enzyme (IDE) is a zinc metalloprotease which has been shown to have the capability of degrading a variety of small proteins including insulin, insulin-like growth factor-2 (IGF-2) and amyloid β [45,46]. The gene has received considerable attention as a putative candidate in the etiology of Alzheimer disease due its strong biochemical activity in being able to degrade the hallmark peptide, amyloid β,

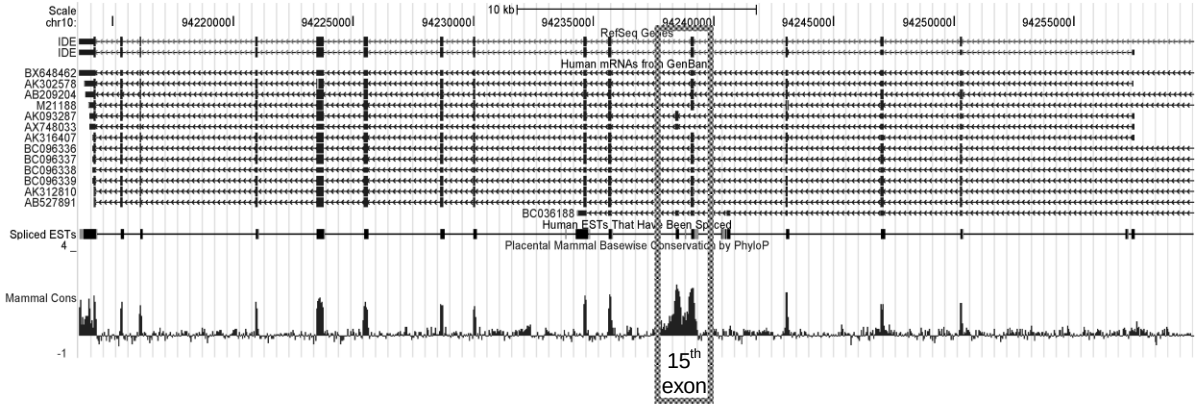

Figure 4. Identified mRNAs of the IDE gene and the conserved exons among placental mammals displayed by UCSC genome browser [1,2]. The second half of the longest transcript variant of IDE (NCBI Reference Sequence: NM_004969.3) including the 15th exon is only shown here. A scale bar, chromosomal position, and a couple of known transcripts in NCBI RNA reference sequence collection(RefSeq) are displayed above human mRNAs from GenBank of NIH. Accession numbers of mRNA are shown on the left. Exons were drawn in blocks, while lines and arrows represent introns and direction of transcription, respectively. The browser was accessed on 2010-09-29.

that is considered to play a causal role in Alzheimer disease. Additional evidence possibly connecting *IDE* to AD was obtained from genetic linkage studies showing the linkage to chromosome 10q [47], more precisely the 10q23-q25 region that includes *IDE*. Further studies also obtained evidence of linkage of $A\beta_{42}$ level in plasma to this same 10q region [48].

Several splice-form variants of the mRNA sequence of *IDE* have been identified thus far (Figure 4). In terms of gene location, this diversity can be observed in the form of different transcription initiation sites, variable lengths of the 3'UTR, and alternative versions of the 15th exon, for which numbering is derived by counting along the longest splice-form that is composed of 25 exons, NM_004969. The alternatively spliced form of transcript at the 15th exon which had been initially identified only in testis samples, was also observed in brain (cerebral cortex and cerebellum). The isoform translated from the transcript including the 15b exon was shown to have less activity in the degradation of $A\beta_{40}$ and $A\beta_{42}$, again supporting the possible role of IDE in the development of AD, and providing evidence that splice-form variation might be a contributing factor [49]. In another study, an alternative initiation site of translation was observed, which can putatively lead to the protein being trafficked to the mitochondria by using a specific targeting sequence located in the 5' end of the protein [50].

5.3 CHOLESTEROL METABOLISM PATHWAY

Cholesterol plays an important role in maintenance of plasticity and function of neurons [51]. Several studies have shown that high cholesterol level in serum or plasma is associated with susceptibility to Alzheimer's disease (AD) [52,53]. An increased in the degree of influx and efflux of cholesterols over the blood-brain barrier in AD patients has also observed [52,54]. One of the most important proteins that are involved in cholesterol transport in the brain is apolipoprotein E (APOE), of which genetic association with AD is well established [55]. APOE also strongly affects amyloid β (Aβ) deposition in the brain [56], suggesting a connection between Aβ, the main component of plaques, and apolipoprotein metabolism. In further support of this, another prominent phenotypic effect of APOE variation is upon cerebrospinal fluid (CSF) levels of the 42 amino acid fragment of amyloid β ($A\beta_{42}$) [57].

6 PATHWAY APPROACH USING GENOME-WIDE DATA

6.1 GENOME-WIDE ASSOCIATION STUDY

Until only a few years ago, most genotyping methods were designed to read a single or perhaps only a handful of SNP markers [58]. The cost of the experiment reagents, time required for preparation/labour, and the quantity of DNA required for genotyping multiple SNPs were all proportional to the number of markers that were to be tested. Limited resources and technologies constrained genetic association studies to be performed only with a finite number of markers around intriguing regions typically of single candidate genes, on the basis of previous molecular biological evidence and/or linkage study results [58,59]. Assisted by this prior knowledge, many genetic markers around the putative genes were claimed to be associated with diseases of interest by such approaches [6,60]. But, since this relied on testing a fairly limited hypothesis (usually one variant in a large gene and usually conducted one gene at a time), many of the studies that attempted replication simply failed to observe the earlier claimed effects. Even for the few replicated loci, the general theme was that the discovered variants simply could not explain substantial fraction of genetic risk of common diseases, such as diabetes and Alzheimer disease. The explanations the research community proposed were lack of statistical power and a general sense that perhaps traits, the development of which result from multiple genetic components interacting together with numerous environmental factors, were just too complex [6,59].

One proposed solution was to turn to high-throughput strategies to test the entire genome in the search for genetic association, a natural extension from the earlier successes from linkage studies, but taken to the level of common genetic variation. Overcoming this tremendous logistical obstacle, the first study was attempted to test associations between a complex disease and SNPs across whole human genome without targeting specific genes using a newly developed high throughput genotyping technique [61,62]. Thus, with the availability of the microarray technology for SNP genotyping, it has become commonplace to acquire the genotypes of hundreds of thousands of SNPs [63,64]. The method has since been perfected and at present requires relatively short experiment time and small quantities of DNA samples. To date more than 400 diseases and traits have been tested, the results of which can be readily surveyed in ‘A Catalog of Published Genome-Wide Association Studies’ available at: www.genome.gov/gwastudies [65].

All GWAS conducted to date are indebted to the development of bioinformatics resources related to the human genome project. One such derivative project however that more than any other has facilitated GWAS is the HapMap program [66]. The goal of the International HapMap consortium was to achieve dense enough genotyping data to produce a haplotype map across human genome [67]. The project found that the genotypes of ~300,000 markers were enough to capture ~800,000 common (MAF $\geq$

0.05) SNPs at the threshold $r^2 \geq 0.8$ across whole genome of European samples that were proved to be polymorphic in phase I of the project [17]. Based on that study and following studies, commercial microarray-based technologies have been designed to capture most of known common SNPs across human genome in a single chip [64,68].

6.2 PATHWAY-BASED APPROACH

Unlike classic genetic association studies with single target genes and presuppositions about molecular mechanisms, genome-wide association studies (GWAS) often entail hundreds of thousands of hypothesis tests (often > 500,000) for association between disease status or other phenotype of interest and numerous markers. A very small proportion (usually less than 50, ~0.01%) of all the tested markers receives attention since most of the others simply fail to pass through the filter that controls the multiple testing problems. In addition, in essentially all cases, identified loci also exhibit small effect sizes and together only explain a small fraction of the total trait variance [69]. Thus, the "pores" of the selection filter are often extremely small to ensure that refined test results contain only true positives that can then be replicated in additional samples. The standard paradigm is thus to conduct an initial GWAS in a small sample and then attempt replication of a handful of markers in much larger populations. From gene expression studies, specific gene-oriented questions could be phrased about a disease under study, but researchers quickly turned to questions about broad biological themes that could be seen in the transcriptome-based results. Thus, as an alternative approach, employing prior knowledge about genes with similar functions (pathway) was proposed to be applied to GWAS, since the data structure has much in common with the earlier expression-based studies [70].

6.3 PATHWAY DATABASES

6.3.1 Gene Ontology

The Gene Ontology (GO) project represents a cooperative effort pursuing formalized vocabularies of gene products stored in separate databases for various organisms [71]. The project was initiated to assist biologists in finding biological roles of unknown genes of which sequences were conserved in different organisms. Since it was a daunting obstacle that the terminologies were not equivalent in the databases of different model organisms, a more standardized dictionary was strongly desired [72]. As research at the genome scale continued to produced volumes of biological data, typically involving thousands of genes, the standardized descriptions of GO attracted great attention and concern [73]. At its core, the GO can be used to group multiple genes by similarities at different levels.

The GO structure is composed of two parts. One is the ontology terms themselves (GO ontology). The database includes the relationships between GO terms. The other is the links between the ontologies and gene products (GO annotation). GO ontology provides systematically structured species-independent terms (ontologies) in the three aspects of a gene product, which are “cellular component”, “molecular function”, and “biological process”. The “cellular component” indicates where a gene product acts in the cell or around the cell, the “molecular function” explains biochemical activity, and the “biological process” describes the function of the gene product for the cells, tissues, organs, and organisms [71]. The ontologies in each category are linked as nodes in a directed acyclic graph form. An example showing terms and relations from a term “meiosis” to the root “biological process” are depicted in Figure 5. The logical relation between nodes is often called a parent-child relationship, in which the node close to root is parent and the distant node is child.

There are three different kinds of relationships between ontologies, which are “is a”, “part of”, and “regulates” relations [71]. Among them, the “is a” and “part of” relations are transitive. For example, if A is a B and B is a C, then A is a C. In those relationship, the genes annotated to the child node is the subset of the genes of the parent node [73]. As an example in the Figure 5, the genes assigned to “meiosis”, in other words the genes which encode the proteins acting in meiosis, function in cell cycle phase.

The links in GO annotation database are generated by curators [71]. GO terms are assigned to genes based on not only the experimental but also the computational evidence such as sequence similarity with the genes of another organisms of which functions have been identified. In reality, quite a large proportion of human genes have been annotated using information other than raw experimental evidence. The GO continues to grow in contents [74]. More genes are annotated to more specific function every day and the number of annotations with experimental evidence is also increasing. The human data last updated 24 January 2011 contains 18 216 annotated genes.

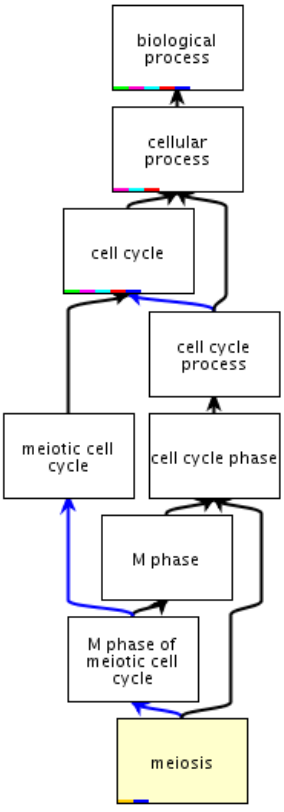

Figure 5. GO ancestor chart of the term "meiosis” generated by QuickGO [4]

6.3.2 KEGG pathway

The KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway database is the collection of interacting information of genes and molecules that has been obtained by manual survey of published literature [75]. It has been

designed to complement the information stored in other contemporary databases that focus on individual genes or molecules [76].

6.4 PATHWAY ANALYSIS TOOLS

All tools described here were originally designed for the analysis and interpretation of genome-wide expression profile data. Because those were designed around genes, it is appropriate and trivial to apply them to the study of genes in the context of GWAS. This is the list of tools that have been used in this thesis, but it should be acknowledged that numerous other software programs exist that perform similar functions.

6.4.1 Database for Annotation, Visualization and Integrated Discovery (DAVID)

DAVID is one of the most popular pathway analysis tools that was designed by academics and is freely available. It allows the testing for the overrepresentation of gene pathways, making use of gene annotation information stored in about 40 different databases. It allows for the use of a unique function termed "functional annotation clustering", which creates a summary of the output that usually includes an extensive number of redundant terms. Thus, it provides additional evidence that may implicate pathways that are not clear from the analysis of the individual terms. The significance of the enrichment is obtained by a slightly modified Fisher's exact test [77].

6.4.2 Gene Set Enrichment Analysis (GSEA)

GSEA, which has been efficiently implemented in JAVA, scrutinizes whether a particular pathway is randomly scattered across an input gene list. The gene list is sorted by the corresponding significance value for each gene, which typically is represented by the negative log of P value for the gene, and this was the primary statistic used in this thesis. For each gene in the gene list, it calculates a walking enrichment score (ES) that increases when the gene is annotated to the pathway and decrease when it is not. The increment is the assigned value for the gene. The final ES for the list is the maximum value of walking ES. The statistical significance of ES is estimated by permutation. Since multiple pathways are tested, the significance shown in output is FDR corrected for multiple testing [78].

6.4.3 GeneCodis

GeneCodis is a freely available web application that searches for enriched pathways among a subset of genes from a full contingent of genes. It provides a unique function to enable to identification of overrepresented combinations of pathways in the separate databases, for example, GO biological process and GO cellular component [79]. This is comparable to the function provided in DAVID, and also has the goal of providing possibly new biological insights that might not be evident from the single term analyses.

7 STATISTICAL METHODS

7.1 ASSOCIATION TEST

7.1.1 Linear regression

Often two variables can be expected to be correlated in a simple linear equation. One of the obvious examples is the pair of variable consisting of a) travelled distance and b) speed in an hour. The relationship can be expressed by a simple linear regression model as described:

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon$$

where x_1 is an explanatory variable (independent variable), Y a response variable (dependent variable), β_1 the effect (slope parameter), β_0 intercept parameter, and ε random error [80]. Under this model, the random error is assumed to follow normal distribution. Whether there exists a true linear relationship or not is tested by the test of the hypothesis the effect β_1 is equal to 0, in which the test statistic that follows t-distribution is often used.

In genetics, allelic additive effects can be examined in the linear regression model. As an independent variable, the number of minor alleles which varies by 0, 1, and 2 is commonly used (this also reflects the genotypes of an individual for 0 would represent homozygosity for the common allele, 1 heterozygosity, and 2 homozygosity for the rare allele).

When some traits in the experiment (e.g. temperature, experiment date) are suspected to affect the response in a linear manner, variables for the traits can be added to examine how strongly they influence the general model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

where x_2 and x_3 are the additional covariates.

7.1.2 Analysis of variance (ANOVA)

The test for the question ‘if two or more different groups are same comparing the observed values’ is often performed by analysis of variance (ANOVA), by converting the question to ‘if the values observed in two or more groups are randomly sampled from a single normal distribution’ or ‘the variance between the groups is significantly larger than the variance within each group’. Practically, it tests the null hypothesis that the means of each group are identical. Thus, if the number of groups is larger than 2, rejecting the null means that there was significant inequality of means between any pair of groups.

There are a few assumptions in ANOVA to be considered in its general application. One is the errors of observations should follow a normal distribution and be independent from the errors of the other observations. One example that is not under the latter assumption is the error increase along with the prolonged usage of a single machine. The other assumption is that the variances in each groups should not differ by a large margin.

7.1.3 Logistic regression

Logistic regression is often applied to the questions that involve a binary response in contrast with simple linear regression. The model is expressed:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 x_1$$

where $P(Y=1)$ is the probability of one state of response and the other variables are same in 7.1.1. The association is tested with the null hypothesis, $\beta_1 = 0$. The P-value of the test is obtained by a Wald test or Likelihood ratio test. The logistic model is one of the generalized linear models. It can be flexibly adapted to include more traits of interest by adding the terms for the traits like linear regression.

The estimate for β_1 in this model especially with binary explanatory variable represents the log of odds ratio (OR), which is often interpreted as relative risk with the implicit assumption that the outcome is rare.

7.1.4 Chi-square test (Contingency table)

When two variables of interest are categorical, contingency tables can be created. Within the table, the association between the variables is tested assuming the independence between the two. Comparing expected values under the assumption and observed values in the table, a Pearson's test statistics is calculated:

$$\mathrm{X}^2 = \sum_{\text{all cells}} \frac{(O-E)^2}{E}$$

where *O* is observed frequency and *E* is expected frequency in the table. The X^2 follows χ^2-distribution [81]. This test is not appropriate when the number of samples is too small. In such cases, Fisher's exact test should be applied, which is explained in the section 7.4.1.

7.1.5 Multinomial logistic regression

With a specific application to the study of human longevity, unlike many other methods used to test genetic association, of which explanatory variables are genotype or allele,

this model inverted the question (variables) by examining genotype frequency change as a function of advancing age. The model is expressed:

$$\ln\left(\frac{\pi_{i,j}(x)}{\pi_{w,w}(x)}\right) = \begin{cases}\alpha_{i,j} + \beta_{i,j}x & i = j \\ \ln 2 + \alpha_{i,j} + \beta_{i,j}x & i < j\end{cases}$$

$$\alpha_{w,w} = 0, \quad \beta_{w,w} = 0 \qquad i, j = 1,2, \ldots \ldots w$$

where *x* is age, *w* indicates the number of alleles (2 for bi-allelic SNP) and $\pi_{i,j}(x)$ is the frequency of genotype *i*, *j* at age *x* [82].

7.2 NORMALITY TEST

7.2.1 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov addressed the question of if the underlying distributions of observed values in two groups are identical without any parameter in model when the values are continuous. Selecting normal distribution as a reference, the test can be used to check how well the observations fit to the normal. Since the formula for the Kolmogorov-Smirnov statistic takes the maximum distance between two distribution functions, the test results can be influenced by outliers [83,84].

7.2.2 Shapiro-Wilk W test

The Shapiro-Wilk test checks whether the distribution of a variable is normal [85]. The W test statistic has typically been shown to have better performance than the Kolmogorov-Smirnov statistic [86,87].

7.3 MULTIPLE TESTING PROBLEM

7.3.1 Bonferroni correction

This method is based on the Bonferroni inequality (or Booles inequality), which states that the probability that at least one event is true is equal to or smaller than the summation of all probabilities that each event is true [88]. Bounded by the inequality, it is enough to declare the association significant when there were multiple hypothesis tests and the estimated P-value was not greater than the significance threshold (usually named α level) divided by the number of tests.

If the multiple tests were independent each other, the Bonferroni method is appropriate. However, in most cases (especially genetic association tests), they are more or less correlated since LD creates extensive correlation between variables. Thus, the correction is too stringent, which may result in many false negatives in the studies. There are thus special requirements to resolve this problem in genome-wide association studies that typically entail in excess of 500,000 tests.

7.3.2 False discovery rate

The false discovery rate is the estimated proportion of the truly negatives among the findings declared positive at a particular threshold. It is a useful method to estimate how many false associations are included among the declared positives acquired by large scale multiple tests such as those included in GWAS [89,90].

7.4 ENRICHMENT TEST

7.4.1 Hypergeometric test (Fisher's exact test)

The hypergeometric probability calculates the probability of observation of a particular combination of subsets from the full set of objects classified into two or more different classes. For example, it calculates the probability to observe two black and three white balls when 5 balls have been picked without replacement from a covered container that contains 20 black and 10 white balls. Since the question of enrichment is same as that of the hypergeometric, it is often used to check overrepresentation by calculating a P-value [91]:

$$\text{P-value} = 1 - \sum_{i=0}^{x} \frac{\binom{M}{i}\binom{N-M}{K-i}}{\binom{N}{K}}$$

where $\binom{b}{a}$ is the number of possible combination of b different objects when the number a has been picked. In terms of genetics application, there are x genes assigned to a pathway within the selected group of K genes. In total there are N genes. The pathway contains M genes.

7.4.2 Mann-Whitney U rank test

The Mann-Whitney U test (or Wilcoxon rank-sum test) is the non-parametric test to explore for evidence that the values in two groups are drawn from the populations with different distributions [92-94]. The underlying assumption is that the values are independent and continuous. It derives the test result by comparing the ranks of the components in two testing groups.

8 ALZHEIMER DISEASE AND AGING

8.1 ALZHEIMER DISEASE (AD)

Alzheimer disease (AD) is the most common cause of dementia in the elderly worldwide. Both dementia and AD are rather common and their prevalences increase exponentially with advancing age [95-98]. In contrast to broadly defined dementias, the slope for AD prevalence is somewhat steeper [97,98]. [97,98]. The prevalence of dementia in the age group of 65-69 is about 1%. It exceeds 5% in the group of 75-79 and continues to increase [99]. A recent study showed that about 40% of individuals of the age of 90 or more suffer from dementia and of these, approximately 75% of had a more strict AD diagnosis [98]. As the expected lifespan gets higher around the world, the number of AD patients is expected to surge [99]. AD sufferers require considerable care due to the chronic nature of the disorder, which will be even a more serious burden to the individuals themselves and families in the coming decades [100-102].

Clinically, AD is characterized by a progressive cognitive and functional impairment. Neuropathologically, it is discriminated by two lesions, "plaques" and "tangles". The first is the aggregated amyloid-β (Aβ) peptide observed in the extracellular matrix of brain tissue. The second is the neurofibrillary depostion of hyperphosphorylated τ (tau) protein in the intracellular matrix [103]. Genetically, it is classified into two forms, early-onset familial AD (EOFAD) and late-onset AD (LOAD) by the patient's age [104]. The latter is much more common (95% of all cases) and does not show as obvious familial segregation as EOFAD does.

8.1.1 Amyloid-β in AD

The amyloid-β peptide is generated by the cleavage of amyloid-β precursor protein (APP) by β-secretase APP-cleaving enzyme (BACE) followed by γ-secretase cleavage, competing with an alternative non-amyloidogenic pathway that begins with cutting APP by α-secretase activity [105,106]. Consistent with the Aβ genesis pathway, mutations in the genes of *APP*, presenilin 1 (*PSEN1*) and presenilin 2 (*PSEN2*) induce EOFAD with full penetrance [107-109], where the presenilins are highly homologous genes encoding one of four components of γ-secretase complex [110]. In the brain of healthy individual, the Aβ of which neurotoxicity is observed [111,112] is degraded by the insulin-degrading enzyme (IDE), neprilysin, endothelin-converting enzyme (ECE-1), and angiotensin-converting enzyme (ACE) [113-115]. It is also eliminated by the efflux mediated by low-density lipoprotein receptor-related protein from the brain [110,116]. Under the Aβ hypothesis, an imbalance that disturbs the equilibrium status between production and clearance is thought to be a core mechanism that leads to the development of AD [105,110,117].

8.1.2 Genetic studies on late-onset AD

The patients that suffer from both LOAD as well as EOFAD forms of AD are usually clustered in families, even if EOFAD carries "familial" in its definition since it is transmitted in a strict Mendelian manner. EOFAD is rare, with only about 1% or less of the AD population having an onset before the age of 65 (the typically threshold used to define early and late onset forms). About 60-80% of LOAD been shown to be determined by genetic factors, showing that there is an extensive genetic component, much of which is still unknown [95]. Towards the discovery of genetic variants that account for the late-onset form of AD, genetic association studies have over the past two decades primarily focused on putative candidate genes, defined subjectively by their possible involvement in biological processes thought to be of relevance to dementia. For instance, many of the investigated candidates over the years have been

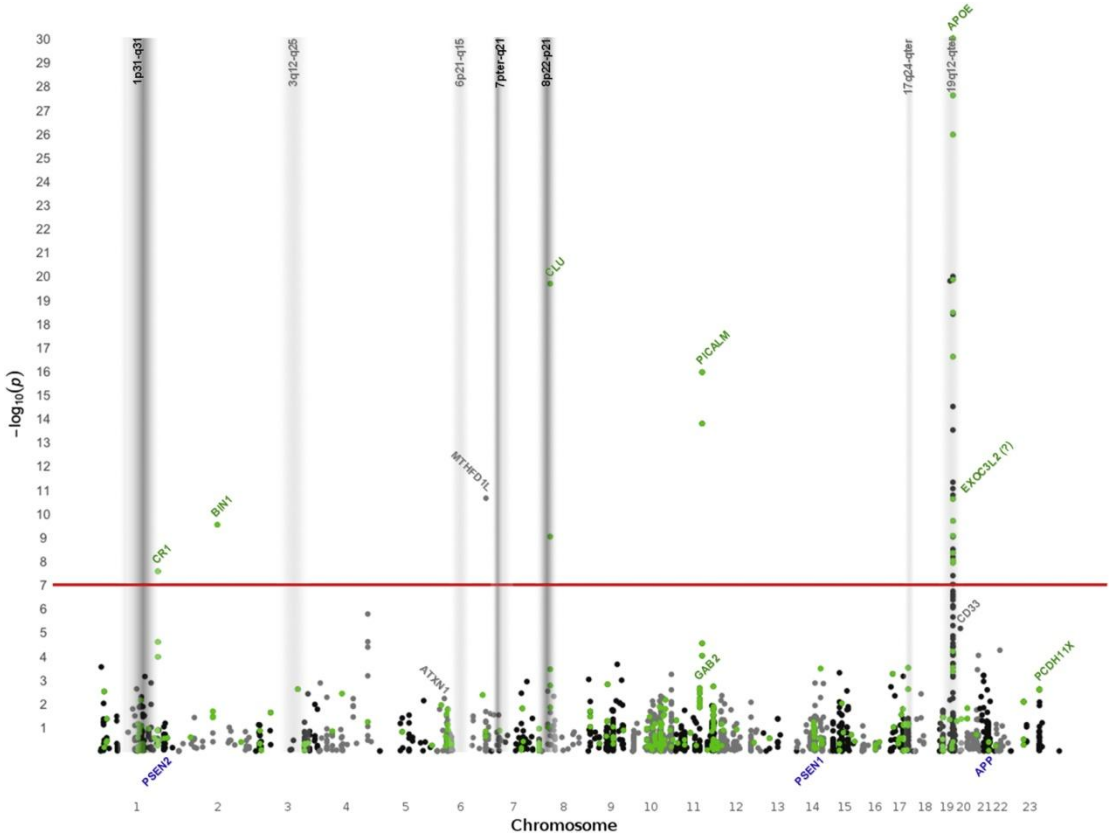

Figure 6. Manhattan plot of the recent genetic association data on the AlzGene database (http://www.alzgene.org. accessed on 2010-09-27)[5]. A total of 2033 genetic variants are shown here. The results from meta-analyses of 4 or more independent data sets are shown in green dots. The results from single-studies or meta-analyses of less than 4 separate studies are shown in black or gray dots. Note that the dot for APOE should be much higher in the plot, since P-value < 1×10^{-50}. The dark columns indicate the locations that showed "genome-wide suggestive" evidence of linkage in a meta-analysis of linkage studies on LOAD. The light columns are for "genome-wide nominal" evidence. Genes in blue at the bottom are those associated with EOFAD.

Reprinted from Neuron, 68, L.Bertram, et al., The Genetics of Alzheimer Disease: Back to the Future, 271, Copyright (2010), with permission from Elsevier

the genes that encode proteins that involved in the metabolism of Aβ. Despite these efforts, before the recent era of GWAS, only a single non-synonymous (amino-acid changing) SNP in Apolipoprotein E (*APOE*) had been confirmed that was associated with LOAD, replicated extensively by multiple research groups and in multiple human populations since it was discovered in 1993 [104,118]. For the large number of other candidate gene studies that claimed association of their target genes with AD, essentially all have failed to be replicated in following studies in independent populations.

Following the technical developments necessary to facilitate genome-scale association studies, a few GWAS on AD, more precisely LOAD, with large samples from various world populations have been conducted, and the summary of these is presented in Table 1. Up until the writing of this these, in total 14 studies have been published, of which three analyzed relatively large samples derived from more than 2000 cases and comparable controls in their primary analysis. By such extensive efforts, the associations of the genes *PICALM*, *CLU*, *CR1*, and *BIN1*, together with the well-known *APOE*, have been reliably replicated in multiple studies. As an overview, the significances of all the markers in the meta-analyses for AD available from the AlzGene database [5] are illustrated in Figure 6 [104]. Supporting some of the findings, the results from genome-wide linkage studies are included in the illustration with gray columns. Despite the successes, it should be noted that the estimated effects of the replicated loci excluding *APOE* were still low (allelic ORs ~1.15 referring to OR ~4 for *APOE* ε4) [104].

Table 1. Description of overall GWAS in AD on AlzGene database

GWAS	Design	Population	No. SNPs	No. AD GWAS (Follow-up)[b]	No. CTRL GWAS (Follow-up)[b]	"Featured" Genes[a]
Grupe et al., 2007[119]	Case-control	USA & UK	17,343	380 (1428)	396 (1666)	***APOE***, *ACAN, BCR, CTSS, EBF3, FAM63A*[**], *GALP, GWA_14q32.13, GWA_7p15.2, LMNA, LOC651924, MYH13, PCK1, PGBD1, TNK1, TRAK2, UBD*
Coon et al., 2007[120] &	Case-control	USA, Netherlands#	502,627	446 (415)	290 (260)	***APOE***, ***GAB2***
Li et al., 2008[122]	Case-control	Canada & UK	469,438	753 (418)	736 (249)	***APOE***, *GOLM1, GWA_15q21.2, GWA_9p24.3*
Poduslo et al., 2009[123]	Family-based &Case-control	USA	489,218	9 (199)	10 (225)	*TRPC4AP*
Abraham et al., 2008[124]	Case-control	UK‡	561,494	1082 (-)	1239 (1400)	***APOE***, *LRAT*
Bertram et al., 2008[125]	Family-based	USA	484,522	941 (1767)	404 (838)	***APOE***, ***ATXN1***, ***CD33***, ***GWA_14q31***
Beecham et al., 2009[126]	Case-control	USAˆ	532,000	492 (238)	496 (220)	***APOE***, *FAM113B*
Carrasquillo et al., 2009[127]	Case-control	USA●	313,504	844 (1547)	1255 (1209)	***APOE***, ***PCDH11X***
Lambert et al., 2009[128]	Case-control	Europe‡	~540,000	2035 (3978)	5328 (3297)	***APOE***, ***CLU*** (***APOJ***), ***CR1***
Harold et al., 2009[129]	Case-control	USA & Europe●‡	~610,000	3941 (2023)	7848 (2340)	***APOE***, ***CLU*** (***APOJ***), ***PICALM***
Heinzen et al., 2009[130] (CNV)	Case-control	USAˆ	n.g.	331 (-)	368 (-)	***APOE***, *CHRNA7*
Potkin et al., 2009[131]	Case-control	USA (ADNI)†	516,645	172 (-)	209 (-)	***APOE***, *ARSB, CAND1, EFNA5, MAGI2, PRUNE2*

GWAS	Design	Population	No. SNPs	No. AD GWAS (Follow-up)[b]	No. CTRL GWAS (Follow-up)[b]	"Featured" Genes[a]
Seshadri et al., 2010[132]	Case-control	Europe & USA●‡#	~2,540,000	3006 (6505)	22604 (13532)	***APOE***, ***BIN1***, ***CLU (APOJ)***, ***EXOC3L2***, ***PICALM***
Naj et al., 2010[133]	Case-control	USA & Europe†#ˆ	483,399	931 (1338)	1104 (2003)	***APOE***, ***MTHFD1L***

The data was achieved from AlzGene database [5] on 2010-09-27. [a] the genes affirmed to be associated with AD by the authors for the original study. The genes in bold font are those that showed study-wide "genome-wide significant" association. [b] follow-up study data set. The symbols (●,‡,#,†,ˆ) indicate there are common samples in different studies. Note that the well-known genetic variants in *APOE* were in many studies acquired by genotyping the proxy SNPs.

Reprinted from Neuron, 68, L.Bertram, et al., The Genetics of Alzheimer Disease: Back to the Future, 271, Copyright (2010), with permission from Elsevier

8.2 AGING

Age is the single most important factor in AD development. Upon neuropathological examination, normal individuals at advanced age may exhibit evidence of neurodegeneration, but at a level considerably more mild than AD patients undergo [134]. As the brain ages, various type of cognitive decline are typically observable, such as a reduced memory capacity, slowed response to external stimuli [135], and diminished creativity [136]. However, the knowledge and skills obtained at a younger age (typically referred to as wisdom) tend to remain intact for a longer time [137]. Microscopically, the number of neocortical neurons decreases by approximately 10% through 20 to 90 years while the number of glial cells tends to be invariant [138]. Neuronal loss in hippocampus is usually not significant in normal elderly individuals [139] whereas severe shrinkage of the hippocampus has been observed in AD cases [140].

The theories that pursue explanations of aging are numerous [141]. The two central branches of these are the "programmed" and "error" theories. The programmed aging theories have originated from the serial observations of the limited lifespan of explanted human cells [142], the shortening of telomeres that occurs with each successive cell cycle [143,144], and DNA damage response (DDR), of which accumulation leads cells to apoptosis or senescence [145]. Thus, most cells are destined to death due to the existence of a molecular clock at the tip of each chromosome. The programmed theory states that aging is controlled by a pre-programmed biological clock. The theory is supported by the observations in model organisms, which have showed that some mutants have increased longevity [146]. Most of the genes that have been mutated in such long-lived organism have been related to the pathways that regulate basic cellular functions such as growth, energy metabolism, and reproduction [147]. The evolutionary existence of programmed aging is explained by two theories of "antagonistic pleiotropy" and "disposable soma theory". The latter posits that lifespan is determined by the balance between growth/reproduction and somatic maintenance [147]. The former describes that the genes that are deleterious in old age remained in the genome because those same versions of the genes are beneficial in young age [141]. The "error" theory is that a decrease in vitality with advancing age is due to the accumulation of environmental attacks such as reactive oxygen species (ROS) in mitochondria inducing biological damages in cell or organism [148].

9 PRESENT INVESTIGATIONS

9.1 AIMS

- To investigate the association of genetic markers in *IDE* with age and with molecular levels in intermediate biological processes
- To examine the association of genetic variants of the genes in the cholesterol metabolism pathway
- To identify the biological pathways that are overrepresented among genes associated with age
- To develop a bioinformatics tool for pathway enrichment analysis and to address emergent issues in the analysis
- To identify pathways that are enriched among genes associated with Alzheimer disease

9.2 PAPER I

9.2.1 Materials and Methods

Human samples are briefly described in Table 2. Detailed descriptions of some of the samples are available in Ulrika K. Eriksson's thesis [149] and the original publications contained therein.

Table 2. Brief descriptions of human samples

Sample	*Sample Size (M/F)*	*Origin*	*Description*
Sample 1	601 (290/311)	Swedish	AD-free controls (Harmony)
Sample 2	321 (116/205)	Australian	PD-free controls
Sample 3	724 (411/313)	Swedish	Non-diabetic controls
Sample 4	178 (77/101)	Swedish	AD-free controls
Sample 5	539 (181/358)	Swedish	Random Population (OCTO)
Sample 6	590 (249/341)	Swedish	Random Population (SATSA)
Sample 7	2703 (1862/841)	Swedish	MI case-control
Sample 8	40 (14/26)	Swedish	Sequencing
Sample 9	178 (83/95)	English	DNA/RNA Brain

M = male, F = female.
Sample size represents the total sample for which genotyping was performed.

Genotyping of SNPs was conducted with dynamic allele specific hybridization (DASH) [150]. The novel spliceform of IDE was identified by PCR with a primer pair targeting both 15a and 15b-exons. The quantification of three different spliceforms was performed by qPCR using fluorescent labelled probes and exon-boundary spanning primers. The reactions were run in duplicate, and Ct values were averaged. Sequencing for identifying novel polymorphism was performed using a standard capillary electrophoresis method.

Multinomial logistic models were used to investigate the dependency of genotype on age [82]. The age difference between genotype group was assessed by ANOVA. Cox models were applied for the survival analysis of samples 5 and 6. Expression level analysis was performed using ANOVA with Ct values. HWE was tested with a χ^2 statistic. Normality was assessed using a Kolgomorov-Smirnov test.

9.2.2 Results

Among tested markers, the SNPs, rs1887922 and rs2251101 were uniformly associated with the observed traits in only men across the different populations. Intriguingly, heterozygotes comparing to homozygotes regardless allele itself had significantly lower age-at-sampling, shorter life span, higher insulin levels in plasma, and higher mRNA expression in brain. The association of both markers with age-at-sampling of men from 4 different samples was significant ($P=2.2\times10^{-7}$ and 5.1×10^{-5}), contrasting no evidence of association in women. The survival analyses with two sample groups showed significantly different mortality across the genotype classes of rs1887922 and rs2251101 with both the COX model and ANOVA, respectively. The genotype frequency of rs1887922 between young and old sample groups was significantly different (P=0.0052), but not for rs2251101 (P=0.17). Increased mortality among heterozygotes was observed in another sample for both markers (P=0.0064, 0.018). Genotypes of rs2251101 in men was strongly associated with insulin levels in plasma that were obtained at a 10 year interval. This marker was significantly associated with both the 15A and 15B spliceform expression levels (P=0.014, 0.0032)

9.3 PAPER II

9.3.1 Materials and Methods

The samples used in this paper were 1567 Swedish dementia cases and 2003 controls from four aging twin studies (SATSA, OCTO-Twin, GENDER, HARMONY) and a non-twin Alzheimer disease case-control study. The descriptions of the samples in more detail can be obtained in the previous publications for individual studies [95,151-154] and in Ulrika K. Eriksson's thesis [149]. For the secondary analyses, the genotype

data of Swedish males from the CAPS(Cancer Prostate in Sweden) and the expression data from two previous studies were obtained [155,156].

The 25 genes in cholesterol metabolism pathway were selected by literature search. Prioritizing the markers with previous finding, functional candidate, and LD, in total 506 SNPs in those genes were chosen and genotyped to investigate the association with dementia. CSF samples were obtained in Swedish AD case-control study [157].

HWE was tested with Pearson χ^2 statistic. Initial association tests between individual markers and dementia were conducted using logistic regression. Alternating logistic regression was applied to account for pair dependency of twins [158]. For the network analysis, FunCoup was employed [159]

9.3.2 Results

Among the observed marker, rs2230805 in *ABCA1* showed the most significant association following the *APOE* genetic variant. The third markers was located close to *SREBF1* ($P=8.5\times10^{-6}$) , which is in a large LD block spanning 7 genes. Applying F-SNP database tool, we found two potentially functional markers in high LD with the associated marker near *SREBF1*. Further investigating the association of gene expression with proxy SNPs of the *SREBF1* marker produced one strong correlation between genotype of the proxy and *ATPAF2* expression trait. In the network analysis, only *TOM1L2* could build a network with well known AD genes among the seven genes.

9.4 PAPER III

9.4.1 Materials and Methods

The first sample set consisted of 191 individuals who were free from neurological disease and whose brain tissues were collected at autopsy after deaths at ages 65-100. The second samples consisted of 1240 individuals at the ages of 15 to 94 years. The detailed descriptions of the human samples and the expression measurement protocols in detail are provided in the original paper [156,160].

After outlier removal by the Šidák's method, the associations between age-at-death and log transformed expression levels of brain samples were examined for brain samples by a linear regression model with global expression as a covariate. For the lymphocyte samples, similar linear regression model without a global expression term was applied.

For the pathway analyses, DAVID was applied to test the overrepresentation of Gene Ontology terms or KEGG pathway descriptors among the associated genes with advancing age [72,76,77].

9.4.2 Results

For the brain sample, the linear regression of 14 078 individual transcripts produced 54 significantly associated genes with age-at-death adjusted by the Bonferroni method. A validation of the association was confirmed by comparing the 54 genes with the genes identified in a previous study [161]. By pathway analysis of the associated genes (unadjusted P<0.05) with age, 'mitochondrion', 'synaptic transmission', and several more terms in GO were observed to be enriched among negatively correlated genes, and 'DNA binding', 'regulation of transcription, DNA-dependent', and more term in GO were identified among positively correlated genes.

Applying same strategy to additional samples of lymphocytes, the linear regression on each of 19 648 individual transcripts versus age produced a gene list of 1080 (612 negative, 468 positive) significantly associated transcripts. Dividing this list into positive and negative groups of genes as was done in the brain study, pathway analysis was performed with the top genes (unadjusted P<0.01). The terms'mitochondrion', 'nucleic acid binding', and several others were enriched among the negatively expressed genes and 'plasma membrane', 'signal transduction' and several additional weaker terms were overrepresented among the positively regulated genes. In the gene architecture comparison study, the negatively associated genes had significantly smaller intron to exon length ratio.

9.5 PAPER IV

9.5.1 ProxyGeneLD.pl

The program was designed to automate the processes of assigning each tested SNP marker to genes and to subsequently compute and assign gene-wide P-values. Considering the correlation structure among the SNPs (Linkage disequilibrium; LD), markers have the potential to be assigned to genes which are located far from them in terms of genomic distance. This is an important characteristic of the program in that it reflects the fact that a single marker can be assigned to multiple genes, each of which could contribute to a disease. The software was specifically designed to also take into consideration the inflation of significance due to multiple single marker tests in the calculation of P-value for each gene.

The process begins with investigating the LD structures using analogous populations with denser genotyping data. In this thesis, all analyses employing this program adopted HapMap CEU data. The software then continues to read the genetic marker lists derived from a genome-wide association study, with the core input being the P-values (usually generated from a simple allele test between cases and controls). All the tested markers are assigned to the genes based on genomic position. The markers within a region of a 1kbp upstream from 5'end in the direction of transcription but not

including any sequence downstream of 3'end of the longest known splice variant of a gene are assigned to the gene in this thesis. This parameter is readily adjustable by a series of command line options. On the grounds that the SNPs that are not genotyped but are strongly correlated with typed SNP (proxies) can be inferred from the observed markers [21], the typed markers are assigned to the genes of their proxies. In order to adjust P-value to account for the length of the gene and marker density spanning each gene region, the program counts the effective number of markers which can be regarded as independent genetic variations assigned to same gene. The numbers are then multiplied by the lowest P-values of the markers assigned to the genes. An example to illustrate the processes is shown in Figure 7.

The Perl programming environment was chosen as the language to build this program. Perl has excellent functions dealing with text-related manipulation compared with other popular programming languages such as JAVA, C++. Importantly, a single run of ProxyGeneLD per GWAS was generally regarded as sufficient, since the only parameters that change the output are the boundaries that define gene regions and LD cut-off thresholds, both of which were standardized. So, running time, the greatest limitation of Perl is not a serious obstacle. A standard run of the software on a GWAS with perhaps 500,000 SNPs might take on the order of 15 minutes on a standard PC.

In order to find the best gene identifier for the following analysis after running this program, several gene identifiers in different databases were tested. We compared how

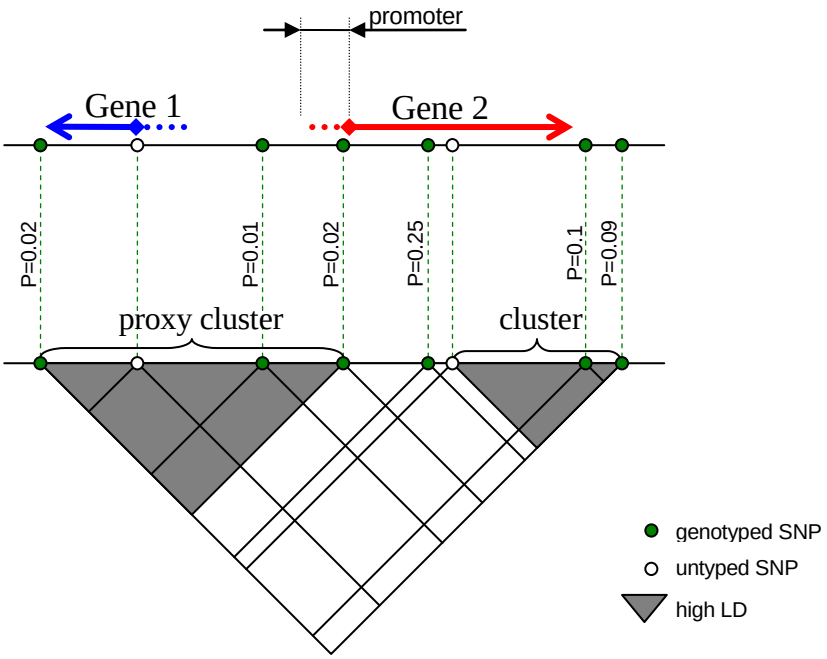

Figure 7. The illustration of how the program calculates gene-wide P-values with an artificial example. There are two genes and 8 SNPs of which six are genotyped in this study and two are proxies. Four SNPs on the left of the figure are in high LD and so are the three on the right. P-value for each genes are computed by the following formulas.
P-value for Gene1 = min(0.02, 0.01, 0.02) × 1 (the effective number) = 0.01
P-value for Gene2 = min(0.02, 0.25, min(0.1, 0.09)) × 3 = 0.06

many genes could be recognized by DAVID. Among the tested identifiers, GeneID of NCBI Entrez performed best and was selected as the standard identifier in the processes and output of ProxyGeneLD.

9.5.2 Pathway analysis

Three different applications were employed to find overrepresented pathways among a subset of a gene lists against full set. One of them was gene set enrichment analysis (GSEA) [78]. It (GSEA v2.0) was applied with slight modification of the output of ProxyGeneLD, since it requires a weight for each gene instead of a P-value, which was calculated by negative log of adjusted gene-wide P-value on the basis of 10. When negative weights appeared due to the gene-based adjustments, entire weights were adjusted by shifting to all positives and single zero. The application ran with 5000 permutations.

Another application was the database for annotation, visualization and integrated discovery (DAVID). The analyses in paper IV were performed with a March 2008 GO annotation update. The other was the Ingenuity pathway analysis (IPA; Ingenuity Systems), which is a commercial web-delivered application with its own annotation database.

9.5.3 Materials and methods

The resultant data from genome-wide association studies was obtained from the four different previous studies of Willer et al., 2008 [162], Zeggini et al., 2008 [163], Barrett et al., [164] and Aulchenko et al., 2009 [165]. The analysis of gene ranks was performed with the Mann-Whitney U test.

9.5.4 Results

To investigate the bias introduced by gene length, pathway analysis of one GWAS dataset was performed using ProxyGeneLD and DAVID. The overrepresented terms among the genes with high unadjusted P-value were similar to the enriched terms among the longest genes. To observe local clustering of functionally similar genes, taking equally distributed blocks along every chromosome, enrichment statistics were recorded. The observation revealed that there exists such clustering, which should be taken into account in any pathway analysis. To examine if study results can be influence by imputation, we compared the pathway analysis results that were performed with imputed marker data and non-imputed marker data. Most genes retained similar rank regardless of imputation, but a few outliers were observed. By the

pathway analyses with several real GWAS data, several GO term enrichment was observed.

9.6 PAPER V

9.6.1 Materials and Methods

We obtained the marker-based results data of the genome-wide association study on Alzheimer disease performed with 2032 AD cases and 5328 controls from a French population [128]. Applying ProxyGeneLD program, the single marker test results for every the SNPs were converted into a list of 16 503 gene-wide P-values. The genes that had an adjusted P-value of 0.05 or less were tested for enrichment of GO terms against the full set of the converted genes using DAVID [166] and Genecodis [79].

Examining the genomic locations of the top genes, we found that four genes of the top five are clustered in a large LD block. Applying FunCoup [159], each of the four genes (*APOE, TOMM40, PVRL2* and *BCL3*) was tested for connections in a network with the genes annotated to the overrepresented GO term found in the pathway analysis.

9.6.2 Results

By the pathway analysis, 'intracellular transmembrane protein transport' term in GO was found to be overrepresented among genes that were highly associated with AD ($P<0.05$). Since the most significantly associated gene, *TOMM40* is located in an LD block spanning *TOMM40*, *PVRL*, *BCL3*, *APOE*, the latter being the most well-known AD predisposing gene, a network analysis was performed. Among the four genes, only *TOMM40* could build a network with the top genes with the identified enriched term.

10 DISCUSSIONS

10.1 METHODOLOGICAL ASPECTS

This thesis is represented by five studies that have applied three different molecular and statistical approaches on different scales; candidate gene, target pathway, and genome-wide pathway approaches. This change in scale of my studies partially reflects the history of genetic association studies in the genetic epidemiology community. There is one missing piece to fit the complete history, which is a genome-wide association study in our own Swedish samples. Thus, two papers in this thesis deal specifically with GWAS data, but we were unable to complete a GWAS in sufficient time to be included here.

The candidate gene study in this thesis identified association of several SNPs spanning *IDE* with age-at-sampling and age-at-death, together indicating the specific genotypes of *IDE* confer increased mortality that can be seen in the population. The study proceeded to examine this finding at the molecular level, targeting *IDE* mRNA expression and insulin levels, which are the direct product of the gene and perhaps the most important substrate of the protein (IDE), respectively. This represents a fairly unique strategy since it is uncommon for genetic association findings to be followed up at the molecular level at all. Thus, most researchers tend to only present the nominal genetic association statistics with their target disease phenotype. There exist a few studies that examine the association of a genetic marker with a "molecular phenotype" representing the putative intermediate biological process that leads from genotype to disease, in order to support the epidemiological finding [167-169]. However, the concept of examining intermediate phenotypes was proposed long ago [170] and interestingly, the method is particularly popular in psychiatry studies [171]. The strategy at its core is based upon the assumption that the possible mechanism leading from genetic variant to molecular phenotype is simpler than the mechanism to explain the association at the far end of the scale, the clinically defined disease state [167]. In the study on *IDE*, the association with the various molecular phenotypes provided strong support for initial observation at two levels. In the first, obtaining association with a molecular phenotype enhances the evidence that gene under study does in fact contain functional variation, and is especially meaningful if it is the same genetic markers. For the second, the kind of association can lead to insight about the potential mechanism, where a genotype that confers increased risk for disease, can be linked to either increased or decreased expression, and this in turn tied to how it affects an intermediate trait (in our case plasma insulin level). The study on *IDE* was also particularly powerful since we were able to replicate the findings in a number of additional samples.

Candidate gene studies have their own particular advantages. The probes and assays designed for a specific gene and markers typically are highly validated. Because cost is

usually low and the hypothesis specific, redundant probes can be added to search for genotyping errors as well as to consider possibly different LD patterns from the population that has been employed for the original design that used either custom genotyping technology or a commercial genotyping chip. Put another way, to search for a paragraph in a book, we don't have to search every book in a library, if we know what book may have the paragraph. In retrospect however, the majority of genetic association studies that have targeted specific candidate genes have failed to be replicated by independent research groups using either further candidate gene studies or GWAS [172-174]. One emerging theme from the thriving GWAS community suggests the phenotypes that we are typically interested in are associated with multiple variants with small effect sizes [69,175-177]. In other words, small phrases of the paragraph have been scattered into many books. Thus candidate gene studies have evolved to be applied in the validation step of GWAS as shown in Table 1 as 'follow-up study' [6,104]. If an association of a genetic marker has been confirmed, zooming-out of the nearby region from that marker, a candidate gene study can be a fruitful avenue to genotype the region at much higher marker density using more probes for rarer SNPs. It can also be an important guide for sequencing or measuring potential intermediate molecular phenotypes [169].

The candidate pathway approach is to some extent analogous to the candidate gene approach. It has similar pros and cons, especially in comparison with GWAS. One application of the method that has to my knowledge not been tested yet is as a secondary validating study focusing on the overrepresented pathways found by a pathway analysis. Thus, establishing lists of genes from pathway analysis of GWAS to perform dense genotyping may be a fruitful avenue to pursue.

GWAS has been remarkably successful in the identification of predisposing genetic variants to a variety of human diseases and traits. However, the effect sizes of the associated variants as estimated are typically too small to explain a large fraction of heritability of the disease [69,175]. Numerous approaches have been developed to resolve this emerging problem [178]. Some of these involve testing a "set" of SNP markers [179,180], integrating with linkage studies to enhance evidence [181], and analyzing overrepresented pathways [70], the latter being upon which this thesis focuses. The putative signals in the multiple marker tests often are confounded by the surrounding noise, thus watering down the statistical significance if the investigated regions are large. The method also doesn't employ any additional information. Studies that attempt to incorporate association findings with linkage results require that samples are collected under special conditions. In contrast, the pathway approach takes advantage of biological knowledge as an additional source of information and is applicable to the vast and growing amount of GWAS data directly.

Pathway-based analyses has been most widely used in genome-wide (transcriptome-wide) expression profiling studies [182]. As the history of the use of expression

microarrays [41] is longer than for the high-throughput genotyping approaches, the various tools that have been created have had longer development time and consequently been tested extensively [77,78,183,184]. It has only relatively recently become possible to test the potential of the strategy to analyze GWAS data [70]. This lack of proper tools to handle SNP data for pathway analysis motivated study IV in this thesis. The availability of an extensive array of free applications available to explore for overrepresented pathways from significant gene lists, enabled the development of our tool for the process converting SNP to gene data.

Paper IV describes in detail how ProxyGeneLD was developed and reflects our general satisfaction with the program. We extended the use of the program to another study in the paper V specifically targeting Alzheimer disease. In terms of discoveries using the software, one of the most important is related to apparent (and false) enrichment of longer genes, which has been ignored by some of published papers that included pathway analysis strategies. The program also assists in the mapping genotyped SNPs to "proxy" genes, which is an area that is still neglected in most association studies [178]. One core premise in ProxyGeneLD that may be questioned is that the LD structure of a study population is comparable to that of the analogous cohort that has denser genotyping data such as HapMap CEU. Thus, the HapMap cohort may not reflect the LD pattern of the population under investigation, but it remains the best available and is a sufficiently valid surrogate of which data have been considered in the design of the commercial genotyping chips [185]. Numerous programs for assisting with pathway analysis of GWAS are available at present [178]. One that has recently been developed and stands out from the rest is VEGAS [186]. VEGAS adjusts gene-wide P-values considering LD using a relatively advanced statistical algorithm involving permutation, which can be implemented in our program by replacing the step of classifying "proxy cluster" with an ad-hoc threshold for high LD (r^2>0.8) [186].

The pathway approach as applied to gene expression results in this thesis demonstrated that such data can produce new insight into general biological themes in the transition youth to advanced age. There were numerous novel findings including evidence that gene structure may influence gene regulation with advancing age, as well as both up-regulated and down-regulated biological pathways that have not previously been observed. One observation, the down-regulation of mitochondrial genes was not novel, having been seen on two other occasions [187,188]. However, the successful replication of an initial finding providing vastly stronger evidence is an important indication of the potential of the approach in paper III that be applied for novel findings in future studies.

The pathway analyses in this thesis were conducted primarily with gene annotation data using the GO database. However, it should be noted that the annotation process in GO is still incomplete. The number of annotated gene is still growing. Additionally genes that have been annotated by the sequence similarity to the genes with known function

will be annotated more precisely based on biochemical study results in the near future. As GO accumulates more and more annotations, it will likely provide a greater opportunity to observe more meaningful pathway discoveries, especially if applied to rapidly expanding GWAS data in various human populations.

10.2 BIOLOGICAL ASPECTS

Each study in this thesis is a distinct entity and together may appear as unrelated biological findings. One contributing factor to this is that three out of five studies didn't have specific hypothesis, involving instead very general concepts such as the study of aging or AD. With no specific hypothesis and thousands of tests, the produced findings are thus difficult to inter-linked and seen in a unifying context. Thus, many of the biological discussions that are related primarily to the individual study are addressed in the discussion section of each constituent paper of this thesis.

The initial biological question of this thesis was "What causes Alzheimer disease in the elderly?" Because of the strong correlation between AD and age, the question was expanded to include another related scientific question "What is aging?" As a metaphor, these concepts can be partially related to the idea from hypothetical analogous research on other object, the automobile. If AD corresponds to engine failure, the first question will be "What causes engine failure?" As we are generally well aware, it can break down for numerous reasons. For example, failure in a shaft, piston, valves, and so on. But, a more fundamental cause is simply aging. The broken down components were aged chemically or mechanically. For the car, to avoid engine failure, we should understand aging first and maintain our cars to lower the speed of the aging. Likewise, it is natural that the study of AD should be accompanied by the study of aging itself.

There are two transcript variants of *IDE* registered in the NCBI RNA reference sequence collection. One is relatively long with ~120kbp of pre-mRNA (the average was ~60kbp in paper III). The other is shorter with a ~46kbp long pre-mRNA. The first has large ratio of intron to exon length (the ratio is ~32, average was ~21 in paper III). The second has a smaller intron to exon ratio of ~10. Taking into consideration that premature aging leads to a shorter life span, the observation in Paper III that the positively associated genes were compact is consistent with the expression difference in the study of IDE assuming the shorter *IDE* transcript variant was predominantly measured. Since the ages of the samples for the *IDE* expression study were rather old (~80 years) and uniform, it can be a possible mechanism that the brain cells of heterozygotes, who were not nominally more aged but biologically more aged, expressed higher levels of the shorter *IDE* transcript.

11 CONCLUSIONS

- Pathway analysis applied to genome-wide association studies has the potential to produce new biological insights
- Pathway analysis in genome-wide expression studies is a useful method to address general biological questions related to aging.
- ProxyGeneLD we developed is a useful tool for pathway analysis, especially in its ability assign SNP to "proxy" gene with high fidelity.
- Genetic variants in *IDE* are associated with age in a manner that may reflect changes in mRNA expression and plasma insulin levels
- Genetic association of an LD block including *SREBF1* with dementia.
- Genes that play a role in mitochondrial function were overrepresented among genes that have lower expression levels at advanced age.
- Negatively associated genes with increasing age tend to be non-compact, indicating that gene structure plays a role in the age-dependent regulation of transcription.
- From pathway and network analyses, *TOMM40* may play a role in Alzheimer disease development.

12 FUTURE PERSPECTIVES

The current version of ProxyGeneLD requires an additional tool such as DAVID to test overrepresentation of pathways by retrieving data in the GO or KEGG pathway databases. Due to the lack of such a function specifically designed in ProxyGeneLD, when genes with similar function are clustered in nearby chromosomal positions, enrichment statistics for pathways may indicate spurious findings. The solution introduced in paper IV has involved manual exclusion of such genes from the gene lists, which is often a tedious process. If the program has a function of directly accessing the GO database to enable the calculation of enrichment statistics by itself, then such a problem can be resolved by simple modification at the counting step where the number of genes is assigned to a pathway. Thus, regardless of how many genes are really assigned, the number of the clustered genes, even if SNPs in the genes are in high LD, is forced to be counted as 1.

As shown in Paper I, the investigation of molecular targets for genetic association produces evidence that can represent an important corroboration of the epidemiological findings. Applying a similar strategy at the genome-wide scale has a great deal of potential. Observing association of intermediate molecules in a biological metabolic pathway from DNA to RNA, to protein, and eventually to metabolites and disease, can provide evidence of the exact molecular mechanism that leads to disease. Investigating the association of genetic variants with metabolites (via "metabolomics") is an ongoing project, in which LC-MS and GC-MS were employed to measure thousands of metabolites simultaneously. It will hopefully also lead to new biological insights that can link known genetic associations to the underlying mechanisms.

13 ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to everyone who helped and supported me during this long enjoyable journey.

First of all, I would like to express deepest thanks to my supervisor, **Dr. Jonathan A. Prince**. Your guide and encouragement definitely helped me get through a lot of emerged obstacles. When I should make a choice, your brilliant idea led me to choose the right one, which turned out to be the more possible and scientifically interesting choice. Your questions and our discussions at least 5 times a day made me think more, which let me improve myself.

I am heartily grateful to my co-supervisor, **Prof. Nancy L. Pedersen**, for her welcome to MEB, to a pot-luck party, and to her office.

I would like to also thank to:

Dr. Patrik Magnusson, Dr. Fredrik Wiklund, Dr. Anna Bennet, Prof. Yudi Pawitan, Prof. Erik Ingelsson, and Dr. Woojoo Lee, for interesting scientific discussions and funny jokes

I am indebted to my friends and colleagues

in former CGB,

Fredrik, Anna, Linda, Lisa, Hagit, Mia, Vivian, Marcela, Hong, Shane, Abbas, Omid, and David

in MEB,

Martin, Iffat, Robert Karlsson, Robert Szulkin, Ralf, Therese Andersson, Therese Moberg, Fatima, Zheng, Sara Christensen, Edoardo, Hatef, Sandra, Louise, Adina, Christina, Thomas, Stefan, Jiaqi, Lovisa, Katherine, Amy, Jingmei, Elisabeth, Kaavya, Iffat, Ulrika, Alexander, Christin, Maria, Ci, Karin, Shu mei, Qian, Zongli, Sara Öberg, Anne, Mikael, Myeongjee, Michael, Zack, Marie Krushammar, Marie Jansson, Ami, Gunilla, Camilla, Ove, Åsa, Erika, Connie, Emil, Ruslan, Arvid, Johan, Fang, Andrey, Prof. Kamila Czene, and Prof. Marie Reilly

I am very grateful to friends in Korea, England, and Sweden.

I would also thank

My relatives, Ji-Yeon, Seung-Bin, You-Hyun, Dong-Min, and Rainbow(Ye-Won)

Sung Geun (father-in-law) and Kyoung Hwe (mother-in-law) for continuous attention to my works and excellent food

Hee-Soon (mother) and SangMan (father), for their endless support, love, and teaching how to live

Joong-Won and Joong-Seop, for making me laugh, happy and forget what is boring time

So-Hyun, for being the one-eye fish who see the other side of what I see. You let me have a happier life.

14 REFERENCES

1. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The human genome browser at UCSC. Genome Research 12: 996-1006.
2. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A (2010) Detection of nonneutral substitution rates on mammalian phylogenies. Genome research 20: 110-121.
3. Kruglyak L (2008) The road to genome-wide association studies. Nature Reviews Genetics 9: 314-318.
4. Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, et al. (2009) QuickGO: a web-based tool for Gene Ontology searching. Bioinformatics 25: 3045-3046.
5. Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE (2007) Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. Nature genetics 39: 17-23.
6. Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics 6: 95-108.
7. Przeworski M, Hudson RR, Di Rienzo A (2000) Adjusting the focus on human variation. Trends in Genetics 16: 296-302.
8. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, et al. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409: 928-933.
9. The 1000 Genome Project Consortium, Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.
10. Li WH, Sadler LA (1991) Low nucleotide diversity in man. Genetics 129: 513-523.
11. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, et al. (2002) A high-resolution recombination map of the human genome. Nature genetics 31: 241-247.
12. The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861.
13. Slatkin M (2008) Linkage disequilibrium - understanding the evolutionary past and mapping the medical future. Nature Reviews Genetics 9: 477-485.
14. Nei M, Li WH (1973) Linkage Disequilibrium in Subdivided Populations. Genetics 75: 213-219.
15. Weir BS, Cockerha.Cc (1969) Group Inbreeding with 2 Linked Loci. Genetics 63: 711-&.
16. Golding GB, Strobeck C (1980) Linkage Disequilibrium in a Finite Population That Is Partially Selfing. Genetics 94: 777-789.
17. The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437: 1299-1320.
18. Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: Models and data. American journal of human genetics 69: 1-14.
19. Hill WG, Robertson A (1968) Linkage Disequilibrium in Finite Populations. Theoretical and Applied Genetics 38: 226-231.
20. Burton PR, Tobin MD, Hopper JL (2005) Key concepts in genetic epidemiology. Lancet 366: 941-951.
21. de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, et al. (2005) Efficiency and power in genetic association studies. Nature Genetics 37: 1217-1223.
22. Alberts B, Bray D, Johnson A, Lewis J, Raff M, et al. (1998) Genetic Variation. Essential Cell Biology. Newyork: Garland. pp. 277-312.
23. Hardy GH (1908) Mendelian Proportions in a Mixed Population. Science (New York, NY) 28: 49-50.
24. Weinberg W (1908) On the demonstration of heredity in man. In: Boyer SH, translator. (1963) Papers on Human Genetics. Englewood Cliffs, NJ: Prentice-Hall.
25. Stern C (1943) The Hardy-Weinberg Law. Science (New York, NY) 97: 137-138.
26. Guo SW, Thompson EA (1992) Performing the Exact Test of Hardy-Weinberg Proportion for Multiple Alleles. Biometrics 48: 361-372.

27. Wigginton JE, Cutler DJ, Abecasis GR (2005) A note on exact tests of Hardy-Weinberg equilibrium. American journal of human genetics 76: 887-893.
28. Cardon LR, Bell JI (2001) Association study designs for complex diseases. Nature reviews Genetics 2: 91-99.
29. Cordell HJ, Clayton DG (2005) Genetic association studies. Lancet 366: 1121-1131.
30. Brookes AJ (1999) The essence of SNPs. Gene 234: 177-186.
31. Altshuler D, Daly M, Kruglyak L (2000) Guilt by association. Nature genetics 26: 135-137.
32. Srinivasan M, Sedmak D, Jewell S (2002) Effect of fixatives and tissue processing on the content and integrity of nucleic acids. American Journal of Pathology 161: 1961-1971.
33. Micke P, Ohshima M, Tahmasebpoor S, Ren ZP, Ostman A, et al. (2006) Biobanking of fresh frozen tissue: RNA is stable in nonfixed surgical specimens. Laboratory Investigation 86: 202-211.
34. Harrison PJ, Heath PR, Eastwood SL, Burnet PWJ, McDonald B, et al. (1995) The relative importance of premortem acidosis and postmortem interval for human brain gene expression studies: Selective mRNA vulnerability and comparison with their encoded proteins. Neuroscience letters 200: 151-154.
35. Barton AJ, Pearson RC, Najlerahim A, Harrison PJ (1993) Pre- and postmortem influences on brain RNA. J Neurochem 61: 1-11.
36. Livak KJ, Flood SJA, Marmaro J, Giusti W, Deetz K (1995) Oligonucleotides with Fluorescent Dyes at Opposite Ends Provide a Quenched Probe System Useful for Detecting Pcr Product and Nucleic-Acid Hybridization. Pcr-Methods and Applications 4: 357-362.
37. Heid CA, Stevens J, Livak KJ, Williams PM (1996) Real time quantitative PCR. Genome research 6: 986-994.
38. Pfaffl MW (2001) A new mathematical model for relative quantification in real-time RT-PCR. Nucleic acids research 29: -.
39. Bioinformatics & relative quantification using real time PCR. Available: http://www.gene-quantification.de/relative.html.
40. Gutala RV, Reddy PH (2004) The use of real-time PCR analysis in a gene expression study of Alzheimer's disease post-mortem brains. Journal of Neuroscience Methods 132: 101-107.
41. Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative Monitoring of Gene-Expression Patterns with a Complementary-DNA Microarray. Science (New York, NY) 270: 467-470.
42. Shalon D, Smith SJ, Brown PO (1996) A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome research 6: 639-645.
43. Lockhart DJ, Dong HL, Byrne MC, Follettie MT, Gallo MV, et al. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature Biotechnology 14: 1675-1680.
44. Hoheisel JD (2006) Microarray technology: beyond transcript profiling and genotype analysis. Nature Reviews Genetics 7: 200-210.
45. Duckworth WC, Bennett RG, Hamel FG (1998) Insulin degradation: progress and potential. Endocrine reviews 19: 608-624.
46. Kurochkin IV, Goto S (1994) Alzheimer's beta-amyloid peptide specifically interacts with and is degraded by insulin degrading enzyme. FEBS letters 345: 33-37.
47. Bertram L, Blacker D, Mullin K, Keeney D, Jones J, et al. (2000) Evidence for genetic linkage of Alzheimer's disease to chromosome 10q. Science (New York, NY) 290: 2302-+.
48. Ertekin-Taner N, Graff-Radford N, Younkin LH, Eckman C, Baker M, et al. (2000) Linkage of plasma A beta 42 to a quantitative locus on chromosome 10 in late-onset Alzheimer's disease pedigrees. Science (New York, NY) 290: 2303-+.
49. Farris W, Leissring MA, Hemming ML, Chang AY, Selkoe DJ (2005) Alternative splicing of human insulin-degrading enzyme yields a novel isoform with a

decreased ability to degrade insulin and amyloid beta-protein. Biochemistry 44: 6513-6525.
50. Leissring MA, Farris W, Wu XN, Christodoulou DC, Haigis MC, et al. (2004) Alternative translation initiation generates a novel isoform of insulin-degrading enzyme targeted to mitochondria. Biochemical Journal 383: 439-446.
51. Pfrieger FW (2003) Cholesterol homeostasis and function in neurons of the central nervous system. Cellular and Molecular Life Sciences 60: 1158-1171.
52. Papassotiropoulos A, Lutjohann D, Bagli M, Locatelli S, Jessen F, et al. (2002) 24S-hydroxycholesterol in cerebrospinal fluid is elevated in early stages of dementia. Journal of Psychiatric Research 36: 27-32.
53. Schonknecht P, Lutjohann D, Pantel J, Bardenheuer H, Hartmann T, et al. (2002) Cerebrospinal fluid 24S-hydroxycholesterol is increased in patients with Alzheimer's disease compared to healthy controls. Neuroscience letters 324: 83-85.
54. Heverin M, Bogdanovic N, Lutjohann D, Bayer T, Pikuleva I, et al. (2004) Changes in the levels of cerebral and extracerebral sterols in the brain of patients with Alzheimer's disease. Journal of Lipid Research 45: 186-193.
55. Strittmatter WJ, Saunders AM, Schmechel D, Pericak-Vance M, Enghild J, et al. (1993) Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. Proc Natl Acad Sci U S A 90: 1977-1981.
56. Beffert U, Aumont N, Dea D, Lussier-Cacan S, Davignon J, et al. (1999) Apolipoprotein E isoform-specific reduction of extracellular amyloid in neuronal cultures. Brain Res Mol Brain Res 68: 181-185.
57. Prince JA, Zetterberg H, Andreasen N, Marcusson J, Blennow K (2004) APOE epsilon4 allele is associated with reduced cerebrospinal fluid levels of Abeta42. Neurology 62: 2116-2118.
58. Syvänen A-C (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. Nature reviews Genetics 2: 930-942.
59. Tabor HK, Risch NJ, Myers RM (2002) Candidate-gene approaches for studying complex genetic traits: practical considerations. Nature Reviews Genetics 3: 391-397.
60. Wang WYS, Barratt BJ, Clayton DG, Todd JA (2005) Genome-wide association studies: theoretical and practical concerns. Nature Reviews Genetics 6: 109-118.
61. Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, et al. (2002) Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nature Genetics 32: 650-654.
62. Ohnishi Y, Tanaka T, Ozaki K, Yamada R, Suzuki H, et al. (2001) A high-throughput SNP typing system for genome-wide association studies. Journal of Human Genetics 46: 471-477.
63. Matsuzaki H, Dong SL, Loi H, Di XJ, Liu GY, et al. (2004) Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. Nature Methods 1: 109-111.
64. Gunderson KL, Steemers FJ, Lee G, Mendoza LG, Chee MS (2005) A genome-wide scalable SNP genotyping assay using microarray technology. Nature genetics 37: 549-554.
65. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America 106: 9362-9367.
66. Tsui C, Coleman LE, Griffith JL, Bennett EA, Goodson SG, et al. (2003) Single nucleotide polymorphisms (SNPs) that map to gaps in the human SNP map. Nucleic acids research 31: 4910-4916.
67. The International HapMap Consortium, Gibbs RA, Belmont JW, Hardenbol P, Willis TD, et al. (2003) The International HapMap Project. Nature 426: 789-796.
68. Hao K (2007) Genome-wide selection of tag SNPs using multiple-marker correlation. Bioinformatics 23: 3178-3184.

69. Ku CS, Loy EY, Pawitan Y, Chia KS (2010) The pursuit of genome-wide association studies: where are we now? Journal of Human Genetics 55: 195-206.
70. Wang K, Li MY, Bucan M (2007) Pathway-based approaches for analysis of genomewide association studies. American journal of human genetics 81: 1278-1283.
71. The Gene Ontology Consortium (2010) Gene Ontology Documentation. Available: http://www.geneontology.org/GO.contents.doc.shtml. Accessed 2010-09-14.
72. The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. Nature genetics 25: 25-29.
73. Rhee SY, Wood V, Dolinski K, Draghici S (2008) Use and misuse of the gene ontology annotations. Nature reviews Genetics 9: 509-515.
74. The Gene Ontology Consortium (2010) The Gene Ontology in 2010: extensions and refinements. Nucleic Acids Res 38: D331-335.
75. Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic acids research 28: 27-30.
76. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, et al. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic acids research 27: 29-34.
77. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, et al. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biology 4: P3.
78. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America 102: 15545-15550.
79. Nogales-Cadenas R, Carmona-Saez P, Vazquez M, Vicente C, Yang X, et al. (2009) GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. Nucleic acids research 37: W317-322.
80. Walpole RE, Myers RH, Myers S, Ye K (2002) Simple linear regression. In: Yagan S, editor Probability and statistics for engineers and scientists (7th edition). Upper Saddle River, NJ: Prentice Hall. pp. 351-354.
81. Bland M (2000) The analysis of cross-tabulations. An introduction to medical statistics. Oxford: Oxford University Press. pp. 230-256.
82. Tan QH, Bathum L, Christiansen L, De Benedictis G, Dahlgaard J, et al. (2003) Logistic regression models for polymorphic and antagonistic pleiotropic gene action on human aging and longevity. Annals of human genetics 67: 598-607.
83. Massey FJ (1951) The Kolmogorov-Smirnov Test for Goodness of Fit. Journal of the American Statistical Association 46: 68-78.
84. (1998) Nonparametrics. StatView reference: SAS Institute. pp. 119-130.
85. Shapiro SS, Wilk MB (1965) An Analysis of Variance Test for Normality (Complete Samples). Biometrika 52: 591-&.
86. DYER AR (1974) Comparisons of tests for normality with a cautionary note. Biometrika 61: 185-189.
87. Shapiro SS, Wilk MB, Chen HJ (1968) A Comparative Study of Various Tests for Normality. Journal of the American Statistical Association 63: 1343-&.
88. Weisstein EW "Bonferroni Inequalities". From MathWorld: A Wolfram Web Resource. http://mathworld.wolfram.com/BonferroniInequalities.html.
89. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America 100: 9440-9445.
90. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological 57: 289-300.
91. Draghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA (2003) Global functional profiling of gene expression. Genomics 81: 98-104.
92. Bland M (2000) Methods based on rank order. An introduction to medical statistics. Oxford: Oxford University Press. pp. 210-229.

93. Wilcoxon F (1945) Individual Comparisons by Ranking Methods. Biometrics Bulletin 1: 80-83.
94. Mann HB, Whitney DR (1947) On a Test of Whether One of 2 Random Variables Is Stochastically Larger Than the Other. Annals of Mathematical Statistics 18: 50-60.
95. Gatz M, Fratiglioni L, Johansson B, Berg S, Mortimer JA, et al. (2005) Complete ascertainment of dementia in the Swedish Twin Registry: the HARMONY study. Neurobiology of aging 26: 439-447.
96. Hendrie HC (1998) Epidemiology of dementia and Alzheimer's disease. American Journal of Geriatric Psychiatry 6: S3-S18.
97. Ott A, Breteler MM, van Harskamp F, Claus JJ, van der Cammen TJ, et al. (1995) Prevalence of Alzheimer's disease and vascular dementia: association with education. The Rotterdam study. BMJ 310: 970-973.
98. Brookmeyer R, Evans DA, Hebert L, Langa KM, Heeringa SG, et al. (2011) National estimates of the prevalence of Alzheimer's disease in the United States. Alzheimer's and Dementia 7: 61-73.
99. Ferri CP, Prince M, Brayne C, Brodaty H, Fratiglioni L, et al. (2005) Global prevalence of dementia: a Delphi consensus study. Lancet 366: 2112-2117.
100. Kalaria RN, Maestre GE, Arizaga R, Friedland RP, Galasko D, et al. (2008) Alzheimer's disease and vascular dementia in developing countries: prevalence, management, and risk factors. Lancet Neurology 7: 812-826.
101. Jönsson L (2004) Economic evidence in dementia: a review. Eur J Health Econ 5 Suppl 1: S30-35.
102. Wimo A, Winblad B, Jönsson L (2010) The worldwide societal costs of dementia: Estimates for 2009. Alzheimer's & Dementia 6: 98-103.
103. Selkoe DJ (1991) The molecular pathology of Alzheimer's disease. Neuron 6: 487-498.
104. Bertram L, Lill CM, Tanzi RE (2010) The genetics of Alzheimer disease: back to the future. Neuron 68: 270-281.
105. Mattson MP (2004) Pathways towards and away from Alzheimer's disease. Nature 430: 631-639.
106. Tanzi RE, Bertram L (2005) Twenty years of the Alzheimer's disease amyloid hypothesis: a genetic perspective. Cell 120: 545-555.
107. Sherrington R, Rogaev EI, Liang Y, Rogaeva EA, Levesque G, et al. (1995) Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. Nature 375: 754-760.
108. Levy-Lahad E, Wasco W, Poorkaj P, Romano DM, Oshima J, et al. (1995) Candidate gene for the chromosome 1 familial Alzheimer's disease locus. Science (New York, NY) 269: 973-977.
109. Bertram L, Tanzi RE (2004) Alzheimer's disease: one disorder, too many genes? Hum Mol Genet 13 Spec No 1: R135-141.
110. Blennow K, de Leon MJ, Zetterberg H (2006) Alzheimer's disease. Lancet 368: 387-403.
111. Walsh DM, Selkoe DJ (2004) Deciphering the molecular basis of memory failure in Alzheimer's disease. Neuron 44: 181-193.
112. Kobayashi DT, Chen KS (2005) Behavioral phenotypes of amyloid-based genetically modified mouse models of Alzheimer's disease. Genes Brain Behav 4: 173-196.
113. Vardy ER, Catto AJ, Hooper NM (2005) Proteolytic mechanisms in amyloid-beta metabolism: therapeutic implications for Alzheimer's disease. Trends Mol Med 11: 464-472.
114. Caccamo A, Oddo S, Sugarman MC, Akbari Y, LaFerla FM (2005) Age- and region-dependent alterations in Abeta-degrading enzymes: implications for Abeta-induced disorders. Neurobiol Aging 26: 645-654.
115. Carson JA, Turner AJ (2002) Beta-amyloid catabolism: roles for neprilysin (NEP) and other metallopeptidases? J Neurochem 81: 1-8.
116. Tanzi RE, Moir RD, Wagner SL (2004) Clearance of Alzheimer's Abeta peptide: the many roads to perdition. Neuron 43: 605-608.

117. Kim J, Basak JM, Holtzman DM (2009) The role of apolipoprotein E in Alzheimer's disease. Neuron 63: 287-303.
118. Strittmatter WJ, Saunders AM, Schmechel D, Pericak-Vance M, Enghild J, et al. (1993) Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. Proceedings of the National Academy of Sciences of the United States of America 90: 1977-1981.
119. Grupe A, Abraham R, Li Y, Rowland C, Hollingworth P, et al. (2007) Evidence for novel susceptibility genes for late-onset Alzheimer's disease from a genome-wide association study of putative functional variants. Human molecular genetics 16: 865-873.
120. Coon KD, Myers AJ, Craig DW, Webster JA, Pearson JV, et al. (2007) A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. Journal of Clinical Psychiatry 68: 613-618.
121. Reiman EM, Webster JA, Myers AJ, Hardy J, Dunckley T, et al. (2007) GAB2 alleles modify Alzheimer's risk in APOE epsilon4 carriers. Neuron 54: 713-720.
122. Li H, Wetten S, Li L, Jean PLS, Upmanyu R, et al. (2008) Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. Archives of neurology 65: 45-53.
123. Poduslo SE, Huang R, Huang J, Smith S (2009) Genome Screen of Late-Onset Alzheimer's Extended Pedigrees Identifies TRPC4AP by Haplotype Analysis. American Journal of Medical Genetics Part B-Neuropsychiatric Genetics 150B: 50-55.
124. Abraham R, Moskvina V, Sims R, Hollingworth P, Morgan A, et al. (2008) A genome-wide association study for late-onset Alzheimer's disease using DNA pooling. Bmc Medical Genomics 1: -.
125. Bertram L, Lange C, Mullin K, Parkinson M, Hsiao M, et al. (2008) Genome-wide Association Analysis Reveals Putative Alzheimer's Disease Susceptibility Loci in Addition to APOE. American journal of human genetics 83: 623-632.
126. Beecham GW, Martin ER, Li YJ, Slifer MA, Gilbert JR, et al. (2009) Genome-wide Association Study Implicates a Chromosome 12 Risk Locus for Late-Onset Alzheimer Disease. American journal of human genetics 84: 35-43.
127. Carrasquillo MM, Zou FG, Pankratz VS, Wilcox SL, Ma L, et al. (2009) Genetic variation in PCDH11X is associated with susceptibility to late-onset Alzheimer's disease. Nature genetics 41: 192-198.
128. Lambert JC, Heath S, Even G, Campion D, Sleegers K, et al. (2009) Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nature genetics 41: 1094-1099.
129. Harold D, Abraham R, Hollingworth P, Sims R, Gerrish A, et al. (2009) Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nature genetics 41: 1088-1093.
130. Heinzen EL, Need AC, Hayden KM, Chiba-Falek O, Roses AD, et al. (2010) Genome-Wide Scan of Copy Number Variation in Late-Onset Alzheimer's Disease. Journal of Alzheimers Disease 19: 69-77.
131. Potkin SG, Guffanti G, Lakatos A, Turner JA, Kruggel F, et al. (2009) Hippocampal Atrophy as a Quantitative Trait in a Genome-Wide Association Study Identifying Novel Susceptibility Genes for Alzheimer's Disease. PloS one 4: -.
132. Seshadri S, Fitzpatrick AL, Ikram MA, DeStefano AL, Gudnason V, et al. (2010) Genome-wide Analysis of Genetic Loci Associated With Alzheimer Disease. Jama-Journal of the American Medical Association 303: 1832-1840.
133. Naj AC, Beecham GW, Martin ER, Gallins PJ, Powell EH, et al. (2010) Dementia Revealed: Novel Chromosome 6 Locus for Late-Onset Alzheimer Disease Provides Genetic Evidence for Folate-Pathway Abnormalities. PLoS genetics 6: -.
134. Drachman DA (2006) Aging of the brain, entropy, and Alzheimer disease. Neurology 67: 1340-1352.

135. Salthouse TA (1996) The processing-speed theory of adult age differences in cognition. Psychological Review 103: 403-428.
136. Lehman HC (1966) Most Creative Years of Engineers and Other Technologists. Journal of Genetic Psychology 108: 263-&.
137. Salthouse TA (1999) Theories of cognition. In: Bengtson V, Schaie, KW, editor Handbook of theories of aging. New York: Springer. pp. 196–208.
138. Pakkenberg B, Pelvig D, Marner L, Bundgaard MJ, Gundersen HJG, et al. (2003) Aging and the human neocortex. Experimental Gerontology 38: 95-99.
139. West MJ, Coleman PD, Flood DG, Troncoso JC (1995) Differential neuronal loss in the hippocampus in normal aging and in patients with Alzheimer disease. Ugeskr Laeger 157: 3190-3193.
140. West MJ, Kawas CH, Martin LJ, Troncoso JC (2000) The CA1 region of the human hippocampus is a hot spot in Alzheimer's disease. Molecular and Cellular Gerontology 908: 255-259.
141. Weinert BT, Timiras PS (2003) Theories of aging. Journal of Applied Physiology 95: 1706-1716.
142. Hayflick L (1965) Limited in Vitro Lifetime of Human Diploid Cell Strains. Experimental Cell Research 37: 614-&.
143. Olovniko.Am (1971) Principle of Marginotomy in Template Synthesis of Polynucleotides. Doklady Akademii Nauk Sssr 201: 1496-&.
144. Watson JD (1972) Origin of Concatemeric T7 DNA. Nature-New Biology 239: 197-&.
145. Kuilman T, Michaloglou C, Mooi WJ, Peeper DS (2010) The essence of senescence. Genes & Development 24: 2463-2479.
146. Kenyon C (2005) The plasticity of aging: Insights from long-lived mutants. Cell 120: 449-460.
147. Vijg J, Campisi J (2008) Puzzles, promises and a cure for ageing. Nature 454: 1065-1071.
148. Finkel T (2003) Oxidant signals and oxidative stress. Current Opinion in Cell Biology 15: 247-254.
149. Eriksson UK (2010) Inflammation-associated risk factors for Alzheimer's disease and dementia. Stockholm: Karolinska Institutet.
150. Howell WM, Jobs M, Brookes AJ (2002) iFRET: an improved fluorescence system for DNA-melting analysis. Genome research 12: 1401-1407.
151. Reynolds CA, Hong MG, Eriksson UK, Blennow K, Bennet AM, et al. (2009) A Survey of ABCA1 Sequence Variation Confirms Association with Dementia. Human mutation 30: 1348-1354.
152. Pedersen NL, Friberg L, Floderus-Myrhed B, McClearn GE, Plomin R (1984) Swedish early separated twins: identification and characterization. Acta geneticae medicae et gemellologiae 33: 243-250.
153. McClearn GE, Johansson B, Berg S, Pedersen NL, Ahern F, et al. (1997) Substantial genetic influence on cognitive abilities in twins 80 or more years old. Science (New York, NY) 276: 1560-1563.
154. Gold CH, Malmberg B, McClearn GE, Pedersen NL, Berg S (2002) Gender and health: a study of older unlike-sex twins. The journals of gerontology Series B, Psychological sciences and social sciences 57: S168-176.
155. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, et al. (2007) A genome-wide association study of global gene expression. Nature genetics 39: 1202-1207.
156. Myers AJ, Gibbs JR, Webster JA, Rohrer K, Zhao A, et al. (2007) A survey of genetic human cortical gene expression. Nature genetics 39: 1494-1499.
157. Andreasen N, Minthon L, Clarberg A, Davidsson P, Gottfries J, et al. (1999) Sensitivity, specificity, and stability of CSF-tau in AD in a community-based patient sample. Neurology 53: 1488-1494.
158. Carey V, Zeger SL, Diggle P (1993) Modelling multivariate binary data with alternating logistic regressions. Biometrika 80: 517-526.
159. Alexeyenko A, Sonnhammer EL (2009) Global networks of functional coupling in eukaryotes from comprehensive data integration. Genome research 19: 1107-1116.

160. Göring HH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, et al. (2007) Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nature genetics 39: 1208-1216.
161. Lu T, Pan Y, Kao SY, Li C, Kohane I, et al. (2004) Gene regulation and DNA damage in the ageing human brain. Nature 429: 883-891.
162. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, et al. (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nature genetics 40: 161-169.
163. Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, et al. (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nature genetics 40: 638-645.
164. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, et al. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nature genetics 40: 955-962.
165. Aulchenko YS, Ripatti S, Lindqvist I, Boomsma D, Heid IM, et al. (2009) Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. Nature genetics 41: 47-55.
166. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols 4: 44-57.
167. Becker KG, Barnes KC, Bright TJ, Wang SA (2004) The Genetic Association Database. Nature genetics 36: 431-432.
168. Lazarus R, Raby BA, Lange C, Silverman EK, Kwiatkowski DJ, et al. (2004) TOLL-like receptor 10 genetic variation is associated with asthma in two independent samples. American Journal of Respiratory and Critical Care Medicine 170: 594-600.
169. Ozaki K, Sato H, Iida A, Mizuno H, Nakamura T, et al. (2006) A functional SNP in PSMA6 confers risk of myocardial infarction in the Japanese population. Nature genetics 38: 921-925.
170. Flint J, Munafo MR (2007) The endophenotype concept in psychiatric genetics. Psychological Medicine 37: 163-180.
171. Gottesman II, Gould TD (2003) The endophenotype concept in psychiatry: Etymology and strategic intentions. American Journal of Psychiatry 160: 636-645.
172. Ioannidis JPA, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG (2001) Replication validity of genetic association studies. Nature genetics 29: 306-309.
173. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. Genetics in Medicine 4: 45-61.
174. Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, et al. (2007) Replicating genotype-phenotype associations. Nature 447: 655-660.
175. Hirschhorn JN, Gajdos ZK (2011) Genome-wide association studies: results from the first few years and potential implications for clinical medicine. Annu Rev Med 62: 11-24.
176. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, et al. (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460: 748-752.
177. Schork NJ, Murray SS, Frazer KA, Topol EJ (2009) Common vs. rare allele hypotheses for complex diseases. Current Opinion in Genetics & Development 19: 212-219.
178. Wang K, Li MY, Hakonarson H (2010) Analysing biological pathways in genome-wide association studies. Nature Reviews Genetics 11: 843-854.
179. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, et al. (2010) Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. American journal of human genetics 86: 929-942.
180. Gauderman WJ, Murcray C, Gilliland F, Conti DV (2007) Testing association between disease and multiple SNPs in a candidate gene. Genetic epidemiology 31: 383-395.

181. Roeder K, Bacanu SA, Wasserman L, Devlin B (2006) Using linkage genome scans to improve power of association in genome scans. American journal of human genetics 78: 243-252.
182. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America 95: 14863-14868.
183. Grosu P, Townsend JP, Hartl DL, Cavalieri D (2002) Pathway processor: A tool for integrating whole-genome expression results into metabolic networks. Genome research 12: 1121-1126.
184. Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, et al. (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biology 4: -.
185. illumina (2010) Genome-Wide DNA Analysis BeadChips.
186. Liu JZ, Mcrae AF, Nyholt DR, Medland SE, Wray NR, et al. (2010) A Versatile Gene-Based Test for Genome-wide Association Studies. American journal of human genetics 87: 139-145.
187. Zahn JM, Sonu R, Vogel H, Crane E, Mazan-Mamczarz K, et al. (2006) Transcriptional profiling of aging in human muscle reveals a common aging signature. PLoS genetics 2: 1058-1069.
188. Miller JA, Oldham MC, Geschwind DH (2008) A systems level analysis of transcriptional changes in Alzheimer's disease and normal aging. Journal of Neuroscience 28: 1410-1420.