From CENTER FOR GENOMICS AND BIOINFORMATICS
Karolinska Institutet, Stockholm, Sweden

SNP BASED STRATEGIES TO STUDY CANDIDATE GENES FOR ALZHEIMER'S DISEASE

Lars Feuk

KONGL CAROLINSKA MEDICO CHIRURGISKA INSTITUTET

Stockholm 2002

Repro Print AB, Stockholm, Sweden
Box 200, SE-171 77 Stockholm, Sweden

ISBN 91-7349-334-1

ABSTRACT

Alzheimer's disease (AD) is the most common form of dementia in the elderly. It is a genetically heterogeneous disease characterized by progressive cognitive decline and memory impairment. The rare familial form of AD is caused by three different genes called APP, PSEN1 and PSEN2. However, the predominant form of AD is a genetically complex disorder involving a combination of genetic factors. To date, the only risk factor identified for the complex form of AD is the *APOE-ε4* allele, but several susceptibility genes remain to be found.

This thesis outlines different strategies to use common genetic variation, in the form of single nucleotide polymorphisms (SNPs), to examine candidate genes and candidate regions for AD. Large-scale genotyping is a prerequisite for performing complex disease studies using SNPs. The validity and accuracy of a newly developed genotyping assay called Dynamic allele specific hybridization (DASH) was therefore investigated. DASH was shown to be a robust genotyping method, and was proven to work as well or better than several other available methods. The method was first implemented for a candidate gene association study of a promoter polymorphism in the *TNFRSF6* gene. Significant association was found between this variant and early onset AD, indicating its possible role in disease etiology.

A large candidate pathway association study effort was then started testing for association between AD and 60 different SNPs. Genes were picked from four different pathways related to AD; oxidation, inflammation/apoptosis, amyloid interacting genes and a group of candidate genes previously showing significant association with AD. None of the markers showed significant disease association after correction for multiple testing. Although largely negative, these results high-lighted several methodological and study design issues related to association studies in general.

The most successful approach yet in dissecting complex disease using genetic variation has been to perform high resolution linkage disequilibrium (LD) mapping of regions indicated by linkage. Several independent research groups recently reported linkage peaks for AD on chromosome 10q. We choose two regions under the 10q linkage peak for LD mapping studies. The first region contained the previously associated *TNFRSF6* gene, and the other region included the insulin-degrading enzyme (*IDE*) gene, which has been shown to be involved in clearance of amyloid-beta. LD maps were created for all pair-wise markers in the two regions to determine the genetic LD structure. Haplotypes were estimated and haplotype tagging markers were chosen for further analysis. Association analyses were performed for both single markers and haplotypes for case/control status as well as for quantitative traits related to the AD phenotype. Only weak significant signals were found for the *TNFRSF6* gene. However, several significant associations were found for a large LD block including the *IDE* gene. The same haplotypes were always over-represented in cases compared to controls, or with more severe AD within the patient groups. These results indicate a role in AD for one of the three genes situated within the 276kb LD block including the *IDE*, *KNSL1* and *HHEX* genes. Further studies will now be required to identify the underlying risk alleles within the region.

LIST OF PUBLICATIONS

This thesis is based on the following papers, which will be referred to in the text by their Roman numerals.

I **Feuk L**, Prince JA, Breen G, Emahazion T, Carothers A, St Clair D, Brookes AJ
Apolipoprotein-E dependent role for the FAS receptor in early onset Alzheimer's disease: finding of a positive association for a polymorphism in the *TNFRSF6* gene. *Hum Genet. 2000 Oct; 107(4):391-6.*

II Prince JA, **Feuk L**, Howell WM, Jobs M, Emahazion T, Blennow K, Brookes AJ
Robust and accurate single nucleotide polymorphism genotyping by dynamic allele-specific hybridization (DASH): design criteria and assay validation.
Genome Res. 2001 Jan;11(1):152-62.

III Emahazion T*, **Feuk L***, Jobs M, Sawyer SL, Fredman D, St Clair D, Prince JA, Brookes AJ
SNP association studies in Alzheimer's disease highlight problems for complex disease analysis.
Trends Genet. 2001 Jul;17(7):407-13.

IV **Feuk L**, Prince JA, Blennow K, Brookes AJ
Further evidence for role of a promoter variant in the *TNFRSF6* gene in Alzheimer's disease.
Hum Mut. In press.

V Prince JA, **Feuk L**, Gu HF, Margaret Gatz, Blennow K, Brookes AJ
Genetic variation in a haplotype block spanning *IDE, KNSL1* and *HHEX* influences Alzheimer's Disease.
Manuscript.

* These authors contributed equally to the project

CONTENTS

LIST OF ABBREVIATIONS

Aβ	Amyloid-beta
AD	Alzheimer's disease
APOE	Apolipoprotein E
APOE-e4	Apolipoprotein E-ε4
APP	Amyloid precursor protein
ASOH	Allele specific oligonucleotide hybridization
Bp	Base pair
cDNA	Coding DNA
CSF	Cerebrospinal fluid
DASH	Dynamic Allele-Specific Hybridization
DNA	Deoxyribonucleic acid
EOAD	Early onset Alzheimer's disease
FAD	Familial Alzheimer's disease
FRET	Fluorescence resonance energy transfer
htSNP	Haplotype tagging SNP
IDE	Insulin-degrading enzyme
Kb	Kilobases
KNSL1	Kinesin-like 1
LD	Linkage disequilibrium
LNA	Locked nucleic acid
LOAD	Late onset Alzheimer's disease
MCI	Mild cognitive impairment
MMSE	Mini mental state examination
PNA	Peptide nucleic acid
PSEN1	Presenilin 1
PSEN2	Presenilin 2
RFLP	Restriction fragment length polymorphism
SNP	Single nucleotide polymorphism
SP-NFT	Senile plaque and neurofibrillary tangle density
Tm	Melting temperature
TNFRSF6	Tumor necrosis factor receptor superfamily, member 6

1 INTRODUCTION

When the structure of DNA was described by Watson and Crick in 1953[1], it marked the beginning of a rapid development of the field of molecular genetics. We know now that the human DNA sequence (the human genome) contains approximately 3.3 billion "letters" of genetic code. The DNA sequence is made up of simple molecules called nucleotides. There are four different nucleotides called adenine, guanine, cytosine and thymine and they are therefore usually referred to by their abbreviations; A, G, C and T.

The scientific progress and the enthusiasm in the research community ultimately led to the start of the Human Genome Project in 1990. Originally, a 15 year project was planned that included not only sequencing of the human genome, but also to sequence a number of model organisms, to develop better technology for sequencing, to create maps with markers covering the human genome, to develop bioinformatics for handling and interpreting sequence data and to establish ethical guidelines regarding genetic information. The progress of the project was much more rapid than anticipated, and the time plan had to be revised in 1993, and then again in 1998[2]. A draft sequence of the human genome was presented in June 2000, years ahead of the original schedule[3]. Now that we have a draft sequence, there are several challenges ahead. First, there is a need to understand the function of the human DNA sequence. Where are the genes in the genome, what is the product and ultimate function of all genes, how are they regulated and what is the purpose of inter-genic DNA sequences? Second, it is important to identify variation in the genetic code in order to understand why people are different, especially with regards to susceptibility to disease and response to pharmaceutical treatment. Identifying genetic differences within and between populations will also help us decipher human evolution and how people spread throughout the world. The field of genetics is now at a point where millions of common genetic variants have been identified in the human genome, but the strategies to use this information for studies of human disease are still underdeveloped. This thesis describes some of the approaches that make use of human genetic variation to study common disease.

1.1 INHERITED DISEASE

An understanding of the inheritance of traits is evident from old records of animal breeding and cross-hybridization of plants, and the mechanisms of how traits could be transferred from parent to offspring were discussed by the famous philosophers of Ancient Greece[4]. However, knowledge about the molecular basis of genetically inherited traits was not achieved until the age of modern molecular biology.

We now know that humans carry 46 chromosomes, 23 from the mother and 23 from the father. Each parent contributes 22 autosomes and one sex chromosome. The autosomal chromosomes are numbered 1-22 and the two sex chromosomes are called X and Y. A person carrying two X chromosomes is female and a person carrying one X and one Y is male. One of the two copies of a chromosome in the parent is passed on to the child, and it is random which of the two copies that is transmitted (referred to as random segregation). Errors in the genetic code can lead to gene dysfunction and cause disease. The genetic error, and thereby the disease, can then be passed on from parent to child. The genetic diseases are usually grouped into monogenic disease, where mutations in one gene is enough to cause disease, and polygenic or complex genetic disease where alterations in several genes, often with environmental influence, lead to the disease phenotype.

1.1.1 Monogenic disease

Monogenic diseases are genetically simple in that they are caused by aberrations in a single gene, and they usually segregate in families according to Mendel's law of inheritance. The genetic alterations can be any kind of sequence changes such as deletions, insertions, point mutations or repeat expansions. According to the laws of Mendel, traits segregate either in a dominant fashion (only one of a pair of chromosomes need to carry the disease allele for the disease phenotype to be expressed), or in a recessive manner (the phenotype is expressed only when both copies of a pair of chromosomes carry the disease allele). There are now more than 4000 diseases described, and about 450 of these have no known associated gene[5,6]. More than half of all the phenotypes described exhibit autosomal dominant inheritance. Examples of diseases with dominant inheritance are Huntington's disease and certain types of Alzheimer's disease. Examples of recessive disorders are cystic fibrosis and Tay-Sachs disease. Diseases can also be caused by mutations in genes on the sex chromosomes. The Y chromosome harbors very few genes, thus Y-linked diseases are rare. Most of the genes on the Y chromosome are involved in male reproduction, and deleterious mutations in those genes are therefore generally not passed on to the next generation. However, should they be passed on, they are obviously inherited only in male lineages. The X chromosome contains a large number of genes and there are several examples of X-linked diseases, e.g. hemophilia and red/green color blindness. Recessive X-linked disorders typically affect male offspring, since one mutant version of the gene is enough for the hemizygous male to express the trait. For Mendelian monogenic disorders, there is usually a very strong correlation between genotype and phenotype, i.e. the penetrance is 100%. Deleterious mutations in one specific gene are normally both necessary and sufficient for the trait to be expressed. However, there are examples where a mutation in one of several genes is necessary cause disease, but a deleterious mutation in any of those genes is sufficient to cause the disease. For monogenic diseases with high penetrance and with a known family history of disease, it is often possible to determine the mode of inheritance by looking at the pedigree. Figure 1 shows examples of some typical pedigrees.

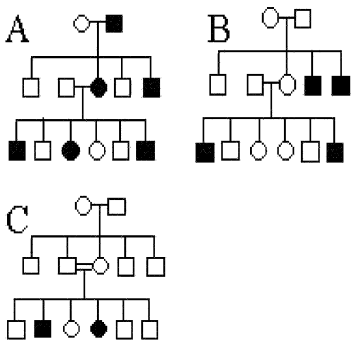

Figure 1. The figure shows three typical pedigree patterns. **A.** The disease is passed on to every generation. Both men and women get the disease. This pattern is typical for autosomal dominant inheritance. **B.** A typical pedigree for X-linked recessive disorder. Only men are affected, and the disease is transmitted by an unaffected mother. **C.** Autosomal recessive inheritance. The disease appears without prior history of disease in the family. Chances of recessive disorders are increased by inbreeding. An extreme case of incestuous inbreeding is shown in fig 1C.

There are also exceptions to Mendel's laws, where monogenic disease is inherited in a non-mendelian pattern. Three examples of mechanisms giving rise to non-mendelian segregation patterns are mitochondrial inheritance, genomic imprinting, and uniparental disomy.

1.1.2 Complex disease

Diseases that cluster in families but do not show a clear or consistent pattern of inheritance are usually referred to as complex diseases. The lack of a clear inheritance pattern indicates that a combination of several factors is needed to give rise to the phenotype. It may be a combination of several gene products, in which case the disease is called polygenic. It can also be a combined effect of genetic factors and the environment, usually referred to as multifactorial inheritance. This means that one single genetic factor is neither necessary nor sufficient to give rise to the disease phenotype. Complex diseases include many of the common disorders in society today, with millions of people affected worldwide. Examples are type II diabetes, cardiovascular disease, rheumatoid arthritis and Alzheimer's disease. However, the extent to which different diseases can be explained by genetic components varies from 0% (purely environmental) to 100% in some monogenic diseases. Most of the disorders mentioned in the context of complex disease have a genetic component of 30-95%. The genetic and environmental components of a disease are usually estimated from twin studies, comparing twins reared together with twins reared apart[7]. It is difficult to achieve an accurate measure of the genetic contribution to diseases that have a very large environmental component, such as susceptibility to infectious disease, behavioral traits etc.

A majority of the complex diseases have a higher age of onset compared to most monogenic diseases. This may be a reflection of an accumulation of environmental factors or of long-term gene interaction effects. Many of the diseases that have become widespread due to modern life style are considered complex, and include diabetes, obesity and cardiovascular disease. In the case of obesity, it was not even considered a disease until very recently. Diseases with an onset after 60 years of age are to large extent also a feature of modern society. There has been a dramatic increase in the number of elderly people in society. During the last 250 years the average life expectancy for a newborn in Sweden has increased from 35 to almost 80 years (Statistiska Centralbyrån). Along with this increase has come a great health burden due to diseases with late onset, e.g. Alzheimer's disease. There is no obvious selection against late-onset diseases, as they occur long after reproduction. Risk factors for a number of late onset diseases may be an advantage early in life[8], although these assumptions are difficult to validate. Selection for specific alleles may actually be a contributing factor to many of the disorders associated with diseases in modern society. When considering risk factors for disease and the environment in which they have increased in frequency it is important to think in terms of thousands or millions of years. The environment for most human beings has changed fundamentally during the last centuries, especially in the industrialized countries. The environment most people live in today is therefore not at all representative of the environment in which humans have evolved. Allelic variants that have risen in frequency due to positive selection may today be the very variants causing disease. One example is obesity, which is a huge health problem in modern society, but the very same alleles contributing to the genetic component of these traits are likely valuable in an environment where food is scarce.

1.2 GENETIC VARIATION

All people are different. The phenotypic differences range from subtle variation to traits that are detrimental for the health of the individual, which is called disease. Although our phenotypes do change due to environmental influence during the course of our lives, it is mainly the underlying variation within the genome that makes people look different, respond differently to medical treatment or have an increased susceptibility to common disease.

There are many types of genetic alterations. The first descriptions of genetic differences between people were large chromosomal aberrations because these can be readily identified using a microscope. The common types of chromosomal rearrangements include deletions, duplications, inversions, translocations and addition or lack of whole chromosomes. A few of these changes are totally neutral and cause no visible phenotypic change, for example certain types of balanced translocations. However, many well-known diseases such as Down's syndrome, Turner syndrome and cri du chat syndrome are due to these types of chromosomal changes.

There are common types of genetic differences between people that are less obvious than chromosomal rearrangements. Single nucleotide changes, small insertion/deletions and variation in length of repeat sequences makes up almost all genetic variation that is found in people. These small changes in the genome, most of which are neutral, are what make people different, and it is the specific combination of these differences that makes each person unique. These genetic differences are also very important for genetic studies of human health and disease. A lot of effort has been put into finding variable sites, usually called markers, in the human genome. On average, two chromosomes differ at 1/1250 base pairs (bp)[9]. Considering that there are 3.3×10^9 bases in the human genome, there are more than 2.5 million variant sites between any two chromosomes randomly chosen from the population. This may sound like a high number, yet the intraspecies diversity in primates is much larger[10]. When aligning human sequence with primate DNA, variation is approximately tenfold more common (1/100 bp) than within human sequence[11].

1.2.1 Single nucleotide polymorphism

Over 95% of all genetic variation is in the form of single nucleotide changes. The average mutation rate in the genome is ~2.5 x 10^{-8} per nucleotide site, which correlates to ~175 new mutations per generation in the diploid genome[12]. Sometimes these mutations increase in frequency in future generations and eventually become a variation commonly found in the population. A variable site for which the most common variant has a frequency of 99% or less is usually referred to as a polymorphism. Single nucleotide polymorphism (SNP), variation involving only one base in the DNA sequence, is the most common form of variation found in the human genome[13]. In principle, there can be up to four different alleles for a variant position, but almost all known SNPs have two alleles (also referred to as di- or bi-allelic). The low mutation rate per generation and the size of the human genome indicates that there is a very low probability of two mutations occurring at the same position[14]. There are four types of genetic variants possible for a single position, one transition C⇔T (A⇔G), and three transversions C⇔A (G⇔T), C⇔G (G⇔C), and T⇔A (A⇔T) (the bases of the opposite strand are shown in parenthesis). Around 2/3 of all SNPs in the genome are C⇔T transitions, probably due to the high frequency of 5-methylcytosine deamination reactions known to occur in the genome[15].

It is estimated that there are around 11 million SNPs in the human genome, corresponding to one polymorphism every 300bp when looking at the population as a whole[16]. Large efforts by industry and academia during the last few years have led to a significant increase in the number of publicly available SNPs[17-19], from a few thousand in the late 1990's to around 3 million today (dbSNP, HGVbase). In addition, there are several companies with private SNP collections that have not yet been released to the public. However, it should also be noted that a large fraction of SNPs deposited in public databases are not real, but seem to be false predictions from aligned sequences[19]. Sequencing errors, sequence misalignments, genome duplications and rare mutations are examples of sources of "false" SNPs. Unfortunately, this problem is exaggerated in coding regions[20], since many of the SNPs predicted for those regions are based on low quality EST sequences.

There is a higher frequency of variation in non-coding sequence as compared to coding DNA because of the natural selection pressure to preserve the reading frame of genes and the biological function of gene products. The actual number of non-coding sequence variants is also larger because over 95% of the human genome is non-coding. Previous studies have found an average of four coding SNPs per gene in the human genome[21,22]. Current estimates of the number of genes in the genome is ~30 000[3,23-25], and subsequently around 120 000 coding SNPs are expected to exist in the genome. There are examples of non-coding regions that are extremely conserved, and these are usually functional in some way, e.g. in transcriptional regulation or part of important structural features. Coding sequence variants can be divided into synonymous (or silent) variants that do not lead to a change of sequence at the protein level, and non-synonymous variants, which lead to substitution of one amino acid for another. Some non-synonymous variants are called conservative changes, because they lead to an exchange of one amino acid for a very similar one that is likely able to perform the same function in the protein. Around 40% of the coding variants in the genome are non-synonymous[21,22]. There is likely a strong selective pressure against non-synonymous changes. Two extreme cases of coding sequence changes are those that create a stop codon and those causing a shift in the reading frame. Both these types of variation likely lead to a non-functional copy of the gene product. There has been a lot of speculation about which types of genetic changes will be responsible for complex diseases. In monogenic diseases, the most common form of mutations is non-synonymous changes[26]. Most researchers think that this will apply also to complex disease. Indeed, the few examples where risk alleles for complex disease have been identified have generally been non-synonymous changes.

Between 80-95% of SNPs are estimated to be present in all major population groups, although reported numbers vary widely between studies[27,28]. These differences may indicate bias in SNP selection, or that different regions of the genome have mutated and recombined at different rates. One reason that such a high fraction of polymorphism is shared between populations, even though the population is large and spread throughout the world, is that human dispersal and expansion is a recent event in human evolution. Humans as a species diverged from the great apes in Africa around 5 million years ago[10]. The first hominids are thought to have left Africa around 1.7 million years ago[29]. However, the dispersal out of Africa with subsequent population expansion took place only 50-150 000 years ago, with modern humans replacing the archaic hominid populations[30]. Most of the variation found in the world today existed before the dispersal and is therefore common to all populations worldwide. Since only a subset of all people left Africa, there is greater genetic diversity on the African continent than in the rest of the world[31]. There has been little time from an evolutionary perspective for new mutations to occur and rise in frequency after humans left Africa. The increase in frequency from a single mutational event to the point where the variation is commonly found in the population can be due to a number of factors. The frequency can increase by chance alone, a phenomenon known as genetic drift. The increase can also be due to selection, when the carrier of the variant has an advantage over non-carriers. Population dynamics such as bottlenecks, where a rapid decrease in the number of individuals is followed by a population expansion, can make rare variants rise quickly in frequency, while other variants go to fixation.

1.3 LINKAGE DISEQUILIBRIUM

As gametes are formed in meioses, homologous chromosomes line up and recombination occurs between them, creating a new unique combination of genetic material that is passed on from the parent to the child. A specific combination of alleles along a chromosome is called a haplotype and new haplotypes may therefore be created by recombination. Several recombinational events may occur for each chromosome. The extent of linkage between two loci can be measured by the recombination fraction. This is called genetic distance and is measured in Morgans (M). The longer the distance between two loci, the greater the chance that recombination will occur between them. However, there is no absolute correlation between genetic and physical distance and it varies between different regions of the genome. One cM roughly corresponds to one mega-base of genetic sequence.

Alleles of neighboring loci tend to be inherited together. This leads to non-random association between alleles in a population and this association of alleles in the human genome is called linkage disequilibrium (LD). For two markers this means that a specific combination of alleles is seen more often than would be expected by chance (under the assumption of random segregation). That is, the presence of one variant provides information about the presence of other variants. Consider a scenario where two C/T polymorphisms are located next to each other along a chromosome. The frequency of the C allele is 75% at both positions, and thus the frequency of the T allele is 25% for both positions. If these markers were not at all physically linked, the combination of alleles from the two loci would be expected to be random and the frequency of each two-marker haplotype could be deduced using the frequencies of the alleles. Hence when studying the two-marker haplotype the frequency of C-C would be expected to be 0.75 x 0.75 (56.25%), the C-T and T-C to be 0.75 x 0.25 (18.25%) respectively, and the T-T to be 0.25 x 0.25 (6.25%). These markers would then be in linkage equilibrium. However, for markers situated close together along the chromosome there is often a strong deviation from the expected frequencies. In the extreme case, when there is perfect LD, the frequencies for the same two markers would be C-C at 75% and T-T at 25%, with a total lack of C-T and T-C. The existence of LD can more easily be understood by thinking about how polymorphism in the region was created (see figure 2). Each new mutation has to occur on an existing haplotype. If no recombination occurs near the new mutation, and the mutation increases in frequency, it will still be perfectly linked to the alleles around it on the same haplotype. Common variants are generally old, and therefore it is more likely that recombination has occurred between them, leading to a reduction of LD.

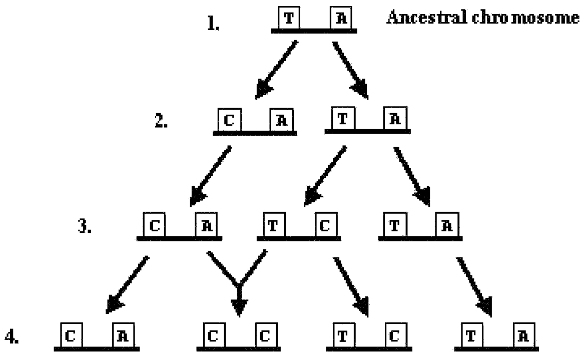

Figure 2 In stage 1 there is only the ancestral chromosome. In stage 2 a mutation occurs in the first position, creating a C/T polymorphism at that site. In the third step a mutational event creates a new polymorphic site in the second position. This change takes place on the ancestral T-A haplotype. There is then a total of three haplotypes in the population (C-A, T-C and T-A). For two polymorphic sites, measures of LD can be used to describe the association of alleles at the two loci. In this case, there is strong LD, since the C-C haplotype is missing in the population. No recombination has taken place between the two polymorphic sites. In step 4 there is a recombination between haplotype C-A and T-C, leading to the creation of the missing haplotype C-C, resulting in a reduction of LD.

Recombination and mutation are the major forces that influence the erosion and creation of LD in a population. However, several demographic factors may have an indirect influence on LD. The change in haplotype frequency that occurs due to the random transmission of gametes is referred to as genetic drift. This force is inversely correlated to population size. Observed LD is increased by genetic drift in small populations due to loss of haplotypes from the gene pool. Conversely, rapid population growth leads to a decrease in LD by a reduction in genetic drift. Another factor influencing LD is natural selection. Positive selection for one marker will affect nearby markers, which thereby stay in the population and rise in frequency. This is known as the hitchhiking effect. Negative selection can also lead to an increase in LD, as entire haplotypes carrying the deleterious alleles are lost from the population. A mechanism that still is poorly understood in the human genome is gene conversion, which denotes a non-reciprocal transfer of one stretch of DNA to another chromosome. The effect is similar to two recombination events very close to each other, and can therefore lead to a loss of LD. Recent data suggests that gene conversion may be more prevalent in the human genome than previously expected, and would then be an important factor influencing short range LD[32,33]. Many of the mechanisms involved in population dynamics thus have an effect on LD in the human genome.

1.4 ISOLATION OF HUMAN DISEASE GENES

Large efforts are required to uncover the genetic causes of a disease. It is more complicated to dissect the causes of disease when many genetic and environmental factors are involved in disease causation. The simplest case is when one specific mutation in one specific gene is sufficient and necessary to cause disease. However, this scenario is rare. Usually there is some type of heterogeneity involved in disease causation. There can either be allelic heterogeneity, in which case many different alleles in one gene can cause the disease. An example of allelic heterogeneity is seen in cystic fibrosis, where more than 1000 disease causing mutations in the large cystic fibrosis gene have been reported (Cystic Fibrosis Mutation Database). Another type of heterogeneity is genetic heterogeneity meaning that one mutation in any of several different genes can give rise to the same phenotype. A good example of this is retinitis pigmentosa, where mutations in at least 20 different genes can cause an identical phenotype[34]. Complex diseases are assumed to include a mixture of genetic and allelic heterogeneity, although very few complex diseases in humans have as yet been totally understood. Other complicating factors for complex disease dissection are interaction and epistasis. Genetic interaction, or gene-gene interaction, means that combinations of alleles from different loci interact to give rise to the phenotype, but none of the alleles are by themselves sufficient to induce this change. Epistasis originally meant that mutations in one gene mask the appearance of a trait that would otherwise be displayed due to a second gene, but the term is nowadays used to mean gene interaction in general[35].

Two of the approaches that have been successful in isolating human disease genes are called forward and reverse genetics[36]. The first approach to be developed was the forward genetics approach. This strategy is used when the causes for disease are known at the protein level. Knowledge about the protein can be used to isolate the gene. The reverse genetics approach, also referred to as positional cloning, requires no prior knowledge about the biochemical basis of the disorder. This has been a very successful approach for rare monogenic diseases. One of the first successful examples using this approach was the cloning of the cystic fibrosis gene in 1989[37]. The starting point is to isolate the chromosomal region that harbors the disease gene. The region is normally isolated using linkage analysis (see below). Disease specific chromosomal aberrations such as deletions, translocations and duplications can often yield further information about the location of the gene. Originally, extensive subcloning was required to identify the genes in the limited region and finally sequencing to identify the mutations. However, the information from the draft of the human genome sequence has simplified the subcloning process. The positional cloning approach has been very successful and there are now very few monogenic diseases left to investigate. Unfortunately, similar strategies have worked less well for complex diseases.

1.4.1 Linkage studies

Hundreds of genes causing monogenic disease have now been cloned. Many of these have been identified using linkage study approaches, which are used for the reverse genetics approach described above. The purpose of linkage studies is to see how genetic material is transmitted through a pedigree. When a linkage study is performed, markers covering the region of interest, usually repeat polymorphisms (microsatellites), are used to unravel which genetic material was passed on from parent to child. By determining from which parent the alleles of polymorphic markers originated, it is possible to gain an understanding of the recombination events that have taken place. Information from many families can reveal if there is a certain region of the genome that is transmitted together with the disease in the pedigree. A single pedigree rarely contains enough informative meioses to identify regions of interest. Therefore data from several pedigrees are added together in order to get a better statistical measure. For recessive disorders and for complex diseases, it is not as informative to use large family pedigrees as for dominant monogenic disease. It is also difficult to collect genetic material from large families, especially for late onset disorders. One common approach is therefore to collect genetic material from affected sib-pairs. Siblings share 50% of their genes on average through common genetic descent. The observed allele sharing for the markers tested is then compared to the expected. In order to delimit the size of the region even further, the region indicated can be fine-mapped by testing additional markers.

Linkage studies have worked extremely well for the monogenic diseases. There has been less success in the dissection of complex disease using this approach[38]. One reason for this may be that linkage studies are not sensitive enough to find loci that contribute to only a small fraction of the total genetic etiology. Usually several broad regions are indicated when linkage studies are performed for complex diseases, and the results rarely reach high statistical significance. However, this does not mean that linkage studies should not be used for complex disease. The few examples where susceptibility markers for complex disease have been isolated, such as Alzheimer's disease[39] and Crohn's disease[40], have usually started with a linkage study indicating a specific chromosomal region.

1.4.2 Association studies

An approach that differs from linkage studies to some extent is the association study approach. The main differences are that association studies are used for higher resolution mapping and are normally not based on family materials, but rather on large groups of unrelated patients and controls. Given the right circumstances, association studies can be more sensitive than linkage studies and provide better mapping resolution[41]. It has therefore been a popular tool in attempts to find genes involved in complex disease. For an association study, markers are chosen from the gene or region of interest. These are then genotyped in the patient and control groups and the genotype frequencies are compared between the two groups. A large difference in genotype frequency is an indication that the gene may be involved in disease. One of the major decisions to make for the design of an association study is which markers to choose for the study. If a single SNP increases the risk for disease, association may be found either directly with that marker or with other markers that are in LD with the actual risk allele. The power to detect the association is best if the actual risk allele is tested. Because of this, efforts have been made to rank SNP after how deleterious they are predicted to be for the resulting protein[42-44], and this may guide the selection of SNPs for association studies. There are three major types of association studies, each with a different study design:

1. The candidate gene approach
2. The candidate region approach
3. The whole genome approach.

The most common type of association study is still the candidate gene approach. Using this approach, genes of interest are chosen based on their biological function. This requires a prior hypothesis about the disease mechanism. Polymorphic markers are chosen from the gene of interest and genotyped in the patient and control cohorts. The approach may work well once the mechanism of the disease has been thoroughly studied, and the candidates can be limited to a few select genes. However, for many complex diseases this is not the case. Another strategy is to scan a region of interest due to a prior linkage result. For complex diseases, linkage studies usually generate a number of regions with potential interest. These regions are typically large, spanning tens of mega-bases of sequence. In the candidate region approach, markers covering either a few genes within the region or markers evenly spaced across the whole region are chosen for study. This approach has been successful in finding risk factors for complex disease. The third strategy is the whole genome approach. The idea is that a large number of markers covering the whole genome are chosen. These markers are then genotyped with the expectations that some of the markers will be in LD with existing risk alleles. So far, no serious attempts have been made using this approach (at least in academic institutions) and there is a lot of debate as to whether it will work.

It is important to remember that all an association study can ever show is an over-representation in marker frequency in the patients studied compared to the frequency of the same marker in the control population. An association does not necessarily mean that the genetic marker is involved in disease causation. There are numerous reasons why an association can be found when there is no actual involvement of the tested marker in the disease. The most obvious source of false association is that it is a chance finding, a type I error. If enough factors are studied, spurious associations will be found by chance. There may be a correlation between the number of children born in Sweden and the number of storks sighted. This does not mean that the number of storks sighted actually has something to do with the number of children born. An association result should always be considered in its context. Association findings may also be due to secondary effects. Lung cancer may be over-represented among coffee drinkers. However, this does not necessarily mean that coffee drinking leads to lung cancer. The effect could be seen because coffee drinkers are more likely to be smokers, and smoking increases the risk for lung cancer.

It is very important that the control samples are well matched with the patient group in genetic association studies. The only difference between the two groups should be the existence of a specific phenotype. If the ancestry of the control and patient materials is different, there is likely to be a difference in the frequency of alleles for markers in the genome. Such population stratification can lead to associations with markers that have nothing to do with the disease phenotype. By genotyping multiple markers it is possible to test your cases and controls for population stratification[45]. It is also possible to correct for population substructure to a certain extent[46]. The best way avoid these problems is to use family based materials. Many study designs and different types of statistical analysis have therefore been developed for samples of parent-child trios, sib-pairs etc[47]. However, the family based association study approach is more expensive (less information per genotype) and may be impractical in terms of sample collection, especially for late onset disease.

1.4.3 LD mapping considerations

There has been considerable debate over the last few years regarding the optimal strategy to use LD for large-scale association study efforts in complex disease. The questions abound: What is the extent of LD in the human genome? Are patterns of LD common to all populations? Are common alleles responsible for common disease? Is it worth the effort and enormous amounts of money to create a genome wide haplotype map? How many markers are needed for whole genome association approaches? There has been a lot of speculation, and recently some experimental data, that have addressed these issues. It is clear that LD is highly variable in different genomic regions. Although there is a general correlation between LD and physical distance on a larger scale, LD is not a strict function of physical distance. Markers within a few kb may show very weak LD[32,48-54], while other regions show strong LD over hundreds of kb[55-65]. Recombination obviously varies widely across the genome[66,67], and there seem to be regions that are "hot spots" for recombination[68-70]. These recombination hot spots are separated by extended regions of strong LD where recombination events are rare. Very little is known about the factors influencing the distribution of LD in the genome, and there is currently no way to predict the extent of LD in a region[62-64]. However, the distribution of LD is of crucial importance for genome-wide and region-wide association studies. Estimations of the level of LD in the genome have been made both through simulations[71] and more recently from large efforts of SNP genotyping[27,62,72-74]. One of the problems has been to define the level of LD that is needed for association studies with markers in a region. Expressions such as "half-length of LD" (the distance at which the average |D'| drops below 0.5)[62] and "useful LD" (no specific definition, except d^2 significantly >0.1)[71] have confused the debate.

Extended regions of strong LD are commonly referred to as haplotype blocks or LD blocks[72,73]. The haplotype diversity within a LD block is low, and 2-5 haplotypes account for a large fraction of all chromosomes[27,72,75]. The size of LD blocks has been a subject of considerable debate. There is no current consensus on the exact definition of a haplotype block. Recent large-scale investigations of whole chromosome haplotype patterns indicate that the haplotype blocks extend between a few kb up to hundreds of kb[72-74]. It has been suggested that block-like patterns of LD should be expected as a result of stochastic variations even when the recombination rate is uniform across the genome[76], and further analyses along those lines are warranted. Although the existing genetic maps correlate with LD patterns across whole chromosomes, the resolution of the genetic map is very poor. The only way to know the distribution of LD in a region is to get empirical data. It has therefore been suggested that a genome-wide haplotype map should be created[77]. If the extent of LD were known for all regions across the genome, it would be easier to decide how many markers would be needed to perform adequate association studies in a region, or indeed across the whole genome. If LD is strong in a region, a few select markers may suffice to study the whole region, while regions with little or no LD would require a higher density of markers. The markers that define the major haplotypes in a region are usually referred to as haplotype tagging markers (htSNPs)[75]. The approach sounds straight forward, but many questions remain unanswered. Will this proposed haplotype map be applicable to all populations? Recent data suggests that although there is a lot of overlap in the location of LD blocks between populations, the blocks are generally smaller, with more haplotypic diversity, in African populations[27,62,78]. Another important consideration is the level of LD that is sufficient to enable the finding of the susceptibility alleles. Even if the pattern of LD is known for a region it must be decided how strong the LD needs to be in order to identify a risk allele in the region. If the contribution to the risk of disease is small, then a further dilution of that signal (due to less than perfect LD) could make it impossible to detect. The current proposal is to create the haplotype map from common variants (>20% frequency). These common variants will define the major haplotypes of the region. Then it must be questioned what the frequency of risk alleles are likely to be. It has been hypothesized that common variants will cause common disease (the CVCD hypothesis)[79]. This would mean that the variants are old, with little or no selective pressure against them or that they have risen rapidly in frequency due to high selection pressure. If this is the case, the risk alleles are likely to be represented by the high frequency haplotypes in the region studied and then the approach may work. However, others suggest that complex disease will likely be caused by multiple rare alleles from genes showing extensive allelic heterogeneity[80,81]. If this is the case, the proposed haplotype map will not work very well for finding risk alleles, and it will generally be more difficult to find the causes for complex diseases. However, rare alleles are younger than common alleles and LD around them would therefore be expected to extend further. It may then still be possible to isolate the susceptibility alleles, although not using the proposed haplotype map. Even so, the haplotype map will be a valuable asset, because it will be very informative for population geneticists. It will give insight into the migration and expansion of human populations and the ancestry of modern humans.

So far, little experimental data is available regarding the nature of genetic variation underlying complex disease in humans. There are only a few examples where risk alleles for complex disease have been isolated and replicated in subsequent studies. Some of the classical examples are shown in table 1[40,82-85].

Table 1

Disease	Susceptibility gene	Polymorphism
Alzheimer's disease	*APOE*	1 non-synonymous
Crohn's disease	*NOD2*	2 non-synonymous, 1 frameshift
Diabetes	*PPAR-γ*(protective)	1 non-synonymous
HIV-1 infection	*CCR5* (protective)	32 bp deletion
Venous Thrombosis	*Factor V*	1 non-synonymous

These few examples indicate that one cannot generalize regarding the nature of variation and patterns of LD for complex disease susceptibility alleles, and each region has its own unique features. The *APOE-e4* allele would be difficult to find using LD mapping approaches, since LD around this variant is generally low[86,87]. Unless the actual *APOE-e4* position is used as one of the markers, the disease association is unlikely to be detected. The opposite is true for the association between Crohn's disease and the cytokine cluster on chromosome 5. LD in the region is strong over a large LD block of ~250 kb, and more than 10 markers on the same risk haplotype in that block gives equally strong association with disease[88]. In such a case the resolution of LD mapping is relatively poor (although still much better than most linkage studies for complex disease) and it is impossible to isolate the actual risk alleles using genetic evidence alone. It would then be a possibility to use isolated populations for initial mapping and populations displaying less LD for further fine mapping[31,62,89-91]. However, the theory of increased LD in isolated populations has been challenged and is currently a matter of considerable debate[60,71,92,93]. The approach was successfully used for the ACE gene, where a quantitative trait locus causing elevated plasma levels of ACE was mapped to a 13kb region in a German cohort. The region could subsequently be narrowed to 3kb in a Jamaican sample[94]. More experimental data is needed to show that the approach may work as a general strategy. Once specific risk haplotypes have been isolated within a defined LD block, it is easier to search for variants with possible functionality. Functional studies are then required to show which allele/s/ on the risk haplotype that increase susceptibility to disease.

1.4.4 Measures of LD

There is currently no consensus for which LD measure to use when describing how two markers are linked[95], or when describing a whole region[96]. The two most commonly used measures are Lewontin's |D'| and r^2 [81]. Both of these metrics describe LD between pairs of markers and can easily be calculated using the allele frequencies and the pair-wise haplotype frequency. To estimate the level of LD between two SNPs (with alleles A, a and B, b, respectively) with frequencies p_A, p_a, p_B and p_b, the first step is to calculate the LD parameter D, which is the difference of the observed and the expected values:

$D = p_{AB} - p_A * p_B$

This value is then used to calculate D' and r^2 according to

$D' = D/D_{max}$ where D_{max} is the lesser of $P_A * P_b$ or $P_a * P_B$

and

$r^2 = D^2 / p_A * p_a * p_B * p_b$

The sign of D' depends on the arbitrary labeling of alleles, and it is therefore common to use the absolute value |D'|. The main difference between the two LD measures is that D' is largely independent of allele frequency, while r^2 is not. Both metrics works in the range from 0 to 1 (where 0 means no LD, random association of alleles and 1 equals complete LD).

For r^2 to reach a value of one, allele frequencies at both markers must be identical, and only two out of the four possible allele pairs can be observed. The r^2 value is directly correlated with the χ^2 value when testing for association between the two alleles (under the assumption of no LD), and the χ^2 can be achieved by multiplying the r^2 value by the sample size[81,96]. The inherent sensitivity to differences in allele frequency also makes r^2 a measure of how good one marker is as a substitute for another. The r^2 value is inversely proportional to the sample size required to find the same association result using a substitute marker as using the actual risk allele. To achieve approximately the same power at the marker locus as is achieved by the susceptibility locus, the sample size must be increased by $1/r^2$. Due to these properties the r^2 measure has been proposed to be the LD measure of choice for LD mapping purposes.

D' is scaled to remove effects of allele frequency differences. Because of this, D' is biased when small sample sizes are used. If allele A at SNP 1 is always linked to allele B at SNP 2, then D' will be 1, indicating complete LD. However, for small sample sizes, this scenario is likely to happen by chance, and D' is therefore biased upward. The same is true when comparing low frequency markers to common markers.

There are still no good measures developed to describe the LD across a region where many markers have been investigated. One common approach is to create a map using all pair-wise LD values. If markers with low allele frequency are excluded when creating a LD map, D' and r^2 will provide a similar picture of the LD in the region. Generally, r^2 values are lower than D' values due to allele frequency differences between markers in the region studied. When defining blocks of LD it is preferable to use a map based on D'. A block is a region where recombination is rare. D' is a better measure of historical recombination between markers than is r^2. It is impossible to draw conclusions about recombination from a low r^2 value, as it may just indicate a difference in allele frequency. Neither r^2 nor D' are well suited to compare different regions of the genome. There have also been suggestions to use scaled recombination rates to describe across regions[96], and development of such metrics may be needed to accurately compare different gene regions. In conclusion, both measures of LD are valuable depending on the purpose for which they are used, but there is presently no consensus for which measure to use, or how to use the existing measures optimally.

1.5 ALZHEIMER'S DISEASE

The clinical symptoms of Alzheimer's disease were first described by a Bavarian psychiatrist, Alois Alzheimer, in 1907[97]. He described a patient, Auguste D, who suffered from several of the cardinal features of the disorder that are also used today to characterize the disease. These include progressive memory impairment, disordered cognitive function, paranoia and decline in language function. Alzheimer's disease (AD) is primarily a disease of the elderly, affecting around 5% of all people over 65 years of age. Extreme cases of the disease have been reported with an onset at around 30 years, but this is rare. After the age of 65 there is a higher incidence with increasing age and around 20% of all people over 80 years show some symptoms of the disease[98]. The average life span in Sweden has increased from around 55 years in the year 1900 to over 80 years in year 2000 (Statistiska Centralbyrån). Consequently, there has been a dramatic increase in the number of affected people during the 20th century, and this increase is projected to continue[99].

1.5.1 Clinical symptoms

Clinically, there are different scales used to characterize the different stages of AD. However, it must be emphasized that it is really a progressive, continuous deterioration. The early stages of AD cannot be differentiated from some characteristics associated with normal aging. More than half of the population over the age of 65 experiences what is referred to as normal aged forgetfulness[100]. The symptoms are quite subjective and persons with these symptoms experience difficulties in concentration, in finding words or remembering where things have been placed. The next stage is known as mild cognitive impairment (MCI)[101]. At this stage it becomes more difficult to learn new skills, executive functions become compromised and concentration deficits may be displayed. Many people with MCI do not decline further, but a majority of patients will show symptoms of dementia within four years. As cognition deteriorates further, patients are diagnosed with mild AD[100]. At this stage, patients forget recent events, have problems recalling the day of the week and which year it is. Mild AD patients can still perform most tasks in their daily life, but need help with complex routines such as finances and signing documents. The next two stages are referred to as moderate AD and moderately severe AD. Loss of memory and cognition continues. In the latter stage, patients can no longer manage basic activities of daily life on their own. Tasks such as getting dressed and taking care of personal hygiene cannot be performed independently. The last phase is called severe AD. Patients require continuous assistance. Speech becomes limited to a few intelligible words. If patients survive, deterioration continues to a point where they cannot sit up, lose the ability to smile, show increased physical rigidity and develop reflexes typical for infants such as grip and sucking reflexes. The most frequent cause of death is pneumonia[102], but patients with severe AD appear to be more susceptible to a number of common causes of death such as stroke, cancer and heart disease. A few of the most severe patients show no other causes of death than AD[102].

1.5.2 Alzheimer's disease pathology

The major neuropathological lesions found in the brains of Alzheimer's patients are neuritic plaques and neurofibrillary tangles. The first attempts to characterize the biochemical components of these lesions were generally considered a waste of time, because the lesions were thought to be the final result of the pathological process, and therefore mere tombstones of other relevant processes[103]. In light of current knowledge, the experiment was highly justified. Neuritic plaques are spherical lesions that contain extracellular deposits of amyloid-β protein (Aβ) both in a fibrillar and non-fibrillar form. The size of plaques is variable, but is usually in the range of 10-150 μm in diameter[98]. Neuritic plaques have degenerating dendrites and axons within and around the amyloid deposit. These plaques normally contain activated microglia expressing surface antigens associated with activation. There are two forms of Aβ commonly found in AD brains. The slightly longer form ending at amino acid 42, referred to as $A\beta_{42}$, is particularly prone to aggregation and is the main constituent of the plaque core[104]. The other form, called $A\beta_{40}$, is the most abundant form making up more than 90% of the secreted Aβ, and is found co-localized with $A\beta_{42}$ in the plaque[98]. A second type of plaque that lacks activated microglia has been found using antibodies against Aβ in brains from AD patients. These plaques are exclusively made up of the $A\beta_{42}$ form, and are referred to as diffuse plaques[105]. The diffuse plaques are generally more widespread in the brain, and are also found in healthy aged people lacking symptoms of AD and dementia[106]. It is therefore hypothesized that these diffuse plaques represent immature lesions that are precursors to mature amyloid plaques.

Neurofibrillary tangles are intra-neuronal fibers made up of pairs of helical 10 nm filaments. These helical filaments are composed of microtubule-associated protein tau[107]. The form of tau isolated from neurofibrillary tangles has been rendered insoluble by hyperphosphorylation, in contrast to normal tau that is found in the cytosol and is highly soluble[108]. It is still not clear whether it is one or several kinases that are responsible for the phosphorylation of tau in vivo, that leads to its dissociation from microtubules and aggregation into the helical filaments. The two types of lesions in Alzheimer's disease, amyloid plaques and neurofibrillary tangles, can occur independently of each other, but both are normally found in AD brains. Furthermore, neurofibrillary tangles are found in a variety of neurodegenerative diseases, the most common being frontotemporal dementia[109]. The extreme tangle formation in certain forms of frontotemporal dementia is caused by mutations in the tau gene[110].

Besides the two types of neuropathological lesions typical for AD, other pathological processes are also evident. There are signs of early inflammatory changes in brains of AD patients[111]. Microglia surrounding the plaques typically express a number of cell surface markers indicating an activated inflammatory response. There are also signs of an active apoptotic machinery in AD[112]. Neurons often display an activation of caspases in late stages of the disease. However, it is a matter of debate whether most neurons die from necrosis or apoptosis. Other processes evident in AD brains include an excessive generation of free radicals and oxidative injury to proteins, lipids and other macromolecules[113].

1.5.3 Genetics of Alzheimer's disease

Alzheimer's disease is a good example of the success in using genetic approaches to understand the disease etiology (using approaches referred to as "reverse genetics" above). Genes related to the disease have been cloned without prior knowledge of the biochemistry or proteins involved. Genetically there are two distinct forms of AD. One form, called familial AD (FAD), shows autosomal dominant inheritance and typically has an age of onset before the age of 65. FAD makes up only ~5% of all AD cases. The remaining 95% constitutes a complex form of the disease with no clear pattern of inheritance, but with a definite genetic component. The complex form of AD usually has an age of onset over 65 years of age and is commonly referred to as late onset AD (LOAD). The genetic component for LOAD has been estimated from twin studies to be around 75%[114].

1.5.4 Familial Alzheimer's disease

The fact that patients with Down's syndrome develop signs of classical AD neuropathology was the first clue as to where in the genome a gene for AD may be located[115]. It was therefore expected that a gene for AD would be situated on chromosome 21. When linkage studies were performed in familial AD families (FAD), conflicting results were reached. Although some families clearly showed evidence for a chromosome 21 locus, others did not. Further studies led to cloning of the amyloid precursor protein gene (*APP*) on chromosome 21[116]. At this time the amyloid component of the neuritic plaques had been identified[117] and the finding that mutations in the gene encoding amyloid caused FAD was therefore not unexpected. Mutations in

APP are however a rare cause of FAD[118] and only around 25 families that carry mutations in this gene have been identified worldwide.

When it was realized that FAD is a genetically heterogeneous disease, an intensive search for genes other than APP was started. Linkage studies soon pointed to a locus on chromosome 14[119], which ultimately led to the cloning of presenilin-1 (PSEN1)[120]. Several mutations were found in this gene, generally in families with very early onset and rapid progression. Today more than 75 different mutations have been found in *PSEN1*, and mutation in this gene is by far the most common cause for FAD[121]. Shortly after the discovery of *PSEN1* a homologous gene was found on chromosome 1[122]. The gene was then called presenilin-2 (*PSEN2*) and so far, three different mutations causing FAD have been found in *PSEN2*.

Finding of the *APP*, *PSEN1* and *PSEN2* genes in the early 90's have been very enlightening for studies regarding disease mechanisms. However, it should be noted that there are still several families where no mutations have been found in any of these three genes[123]. This suggests that additional FAD genes remain to be found. Another gene suggested to be involved in FAD is the nicastrin gene[124]. More evidence is needed for the general acceptance of these findings. After the cloning of *PSEN2* in 1995[122] the focus of the research community turned towards the dissection of LOAD.

1.5.5 Late onset Alzheimer's disease

Linkage studies performed on AD families in the late 80's and early 90's pointed to a disease susceptibility region on chromosome 19[125,126]. At the same time, experiments were performed based on knowledge about the biological role of Aβ. Proteins that would bind to immobilized Aβ peptides were isolated from cerebrospinal fluid (CSF) from AD patients. One of the proteins that were identified was the apolipoprotein-E (*APOE*)[127]. The *APOE* gene was already mapped to 19q13, and was therefore an obvious candidate in the previously identified linkage region. Two different non-synonymous polymorphisms had previously been identified in the gene[128], giving rise to a total of three different haplotypes. These were called *APOE -ε2*, *APOE -ε3* and *APOE -ε4* (*APOE-e4*). Further genetic analysis showed that the *APOE-e4* allele is over-represented in patients with LOAD compared to the general population[39,83]. This means that the *APOE-e4* allele is a risk allele for AD and not a causative mutation[129]. It is neither necessary nor sufficient to cause AD. One study estimates that *APOE* accounts for 50% of the total genetic effect[130], whereas others estimate 10-30%[131,132].

After the identification of *APOE-e4*, multiple studies were performed on candidate genes and genes encoding proteins that were co-localized with the amyloid plaques. Most of these studies were association studies in small cohorts of cases and controls. More than 15 genes have been suggested to be involved in AD due to positive associations. Unfortunately, many of these signals have not been replicated in subsequent studies. The lack of convincing results from candidate association studies led to new interest in performing linkage studies. Genome scans in LOAD sib-pairs were performed in 1999 and 2000[133,134]. One genome scan was also performed in five LOAD pedigrees with extremely high plasma Aβ-levels[135]. Linkage studies of smaller regions have also been performed for chromosome 9[130] and 12[136]. The most overwhelming result from the whole genome scans was a clear linkage peak in all studies on chromosome 10q[133,134,137,138]. However, the region under the peak was large, and different groups indicated different region of chromosome 10q[139]. An additional genome scan using age-of-onset in AD and Parkinson disease also gave a strong linkage signal on chromosome 10q[140]. These results gave rise to an intensive search for genes of interest in this region. Positive results have so far been reported for at least seven different genes on chromosome 10q[141]. Further screening and replication of these preliminary results are now needed in order to establish whether one or several of the genes actually are involved in increasing the risk for disease.

1.5.6 Theories about disease mechanisms

It has been speculated vividly through the years regarding the mechanisms leading to disease pathology. One matter of debate has been whether the production of Aâ is causative of disease, or simply a downstream side effect of disease mechanisms. There has also been speculation regarding the causes of neuronal death with apoptosis, oxidative damage and inflammation as candidates. It is now widely accepted that Aβ plaque formation is an early event in disease etiology[103]. According to the “amyloid hypothesis” AD is caused by the deposition of Aβ peptides in plaques in brain tissue[142]. The events involved in the cleavage of the amyloid precursor protein (APP) into the short peptides commonly found in amyloid plaques are not yet fully understood. APP can be cleaved in a number of positions by proteases called secretases[143]. The enzyme that gives rise to the Aβ peptide is called γ-secretase[144]. It has also been established that the presenilin proteins are required for the γ-secretase cleavage to occur, with some researchers claiming that the presenilins indeed are the γ-secretase[145]. This matter of heated debate remains to be resolved.

In the early 90's it was thought that the mere production of Aβ was a pathological event. However, more sensitive methods led to the discovery that Aβ is produced in healthy individuals throughout life[146-148]. Several lines of evidence now suggest that excess production of Aβ is enough to cause AD, proving the amyloid hypothesis to be at least partly true[142]. The mutations in *APP* as well as in the presenilins all lead to an increased production of Aβ by favoring proteolytic processing of APP by secretases[98,149]. Patients suffering from Down's syndrome, who carry an extra copy of chromosome 21 and thereby an extra copy of the *APP* gene, show AD symptoms early in life[115]. The neurofibrillary tangles are now thought to occur after the formation of the immature plaques, and are not a primary cause of disease.

It has been more complex to unravel the mechanisms of the late onset forms of AD. In LOAD patients there is no clear indication of Aβ overproduction as is seen in the early onset familial cases. Instead it is hypothesized that it is the clearance of Aβ from the brain that is less functional. The *APOE-e4* allele does not lead to an increased production of Aβ[150]. However, steady-state levels of Aβ, especially $A\beta_{40}$, is higher in animal models of AD homozygous for the *APOE-e4* allele[151]. Crossing *APP* transgenic mice with *APOE* deficient mice leads to a markedly reduced amount of cerebral Aβ deposition in the offspring[152]. Together, this indicates that it is a decrease in the removal of Aβ rather then the production of Aβ that is dysfunctional. Suggestions have now been raised that proteins involved in Aβ cleavage and Aβ clearance may contain risk alleles for the disease[153]. The list of such genes include insulin-degrading enzyme (*IDE*)[154], neprilysin (*NEP*)[155], plasmin (*PLG*)[156] and urokinase plasminogen activator (*PLAU*)[157]. Both *IDE* and *PLAU* are situated on chromosome 10q and have therefore been suggested to play a role the linkage findings in the region.

1.6 OVERVIEW OF SNP METHODOLOGY

During the last ten years there has been an increasing interest in studies of genetic variation. This has led to a great need for new and better for genotyping. The research community is moving towards studying more SNPs in larger sample sets, thus creating a constant demand for faster and cheaper genotyping methods. With the prospect of whole genome mapping in the near future, the demand for throughput is in the range of 100-fold faster and cheaper than what is currently available. Unlike sequencing, where technology development has entailed the improvement of the same basic concept, the current SNP genotyping methods have evolved in parallel from several different sources. Although principles for detection of single base changes were available over 20 years ago, involving either allele-specific oligonucleotides[158] or allele-specific restriction enzyme cleavage, it was not until the discovery of PCR[159] that genotyping of SNPs could be conducted on a larger scale. Most of the SNP scoring methods available today are dependent upon PCR amplification. There are alternatives, but those require more genomic DNA and may deplete limited amounts of valuable patient DNA. There are today four reaction principles that are commonly used for genotyping of SNPs. These are:

1. Hybridization methods
2. Primer extension methods
3. Oligonucleotide ligation methods
4. Enzymatic cleavage methods

These reaction principles are then combined with different detection methods. The most common detection methods are based on mass spectrometry, gel-based systems or different types of light emission detection. The latter category includes a wide range of systems, e.g. fluorescence, chemiluminescence, fluorescence resonance energy transfer (FRET) and fluorescence polarization.

1.6.1 Hybridization methods

The principle of hybridization for allele discrimination is based on the fact that complementary DNA molecules forming double stranded DNA will dissociate at different temperatures depending on the binding energies between the two strands. A short oligonucleotide hybridized to target DNA will have a higher melting temperature if it is fully matched to its target than if there is a one base-pair mismatch. When this approach was first used for allele discrimination it was referred to as hybridization with allele specific oligonucleotides (or ASOH)[158]. Although the early ASOH experiments were carried out by Southern blot analysis[158,160], all modern methods that uses hybridization for allele discrimination are dependent upon first performing a PCR. Detection temperature for discrimination between alleles can either be within the interval where the mismatch probe has reached its melting temperature and the fully matched probes are still hybridized (dot blots[161], Affymetrix[162]), or it can be carried out dynamically over a wide temperature range, as in Dynamic Allele Specific Hybridization (DASH)[163]. There are also attempts to use nucleotide analogs such as LNA and PNA in order to increase the discrimination between alleles[164-166]. Most of the hybridization methods use fluorescence or FRET based detection systems.

1.6.2 Primer extension methods

There are two main types of primer extension methods. One is based on allele-specific PCR amplification and the other on allele-specific nucleotide incorporation. Allele-specific PCR (also called amplification refractory mutation system) is based on the fact that extension of the primer in the PCR amplification is dependent on the 3' end of the primer being perfectly matched to the target[167]. Allele specific PCR often requires substantial optimization and although the traditional gel-based analysis now has been replaced by real-time fluorescence based detection, it is not widely used for high throughput genotyping. It can however be used for creating long-range chromosome specific amplification and therefore be valuable for molecular haplotype analysis[168].

In the allele-specific nucleotide incorporation approach an oligonucleotide probe is hybridized next to the SNP position. A nucleotide complementary to the SNP position in the target DNA is then incorporated by DNA-polymerase[169]. The primer extension product can be analyzed in a number of ways including mass spectrometry[170], detection of fluorescently labeled terminating nucleotide analogues[171] or luminometric detection of pyrophosphate[172].

1.6.3 Oligonucleotide ligation

DNA ligase is an enzyme that repairs nicks in DNA molecules with high specificity. For oligonucleotide ligation (OLA), two adjacent nucleotides are ligated only when the bases at the ligation site of each oligonucleotide are complementary to the target DNA[173]. Oligonucleotides not perfectly matched to the template at the ligation junction are not ligated. Scoring is then based on whether ligation has occurred (thus creating one longer fragment), using fluorescently labeled ligation probes or using gel-based systems. A variant of the OLA is the padlock probe method[174], in which one single molecule is circularized by ligation. The ligated circular DNA can then be amplified by rolling circle amplification[175].

1.6.4 Enzymatic cleavage

One of the most widely used methods for SNP genotyping is restriction fragment length polymorphism (RFLP), which is based on sequence specific restriction enzyme cleavage.

A restriction enzyme is chosen that cuts one allele but not the other, resulting in different lengths of digestion products. The common detection method for RFLP is to run it on an agarose gel. However, melting curve analysis works equally well[176]. Another method based on enzymatic cleavage is the Invader assay[177], which utilizes invasive cleavage by a flap endonuclease. Two probes are utilized, an Invader oligonucleotide and an allele-specific primary probe with a 5´ flap sequence. Cleavase exonuclease releases the 5´-arm only if there is a perfect complementarity at the overlap (the SNP position). The released 5'-arm is then used as an Invader probe in a secondary detection step, where it binds to a specific FRET cassette and leads to a second cleavase exonuclease reaction separating donor and quenching acceptor fluorophore, thereby giving rise to a fluorescence signal.

2 PRESENT INVESTIGATIONS

2.1 AIMS

The principal aims of the work presented in this thesis have been to explore different approaches for genetic studies of Alzheimer's disease using SNP based strategies.

Paper I

Apoptosis is one mechanism suggested to cause neuronal death in AD. In this study, a promoter variant in a gene called tumor necrosis factor receptor, super family 6 (*TNFRSF6*) encoding an apoptosis inducing death receptor was studied in relation to AD. The specific aim of the study was to test the genotyping method in an association study and at the same time investigate the role of the *TNFRSF6* gene in relation to AD.

Paper II

A robust and valid genotyping method is of utmost importance for large-scale genotyping projects. Erroneous genotyping and missed genotype calls will dramatically lower the power of the study, and may induce bias in the results. The aim of paper II was to investigate the reliability of the Dynamic Allele Specific Hybridization (DASH) method, and to formulate strict assay design rules to increase the robustness of the method.

Paper III

Several different cellular processes have been implicated in AD pathology, including apoptosis, inflammation, oxidative damage and processing of amyloid. In this large-scale project including 60 markers, a candidate pathway approach was used to perform SNP based association studies. Large studies lead to new challenges in terms of assay design and analysis of results, and the potential problems of association studies employing multiple markers were investigated and discussed.

Paper IV

Linkage studies have indicated an AD susceptibility locus on chromosome 10q. The 10q region includes the *TNFRSF6* gene, for which a previous positive association result was found. However, the positive study was small using only one marker. The aim of this project was to thoroughly explore the *TNFRSF6* region using an LD based association study approach. Haplotype tagging markers were defined in a small sample and then genotyped in a set of Swedish LOAD cases and controls. Haplotypes and single markers were tested for association with AD.

Paper V

The purpose of this study was to explore the linkage region on chromosome 10q using an SNP based LD mapping approach, with further detailed scanning in regions of potential interest.

2.2 METHOD

The method used for most of the work in this thesis is the genotyping method Dynamic Allele Specific Hybridization (DASH)[163,178]. The DASH method is used for the discrimination of SNPs and small deletions (figure 3). DASH is a PCR based method, utilizing one biotinylated primer in the PCR reaction. This makes possible the binding of the PCR product to a streptavidin coated microtiter plate. The DNA is rendered single stranded with an alkali rinse. An allele-specific probe (15-17 bp) is then hybridized over the SNP position, and a double-stranded DNA specific dye, called Sybr Green 1, is added. The dye fluoresces as long as the probe remains hybridized to the target. The plate is then heated slowly from 35°C up to 85°C while fluorescence is monitored. At a specific temperature the probe-target duplex will reach its melting point and the probe falls off, leading to a sharp drop in fluorescence. The probe falls off at different temperatures depending on whether it is fully matched to the target or if there is a one base pair mismatch present. Plotting the negative derivative of the fluorescence data will then display distinct peaks at differing temperatures for the match and the mismatch (figure 3). A heterozygous sample will give rise to a double peak. The method can be performed in any machine that facilitates heating while monitoring fluorescence, but a specific DASH machine is commercially available.

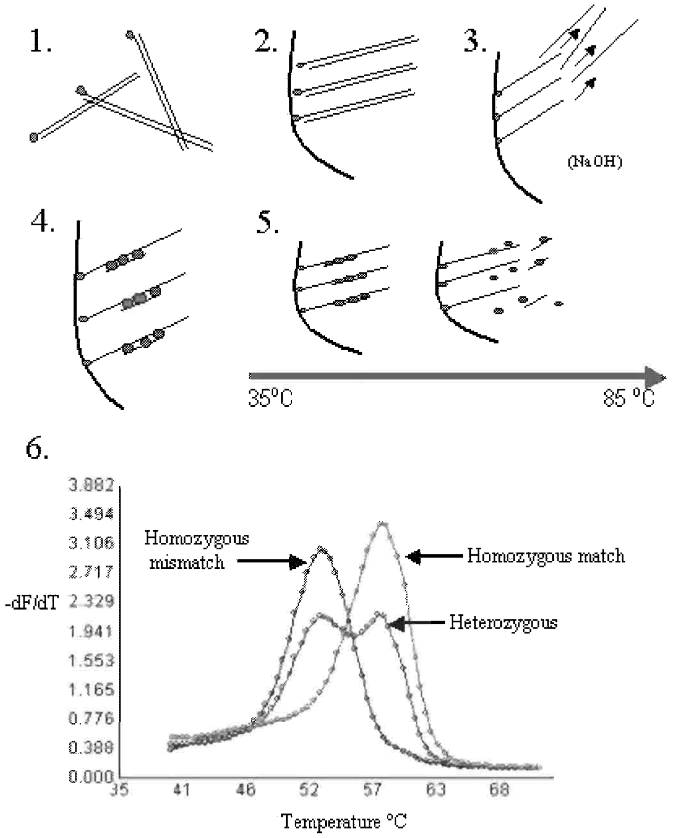

Figure 3 Overview of the DASH procedure. A PCR is run with one primer biotinylated (1). The biotinylated products are bound to a streptavidin coated well (2), and then rendered single stranded with an alkali rinse (3). An allele-specific probe is added together with a fluorescent dye specific for double-stranded DNA (4). The plate is then slowly heated while fluorescence is monitored (5). The dye fluoresces as long as the probe is bound to the target, and a drop in fluorescence is seen when the probe-target duplex reaches its melting temperature and the probe falls off. The negative derivative of the fluorescence is then plotted for the entire temperature gradient for each well (6).

Recently, a new version of the DASH method was developed, and this was used for the genotyping in paper V. The new DASH method is based on a concept utilizing fluorescence resonance energy transfer (FRET) together with an intercalating dye, and is called induced FRET (iFRET)[179]. The difference from the original DASH is that the probe carries a ROX moiety that is used as an acceptor molecule. Fluorescence energy is transferred from the intercalating dye (Sybr Green 1) to the acceptor molecule, which then emits fluorescence at a different wavelength that is monitored. When the probe-target duplex reaches its melting temperature, there will be a loss of fluorescence from the intercalating dye and subsequently a drop in acceptor fluorescence. The benefit of this new version is reduced background and increased signal strength. Only double stranded DNA that is in proximity to the probe will affect the emission from the ROX molecule, while secondary structures and other non-specific double stranded DNA formations will not be detected. The iFRET system also makes multiplexing possible by using different acceptor molecules.

2.3 RESULTS

2.3.1 Paper I

Apolipoprotein-E dependent role for the FAS receptor in early onset Alzheimer's disease: finding of a positive association for a polymorphism in the *TNFRSF6* gene

The *TNFRSF6* gene is located at chromosomal position 10q24, and it encodes the FAS receptor. FAS is a cell surface receptor involved in initiation of a downstream cascade of caspase activation, ultimately leading to cell death. Increased level of this protein has been detected in patients with AD[180,181] and Down's syndrome[182]. Apoptosis-like processes have been implicated in neuronal death in brains of AD patients[183,184], and this made *TNFRSF6* a good candidate gene for AD. Furthermore, a previously reported polymorphism situated in a potential transcription factor binding site[185,186] in the promoter of *TNFRSF6* made this gene an attractive choice for a single marker association study.

The promoter polymorphism was first tested for association with AD in a small sample of Scottish EOAD patients and controls. The results were not significant (p=0.09), but there was a trend showing an over-representation of GG-homozygotes in the AD patients. This trend was even more evident in the *APOE-e4* carriers. However, the sample size was small, especially after *APOE-e4* stratification. Another set of materials was therefore acquired from Scotland, doubling the total number of cases and controls. This sample set yielded similar, but more significant results (p=0.005). Again, the most striking difference was found in the *APOE-e4* carriers (figure 4). At this time, only one other polymorphism had been described in the *TNFRSF6* gene. This exonic synonymous polymorphism was therefore tested, but was found at very low frequency in both cases and control cohorts (<5% frequency), with no significant difference between the two groups.

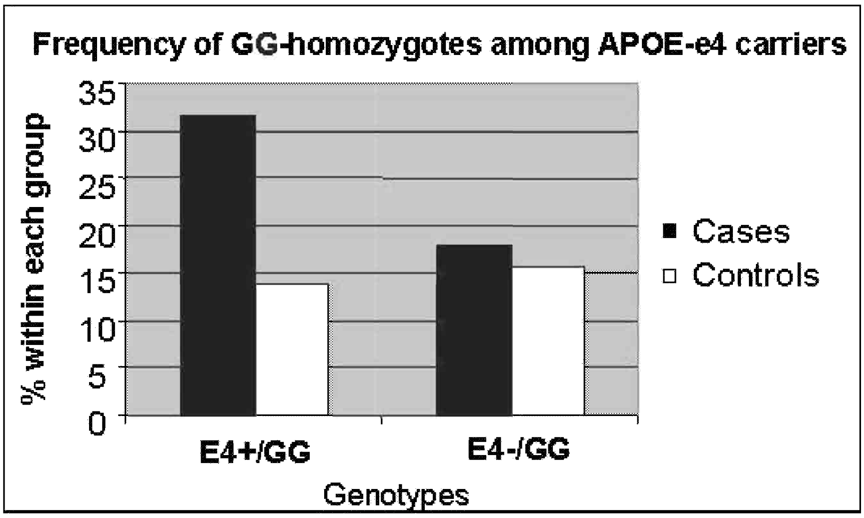

Figure 4 The graph shows the distribution of homozygotes for the G-allele in *APOE-e4* carriers and non-carriers within case and control groups. There is a significant overrepresentation of G-allele homozygotes in the patient group and that difference is largely found in the *APOE-e4* carriers (p=0.0016).

2.3.2 Paper II

Robust and accurate single nucleotide polymorphism genotyping by dynamic allele-specific hybridization (DASH): design criteria and assay validation

This investigation provides an evaluation of the DASH method. When DASH was developed we wanted to compare it to other methods, and get empirical data on reliability, reproducibility and source of errors for the method. 92 assays were therefore designed, including all types of SNPs. The SNPs were chosen from prior publications and from findings in the lab. All sequences were checked for pseudogenes and repeat sequences prior to inclusion in the study. All assays were designed in a crude way with regard only to create functional PCRs, and not with regard to the subsequent DASH experiment. Successful PCR is crucial to evaluate DASH performance. Out of the 92 assays, successful PCRs (giving one distinct band when examined on a gel) were attained for 89.

Each assay was tested upon five genomic DNAs in three steps:

1. DASH without inclusion of a probe
2. DASH using a probe matching the publicly available consensus sequence, usually corresponding to the major allele
3. DASH using a probe matching the minor allele

The first step is an estimation of the template background fluorescence of the assay. Since DASH utilizes a dye specific for double stranded DNA, it is important to measure the signal arising from double stranded DNA that is not part of the probe-target duplex. This includes non-specific probe hybridization and DNA secondary structure. The results from the remaining two steps were included to determine the success of the DASH assay with each of the allele-specific probes.

The results of the 89 assays were divided into three categories. They were scored as ideal, successful or failed. An ideal result indicates that the result was possible to score using any of the two probes independently. A successful result means that the assay was scoreable using one of the probes, or both probes together, but not both probes independently. A failed assay was not possible to score. Assays were found to be ideal for 66% of the 89 assays, successful for 23% and 11% of the assays failed. The results were then examined for differences between the three groups. The strongest factor influencing whether an assay was ideal or not, was the secondary structure of the single stranded DNA target molecule. The same result was reached both using the experimentally derived Tm of the target secondary structure and using computationally predicted ΔG values. The role of secondary structure for DASH success was validated by redesigning six of the assays that failed when using the "crude" design. The secondary structure in the single stranded biotinylated DNA product was decreased by inducing changes in the DNA sequence. This was achieved by changing a few select bases in the primers used in the PCR reaction (figure 5). All the six previously failed assays gave ideal results after redesign.

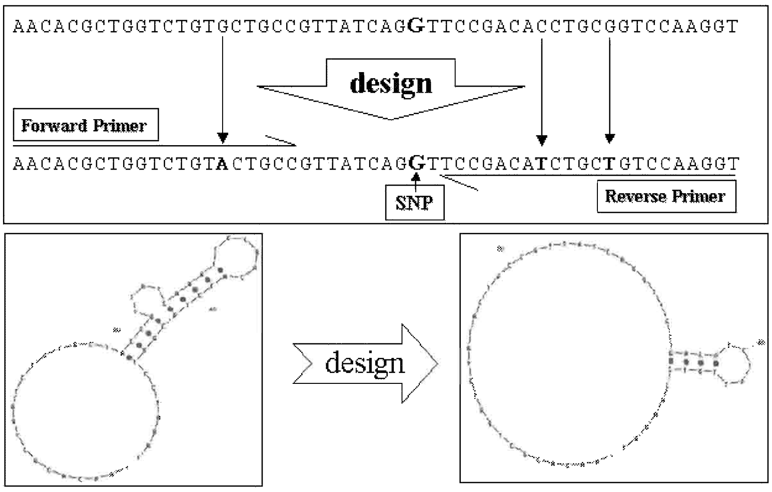
AACACGCTGGTCTGTGCTGCCGTTATCAGGTTCCGACACCTGCGGTCCAAGGT
design
Forward Primer
AACACGCTGGTCTGTACTGCCGTTATCAGGTTCCGACATCTGCTGTCCAAGGT
SNP
Reverse Primer
design

Figure 5 Overview of the DASH design strategy. Secondary structures are overcome by changing specific bases in the PCR primers. Only weak secondary structures remain in the final product and the template is accessible for probe binding.

The accuracy and reproducibility of DASH were also evaluated. The reproducibility was measured using six different DASH assays. Equal amounts of product from the same PCR reactions were bound to separate microtiter plates. DASH assays were then run by different people and scored blindly. A total of 733 genotypes were scored in duplicate, with total agreement in genotype assignment.

Accuracy of DASH was evaluated by comparing it to other genotyping methods. DASH was first compared to PCR-RFLP for two ideal assays and one successful assay. For the ideal assay there were four discrepancies in genotype assignment out of 546, all of which were attributed to PCR-RFLP (as assessed by repeating the genotyping of the discrepant samples using both methods). For the successful assay, there were four discrepancies out of 273 genotypes. Two of these errors were attributed to DASH. The reasons for these were most likely due to a contaminated sample in one case, and a weak denaturation signal in the other. A further comparison was made with the minisequencing method[171], for an assay that was in the "successful" category for DASH. Genotyping results from 381 samples were compared and 13 differences were identified. 11 of these were determined to be due to erroneous genotyping by the minisequencing method. The correct genotype for the remaining two samples could not be determined even after repetition of the experiment using both methods, possibly indicating differences in sensitivity to low levels of contamination.

In conclusion, DASH was validated to be an effective method for SNP genotyping, perhaps better than most other genotyping methods available at that time. The analysis of the crudely designed assays, many of which failed, also allowed us to identify the factors most important for successful DASH design. These results led to the creation of specific guidelines for assay design.

2.3.3 **Paper III**

SNP association studies in Alzheimer's disease highlight problems for complex disease analysis

The potential success of association studies for the dissection of human complex disease has been a matter of debate for years[187]. The association studies reported between 1998 and 2001 were generally small-scale and performed using one or a few markers in small sample sets. Many of the published studies showed positive results. However, replication attempts of these positive associations generally gave mixed results[188]. We therefore instigated a large association study effort, using 60 polymorphic markers. The genes chosen for this study can be divided into four different categories, all of which have been indicated to play a role in Alzheimer’s disease. The categories were:

1. Amyloid related genes
2. Apoptosis/inflammation
3. Oxidative stress
4. Previously claimed AD association

These were all regarded as candidate categories for AD. The prior candidate group was chosen not only to validate previous findings, but also as a comparison to the other groups. The markers were first tested in a set of Scottish EOAD cases and controls. Any marker yielding positive association was further tested in a replication set of Scottish EOAD samples.

When 60 markers are tested, a few of these would be expected to result h significant association by chance. This can be corrected for by correction for multiple testing. However, tests were also performed in stratified materials, e.g. in *APOE-e4* carriers. Tests on subgroups of the total cohort are not totally independent of the test performed on the whole case/control cohort. To correct for yet another fully independent test would therefore be too conservative. How to correctly adjust for multiple dependent and independent tests is an important issue. Another problem regarding multiple test correction is that actual positives may become false negatives. Since the increase in risk conferred by a single marker is not expected to be strong in complex disease, a correction that is too strict may hide real association signals. Most validated susceptibility variants would not survive a multiple test correction after genotyping one thousand markers even in a large set of samples, yet this is the direction in which the genotyping field is headed.

The results of our study showed a spread of p-values, most of which were not significant. Out of the markers yielding a significant association in the first sample set, only one marker gave a significant p-value also in the second set of samples (in the AGER gene). A multiple test correction was developed based on permutation. The case/control status was randomized multiple times and a test of significance was performed. The p-value of each randomization was plotted, creating a distribution of p-values. The significance value from the observed data was then compared to the distribution of p-values achieved from 10 000 iterations, and the actual p-value from the experiment was estimated based on the fraction of values exceeding the observed value. In order to correct for multiple testing, p-values were calculated for each test statistic in 10 000 replicates and the minimum p-value from each test was plotted. The proportion of replicates with a minimum p-value lower than that found in the observed data was taken as the “per marker” corrected p-value. The correction across all 60 markers was performed using the same strategy. However, not even the AGER marker (which was positively associated in both sample groups) survived multiple testing for all 60 markers included in the study. That does not mean that the association signal is false. The only way to validate whether the association is real or not by genetic means, is by replication in independent sample sets. Interestingly, none of the previously implicated AD genes gave a significant result in our study. There are several possible explanations for this:

- The original finding was done in a different population
- Most other studies have been performed in LOAD while our study used EOAD samples
- Some of the original studies had better power to detect risk alleles
- There is usually an upward bias in effect size in original positive findings compared to replication studies[189]

- The original reports may be due to false positive (Type I error) association signals
- The result in this study could be a false negative (Type II error)

It is not immediately obvious which of the above factors that play a role for the examples presented in this study. It is possibly a combination of several factors. There is a clear publication bias in the field of association studies, with a tendency for positive association results to be published, while negative results remain unreported. Therefore, a large fraction of reported disease associations are likely to be false positive signals.

Overall, the data presented in this study, and an overview of the association results in the field in general, suggested that improvements in study design and interpretation of results were needed. Single marker studies in small sample sets without a clear prior hypothesis will only provide a large collection of association signals, many of which will be false positive. A number of strategies were therefore suggested in light of these problems. Case-control samples must be increased in size and clinical definitions must be improved to give better power to the studies. Family materials should be used whenever available. Positive association signals must be replicated in independent sample sets by independent research groups. That is the only way to give further genetic evidence for disease association. Quantitative traits and sub-phenotypes should be included in analyses whenever possible. Moreover, multiple markers should be studied in each candidate gene, with construction of LD maps and estimation of haplotypes.

2.3.4 Paper IV

Further evidence for role of a promoter variant in the *TNFRSF6* gene in Alzheimer's disease

Four independent research groups have published family studies of LOAD indicating a susceptibility locus on chromosome 10q[133-135,137]. There is also a genome scan for age-at-onset in AD and Parkinson disease that gives a strong linkage signal on 10q[140]. These results have lead to an intensive search for risk alleles in the candidate regions. A problem is that the different linkage studies seem to indicate different regions. Two studies agree on a region close to the centromer[134,135,138], while other studies indicate a region about 60Mb towards the telomer[137,140] (figure 6). It is possible that this is a sign of genetic heterogeneity, and that there are at least two distinct loci that harbor risk alleles exist on 10q.

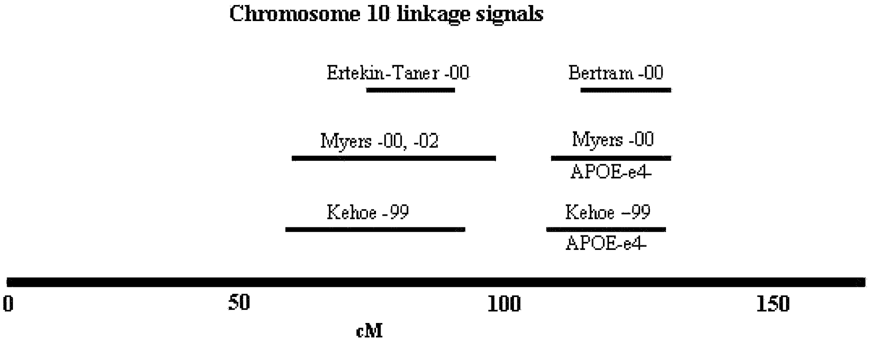

Figure 6 Overview of linkage results from genome scans indicating regions on chromosome 10 in AD. There is some overlap between the study by Kehoe -99 and Myers –00. The study by Ertekin-Taner was performed on AD families with extremely high plasma Aβ levels.

In a previous association study, we found an association with AD for a marker in the *TNFRSF6* gene. This gene is located near the telomeric linkage peak, pointing to a possible role for *TNFRSF6* or a marker in LD with the previously investigated SNP in *TNFRSF6* as part of the linkage signal. A follow-up study was therefore designed. The availability of SNPs has increased rapidly in the last few years. For the *TNFRSF6* gene where only two markers were known in 1999, there are now more than 40 SNPs in the gene available from public databases. For the present study a total of 50 SNPs were chosen from a 175kb region around the *TNFRSF6* gene. A higher density of markers was chosen in the region immediately surrounding the promoter, where the previously associated marker was located. 34 of the 50 markers were polymorphic when tested in 16 Swedish individuals. These 34 markers were genotyped in the small set of Scottish samples (121 cases and 152 controls) that was used for initial testing in the previous association study (Paper I). The genotyping results were used to create a LD map of the region. Measures of pair-wise LD were calculated for all marker pairs. The pair-wise LD map displayed two LD blocks. Haplotypes were estimated for the markers in the two blocks. As would be expected for LD blocks, there were a limited number of haplotypes representing a large fraction of all chromosomes. Since LD was strong within each block, several of the markers were redundant in their information. A limited number of markers could therefore be chosen that defined each of the major haplotypes. These markers are referred to as haplotype tagging markers or haplotype tagging SNPs (htSNPs)[75]. Three markers were enough to define all haplotypes of >5% frequency in the first block, and six markers in the second block. The second block could optimally be tagged by five markers, but the previously associated marker was also included, since that specific SNP could be of functional relevance[186]. The nine htSNPs from the region were then genotyped in a set of Swedish LOAD materials and controls.

Single marker tests were first performed for informative markers in the Scottish sample set. Two markers in high LD gave significant association with disease. Analyses were also performed using results of a cognitive test used in the clinical diagnosis of AD, called Mini Mental State Examination (MMSE), as a quantitative trait. Several of the markers gave significant results in these tests, and the lowest p-value was found for the previously reported promoter polymorphism. Even larger differences in MMSE mean values were found for each genotype in the *APOE-e4* carriers, thus correlating with the previous association results. The results of the htSNP genotypes in the Swedish samples were then analyzed. None of the markers gave a significant case/control association. The markers were then analyzed using MMSE values in the Swedish samples. Again, a positive association was found for the promoter polymorphism in the *APOE-e4* carriers. However, the p-value was higher than in the Scottish samples, and was found only after *APOE-e4* stratification. We next used estimated haplotypes from each of the LD blocks and tested for association with disease and with MMSE (the patient group was split at the median value into "high" and "low" MMSE groups). None of these tests gave any significant results. In conclusion, the *TNFRSF6* gene may play a role in AD, but not likely a major one. The results indicate that the promoter polymorphism, located in a transcription factor binding site, gives the strongest association with disease and could possibly be the susceptibility marker.

2.3.5 Paper V

Genetic variation in a haplotype block spanning *IDE, KNSL1* and *HHEX* influences Alzheimer's Disease

The strong indications of a susceptibility locus for AD on chromosome 10q (see figure 6), led to the start of a large-scale study covering numerous genes in the region. A first preliminary scan of the region included 61 SNPs from 19 different candidate genes. 26 of the 61 markers were polymorphic when tested upon 16 Swedish individuals. Those 26 markers were then genotyped in a total of 662 cases and controls comprising both EOAD and LOAD samples. Several different statistical analyses were performed for these markers in the different sample sets, utilizing case/control status as well as quantitative measures of MMSE, cerebrospinal fluid Tau protein levels (CSF-Tau), a post-mortem index of brain senile plaque and neurofibrillary tangle density (SP-NFT), and age-at-onset. In this first preliminary scan, weak association was found for several of the markers in the region as would be expected when performing multiple tests. However, two markers in the insulin-degrading enzyme gene (*IDE*) yielded significant results in several of the tests performed, with the same alleles consistently associated with case status or quantitative scores indicative of severe disease. Cases/control association was only found in the EOAD samples, and not in the LOAD materials. The results were intriguing because the *IDE* gene is one of the best candidate genes on chromosome 10q, due to its role in Aβ clearance from the brain. Instead of continuing the scan of the entire chromosomal arm, a more detailed study focused on the region around *IDE* was performed. 26 polymorphic markers were investigated in 848 clinical samples. A pair-wise LD map was created for the region. This map indicated a large LD block covering 276kb and 17 markers, spanning three different genes; *IDE*, kinesin-like 1 (*KNSL1*) and hematopoietically expressed homeobox (*HHEX*). 10 haplotypes of >1% frequency were estimated for the block region, accounting for >85% of all chromosomes. Five of the haplotypes has a frequency >5% and these could be defined using three htSNPs. Haplotypes defined by the htSNPs were then used in association tests, evaluating case/control status, CSF-Tau, MMSE and debut of AD. There were not enough patients with a SP-NFT score to use this measure for haplotype tests. Another case/control set was also acquired and genotyped for the three htSNPs. Most of the tests performed gave significant association using the haplotype approach. Case/control associations were found for three out of the four sample sets. All LOAD sets together gave a strong association between cases status and AD. Analysis of CSF-Tau, MMSE and debut also gave positive associations. Although a number of risk haplotypes could be identified, there did not seem to be one single haplotype always giving rise to the association signal. This suggests a complex picture of the underlying genetic factors, possibly due to heterogeneity in the region. A cladogram was constructed and it was clear that the haplotypes that differed most in frequency between cases and controls did not cluster together. This could indicate that many risk alleles have arisen on different haplotypic backgrounds. Unfortunately, the resolution of LD mapping strategies is limited by the size of a haplotype block. By association analyses alone it is very difficult to draw conclusions about where the risk alleles are located within the haplotype block. The genetic evidence therefore indicates that any of the three genes in the LD block may harbor the risk alleles. However, based on the biological data at hand, the *IDE* gene is by far the most convincing AD candidate gene within the block. Other small-scale SNP based association studies have been performed on *IDE* in LOAD materials, but those have not found a positive association[190,191]. These studies employed only a few markers in the region, and thorough haplotype tests were not performed. The only positive association found in the region is for a microsatellite marker located in the *KNSL1* gene, within the haplotype block[137,192]. Another group recently reported that they find association in the *IDE* region using haplotype analysis, but not with single marker tests[193]. We did not find any single marker association in LOAD samples in our study. Single marker tests yielded significant results only in EOAD. It seems that a haplotype approach may be required to detect the association signal. One single haplotype does not seem to explain the association findings. Possible heterogeneity or interaction effects must be taken into account in further analysis of the region.

3 DISCUSSION

Nearly ten years have passed since the discovery of *APOE-e4*, which is the only known risk factor for the complex form of AD that has been repeatedly replicated. Multiple association studies have been performed since then, many yielding positive results, and many of them performed on excellent biological candidate genes. Replication attempts of the suggested risk alleles have produced mixed results. In many instances both the original study and the replication attempts may be criticized for the study design. The rapid increase in the number of markers available and the development of high throughput genotyping methods during the last few years has overwhelmed the field of association studies, leading to confusion regarding study design, statistical analysis of results etc. Many studies were performed using one marker in a gene, often in small patient and control materials. Very few studies have been performed where serious attempts have been made to thoroughly investigate the proposed susceptibility loci. It is quite possible that the mixed results that have been reported from various replication attempts in some cases are due to actual risk alleles. Better studies need to be designed to get a clear picture of the possible involvement of these genes in AD.

In many ways the work described in this thesis is a reflection of how the field of SNP based studies has evolved during the last four years. Four years ago there were only a few thousand SNPs reported, many in the form of RFLPs in the scientific literature. There were no databases specifically designed to describe common human variation and for most candidate genes there were no known polymorphisms available. The methods for looking at single markers were laborious and expensive. However, reports on the possible importance of SNPs and how to find human variation by EST alignment were starting to emerge. There was clearly a need for:

- databases with genetic polymorphisms
- methods for SNP genotyping
- knowledge about study design when using SNP markers for studies of complex human disease
- improved statistics for evaluating tests with multiple markers

Reports about SNPs were at this stage limited to identification of new markers and reports on disease association, usually with single SNPs in candidate genes. Most genes did not have more than one or a few SNPs reported and genotyping of more than a few hundred samples was a very laborious task.

3.1.1 Genotyping

As would be expected when several different combinations of reaction principles and detection schemes are used in the research community, there is no single method that is better than all other. All reaction and detection principles have their own advantages and disadvantages. The three most important factors influencing the choice of method are throughput, reliability and price. Some methods are also more versatile in that they can perform analysis on several adjacent bases, be used for insertion/deletions, or have multiplexing potential.

The majority of genotypes for the studies in this thesis were scored using the DASH method. There are several reasons for this. The most obvious reason is that the method has been developed in our lab, and all changes or developments of the method could therefore rapidly be implemented. Costs of using DASH are lower than most competing methods. We have also found the method to work extremely well. One of the most time consuming steps in getting a perfect DASH result used to be assay design. This step has now been automated, making the assay design a very straightforward procedure. The development of the iFRET system for DASH genotyping has increased the reliability and signal strength by reducing background fluorescence from secondary structure.

3.1.2 The candidate gene approach

The large-scale candidate gene study in paper III was one of the largest association studies ever performed up to that point. The approach was to include candidate genes from different pathways with possible involvement in AD pathology. In light of the results of the study, we bring up a number of discussion points regarding association study design and analysis. The basic conclusion is that single marker candidate gene association studies on a single set of samples will not work very well. No matter if the association is positive or negative, it gives little information as to whether the gene is involved in the disease or not. As mentioned earlier, there have been a number of association studies performed using single markers in candidate genes[194-210] where initial positive association results have been followed by mostly negative results. See table 2 for a list of previously reported associations with AD.

Table 2

Gene Symbol	Gene Name	Location
A2M	alpha-2-macroglobulin	12p13-p12
ACE	angiotensin I converting enzyme	17q23
APBB1	amyloid beta (A4) precursor protein-binding, family B, member 1	11p15
APOE (promoter)	apolipoprotein E	19q13.2
BCHE	butyrylcholinesterase	3q26.1-q26.2
BLMH	bleomycin hydrolase	17q11.2
CTSD	cathepsin D	11p15.5
DLST	dihydrolipoamide S-succinyltransferase	14q23.1
IL1A	interleukin 1, alpha	2q13
IL1B	interleukin 1, beta	2q13
LBP-1c/CP2/LSF	transcription factor CP2	12q13
LRP1	low density lipoprotein-related protein 1	12q13
NOS3	nitric oxide synthase 3	7q35
PSEN1	presenilin 1	14q24.3
SERPINA3	serine proteinase inhibitor, clade A, member 3	14q32.1
TF	transferrin	3q21
TNFa	TNF-alpha	6p21.3

None of the markers in table 2 have been consistently replicated, and only a few are close to regions indicated by the genome scans. As should be expected, and as was shown in our study, a few positive results are found by chance when testing multiple markers. However, it does not matter if one research group tests 50 markers or if 50 groups test one marker; a few positive results will emerge. The only way to prove by genetic means whether a positive association is real or not is to replicate the result in independent materials. Results would also be more reliable if complete investigations of the candidate genes were performed instead of the "one marker per gene" approach. The reason that the single marker approach was used in our own study was mainly that very few markers were available. One could also argue that if all candidate genes were equally valid, the chance of finding a signal is maximized by using one marker in each gene, because there is unlikely to be any LD between the markers tested, and therefore no overlap in information from the markers tested.

3.1.3 The chromosome 10q linkage signals

The linkage results that indicate the chromosome 10q region have sparked new hope in the quest for AD susceptibility loci. Several of the best AD genetics groups in the world are now involved in a race to find the gene or genes giving rise to the linkage results. No convincing data has been published yet, but several preliminary results have been presented. The linkage signals are very convincing, with signal strengths equal to that found for *APOE-e4*. Unfortunately, this does not mean that the susceptibility allele or alleles will be as easy to identify. It is possible that there is more than one AD susceptibility gene on chromosome 10q alone. The linkage regions indicated by different groups do not completely overlap, and the regions are very broad. Hundreds of genes are located under the linkage peaks, making the hunt for susceptibility loci complicated. There are also a large number of predicted genes with unknown function in the region. A scenario where a single SNP is responsible for the linkage signals should probably not be expected. Most genetic diseases show some extent of allelic heterogeneity. If both genetic and allelic heterogeneity does exist in the region, the task to identify the underlying risk alleles will be complex. After some initial candidate association studies in the chromosome 10q region, we decided to choose two regions for more detailed LD mapping approaches. The first region included the *TNFRSF6* gene and the second region included the candidate gene *IDE*.

TNFRSF6

Paper I was a pilot study in our lab for how to perform a case-control associations study using the recently developed SNP scoring method (DASH). The results in the first sample set were not significant, but showed an interesting trend. The second set of samples showed a stronger signal in the same direction. It is important to point out that both of these materials were early onset AD samples, and may therefore represent a different etiology, possibly with less genetic heterogeneity, than late onset cases. Non-familial AD cases with a debut under the age of 65 are extremely rare, and very few sample sets are available in the world. Since the *TNFRSF6* gene is located on chromosome 10q, near one of the regions indicated by linkage, it was an obvious choice for a more detailed study. However, the regional scan around the *TNFRSF6* gene presented in paper IV did not show a case/control association. The patients were in this study Swedish late onset cases as compared to Scottish EOAD cases in the previous study, and may therefore in ancestry and disease etiology. An association with MMSE score was found both in the Scottish EOAD samples and in the Swedish LOAD samples. This may indicate a role for *TNFRSF6* in disease severity rather than causation. Since the gene encodes an apoptosis initiating receptor, this is a plausible scenario. However, MMSE is not an optimal quantitative trait, even after taking disease duration into account.

It is difficult to draw conclusions from these results about whether *TNFRSF6* really does play a role in AD, but they indicate that it does play a role. Although a thorough analysis of the region was performed, the results still point to the functional promoter marker. It is possible that the stronger disease association is seen in the EOAD samples because EOAD represents a less heterogeneous form of the disease. Further replications in other sample sets, preferably in EOAD cases, are needed to validate this association. Functional studies will also be required to elucidate whether the promoter marker specific marker plays a role in transcriptional regulation of the gene in neurons.

The *KNSL1*, *IDE* and *HHEX* haplotype block

The second gene chosen for detailed LD mapping was *IDE*. There were two reasons for this. First, the microsatellite marker that gave the highest LOD score in one of the previous linkage studies is situated only 30 kb from the *IDE* gene[137]. Second, in an initial screen of several genes in the 10q region, we found significant associations for two SNPs in the *IDE* gene. The *IDE* gene is also an excellent biological candidate for AD. Results show that *IDE* is involved in the clearance of extracellular Aβ[211,212]. Furthermore, *IDE* has been implicated in the degradation of the cytoplasmic product from secretase cleavage of the APP intracellular domain[213]. In light of current theories stating that the complex form of AD is due to an impaired clearance of Aβ, rather than an increased production[153], the *IDE* gene is one of the best candidate genes in the 10q region. 26 markers in the region around *IDE* were then analyzed. Single marker case/control tests were only significant in the EOAD samples, and only in *APOE-e4* non-carriers. Concerns may then be raised that there is some substructure in the Scottish materials, since similar results were found for *TNFRSF6* (although in *APOE-e4* carriers). However, since more than 60 association studies have been performed in these materials without any such indications, that is not likely to be the case. The most interesting finding in this study is the result from the haplotype analysis, giving significant results for case/control status as well as for MMSE score, CSF-Tau levels and age-at-onset within patient groups. There does not seem to be one haplotype that give rise to the association signals. Instead, there are consistent differences in haplotype frequencies between the compared groups. It seems that a haplotype approach may be required to detect the association signal. Possible heterogeneity or interaction effects must be taken into account in further analysis of the region. Clearly, this indicates that more than one marker is involved, a situation that should be expected for any disease. The haplotypes that are over-represented in patients are not closely related (separated by several mutational events), suggesting that risk alleles may have occurred on different haplotypic backgrounds. There could also be interaction effects, where polymorphisms have no effect unless they are inherited together. A rat with a diabetes phenotype, called the GK rat, has been shown to have mutations in *IDE* contributing to the phenotype[214]. Two mutations have been found in the rat *IDE* gene, and they have an effect only when they are inherited together. If a similar situation is present in humans, the risk alleles will be very difficult to isolate.

Identification of risk alleles in the haplotype block is complicated by the strong LD in the region. The LD block covers 276kb and includes three genes (*KNSL1*, *IDE* and *HHEX*), as well as a small computationally predicted gene. The resolution of genetic studies is limited in regions of strong LD, and genotyping of more markers in the region may not increase the resolution further, at least in populations of European descent. Genetically, any of the three genes is an equally good candidate. However, when taking the biology into account, *IDE* is the better candidate. There are now two possible ways to go forward. One strategy is to sequence the gene to search for polymorphisms or mutations with a possible functional role. We have now started that study. Another approach is to use functional studies to prove the involvement of the gene in pathology, and possibly find variants of the protein. Functional studies of *IDE* have been going on in different labs for several years due to its capability to cleave Aβ[154]. If *IDE* is the susceptibility gene and the disease mechanism can be elucidated, it is important for the understanding of AD. It could then (depending on the mechanism) be the first evidence that a factor involved in clearance of amyloid-β is important for disease etiology. *IDE* may also be an excellent target for preventive medicine.

3.1.4 Future perspectives

The field of complex disease genetics has evolved rapidly during the last few years. Unfortunately, there has been an extreme lack of success in isolating susceptibility factors. The complex diseases (and our genome) are obviously more complex than initially imagined. There are now excellent tools available for genetic research, both in terms of technology and in terms of information about markers, genes and reliable sequence data. The strategies have been changing accordingly. The LD mapping approaches are now becoming the strategy of choice for isolating disease genes. Another field that has been given more and more attention is population genetics. If there has been any type of selection acting on the genes that harbor risk alleles for complex disease, it will have left a footprint in the genome. This can be used to isolate regions of interest. There is also an increased interest in using various populations in the LD mapping experiments. Homogeneous populations are good because they are likely to carry the same risk alleles. However, the resolution of LD mapping studies may be better in populations of African descent. Homogeneous populations may therefore be used for initial mapping, while fine mapping may be performed on other populations. A prerequisite for this strategy to work is that the same markers confer risk to disease in different populations. There should also be an emphasis on collecting family materials whenever possible. However, for any of this to work well there is a need for large sample collections from different populations, and this can only be achieved by collaborations between research groups.

The AD research field is right now at an exciting stage. It will hopefully not be long before the susceptibility locus on chromosome 10 is isolated, with the possibility of shining new light on the mechanisms of pathology. There is also intensive research focused on the role of *APOE* and the clearance of Aβ from the brain. Finally, but perhaps most importantly, there are better treatments for AD on the way that should vastly change the situation for patients within the next few years.

1 ACKNOWLEDGEMENTS

Many people have contributed to this work in various ways. I would like to thank all my friends and colleagues at the Center for Genomics and Bioinformatics at Karolinska Institute, and my friends and former colleagues at the Medical Genetics department in Uppsala, for good times and an excellent working environment. I would especially like to mention:

Dr. Tony Brookes, for working hard for the group and the science. Thank you for challenges, excellent scientific guidance, for teaching me that science is much more than just science and for your enthusiasm and determination.

Dr. Claes Wahlestedt, for your support and for providing a positive atmosphere at the department.

Dr. Nancy Pedersen, for your generosity and kindness. It has been a valuable experience working on the twin materials.

Dr. Jonathan Prince, for friendship, for interesting scientific discussions and for good collaborative work. I still think it will be hard to prove that there is a genetic component to being hit by lightning…

The students of the Brookes' group:
Mathias Howell, for valuable discussions. We wandered off the track together many times, but it seems we found our way back in the end from Diablo, BG, U and R systems, Copéma, and I guess even from Solbacka somehow…..
Magnus Jobs, you are by far the kindest man on this planet, always supportive. I still don't know how you managed to always stay focused when Mat and I were not….except for a brief moment while playing pool, that is. I will never forget the Canyon trip.
Sarah Sawyer, the best person anyone could share an office with, for always listening (although mostly talking) and for valuable discussions about everything and anything (and still so many issues are unresolved). Good thing I'm not leaving quite yet. Thank you for reading and discussing this thesis.
David (Panda) Fredman, for extremely irrational comments that sometimes almost make sense. It's a giant lizzard! No, it's a plastic bag. You have made life more fun at the CGB.

Thank you to **Marianne Siegfried, Vivianne Daza-Saucedo, Camilla Lagerberg, Linda Strömqvist, Susanna Potter, Sandra Tammer, Daniel Rios, Gill Munns, Daniel Pedersen, Linda Berglind, Torsten Mayr, Hagit Katzov, Anna Lindberg** and **Anna Hammarsjö** for your friendship, help and hard work.

I also want to thank all the nice people who have left the group the last few years: **Dr. Tesfai Emahazion, Anna Beskow, Astrid Lindstedt, Dr. Harvest Gu, Per Ewing, Nina Juntilla, Satu Valtonen, Alex Kovacs** and **Isabelle Dussauge.**

Shane McCarthy for collaborative work, journal clubs and for always showing good spirit.
Dr. Salim Mottagui-Tabar for keeping me in good shape and never failing the quest for the perfect badminton shot.
Mårten Jansson, for working with me on the box called "Richard" and for valuable discussions while aliquoting samples.
Veronica Magnusson, for friendship and kindness the last 5 years.
Cecilia Johansson, for always being a great friend, for your taste in movies and food. It's been nice to have someone to talk to who is in the field, but at another university.

My family in England; **Dr Tore, Ulla and Karin**, you are always encouraging and positive.

My family in Sweden; **Kerstin, Martin and Jezica**, thank you for all your support and for always believing in me.

Linda, my true love, my best friend, thank you for everything.

The work in this thesis was supported by the National Network in Neuroscience and the Swedish Foundation for Strategic Research.

5 REFERENCES

1. Watson, J.D. & Crick, F.H.C. A structure for deoxyribose nucleic acid. *Nature* **171**, 737-738 (1953).
2. Collins, F.S. et al. New goals for the U.S. Human Genome Project: 1998-2003. *Science* **282**, 682-9. (1998).
3. Lander, E.S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921. (2001).
4. Sturtevant, A.H. *A history of Genetics*, (Cold Spring Harbor Laboratory Press, 1965).
5. Perez-Iratxeta, C., Bork, P. & Andrade, M.A. Association of genes to genetically inherited diseases using data mining. *Nat Genet* **31**, 316-9. (2002).
6. Hamosh, A. et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **30**, 52-5. (2002).
7. MacGregor, A.J., Snieder, H., Schork, N.J. & Spector, T.D. Twins. Novel uses to study complex traits and genetic diseases. *Trends Genet* **16**, 131-4. (2000).
8. Toupance, B., Godelle, B., Gouyon, P.H. & Schachter, F. A model for antagonistic pleiotropic gene action for mortality and advanced age. *Am J Hum Genet* **62**, 1525-34. (1998).
9. Reich, D.E. et al. Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat Genet* **32**, 135-42. (2002).
10. Chen, F.C. & Li, W.H. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am J Hum Genet* **68**, 444-56. (2001).
11. Ebersberger, I., Metzler, D., Schwarz, C. & Paabo, S. Genomewide comparison of DNA sequences between humans and chimpanzees. *Am J Hum Genet* **70**, 1490-7. (2002).
12. Nachman, M.W. & Crowell, S.L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297-304. (2000).
13. Brookes, A.J. The essence of SNPs. *Gene* **234**, 177-86. (1999).
14. Nielsen, R. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* **154**, 931-42. (2000).
15. Holliday, R. & Grigg, G.W. DNA methylation and mutation. *Mutat Res* **285**, 61-7. (1993).
16. Kruglyak, L. & Nickerson, D.A. Variation is the spice of life. *Nat Genet* **27**, 234-6. (2001).
17. Altshuler, D. et al. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**, 513-6. (2000).
18. Mullikin, J.C. et al. An SNP map of human chromosome 22. *Nature* **407**, 516-20. (2000).
19. Sachidanandam, R. et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928-33. (2001).
20. Small, K.M., Seman, C.A., Castator, A., Brown, K.M. & Liggett, S.B. False positive non-synonymous polymorphisms of G-protein coupled receptor genes. *FEBS Lett* **516**, 253-6. (2002).
21. Cargill, M. et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* **22**, 231-8. (1999).

22. Halushka, M.K. et al. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat Genet* **22**, 239-47. (1999).

23. Ewing, B. & Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat Genet* **25**, 232-4. (2000).

24. Roest Crollius, H. et al. Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nat Genet* **25**, 235-8. (2000).

25. Das, M., Burge, C.B., Park, E., Colinas, J. & Pelletier, J. Assessment of the total number of human transcription units. *Genomics* **77**, 71-8. (2001).

26. Krawczak, M. et al. Human gene mutation database-a biomedical information and research resource. *Hum Mutat* **15**, 45-51 (2000).

27. Gabriel, S.B. et al. The structure of haplotype blocks in the human genome. *Science* **296**, 2225-9. (2002).

28. Goddard, K.A., Hopkins, P.J., Hall, J.M. & Witte, J.S. Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *Am J Hum Genet* **66**, 216-34. (2000).

29. Gabunia, L. et al. Earliest Pleistocene hominid cranial remains from Dmanisi, Republic of Georgia: taxonomy, geological setting, and age. *Science* **288**, 1019-25. (2000).

30. Templeton, A. Out of Africa again and again. *Nature* **416**, 45-51. (2002).

31. Jorde, L.B., Watkins, W.S. & Bamshad, M.J. Population genomics: a bridge from evolutionary history to genetic medicine. *Hum Mol Genet* **10**, 2199-207. (2001).

32. Ardlie, K. et al. Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. *Am J Hum Genet* **69**, 582-9. (2001).

33. Frisse, L. et al. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Am J Hum Genet* **69**, 831-43. (2001).

34. Farrar, G.J., Kenna, P.F. & Humphries, P. On the genetics of retinitis pigmentosa and on mutation-independent approaches to therapeutic intervention. *Embo J* **21**, 857-64. (2002).

35. Cordell, H.J. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet* **11**, 2463-8. (2002).

36. Brookes, A.J. Rethinking genetic strategies to study complex diseases. *Trends Mol Med* **7**, 512-6. (2001).

37. Kerem, B. et al. Identification of the cystic fibrosis gene: genetic analysis. *Science* **245**, 1073-80. (1989).

38. Altmuller, J., Palmer, L.J., Fischer, G., Scherb, H. & Wjst, M. Genomewide scans of complex human diseases: true linkage is hard to find. *Am J Hum Genet* **69**, 936-50. (2001).

39. Saunders, A.M. et al. Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. *Neurology* **43**, 1467-72. (1993).

40. Hugot, J.P. et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599-603. (2001).

41. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516-7. (1996).

42. Chasman, D. & Adams, R.M. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J Mol Biol* **307**, 683-706. (2001).

43. Ng, P.C. & Henikoff, S. Accounting for human polymorphisms predicted to affect protein function. *Genome Res* **12**, 436-46. (2002).

44. Sunyaev, S. et al. Prediction of deleterious human alleles. *Hum Mol Genet* **10**, 591-7. (2001).

45. Pritchard, J.K. & Rosenberg, N.A. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* **65**, 220-8. (1999).

46. Reich, D.E. & Goldstein, D.B. Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol* **20**, 4-16. (2001).

47. Schulze, T.G. & McMahon, F.J. Genetic association mapping at the crossroads: which test and why? Overview and practical guidelines. *Am J Med Genet* **114**, 1-11. (2002).

48. Clark, A.G. et al. Haplotype structure and population genetic inferences from nucleotide- sequence variation in human lipoprotein lipase. *Am J Hum Genet* **63**, 595-612. (1998).

49. Rieder, M.J., Taylor, S.L., Clark, A.G. & Nickerson, D.A. Sequence variation in the human angiotensin converting enzyme. *Nat Genet* **22**, 59-62. (1999).

50. Moffatt, M.F., Traherne, J.A., Abecasis, G.R. & Cookson, W.O. Single nucleotide polymorphism and linkage disequilibrium within the TCR alpha/delta locus. *Hum Mol Genet* **9**, 1011-9. (2000).

51. Dunning, A.M. et al. The extent of linkage disequilibrium in four populations with distinct demographic histories. *Am J Hum Genet* **67**, 1544-54. (2000).

52. Fullerton, S.M. et al. Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. *Am J Hum Genet* **67**, 881-900. (2000).

53. Kidd, J.R. et al. Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, PAH, in a global representation of populations. *Am J Hum Genet* **66**, 1882-99. (2000).

54. Nakajima, T. et al. Nucleotide diversity and haplotype structure of the human angiotensinogen gene in two populations. *Am J Hum Genet* **70**, 108-23. (2002).

55. Laan, M. & Paabo, S. Demographic history and linkage disequilibrium in human populations. *Nat Genet* **17**, 435-8. (1997).

56. Huttley, G.A., Smith, M.W., Carrington, M. & O'Brien, S.J. A scan for linkage disequilibrium across the human genome. *Genetics* **152**, 1711-22. (1999).

57. Collins, A., Lonjou, C. & Morton, N.E. Genetic epidemiology of single-nucleotide polymorphisms. *Proc Natl Acad Sci U S A* **96**, 15173-7. (1999).

58. Bonnen, P.E. et al. Haplotypes at ATM identify coding-sequence variation and indicate a region of extensive linkage disequilibrium. *Am J Hum Genet* **67**, 1437-51. (2000).

59. Gordon, D., Simonic, I. & Ott, J. Significant evidence for linkage disequilibrium over a 5-cM region among Afrikaners. *Genomics* **66**, 87-92. (2000).

60. Taillon-Miller, P. et al. Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nat Genet* **25**, 324-8. (2000).

61. Service, S.K., Ophoff, R.A. & Freimer, N.B. The genome-wide distribution of background linkage disequilibrium in a population isolate. *Hum Mol Genet* **10**, 545-51. (2001).

62. Reich, D.E. et al. Linkage disequilibrium in the human genome. *Nature* **411**, 199-204. (2001).

63. Stephens, J.C. et al. Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293**, 489-93. (2001).

64. Abecasis, G.R. et al. Extent and distribution of linkage disequilibrium in three genomic regions. *Am J Hum Genet* **68**, 191-197. (2001).

65. Mohlke, K.L. et al. Linkage disequilibrium between microsatellite markers extends beyond 1 cM on chromosome 20 in Finns. *Genome Res* **11**, 1221-6. (2001).

66. Payseur, B.A. & Nachman, M.W. Microsatellite variation and recombination rate in the human genome. *Genetics* **156**, 1285-98. (2000).

67. Yu, A. et al. Comparison of human genetic and sequence-based physical maps. *Nature* **409**, 951-3. (2001).

68. Templeton, A.R. et al. Recombinational and mutational hotspots within the human lipoprotein lipase gene. *Am J Hum Genet* **66**, 69-83. (2000).

69. Goldstein, D.B. Islands of linkage disequilibrium. *Nat Genet* **29**, 109-11. (2001).

70. Jeffreys, A.J., Kauppi, L. & Neumann, R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet* **29**, 217-22. (2001).

71. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet* **22**, 139-44. (1999).

72. Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. & Lander, E.S. High-resolution haplotype structure in the human genome. *Nat Genet* **29**, 229-32. (2001).

73. Patil, N. et al. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719-23. (2001).

74. Dawson, E. et al. A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **418**, 544-8. (2002).

75. Johnson, G.C. et al. Haplotype tagging for the identification of common disease genes. *Nat Genet* **29**, 233-7. (2001).

76. Subrahmanyan, L., Eberle, M.A., Clark, A.G., Kruglyak, L. & Nickerson, D.A. Sequence variation and linkage disequilibrium in the human T-cell receptor beta (TCRB) locus. *Am J Hum Genet* **69**, 381-95. (2001).

77. Couzin, J. Genomics. New mapping project splits the community. *Science* **296**, 1391-3. (2002).

78. Tishkoff, S.A. et al. A global haplotype analysis of the myotonic dystrophy locus: implications for the evolution of modern humans and for the origin of myotonic dystrophy mutations. *Am J Hum Genet* **62**, 1389-402. (1998).

79. Reich, D.E. & Lander, E.S. On the allelic spectrum of human disease. *Trends Genet* **17**, 502-10. (2001).

80. Pritchard, J.K. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* **69**, 124-37. (2001).

81. Weiss, K.M. & Clark, A.G. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet* **18**, 19-24. (2002).

82. Dean, M. et al. Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE Study. *Science* **273**, 1856-62. (1996).

83. Strittmatter, W.J. et al. 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**, 1977-81. (1993).

84. Altshuler, D. et al. The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat Genet* **26**, 76-80. (2000).

85. Ogura, Y. et al. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411**, 603-6. (2001).

86. Martin, E.R. et al. SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease. *Am J Hum Genet* **67**, 383-94. (2000).

87. Nickerson, D.A. et al. Sequence diversity and large-scale typing of SNPs in the human apolipoprotein E gene. *Genome Res* **10**, 1532-45. (2000).

88. Rioux, J.D. et al. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nat Genet* **29**, 223-8. (2001).

89. Wright, A.F., Carothers, A.D. & Pirastu, M. Population choice in mapping genes for complex diseases. *Nat Genet* **23**, 397-404. (1999).

90. Shifman, S. & Darvasi, A. The value of isolated populations. *Nat Genet* **28**, 309-10. (2001).

91. Kaessmann, H. et al. Extensive linkage disequilibrium in small human populations in Eurasia. *Am J Hum Genet* **70**, 673-85. (2002).

92. Jorde, L.B., Watkins, W.S., Kere, J., Nyman, D. & Eriksson, A.W. Gene mapping in isolated populations: new roles for old friends? *Hum Hered* **50**, 57-65. (2000).

93. Eaves, I.A. et al. The genetically isolated populations of Finland and sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. *Nat Genet* **25**, 320-3. (2000).

94. Zhu, X. et al. Localization of a small genomic region associated with elevated ACE. *Am J Hum Genet* **67**, 1144-53. (2000).

95. Devlin, B. & Risch, N. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311-22. (1995).

96. Pritchard, J.K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am J Hum Genet* **69**, 1-14. (2001).

97. Alzheimer, A. *Allgemeine Zeitschrift für Psychiatrie*, 146-148 (1907).

98. Selkoe, D.J. Alzheimer's disease: genes, proteins, and therapy. *Physiol Rev* **81**, 741-66. (2001).

99. Brookmeyer, R. & Gray, S. Methods for projecting the incidence and prevalence of chronic diseases in aging populations: application to Alzheimer's disease. *Stat Med* **19**, 1481-93. (2000).

100. Leon, M.d. *An atlas of Alzheimer's disease*, (The Parthenon Publishing Group Inc., 1999).

101. Petersen, R.C. et al. Mild cognitive impairment: clinical characterization and outcome. *Arch Neurol* **56**, 303-8. (1999).

102. Keene, J., Hope, T., Fairburn, C.G. & Jacoby, R. Death and dementia. *Int J Geriatr Psychiatry* **16**, 969-74. (2001).

103. Selkoe, D.J. Translating cell biology into therapeutic advances in Alzheimer's disease. *Nature* **399**, A23-31. (1999).

104. Sisodia, S.S. & St George-Hyslop, P.H. gamma-Secretase, Notch, Abeta and Alzheimer's disease: where do the presenilins fit in? *Nat Rev Neurosci* **3**, 281-90. (2002).

105. Joachim, C.L., Morris, J.H. & Selkoe, D.J. Diffuse senile plaques occur commonly in the cerebellum in Alzheimer's disease. *Am J Pathol* **135**, 309-19. (1989).

106. Tagliavini, F., Giaccone, G., Frangione, B. & Bugiani, O. Preamyloid deposits in the cerebral cortex of patients with Alzheimer's disease and nondemented individuals. *Neurosci Lett* **93**, 191-6. (1988).

107. Wood, J.G., Mirra, S.S., Pollock, N.J. & Binder, L.I. Neurofibrillary tangles of Alzheimer disease share antigenic determinants with the axonal microtubule-associated protein tau (tau). *Proc Natl Acad Sci U S A* **83**, 4040-3. (1986).

108. Grundke-Iqbal, I. et al. Abnormal phosphorylation of the microtubule-associated protein tau (tau) in Alzheimer cytoskeletal pathology. *Proc Natl Acad Sci U S A* **83**, 4913-7. (1986).

109. Arnold, S.E., Han, L.Y., Clark, C.M., Grossman, M. & Trojanowski, J.Q. Quantitative neurohistological features of frontotemporal degeneration. *Neurobiol Aging* **21**, 913-9. (2000).

110. Hutton, M. et al. Association of missense and 5'-splice-site mutations in tau with the inherited dementia FTDP-17. *Nature* **393**, 702-5. (1998).

111. McGeer, P.L. & McGeer, E.G. Inflammation, autotoxicity and Alzheimer disease. *Neurobiol Aging* **22**, 799-809. (2001).

112. Yuan, J. & Yankner, B.A. Apoptosis in the nervous system. *Nature* **407**, 802-9. (2000).

113. Butterfield, D.A., Drake, J., Pocernich, C. & Castegna, A. Evidence of oxidative damage in Alzheimer's disease brain: central role for amyloid beta-peptide. *Trends Mol Med* **7**, 548-54. (2001).

114. Gatz, M. et al. Heritability for Alzheimer's disease: the study of dementia in Swedish twins. *J Gerontol A Biol Sci Med Sci* **52**, M117-25. (1997).

115. Olson, M.I. & Shaw, C.M. Presenile dementia and Alzheimer's disease in mongolism. *Brain* **92**, 147-56. (1969).

116. Goate, A. et al. Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature* **349**, 704-6. (1991).

117. Glenner, G.G. & Wong, C.W. Alzheimer's disease: initial report of the purification and characterization of a novel cerebrovascular amyloid protein. *Biochem Biophys Res Commun* **120**, 885-90. (1984).

118. Tanzi, R. et al. Genetic heterogeneity of gene defects responsible for familial Alzheimer disease. *Genetica* **91**, 255-63. (1993).

119. Schellenberg, G.D. et al. Genetic linkage evidence for a familial Alzheimer's disease locus on chromosome 14. *Science* **258**, 668-71. (1992).

120. Sherrington, R. et al. Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature* **375**, 754-60. (1995).

121. Hardy, J. The Alzheimer family of diseases: many etiologies, one pathogenesis? *Proc Natl Acad Sci U S A* **94**, 2095-7. (1997).

122. Levy-Lahad, E. et al. A familial Alzheimer's disease locus on chromosome 1. *Science* **269**, 970-3. (1995).

123. Campion, D. et al. Early-onset autosomal dominant Alzheimer disease: prevalence, genetic heterogeneity, and mutation spectrum. *Am J Hum Genet* **65**, 664-70. (1999).

124. Dermaut, B. et al. The gene encoding nicastrin, a major gamma-secretase component, modifies risk for familial early-onset Alzheimer disease in a Dutch population-based sample. *Am J Hum Genet* **70**, 1568-74. (2002).

125. Pericak-Vance, M.A. et al. Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage. *Am J Hum Genet* **48**, 1034-50. (1991).

126. van Duijn, C.M. et al. A population-based study of familial Alzheimer disease: linkage to chromosomes 14, 19, and 21. *Am J Hum Genet* **55**, 714-27. (1994).

127. Wisniewski, T., Golabek, A., Matsubara, E., Ghiso, J. & Frangione, B. Apolipoprotein E: binding to soluble Alzheimer's beta-amyloid. *Biochem Biophys Res Commun* **192**, 359-65. (1993).

128. Weisgraber, K.H., Rall, S.C., Jr. & Mahley, R.W. Human E apoprotein heterogeneity. Cysteine-arginine interchanges in the amino acid sequence of the apo-E isoforms. *J Biol Chem* **256**, 9077-83. (1981).

129. Corder, E.H. et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261**, 921-3. (1993).

130. Pericak-Vance, M.A. et al. Identification of novel genes in late-onset Alzheimer's disease. *Exp Gerontol* **35**, 1343-52. (2000).

131. Warwick Daw, E. et al. The number of trait loci in late-onset Alzheimer disease. *Am J Hum Genet* **66**, 196-204. (2000).

132. Slooter, A.J. et al. Risk estimates of dementia by apolipoprotein E genotypes from a population-based incidence study: the Rotterdam Study. *Arch Neurol* **55**, 964-8. (1998).

133. Kehoe, P. et al. A full genome scan for late onset Alzheimer's disease. *Hum Mol Genet* **8**, 237-45. (1999).

134. Myers, A. et al. Susceptibility locus for Alzheimer's disease on chromosome 10. *Science* **290**, 2304-5. (2000).

135. Ertekin-Taner, N. et al. Linkage of plasma Abeta42 to a quantitative locus on chromosome 10 in late-onset Alzheimer's disease pedigrees. *Science* **290**, 2303-4. (2000).

136. Pericak-Vance, M.A. et al. Complete genomic screen in late-onset familial Alzheimer disease. Evidence for a new locus on chromosome 12. *Jama* **278**, 1237-41. (1997).

137. Bertram, L. et al. Evidence for genetic linkage of Alzheimer's disease to chromosome 10q. *Science* **290**, 2302-3. (2000).

138. Myers, A. et al. Full genome screen for Alzheimer disease: Stage II analysis. *Am J Med Genet* **114**, 235-44. (2002).

139. Lendon, C. & Craddock, N. Susceptibility gene(s) for Alzheimer's disease on chromosome 10. *Trends Neurosci* **24**, 557-9. (2001).

140. Li, Y.J. et al. Age at onset in two common neurodegenerative diseases is genetically controlled. *Am J Hum Genet* **70**, 985-93. (2002).

141. Abstracts from the 8th International Conference on Alzheimer's Disease and Related Disorders. July 20-25, 2002. Stockholm, Sweden. *Neurobiol Aging* **23**, S1-606. (2002).

142. Hardy, J. & Selkoe, D.J. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. *Science* **297**, 353-6. (2002).

143. Walter, J., Kaether, C., Steiner, H. & Haass, C. The cell biology of Alzheimer's disease: uncovering the secrets of secretases. *Curr Opin Neurobiol* **11**, 585-90. (2001).

144. Steiner, H. & Haass, C. Intramembrane proteolysis by presenilins. *Nat Rev Mol Cell Biol* **1**, 217-24. (2000).

145. Wolfe, M.S. et al. Two transmembrane aspartates in presenilin-1 required for presenilin endoproteolysis and gamma-secretase activity. *Nature* **398**, 513-7. (1999).

146. Haass, C. et al. Amyloid beta-peptide is produced by cultured cells during normal metabolism. *Nature* **359**, 322-5. (1992).

147. Seubert, P. et al. Isolation and quantification of soluble Alzheimer's beta-peptide from biological fluids. *Nature* **359**, 325-7. (1992).

148. Shoji, M. et al. Production of the Alzheimer amyloid beta protein by normal proteolytic processing. *Science* **258**, 126-9. (1992).

149. Scheuner, D. et al. Secreted amyloid beta-protein similar to that in the senile plaques of Alzheimer's disease is increased in vivo by the presenilin 1 and 2 and APP mutations linked to familial Alzheimer's disease. *Nat Med* **2**, 864-70. (1996).

150. Biere, A.L. et al. Co-expression of beta-amyloid precursor protein (betaAPP) and apolipoprotein E in cell culture: analysis of betaAPP processing. *Neurobiol Dis* **2**, 177-87. (1995).

151. Gearing, M., Mori, H. & Mirra, S.S. Abeta-peptide length and apolipoprotein E genotype in Alzheimer's disease. *Ann Neurol* **39**, 395-9. (1996).

152. Bales, K.R. et al. Lack of apolipoprotein E dramatically reduces amyloid beta-peptide deposition. *Nat Genet* **17**, 263-4. (1997).

153. Selkoe, D.J. Clearing the brain's amyloid cobwebs. *Neuron* **32**, 177-80. (2001).

154. Kurochkin, I.V. & Goto, S. Alzheimer's beta-amyloid peptide specifically interacts with and is degraded by insulin degrading enzyme. *FEBS Lett* **345**, 33-7. (1994).

155. Howell, S., Nalbantoglu, J. & Crine, P. Neutral endopeptidase can hydrolyze beta-amyloid(1-40) but shows no effect on beta-amyloid precursor protein metabolism. *Peptides* **16**, 647-52. (1995).

156. Exley, C. & Korchazhkina, O.V. Plasmin cleaves Abeta42 in vitro and prevents its aggregation into beta-pleated sheet structures. *Neuroreport* **12**, 2967-70. (2001).

157. Tucker, H.M., Kihiko-Ehmann, M. & Estus, S. Urokinase-type plasminogen activator inhibits amyloid-beta neurotoxicity and fibrillogenesis via plasminogen. *J Neurosci Res* **70**, 249-55. (2002).

158. Wallace, R.B. et al. Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch. *Nucleic Acids Res* **6**, 3543-57. (1979).

159. Mullis, K.B. & Faloona, F.A. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol* **155**, 335-50 (1987).

160. Conner, B.J. et al. Detection of sickle cell beta S-globin allele by hybridization with synthetic oligonucleotides. *Proc Natl Acad Sci U S A* **80**, 278-82. (1983).

161. Saiki, R.K. et al. Diagnosis of sickle cell anemia and beta-thalassemia with enzymatically amplified DNA and nonradioactive allele-specific oligonucleotide probes. *N Engl J Med* **319**, 537-41. (1988).

162. Chee, M. et al. Accessing genetic information with high-density DNA arrays. *Science* **274**, 610-4. (1996).

163. Howell, W.M., Jobs, M., Gyllensten, U. & Brookes, A.J. Dynamic allele-specific hybridization. A new method for scoring single nucleotide polymorphisms. *Nat Biotechnol* **17**, 87-8. (1999).

164. Griffin, T.J., Tang, W. & Smith, L.M. Genetic analysis by peptide nucleic acid affinity MALDI-TOF mass spectrometry. *Nat Biotechnol* **15**, 1368-72. (1997).

165. Ross, P.L., Lee, K. & Belgrader, P. Discrimination of single-nucleotide polymorphisms in human DNA using peptide nucleic acid probes detected by MALDI-TOF mass spectrometry. *Anal Chem* **69**, 4197-202. (1997).

166. Orum, H., Jakobsen, M.H., Koch, T., Vuust, J. & Borre, M.B. Detection of the factor V Leiden mutation by direct allele-specific hybridization of PCR amplicons to photoimmobilized locked nucleic acids. *Clin Chem* **45**, 1898-905. (1999).

167. Newton, C.R. et al. Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). *Nucleic Acids Res* **17**, 2503-16. (1989).

168. Michalatos-Beloin, S., Tishkoff, S.A., Bentley, K.L., Kidd, K.K. & Ruano, G. Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR. *Nucleic Acids Res* **24**, 4841-3. (1996).

169. Syvanen, A.C., Aalto-Setala, K., Harju, L., Kontula, K. & Soderlund, H. A primer-guided nucleotide incorporation assay in the genotyping of apolipoprotein E. *Genomics* **8**, 684-92. (1990).

170. Ross, P., Hall, L., Smirnov, I. & Haff, L. High level multiplex genotyping by MALDI-TOF mass spectrometry. *Nat Biotechnol* **16**, 1347-51. (1998).

171. Pastinen, T., Kurg, A., Metspalu, A., Peltonen, L. & Syvanen, A.C. Minisequencing: a specific tool for DNA analysis and diagnostics on oligonucleotide arrays. *Genome Res* **7**, 606-14. (1997).

172. Nyren, P., Pettersson, B. & Uhlen, M. Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Anal Biochem* **208**, 171-5. (1993).

173. Landegren, U., Kaiser, R., Sanders, J. & Hood, L. A ligase-mediated gene detection technique. *Science* **241**, 1077-80. (1988).

174. Nilsson, M. et al. Padlock probes: circularizing oligonucleotides for localized DNA detection. *Science* **265**, 2085-8. (1994).

175. Baner, J., Nilsson, M., Mendel-Hartvig, M. & Landegren, U. Signal amplification of padlock probes by rolling circle replication. *Nucleic Acids Res* **26**, 5073-8. (1998).

176. Akey, J.M. et al. Melting curve analysis of SNPs (McSNP): a gel-free and inexpensive approach for SNP genotyping. *Biotechniques* **30**, 358-62, 364, 366-7. (2001).

177. Lyamichev, V. et al. Polymorphism identification and quantitative detection of genomic DNA by invasive cleavage of oligonucleotide probes. *Nat Biotechnol* **17**, 292-6. (1999).

178. Prince, J.A. et al. Robust and accurate single nucleotide polymorphism genotyping by dynamic allele-specific hybridization (DASH): design criteria and assay validation. *Genome Res* **11**, 152-62. (2001).

179. Howell, W.M., Jobs, M. & Brookes, A.J. iFRET: An Improved Fluorescence System for DNA-Melting Analysis. *Genome Res* **12**, 1401-7. (2002).

180. de la Monte, S.M., Sohn, Y.K. & Wands, J.R. Correlates of p53- and Fas (CD95)-mediated apoptosis in Alzheimer's disease. *J Neurol Sci* **152**, 73-83. (1997).

181. Martinez, M., Fernandez-Vivancos, E., Frank, A., De la Fuente, M. & Hernanz, A. Increased cerebrospinal fluid fas (Apo-1) levels in Alzheimer's disease. Relationship with IL-6 concentrations. *Brain Res* **869**, 216-9. (2000).

182. Seidl, R., Fang-Kircher, S., Bidmon, B., Cairns, N. & Lubec, G. Apoptosis-associated proteins p53 and APO-1/Fas (CD95) in brains of adult patients with Down syndrome. *Neurosci Lett* **260**, 9-12. (1999).

183. Cotman, C.W. & Anderson, A.J. A potential role for apoptosis in neurodegeneration and Alzheimer's disease. *Mol Neurobiol* **10**, 19-45. (1995).

184. Behl, C. Apoptosis and Alzheimer's disease. *J Neural Transm* **107**, 1325-44 (2000).

185. Huang, Q.R., Morris, D. & Manolios, N. Identification and characterization of polymorphisms in the promoter region of the human Apo-1/Fas (CD95) gene. *Mol Immunol* **34**, 577-82. (1997).

186. Kanemitsu, S. et al. A functional polymorphism in fas (CD95/APO-1) gene promoter associated with systemic lupus erythematosus. *J Rheumatol* **29**, 1183-8. (2002).

187. Weiss, K.M. & Terwilliger, J.D. How many diseases does it take to map a gene with SNPs? *Nat Genet* **26**, 151-7. (2000).

188. Hirschhorn, J.N., Lohmueller, K., Byrne, E. & Hirschhorn, K. A comprehensive review of genetic association studies. *Genet Med* **4**, 45-61. (2002).

189. Goring, H.H., Terwilliger, J.D. & Blangero, J. Large upward bias in estimation of locus-specific effects from genomewide scans. *Am J Hum Genet* **69**, 1357-69. (2001).

190. Abraham, R. et al. Substantial linkage disequilibrium across the insulin-degrading enzyme locus but no association with late-onset Alzheimer's disease. *Hum Genet* **109**, 646-52. (2001).

191. Boussaha, M. et al. Polymorphisms of insulin degrading enzyme gene are not associated with Alzheimer's disease. *Neurosci Lett* **329**, 121. (2002).

192. Ait-Ghezala, G. et al. Confirmation of association between D10S583 and Alzheimer's disease in a case--control sample. *Neurosci Lett* **325**, 87-90. (2002).

193. Tanzi, R. Abstracts from the 8th International Conference on Alzheimer's Disease and Related Disorders. July 20-25, 2002. Stockholm, Sweden. *Neurobiol Aging* **23**, S1-606. (2002).

194. Blacker, D. et al. Alpha-2 macroglobulin is genetically associated with Alzheimer disease. *Nat Genet* **19**, 357-60. (1998).

195. Kehoe, P.G. et al. Variation in DCP1, encoding ACE, is associated with susceptibility to Alzheimer disease. *Nat Genet* **21**, 71-2. (1999).

196. Hu, Q. et al. The human FE65 gene: genomic structure and an intronic biallelic polymorphism associated with sporadic dementia of the Alzheimer type. *Hum Genet* **103**, 295-303. (1998).

197. Lambert, J.C. et al. A new polymorphism in the APOE promoter associated with risk of developing Alzheimer's disease. *Hum Mol Genet* **7**, 533-40. (1998).

198. Lehmann, D.J., Johnston, C. & Smith, A.D. Synergy between the genes for butyrylcholinesterase K variant and apolipoprotein E4 in late-onset confirmed Alzheimer's disease. *Hum Mol Genet* **6**, 1933-6. (1997).

199. Montoya, S.E. et al. Bleomycin hydrolase is associated with risk of sporadic Alzheimer's disease. *Nat Genet* **18**, 211-2. (1998).

200. Papassotiropoulos, A. et al. A genetic variation of cathepsin D is a major risk factor for Alzheimer's disease. *Ann Neurol* **47**, 399-403. (2000).

201. Nakano, K., Ohta, S., Nishimaki, K., Miki, T. & Matuda, S. Alzheimer's disease and DLST genotype. *Lancet* **350**, 1367-8. (1997).

202. Grimaldi, L.M. et al. Association of early-onset Alzheimer's disease with an interleukin- 1alpha gene polymorphism. *Ann Neurol* **47**, 361-5. (2000).

203. Nicoll, J.A. et al. Association of interleukin-1 gene polymorphisms with Alzheimer's disease. *Ann Neurol* **47**, 365-8. (2000).

204. Lambert, J.C. et al. The transcriptional factor LBP-1c/CP2/LSF gene on chromosome 12 is a genetic determinant of Alzheimer's disease. *Hum Mol Genet* **9**, 2275-80. (2000).

205. Kang, D.E. et al. Genetic association of the low-density lipoprotein receptor-related protein gene (LRP), an apolipoprotein E receptor, with late-onset Alzheimer's disease. *Neurology* **49**, 56-61. (1997).

206. Dahiyat, M. et al. Association between Alzheimer's disease and the NOS3 gene. *Ann Neurol* **46**, 664-7. (1999).

207. van Duijn, C.M. et al. Genetic association of the presenilin-1 regulatory region with early- onset Alzheimer's disease in a population-based sample. *Eur J Hum Genet* **7**, 801-6. (1999).

208. Kamboh, M.I., Sanghera, D.K., Ferrell, R.E. & DeKosky, S.T. APOE*4-associated Alzheimer's disease risk is modified by alpha 1-antichymotrypsin polymorphism. *Nat Genet* **10**, 486-8. (1995).

209. Namekata, K. et al. Association of transferrin C2 allele with late-onset Alzheimer's disease. *Hum Genet* **101**, 126-9. (1997).

210. Collins, J.S. et al. Association of a haplotype for tumor necrosis factor in siblings with late-onset Alzheimer disease: the NIMH Alzheimer Disease Genetics Initiative. *Am J Med Genet* **96**, 823-30. (2000).

211. Qiu, W.Q. et al. Insulin-degrading enzyme regulates extracellular levels of amyloid beta- protein by degradation. *J Biol Chem* **273**, 32730-8. (1998).

212. Vekrellis, K. et al. Neurons regulate extracellular levels of amyloid beta-protein via proteolysis by insulin-degrading enzyme. *J Neurosci* **20**, 1657-65. (2000).

213. Edbauer, D., Willem, M., Lammich, S., Steiner, H. & Haass, C. Insulin-degrading enzyme rapidly removes the beta-amyloid precursor protein intracellular domain (AICD). *J Biol Chem* **277**, 13389-93. (2002).

214. Fakhrai-Rad, H. et al. Insulin-degrading enzyme identified as a candidate diabetes susceptibility gene in GK rats. *Hum Mol Genet* **9**, 2149-58. (2000).

5.1 ON-LINE REFERENCES

1. Statistiska Centralbyrån	www.scb.se
2. HGVbase	www.hgvbase.cgr.ki.se
3. dbSNP	www.ncbi.nlm.nih.gov/SNP/
4. Cystic Fibrosis Mutation Database	www.genet.sickkids.on.ca/cftr/