Department of Neurobiology, Care Sciences and Society (NVS),
KI-Alzheimer disease research center
Karolinska Institutet, Stockholm, Sweden

ASSOCIATION STUDIES ON SUSCEPTIBILITY GENES IN ALZHEIMER DISEASE

Behnosh Fakhri Björk

Karolinska Institutet

Stockholm 2008

Published by Karolinska Institutet. Printed by Laserics Digital Print AB

ISBN 978-91-7409-176-2

To my family

ABSTRACT

Alzheimer disease (AD) is the most common form of dementia in the elderly. Due to the complexity of AD, it has been difficult to find genetic risk factors predisposing to disease. To date, three genes (*APP*, *PSEN1* and *PSEN2*) with disease causing genetic variants have been reported for the rare early onset monogenic forms of AD. For the more prevalent, late onset Alzheimer disease (LOAD), the ε4 allele of the *APOE* gene, is the only confirmed genetic risk factor. However, since functional studies on AD cases with known mutations have shed light on the molecular mechanisms of the disease and linkage studies have highlighted additional candidate regions for AD in our genome, it is of great importance to identify additional susceptibility genes for AD.

All studies included in this thesis use single nucleotide variation (SNV) genotyping as a method for identifying new susceptibility genes involved in the sporadic, late onset forms of AD. Studies I and II are replication studies, investigating the role of genetic variants in the genes *IDE* and *TFAM* in case-control populations from Sweden. These genes are located in a reported linkage area in chromosome region 10q. Studies III and IV are original association studies on the genes *COL25A1* and *PITRM1* encoding two recently reported proteins, CLAC and hPreP, respectively. The gene encoding CLAC is located in chromosome region 4q24-25, which is a region associated with increased allele sharing in two linkage studies on Swedish AD families. These findings, together with results from functional studies on this protein, prompted us to test the role of genetic variants in *COL25A1* with respect to risk for AD. The study on *PITRM1* was initiated by published results from functional studies on hPreP in post-mortem brain tissue from AD cases. We used a gene-wide association approach to test whether variants in this gene affect the risk for developing AD.

Our results support the involvement of genetic variants in *IDE*, *TFAM* and *COL25A1* in AD cases from Sweden; whereas, it does not for *PITRM1*.

LIST OF PUBLICATIONS

This thesis is based on the following studies and will be referred to by their Roman numerals in the text.

I. **B.F. Björk,** H . Katzov, P . Kehoe, L . Fratiglioni, B . Winblad, J. Prince, C. Graff. Positive association between risk for late-onset Alzheimer disease and genetic variation in *IDE*. *Neurobiol Aging 28 (2007), Issue 9, 1374 - 1380*.

II. A.C. Belin*****, **B.F. Björk***, M. Westerlund, D. Galter, O. Sydow, C. Lind, K. Pernold, L. Rosvall, A. Hakansson, B. Winblad, H. Nissbrandt, C. Graff, L. Olson. Association study of two genetic variants in mitochondrial transcription factor A (TFAM) in Alzheimer's and Parkinson's disease. *Neurosci Lett 420 (2007) 257-262*.

III. C. Forsell***, B.F. Björk*,** L. Lilius, K. Axelman, S.F. Fabre, L. Fratiglioni, B. Winblad, C. Graff. Genetic association to the amyloid plaque associated protein gene *COL25A1* in Alzheimer's disease. *Neurobiol Aging (2008)*. doi:10.1016/j.neurobiolaging.2008.04.009

IV. **B.F. Björk,** E. Glaser, L. Fratiglioni, C. Graff. Association study on the gene encoding a novel mitochondrial metalloprotease (hPreP), *PITRM1*, in Alzheimer disease. *Submitted manuscript*

CONTENTS

LIST OF ABBREVIATIONS

Aβ	Amyloid β-peptide
AD	Alzheimer disease
AICD	APP intracellular domain
APOE	Apolipoprotein E
APP	Amyloid precursor protein
CERAD	The Consortium to Establish a Registry for Alzheimer's Disease
CI	Confidence interval
COL25A1	Collagen type XXV alpha 1
CLAC	Collagenous Alzheimer amyloid plaque component
DASH	Dynamic allele specific hybridisation
DNA	Deoxyribonucleic acid
dATP	2'-deoxyadenosine 5'-triphosphate
dCTP	2'-deoxycytidine 5'-triphosphate
dGTP	2'-deoxyguanosine 5'-triphosphate
dTTP	2'-deoxythymidine 5'-triphosphate
dNTP	Deoxynucleotide triphosphates
FAD	Familial Alzheimer disease
GWAS	Genome-wide association study
HMG	High mobility group
hPreP	Human pre-sequence protease
HWE	Hardy Weinberg Equilibrium
IDE	Insulin degrading enzyme
kb	Kilo base-pair
KP	The Kungsholmen Project
LD	Linkage disequilibrium
LOAD	Late onset Alzheimer disease
MCI	Mild cognitive impairment
MZ	Monozygotic
mtDNA	Mitochondrial Deoxyribonucleic acid
NINCDS-ADRDA	National Institute of Neurological and Communicative Disorders and stroke–Alzheimer's Disease and Related Disorders Association
OR	Odds ratio
PCR	Polymerase chain reaction
PD	Parkinson disease
PITRM1	Pitrilysin metallopeptidase 1
PSEN	Presenilin
SNACK	The Swedish National Study on Aging and Care in Kungsholmen
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variation
TFAM	Transcription factor A, mitochondrial

INTRODUCTION

The discovery of the double helix structure of the DNA (Deoxyribonucleic acid) molecule by Watson and Crick in 1953[1] had a major impact on the field of molecular genetics. Today, we know that the DNA sequence is composed of over 3 billion chemical building blocks called nucleotides. Each nucleotide consists of three structures: a nitrogenous base, a sugar and a phosphate group. There are four types of bases: adenine (A), guanine (G), cytosine (C) and thymine (T) which can be subdivided in two classes, purins (A and G) and pyrimidines (C and T).

The DNA molecules contain "recipes" with instructions needed to compose other cell components such as proteins. These recipes are called *genes* and they contain coding regions called *exons*. The exons are composed of tri-nucleotide sequences called *codons*, and each codon codes for one of the 20 amino acids required to form a protein. Interestingly, the exons only cover about 1.5% of the human genome. The rest is comprised of RNA-genes, non-protein coding sections (*introns*), non-coding repetitive sequences that are involved in the regulation of gene expression and sequences with no known function. To date, the number of identified protein coding genes in our genome is approximately 20,000–25,000 (http://www.ebi.ac.uk/genomes/).

On average, two unrelated individuals share about 99.9% of their DNA sequence[2], thereby resulting in 0.1% of inter-individual variation in the genome. DNA sequence variation can occur in several forms, which will be discussed in more detail later, and each possibility for a specific variant is called an *allel*e.

The majority of DNA variations are harmless; but in some rare cases they alter the code in a gene in a way that it affects the proper production of its specific product which in turn may result in disease. However, it is important to mention that some of the variations that occur during gamete formation (meiosis) are actually beneficial and favored by *natural selection* contributing to human evolution.

DNA sequence variants are used in all forms of genetic investigation including linkage and association studies for mapping genes predisposing to disease. This thesis describes four genetic association studies in which the effects of genetic sequence variants of four candidate genes (*IDE*, *TFAM*, *COL25A1* and *PITRM1*) in AD are investigated in cases and controls from Sweden.

MUTATION, POLYMORPHISM AND VARIATION

While all three terms in the title of this section refer to alterations in our DNA sequence, the first two are hard to distinguish sometimes and can cause a great deal of confusion. The term *mutation* is often related to harmful changes and usually implies causation both in clinical practice and by the public. However, this interpretation is not entirely fair since the majority of mutations in the coding regions of the genome are "silent" (also called synonymous) and do not change the amino acid of the gene product. Also the usage of the expression *polymorphism* is misleading because it is commonly used for both harmless changes in the genome as well as the traditional definition that refers to variations which are present in >1% of the general population[3]. Therefore, to avoid confusion in the first part of the thesis, the neutral term *variation* will be used when referring to changes in the genomic sequence.

GENETIC VARIATION

In order to gain better understanding of why people exhibit difference in both susceptibility to disease and the response to pharmaceutical treatment, it is important to identify genetic variants mediating these effects.

Genetic variation is defined as the presence of alleles at a specific locus in the genome, which may be distributed with different frequencies between individuals or populations. Single nucleotide variants, small insertions/deletions (presence or absence of one or more nucleotides), and repetitive sequences (microsatellites) are the most common types of genetic variation in the human genome. These changes may occur spontaneously and may have no effect on the way the genomic message is read, or they can change the genomic code in a way that the message is not read correctly or not read at all.

There are a variety of reasons to why changes occur in our genomic sequence (DNA). The most common cause is believed to be aging. The majority of our cells have a limited life span; and as time goes by, these cells need to be replaced and are copied over and over again in a process known as cell division or *mitosis*. Random mistakes may occur during these cell divisions and may result in the creation of new genetic variants. On average, approximately 10^{17} cell divisions take place during a human's lifetime; and 10^{-8} spontaneous variants are expected to occur during each cell generation[4].

The international HapMap Project is an organization with the goal to develop a complete map of the human genome, which will describe the common patterns of

human genetic variation. All information produced by this project is available to researchers around the world in order to facilitate identification of common genetic variants affecting health, disease, responses to drugs and environmental factors as well as studying ethnical differences and human evolution.

Single nucleotide variation

As the name implies, single nucleotide variation (SNV), also known as single nucleotide polymorphism (SNP), refers to the alteration of a single base (A, C, G or T) in the genome. An example of which is when a DNA sequence AA**A**GGTCA is changed to AA**T**GGTCA. SNVs account for approximately 90% of all common types of genetic variations in humans and occur every 100-300 bases (http://www.ncbi.nlm.nih.gov/SNP/).

The variations can occur as either an insertion or deletion (indels) of a single base or as a substitution of one base for another. Theoretically, for each such genomic variant there should be four different alternatives, or alleles, (A, C, G or T). However, almost all reported SNVs have only two alleles at each position. SNVs are therefore generally referred to as di-allelic variants.

Substitutions are the most frequent form of SNVs and they can be subdivided in two different groups: *transitions* and *transversions*.

Transition is the substitution of a C↔T (amino acid change: pyrimidine ↔ pyrimidine) or A↔G (purine ↔ purine). Transversion, however, is the substitution of a C ↔ A, C ↔ G, G ↔ T or A ↔ T (purine↔pyrimidine), hence there are two possible choices at each transversion. Therefore, one might assume that these types of substitutions are twice as frequent as transition; however, this is not the case. In fact, the transition rate is more than 30 % higher than for transversions in humans. Also, C↔T transitions outnumber G↔A transitions, which is believed to be due to cytosine (C) having a much higher tendency to be methylated at carbon atom 5 which makes it more susceptible to spontaneous deamination to result in a thymine (T)[5].

During this last decade, the number of reported SNVs that are publicly available has increased from several thousands to nearly 4 million (http://www.NCBI.nlm.nih.gov/SNP/). Though this may seem like a large number, it is not due to the total number of SNVs is estimated to about 10 million which leaves several millions of SNVs yet to be discovered.

SNVs are not randomly spaced across the genome. They occur more often in non-protein-coding regions of genes. This is reasonable, since the coding regions contain

sequences that need to be preserved for normal biological functions. There are three classes of single base substitutions found in coding regions: *silent* (synonymous) substitutions which result in an alternative codon for the same amino acid, *nonsense* (non-synonymous) variations which change the codon for an amino acid to a stop codon (these variations have quite a drastic effect on the gene function), and *missense* variations that are also non-synonymous coding changes, but in this case, the codon will encode a different amino acid.

GENETIC INHERITANCE

Genetic inheritance was first described by Gregor Johann Mendel, also called the "father of genetics" for his studies on inheritance in pea plants in the mid 1860's. By selectively cross breeding pea plants with particular traits and observing the outcome over many generations, Mendel was able to see that some traits were inherited in predictable patterns. Mendel's observations led him to four important conclusions that came to form the basis of genetic inheritance. Today, Mendel's observations are known as Mendel's laws of inheritance:

- Each trait is represented by a gene that can be passed on to an offspring unchanged.
- An individual inherits a gene from each parent for each trait and the inheritance pattern of one trait will not affect the inheritance pattern of another.
- A trait may not show up in an individual but can still be passed on to the next generation.
- The genes coding for each trait segregate themselves during gamete (germ cell) production.

Our genetic code (DNA) is packed into organized structures called *chromosomes*. Humans carry 46 chromosomes grouped into 23 pairs. Since the chromosomes come in pairs, the genes they code for also come in pairs. Each parent passes on 22 chromosomes (called *autosomes*) and one sex chromosome. The autosomal chromosomes are numbered 1-22 according to their size from the largest to the smallest, and the sex-determining chromosomes are called X (female) and Y (male). During gamete production one of the two copies of each autosome is transmitted from the parent to the offspring, including an X chromosome from the mother and an X or a Y chromosome from the father. This happens in a random manner, known as *random*

segregation. However, it should be mentioned that chromosome inheritance from parents during gamete formation is more complicated than just inheriting a whole ancestral copy of a chromosome from each parent. Instead, the gamete may receive a mixture of the two homologous chromosomes because of events called *crossovers* (Figure 1). Such events force two alleles that lie together on a common parental chromosome to split and therefore results in a new combination of alleles originating from different grandparents on the same chromosome. The *recombination fraction* (θ) is an indication of the proportion of crossovers or recombinations that occur during meiosis between two loci. For loci that are far apart, recombination will occur by chance in 50% of the meiosis (θ=0.5). The closer the recombination fraction is to zero the more tightly linked are the loci. If two loci frequently co-segregate during meiosis and are found on the same *haplotype* (a combination of alleles in a single chromosome) more often than expected based on allele frequencies, they are said to be in *linkage disequilibrium* (LD; see section: Linkage disequilibrium). LD is weakened (closer to zero) for each recombination that occurs.

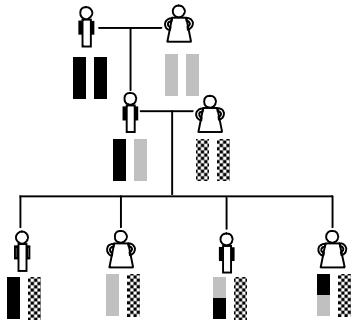

Figure 1. An example of recombinant chromosomes caused by crossover events in gametes.

DISEASE INHERITANCE

Genetic variations that occur in any of our genes may be disease causing if they inhibit the proper production of the specific product (protein) coded by that particular gene. This can mean that the protein will be produced in the wrong amounts, not produced at all, or that an alteration (gain or loss of function) of the protein function will occur, all of which may cause disease. Such errors, and thereby the disease, can be genetically inherited (passed on from one generation to another). The observed effect of a gene, or in this case the appearance of a disease, is called a *phenotype*. In those cases where an individual needs two copies of a disease causing variant to develop a certain phenotype (e.g. a disease), the disease is said to be *recessive*. If, on the other hand, the disease occurs when only one copy of the disease causing variant is sufficient to cause disease, the disease is referred to as *dominant*.

Therefore, genetic diseases can behave as single gene diseases (*monogenic*) when inherited in a genetically simple pattern according to Mendel's laws of inheritance (also referred to as Mendelian diseases); or they may be multifactorial, where multiple genes are likely to be involved in combination with lifestyle and environmental factors. These forms of diseases are referred to as *complex diseases*.

Single gene disease

Diseases caused by a single genetic defect are characterized by how they are passed down in families according to the laws of Mendelian inheritance and are classified as autosomal dominant, autosomal recessive, or sex chromosome (X/Y)-linked (Figure 2). To date, over 2400 diseases where the phenotype is described and the molecular basis is known have been identified (http://www.ncbi.nlm.nih.gov/omim/). Although monogenic diseases are relatively rare, it is of great importance to identify and understand the nature of their disease-causing product, since it will help us to develop better diagnostic tests for affected individuals or individuals at risk, as well as enhance our understanding of the function of different proteins in humans. Today genetic testing is not only used for diagnostic purposes, but also for the study of the pathoetiology of inherited diseases in order to understand their underlying mechanisms and provide potential targets for intervention and drug development.

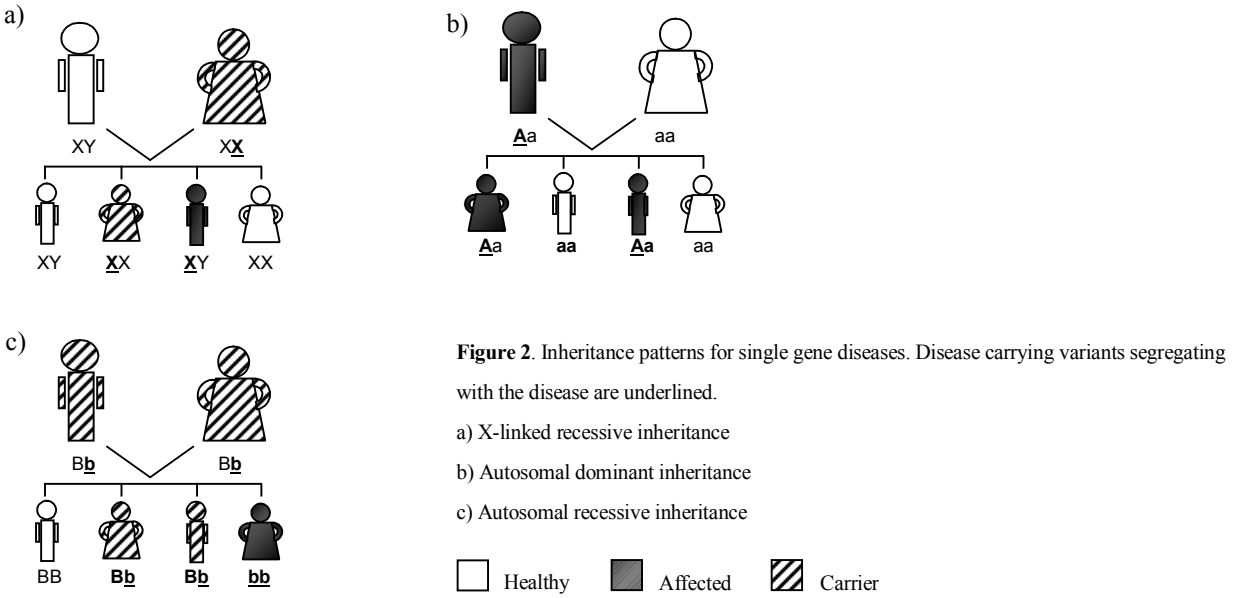

Figure 2. Inheritance patterns for single gene diseases. Disease carrying variants segregating with the disease are underlined.

a) X-linked recessive inheritance

b) Autosomal dominant inheritance

c) Autosomal recessive inheritance

Autosomal dominant inheritance

In this type of inheritance, only one defective copy of a gene located on any of our autosomes (the chromosomes numbered 1-22) is enough to cause disease. Since there are two copies of each gene to choose from during meiosis (each parent carries two copies of each gene), a child that has one affected parent has a 50% chance to inherit the disease causing variant of the gene. Males and females run an equal risk of being affected.

Autosomal recessive inheritance

In the case of autosomal recessive inheritance, two copies of a disease causing gene must be inherited for a person to develop disease. The affected child usually has unaffected parents who each carry a single copy of the disease gene. The parents are referred to as *carriers*. There is a 25% chance for a child to be affected if both parents are carriers. Because autosomal recessive diseases are very rare, the risk of being affected increases significantly with inbreeding.

Sex-chromosome linked inheritance

X-linked diseases are very rare and caused by mutations in genes on the X chromosome. They can occur in both recessive and dominant inheritance patterns. For the dominant form, only one copy of the defective gene is enough for the disease to appear in females. Since males only have one X chromosome, an affected male always passes on the disease to his daughters; whereas affected females have a 50% chance of passing on the disease causing gene to both their sons and daughters at each pregnancy.

The recessive form of an X-linked disease affects males more frequently than females, and the chance of passing on the disease differs between the two sexes. The daughters of an affected male will all be carriers of the disease but not affected, whereas 50% of sons of female carriers will be affected.

Only a few numbers of genes are located on the Y chromosome. Since only males carry a Y-chromosome, diseases caused by variation in these genes can only be transmitted from father to son.

One of the characteristics of monogenic diseases is that the expressivity may vary between affected individuals. This means that some have severe symptoms while others

have mild. Other factors that can vary are which systems of the body this genetic variation affects and the age at which the disease starts. These variabilities occur even within the same family. Another important characteristic of this form of genetic inheritance is the occurrence of *reduced penetrance*. This means that sometimes a person can have a dominant mutation without showing any clinical symptoms. The etiology of reduced penetrance has not been established, although there are theories suggesting that reduced penetrance may be a reflection of the combined effect of genetic, environmental, and lifestyle factors.

Complex diseases

Complex diseases (also referred to as multifactorial diseases) often cluster in families; however, they do not have a clear inheritance pattern. As opposed to monogenic diseases where modification in a single gene causes the disease, complex diseases may be caused by the combined effect of multiple genes (polygenic). In addition, the additive effects of genetic and environmental risk factors may also cause complex diseases. The chance of inheriting a complex disease is less than the chance for a single gene disease and is related to the extent of common environmental factors such as the number of genes a person shares in common with the affected individual (the closer two individuals are related, the more genes do they have in common) and the severity of the disease. One strong indication of the involvement of non-DNA-coded genetic factors in complex diseases is the phenotypic differences between monozygotic (MZ) twins. These differences are believed to be caused by environmental factors and alteration in epigenetic regulation (heritable changes which cause activation or inactivation of genes without changes in DNA sequence)[6] of gene activity, since MZ twins share 100% of their DNA[7].

Complex forms of genetically inherited diseases, such as AD, generally have a higher age of onset compared to the monogenic forms of the disease, which may reflect the importance of the extent of exposure to environmental factors as well as long-term gene interaction effects.

This thesis includes studies focusing on the impact of different genetic risk factors involved in the complex forms of AD, and the methods for how this was performed are explained in more detail in the next section.

GENETIC EPIDEMIOLOGY

The most accepted definition of genetic epidemiology was published by Newton Morton in the early 80's[8]. It states that "A science which deals with the etiology, distribution, and control of disease in groups of relatives and with inherited causes of disease in populations".

The term genetic epidemiology refers to the study of the correlation between a phenotypic trend (e.g. a disease) and genetic variation in a population, as well as the application of the results from such studies to control health problems. The aims of genetic epidemiology studies are to detect the inheritance pattern of a particular disease, localize the disease causing gene(s) in our genome, and find genetic marker(s) associated with disease susceptibility. Such studies also consider the occurrence of gene-gene interaction as well as the impact of environmental factors on genes.

Today, there are two main approaches used to map disease associated genes: *Linkage* and *association* studies.

Genetic linkage analyses in humans are basically used for mapping genomic regions that contain genes predisposing to disease. Therefore, this strategy is widely used in genetic mapping of families with multiple affected individuals and works extremely well for mapping monogenic diseases. In linkage studies two loci are taken into consideration: the locus for the disease trait, which is unknown, and the locus for a genotyped marker in the mapping process, which is known. This way, when a marker allele appears often with the disease trait (i.e. co-segregate), it is likely that they are located close to each other on the same chromosome.

Association studies

The genetic association study approach is the one used in all four studies presented in this thesis which is why a special section for this approach has been included.

Association studies differ from linkage studies in that linkage studies focus on the segregation of a particular gene or genetic variation and the disease in families while association studies consider the segregation of an allele and the disease across the whole population.

It is difficult to find genetic risk factors for complex genetic diseases, such as AD, by using linkage analysis alone, since many different genetic risk factors, each with a small effect may cause the disease. Association studies are therefore used as a complementary approach since they are more likely to detect genetic risk factors with

smaller effects on disease development. In this respect, association studies have greater power than family based linkage studies. There are, however, limitations to the association approach as well. For example, in order for the test to be performed, genetic variants predisposing to disease or in LD (see section: Linkage disequilibrium) with disease causing variants must be available for genotyping. Also, there is an economic aspect which limits the extent of an association study, since such studies may require a large number of markers in order to cover the genomic region of interest.

The majority of association studies conducted thus far have been the so called candidate gene studies based on biological data from functional studies on the molecular pathology of a disease. A common approach is to select a subset of unrelated, affected (cases), and unaffected (controls) individuals from a population (Figure 3) and genotype a number of genetic variants (usually SNVs) in the region or gene of interest. Then the frequencies of alleles from these variants are compared between cases and controls to determine whether they differ between the two groups, that is if one group shares a specific allele more often than the other. However, association studies are not only performed on single biological or positional candidate genes predisposing to disease, but also on candidate regions. The latest approach in association analysis is the genome-wide association study (GWAS). GWAS is designed to study genetic variants across the entire human genome. It is a powerful, cost-effective approach that with the help of modern, high-throughput genotyping methods, is able to map genetic variation across the entire genome of one or many individuals. The association results from these studies are used to better understand the molecular background of common, complex diseases such as Alzheimer disease and hence aid in the development of better diagnostic, treatment, and prevention strategies for such diseases.

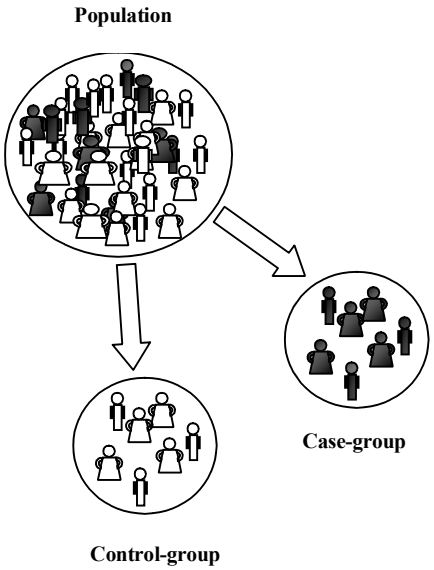

Figure 3. Case-control design in association studies

LINKAGE DISEQUILIBRIUM

Linkage disequilibrium (LD) is observed when a particular allele at one locus (SNV1) is likely to co-segregate with a specific allele at a neighboring locus (SNV2) on the same chromosome more often than expected by random segregation in a population. Thus LD is a measure of the association, correlation, or co-segregation of two separate loci in a population. LD between two sites is the result of a low number of early (ancestral) recombination events and is thus important in evolutionary biology and human genetics. While LD throughout the genome gives us information about the history, inbreeding and the geographic subdivision of a population, LD in each individual DNA region gives us information about events causing gene-frequency evolution, such as natural selection, gene conversion and variation. How and to what extent these factors affect LD between two loci depends on the number of recombination events occurring in that specific region during meiosis.

Meiotic recombination events are distributed unevenly throughout the genome[9]. Those regions with a high number of recombinations are called recombination "hot spots" and are separated by extended regions of higher levels of LD with limited recombination events. Each such region of strong LD is called a *haplotype block* or *LD block* and the size of these blocks in humans has been shown to vary from just a few kilo base-pairs (kb) to more than 100 kb[10].

Several studies have suggested a block-like pattern for different genomic regions in humans[11-16] and the discovery of these blocks initiated the International HapMap project which aims to construct genome-wide maps of LD patterns in the human genome in four ethnically different populations (European, Han Chinese, Japanese, and Yoruban). This project includes a total of 1301 individuals from 11 populations and has so far characterized over 4 million SNVs (http://www.hapmap.org).

Due to limited diversity within the haplotype blocks, it is suggested that genotyping only one or a few SNVs in each block (also called haplotype "tagging SNVs") should be sufficient to gain indirect information about most of the genetic variation in that same block and may not lead to much, if any loss of information. The HapMap project and the tagged SNV approach will therefore facilitate the identification of susceptibility genes for common complex diseases in association studies[10], as well as maximize the cost-effectiveness, especially when performing large scale GWASs. Furthermore, the discovery of the haplotype blocks has been of great importance in evolutionary biology for the reconstruction of the evolution of our genome.

The LD block structure hypothesis is however still under debate, partly because many different measures have been proposed for LD and partly because the human LD structure is very complex and unpredictable. For example, two loci (e.g. SNVs) located in close proximity may be in weak LD while two loci that are several kilo bases apart may be in strong (or complete) LD with each other[10]. In addition there are studies showing that some genetic regions lack a block like structure[17].

How to measure LD

There is a variety of measures available for LD[18]. Two of the most commonly used pair-wise LD measures (LD between alleles at two different loci) are r^2 and D'[10] and are both dependent on the linkage disequilibrium coefficient *D*.

Let us consider two SNVs (SNV1 and SNV2) each segregating with two alleles: A, a and B, b respectively with allele frequencies p_A, p_a, p_B and p_b.

If the alleles have equal frequencies and are randomly segregating during meiosis then there will be four possible gametes (or haplotypes) present in the population: AB, Ab, aB and ab at equal frequencies (see table). However, if these two SNVs are completely independent of each other (i.e. they are in complete linkage equilibrium) then the frequencies of the combined genotypes correspond to the product of the frequencies of the alleles they carry (p_A*p_B, p_A*p_b, p_a*p_B and pa*pb), see table.

			Alleles for SNV1	
			A	**a**
Alleles for SNV2	**B**		AB	aB
		Observed frequencies	pAB	paB
		Expected frequencies	pA*pB	Pa*pB
	b		Ab	ab
		Observed frequencies	pAb	pab
		Expected frequencies	pA*pb	Pa*pb

D is the difference between the observed and the expected gametic frequencies of alleles at two loci and defined by:

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

The D-value from this equation is then used for calculation of both D' and r^2 according to:

$D' = D/D_{max}$ where D_{max} is the theoretical maximum for the observed allele frequencies and is equal to the smaller of p_A*p_b or p_a*p_B.

and

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

The key difference between these two measures is that r^2 is much more sensitive to allele frequencies compared to D'. Both measures range from 0 (no disequilibrium) to 1 (complete or perfect disequilibrium) but as shown by their definitions, the interpretation of their values differs. If D' is equal to 1 (complete LD) it means that two or three of the four possible haplotypes are present in the population. On the other hand if D' is less than 1 it means that all four of the possible haplotypes are present which in turn indicates that recombination has occurred between these two sites. Therefore D' is the most useful tool when assessing the probability of historical recombination in a population[19]. However, D' encounters some difficulties when using SNVs with low minor allele frequencies in a small population, where the rare haplotype carrying the minor allele may not be observed. Such cases may generate a D' value of 1 and falsely suggest complete LD between two SNVs.

In order for r^2 to take the value of 1 between two markers (perfect LD), only two of the four possible haplotypes can be present and the allele frequencies must be identical. r^2 is suggested to be the best measure for association mapping because there is a simple linear relationship between r^2 and the sample size required to detect association between two loci[19]. r^2 is also more useful for identifying "tagged" SNVs" in large scale association studies in order to select the minimum set of SNVs to represent the haplotypes in a block.

Despite the popularity of these pair-wise measures, they do not provide a direct measure for multi-locus LD across a chromosomal region with several markers. In fact, there is still no optimal measure developed for the description of LD in large regions[10]. However, one common approach in multi-marker association studies of a chromosomal region, is to create a map of the combined analysis of all pair-wise LD measures between SNVs in that region. If only SNVs with moderate to high allele frequencies are included in the analysis, then D' and r^2 should provide a reliable LD map in that region.

STATISTICS

Hardy Weinberg equilibrium

According to the Hardy-Weinberg principle, the genotype and allele frequencies will remain constant or in equilibrium between generations in a randomly mating population. Consequently, HWE statistics, which is a chi-square analysis (see section: Chi-square analysis), is used to determine whether the observed genotype and allele frequencies in a study population fit what would be expected from a population in Hardy-Weinberg equilibrium. Deviation from this law may be caused by outside evolutionary forces such as genetic migration, natural selection and variation. Therefore, deviation from HWE is expected among the cases in an association study if the genetic marker that is investigated is the disease causing variant itself.
Other, more common causes of deviation from HWE are:

- Inbreeding, which increases homozygosity for genes.
- Assortative mating, which also causes increased homozygosity, but only for a specific trait and other traits in LD with them.
- A small population size, which has the tendency to cause random changes in genotype frequencies.

In most genetic studies, deviation from HWE for a genetic variant is a strong indication of genotyping error; and in those cases, that specific marker is usually excluded from the study.

Chi-square analysis

Chi-square analysis is a model-free, goodness of fit test that uses bivariate tables (crosstabs) to look for statistical significance. If the test is correctly performed, the chi-square analysis lets you know whether you should reject or accept your hypothesis by testing the deviation of observed frequencies from expected frequencies. In the case of genetic association studies, chi-square analysis compares the genotype and allele frequencies between cases and controls to determine whether there is a significant difference between these groups. For a single test, a p-value below 0.05 is considered to be significant. In genetic association studies, chi-square analysis is usually performed as a first step in genotypic and allelic association testing.

Logistic regression analysis

The next step in association testing (those including more than one independent variable) is to use logistic regression analysis that describes the relationship of several independent variables to a dependent variable such as the disease outcome (i.e. healthy or sick). The results from these analyses are presented as p-values, odds ratios (OR), and confidence intervals (CI).

In genetic association studies, an OR value represents the increased risk (when OR > 1) or decreased risk (when OR <1) of developing disease among cases who are carriers of a specific variant (allele) compared to controls. CI represents the upper and lower intervals in which the true mean of a population will fall. Hence, confidence intervals are used to indicate the reliability of an estimate. CI is presented as a percentage and usually set to 95%. This means that if 100 random samples were to be taken from a whole population then we would expect 95 of these (95%) to fall within the estimated CI.

COX-regression analysis

Cox regression is similar to logistic regression, but it is used for investigating the effect of several variables upon the **time** a specified event takes place. This type of regression is used for survival analysis to explore the relationship between the survival of a patient and the predictor variables.

Haplotype analysis

A haplotype for an individual is defined as a combination of alleles that are located closely together on the same chromosome and therefore tend to be inherited together. In association studies, haplotype analyses are often used in the detection and fine mapping of genomic regions associated to disease.

In linkage studies, analysis of haplotypes help to identify the minimal region of interest, limited by recombination events, and may also help determine phase when parents are unavailable for genotyping.

Multiple testing correction

A major limitation of the multiple testing approach is that of significance. When performing genetic studies, several independent tests may be performed in order to test many different hypotheses. For example, a single study may include a great number of genetic variants and statistical approaches. This may result in a type I error (false

positive, when rejecting a true null hypothesis) and can easily be avoided by adjusting the p-value for multiple comparisons. Very often, the Bonferroni correction method, which is the simplest and most conservative approach, is used. One simply multiplies the p-value of each test with the number of independent tests performed in the study to see if the significance level of the p-value ($p<0.05$) is retained.

ALZHEIMER DISEASE

Alzheimer disease (AD) is a complex, neurodegenerative disease. It is the most common form of late-life dementia, accounting for approximately 70 % of all dementia cases [20]. In Sweden, there are approximately 80.000 Alzheimer patients. This number is expected to increase to 100.000 within two years, due to increased life expectancy (http://www.alzheimerforeningen.se).

The vast majority of AD cases do not have a clear inheritance pattern and tend to strike people at the age of 65 years or older. There are however some rare heritable forms of the disease reported in individuals as young as ~30 years of age[21].

CLINICAL SYMPTOMS AND DIAGNOSIS

Alzheimer disease is named after the Bavarian psychiatrist Alois Alzheimer, who was the first to describe the clinical symptoms of AD in 1906 and later published his findings in 1907[22].

During a five year period he had observed a female patient who in his words, suffered from a "peculiar" disease. The clinical features Alzheimer described then are still applicable in modern diagnostics of AD. These include: short-term memory impairment, disorientation, general cognitive decline, paranoia, language difficulties and mood swings. In the advanced disease stage, motor and spatial abilities are affected.

At autopsy, he also discovered abnormal neuropathological changes in the brain tissue of this patient. These abnormalities are today known as amyloid plaques (extracellular deposits of amyloid peptides), intracellular neurofibrillary tangles (intraneuronal aggregations of hyperphosphorylated tau protein), decreased neuronal mass and brain volume (atrophy), and synaptic loss.

The early clinical symptoms of AD, such as forgetfulness or confusion, may be hard to distinguish from signs of normal aging, acute confusional states in the elderly[23] or early stages of other types of dementia. At this stage the patients suffer from memory impairment, have difficulties to remember recently learned tasks as well as having concentration difficulties. Otherwise they are functioning well and do not meet the clinical criteria for dementia. Some of these patients get the clinical diagnosis of mild cognitive impairment (MCI)[24]. With time, as the cognitive decline advances, the patients fulfill the diagnostic criteria for the different stages of AD. All AD patients

included in this thesis (studies I-IV) were clinically evaluated and diagnosed according to NINCDS-ADRDA (the National Institute of Neurological and Communicative Disorders and Stroke and Alzheimer's Disease and Related Disorders Association) criteria as possible, probable and definite AD[25]. The first two diagnoses are determined after mental and physical examination of the patients, whereas a definite AD diagnosis has to be confirmed with a neuropathological examination[26]. NINCDS-ADRDA is one of several diagnostic criteria developed for a more homogenous clinical diagnosis worldwide.

AMYLOID CASCADE HYPOTHESIS

There are several theories as to what mechanisms cause the neuronal cell death in AD, but the most prominent is still the amyloid cascade hypothesis[27]. It was first proposed more than 15 years ago and states that the accumulation of the amyloid β-peptide (Aβ) in the brain is the main cause of disease development and that neuronal cell death, tangle formation, inflammatory responses and vascular damage are secondary responses to this accumulation[27]. Aβ is a proteolytic product of the amyloid precursor protein (APP), a protein found throughout the body[28] whose normal function is unknown. The amyloid cascade hypothesis was based on the increased Aβ deposition in the brain of AD patients as a result of mutations in the APP gene and the fact that some Down syndrome patients who have an extra copy of the APP gene (trisomy 21) develop AD-like symptoms with age. Therefore, it has been of great importance to understand the regulatory mechanisms of APP processing during the years.

APP is a cell surface protein provided with a large extracellular N-terminal domain, a transmembrane segment, and a C-terminal cytoplasmic domain. Two main pathways have been identified for APP processing; the amyloidogenic pathway, which results in the production of intact Aβ and the non-amyloidogenic pathway which precludes the formation of amyloidogenic peptides. The first cleavage of the APP in this pathway is performed by β-secretase which cleaves the APP molecule at the N-terminal. This results in a soluble extracellular fragment of APP (sAPP-β) and a membrane-bound 99 amino acid residue C-terminal fragment (C99). In the non- amyloidogenic pathway, APP is first cleaved by α-secretase which results in the soluble extracellular fragment sAPP-α and a membrane-bound 83 amino acid residue fragment (C83). C99 and C83 are further processed by γ-secretase which is located in the APP transmembrane domain. This cleavage generates Aβ-peptides of different lengths (39-42 amino acid

residues) in the amyloidogenic pathway. Two of these, Aβ40 and Aβ42, are found in AD brains. The most common form of Aβ, Aβ40, comprises 90% of all secreted Aβ whereas Aβ42 accounts for only 10% of all produced Aβ[29, 30]. Aβ42 has been shown to have a higher tendency to aggregate and is the main constituent of senile plaques[31]. The γ-secretase cleavage in the non-amyloidogenic pathway generates a short protein fragment referred to as p3. In addition, a soluble cytosolic fragment, the APP intracellular domain (AICD), is generated in both pathways[32].

GENETICS

The identification of genetic risk-factors for complex diseases such as AD has been very challenging. A small fraction of all AD cases (~5%) are classified as familial Alzheimer disease (FAD), some of which are inherited in an early onset autosomal dominant fashion. This form of the disease often affects individuals before the age of 65 years. In our studies we define FAD as AD cases with a history of dementia in at least one first degree relative and/or at least one, second degree relative (including half siblings). Although FAD only covers a small percentage of all AD cases, linkage studies on FAD have helped to enlighten many aspects of AD. The literature reports disease causing variants in three genes for early onset autosomal dominant AD; amyloid precursor protein (*APP*; MIM# 104760)[33], presenilin 1 (*PSEN1*; MIM# 104311)[34], and presenilin 2 (*PSEN2*; MIM# 633044)[35] on chromosomes 21, 14 and 1 respectively.

The remaining 95% of all AD cases, however, do not exhibit familial inheritance and are referred to as sporadic late onset AD (LOAD) with an age of onset of ≥ 65 years. The only confirmed genetic risk factor for this more common form of AD is the ε4 allele of the gene encoding apolipoprotein E (*APOE* ε4; MIM# 107741.0016). Although this gene is neither sufficient nor necessary for a carrier to develop AD, it has been reported to reduce age of onset and increase the risk for disease development in a dose dependant manner[36-38].

In order to fully understand the disease causing mechanisms in AD, it is of great importance to identify additional genetic risk variants predisposing to the disease.

Several linkage studies have provided evidence for the existence of additional susceptibility genes on chromosomes 9, 10 and 12[39-44]. The genes investigated in studies I and II (*IDE* and *TFAM* respectively) in this thesis are two of the strongest biological candidates in a linkage region of chromosome 10q. However, it should be

mentioned that neither of the two genome scans on Swedish AD families reported by our group were able to replicate the previously reported linkage signals in chromosome 10[45, 46] .

At the time this thesis was written, over 500 different genes had been evaluated for association with late-onset sporadic AD (http://www.alzgene.org). Although approximately 30 of these genes have been suggested as susceptibility genes for AD, still none of these genes have been as widely accepted as a genetic risk factors as the ε4 allele of *APOE*.

However, the results from functional studies on AD, previously reported linkage studies and large meta-analysis on AD susceptibility genes (http://www.alzgene.org) all suggest the involvement of additional genetic components in AD. Therefore, it is of great interest to further investigate the role of genetic risk factors in AD.

Increasing attention has been directed towards the genome wide association study (GWAS) approach which is suggested as the new, high-throughput, genotyping tool for identifying genetic risk factors in complex genetic diseases such as AD. So far, encouraging results have been presented in early reports from four GWASs on AD [47-50].

PRESENT INVESTIGATIONS

AIMS

The overall aim of studies I-IV was to investigate the possible involvement of four different candidate genes (*IDE*, *TFAM*, *COL25A1* and *PITRM1*) in Alzheimer disease (AD), by performing case-control association studies and using single nucleotide variations (SNVs).

Study I. Positive association between risk for late-onset Alzheimer disease and genetic variation in *IDE*

Insulin degrading enzyme (IDE) has been suggested to be involved in the degradation of the extracellular amyloid-β peptide (Aβ). The gene encoding IDE was previously associated to AD in a population from Sweden[51]. Here we performed an association study, using four tagged SNVs, in an attempt to replicate the reported association results presented in Prince et al. 2003.

Study II. Association study of two genetic variants in mitochondrial transcription factor A (TFAM) in Alzheimer's and Parkinson's disease

Mitochondrial transcription factor A (TFAM) has been implicated to be involved in both Alzheimer and Parkinson Disease[52-55]. In this study we tested two previously reported SNVs for association in an AD and a PD case-control sample set.

Study III. Genetic association to the amyloid plaque associated protein gene *COL25A1* in Alzheimer's disease

The collagenous amyloid plaque component (CLAC) has been suggested to be involved in AD pathogenesis[56, 57]. We performed genetic association study in three different sample sets from Sweden, by genotyping a subset of SNVs in the gene encoding CLAC, namely *COL25A1*.

Study IV. Association study on the gene encoding a novel mitochondrial metalloprotease (hPreP), *PITRM1*, in Alzheimer disease

The recently identified human presequence protein (hPreP) is suggested to have the ability to degrade intracellular Aβ[79]. We tested this gene for genetic association by genotyping 18 SNVs in two case-control sample sets from Sweden.

SNV GENOTYPING METHODS

The vast majority of the SNV genotyping methods are based on the polymerase chain reaction (PCR) which amplifies the genomic region were the SNV of interest is located. The increased interest in SNV genotyping in genetic studies has led to the development of a large number of SNV genotyping methods. In this thesis I have used four different genotyping methods and the practical advantages and disadvantages of each method will be discussed in this section.

Dynamic allele specific hybridization (DASH)

DASH [58, 59] is used in study I. DASH is a robust genotyping method which scores DNA variation in the DNA sequence by taking advantage of the differences in melting temperature (Tm) between perfectly matched and mismatched base pairs in the DNA, and is therefore capable of measuring all types of genetic variations, not only SNVs. The DASH genotyping method is however more time consuming than the TaqMan (http://www.appliedbiosystems.com) and the iPLEXMassARRAY systems (Sequenome, Inc. http://www.sequenom.com) used in study IV, because of the pre-amplification step (PCR) and some intermediate steps.

Pyrosequencing

Pyrosequencing is a real-time "sequencing by synthesis" method used for detecting variations in a DNA sequence up to 20 bases and was initially developed as a DNA sequencing method. This method includes a cascade of enzymatic reactions and the detection of a SNV is based on the release of pyrophosphate during the incorporation of the four possible deoxynucleotide triphosphates (dATP, dCTP, dGTP and dTTP). These incorporations result in emission of light, which is proportional to the number of incorporated nucleotides. One major advantage of this method is that, if needed, the flanking regions around the genotyped SNV can be determined. However, the separate pre-amplification step (PCR) and intermediate steps before the actual sequencing process makes this method somewhat time consuming in comparison to TagMan and iPLEXMassARRAY system, which do not require these steps.

TaqMan

The TaqMan detection system, relies on allele-specific hybridization of oligonucleotides during PCR for allele discrimination. Genomic DNA and all necessary reagents are mixed in the first step, and the fluorescent signal is read after a

thermocycling step. Because there is no separate PCR step, or any other intermediate steps, this method provides one the simplest assay formats possible at an average laboratory.

iPLEXMassARRAY system

Sequenom iPLEXMassARRAY technology was used in study IV and is based on the single base extension of an extension primer into the region of the DNA where the variation of interest is located. Primarily, a locus-specific PCR reaction takes place followed by a locus-specific primer extension reaction, in which an oligonucleotide primer anneals immediately upstream of the polymorphic site being genotyped. This extension cocktail is spotted onto a chip, which contains 384 matrix pads. The mass difference between the extension products allows differentiation between SNV alleles. The iPLEXMassARRAY genotyping method was used for study IV, because it has the capacity of detecting a large number of SNVs in one assay (up to 40 SNVs). In our study, the maximum number of SNVs in one assay was 15, since some primer pairs generated indistinguishable PCR products and could not be included in the same assay. The genotyping process of study IV was outsourced to Mutational Analysis Facility (MAF) at Karolinska Institutet.

BACKGROUND AND RESULTS

Study I

Insulin degrading enzyme (IDE), a 110-kDa thiolmetalloendopeptidase in chromosome region 10q23-q25, is known to cleave small proteins such as Aβ. Several studies have suggested a role for this enzyme in the degradation and clearance process of Aβ in the brain[60-62]. IDE is also involved in the degradation process of the γ-secretase generated APP intracellular domain (AICD) and should therefore affect the nuclear signalling function suggested for the AICD fragment[63]. The involvement of IDE in Aβ deposition in the brain has been studied in mouse models with homozygous deletion and partial loss-of-function variants of the *IDE* gene with increased cerebral Aβ levels as a result[60, 64, 65]. The effect of genetic variation in *IDE* has also been studied in neuropathological cases, with results suggesting that *IDE* may be involved in Aβ deposition in the human brain[62].

The results from previously performed association studies on genetic variants in the *IDE* gene and the risk for AD are inconclusive (www.alzgene.org). However, variants in this gene were previously associated with AD in a Swedish sample set[51]. Therefore,

we decided to perform a replication study in an unreported case-control sample set from Sweden totalling **423** cases and **464** controls. We also included **356** cases and **110** controls from the United Kingdom (UK) and data, including one novel SNV, from the previously reported Swedish sample-set (**490** cases and **406** controls) in Prince et al. 2003. We found genetic association in two of the individual sample sets as well as the total combined genotype data from all included individuals in the study (**1269** cases and **980** controls). SPSS version 14.0 was used to perform all regression analyses in study I. The findings in this study support the suggested involvement of variants in *IDE* in AD pathogenesis.

Study II

Mitochondrial transcriptionfactor A (TFAM) is a high-mobility-group (HMG)-box protein[66] that is about 25 kDa and required for both transcription and replication of mammalian mitochondrial DNA (mtDNA). Mitochondria are cellular organelles known to be responsible for lipid metabolism, apoptosis, oxidative phosphorylation and energy supply of the cell.

Previous *in vivo* studies on Alzheimer disease brains have shown significantly reduced oxygen and glucose metabolism compared to control brains[67], thus disturbance of the mitochondrial capacity, due to variation in mtDNA, may be of importance in neurodegenerative diseases such as AD. Also, post-mortem analyses suggest reduced metabolism of oxygen and glucose in both Alzheimer and Parkinson disease (PD)[68]. Reduced mitochondrial function may therefore underlie chronic, progressive, late-onset neurodegeneration such as PD and AD.

The human *TFAM* gene consists of seven exons, six introns and covers 11.7 kb in chromosome region 10q21.1[69, 70]. Two variants in *TFAM* were recently investigated in a Caucasian AD case-control material collected from Germany, Switzerland and Italy[55]. One of these was also reported to be associated with AD in a pooled Scottish-Swedish AD case control material[53].

In study II, these two genetic variants in *TFAM* were genotyped in an AD case-control material (**433** cases and **348** controls) and a PD case-control material (**301** cases and **255** controls) both from Sweden, to further investigate the role of *TFAM* in two of the most prevalent neurodegenerative diseases. We found genetic association between AD and one of the genotyped SNVs. We did not observe any association between PD and the investigated variants in *TFAM*. Genotype distributions and allele frequencies were

compared between cases and controls in the two different sample sets using Chi-square analysis.

Study III

Collagenous Alzheimer amyloid plaque component (CLAC) is a recently identified plaque component. CLAC is a soluble proteolytic cleavage product of the neuron-specific type II transmembrane protein CLAC-P[56, 57] and believed to be involved in Aβ accumulation in the brain of AD patients. Although the actual function of this protein is not known, results from *in vitro* studies suggest that CLAC binds to Aβ fibrils and inhibits fibril-elongation[71, 72]. Also, Aβ fibrils seem to be more resistant to proteolytic degradation in the presence of recombinant CLAC[73]. Further studies are required for a more clear picture of the impact of CLAC binding to Aβ and to clarify the disagreement in the literature as to where the Aβ-binding motif is localized in the CLAC protein[72, 74].

Study III presents results from an association study between genetic variants in two haplotype-blocks in the *CLAC* gene, *COL25A1*, located in chromosome 4q25 and risk for AD.

Significant association between AD and four out of five genotyped SNVs was observed when comparing genotype and allele frequencies in **290** AD cases with known family history and **364** cognitive controls. These results were primarily replicated in a second case group including **527** clinical AD cases that did not fulfill the criteria for familial AD, using the same control group. The association was replicated for three of the originally associated SNVs when combining the two case groups. These results were obtained by performing logistic regression analyses. As a third step, we decided to test the effect of the associated SNVs in the general population and included a longitudinal population based study, consisting of **926** individuals. The associations to two of the SNVs were confirmed using COX-regression analysis. All of the above statistical analyses were performed using the SPSS statistical program version 14.0.

These results indicate an involvement of genetic variants in the *COL25A1* gene in AD.

Study IV

Recent studies have directed attention toward the intracellular accumulation of the Aβ-peptide and its presence in mitochondria[75]. It has been suggested that Aβ accumulation in mitochondria may induce cellular energy impairment, which in turn may contribute to cellular dysfunction in the brain of AD patients[76].

PreP is a mitochondrial protease first identified in plants, with the ability to degrade mitochondrial pre-sequences[77, 78]. The PreP homologue in humans is called hPreP and just like IDE, hPreP is a metallopeptidase and has been suggested to degrade mitochondrial Aβ and may therefore be of importance in AD pathology[79]. Study IV describes a gene-wide association study on the gene encoding hPreP, *PITRM1*, using **673** AD cases and **649** cognitive controls and genotyping 8 coding and 10 intronic SNVs. The results did not indicate any positive genotype, allelic or haplotypic association between any of the genotyped variants in *PITRM1* and AD.

GENERAL DISCUSSION AND FUTURE PERSPECTIVES

Common complex diseases with adult onset, such as Alzheimer disease (AD), tend to cluster in families but do not follow Mendelian inheritance patterns. Instead, it appears as complex diseases are triggered by genetic variations in multiple genes in combination with various environmental factors. Many association studies have been performed to relate genetic sequence variants (commonly SNVs) to AD (http://www.Alzgene.org). However, the identification of disease predisposing genes for this complex form of AD has proven to be difficult with relatively infrequent replicable results (http://www.Alzgene.org). One exception is the ε4 allele of the *APOE* gene, which has been repeatedly associated with increased risk in sporadic AD. Non-replicable, false positive results from association studies are often caused by small sample sizes hence these studies lack the power to identify disease predisposing genes with small effects. Therefore, large sample sets are needed in order to elucidate the involvement of such genetic risk factors when studying complex diseases. In this thesis the role of four different susceptibility genes have been investigated in regards to AD by using the case-control association study approach. Although positive associations to several genetic variants in three of the investigated genes have been reported, the mechanisms or even the causal genetic variants cannot be discerned from these studies and require further investigations.

The Genome-wide association study (GWAS) approach for mapping disease genes has been heralded as the future design for association studies. Because GWAS include many SNVs across the genome, they represent a promising way to study complex genetic diseases where many genes are likely to be involved in disease development. An ideal GWAS design for AD would include a large population including clinically well characterized cases and controls and a sufficient number of SNVs to cover the information in the entire genome. When or if significant association is observed between AD and SNVs, then a second, larger sample would be used in order to genotype the most significant SNVs for replication. Those SNVs which still are significant in this second step of the analysis would then be tested for replication in an even larger population. Finally, the most significant SNVs can be further investigated by fine mapping or sequencing the region in which they are located. If any of the SNVs are believed to be a causal variant, then we would investigate the

role of these variants in functional studies.

One major advantage of the GWAS approach is that it looks for new susceptibility genes in an unbiased manner and is therefore able to identify new genes associated to disease, as opposed to the candidate gene approach, which usually are based on already reported genes or the functional studies on gene products.

However, the current GWAS design is far from fully developed and has many unsolved problems; such as population stratification, the impact of epigenetics and environmental factors on disease development, multiple comparison tests and study design to achieve maximum power. Whether this approach holds the key to the understanding and identifying genetic risk factors for complex diseases remains to be seen.

ACKNOWLEDGEMENTS

I would like finish by thanking all the great people that have supported, encouraged and helped me with my thesis work during these years at the department of NVS (former NEUROTEC) I would specially like to thank:

My main supervisor
Ass. Prof. Caroline Graff
for so many things that I don't even know where to begin. You took me on as a PhD student, believed in me, and inspired me with your passion for science! Thank you for financial support, scientific guidance and for challenging me throughout my thesis work. I will miss our crazy brainstorming meetings, after which I was so full of energy that I felt like a "Duracell-Kanin" ☺

My co-supervisor
Prof. Bengt Winblad
for financial support, encouragement and your great work as an ambassador for our section, KI-Alzheimer disease research center

All past and present members of the "Team Graff"
for interesting scientific discussions and great times at the lab!
Lena Lilius, Annica Rönnbäck, Lotta Forsell, Susanne Froelich-Fabre, Lina Rosvall, Karin Dillner, Huei-Hsin Chiang, Karin Axelman, Håkan Thornberg, Jenny Björkström, Marie Fallström, Shunwei Zhu, Anna Sillén, Jesper Brohede and Tiina Robins

Lotta Forsell
for great collaboration and for putting up with my questions during these years (all millions of them!!!)

Dr. Annica Rönnbäck
for friendship, for always listening and always smiling! You are one of the kindest people I know!

Our collaborators
Andrea Carmine Belin, Lars Olson, Marie Westerlund, Laura Fratiglioni, Dagmar Galter, Karin Pernold, Jonathan A. Prince, Hagit Katzov, John Hardy, Amanda Myers, Patrick Kehoe and Elzbieta Glaser

Dr. Andrea Carmine Belin
I knew from the first day of our collaboration that we would become friends ☺! Thank you for a great collaboration, for being my Friskis companion (especially during our pregnency), for the Tuesday dinners and for our "mamma" lunches during our maternity leave!!!

All past end present senior scientists and everyone working in the labs at NEUROTEC/NVS, for providing help, support and a terrific scientific atmosphere

Dr. Helena Karlström
for your friendship, support, guidance, fun conversations and for your excellent taste in shoes ☺! You have been a rock to lean on!

Ass. Prof. Maria Ankarcrona
for friendship and good times with our boys!

All past and present PhD students who created a fun working environment and made conferences extra fun!

Erik Hjort, Eric Westman, Anna Sandebring and **Nodi Dehvari** for friendship and fun conversations at lunches and fikas!

Stefan Spulber for fun late night conversations, for being very helpful no matter what and for being a walking brain atlas when I studied for the Kandel exam!

My dear friends **Camilla Hansson-Petersen** and **Susanne Akterin**
You have become two of my closest friends. Thank you for always being there when I needed you. We have shared both good and bad times together and I honestly don't know what I would have done without the two of you during these years!

I would also like to thank **Gunilla Johansson**, **Balbir Kahur** and **Maria Roos**
for great organization and support to all the administrative work that I have had to deal with during these years.

All my friends outside of Karolinska Institutet, you know who you are ☺.

Åke, Lill and **Jonas** for always showing so much interest in my work.

Anna "Hallon" Hallbergson-Lewis and **Erika "Lemon" Gustafsson**
I love you my sweet "bästisar". Thank you for being such great and true friends and for all your support and involvement in my life! I will never stop hoping for the day when the three of us live in the same city or at least in the same country!!! I'm looking forward to our Boston week in December! Shopping patrullen, here we come!!!

Mike for being like a brother to me and always being there for me! Thank you for always making me laugh!!! And **Mathias** for interesting scientific discussions at the dinner table and for bringing a little bit of Skåne into our lives☺!

Melina and **Marlon** my sweet nieces for bringing so much joy into our lives☺.

My two favourite sisters, **Behnaz** and **Bita** ☺. For your support, love, patients and understanding!!! It has meant the world to me! I love you a WHOLE lot and I couldn't have asked for better sisters! I will have much more time on my hands from now on…I promise!!!!

Världens bästa **Mamma** och **Pappa**, for being such FANTASTIC parents and for your endless love and support in all my decisions in life! You have always been there for me and I love you for that!

Christoffer, the light of my life! Thank you for bringing so much joy into our home! Luckily you won't remember that I've been absent a lot during the time of writing my thesis. I love you!

Daniel, the love of my life. Thank you for being you and for making me the happiest person on earth! I couldn't have done this without you!

♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦

This thesis was supported by the Swedish Brain Power Initiative, Gun and Bertil Stohne's Foundation, Foundation for Old Servants, Alzheimer Foundation and LIONS Foundation for Research of Age Related Disorders.

I would also like to thank Karin Axelman at Karolinska University Hospital, Huddinge, Sweden for her contacts with familial AD cases, and Nenad Bogdanovic and Inger Nennesmo for the neuropathology of the Swedish AD samples from Huddinge Brain Bank. Finally I would like to express my sincere gratitude to all the participants for their cooperation and support of this study.

REFERENCES

1. Watson JD, Crick FH. The structure of DNA. Cold Spring Harb Symp Quant Biol. 1953;18:123-131

2. Levy S, Sutton G, Ng PC et al. The diploid genome sequence of an individual human. PLoS Biol. 2007;5:e254

3. Cotton RG, Scriver CR. Proof of "disease causing" mutation. Hum Mutat. 1998;12:1-3

4. Kondrashov AS. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. Hum Mutat. 2003;21:12-27

5. Holliday R, Grigg GW. DNA methylation and mutation. Mutat Res. 1993;285:61-67

6. Gopalakrishnan S, Van Emburgh BO, Robertson KD. DNA methylation in development and human disease. Mutat Res. 2008

7. MacGregor AJ, Snieder H, Schork NJ, Spector TD. Twins. Novel uses to study complex traits and genetic diseases. Trends Genet. 2000;16:131-134

8. Morton NE. Outline of Genetic Epidemiology, 1982

9. de Massy B. Distribution of meiotic recombination sites. Trends Genet. 2003;19:514-522

10. Slatkin M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. Nat Rev Genet. 2008;9:477-485

11. Daly MJ, Rioux JD, Schaffner SF et al. High-resolution haplotype structure in the human genome. Nat Genet. 2001;29:229-232

12. Dawson E, Abecasis GR, Bumpstead S et al. A first-generation linkage disequilibrium map of human chromosome 22. Nature. 2002;418:544-548

13. Gabriel SB, Schaffner SF, Nguyen H et al. The structure of haplotype blocks in the human genome. Science. 2002;296:2225-2229

14. Jeffreys AJ, Kauppi L, Neumann R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet. 2001;29:217-222

15. Johnson GC, Esposito L, Barratt BJ et al. Haplotype tagging for the identification of common disease genes. Nat Genet. 2001;29:233-237

16. Patil N, Berno AJ, Hinds DA et al. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science. 2001;294:1719-1723

17. Phillips MS, Lawrence R, Sachidanandam R et al. Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. Nat Genet. 2003;33:382-387

18. Devlin B, Risch N. A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics. 1995;29:311-322

19. Wall JD, Pritchard JK. Haplotype blocks and linkage disequilibrium in the human genome. Nat Rev Genet. 2003;4:587-597

20. Kukull WA, Higdon R, Bowen JD et al. Dementia and Alzheimer disease incidence: a prospective cohort study. Arch Neurol. 2002;59:1737-1746

21. Filley CM, Rollins YD, Anderson CA et al. The genetics of very early onset Alzheimer disease. Cogn Behav Neurol. 2007;20:149-156

22. Alzheimer A. Allgemeine Zeitschrift für Psychiatrie. 1907:146-148

23. Petersen RC. Acute confusional state. Don't mistake it for dementia. Postgrad Med. 1992;92:141-148

24. Winblad B, Palmer K, Kivipelto M et al. Mild cognitive impairment--beyond controversies, towards a consensus: report of the International Working Group on Mild Cognitive Impairment. J Intern Med. 2004;256:240-246

25. McKhann G, Drachman D, Folstein M et al. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology. 1984;34:939-944

26. Mirra SS, Heyman A, McKeel D et al. The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part II. Standardization of the neuropathologic assessment of Alzheimer's disease. Neurology. 1991;41:479-486

27. Hardy JA, Higgins GA. Alzheimer's disease: the amyloid cascade hypothesis. Science. 1992;256:184-185

28. Selkoe DJ. The cell biology of beta-amyloid precursor protein and presenilin in Alzheimer's disease. Trends Cell Biol. 1998;8:447-453

29. Haass C, Schlossmacher MG, Hung AY et al. Amyloid beta-peptide is produced by cultured cells during normal metabolism. Nature. 1992;359:322-325

30. Seubert P, Vigo-Pelfrey C, Esch F et al. Isolation and quantification of soluble Alzheimer's beta-peptide from biological fluids. Nature. 1992;359:325-327

31. Sisodia SS, St George-Hyslop PH. gamma-Secretase, Notch, Abeta and Alzheimer's disease: where do the presenilins fit in? Nat Rev Neurosci. 2002;3:281-290

32. Mills J, Reiner PB. Regulation of amyloid precursor protein cleavage. J Neurochem. 1999;72:443-460

33. Goate A, Chartier-Harlin M-C, Mullan M et al. Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. Nature. 1991;349:704-706

34. Sherrington R, Rogaev EI, Liang Y et al. Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. Nature. 1995;375:754-760

35. Rogaev EI, Sherrington R, Rogaeva EA et al. Familial Alzheimer's disease in kindreds with missense mutations in a gene on chromosome 1 related to the Alzheimer's disease type 3 gene. Nature. 1995;376:775-778

36. Saunders AM, Strittmatter WJ, Schmechel D et al. Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. Neurology. 1993;43:1467-1472

37. Strittmatter WJ, Saunders AM, Schmechel D 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. 1993;90:1977-1981

38. Corder EH, Saunders AM, Strittmatter WJ et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science. 1993;261:921-923

39. Myers A, Wavrant De-Vrieze F, Holmans P et al. Full genome screen for Alzheimer disease: stage II analysis. Am J Med Genet. 2002;114:235-244

40. Kehoe P, Wavrant-De Vrieze F, Crook R et al. A full genome scan for late onset Alzheimer's disease. Hum Mol Genet. 1999;8:237-245

41. Pericak-Vance MA, Bass MP, Yamaoka LH et al. Complete genomic screen in late-onset familial Alzheimer disease. Evidence for a new locus on chromosome 12. Jama. 1997;278:1237-1241

42. Hiltunen M, Mannermaa A, Thompson D et al. Genome-wide linkage disequilibrium mapping of late-onset Alzheimer's disease in Finland. Neurology. 2001;57:1663-1668

43. Blacker D, Bertram L, Saunders AJ et al. Results of a high-resolution genome screen of 437 Alzheimer's disease families. Hum Mol Genet. 2003;12:23-32

44. Farrer LA, Bowirrat A, Friedland RP et al. Identification of multiple loci for Alzheimer disease in a consanguineous Israeli-Arab community. Hum Mol Genet. 2003;12:415-422

45. Sillén A, Andrade J, Lilius L et al. Expanded high-resolution genetic study of 109 Swedish families with Alzheimer's disease. Eur J Hum Genet. 2008;16:202-208

46. Sillén A, Forsell C, Lilius L et al. Genome scan on Swedish Alzheimer's disease families. Mol Psychiatry. 2006;11:182-186

47. Coon KD, Myers AJ, Craig DW et al. A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. J Clin Psychiatry. 2007;68:613-618

48. Grupe A, Abraham R, Li Y et al. Evidence for novel susceptibility genes for late-onset Alzheimer's disease from a genome-wide association study of putative functional variants. Hum Mol Genet. 2007;16:865-873

49. Liu F, Park PJ, Lai W et al. A genome-wide screen reveals functional gene clusters in the cancer genome and identifies EphA2 as a mitogen in glioblastoma. Cancer Res. 2006;66:10815-10823

50. Reiman EM, Webster JA, Myers AJ et al. GAB2 alleles modify Alzheimer's risk in APOE epsilon4 carriers. Neuron. 2007;54:713-720

51. Prince JA, Feuk L, Gu HF et al. Genetic variation in a haplotype block spanning IDE influences Alzheimer disease. Hum Mutat. 2003;22:363-371

52. Bertram L, McQueen MB, Mullin K et al. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. Nat Genet. 2007;39:17-23

53. Blomqvist ME, Reynolds C, Katzov H et al. Towards compendia of negative genetic association studies: an example for Alzheimer disease. Hum Genet. 2006;119:29-37

54. Ekstrand MI, Terzioglu M, Galter D et al. Progressive parkinsonism in mice with respiratory-chain-deficient dopamine neurons. Proc Natl Acad Sci U S A. 2007;104:1325-1330

55. Gunther C, von Hadeln K, Muller-Thomsen T et al. Possible association of mitochondrial transcription factor A (TFAM) genotype with sporadic Alzheimer disease. Neurosci Lett. 2004;369:219-223

56. Hashimoto T, Wakabayashi T, Watanabe A et al. CLAC: a novel Alzheimer amyloid plaque component derived from a transmembrane precursor, CLAC-P/collagen type XXV. Embo J. 2002;21:1524-1534

57. Söderberg L, Zhukareva V, Bogdanovic N et al. Molecular identification of AMY, an Alzheimer disease amyloid-associated protein. J Neuropathol Exp Neurol. 2003;62:1108-1117

58. Prince JA, Brookes AJ. Towards high-throughput genotyping of SNPs by dynamic allele-specific hybridization. Expert Rev Mol Diagn. 2001;1:352-358

59. Howell WM, Jobs M, Gyllensten U, Brookes AJ. Dynamic allele-specific hybridization. A new method for scoring single nucleotide polymorphisms. Nat Biotechnol. 1999;17:87-88

60. Farris W, Mansourian S, Chang Y et al. Insulin-degrading enzyme regulates the levels of insulin, amyloid beta-protein, and the beta-amyloid precursor protein intracellular domain in vivo. Proc Natl Acad Sci U S A. 2003;100:4162-4167

61. Kurochkin IV, Goto S. Alzheimer's beta-amyloid peptide specifically interacts with and is degraded by insulin degrading enzyme. FEBS Lett. 1994;345:33-37

62. Blomqvist ME, Chalmers K, Andreasen N et al. Sequence variants of IDE are associated with the extent of beta-amyloid deposition in the Alzheimer's disease brain. Neurobiol Aging. 2005;26:795-802

63. Edbauer D, Willem M, Lammich S et al. Insulin-degrading enzyme rapidly removes the beta-amyloid precursor protein intracellular domain (AICD). J Biol Chem. 2002;277:13389-13393

64. Miller BC, Eckman EA, Sambamurti K et al. Amyloid-beta peptide levels in brain are inversely correlated with insulysin activity levels in vivo. Proc Natl Acad Sci U S A. 2003;100:6221-6226

65. Farris W, Mansourian S, Leissring MA et al. Partial loss-of-function mutations in insulin-degrading enzyme that induce diabetes also impair degradation of amyloid beta-protein. Am J Pathol. 2004;164:1425-1434

66. Parisi MA, Clayton DA. Similarity of human mitochondrial transcription factor 1 to high mobility group proteins. Science. 1991;252:965-969

67. Sims NR, Finegan JM, Blass JP et al. Mitochondrial function in brain tissue in primary degenerative dementia. Brain Res. 1987;436:30-38

68. Orth M, Schapira AH. Mitochondria and degenerative disorders. Am J Med Genet. 2001;106:27-36

69. Reyes A, Mezzina M, Gadaleta G. Human mitochondrial transcription factor A (mtTFA): gene structure and characterization of related pseudogenes. Gene. 2002;291:223-232

70. Tiranti V, Rossi E, Ruiz-Carrillo A et al. Chromosomal localization of mitochondrial transcription factor A (TCF6), single-stranded DNA-binding protein (SSBP), and endonuclease G (ENDOG), three human housekeeping genes involved in mitochondrial biogenesis. Genomics. 1995;25:559-564

71. Kakuyama H, Soderberg L, Horigome K et al. CLAC binds to aggregated Abeta and Abeta fragments, and attenuates fibril elongation. Biochemistry. 2005;44:15602-15609

72. Osada Y, Hashimoto T, Nishimura A et al. CLAC binds to amyloid beta peptides through the positively charged amino acid cluster within the collagenous domain 1 and inhibits formation of amyloid fibrils. J Biol Chem. 2005;280:8596-8605

73. Söderberg L, Dahlqvist C, Kakuyama H et al. Collagenous Alzheimer amyloid plaque component assembles amyloid fibrils into protease resistant aggregates. Febs J. 2005;272:2231-2236

74. Söderberg L, Kakuyama H, Möller A et al. Characterization of the Alzheimer's disease-associated CLAC protein and identification of an amyloid b-peptid-binding site. Journal of Biological Chemistry. 2005;280:1007-1015

75. Anandatheerthavarada HK, Devi L. Amyloid precursor protein and mitochondrial dysfunction in Alzheimer's disease. Neuroscientist. 2007;13:626-638

76. Caspersen C, Wang N, Yao J et al. Mitochondrial Abeta: a potential focal point for neuronal metabolic dysfunction in Alzheimer's disease. Faseb J. 2005;19:2040-2041

77. Moberg P, Stahl A, Bhushan S et al. Characterization of a novel zinc metalloprotease involved in degrading targeting peptides in mitochondria and chloroplasts. Plant J. 2003;36:616-628

78. Ståhl A, Moberg P, Ytterberg J et al. Isolation and identification of a novel mitochondrial metalloprotease (PreP) that degrades targeting presequences in plants. J Biol Chem. 2002;277:41931-41939

79. Falkevall A, Alikhani N, Bhushan S et al. Degradation of the amyloid beta-protein by the novel mitochondrial peptidasome, PreP. J Biol Chem. 2006;281:29096-29104

ON-LINE REFERENCES

1. http://www.ebi.ac.uk/genomes/ Assessed September 2008
2. http://www.ncbi.nlm.nih.gov/SNP/ Assessed Oktober 2008
3. http://www.ncbi.nlm.nih.gov/omim/ Assessed September 2008
4. http://www.hapmap.org Assessed Oktober 2008
5. http://www.appliedbiosystems.com
6. http://www.sequenom.com
7. http://www.Alzgene.org Assessed September 2008