INVESTIGATION OF GENETIC FACTORS
IN MULTIPLE SCLEROSIS

Izaura Lima Bomfim
This thesis is dedicated with great love to

Kristina, Erik
Lars, Ted
Maria Izaura, Vollmer
Solveig, Börje
my de Andrade Lima/Bomfim family

We can’t afford to be paralyzed by uncertainty.
ABSTRACT

Multiple sclerosis (MS) is a chronic disease where the transmission of signals in the central nervous system is affected leading to a broad range of symptoms. The aetiology of the disease is unknown but multiple genetic and environmental factors are believed to play a part. While no environmental factor has been unequivocally established a lot has happened with regard to our knowledge of the genetic component of MS. Besides the consistent replication of associations with the HLA Class II region, association analyses have disclosed several genes of interest in MS including IL7R, IL2RA, CLEC16A, CD58 and RPL5. The HLA association shows the strongest effect on risk for MS, odds ratio for the other genes are in the range 1.1–1.4. It is not known how many genes that are implicated in MS pathogenesis, an estimation is that there will be 20-100 genes, thus there remain genetic variants yet uncovered. The main aim of this thesis was to identify genetic variants that affect susceptibility to MS. Such knowledge could aid in the development of better therapies whilst current therapy highlights potential candidate genes to be studied for association with disease. We investigated eight genes for association with susceptibility to MS – in total 113 genetic markers were evaluated using a case-control candidate gene association approach. Over 4,700 patients and approximately 5,000 controls contributed to this research where single nucleotide polymorphisms (SNPs) in the following genes were studied: ITGA4, the gene coding for the target of the MS-drug Tysabri®; IL23R, a gene associated with the autoimmune disease inflammatory bowel disease (IBD); TRAF1/CS, a region reported to be associated with the autoimmune disease rheumatoid arthritis (RA); NGFR, RTN4R, LINGO1, TNFRSF19; four genes which codes for the Nogo receptor complex, a potential therapeutical target, and IRF5, a gene associated with other autoimmune diseases including systemic lupus erythematosus (SLE) and RA. We found two SNPs (rs4728142 and rs3807306) and one insertion/deletion polymorphism (CGGGG) located in the promoter and first intron of IRF5 to be associated with MS susceptibility (OR=1.1 for all three markers, P=10⁻⁵ for the two SNPs and P=10⁻⁴ for CGGGG) based on a combined analysis of association results from a Spanish, a Finnish and a Scandinavian cohort. 9,125 individuals contributed to the finding including 3,847 patients, 3,745 controls and 511 trio families (one patient and both parents). Furthermore, one SNP (rs741072) located in exon 6 of NGFR was found to be associated with risk of MS (OR=1.16, P=0.001) based on a dataset consisting of 2,108 patients and 1,871 controls of Scandinavian ancestry. Further studies are needed to verify the role of IRF5 and NGFR in the pathogenesis of MS.
LIST OF PUBLICATIONS

This thesis is based on the following articles that will be referred to by their Roman numerals:

I. ITGA4 polymorphisms and susceptibility to multiple sclerosis.
   Catherine O'Doherty*, IZAURA M. ROOS*, Alfredo Antiguedad, Ana M. Aransay, Jan Hillert, Koen Vandenbroeck.
   * Authors contributed equally.

II. The interleukin 23 receptor gene in multiple sclerosis: a case-control study.
    IZAURA M. ROOS, Ingrid Kockum, Jan Hillert.

III. TRAF1/C5 genetic variants in sporadic and familial multiple sclerosis.
     IZAURA LIMA BOMFIM, Helena Modin, Kristina Duvefelt, Cecilia M. Lindgren, Marco Zucchelli, Juha Kere, Ingrid Kockum and Jan Hillert
     Manuscript.

IV. The nerve growth factor receptor gene affects susceptibility to multiple sclerosis.
    IZAURA LIMA BOMFIM, Kristina Duvefelt, Lars Alfredsson, Tomas Olsson, Ingrid Kockum and Jan Hillert.
    Manuscript.

V. Interferon regulatory factor 5 (IRF5) gene variants are associated with multiple sclerosis in three distinct populations.

VI. IRF5 implicated in the pathogenesis of multiple sclerosis.
    IZAURA LIMA BOMFIM, Chuan Wang, Mohsen Khademi, Johanna Sandling, Boel Brynedal, Åslaug R. Lorentzen, Helle Bach Søndergaard, Annette B. Oturai, Elisabeth Gulowsen Celius, Lars Alfredsson, Ann-Christine Syvänen, Ingrid Kockum, Tomas Olsson and Jan Hillert
    Manuscript.
CONTENTS

1 Aims of the thesis.................................................................1
2 Genetics of complex diseases.............................................2
   2.1 Finding genetic factors involved in disease..................4
      2.1.1 Linkage analysis...............................................4
      2.1.2 Association analysis.........................................4
3 Multiple sclerosis ..............................................................9
   3.1 Clinical outcome measures .......................................9
   3.2 Diagnostic criteria..................................................10
   3.3 Treatment ................................................................10
   3.4 Pathogenesis ..........................................................10
   3.5 A complex disease..................................................12
      3.5.1 The genetic component ...................................12
      3.5.2 The environmental component..........................12
   3.6 MS genetics............................................................13
      3.6.1 MS genes.........................................................13
4 Materials and methods ......................................................15
   4.1 Patients and controls ..............................................15
   4.2 DNA extraction......................................................16
   4.3 Genotyping.............................................................16
      4.3.1 Pyrosequencing..................................................16
      4.3.2 TaqMan based allelic discrimination .................16
      4.3.3 MALDI-TOF mass spectrometry .......................16
      4.3.4 SNPstream and FP-TDI ....................................17
      4.3.5 Fragment analysis ...........................................17
   4.4 Sequencing..................................................................17
   4.5 Electrophoretic mobility shift assay (EMSA) ............18
   4.6 Proximity ligation assay (PLA)................................18
   4.7 Expression analysis..................................................18
      4.7.1 Preparation of PBMC and CSF-MC....................18
      4.7.2 mRNA and cDNA preparation..........................19
      4.7.3 Quantitative real-time PCR...............................19
   4.8 Statistical analyses...................................................19
      4.8.1 Posterior odds and false positive report probability19
      4.8.2 Allelic association measures..............................20
      4.8.3 Power..................................................................21
      4.8.4 Single point association analysis.......................21
      4.8.5 Adjusting for other factors...............................21
      4.8.6 Combining P-values.........................................22
      4.8.7 Haplotype association analysis..........................22
      4.8.8 Disease severity association analysis..................22
      4.8.9 Comparing expression levels............................22
      4.8.10 Genotype-phenotype correlation......................22
5 Results and discussion ......................................................23
   5.1 PAPER I .................................................................23
   5.2 PAPER II .................................................................26
5.3  PAPER III.................................................................................................................27
5.4  PAPER IV...................................................................................................................29
  5.4.1  A false positive finding?.............................................................................30
  5.4.2  NGFR in MS ...............................................................................................32
5.5  PAPERS V-VI..........................................................................................................33
6   Concluding remarks and future perspective.................................................39
7   Acknowledgements .............................................................................................41
8   References .............................................................................................................44
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBB</td>
<td>blood brain barrier</td>
</tr>
<tr>
<td>C5</td>
<td>complement 5</td>
</tr>
<tr>
<td>CDCV</td>
<td>common disease common variant</td>
</tr>
<tr>
<td>CNS</td>
<td>central nervous system</td>
</tr>
<tr>
<td>CSF</td>
<td>cerebrospinal fluid</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxiribonucleic acid</td>
</tr>
<tr>
<td>EAE</td>
<td>experimental autoimmune encephalomyelitis</td>
</tr>
<tr>
<td>EDSS</td>
<td>expanded disability status scale</td>
</tr>
<tr>
<td>EMSA</td>
<td>electrophoretic mobility shift assay</td>
</tr>
<tr>
<td>FPRP</td>
<td>false positive report probability</td>
</tr>
<tr>
<td>FP-TDI</td>
<td>fluorescent polarization template directed incorporation</td>
</tr>
<tr>
<td>HLA</td>
<td>human leukocyte antigen</td>
</tr>
<tr>
<td>HPLC</td>
<td>high pressure liquid chromatography</td>
</tr>
<tr>
<td>HW</td>
<td>Hardy-Weinberg</td>
</tr>
<tr>
<td>IL23R</td>
<td>interleukin 23 receptor</td>
</tr>
<tr>
<td>IMSGC</td>
<td>International Multiple Sclerosis Genetics Consortium</td>
</tr>
<tr>
<td>IRF5</td>
<td>interferon regulatory factor 5</td>
</tr>
<tr>
<td>ITGA4</td>
<td>α4 integrin</td>
</tr>
<tr>
<td>LD</td>
<td>linkage disequilibrium</td>
</tr>
<tr>
<td>LINGO1</td>
<td>leucine rich repeat and Ig domain containing 1</td>
</tr>
<tr>
<td>MALDI-TOF</td>
<td>matrix-assisted laser/desorption ionization - time of flight</td>
</tr>
<tr>
<td>MS</td>
<td>multiple sclerosis</td>
</tr>
<tr>
<td>MSSS</td>
<td>multiple sclerosis severity score</td>
</tr>
<tr>
<td>NGFR</td>
<td>nerve growth factor receptor</td>
</tr>
<tr>
<td>OND</td>
<td>other, non-inflammatory, neurological disease</td>
</tr>
<tr>
<td>OND.INF</td>
<td>other, inflammatory, neurological disease</td>
</tr>
<tr>
<td>OR</td>
<td>odds ratio</td>
</tr>
<tr>
<td>P</td>
<td>P-value</td>
</tr>
<tr>
<td>PBMC</td>
<td>peripheral blood mononuclear cells</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>PPMS</td>
<td>primary progressive MS</td>
</tr>
<tr>
<td>RA</td>
<td>rheumatoid arthritis</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>RRMS</td>
<td>relapsing-remitting MS</td>
</tr>
<tr>
<td>RTN4R</td>
<td>reticulon 4 receptor</td>
</tr>
<tr>
<td>SLE</td>
<td>systemic lupus erythematosus</td>
</tr>
<tr>
<td>SNP</td>
<td>single nucleotide polymorphism</td>
</tr>
<tr>
<td>SPMS</td>
<td>secondary progressive MS</td>
</tr>
<tr>
<td>TNFRSF19</td>
<td>tumor necrosis factor receptor superfamily, member 19</td>
</tr>
<tr>
<td>TRAF1</td>
<td>TNF receptor-associated factor 1</td>
</tr>
<tr>
<td>VLA-4</td>
<td>very late antigen 4</td>
</tr>
<tr>
<td>WTCCC</td>
<td>Wellcome Trust Case-Control Consortium</td>
</tr>
</tbody>
</table>
1 Aims of the thesis

Multiple sclerosis (MS) is a disease characterized by two or more neurological signs occurring in different parts of the central nervous system at at least two different occasions. MS affects women more often than men and is unevenly distributed worldwide, Scandinavia being one of the high risk areas. We do not expect there to be one unique cause of MS, rather it is multifactorial with the involvement of genes as well as environmental factors. The present study aims at identifying genetic factors associated with disease by comparing frequencies of genetic variants between unrelated patients and controls.

The specific aims were as follow:

PAPER I
To determine if polymorphisms in the gene coding for the α4 subunit of the VLA-4 receptor, ITGA4 (2q31.3) influence susceptibility to multiple sclerosis.

PAPER II
To evaluate the potential involvement of IL23R (1p31.3) in MS by analysing as many as 32 SNPs within and surrounding the gene, in a Scandinavian dataset.

PAPER III
To investigate whether the TRAF1/C5 (9q33-q34) region is a common autoimmune region for RA and MS and address the potential role of the anaphylatoxin molecule C5a by mutation analysis in family members of a consanguineous pedigree.

PAPER IV
To test the hypothesis that genetic variants in four genes coding for the Nogo receptor complex influence the risk of MS in a Scandinavian dataset. The investigated genes were: NGFR (17q21-q22), RTN4R (22q11.21), LINGO1 (15q24.3) and TNFRSF19 (13q12.11-q12.3)

PAPER V
To compare MS patients and controls from three distinct populations with regard to 10 markers located near or within the IRF5 gene (7q32) in order to determine if these variants are involved in MS susceptibility as they were implicated in other autoimmune diseases.

PAPER VI
To validate the results of paper V where two SNPs (rs4728142 and rs3807306) and one insertion/deletion polymorphism (CGGGG) were found to be associated with MS in a larger material and, in addition, to determine whether there were differential expression of IRF5 between the patient and control groups in PBMC as well as CSF.
2 Genetics of complex diseases

Complex diseases are multifactorial (both genetic and environmental factors are involved), polygenic (involves several genes) and characterized by reduced penetrance (a given genotype does not always imply disease) as well as genetic heterogeneity (disease develops due to different genes and/or alleles in the same pathway or due to different pathways altogether; reviewed in 1).

These concepts can be visualized with the aid of the causal pie model\(^2\) where all components necessary for a disease to occur are depicted as pieces of a pie (component causes). Whenever an individual has ALL the component causes in a pie he or she will have the disease – even if there is just one component cause missing, disease will not occur. One full pie represents a causal mechanism, one sufficient cause. In MS there is more than one sufficient cause, we don’t know yet how many. A hypothetical example is shown in figure 1.

![Figure 1: A hypothetical example using the pie-model\(^2\) incorporating the following concepts: multifactorial disease, neither-necessary-nor-sufficient factors, reduced penetrance, genetic heterogeneity, polygenic disease, biologic interaction. The three sufficient causes are analysed in seven individuals. Individuals 1 and 2 are affected due to sufficient cause B; individual 3 is affected due to sufficient cause C; individual 4 is affected due to either sufficient cause A or C, we can’t tell which. Individuals 5, 6 and 7 are not affected since there is no full pie.](image)

Each piece of pie can depict either a genetic factor or an environmental factor. Complex diseases are expected to contain both types – they are multifactorial. Furthermore, in complex diseases one piece of pie (one component cause/one factor) is neither necessary nor sufficient for disease to occur as depicted in figure 1 by no piece of pie being present in every pie (no necessary factor) and by all pies containing more than one component cause (no sufficient factor). The concept of reduced penetrance (meaning that the probability of getting disease given that one individual have a risk factor is less than 1; in other words there are unaffected individuals that carry the risk factor) implies that the pie that holds the
risk factor is composed of more than one component: individuals carrying the factor are unaffected because some other component is missing, the factor is not sufficient for disease to occur – in some instances that other component might be called “chance”. Reduced penetrance can be affected by the number of sufficient causes (pies) present in a given disease – the more pies, the more probable it is that a given factor is present in more than one pie; the more pies that an individual has “started to fill” the higher the risk of getting disease since there is more than one way of contracting disease; thus even if penetrance is reduced for a given factor at the level of one pie, since the factor is present in several pies, that reduction won’t be as noticeable at the population level (when all pies are considered simultaneously).

The presence of more than one pie implies etiologic heterogeneity – there are several ways of getting disease. When considering genetic factors the term genetic heterogeneity is used instead. There are two types of genetic heterogeneity: if alleles at the same locus are present in different pies there is allelic heterogeneity; when different genes (loci) are involved in different sufficient causes there is locus heterogeneity. The level of genetic heterogeneity present in a complex disease is dependent on the origins of disease, the common-disease-common-variant (CDCV; see section 2.1.2.3.1) hypothesis implies less genetic heterogeneity compared to the common-disease-rare-variant hypothesis.

When there are different alleles, genes or factors in general in the same pie there is biologic interaction between them. Interaction means that all factors act to produce disease – their action might occur at the same time, as when molecules physically interact, or they might occur separate in time – one factor “laying the grounds” for the actions of another factor. Figure 2 allows me to visualize that absence of a factor can be a component cause just as presence of a factor is and introduce the concepts of synergistic and antagonistic biologic interaction.

![Figure 2: a sufficient cause which consists of four component causes: presence of X, presence of Z, absence of Q and presence of R.](image)

Factor X and factor Z are both needed in this sufficient cause for disease to occur – if one is removed there is no disease: there is synergistic biologic interaction between factors X and Z. On the other hand, for disease to occur because of the causal mechanism presented in figure 2 there has to be an absence of factor Q – if Q is present there is no disease due to this
sufficient cause. When evaluating biologic interaction between say factor Z and factor Q (or between Q and any of the three other factors present in this example), if both factors are present, there is no disease thanks to an antagonistic biologic interaction between Q and Z. When no case is caused or prevented by simultaneous presence of two factors the effects of these factors are independent.

2.1 FINDING GENETIC FACTORS INVOLVED IN DISEASE

The human genome consists of 3.02 billion bases – where should we start looking for a disease contributing factor? Technology is now in place at reasonable cost to allow studies to be performed without having to answer that question, genome-wide association studies (see section 2.1.2.2) on several diseases have been published and there are several ongoing projects. This thesis reflects a period when this technology was not yet accessible at a reasonable cost, instead a candidate gene association design (see section 2.1.2.2) was chosen. There are several approaches for selecting a particular region or gene, most often a combination of evidence will be used. In Mendelian genetics, linkage analysis (see section 2.1.1) has provided regions for further investigation. In complex genetics, attempts at finding regions containing a disease gene by linkage analysis in humans have provided few regions of interest while linkage analyses in animal models have delivered interesting regions, some of which have proved to contain disease genes when investigated in human datasets. Other sources of evidence for the potential involvement of specific genes are expression studies – conducted either on few genes or more extensive designs with investigation of thousands of transcripts (expression profiling) – and functional studies. After identifying an allele as associated with disease further work is necessary to verify that the allele is implicated in the pathogenesis of a disease, and not just a confounder, as well as the mechanisms of action of the allele in disease.

2.1.1 Linkage analysis

In linkage analysis deviations from Mendel’s second law, the law of independent assortment, is used to find where the disease allele is located. Thus, linkage analysis does not aim at identifying the disease allele per se but rather identify a locus that contains the disease allele. Usually linkage studies are performed on a genomewide scale, utilizing microsatellites and, more recently, SNPs located throughout the genome in related individuals. In complex genetics, where large extended pedigrees are rare, linkage analysis is commonly based on allele sharing between affected relative pairs such as affected siblings. However, power is low compared to association approaches at least for modest effects.

2.1.2 Association analysis

Association is a statistical statement about the co-occurrence of factors. In medical genetics, the factor sought to be associated with disease might be one or more bases in the DNA molecule (a sequence variant), a structural or an epigenetic variant. This thesis focuses
exclusively on association of sequence variants, mainly single nucleotide polymorphisms (SNP), with disease, although it doesn't exclude that the actual causal variant ultimately found would be of another category.

2.1.2.1 Direct and indirect association studies
A functional variant (be it a SNP or another type of genetic variant) is a potential disease-causing unit, thought to be found in coding regions or regulatory regions surrounding these coding regions – thus studies that target functional variants are called direct association studies. In an indirect association approach markers, commonly SNPs, are selected based on their tagging abilities (high linkage disequilibrium (LD) with other markers) or just randomly. Both approaches, direct and indirect association, are commonly undertaken in parallel.

2.1.2.2 Candidate gene and genomewide association studies
Direct and indirect association can be conducted either in candidate genes or regions or with thousands of markers throughout the genome in genomewide association studies. Statistically significant differences in allele or haplotype frequencies between cases and controls, or distortion of transmission in trio families (parents and one affected child), are taken as an indication of association with disease.

Candidate gene association studies are based on current knowledge of disease biology as well as evidence for regions of special interest produced by previous linkage or association analysis. This is a strength as well as limitation of candidate gene studies – it provides a higher prior probability and thereby affects the probability of reporting false positives (see Material and Methods) but it requires that the gene is suspected to be involved in disease, which could lead to important pathways being missed.

A limitation common to both candidate gene and genomewide approaches is that what can be discovered is critically dependent on patterns of linkage disequilibrium, since not all genetic variation is tested but rather tag SNPs and/or randomly selected SNPs are included.

2.1.2.3 Underlying assumptions
2.1.2.3.1 Common-disease-common-variant hypothesis (CDCV)
How many sufficient causes (pies) are there in complex diseases? The CDCV-hypothesis reasons that common diseases are common both with regard to prevalence as to distribution, they are frequent and widespread. These characteristics could be the result of susceptibility alleles common in a founding population of modern humans (Homo sapiens sapiens) that became distributed with human global dispersion (based on the recent African origin/replacement hypothesis; reviewed in 6) – in this scenario common variant has the additional meaning of shared variant, all populations share these susceptibility variants. The opposing view states that mutations giving rise to susceptibility alleles occurred several
times in different populations and thus are likely to be rare and not shared among populations. The true picture probably contains both types of disease alleles – those that are common in frequency and shared among populations and those that are rare and population-specific. The main study design used in all papers in this thesis (case-control association analysis) aims at identifying the common variants in accordance to the CDCV-hypothesis but we do not claim that rare variants don’t exist.

2.1.2.3.2 Ancestral haplotypes
As a consequence of the CDCV-hypothesis, we expect that the haplotype a particular susceptibility allele arose on is the same (identical by descent) in all affected individuals living today. That haplotype is expected to have been broken down by recombination during meiosis and its size is a function of the recombination fraction and the number of generations since the mutation arose. The presence of this common, ancestral, haplotype makes indirect association as well as haplotype analysis sensible and possible as long as a detectable percentage of cases share this ancestral haplotype (reviewed in 7). The tools (the LD metrics D’, r²; see Material and Methods) we use to detect indirect association do not inherently imply that alleles are present on the same haplotype through. Rather, allelic association is measured i.e. whether alleles are found together in individuals more often than expected by chance either genetically linked or unlinked. In order to overcome that, algorithms that estimates haplotypes (e.g. the EM algorithm) are used and LD patterns are visualised through calculations of D’ or r² between alleles on these estimated haplotypes – inference on the possible source of association signal can then be made.

2.1.2.4 Ascertainment of cases and controls
Epidemiologists have demonstrated the importance of cases and controls coming from the same source population and that the control group should reflect the relative size of exposed and unexposed components of that source population2. For genetic epidemiology, a special concern regards the ancestry of the groups being compared – if cases and controls don’t share ethnicity then markers associated with ethnicity will appear to be markers associated with disease. The impact of population stratification on spurious association results is a topic of debate8–12. Empirical evidence suggest that population stratification might not be a problem within one population (such as the British) while avoidance of non-European ancestry when studying a European population still is warranted13.

2.1.2.5 Selection of genes and markers
Genomewide association studies are based on markers that cover much of the genetic variation and/or are selected more randomly with consideration taken to genotyping technology (reviewed in 14, 15). While not all genes will be covered by genomewide studies on current platforms, a wide spectrum of genes are tagged/randomly picked, thus it is considered a hypothesis-free design with regard to genes. Selection of genes for candidate gene studies is based on plausibility due to biological function, evidence from animal
studies, expression or linkage analysis, preferably by genomic convergence i.e. from as many types of evidence as possible (e.g. linkage AND expression AND animal studies).

2.1.2.6 Interpreting results

2.1.2.6.1 Negative studies

It is hard to predict LD patterns at the local scale, inclusion of more markers might change the picture – thus a gene cannot be excluded based on association studies as they are designed today, even if all studied markers showed no evidence of association, there might be an unobserved susceptibility allele which was not captured by the analysed markers. In the future, when whole genome sequencing becomes available at reasonable costs, the problem won’t be of uncertainty of coverage but rather a matter of power.

2.1.2.6.2 Positive studies

When a genetic variant is found to be more common among patients than among controls it could be because it is the susceptibility factor but it could also be the effect of confounding (the associated variant is in linkage disequilibrium with the susceptibility factor alternatively there is population stratification) or due to chance (sampling error) (reviewed in 7). Further investigation is thus necessary in order to clarify which of these explanations lie behind the positive finding as it has bearing on which conclusion to make about the disease-causing mechanisms.

Another issue with positive genetic associations in complex disease is their effect size. Are effect sizes (OR) in the order of 1.1-1.5 relevant? A small relative effect size doesn’t automatically translate to a small absolute effect, if the variant is common there could still be many individuals whose disease could be attributed to such a variant. More importantly, information on the biological mechanism is provided regardless of the size of the effect.

2.1.2.7 Follow-up of findings in association studies

Initial association studies, while performed in better powered studies today still need to be replicated in independent samples. After confirmation, other variants that could be linked to the marker should be sought for either by further fine-mapping of the region or, preferably, by sequencing. In a near future there will be sequencing data available from 1,200 genomes delivered through the 1000 genomes project (www.1000genomes.org), providing a comprehensive database of genetic variation and enabling visualisation of high-resolution LD patterns. Additional sequencing might be warranted in the search for population specific variation and for variation found only among affected individuals. If no other associated variants are found, there could still be a variant further away in LD with the associated marker as the existence of long distance LD has been reported16; until we are able to sequence the whole genome that’s an uncertainty we need to comply with. For regions with high LD, further pinpointing of an association signal becomes difficult by genetic analyses. In any case, several strategies for identifying the mechanisms of action of
the associated variant will probably be necessary such as gene-gene and gene-environment interaction analyses, expression studies as well as *in vitro* and *in vivo* functional studies.
Multiple sclerosis

Multiple sclerosis (MS) is a chronic inflammatory disease of the central nervous system (CNS) with autoimmune and neurodegenerative features. In MS, there is scar formation in the brain and/or spinal cord leading to various symptoms. The aetiology of MS is not known although several genes, mostly immune genes, have been found to be associated with the disease. MS was first described by J M Charcot in 1868 and has been traced back to 1822 (reviewed in 18) although symptoms associated with MS may be found in the Icelandic sagas: a woman named Halldora, in late 11th century, got cured from symptoms of paralysis and pain – a miracle attributed to Bishop Thorlak (reviewed in 19).

There is a broad spectrum of symptoms in MS, with great variation within as well as between individuals. Symptoms include visual disturbance, motor weakness, fatigue, cognitive impairment, coordination problems, bladder and bowel dysfunction and sensory abnormalities. Women are more likely to get MS than men (the female to male ratio is at least 2:1) a characteristic shared by other autoimmune diseases such as SLE and RA. Onset commonly occurs between 20 and 40 years of age and 80% of patients have a relapsing (disease bouts) and remitting (periods of recovery) form of MS (RRMS) at onset which over time usually moves into a progressive form of MS, secondary progressive MS (SPMS). In some instances, primary progressive MS (PPMS), progression without bouts is present at onset. Progression is dependent on age rather than initial course of disease based on the observation that there is no difference between SPMS and PPMS regarding age at onset of progression. Progression was defined as continuous worsening of neurologic symptoms with no connection to relapses for at least one year.

3.1 CLINICAL OUTCOME MEASURES

There is no outcome measure that doesn’t display some disadvantages, thus there are several measures in use. Amato and Portaccio categorized different measures of clinical outcomes into four main categories: objective neurological examination performed by neurologists, quantitative tests of neurological function, patient-oriented measures and hybrids. The first category (objective neurological examination) includes measures such as the Expanded Disability Status Scale (EDSS), the Scripps Neurological Rating Scale (SNRS) and the MS Impairment Scale (MSIS). The second category (quantitative tests) includes the MS Functional Composite (MSFC) while the third category provides the patients perspective and include measures such as Incapacity and Environmental Status Scale (ISS and ESS) and a wide range of quality of life (QOL) measures. The fourth category (hybrids) includes the Ambulation Index (AI), the Cambridge MS Basic Score (CAMBS) and counting relapses.

EDSS, the most widely used measure of disability, consists of scoring the findings of the neurological examination on eight mutually exclusive subscales, the Functional Systems (FS): pyramidal, cerebellar, brain stem, sensory, bowel and bladder, visual, cerebral and
The main advantage of this scale is its widespread use and thus familiarity, disadvantages included inter-examiner variability, intervals in the ordinal scale not being equidistant (mean staying time differs between levels of the scale), non-responsiveness to clinical change and focus mainly on ambulation while it doesn’t capture cognitive impairment or upper limb dysfunctions equally well (reviewed in 24).

The EDSS score of patients with comparable disease durations was assessed in nearly 10,000 MS patients and the distribution of EDSS was plotted according to an algorithm, the MS Severity Score (MSSS)27 producing a global MSSS score and allowing individual MS patients to be compared with others with regard to disability adjusted for duration. A patient that has a moderate EDSS score and short duration gets a similar MSSS score to one that has a high EDSS score but a longer duration.

3.2 DIAGNOSTIC CRITERIA

The criteria for making a MS diagnosis was updated in 200528, the McDonald criteria, and states that if the patient has had two or more clinical attacks (indicating dissemination in time) with symptoms indicating the presence of two or more lesions disseminated in space no further evidence is necessary albeit welcomed. Further evidence would be the presence of oligoclonal bands in cerebrospinal fluid (CSF) in conjunction with absence of these bands in serum, a high IgG index, positive visual evoked potentials and visualization of lesions with magnetic resonance imaging (MRI). This evidence is especially valuable when there are not two separate attacks or the attacks do not indicate dissemination in space, in accordance to specific recommendations28– still allowing for a diagnosis to be made.

The McDonald criteria differ from the previous diagnostic criteria, the Poser criteria, mainly in the incorporation of defined magnet resonance imaging (MRI) criteria and the addition of guidelines for diagnosis of PPMS29 but the basis for diagnosis remains dissemination in time and space of neurological symptoms typical for MS and exclusion of differential diagnosis.

3.3 TREATMENT

The main treatments offered to patients today are β interferons (Avonex, Betaferon, Rebif), glatiramer acetate (Copaxone) and the monoclonal antibody natalizumab (Tysabri); while not offering a cure for MS, these disease modifying drugs reduces the number and severity of clinical bouts and MRI lesions. As the effect of the treatment is on relapses these medications are offered to RRMS patients, currently there is no option available for treatment of the progressive forms of the disease, SPMS and PPMS. Future therapies might include oral immunomodulatory and neuroprotective drugs30.

3.4 PATHOGENESIS

MS is a chronic disease characterised by both autoimmune and neurodegenerative aspects. Symptoms arise as axons lose their insulating sheet, the myelin, in the central but not
peripheral nervous system. Within the CNS it appears that disease starts in the white matter and only later affects the cortex as cortical thickness is reduced primarily in patients with long disease duration\textsuperscript{31, 32}. The current view states that axonal loss is an early event that accumulates, with transition from RRMS to SPMS occurring when the CNS no longer is able to compensate for loss of function\textsuperscript{33}.

Autoimmunity is present in affected as well as unaffected individuals but in some individuals it’s associated with pathology and instead of autoimmunity there is autoimmune disease. Autoreactive T cells from MS patients have provided evidence of lower antigen specificity due to unconventional binding of the T-cell receptor (TCR) to myelin basic protein (MBP) in complex with human leukocyte antigen (HLA)-DR (reviewed in \textsuperscript{34}). This unconventional binding leads to cross-reactivity with other myelin antigens and provides a possible mechanism for molecular mimicry – one of the hypotheses behind the loss of tolerance leading to autoimmune disease. It is not clear where activation of these autoreactive myelin-specific T cells occur although it is often stated that activation starts in the periphery, enables the activated T-cells to cross the blood-brain-barrier (BBB) and proliferate in the CNS. Following production of pro-inflammatory cytokines microglia, macrophages and astrocytes are activated in the CNS and B-cells are recruited. This inflammatory process results in damage of myelin, oligodendrocytes and axons (reviewed in \textsuperscript{35}). Another view is that neurodegeneration is the primary process; oligodendrocytes, myelin and axons could be damaged due to glutamate toxicity or viral infection, asymptomatic to the host but with cytopathic effects (changes of morphology and/or metabolism) on target cells that could lead to prolonged exposure of neural antigens and consequent induction of an inflammatory response\textsuperscript{33}.

After demyelination there is a spontaneous effort to recover – there is no evidence that damaged myelin is “mended” rather, the tissue is often restored almost completely by the process of remyelination (the new myelin sheet is thinner and shorter than the original) (reviewed in \textsuperscript{36}). Interestingly, when demyelination is induced in an environment where the adaptive immune response is not activated (such as demyelination in the cuprizone-diet animal model of demyelination or following delivery of toxins such as ethidium bromide) remyelination is utterly effective. When the adaptive immune response is activated on the other hand chronic demyelination of axons might persist as a consequence of a, for oligodendrocytes, hostile environment; the process of remyelination is a process where new mature oligodendrocytes are generated from adult CNS oligodendrocyte progenitor cells (OPCs). That said, complete remyelination has been shown in MS and in experimental autoimmune encephalomyelitis (EAE), an animal model of MS, and the question of whether MS patients have impaired remyelinating capacities is not yet resolved.
3.5 A COMPLEX DISEASE

Despite family aggregation there is no apparent Mendelian inheritance pattern – that defines MS as a complex disease.

3.5.1 The genetic component

The existence of a genetic component in MS has been established by evidence that the recurrence risk declines as a function of decreased genetic sharing. Comparison of concordance rates between monozygotic (approximately 30 %) and dizygotic twins (approximately 5 %) indicates that genes are involved in MS\textsuperscript{37-40}. The increased relative risk among siblings is approximately 3 %, corresponding to a $\lambda_s$ of 15-20, corrected for age and with a prevalence of 0.1 %\textsuperscript{41,42}. While the results from studies of twins and sibling are in line with a heritable component, there is still the possibility that shared environment, rather than shared genes, explains the higher risk. Studies showing that 2\textsuperscript{nd} and 3\textsuperscript{rd} degree relatives are at higher risk of getting MS than the general population add to the picture from twins and siblings and makes an explanation by shared environment unlikely\textsuperscript{43}. Additional studies of adoptees, half-siblings and spouses\textsuperscript{44-46} make it clear that there is in fact a genetic component in MS. In addition, some ethnic groups, such as Sami and Norwegians, living in close proximity have strikingly different risks of disease\textsuperscript{47,48} although one might argue that it could be an argument for cultural factors (shared environment as a consequence of culture) rather than genes.

In conclusion, there is compelling evidence that genes do in fact influence susceptibility to MS.

3.5.2 The environmental component

The concordance rate among monozygotic twins even though higher than among dizygotic twins is still only about 30 % leaving room for the influence of environmental factors. Other evidence of environmental agents in MS includes migration studies where the risk of disease for individuals who moved from high risk areas to low risk areas before the age of 15 years declined (reviewed in\textsuperscript{49}). Moving from a low risk area to a high risk area, on the other hand, did not increase risk as expected. A possible explanation is that individuals get protected by environmental factors in low risk areas, a protection that remains throughout their lives but is not passed on to their children. Also, there have been several reports on an increased incidence of MS, which could only be explained by environmental factors\textsuperscript{50-53}, although no such increase has been noted in Sweden\textsuperscript{54}. Another reason to include environmental agents in the aetiology of MS is its uneven geographical distribution with an apparent latitude gradient, where northern Europe, southern Australia and North America belong to high prevalence areas while Africa and Asia are among the low prevalent areas – although this distribution could reasonably be attributed to a combination of genetic and environmental factors\textsuperscript{55,56}.
Several environmental agents have been proposed to be implicated in MS both non-infectious (e.g. vitamin D and sun exposure, smoking)\(^{47}\) and infectious (e.g. Epstein-Barr virus; EBV)\(^{49}\).

### 3.6 MS GENETICS

Traditionally patients with complex diseases have been subdivided into familial cases (~20 % of cases)\(^{58}\) and sporadic cases (do not share disease with a relative, as far as it is known). Familial and sporadic MS are believed to be occurrences of the same disease\(^{58}\), although there might be subgroups that differ as a recent study report disease onset of PPMS to occur at younger age in familial compared to sporadic MS\(^{59}\).

Family-based approaches could provide important clues on genetic variants that co-act to produce disease, thus with the potential of identifying genes in a single sufficient cause; this approach is difficult to pursue as there are not many extended pedigrees in MS. Several genomewide linkage analyses have been reported\(^{60-71}\), the best powered among them\(^{71}\) concluded that effect sizes ($\lambda_s$) of MS susceptibility genes are expected to be below 1.2; statistically significant linkage was found only for the most established MS locus, the HLA region ($\lambda_s$ =1.51). As linkage analysis would require unrealistic large sample size to detect such modest effects, association-based methods has been advocated as the method of choice\(^5\).

#### 3.6.1 MS genes

For over thirty years the only established MS gene was $\text{HLA-DRB1}^{72}$, more specifically the association is with a haplotype consisting of the following alleles: $\text{DRB1}^*1501$, $\text{DRB5}^*0101$, $\text{DQA1}^*0102$, $\text{DQB1}^*0602$ (HLA-DR15,DQ6)\(^{73}\). Approximately 60 % of MS patients and 30 % of controls carries the DR15,DQ6 haplotype (assessed through genotyping of the DRB1 locus) in our dataset. MS aside, the HLA-DR15,DQ6 haplotype is associated with risk for cataplectic narcolepsy (reviewed in \(^{74}\)) and protection against type I diabetes (reviewed in \(^{75}\)). In both mentioned diseases $\text{DQB1}^*0602$ is the allele of interest, whereas in MS DRB1 seems to be pivotal\(^{76}\) although a smaller study found evidence for the implication of $\text{DQB1}^*0602$ rather than $\text{DRB1}^*1501$ in MS susceptibility\(^{77}\). A regained interest for HLA Class I molecules have provided evidence for protection by $\text{HLA-A}^*02^{78, 79}$ and a possible signal from the $\text{HLA-C}$ locus\(^{80}\); more studies on the involvement of Class I genes are ongoing.

The first non-HLA association with MS to be widely replicated was $\text{IL7R}^{81-89}$. Mutations in the $\text{IL7R}$ genes leads to severe combined immunodeficiency syndrome (SCID), a condition characterised by failure to produce T lymphocytes, in humans\(^{90}\) as well as in mice\(^{91}\), thus $\text{IL7R}$ is critical for T-cell survival. In MS, evidence points to a functional role for rs6897932, a nonsynonymous SNP located in exon 6 of the $\text{IL7R}$ gene\(^{84, 85}\).
Following the publication of first SNP-based genomewide association study in MS\textsuperscript{86} several associations between immune-related genes and MS have been confirmed including \textit{IL2RA}, \textit{CD58}, \textit{EVI5}, \textit{CD226}, \textit{SH2B3} and \textit{CLEC16A}\textsuperscript{86, 88, 89, 92-97}. \textit{IL2RA}, just as \textit{IL7R}, is important for T-cell survival and a shift in the balance between soluble and membrane bound forms of these receptors has been proposed as the underlying mechanism by which both these genes lead to MS\textsuperscript{85, 98}. \textit{CD58} encodes LFA-3, a co-stimulatory molecule that participates in T cell receptor signalling by binding of CD2, the CD58 receptor\textsuperscript{99}. One possible mechanism of action in MS has been put forward by De Jager et al\textsuperscript{100}: down-regulation of \textit{CD58} expression leads to decreased FoxP3 expression in regulatory T cells with impaired suppression capabilities as a result\textsuperscript{101-104}.

Less is known about the functions and/or functional role of \textit{CLEC16A}, \textit{CD226}, \textit{SH2B3} and \textit{EVI5} in MS. \textit{CLEC16A} (also known as \textit{KIAA0350}) is a gene with a motif shared with C-type lectins expressed on B-cells, dendritic and natural killer cells. C-type lectins functions as adhesion receptors as well as pattern recognition receptors in the immune system\textsuperscript{105}. \textit{CD226} (also known as DNAX accessory molecule 1, \textit{DNAM1}) is involved in natural killer cell-mediated cytotoxicity as well as immune response mediated by Th1 cells\textsuperscript{106, 107}. Onset of the animal model of MS, EAE is delayed and disease is less severe as a result of anti-CD226 treatment\textsuperscript{106}. \textit{SH2B3} (also known as \textit{LNK}) is an adaptor molecule that act as a link between the TCR-CD3 complex and intracellular signalling molecules (MIM*605093, OMIM database at www.ncbi.nlm.nih.gov). \textit{EVI5} is a common site of retroviral integration in T-cell lymphomas of AKXD mice\textsuperscript{108}.

The protein coded by \textit{RPL5}, one of the components of the 60S subunit of ribosomes, has been implicated in other human diseases such as Diamond-Blackfan Anemia where haplo-insufficiency of other ribosomal proteins is believed to lead to apoptotic death of erythroid progenitors cells\textsuperscript{109} – while \textit{RPL5} is considered to be associated with MS\textsuperscript{86, 94}, the mechanisms by which MS susceptibility is affected is yet not known.
4 Materials and methods

4.1 PATIENTS AND CONTROLS

All patients included in this thesis had MS according to the McDonald criteria and/or the Poser criteria for definite MS. All in all 14 clinics were involved and over 4,700 patients and 5,000 controls (including the parents to the trio cases) contributed to this effort.

Cases in the Stockholm dataset are ascertained from the population of individuals in the catchment areas of three hospitals, Karolinska University Hospital Huddinge, Karolinska University Hospital Solna and Danderyds Hospital. Controls were consecutive blood-donors at three blood donation facilities in the Stockholm area. Cases and controls were of Swedish, Norwegian or Danish ancestry.

In paper I two populations were studied: the Stockholm dataset consisting of 1,119 patients and 1,235 controls and a Basque cohort consisting of 352 patients and 235 controls. Cases and controls in the Basque cohort were of Spanish and Basque ancestry.

In paper II the Stockholm dataset consisted of 1,114 patients and 1,235 controls.

In paper III 1,021 patients and 1,215 controls from the Stockholm dataset were included. In addition, one affected and one unaffected family member of a consanguineous pedigree was investigated for mutations in exons 16 and 17 of the C5 gene.

In paper IV 811 patients and 757 controls from the Stockholm dataset were investigated in study I. In study II 1,016 patients and 1,215 controls from the Stockholm dataset, including most individuals from study I, were investigated. In study III 1,168 additional patients and 656 controls from throughout Sweden were included.

In papers IV and VI a subgroup of patients are included in a larger epidemiological project (EIMS) and represent the whole country of Sweden. EIMS controls are geographically, age and sex-matched to the larger EIMS cohort.

In paper V three populations were investigated: 1,166 patients and 1,235 controls from the Stockholm dataset, 660 patients and 833 controls from Spain and 511 trio families from Finland.

In paper VI we collaborated with our Norwegian and Danish colleagues and 542 patients and 525 controls from Norway as well as of 508 patients and 538 controls from Denmark were included. The Swedish dataset consisted of 2,137 patients and 1,849 controls of which 2,197 individuals, 1,016 cases and 1,181 controls had been previously genotyped and were included in paper V. Thus, the Scandinavian cohort consisted of 3,187 patients and 2,912
controls. For each of the studied markers a combined P-value was calculated taking into consideration results from paper V based on a total of 4,358 patients and 3,745 controls.

In this thesis we utilize a case-control strategy where our cases are thought to represent the entire population of MS patients in a given geographical region and thus include both familial as well as sporadic cases. It should be noted however, that we have employed a hospital-based design and thus it is possible that the most benign and most severe cases are underrepresented in our datasets.

4.2 DNA EXTRACTION
Genomic DNA was extracted from leukocytes by one of three methods: salting out\textsuperscript{110}, QiAMP DNA Blood Maxi kit (Qiagen GmbH, Germany) or PureGene (Qiagen).

4.3 GENOTYPING
The genotyping methods (reviewed in \textsuperscript{111}) utilized in this thesis were either based on allele-specific primer extension (MALDI-TOF mass spectrometry, SNPstream and FP-TDI), allele-specific primer hybridization (TaqMan), a sequencing/enzymatic based method (pyrosequencing) or on size separation (fragment analysis).

4.3.1 Pyrosequencing
The genotyping performed in paper I for the Basque cohort utilized the pyrosequencing method for all but one marker according to the protocol provided by the manufacturer (Biotage, Charlottesville, VA). Primers and probes were designed using Biotage PSQ Assay Design software. Pyrosequencing involves an enzymatic cascade leading to emission and detection of light proportional to the number of nucleotides incorporated, only one type of nucleotide is added at a time.

4.3.2 TaqMan based allelic discrimination
In paper I rs155141 was genotyped by the TaqMan Assay-on-Demand kit (C\textunderscore 1276262\textunderscore 10) according to the manufacturers' instruction (ABI, Foster City, USA) in the Basque cohort. TaqMan allelic discrimination performed as in\textsuperscript{110} was the method of choice for genotyping of rs741072 and rs701421 for the expansion of the dataset (study III) in paper IV; the same applies to the two SNPs in paper VI.

4.3.3 MALDI-TOF mass spectrometry
Genotyping was performed at the Mutations Analysis Facility core facility at Karolinska Institutet using MALDI-TOF mass spectrometry (Sequenom Inc., San Diego, USA) of allele-specific primer extension products for the Scandinavian cohort in papers I – IV. Genotyping in papers I, II and IV (study I) was based on hME chemistry, that applies to four markers in study III as well. The iPLEX chemistry was used for nine markers in paper III and markers in
The spectroDESIGNER software was used to design multiplex SNP assays i.e. PCR and allele-specific extension primers. MassEXTEND reagents kit and allele-specific extension primers were used to produce the allele-specific extension primers that was analysed using a massARRAY mass spectrometer. The resulting mass spectra were processed and analysed using the spectroTYPE software. Calls were manually read by two persons independently. In papers III and IV assay validation was improved by addition of 14 trio families (42 individuals) which enabled check for Mendelian inconsistencies and concordance tests with published HapMap data. Internal concordance tests were also performed by genotyping a subset of individuals more than once.

4.3.4 SNPstream and FP-TDI
Allelic discrimination on the SNPstream platform (Beckman Coulter) is based on allele-specific primer extension; the same applies to MALDI-TOF mass spectrometry and fluorescent polarization template directed incorporation (FP-TDI). The main steps in any allele-specific primer extension method are 1) hybridization, 2) extension and 3) detection. SNPstream and FP-TDI are minisequencing methods as a single base is extended. FP-TDI differs from SNPstream in that detection is performed in solution by fluorescent polarization which only allows one SNP to be genotyped at a time. With SNPstream multiplexing is possible as fluorescence detection takes places on arrays; the extension primers are designed to contain tag-sequences complementary to sequences on oligonucleotides on an array allowing capture of extended primers prior to allele detection. rs4728142 was genotyped with the SNPstream as well as the FP-TDI (Analyst AD, Molecular Probes) method, all other SNPs were genotyped on the SNPstream system.

4.3.5 Fragment analysis
In papers V and VI the insertion/deletion polymorphism was genotyped by size separation on agarose gel or by capillary electrophoresis.

4.4 SEQUENCING
Exons 16 and 17 (in total 1,525 bases) of the C5 gene, which codes for C5a, were sequenced in one affected and one healthy member of a family with six affected individuals. PCR products were purified on Microcon 100 columns (Millipore, Bedford, USA) followed by sequencing reactions based on ABI PRISM Big Dye Terminator chemistry (Applied Biosystems, CA, USA). Sequencing products were detected on an ABI 377 sequencer (Applied Biosystems, CA, USA) and analysed using the Sequencing analysis software (PE Biosystems). Alignment was subsequently performed using ClustalW (www.ebi.ac.uk/Tools/sequence.html).
4.5 ELECTROPHORETIC MOBILITY SHIFT ASSAY (EMSA)

EMSA (also called gel shift assay) was performed to assess whether the studied polymorphisms affect protein-DNA interaction\textsuperscript{113}. The gel shift assay demonstrates the ability of proteins to bind to each sequence as DNA-protein complexes migrate more slowly on a gel compared to free DNA molecules. Double-stranded probes representing each allele were produced by allowing a 5’-biotin-labelled strand and an unlabeled strand to anneal. The labeled probes were incubated with nuclear extract prepared from PBMCs. To address specificity of binding we performed competition experiments where a 100-fold molar excess of unlabeled probe was added to the incubation. The DNA-protein complexes produced in the binding reactions were analyzed using electrophoresis on polyacrylamide gels. After the electrophoretic separation, the biotinylated fragments were transferred to membranes, and detected by a chemiluminescent procedure using the LightShift\textsuperscript{®} Chemiluminescent EMSA kit.

4.6 PROXIMITY LIGATION ASSAY (PLA)

The 4x allele of the CGGGG insertion/deletion marker was found to bind more protein compared to the 3x allele. We further explored this protein-DNA binding by performing PLA\textsuperscript{114}. This technique is based on two bi-functional probes: one of the probes is a labeled antibody directed against the protein of interest that has been conjugated to oligonucleotides. Here we used biotinylated polyclonal antibody against the SP1 protein, combined with a streptavidin–oligonucleotide conjugate. SP1 was chosen for study as it was indicated to bind to the CGGGGCGGGG sequence (Yutaka Akiyama: "TFSEARCH: Searching Transcription Factor Binding Sites", http://www.rwcp.or.jp/papia/). The other probe consisted of a partially double-stranded DNA sequence containing the SP1 binding site. Here we used a DNA probe containing the polymorphic CGGGG repeat purified by high pressure liquid chromatography (HPLC). The HPLC-purified probe was made partially double stranded as described by Gustafsdottir et al\textsuperscript{114}. Since the SP1 protein was simultaneously bound by both probes, the oligonucleotides ends of the probes were brought physically close together and were thus able to hybridize to an added connector oligonucleotide. This DNA structure was then covalently joined by enzymatic ligation. The ligated DNA sequence, which serves as a representation of the binding event between SP1 and the CGGGG repeat, was then amplified and detected by real-time PCR.

4.7 EXPRESSION ANALYSIS

In paper VI, gene expression studies were performed on samples (blood and CSF) collected during 2002-2007 at Karolinska University hospital.

4.7.1 Preparation of PBMC and CSF-MC

CSF samples were obtained when lumbar puncture procedure was considered from a clinical perspective, with informed consent from the patients. Samples were collected on siliconized glass tubes and immediately centrifuged; the pellet was recovered and stored at
-70˚C until use. Peripheral blood was collected into sodium citrate-containing cell preparation tubes (Vacutainer CPT, Becton Dickinson and Company). PBMCs were separated by density gradient centrifugation and pellets were stored at -70˚C.

4.7.2 mRNA and cDNA preparation
Total RNA was extracted from lysed cell pellets (PicoPure RNA isolation kit, Arcturus Bioscience, USA) and purity of samples was determined using the Agilent 2100 Bioanalyzer (Agilent Technologies, USA). cDNA was produced by reverse transcription PCR using 1-5 ng (10 μL) total RNA template, 0.1μg random hexamers (Gibco BRL) and 200 U of Superscript Reverse Transcriptase (Gibco BRL).

4.7.3 Quantitative real-time PCR
Primers were designed for the target (IRF5) and the reference gene (GAPDH) with the Primer Express Software (Perkin Elmer). Real-time PCR was performed using a BioRad iQ™ iCycler Detection System (Bio-Rad Laboratories, Ltd). All samples were run in duplicates; a 10-fold dilution standard curve was present on each plate. The PCR efficiency was between 90 and 105 % in all PCR runs. Quantification of the relative amount of RNA was done by a variant of the comparative method, instead of calculating the individual fold-changes ($2^{-ΔΔCt}$) and then comparing groups by fold-change we compared the groups by relative gene expression ($2^{-ΔCt}$) and then calculated fold-change by taking the ratio of the geometric mean in each group to the geometric mean among controls with other neurological diseases (OND).

4.8 STATISTICAL ANALYSES

4.8.1 Posterior odds and false positive report probability
Whenever inference on populations is made through the study of samples there is a risk for false findings. We have applied calculations of false positive report probabilities (FPRP), a Bayesian approach, in an attempt to control the type I error rate (false positive rate). When calculating the FPRP, odds rather than risk is used as it simplifies the mathematical calculations. Posterior odds can easily be converted to FPRP using the formula: $[1/(1+\text{posterior odds})]$. The posterior odds is calculated as $[(\text{power/significance level})\times\text{prior odds}]^{13, 116, 117}$ and is a measure of how certain you are that your findings are true based on several parameters. Three of these parameters are incorporated into the power estimation: sample size, size of the effect conveyed by the variant under study and the risk allele frequency. Consideration is also taken to the plausibility of the inference based on prior knowledge (prior odds) and on how conservative you are when making a statement that a finding is statistically significant (significance level).

We decided that a finding was worth reporting (i.e. stated as a statistical significant finding) when it was more likely to be true than false (FPRP=40 %). We believe this to be reasonable given our sample sizes and the modest effects that we expect to find.
In order to calculate the appropriate significance level to reach a chosen posterior odds we need to make a statement about the prior odds of finding one true association in the investigated genes. It is not possible to know the prior odds, estimations need to be made in accordance to the study design: for candidate gene studies estimations are made based on the strength of evidence from previous studies linking the gene to disease; for genomewide association studies estimation is made based on how many true findings are believed to exist in total in a given disease among the estimated $1 \times 10^{-6}$ independent markers in the genome. It is however very difficult to estimate the prior odds from biological evidence, therefore we investigated the impact of several different prior odds while inference was based on a prior odds of 0.01 – we expected to find 1 true associated genetic variant among 100 markers.

Calculations on significance levels ($\alpha$) were performed by calculating power for an arbitrarily chosen $\alpha$ and verifying what FPRP that would imply for a prior odds of 0.01. The procedure was repeated until a significance level was found which led to approximately 40 % FPRP.

### 4.8.2 Allelic association measures

Allelic association implies that alleles are found together in gametes more often than would be expected under random segregation. That would occur, for instance, when there is no recombination between alleles on the same chromosome which are then passed on to a gamete together – this form of allelic association is called linkage disequilibrium. The degree of LD is influenced by the rate of recombination between loci, but also by genetic drift, mutations, migration, population expansion and selection\(^\text{118}\). The most commonly used measures of linkage disequilibrium\(^\text{119}\) are $D'$ and $r^2$. Both of these measures are based on the LD coefficient, $D$, which is defined as:

$$D = f(AB) - f(A) \times f(B)$$

Where $f(AB)$ represents the observed haplotype frequency and $f(A) \times f(B)$ the expected frequency under random segregation for two biallelic loci with alleles $A$ or $a$ and $B$ or $b$ respectively. Thus, $D$ gives a measure of the excess/deficit of haplotypes containing both $A$ and $B$. The value of $D$ depends on the allele frequencies, to overcome that normalised measures are used – $D'$ and $r^2$ are two such normalised measures.

$$D' = \frac{D}{D_{\text{max}}}$$

$$r^2 = \frac{D^2}{(f(A) \times f(a) \times f(B) \times f(b))}$$

$D_{\text{max}}$ is the maximum of $D$ given the allele frequencies at the two loci, thus $D'$ represents the proportion of the maximum amount of LD between two loci. $D'$ ranges from -1 to +1, when $|D'|=1$ there is no recombination between the least common allele and the other locus. $r^2$ ranges from 0 to 1, when $r^2=1$ the two loci are in perfect LD and one allele at the first locus
exists ONLY with a particular allele at the other locus. When conducting a genetic association study $r^2$ is relevant when selecting markers to type – markers in perfect LD are redundant and only one of them needs to be genotyped. Moreover, an observed marker allele might be in LD with a disease allele, to achieve the same power at the marker locus as if the actual disease mutation had been genotyped the sample size needs to be increased by a factor of $(1/r^2)$.

### 4.8.3 Power
Power calculations were performed using the web-based Genetic Power Calculator\textsuperscript{120} or CaTS Power Calculator\textsuperscript{121}.

### 4.8.4 Single point association analysis
Chi-squared tests or logistic regression was performed to compare genotype distribution, carriage of an allele or allele frequencies between cases and controls as implemented in PLINK\textsuperscript{122}, EpiInfo Statcalc (Centers for Disease Control and Prevention, Atlanta, GA) or the SNPassoc package\textsuperscript{123} in R.

### 4.8.5 Adjusting for other factors
Logistic regression was used to adjust estimated effects (odds ratio) of one marker to previously known risk factors using the glm function in the R software.
4.8.6 Combining P-values
In paper V, P-values from case-control studies were combined with P-values from a family-based (trio families) approach with the analytical formula defined by Lou Joust (www.loujost.com) which is an extension to Fisher’s method.

\[
\text{Combined P-value} = k \sum (-\ln (k))/i!
\]

Where k is the product of the set of P-values and \( \sum \) goes from \( i=0 \) to \( n-1 \).

4.8.7 Haplotype association analysis
Haplotype analysis was performed with the Haploview software\(^{124}\). In paper V a sliding window haplotype analysis was performed using PLINK\(^{122}\).

4.8.8 Disease severity association analysis
Association of genetic variants with disease severity was assessed either with MSSS\(^{27}\) or by survival analysis on time to as well as age at EDSS 6 using Cox regression models as implemented in Stata software version 9.1 (StataCorp, College Station, Texas, USA).

4.8.9 Comparing expression levels
Wilcoxon unpaired test was used to test for differences between different groups in relative amount of \( IRF5 \) cDNA as implemented in the software R. In order to achieve equal variances we used log-transformed data, Bartlett’s test was used to validate our assumption that variances were equal between groups (bartlett.test function in R). Test of adaptation to a normal distribution was performed using both Shapiro-Wilk normality test and D’Agostino & Pearson omnibus normality test, GraphPad Prism software v.5.

4.8.10 Genotype-phenotype correlation
Correlation between genotype and expression levels of \( IRF5 \) was assessed using ordinal logistic regression with the function lrm implemented in the Design package for R software.
5 Results and discussion

Figure 3 shows P-values for all analysed markers in the eight genes investigated in this thesis; the number of included individuals varies among studies.

![Figure 3: Single point association results (P-values) for all markers analysed in this thesis. Note that sample size varies across different studies. The dotted line represents a P-value of 0.05. Markers are not ordered according to physical position.](image)

In interpreting these results the power of the study (which is dependent on estimated effect of the investigated variant, the allele frequency, chosen significance level and number of patients and controls) as well as other evidence in favour or against an association (the prior probability) needs to be taken into account. When considering all these factors I believe that the association of rs741072 (in the NGFR gene) along with the three associated markers in the IRF5 gene (rs4728142, CGGGG and rs3807306) warrants further study as will be discussed in sections 5.4 and 5.5.

5.1 PAPER I

α4 integrin (ITGA4) codes for the α4 subunit of VLA-4, an adhesion molecule composed of two subunits: α4 and β1. ITGA4 is a strong candidate for involvement in MS pathogenesis as evidenced by various publications including animal studies, expression analyses and from the effects of treatment offered to relapsing-remitting MS patients (Tysabri®). Yednock et al. studied how several different antibodies against adhesion molecules affected how leukocytes migrate through inflamed EAE brain vessels in vitro. Only antibodies against the α4 subunit of the VLA-4 receptor inhibited binding of leukocytes (monocytes and lymphocytes) to the EAE vessels. In vivo studies demonstrated that administration of anti-α4 antibodies prevented infiltration of inflammatory cells and development of EAE (paralysis was prevented in 75 % of treated animals). Iglesias et al. performed an expression profiling study on 17 patients and 7 healthy controls and found expression of VLA-4 to be up-regulated in PBMCs of MS patients. Jensen et al. investigated the impact of IFN-β1b treatment on expression of VLA-4 in CD4+ and CD8+ cells and found expression of VLA-4 to be down-regulated by treatment in both CD4+ and CD8+ cells.

In paper I we investigated 11 SNPs in the ITGA4 gene in two populations – a Nordic population consisting of 1,119 patients (73% female) and 1,235 controls (63% female) and a Basque cohort consisting of 352 patients (73% female) and 235 healthy controls (52% female).
Markers were selected based on their distribution throughout the gene; markers with minor allele frequency below 10% were not considered for inclusion.

The most interesting result in this study was an allelic association between the C allele of rs1449263 located in the promoter of ITGA4 and MS, although we lack power to exclude the possibility of a false positive finding.

Several markers displayed deviation from Hardy-Weinberg (HW) proportions, among them rs1449263. For this marker, Basque controls deviate from HW proportions (increased homozygosity) while Basque cases do not. If there were a systematic genotyping error we would expect both cases and controls to deviate from HW proportions as each genotyping plate contained samples from cases as well as controls. In addition, 188 Basque cases and 176 Basque controls were genotyped with two genotyping methods (pyrosequencing and allele-specific primer extension+MALDI-TOF mass spectrometry) and the genotype concordance between these two methods was 99.9% for rs1449663. Genotyping error does thus not seem to be the cause of deviation from HW proportions in controls in the Basque population. Still, in the Nordic dataset we did not observe deviation from HW proportions therefore selective pressure does not seem a likely explanation either, unless genotypes at rs1449263 reflects selective pressure of an environmental agent not common to both Basque and Nordic populations. Another source of deviations from HW proportions is population stratification – in a study of genetic ancestry in Europeans, Basque (n=8) and Valencian Spanish (n=20) individuals cluster next to each other although the two different groups are clearly identifiable. Our Basque dataset includes a mixture of Basque and Spanish ancestry; the mixture of ancestries applies to both cases and controls and we do therefore not expect population stratification to be the cause of deviations from HW proportions, although we have not performed any formal test and thus cannot exclude this explanation. Perhaps the simplest explanation for rs1449263 not displaying HW proportions is the impact of a chance event in this particular sample, i.e. sampling error. In any case, the discussion that now follows is based solely on the results from the Nordic dataset.

The C allele of rs1449263 was found to confer risk of MS with OR=1.14, P=0.04 based on 987 cases and 1,144 controls. The P-value for the genotypic test was 0.098, thus not statistically significant. Still, it seems in line with a potential association that would require larger sample sizes to pass significance cut-offs as the genotypic test provides lower power compared to the allelic test. A haplotype analysis of the markers included in a haplotype block as defined by Gabriel et al indicates that rs1449263, the second marker in the block (in bold in table 1) is the most interesting of the observed markers to consider for further investigation – whenever the C allele is present the haplotype is more frequent in cases compared to controls; the fact that these haplotypes do not differ statistically between cases and controls might be taken as an indication that a single marker, rather than a haplotype could be associated with disease.
Table 1: Haplotype analysis using Haploview v4.1. a) Markers are, from left to right: rs155141, rs1449263 (in bold), rs3770138, rs3770136 and rs3770132. b) Relative haplotype frequency. Bold frequencies indicate that the haplotype is more frequent among cases, underlined frequency indicates that the haplotype is more frequent among controls.

<table>
<thead>
<tr>
<th>Haplotype&lt;sup&gt;a&lt;/sup&gt;</th>
<th>MS&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Controls&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Chi Square</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATGCA</td>
<td>0.54</td>
<td>0.57</td>
<td>3.70</td>
<td>0.05</td>
</tr>
<tr>
<td>ACATA</td>
<td>&lt;b&gt;0.18&lt;/b&gt;</td>
<td>0.18</td>
<td>0.03</td>
<td>0.86</td>
</tr>
<tr>
<td>ACGCA</td>
<td>0.14</td>
<td>0.13</td>
<td>0.78</td>
<td>0.38</td>
</tr>
<tr>
<td>TGGC</td>
<td>&lt;b&gt;0.13&lt;/b&gt;</td>
<td>0.11</td>
<td>2.63</td>
<td>0.11</td>
</tr>
</tbody>
</table>

The P-value of the allelic test, P=0.04, is too modest to be considered a true finding in terms of P-values advocated for candidate gene association studies (10<sup>-5</sup> - 10<sup>-6</sup>). Still, given the strong potential involvement of this gene in MS, larger studies including rs1449263 as well as other markers (figure 4) are warranted.

Figure 4: ITGA4 with 10kb of 5’ flanking region and 10kb of 3’ flanking region, HapMap Phase II+III SNPs with MAF > 0.05 generated in Haploview beta 4.2. LD colour scheme=\(r^2\). A tagging approach based on these set of SNPs would lead to investigation of 36 SNPs capturing 116 alleles (pairwise tagging , \(r^2=0.9\)). Investigated SNPs in paper I; the position of rs1449263 is indicated by a vertical arrow.

Future studies will surely include genome wide association studies with large sample sizes. In fact, four of the SNPs included in paper I were also included in a genomewide association study performed by the International MS Genetics Consortium: rs155141, rs3770132, rs1038034 and rs3770111. None of these SNPs were taken to the second stage of investigation even though the P-value for rs3770111 was 0.099 in one of the datasets (931 cases and 2,431 controls) used in the first stage of the IMSGC study. 9 of the 13 SNPs studied in paper I (one of them being rs1449263) are included in the Illumina 1M Beadchip – the chip of choice of an ongoing international effort to evaluate common
variants in MS – thus it may not be long until we get a more definite answer as to if ITGA4 is implicated in MS.

5.2 PAPER II

Several genes have been found to be associated to more than one autoimmune disease, examples includes several HLA genes, CTLA4, PTPN22 and more recently IL2R, IL7R, CLEC16A, CD226 and SH2B3. The IL23R gene, located on chromosome 1p31.3 was found to be associated with inflammatory bowel disease (IBD)148, an autoimmune disease. In addition, IL23R is one of the genes included in the EAE27 locus, a region found to be associated with EAE (Tomas Olsson, personal communication). We set out to investigate whether IL23R played a role also in MS susceptibility in a dataset consisting of 1,114 MS patients and 1,235 controls.

SNPs were selected based on genotype data from HapMap Phase II using a tag SNP approach including the 10 IBD SNPs and SNPs in the 3’ adjacent gene IL12RB2. After quality control (success rate > 85 %, HWE P>0.0001 in controls) 32 SNPs were included in the analysis.

One SNP, located in intron 1 of the IL12RB2 gene had nominal P-value of 0.02 – although statistically significant the probability of this finding being a false positive is very high for most of the prior odds tested. False positive report probabilities (see Material and Methods) for several prior odds are depicted in table 2 for an OR of 1.2.
Table 2: False positive report probability for various prior probabilities and significance levels (including the P-value in the single point analysis, \( P=0.02 \)).

<table>
<thead>
<tr>
<th>Prior</th>
<th>1.2</th>
<th>1.2</th>
<th>1.2</th>
<th>1.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>26%</td>
<td>16%</td>
<td>3%</td>
<td>1%</td>
</tr>
<tr>
<td>0.1</td>
<td>47%</td>
<td>33%</td>
<td>8%</td>
<td>3%</td>
</tr>
<tr>
<td>0.01</td>
<td>90%</td>
<td>83%</td>
<td>46%</td>
<td>21%</td>
</tr>
<tr>
<td>0.001</td>
<td>99%</td>
<td>98%</td>
<td>89%</td>
<td>73%</td>
</tr>
<tr>
<td>0.0001</td>
<td>100%</td>
<td>100%</td>
<td>99%</td>
<td>96%</td>
</tr>
<tr>
<td>0.00001</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

\[ \alpha \]

\[ \text{power} \]

\[ 0.56 \quad 0.41 \quad 0.12 \quad 0.037 \]

Since no other significant associations were found between the 31 SNPs in \( IL23R \) and MS we conclude that \( IL23R \) is not involved in MS susceptibility in our population, at least with effect sizes matching those in IBD.

5.3 PAPER III

Plenge et al\(^\text{149}\) reported that associations with RA had been found within the \textit{TRAF1/C5} region in a genomewide study of RA. Following the initial finding the region was fine-mapped using nine markers where rs3761847 and rs2900180 showed the strongest association signals with RA. It was not possible to discern whether the association signal came from \textit{TRAF1} or \textit{C5} because of high LD in the region.

In paper III we analysed the nine SNPs used in the fine-mapping by Plenge et al for association with MS. We chose to analyse all nine markers rather than just the ones reported to be associated with RA as there could be different markers involved in MS compared to RA, similar to what has been observed for \( IL2RA \) in MS and type I diabetes\(^\text{150}\). In addition, we sequenced exons 16 and 17 of the \textit{C5} gene, which codes for the anaphylatoxin C5a in two individuals of a consanguineous pedigree, one affected and one unaffected. \( C5 \), one of several genes found under a suggestive linkage peak in this family\(^\text{151}\), drew our attention due to the potential involvement of complement in MS (reviewed in \( ^{152} \)). Complement components have been found in higher concentrations in the cerebrospinal fluid (CSF) of MS patients compared to controls\(^\text{153-158} \). Moreover, complement activation products are present within white matter lesions although not in purely cortical lesions\(^\text{159-162} \). C5a has also been suggested to be involved in neuroprotection as well as apoptosis of neurons (reviewed in \( ^{163,164} \)).

The intracellular protein encoded by the \textit{TRAF1} gene mediates tumor necrosis factor (TNF) signalling. TNF is an inflammatory and pro-survival molecule that is associated with MS\(^\text{165-168} \).
Since none of the nine markers were statistically significant a discussion of the risk of a false negative report is warranted. For a sample size of 1,019 cases and 1,215 controls, \( \alpha=0.05 \) and under a multiplicative model our power to detect an association (given that there was in fact an association to detect) was 99.6 \% for rs3761847 and 99.8 \% for rs2900180. If we instead used \( \alpha=1 \times 10^{-5} \), power would be reduced to 57.8 \% for rs3761847 and 65.4 \% for rs2900180. We found no association even at the 0.05 level - we conclude that rs3761847 and rs2900180 are not associated with MS with effect sizes of 1.32/1.34 respectively or higher. As shown in table 3, power changes only slightly for changes of \( \alpha \) from 0.1 to 0.001 (from 99.9 \% to 90.7 \%) whereas we would only have approximately 58 \% power at \( \alpha=1 \times 10^{-5} \) to detect an association with \( \text{OR}=1.32 \) if the frequency of the risk allele was 47 \% in the population. Applying a significance level of \( \alpha=1 \times 10^{-5} \) would allow us to be confident of a positive finding being true had there been one, which was not the case.

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>Power</th>
<th>N cases for 80% power</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.999</td>
<td>295</td>
</tr>
<tr>
<td>0.05</td>
<td>0.996</td>
<td>375</td>
</tr>
<tr>
<td>0.01</td>
<td>0.979</td>
<td>559</td>
</tr>
<tr>
<td>0.001</td>
<td>0.907</td>
<td>817</td>
</tr>
<tr>
<td>0.0001</td>
<td>0.765</td>
<td>1072</td>
</tr>
<tr>
<td>1.00E-05</td>
<td>0.578</td>
<td>1323</td>
</tr>
</tbody>
</table>

Table 3: Power analysis for risk allele frequency of 0.47 and genotypic relative risk equal to 1.32 under a multiplicative model for 1,019 cases and 1,215 controls. Analysis was undertaken using Genetic Power Calculator\(^{120}\). a) Number of cases needed in order to achieve 80 \% power for given parameters (including controls:cases ratio of 1.19).

As for the other investigated SNPs, five hade allele frequencies below 10 \% (rs10985095, rs12338903, rs10985097, rs7035682 and rs10985112); rs2269066 had an allele frequency of 12 \% in our control group and rs7026551 an allele frequency of 22 \%. Given these allele frequencies and our sample size – what effect sizes did we expect to be able to detect? For risk alleles with a frequency of 10 \% we had at least 80 \% power (power=84 \%) to detect an association with effect size \( \text{OR}=1.32 \), under a multiplicative model and at the 0.05 level; for an allele frequency of 20 \% we would achieve at least 81 \% power to detect an effect size of \( \text{OR}=1.23 \). This gives a point of reference as to what effect sizes that could be detected in this study –leaving room for association with MS with lower effect sizes and/or other genetic models.

As for the sequencing of exons 16 and 17 of the C5 gene – we found no difference between the affected and unaffected family members with regard to sequence of the C5a exons, i.e. no mutation was found and we conclude that C5a sequence variants is not the cause of MS in this family. In fact, in a recent reinvestigation of the family, with higher marker density
and two additional individuals, the suggestive linkage peak in chromosome 9 was not detected (Imrell et al, unpublished).

In conclusion, we found no association of nine SNPs in the TRAF1/C5 region in a case-control association study and no mutation in exons 16 and 17 of C5 (which encode C5a) that explained MS in a family with 6 affected individuals.

5.4 PAPER IV

50 SNPs in four genes (RTN4R, LINGO1, NGFR and TNFRSF19) were analysed for association with MS. All SNPs were investigated in a case-control dataset of Nordic origin although most SNPs in a subset of these individuals: 20 SNPs were analysed in 811 patients and 757 controls (study I), 28 SNPs were analysed in 1,016 patients and 1,215 controls (study II) and 2 SNPs (rs741072 and rs701421) were analysed in 2,108 patients and 1,871 controls (study III). The subsets are not independent; it reflects the growth of our material during my PhD studies.

Analysis after the first stage of genotyping (studies I and II) revealed two markers with P-values below 0.05: rs701421, located in the first and only intron of RTN4R and rs741072, located in the untranslated region of exon 6 in NGFR. No association signal was detected for the remaining 48 markers, neither with single point nor haplotype analysis.

rs701421 and rs741072 were first genotyped in study I and II respectively where significant (α=0.05) associations were found with the T allele of rs741072 (OR=0.82, P=0.001 for the log-additive model; P=0.005 for the co-dominant model) and the T allele of rs701421 (OR=1.23, P=0.005 for the log-additive model; P=0.02 for the co-dominant model). These two SNPs were then genotyped in additionally 1,065 patients and 656 controls. After testing for homogeneity, the genotypes were combined with data from study II (for rs741072) and study I (for rs701421) prior to analysis.

Based on our sample size and the expectation that effect sizes would be modest we decided that a finding that was more likely to be true than false would be noteworthy\(^1\). Therefore, we aimed at a false positive report probability (FPRP)\(^\text{1}^\) no higher than 40 %. The corresponding significance level (α) was assessed by choosing an arbitrary value for α, calculating power and estimating the FPRP for various prior probabilities. Calculations were performed until a significance level, α=0.006, was found that would provide a FPRP ≤ 0.40 for a prior probability of 0.01 (see section 5.4.1).

For rs701421 the association signal was lost in the expanded dataset (P=0.04) while the signal at rs741072 remained statistical significant at α=0.006 (OR=0.86, P=0.001).
The major allele (C) constitutes the risk allele with an OR of 1.18 and P=0.0026 when fitted in a model together with sex, *HLA-DRB1*15, *HLA-A*02, rs6897932 (*IL7R*) and rs12708716 (*CLEC16A*).

The haplotype analysis is consistent with a risk effect of the C allele of rs741072, haplotypes carrying the risk allele (C) are more frequent in cases compared to controls while haplotypes carrying the T allele are more frequent in controls. The association of haplotypes including rs741072 was not stronger than what was observed for rs741072 on its own implying that rs741072 or a marker in LD with it but outside the haplotype block is the causative allele – under the simplifying assumption that haplotype blocks as delineated by the algorithm of Gabriel et al. depicts true haplotype blocks.

### 5.4.1 A false positive finding?

There have been plenty of reports of associations in the history of complex genetics; many of these were never replicated. Several reasons for this failure to replicate have been proposed, the main being lack of power in the initial study. When I started working on this thesis much attention was paid to multiple testing corrections as a way of avoiding false positives, i.e. reports that would fail to be replicated. Many methods of correction were proposed e.g. Bonferroni correction and permutation but no consensus was reached on which method that was preferable or even whether correcting P-values was the right way to go. Had I performed this analysis at that time my conclusion would certainly have been that no associations were found because no association would persist after correcting for multiple testing. In 2007 the WTCCC published their influential report presenting genomewide association results for seven diseases; there they argued that the aim of multiple testing correction is to correct for multiple tests of a single “global” null hypothesis – in genomewide association analyses, as in SNP-based candidate gene association analyses, we don’t test a single global hypothesis instead we test several different hypotheses, one for each SNP (for biallelic SNPs). The Bayesian approach was proposed to be the method of choice for controlling type I error rate. Rather than just choosing an arbitrary threshold (commonly 0.05) which does not provide you with necessary information to interpret your finding, the choice of significance level (P-value threshold) is decided based on the level of certainty with which you want to claim a finding to be a true association rather than a false positive. That level of certainty is represented by the posterior odds – the odds that the finding is true. E.g. a posterior odds of 10:1 translates into a 9 % probability of a reported association being a false positive; posterior odds of 20:1 implies a 5 % risk of a false positive finding. Given our sample size and the modest effect sizes (OR) we expect to find we decided that an association that was more likely to be true than false was worth reporting. Therefore, when setting significance levels, a posterior odds of 1.5:1, which translates into a 40 % risk of a false positive was aimed at.
By allowing us to express the basis for our inferences in terms of parameters (power, significance level, prior odds and posterior odds) we are able to visualize, to ourselves just as much as when discussing with others, the impact of each parameter in relation to the others. For instance, today no one would argue that rs6897932, located in the IL7R gene is not truly associated with MS – even so, in the MS association screen performed by the IMSGC the P-value for the risk allele (C; allele frequency 0.75) in 931 cases and 2,431 controls was just 0.017 (OR=1.17). Contrast this with the association between rs741072 (risk allele C; allele frequency 0.57) located in the NGFR gene and MS, based on 2,021 cases and 1,760 controls and with P-value for the single point analysis equal to 0.001 (OR=1.16). Compared to the IMSGC results on IL7R, how would you judge the probability of NGFR being a true finding? Well, it depends on power and on your prior beliefs of finding an association in the investigated area. Calculating the power (for \( \alpha=0.05 \) to allow for the IL7R result to be significant) of detecting the IL7R association with above settings yields a power of approximately 68 \%; for the NGFR association test power is approximately 89 \%. The false positive report probability is shown in table 4 for various prior odds. If there is equal prior probability of either one marker being associated with disease there won’t be a large difference in the risk of the two findings being a false positive (high in this example as \( \alpha \) was set to 0.05) – yet the IL7R association turned out to be true, in fact incorporation of evidence from previous studies into the prior odds lowers the risk of IL7R being a false positive. The problem is – there are no previous studies on rs741072 and MS. Therefore, our choice of prior probability relies on a belief that the biological functions of NGFR translates into one associated marker out of all available variants in the four investigated genes (expected to be in the order of \( 10^2 \) based on available HapMap data).
### Table 4: Comparison of false positive report probabilities

The IL7R SNP (rs6897932) and the NGFR SNP (rs741072) for various prior probabilities.

<table>
<thead>
<tr>
<th>Prior odds</th>
<th>OR</th>
<th>1.16</th>
<th>1.17</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>5%</td>
<td>7%</td>
</tr>
<tr>
<td>0.1</td>
<td></td>
<td>36%</td>
<td>42%</td>
</tr>
<tr>
<td>0.01</td>
<td></td>
<td>85%</td>
<td>88%</td>
</tr>
<tr>
<td>0.001</td>
<td></td>
<td>98%</td>
<td>99%</td>
</tr>
<tr>
<td>0.0001</td>
<td></td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

- **power**: 0.89 0.68
- **α**: 0.05 0.05
- **allele fqy**: 0.57 0.75
- **N cases**: 2,021 931
- **ctrls:cases**: 0.87 2.61

**In conclusion**, the association of rs741072 with MS is more likely to be true than false (60% probability of a true finding). We don't have the necessary power to make a more conclusive statement, nonetheless such finding have baring on future studies – on future statements of the prior odds of rs741072 being involved in MS.

### 5.4.2 NGFR in MS

Expression of NGFR in glial cells within white-matter lesions of MS has been reported to be up-regulated, although oligodendrocyte death could not be attributed to high NGFR expression\(^\text{172}\). NGFR has also been studied in animal models of MS; in EAE NGFR knockout mice developed more severe or even lethal disease with a higher proportion (doubled compared to EAE wild-type mice) of T-cells in the cellular infiltrate that invades CNS following induction of EAE\(^\text{173}\). In another animal model of MS, the cuprizone-diet model, where oligodendrocyte death is studied in the absence of T-cell infiltration, induction of disease caused expression of NGFR but, as in\(^\text{172}\), oligodendrocyte death seemed independent of NGFR expression. NGFR is a pleiotropic TNF receptor family member, if not acting in MS through oligodendrocyte death, several other mechanisms are possible due to the versatility of this protein. In addition to its role as co-receptor for myelin-associated inhibitors of neurite outgrowth together with nogo receptor (encoded by RTN4R) and leucine-rich repeat neuronal 6A (encoded by LINGO1), NGFR has been implicated both in pro-survival and pro-apoptotic signalling\(^\text{174-176}\) (reviewed in\(^\text{177}\)). Moreover, NGFR is involved in the weakening of synaptic connections, so-called long-term depression (LTD)\(^\text{178,179}\) and has been implicated in the control of tissue fibrosis\(^\text{180}\) (when extravascular fibrinogen is deposited in tissues after vascular rupture). Tissue fibrosis has been reported to not be just a marker of inflammation but also to mediate disease (reviewed in\(^\text{181}\))
The potential to use NGFR as a therapeutic target has been investigated because of its involvement in several patophysiological states (summarized in 182) but also because it is possible that modulation of NGFR expression would promote neurogenesis 183 and thereby be of importance in therapeutic intervention for any disease in which neurons are damaged. Several different nerve growth factor-targeting strategies have been proposed including the use of small modulating molecules 182 and direct targeting of nerve growth factor using peptides (reviewed in 184).

5.5 PAPERS V-VI
Just as was the case for papers II (IL23R) and III (TRAF1/C5 region) in papers V and VI we investigate a candidate gene shown to be involved in other autoimmune diseases. The interferon regulatory 5 (IRF5) gene is associated with systemic lupus erythematosus (SLE) 185 and with rheumatoid arthritis 186; there are also reports on associations with IBD 187 and primary Sjögrens syndrome 188 although sample sizes in these two studies are limited.

We investigated 10 genetic markers for association with MS, nine SNPs and one insertion/deletion polymorphism (CGGGG). Markers were selected based on previous reports of association with other autoimmune diseases and/or possible modulation of IRF5 expression. For rs4728142, rs3807306 (P<0.05 in all three datasets) and CGGGG (P<0.05 in both case-control datasets, P=0.06 in the trio dataset) association were all in the same direction, i.e. the same allele was more frequent in cases compared to controls for the two case-control datasets and that same allele was overtransmitted in the trio families (table 4). Allele frequencies are population specific (varies among controls from different populations) thus association studies of IRF5 are sensitive for population stratification; association signal would be lost when comparing Nordic cases to Spanish controls for CGGGG for instance (table 5).
Table 5: Allele frequencies in the three datasets in paper I. Frequencies in bold indicate that the frequencies differ significantly (P<0.05) between cases and controls within each dataset. + indicates overtransmission of the risk allele, significantly only where in bold. a) P=0.06.

<table>
<thead>
<tr>
<th>Marker</th>
<th>Location</th>
<th>Risk allele</th>
<th>Spanish cases</th>
<th>Spanish controls</th>
<th>Nordic cases</th>
<th>Nordic controls</th>
<th>Finnish trios</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs729302</td>
<td>5'</td>
<td>A</td>
<td>0.70</td>
<td>0.68</td>
<td>0.67</td>
<td>0.66</td>
<td>+</td>
</tr>
<tr>
<td>rs4728142</td>
<td>5'</td>
<td>A</td>
<td>0.52</td>
<td>0.46</td>
<td>0.48</td>
<td>0.44</td>
<td>+</td>
</tr>
<tr>
<td>rs3757385</td>
<td>5'</td>
<td>G</td>
<td>0.70</td>
<td>0.65</td>
<td>0.67</td>
<td>0.66</td>
<td>+</td>
</tr>
<tr>
<td>CGGGG</td>
<td>5'</td>
<td>in</td>
<td>0.53</td>
<td>0.48</td>
<td>0.49</td>
<td>0.45</td>
<td>+ a</td>
</tr>
<tr>
<td>rs2004640</td>
<td>intron 1</td>
<td>T</td>
<td>0.59</td>
<td>0.53</td>
<td>0.55</td>
<td>0.52</td>
<td>+</td>
</tr>
<tr>
<td>rs3807306</td>
<td>intron 1</td>
<td>T</td>
<td>0.57</td>
<td>0.52</td>
<td>0.54</td>
<td>0.51</td>
<td>+</td>
</tr>
<tr>
<td>rs10954213</td>
<td>exon 10</td>
<td>A</td>
<td>0.69</td>
<td>0.65</td>
<td>0.67</td>
<td>0.64</td>
<td>+</td>
</tr>
<tr>
<td>rs11770589</td>
<td>exon 10</td>
<td>A</td>
<td>0.57</td>
<td>0.55</td>
<td>0.51</td>
<td>0.50</td>
<td>+</td>
</tr>
<tr>
<td>rs2280714</td>
<td>3'</td>
<td>T</td>
<td>0.73</td>
<td>0.68</td>
<td>0.71</td>
<td>0.70</td>
<td>+</td>
</tr>
<tr>
<td>rs12539741</td>
<td>3'</td>
<td>T</td>
<td>0.11</td>
<td>0.09</td>
<td>0.15</td>
<td>0.14</td>
<td>+</td>
</tr>
</tbody>
</table>

When combining the P-values using the analytical expression of Lou Jost, an extension of Fisher’s method, the most strongly associated markers were rs4728142 (P=2x10^{-4}), rs3806307 (P=2x10^{-4}) and CGGGG (P=5x10^{-4}). In addition rs3757385 (P=0.04), rs2004640 (P=0.003) and rs2280714 (P=0.02) had P-values below 0.05. It should be noted that the highest P-value was no higher than 0.16 (for rs11770589) and this could be taken as an indication that the association signal affect all tested markers.

Haplotype analysis didn’t reveal any stronger signal compared to single point analysis.

In order to verify that associations found were true rather than false we investigated the three most significant associations in addition individuals of Nordic origin in paper VI. Association was maintained for CGGGG (P=2x10^{-4}) while a ten-fold decrease in P-values was noted for the rs4728142 (P=3x10^{-5}) and rs3807306 (P=4x10^{-5}) with the addition of 2,171 cases and 1,731 Scandinavian controls when combining P-values from the Spanish dataset (paper V), Finnish dataset (paper V) and the expanded Scandinavian dataset (paper VI).

In the MS genomewide association study performed by IMSGC in 2007 P-values for the T allele of rs3807306 were P=0.01 for the dataset consisting of 931 trio families and P=0.14 for the dataset consisting of the 931 cases from the trio families and 2431 unrelated controls. Since the two datasets in the genomewide screen are not independent results from only one at a time can be combined with the P-values for the other datasets (www.loujoust.com); combining the IMSGC case-control, Spanish case-control, expanded Scandinavian case-control and Finnish trios datasets yields a combined P-value of 4x10^{-6}. When combining the IMSGC trios, Spanish case-control, expanded Scandinavian case-control and Finnish trios datasets the combined P-value was 5x10^{-6}.
Having established the genetic association of \textit{IRF5} with MS we next examined expression levels of \textit{IRF5} in MS patients compared to patients with other neurological diseases both non-inflammatory (OND) and inflammatory (OND.INF), in PBMC as well as CSF. There are at least 12 described isoforms of \textit{IRF5}\textsuperscript{189}, subcategorized into three sets depending on which exon1 that is transcribed. Primers were selected as to detect all but one isoform (variant 7) and relative expression levels were detected by quantitative real-time PCR.

Comparison of MS patients to OND with the non-parametric Wilcoxon rank-sum test yielded significant P-values in CSF (P=0.01) and PBMC (P=0.002); we first believed this to reflect a difference in expression levels between the two groups. Visual inspection of boxplots of the data revealed that while medians didn’t seem to vary, the shape of the distributions did. In order to deduce that medians differ between groups analysed by a non-parametric statistical test assumptions need to be made regarding equal shape of distributions and equal variances. While we were able to validate the equal variance assumption by log-transforming the data we were not able to transform the groups as to get equal distribution shapes presumably due to a limited material. Transformation of data does however affect the results and it is not sure that a non-significant finding of transformed data would imply that IRF5 expression levels didn’t vary between groups. On the other hand, we did not detect any correlation between genotype at the three markers and expression levels which would indicate that the mechanism by which genotype at these markers affects probability of disease is not through a shift in expression levels. That seems also, partly, to be the case for the rs2004640 SNP in SLE where increase in expression levels by itself is not associated with a higher SLE risk\textsuperscript{190}.

In a separate study, Brynedal et al (unpublished) found expression of \textit{IRF5} to be 3-fold down regulated compared to OND in CSF. In paper VI we made a network neighbourhood analysis to investigate which other molecules, known to interact with IRF5 that displayed differential expression in the data provided by Brynedal et al. The aim was to generate hypotheses of mechanisms of action of \textit{IRF5} in MS. The network neighbourhood analysis pointed, among others, at molecules involved in the type I interferon system and immune modulation (table 6). The genes that encode those molecules are candidates for interaction studies with \textit{IRF5} in MS compared to controls.
<table>
<thead>
<tr>
<th>Gene</th>
<th>Name</th>
<th>Description</th>
<th>Up or Down-regulated in MS compared to OND (CSF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCL3</td>
<td>chemokine (C-C motif) ligand 3</td>
<td>Cytokine that is involved in the acute inflammatory state in the recruitment and activation of mononuclear, polymorphonuclear leukocytes and NK cells(^{191}).</td>
<td>down</td>
</tr>
<tr>
<td>CCL4</td>
<td>chemokine (C-C motif) ligand 4</td>
<td>Cytokine to which mononuclear, polymorphonuclear leukocytes and NK cells respond. (^{191}).</td>
<td>down</td>
</tr>
<tr>
<td>CCL5</td>
<td>chemokine (C-C motif) ligand 5</td>
<td>The cytokine encoded by this gene functions as a chemoattractant for blood monocytes, memory T helper cells and eosinophils. It causes the release of histamine from basophils and activates eosinophils. This cytokine is one of the major HIV-suppressive factors produced by CD8+ cells.</td>
<td>up</td>
</tr>
<tr>
<td>IRF1</td>
<td>Interferon regulatory factor 1</td>
<td>IRF1 serves as an activator of interferons alpha and beta transcription. IRF1 also functions as a transcription activator of genes induced by interferons alpha, beta, and gamma. Further, IRF1 has been shown to play roles in regulating apoptosis and tumor-suppression.</td>
<td>up</td>
</tr>
<tr>
<td>IRF3</td>
<td>Interferon regulatory factor 3</td>
<td>IRF3 is found in an inactive cytoplasmic form that upon serine/threonine phosphorylation forms a complex with CREBBP. This complex translocates to the nucleus and activates the transcription of interferons alpha and beta, as well as other interferon-induced genes.</td>
<td>up</td>
</tr>
<tr>
<td>SP110</td>
<td>SP110 nuclear body protein</td>
<td>The protein can function as an activator of gene transcription and may serve as a nuclear hormone receptor co-activator. In addition, it has been suggested that the protein may play a role in ribosome biogenesis and in the induction of myeloid cell differentiation.</td>
<td>up</td>
</tr>
<tr>
<td>IFIT2</td>
<td>interferon-induced protein with tetratricopeptide repeats 2</td>
<td>Interferon induced gene</td>
<td>down</td>
</tr>
<tr>
<td>NAMPT</td>
<td>nicotinamide phosphoribosyltransferase</td>
<td>The protein is an adipokine that is localized to the bloodstream and has various functions, including the promotion of vascular smooth muscle cell maturation and inhibition of neutrophil apoptosis. It also activates insulin receptor and has insulin-mimetic effects, lowering blood glucose and improving insulin sensitivity.</td>
<td>down</td>
</tr>
<tr>
<td>PLSCR1</td>
<td>phospholipid scramblase 1</td>
<td>Interferon induced gene, amplifies the IFN response through increased expression of antiviral genes(^{192}).</td>
<td>down</td>
</tr>
<tr>
<td>Gene</td>
<td>Name</td>
<td>Description</td>
<td>Up or Down-regulated in MS compared to OND (CSF)</td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>-------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>TMPO</td>
<td>thymopoietin</td>
<td>Encodes a protein of the nuclear envelope implicated in gene silencing&lt;sup&gt;193&lt;/sup&gt;.</td>
<td>up</td>
</tr>
<tr>
<td>TNFSF10</td>
<td>tumor necrosis factor (ligand) superfamily, member 10</td>
<td>A cytokine that belongs to the tumor necrosis factor (TNF) ligand family. This protein preferentially induces apoptosis in transformed and tumor cells, but does not appear to kill normal cells although it is expressed at a significant level in most normal tissues. The binding of this protein to its receptors has been shown to trigger the activation of MAPK8/JNK, caspase 8, and caspase 3.</td>
<td>up</td>
</tr>
<tr>
<td>CASP3</td>
<td>caspase 3, apoptosis-related cysteine peptidase</td>
<td>This gene encodes a protein which is a member of the cysteine-aspartic acid protease (caspase) family. Sequential activation of caspases plays a central role in the execution-phase of cell apoptosis. It is the predominant caspase involved in the cleavage of amyloid-beta 4A precursor protein, which is associated with neuronal death in Alzheimer’s disease.</td>
<td>up</td>
</tr>
<tr>
<td>TRAF6</td>
<td>TNF receptor-associated factor 6</td>
<td>This protein mediates the signaling not only from the members of the TNF receptor superfamily, but also from the members of the Toll/IL-1 family.</td>
<td>up</td>
</tr>
<tr>
<td>CDKN1A</td>
<td>cyclin-dependent kinase inhibitor 1A (p21, Cip1)</td>
<td>The encoded protein binds to and inhibits the activity of cyclin-CDK2 or -CDK4 complexes, and thus functions as a regulator of cell cycle progression at G1. This protein can interact with proliferating cell nuclear antigen (PCNA), a DNA polymerase accessory factor, and plays a regulatory role in S phase DNA replication and DNA damage repair. This protein was reported to be specifically cleaved by CASP3-like caspases, which thus leads to a dramatic activation of CDK2, and may be instrumental in the execution of apoptosis following caspase activation.</td>
<td>down</td>
</tr>
<tr>
<td>OAS1</td>
<td>2’5’-oligoadenylate synthetase 1, 40/46kDa</td>
<td>Essential proteins involved in the innate immune response to viral infection. The encoded protein is induced by interferons and uses adenosine triphosphate in 2’-specific nucleotidyl transfer reactions to synthesize 2’5’-oligoadenylates (2-5As). Mutations in this gene have been associated with host susceptibility to viral infection.</td>
<td>up</td>
</tr>
</tbody>
</table>
Table 6: Description of function of the genes differentially expressed in MS patients compared to OND in Brynedal et al (unpublished)\textsuperscript{19}. Whenever no reference is given, description is gathered from the gene summary in Entrez Gene (www.ncbi.nlm.nih.gov).

In conclusion, IRF5 is associated with MS and further studies need to be undertaken to clarify the mechanisms of action of this transcription factor in MS.
6 Concluding remarks and future perspective

Gene discovery case-control association approaches in the field of complex genetics is in transition from medium scale candidate gene studies, such as those within this thesis, to genomewide association studies involving several thousands of patients and controls. The process has been gradual as studies with sample sizes of 500 cases and 500 controls was considered large at the start of this project to the realization that effect sizes are smaller than first thought and thus thousands of patients and controls are needed in order to detect association.

Georg Klein once said that a biologist doesn’t just have to learn to accept complexity he also needs to learn to appreciate complexity. There is surely beauty in complexity but to my mind the beauty doesn’t show itself fully until we are able to understand it. In this thesis I have worked on a population-based sample of patients, I believe it to be a group constituted by subgroups that each share a genetic composition that predisposes to MS. In the quest for that genetic composition I’d be better off studying each subgroup separately instead of getting confounded and obscured by the mixture.

However, it is not clear on which basis the subgroups should be separated, interesting research is ongoing to that end\textsuperscript{195,196} and the potential of studying sub-phenotypes such as expression levels might also be further exploited in future MS research\textsuperscript{197,198}. In the meantime, rather than being paralyzed by uncertainty we are trying to retrieve the information we can from the complex picture – in that process we are likely to find, not what is unique for each subgroup, but rather common factors among several groups – that is, in itself, a compelling strategy.

The common genetic variant that is searched for by this strategy can be thought of as component causes, i.e. pieces of pies (see section 2); what the full sufficient causes (pies) are composed of has not been unravelled yet. With ever increasing sample sizes we might be able to address this issue by investigating biologic interactions between associated factors, complementing efforts to answer the same question by investigating less heterogeneous groups\textsuperscript{199}. In that respect, clues might be retrieved by investigating patients for differential gene expression (affected only expression profiling) as have been done for healthy populations\textsuperscript{198}.

A deeper understanding of the patterns of LD in the small scale is another advance in complex genetics that has a great impact on design and interpretation of genetic association studies. We now know that it is not possible to test markers evenly spaced throughout a gene and expect to be able to exclude the gene from the list of candidates given negative results for all tested markers. Rather, effective design of studies requires every region to be examined, in an informative population, for patterns of linkage disequilibrium so that markers are selected in a way that maximizes coverage; the interpretation of results is equally dependent on patterns of LD. Whole genome or regional
sequencing is starting to emerge and the launch of the 1000 genomes project (www.1000genomes.org) brings promise of exclusion or inclusion of genes, rather than markers, as susceptibility candidates. Who knows, we might find out that the gene is not the functional disease-causing unit. In parallel, I hope and expect that the large genomewide association studies, just recently made possible through collaborations between centers in different countries will be possible to conduct in a single population – enabling us to discover population-specific factors. Such a study would possibly include the majority of patients in a country not allowing for an initial finding to be replicated, I would argue that the probability of a false positive report would be minimal in such a large study allowing for confirmation of findings to be undertaken with other methods, e.g. functional studies.

In the last couple of years robust associations have been found between genetic variants and MS, still I would like to go back to Andrew Clark's words from 2004, a period of deep pessimism in the field of complex genetics:

"In the end, it is clear that nature could be perverse and present us with patterns of genetic variation for complex chronic diseases that will completely evade the approaches that we are bringing to bear on the problem. On the other hand, it is highly unlikely that all undiscovered contributions to complex disorders are so recalcitrant, and we can find solace in the early successes in finding the easier, ApoE-like genes first."

While there still remains plenty of work until the MS-associated genetic variants have been linked to biological actions leading to disease, the recent advances in the field of complex genetics reassures me that nature in fact is not perverse and that, if required, we will be able to redirect our research efforts to assure new progress – we need to, for the sake of all present and future patients.
7 Acknowledgements

The years in the “MS genetics group” have given me the pleasure of coming in contact with warm, bright and truly inspiring persons. I’ve been given the opportunity to interact with researchers in various groups, to get acquainted with many scientific fields including genetics, epidemiology, statistics, immunology and evolutionary biology. I’ve had the privilege of working in one environment where we are encouraged to question…everything! I would like to express my sincere gratitude to all of you, friends, colleagues and family who have made this thesis possible. In particular, I would like to thank:

Jan Hillert, my supervisor, for providing the perfect environment so that I could complete this project. For your enormous generosity when you let me take a year of absence and for welcoming me back when I wanted to return. For letting me have my own opinion and for providing me with the best education ever!

Ingrid Kockum, my co-supervisor, for sharing your great knowledge of human genetics with me, for helping me with everything from data file handling to understanding theoretical concepts, for being a friend as much as a co-supervisor.

Ewa Waldebäck, for always being there when I needed help never asking for anything in return, for taking care of my children as if they were your own, for all conversations about everything in life and… for letting me know that Jan was looking for new students 😊

Boel Brynedal, for all the moments we have shared during our years at KI, for your constant support and conversations about science and all other things that matter. Thank you for reading my thesis and for your insightful comments.

Kerstin Imrell, I would never have learned as much as I have if there hasn’t been for you. The way you address science is truly inspiring – there are no “field-boundaries”, just questions that you really want to answer. Thank you for all discussions scientific and non-scientific.

Jenny Link, for wanting to learn everything and by that being a true source of inspiration, for your sense of humor and willingness to share information and for excellent comments on my thesis.

Kristina Duvefelt, for always being there for me when I needed advice, for always finding time to help me even when there was no time. Thank you for your comments on my thesis.

Frida Lundmark (for all discussions about LD and haplotypes and for all other interesting conversations), Wangko Lundström (for spreading joy around you), Helena Modin (for your patience when I was a newbie, I still think you should get back to Academia!), Thomas Masterman (for active listening always and for all knowledge you possess and share), Eva Lindström (for many interesting discussions and for the saying det vill bli bra – a very useful one!).

All "plan 6"-colleagues: Marjan Jahanpanah (for alla djupa och grunda samtal och för fantastisk mat), Annelie Porsborn, Eva Johansson, Virginija Karrenbauer (thanks for helping me to prepare for the “MS-examination” and for your optimism), Eva Greiner, Sverker Johansson, Kristina Gottberg, Christina Sjöstrand.

All colleagues “downstairs”: Malin Lundkvist, Rasmus Gustafsson, Anna Fogdell-Hahn, Cecilia Svarén-Quiding, Anna Mattson, Elin Karlberg, Ingegerd Löfving Arvholm, Merja Kanerva (for helping me with all kinds of issues, always with a smile on your face), Faezeh Vejdani, Leszek Stawiarz (for your enthusiasm regarding everything from rare languages to MRI), Anna Hillert, Gunnel Larsson, Yvonne Sjölind, Jenny Ahlqvist,
Marita Ingemarsson, Anna Nilzén (for efficient and professional help when patient samples needed to be handled), Anny Rydberg, Helena Ytterberg, Anna Aronsson, Madeleine Berg, Lise-Lotte Bengtsson, Helen Hallin, Kosta Kostulas, Sebastian Yakisch (special thanks for translating ethical permits), Lena von Kock, Lotta Widén Holmqvist, Ajith Sominanda, Mathula Thangarajh, Andreia Gomes, Marina Vita, Susanna Mjörnheim, Uros Rot, Yassir Hussein, Vilmantas Gierdatrius.

Tomas Olsson for your enthusiasm for everything that regards science; Mohsen Khademi for always taking your time to help me with practical as well as theoretical “expression” issues; Emelie Sundqvist, Johan Öckinger, Nada Abdelmagid, Ritha Nohra and Magda Lindén for helping me out in the lab and for nice chats.

Magnus Lekman for contributing to interesting discussions and helping me to prepare for the dissertation.

Gudrun J. Bergman for all discussions when we had our statistics seminars and for your ability to talk statistics with biologists!


I would like to thank all patients and blood donors for their contribution to this work and the personnel at the Mutation Analysis Facility for providing a great proportion of genotypes for this thesis.

All members of the Nordic MS genetics group.

Robert Hallin for helping me with travel-arrangements.

My SU-friends Anna Tjämlund, Ariane Rodriguez, Monika Hansson and Karin Lindroth for all moments within and outside of “Immunologen”.

My dear friend Gunilla Niss Jonsson, for your enthusiasm for life, for all laughs over a cup of tea (and sockerkaka of course) and for introducing me to the world of students, classrooms and eternal grading of exams, I really enjoyed it!

“Gradenarna”: Gretel, Urban, Martin, Thomasine, Kerstin, Abbe, Åsa, and Nicolas – for letting me in to your family.

All members of the de Andrade Lima and Bomfim families, especially my grandmothers vovó Maria and Manhula as well as tia Inez, for always making me feel that I belong even though we are so far apart.

Mãe Dulce, obrigada por todo amor e carinho, pelo seu bom humor, tão contagioso, por toda ajuda durante as minhas visitas e pelas memórias fantásticas que os meus filhos têm do Brasil.

My dear sister Julieta, my aunt Tereza and Mats, my brother Vollmer and Helena, my cousin Laura and Robert for always being there for me, in good times and less good times, for your condition-less support. In addition, I would like to thank Julieta, Tereza and Vollmer for planning the party!

Meus queridos pais, Maria Izaura e Vollmer Bomfim, obrigada por todo amor e apoio, por terem me ensinado que é preferível ser e saber do que ter e por acreditar em mim.

Ted Gradén – för din humor och generositet, för alla glada stunder och för ditt intresse över hur det gick med ”skrivandet” – det är så lätt att tycka om dig!

Kristina och Erik Roos – tack för att ni varit generösa och stöttat mig och för att ni är de ni är, ni inspirerar mig, jag ålskar er enormt och bär er ständig i mitt hjärta!

Lars Gradén – thank you for believing in me, for letting me be who I am, for introducing me to the world of Kosters and Violas, I love you for being you!
8 References

13. Wellcome Trust Case Control Consortium W. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007 Jun 7;447(7145):661-78.


Hillert J, Olerup O. Multiple sclerosis is associated with genes within or close to the HLA-DR-DQ subregion on a normal DR15,DQ6,Dw2 haplotype. Neurology. 1993 Jan;43(1):163-8.


Butte MJ, Haines C, Bonilla FA, Puck J. IL-7 receptor deficient SCID with a unique intronic mutation and post-transplant autoimmunity due to chronic GVHD. Clinical Immunology. 2007;125(2):159-64.


Jianya Huan NCLSRBGGBYKCDBSFZHOAAV. Decreased FOXP3 levels in multiple sclerosis patients. 2005:45-52.


Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. 1988:1215-.


120 Purcell S, Cherry SS, Sham PC. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. 2003:149-50.


122 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. 2007:559-75.


Mamidipudi V, Wootten MW. Dual role for p75(NTR) signaling in survival and cell death: can intracellular mediators provide an explanation? J Neurosci Res. 2002 May 15;68(4):373-84.


Roux PP, Bhakar AL, Kennedy TE, Barker PA. The p75 Neurotrophin Receptor Activates Akt (Protein Kinase B) through a Phosphatidylinositol 3-Kinase-dependent Pathway. 2001;23097-104.


Sergey V. Kozyrev SLPMVLBP-EACGTWGCGPJHLCGASM. Structural insertion/deletion variation in IRF5 is associated with a risk haplotype and defines the precise IRF5 isoforms expressed in systemic lupus erythematosus. 2007:1234-41.


